ME280A

Introduction to the Finite Element Method

Panayiotis Papadopoulos

Department of Mechanical Engineering
University of California, Berkeley

2015 edition

Contents

List of Figures

Introduction

This is a set of notes written as part of teaching ME280A, a first-year graduate course on the finite element method, in the Department of Mechanical Engineering at the University of California, Berkeley.

Berkeley, California
August 2015

P.P.

Chapter 1

Introduction to the Finite Element Method

1.1 Historical perspective: the origins of the finite element method

The finite element method constitutes a general tool for the numerical solution of partial differential equations in engineering and applied science. Historically, all major practical advances of the method have taken place since the early 1950s in conjunction with the development of digital computers. However, interest in approximate solutions of field equations dates as far back in time as the development of the classical field theories (e.g. elasticity, electro-magnetism) themselves. The work of Lord Rayleigh[1] (1870) and W. Ritz[2] (1909) on variational methods and the weighted-residual approach taken by B.G. Galerkin[3] (1915) and others form the theoretical framework to the finite element method. With a bit of a stretch, one may even claim that K. Schellbach's approximate solution to Plateau's problem (find a surface of minimum area enclosed by a given closed curve in three dimensions) by triangulation, which dates back to 1851, is a

Figure 1.1. B.G. Galerkin

[1] John William Strutt, 3rd Baron Rayleigh (1842–1919) was a British physicist.
[2] Walther Ritz (1878–1909) was a Swiss theoretical physicist.
[3] Boris Grigoryevich Galerkin (1871-1945) was a Russian mathematician and mechanician.

rudimentary application of the finite element method.

Figure 1.2. R. Courant

Most researchers agree that the era of the finite element method begins with a lecture presented in 1941 by R. Courant[4] to the American Association for the Advancement of Science. In his work, Courant used the Ritz method and introduced the pivotal concept of spatial discretization for the solution of the classical torsion problem. Courant did not pursue his idea further, since computers were still largely unavailable for his research.

More than a decade later, R.W. Clough[5] and his colleagues at Berkeley essentially reinvented the finite element method as a natural extension of matrix structural analysis and published their first work in 1956. Clough had spent the summers of 1952 and 1953 at Boeing working on modeling of the vibration in a wing structure and it is this work that he led to his formulation of finite elements for plate structures. Clough is also credited with coining the term "finite element" in a 1960 paper[6] on the approximate solution of two-dimensional problems in elasticity. An apparently simultaneous effort by J. Argyris[7] at the University of London independently led to another successful introduction of the method. It should come as no surprise that, to a

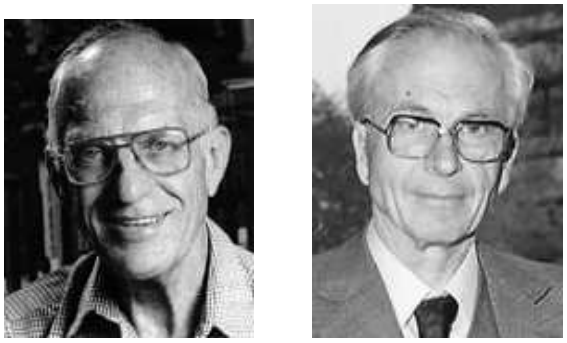


Figure 1.3. R.W. Clough (left) and J. Argyris (right)

large extent, the finite element method appears to owe its reinvention to structural engineers.

[4]Richard Courant (1888-1972) was a German-born American mathematician.

[5]Ray W. Clough, Jr. (1920-) is an American structural engineer.

[6]R.W. Clough. The finite element method in plane stress analysis. In *Proceedings of the 2nd ASCE Conference on Electronic Computation*, Pittsburg, PA, (1960).

[7]John (Hadji)Argyris (1913-2004) was a Greek civil engineer.

In fact, the consideration of a complicated system as an assemblage of simple components (elements) on which the method relies is very natural in the analysis of structural systems.

A few years after its introduction to the engineering community, the finite element method gained the attention of applied mathematicians, particularly those interested in numerical solution of partial differential equations. In 1973, W.G. Strang[8] and G.J. Fix authored the first conclusive treatise on mathematical aspects of the method, focusing exclusively on its application to the solution of problems emanating from standard variational theorems.

The finite element has been subject to intense research, both at the mathematical and technical level, and thousands of scientific articles and hundreds of books about it have been authored. By the beginning of the 1990s, the method clearly dominated the numerical solution of problems in the fields of structural analysis, structural mechanics and solid mechanics. Moreover, the finite element method currently competes in popularity with the finite difference method in the areas of heat transfer and fluid mechanics.

1.2 Introductory remarks on the concept of discretization

The basic goal of discretization is to provide an approximation of an infinite dimensional system by a system that can be fully defined with a finite number of "degrees of freedom". To clarify the notion of dimensionality, consider a deformable body in the three-dimensional

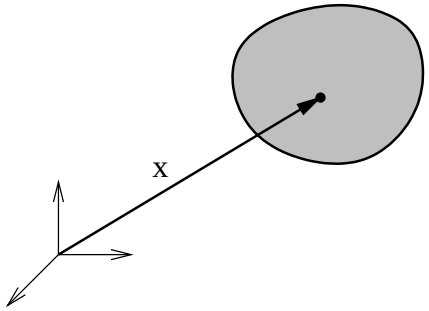


Figure 1.4. *An infinite degree-of-freedom system*

Euclidean space, for which the position of a typical particle with reference to a fixed coordinate system is defined by means of a vector $\mathbf{x}$, as in Figure 1.4. This is an infinite dimensional system with respect to the position of all of its particle points. If the same body

[8] W. Gilbert Strang (1934-) is an American applied mathematician.

is assumed to be rigid, then it is a finite dimensional system with only six degrees of freedom. A dimensional reduction of the above system is accomplished by placing a (somewhat severe) restriction on the admissible motions that the body may undergo.

Finite dimensional approximations are very important from the computational standpoint, because they often allow for analytical and/or numerical solutions to problems that would otherwise be intractable. There exist various methods that can reduce infinite dimensional systems to approximate finite dimensional counterparts. Here, we consider three of those methods, namely the physically motivated structural analogue substitution method, the finite difference method and the finite element method, and also address, in passing, various particle-based methods.

1.2.1 Structural analogue substitution method

Consider the oscillation of a liquid in a manometer. This system can be approximated ("lumped") by means of a single degree-of-freedom mass-spring system, as in Figure 1.5. Clearly, such an approximation is largely intuitive and cannot precisely capture the complexity of the original system (viscosity of the liquid, surface tension effects, geometry of the manometer walls).

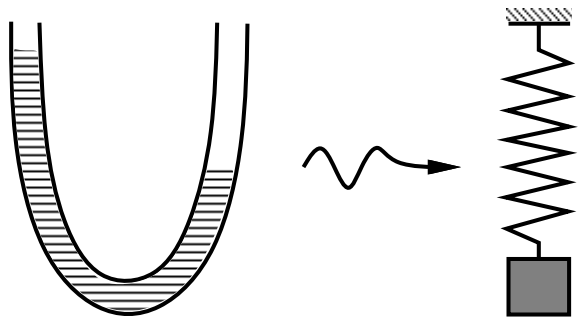


Figure 1.5. *A simple example of the structural analogue method*

The structural analogue substitution method, whenever applicable, generally provides coarse approximations to complex systems. However, its degree of sophistication (hence, also the fidelity of its results) can vary widely. The "network analysis" of G. Kron in the 1930s and 1940s for solving differential equations is generally viewed as a typical example of the structural analogue approach.

1.2.2 Finite difference method

Consider the ordinary differential equation

$$\begin{aligned} k\frac{d^2u}{dx^2} &= f \text{ in } (0,L) \ , \\ u(0) &= u_0 \ , \\ u(L) &= u_L \ , \end{aligned} \tag{1.1}$$

where k is a constant and $f = f(x)$ is a smooth function. Let N points be chosen in the interior of the domain $(0, L)$, each of them equidistant from its immediate neighbors. An algebraic (or "difference") approximation to the second derivative may be computed as

$$\left.\frac{d^2u}{dx^2}\right|_l \doteq \frac{u_{l+1} - 2u_l + u_{l-1}}{\Delta x^2} \ , \tag{1.2}$$

with error $o(\Delta x^2)$. Indeed, assuming that the solution $u(x)$ is at least four times continuously

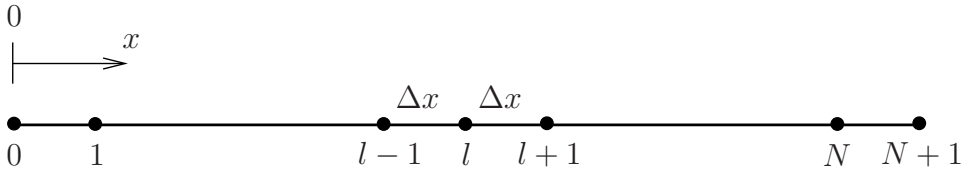


Figure 1.6. *The finite difference method in one dimension*

differentiable and employing twice a Taylor series expansion with remainder around a typical point l in Figure 1.6, write

$$u_{l+1} = u_l + \Delta x \left.\frac{du}{dx}\right|_l + \frac{\Delta x^2}{2!}\left.\frac{d^2u}{dx^2}\right|_l + \frac{\Delta x^3}{3!}\left.\frac{d^3u}{dx^3}\right|_l + \frac{\Delta x^4}{4!}\left.\frac{d^4u}{dx^4}\right|_{l+\theta_1} \ ; \quad 0 \le \theta_1 \le 1 \ , \tag{1.3}$$

$$u_{l-1} = u_l - \Delta x \left.\frac{du}{dx}\right|_l + \frac{\Delta x^2}{2!}\left.\frac{d^2u}{dx^2}\right|_l - \frac{\Delta x^3}{3!}\left.\frac{d^3u}{dx^3}\right|_l + \frac{\Delta x^4}{4!}\left.\frac{d^4u}{dx^4}\right|_{l-\theta_2} \ ; \quad 0 \le \theta_2 \le 1 \ . \tag{1.4}$$

Adding the above equations results in

$$\left.\frac{d^2u}{dx^2}\right|_l = \frac{u_{l+1} - 2u_l + u_{l-1}}{\Delta x^2} - \frac{\Delta x^2}{4!}\left(\left.\frac{d^4u}{dx^4}\right|_{l+\theta_1} + \left.\frac{d^4u}{dx^4}\right|_{l-\theta_2}\right) \ , \tag{1.5}$$

so that ignoring the second term of the right-hand side, the proposed approximation to the second derivative of u is recovered. This approximation becomes increasingly accurate as Δx

approaches zero, that is, as the spacing of the points in the domain $(0, L)$ becomes denser. Applying the difference equation (1.2) to nodal points $1, 2, ..., N$, and accounting for the boundary conditions $(1.1)_{2,3}$ gives rise to a system of N linear algebraic equations

$$\begin{aligned} u_2 - 2u_1 &= \frac{f_1 \Delta x^2}{k} - u_0 \ , \\ u_{l+1} - 2u_l + u_{l-1} &= \frac{f_l \Delta x^2}{k} \quad , \qquad l = 2, ..., N-1 \ , \\ -2u_N + u_{N-1} &= \frac{f_N \Delta x^2}{k} - u_L \ , \end{aligned} \tag{1.6}$$

with unknowns u_l, $l = 1, 2, \ldots, N$. Again, an infinite-dimensional problem with respect to the value of u in the domain $(0, L)$ is transformed by the above method into an N-dimensional problem.

Clearly, the state equations are (approximately) satisfied only at discrete points $1, 2, ..., N$. Also, the boundary conditions are enforced directly when writing the discrete counterparts of the state equations in the nodes that reside next to the boundaries. It is easy to see that finite difference methods run into difficulties when dealing with complex boundaries due to the need for spatial regularity of the grid.

1.2.3 Finite element method

Consider again the problem defined in the previous section and employ the same domain discretization as in Figure 1.3. Here, however, assume that the u varies linearly in each line segment between successive points, and also that it is continuous throughout the domain $(0, L)$. Now, concentrate on a typical line segment between points l and $l + 1$. This is now the domain of the finite element e. In this domain, assume that u varies linearly, as shown in Figure 1.7, and attains values u_l at point l and u_{l+1} at point $l + 1$.

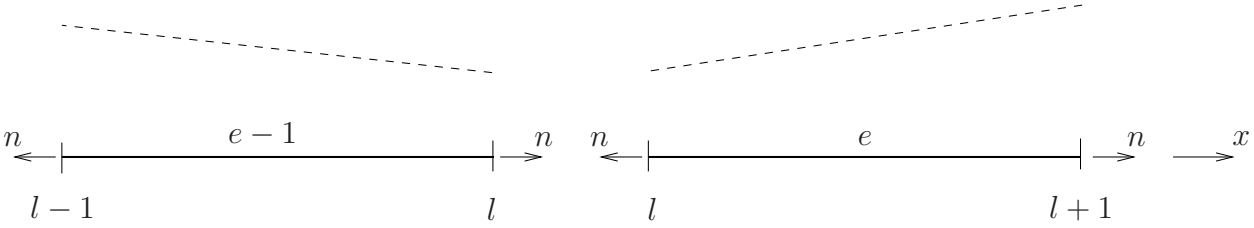


Figure 1.7. *A one-dimensional finite element approximation*

The normal flux $q = -k\dfrac{du}{dn}$, where n denotes the outward unit normal to the element

domain is equal to

$$q_l^e \;=\; \left(-k\frac{du}{dn}\right)_l^e \;=\; -k\frac{u_l - u_{l+1}}{\Delta x} \tag{1.7}$$

and

$$q_{l+1}^e \;=\; \left(-k\frac{du}{dn}\right)_{l+1}^e \;=\; -k\frac{u_{l+1} - u_l}{\Delta x} \tag{1.8}$$

at points l and $l+1$, respectively. These two equations can be written in matrix form as

$$-\frac{k}{\Delta x}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} u_l \\ u_{l+1} \end{bmatrix} \;=\; \begin{bmatrix} q_l^e \\ q_{l+1}^e \end{bmatrix} . \tag{1.9}$$

An analogous matrix equation can be written for element $e-1$, whose domain lies between points $l-1$ and l, and takes the form

$$-\frac{k}{\Delta x}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} u_{l-1} \\ u_l \end{bmatrix} \;=\; \begin{bmatrix} q_{l-1}^{e-1} \\ q_l^{e-1} \end{bmatrix} . \tag{1.10}$$

Now, adding the first equation of (1.9) to the second equation of (1.10) yields

$$\frac{k}{\Delta x}(u_{l+1} - 2u_l + u_{l-1}) \;=\; q_l^e + q_l^{e-1} . \tag{1.11}$$

To approximate the right-hand side of (1.11), first note that the terms q_l^e and q_l^{e-1} represent fluxes on opposite sides of the point l. It follows, upon recalling (1.7) and the relation between the coordinate x and the outward normal n, that

$$q_l^e + q_l^{e-1} \;=\; \left(-k\frac{du}{dn}\right)_l^e + \left(-k\frac{du}{dn}\right)_l^{e-1} \;=\; \left(k\frac{du}{dx}\right)_l^e - \left(k\frac{du}{dx}\right)_l^{e-1} . \tag{1.12}$$

At the same time, one may rewrite the differential equation as $k\dfrac{du}{dx} = \displaystyle\int f\,dx$. Hence, if the total force $f_{total} = \displaystyle\int_0^L f\,dx$ is somehow distributed to the points $0, 1, \ldots, N+1$ so that to point l corresponds a force $\tilde{f}_l$, then the jump in the normal derivative $k\dfrac{du}{dx}$ at l is exactly $\tilde{f}_l$, therefore (1.11) attains the form

$$\frac{k}{\Delta x}(u_{l+1} - 2u_l + u_{l-1}) \;=\; \tilde{f}_l . \tag{1.13}$$

Then, the complete finite element system becomes

$$\begin{aligned} u_2 - 2u_1 \;&=\; \frac{\tilde{f}_1\Delta x}{k} - u_0 , \\ u_{l+1} - 2u_l + u_{l-1} \;&=\; \frac{\tilde{f}_l\Delta x}{k} \qquad , \qquad l = 2, ..., N-1 , \\ -2u_N + u_{N-1} \;&=\; \frac{\tilde{f}_N\Delta x}{k} - u_L , \end{aligned} \tag{1.14}$$

This is the so-called *direct approach* to formulating the finite element equations. Upon comparing (1.6) and (1.14), it is concluded that the two sets of equations are identical to within the definition of the force term. Yet, these equations were derived by way of fundamentally different approximations.

It will be established that in finite element methods the state equations are satisfied in an integral sense over the whole domain with respect to a set of (simple) admissible functions. Also, it will be seen that boundary conditions can be handled trivially.

1.2.4 Particle methods

These are known by many different names, including Smooth Particle Hydrodynamics (SPH), Element-Free Galerkin (EFG), finite point, and hp-cloud methods. The principal idea behind all of these methods is that an interpolation of the dependent variable is constructed using values associated with a given set of points ("particles") in the domain. However, unlike the finite element method, a smooth interpolation function (called a kernel function) is defined for each particle without explicit dependence on the placement of neighboring particles, see Figure 1.8.

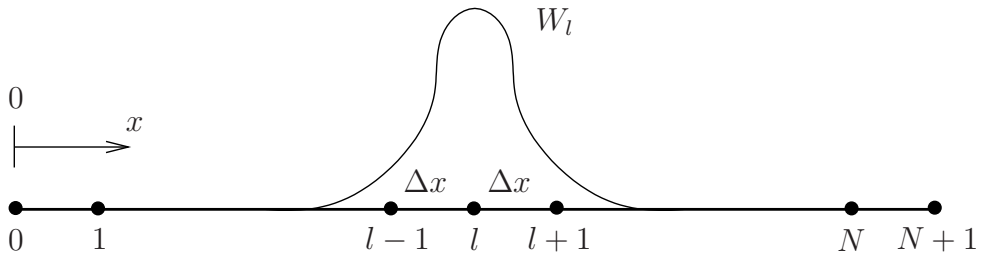


Figure 1.8. *A one-dimensional kernel function W_l associated with a particle method*

The major advantage of particle methods is that they do not require, in principle, a finite element mesh (hence, are often referred to as meshless methods). Challenges with these methods are associated with the enforcement of boundary conditions, the evaluation of integrals associated with the solution of the differential equations, and the selection of suitable kernel functions.

1.3 Classifications of partial differential equations

Consider a scalar partial differential equation (PDE) of the general form

$$F(x, y, ...u, u_{,x}, u_{,y}, ...u_{,xx}, u_{,xy}, u_{,yy}, ...) \;=\; 0 \;, \tag{1.15}$$

where x, y, ... are the *independent variables* (often representing space or time), $u = u(x, y, \ldots)$ is the *dependent* variable, and, also,

$$u_{,x} \;=\; \frac{\partial u}{\partial x} \quad , \quad u_{,xx} \;=\; \frac{\partial^2 u}{\partial x^2} \quad , \quad u_{,xy} \;=\; \frac{\partial^2 u}{\partial x \partial y} \quad , \quad \text{etc. .} \tag{1.16}$$

The *order* of the PDE is defined as the order of the highest derivative of u in (1.15). Also, a PDE is *linear* if the function F is linear in u and in all of its derivatives, with coefficients possibly depending on the independent variables $x, y, \ldots$.

Example 1.3.1: Partial differential equations of different types

(a) $3u_{,x} + u_{,y} - u = 0$ (linear, first order) ,

(b) $u_{,x} + uu_{,y} = 1$ (non-linear, first order) ,

(c) $xu_{,xx} + \frac{1}{y}u_{,yy} - 3u = 0$ (linear, second order) ,

(d) $u_{,xx}^2 + u_{,yy} = 0$ (non-linear, second order) ,

(e) $u_{,x}\, u_{,xxx}^2 + u_{,yy} = 0$ (non-linear, third order) . ◀

For the purpose of the forthcoming developments, consider linear second-order partial differential equations of the general form

$$a u_{,xx} \;+\; b u_{,xy} \;+\; c u_{,yy} \;=\; d \;, \tag{1.17}$$

where not all a, b, c are equal to zero. In addition, let a, b, c be functions of x, y only, whereas d can be a function of $x, y, u, u_{,x}, u_{,y}$.

Equations of the form (1.17) can be categorized as follows:

(a) Elliptic equations ($b^2 - 4ac < 0$)

A typical example of an elliptic equation is the two-dimensional version of the Laplace (Poisson) equation used in modeling various physical phenomena (e.g., heat conduction, electro-statics, torsion of bars), namely

$$u_{,xx} + u_{,yy} \;=\; f \quad ; \quad f \;=\; f(x, y) \;,$$

for which $a = c = 1$ and $b = 0$.

(b) Parabolic equations ($b^2 - 4ac = 0$)

The equation of transient linear heat conduction in one dimension,

$$k u_{,xx} \;=\; u_{,t} \qquad ; \qquad k \;=\; k(x) \;,$$

where $a = k$ and $b = c = 0$, is a representative example of a parabolic equation.

(c) Hyperbolic equations ($b^2 - 4ac > 0$)

The one-dimensional linear wave equation,

$$\alpha^2 u_{,xx} \;-\; u_{,tt} \;=\; 0 \qquad ; \qquad \alpha \;=\; \alpha(x) \;,$$

where $a = \alpha^2$, $b = 0$ and $c = -1$, falls in this class of equations.

Extension of the above classification to more general types of partial differential equations than those of the form (1.17) is not always an easy task. The elliptic, hyperbolic or parabolic nature of a partial differential equation is associated with the particular form of its characteristic curves. These are curves along a partial differential equation becomes ordinary, and its solution can be (theoretically) determined by normal integration.

The type of a partial differential equation determines the overall character of the expected solution. Broadly speaking, elliptic differential equations exhibit solutions which are as smooth as its coefficients allow. On the other hand, the solutions to parabolic differential equations tend to smooth out any initial discontinuities, while the solutions to hyperbolic partial differential equations preserve any initial discontinuities. To a great extent, the type of the partial equation dictates the choice of methodology used in its numerical approximation by the finite element method.

Remarks:

☛ Partial differential equations of mixed type are possible, such as the classical one-dimensional *convection-diffusion* equation of the form

$$u_{,t} + \alpha u_{,x} = \epsilon u_{,xx} \qquad ; \qquad \alpha \geq 0 \quad , \quad \epsilon \geq 0 \; .$$

The above equation is of hyperbolic type if $\epsilon = 0$ and $\alpha > 0$ (that is, when the diffusive term is suppressed), since

$$\begin{aligned} \alpha^2 u_{,xx} &= \alpha(\alpha u_{,x})_{,x} = \alpha(-u_{,t})_{,x} \\ &= \alpha(-u_{,x})_{,t} = -(\alpha u_{,x})_{,t} \\ &= -(-u_{,t})_{,t} = u_{,tt} \end{aligned}$$

implies that it is merely a first-order counterpart of the previously mentioned wave equation. On the other hand, for $\epsilon > 0$ and $\alpha = 0$ the convective part vanishes and the equation is purely parabolic and coincides with the previously mentioned one-dimensional transient heat conduction equation. The dominant character in the convection-diffusion equation is controlled by the relative values of parameters α and ϵ.

☛ The type of a partial differential equation may be spatially dependent, as with the following example:

$$u_{,xx} + x u_{,yy} = 0 \; ,$$

where $a = 1$, $b = 0$ and $c = x$, so that the equation is elliptic for $x > 0$, parabolic for $x = 0$ and hyperbolic for $x < 0$.

1.4 Suggestions for further reading

Section 1.1

[1] C.A. Felippa. An appreciation of R. Courant's 'Variational methods for the solution of problems of equilibrium and vibrations', 1943. *Int. J. Num. Meth. Engr.*, 37:2159–2187, 1994. [*This reference contains the original article on the finite element method by Courant, preceded by an interesting introduction by C. Felippa.*]

[2] R.W. Clough, Jr. The finite element method after twenty-five years: A personal view. *Comp. Struct*, 12:361–370, 1980. [*This reference offers a unique view of the finite element method by one of its inventors*].

[3] P.G. Ciarlet and J.L. Lions, editors. *Finite Element Methods (Part 1)*, volume II of *Handbook of Numerical Analysis*. North-Holland, Amsterdam, 1991. [*The first article in this handbook presents a comprehensive introduction to the history of the finite element method, authored by J.T. Oden*].

Section 1.2

[1] O.C. Zienkiewicz and R.L. Taylor. *The Finite Element Method; Basic Formulation and Linear Problems*, volume 1. McGraw-Hill, London, 4th edition, 1989. [*Chapter 1 of this book is devoted to an introductory discussion of discretization*].

[2] G. Kron. Numerical solutions of ordinary and partial differential equations by means of equivalent circuits. *J. Appl. Phys.*, 16:172–186, 1945. [*This is an interesting use of an electrical circuits analogue method to obtain approximate solutions of differential equations*].

Section 1.3

[1] F. John. *Partial Differential Equations*. Springer-Verlag, New York, 4th edition, 1985. [*Chapter 2 contains a mathematical discussion of the classification of linear second-order partial differential equations in connection with their characteristics*].

Chapter 2

Mathematical Preliminaries

2.1 Sets, linear function spaces, operators and functionals

A *set* $\mathcal{U}$ is a collection of objects, referred to as *elements* or *points*. If u is an element of the set $\mathcal{U}$, one writes $u \in \mathcal{U}$. If not, one writes $u \notin \mathcal{U}$. Let $\mathcal{U}$, $\mathcal{V}$ be two sets. The set $\mathcal{U}$ is a *subset* of the set $\mathcal{V}$ (denoted as $\mathcal{U} \subseteq \mathcal{V}$ or $\mathcal{V} \supseteq \mathcal{U}$) if every element of $\mathcal{U}$ is also an element of $\mathcal{V}$. The set $\mathcal{U}$ is a *proper subset* of the set $\mathcal{V}$ (denoted as $\mathcal{U} \subset \mathcal{V}$ or $\mathcal{V} \supset \mathcal{U}$) if every element of $\mathcal{U}$ is also an element of $\mathcal{V}$, but there exists at least one element of $\mathcal{V}$ that does not belong to $\mathcal{U}$.

Some sets of particular interest in the remainder of these notes are $\mathbb{Z} = \{\text{all integer numbers}\}$ and $\mathbb{R} = \{\text{all real numbers}\}$, as well as $\mathbb{N} = \{\text{all positive integer numbers}\}$ and $\mathbb{R}_0^+ = \{\text{all non-negative real numbers}\}$.

The *union* of sets $\mathcal{U}$ and $\mathcal{V}$ (denoted by $\mathcal{U} \cup \mathcal{V}$) is the set which is comprised of all elements of both sets. The *intersection* of sets $\mathcal{U}$ and $\mathcal{V}$ (denoted by $\mathcal{U} \cap \mathcal{V}$) is a set which includes only the elements common to the two sets. The *empty set* (denoted by $\emptyset$) is a set that contains no elements and is contained in every set, therefore, $\mathcal{U} \cup \emptyset = \mathcal{U}$. The (set-theoretic) *difference* of a set $\mathcal{V}$ from another set $\mathcal{U}$, denoted $\mathcal{U} \setminus \mathcal{V}$, consists of all elements in $\mathcal{U}$ which do not belong to $\mathcal{V}$. The *Cartesian product* $\mathcal{U} \times \mathcal{V}$ of sets $\mathcal{U}$ and $\mathcal{V}$ is a set defined as

$$\mathcal{U} \times \mathcal{V} \;=\; \{(u, v) \text{ such that } u \in \mathcal{U},\; v \in \mathcal{V}\} \;. \tag{2.1}$$

Note that the pair (u, v) in the preceding equation is ordered, that is, the element (u, v) is, in general, not the same as the element (v, u). The notation $\mathcal{U}^2$, $\mathcal{U}^3$, ..., is used to respectively

denote the Cartesian products $\mathcal{U} \times \mathcal{U}$, $\mathcal{U} \times \mathcal{U} \times \mathcal{U}$, ...

Consider a set $\mathcal{V}$ whose members can be scalars, vectors or functions, as visualized in Figure 2.1. Assume that $\mathcal{V}$ is endowed with an addition operation $(+)$ and a scalar multiplication operation $(\cdot)$, which do not necessarily coincide with the classical addition and multiplication for real numbers.

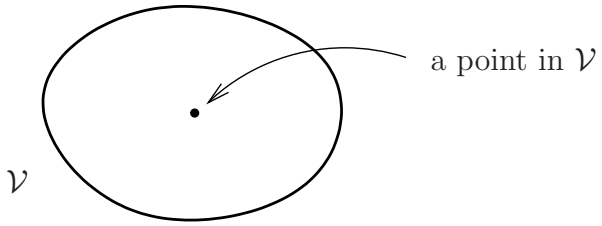


Figure 2.1. *Schematic depiction of a set $\mathcal{V}$*

A *linear* (or *vector*) *space* $\{\mathcal{V}, +; \mathbb{R}, \cdot\}$ is defined by the following properties for any $u, v, w \in \mathcal{V}$ and $\alpha, \beta \in \mathbb{R}$:

(i) $\alpha \cdot u + \beta \cdot v \in \mathcal{V}$ (closure),

(ii) $(u + v) + w = u + (v + w)$ (associativity with respect to $+$),

(iii) $\exists\, 0 \in \mathcal{V} \quad | \quad u + 0 = u$ (existence of null element),

(iv) $\exists\ -u \in \mathcal{V} \quad | \quad u + (-u) = 0$ (existence of negative element),

(v) $u + v = v + u$ (commutativity),

(vi) $(\alpha\beta) \cdot u = \alpha \cdot (\beta \cdot u)$ (associativity with respect to $\cdot$),

(vii) $(\alpha + \beta) \cdot u = \alpha \cdot u + \beta \cdot u$ (distributivity with respect to $\mathbb{R}$),

(viii) $\alpha \cdot (u + v) = \alpha \cdot u + \alpha \cdot v$ (distributivity with respect to $\mathcal{V}$),

(ix) $1 \cdot u = u$ (existence of identity).

Example 2.1.1: Linearity of spaces

(a) $\mathcal{V} = \mathbb{P}_2 = \{$all second degree polynomials $ax^2 + bx + c\}$ with the standard polynomial addition and scalar multiplication.

It can be trivially verified that $\{\mathbb{P}_2, +; \mathbb{R}, \cdot\}$ is a linear function space. $\mathbb{P}_2$ is also "equivalent" to an *ordered* triad $(a, b, c) \in \mathbb{R}^3$.

(b) $\mathcal{V} = M_{m,n}(\mathbb{R})$, where $M_{m,n}(\mathbb{R})$ is the set of all $m \times n$ matrices whose elements are real numbers. This set is a linear space with the usual matrix addition and scalar multiplication operations.

(c) Define $\mathcal{V} = \left\{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\right\}$ with the standard addition and scalar multiplication for vectors. Note that given $u = (x_1, y_1)$ and $v = (x_2, y_2)$ as in Figure 2.2, property (i) is violated,

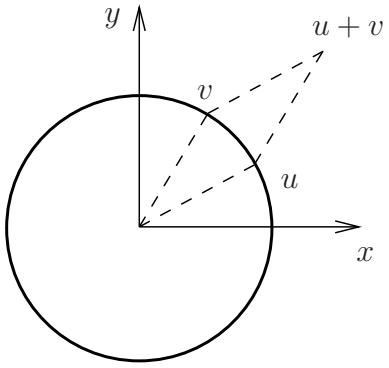


Figure 2.2. *Example of a set that does not form a linear space*

since, in general, for $\alpha = \beta = 1$

$$u + v = (x_1 + x_2 \ , \ y_1 + y_2) \ ,$$

and $(x_1 + x_2)^2 + (y_1 + y_2)^2 \neq 1$. Thus, $\{\mathcal{V}, +; \mathbb{R}, \cdot\}$ is not a linear space. ◀

Consider a linear space $\{\mathcal{V}, +; \mathbb{R}, \cdot\}$ and a subset $\mathcal{U}$ of $\mathcal{V}$. Then $\mathcal{U}$ forms a *linear subspace* of $\mathcal{V}$ with respect to the same operations $(+)$ and $(\cdot)$, if, for any $u, v \in \mathcal{U}$ and $\alpha, \beta, \in \mathbb{R}$

$$\alpha \cdot u \ + \ \beta \cdot v \ \in \ \mathcal{U} \ ,$$

that is, closure is maintained within $\mathcal{U}$.

Example 2.1.2: Subspace of a linear space
Define the set $\mathbb{P}_n$ of all algebraic polynomials of degree smaller or equal to $n > 2$ and consider the linear space $\{\mathbb{P}_n, +; \mathbb{R}, \cdot\}$ with the usual polynomial addition and scalar multiplication. Then, $\mathbb{P}_2$ is a linear subspace of $\{\mathbb{P}_n, +; \mathbb{R}, \cdot\}$. ◀

Let $\mathcal{U}$, $\mathcal{V}$ be two sets and define a *mapping* f from $\mathcal{U}$ to $\mathcal{V}$ as a rule that assigns to each point $u \in \mathcal{U}$ a unique point $f(u) \in \mathcal{V}$, see Figure 2.3. The usual notation for a mapping is:

$$f \ : \ u \in \mathcal{U} \rightarrow f(u) \in \mathcal{V} \ .$$

With reference to the above setting, $\mathcal{U}$ is called the *domain* of f, whereas $\mathcal{V}$ is termed the *range* of f.

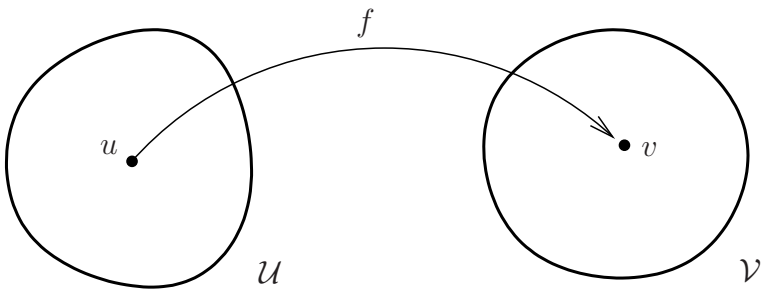


Figure 2.3. *Mapping between two sets*

The above definitions are general in that they apply to completely general types of sets $\mathcal{U}$ and $\mathcal{V}$. By convention, the following special classes of mappings are identified here:

(1) *function*: a mapping from a set with scalar or vector points to scalars or vectors, namely,

$$f \ : \ x \in \mathcal{U} \to f(x) \in \mathbb{R}^m \quad ; \quad \mathcal{U} = \mathbb{R}^n \quad , \quad n, m \in \mathbb{N} \dots \ ,$$

(2) *functional*: a mapping from a set of functions to the real numbers, namely,

$$I \ : \ u \in \mathcal{U} \to I[u] \in \mathcal{V} \subset \mathbb{R} \quad ; \quad \mathcal{U} \text{ a function space} \ .$$

(3) *operator*: a mapping from a set of functions to another set of functions, namely,

$$A \ : \ u \in \mathcal{U} \to A[u] \in \mathcal{V} \quad ; \quad \mathcal{U}, \mathcal{V} \text{ function spaces} \ .$$

The preceding distinction between functions, functionals and operators is largely arbitrary: all of the above mappings can be classified as operators by viewing $\mathbb{R}^n$ as a simple function space. However, the distinction will be observed for didactic purposes.

Example 2.1.3: Functions, functionals and operators

(a) $f(\mathbf{x}) = \sqrt{x_1^2 + x_2^2}$ is a function, where $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$.

(b) $I[u] = \displaystyle\int_0^1 u(x)dx$ is a functional, where u belongs to a function space, say $u(x) \in \mathbb{P}_n$.

(c) $A[u] = \dfrac{d}{dx}u(x)$ is a (differential) operator where $u(x) \in \mathcal{U}$, where $\mathcal{U}$ is a function space. ◀

Given a linear space $\mathcal{U}$, an operator $A : \mathcal{U} \to \mathcal{V}$ is called *linear*, provided that, for all $u_1, u_2 \in \mathcal{U}$ and $\alpha, \beta \in \mathbb{R}$,

$$A[\alpha \cdot u_1 + \beta \cdot u_2] = \alpha \cdot A[u_1] + \beta \cdot A[u_2] \ .$$

Otherwise, the operator is termed *non-linear*.

Example 2.1.4: Linear and non-linear operators

(a) $A[u] = \frac{d}{dx}u(x)$ is a linear differential operator.

(b) $A[u] = u^2(x)$ is a non-linear algebraic operator. ◀

Linear partial differential equations can be formally obtained as mappings of an appropriate function space to another, induced by the action of linear differential operators. For example, consider a linear second-order partial differential equation of the form

$$au_{,xx} + bu_{,x} = c \ ,$$

where a,b and c are functions of x and y only. The operational form of the above equation is

$$A[u] = c \ ,$$

where the linear differential operator A is defined as

$$A[\,\cdot\,] = a(\,\cdot\,)_{,xx} + b(\,\cdot\,)_{,x}$$

over a space of functions $u(x)$ that possess second derivatives in the domain of analysis.

2.2 Continuity and differentiability

Consider a real function $f : \mathcal{U} \to \mathbb{R}$, where $\mathcal{U} \subset \mathbb{R}$. The function f is *continuous at a point* $x = x_0$ if, given any scalar $\epsilon > 0$, there exists a scalar $\delta(\epsilon)$, such that

$$|f(x) - f(x_0)| < \epsilon \ , \tag{2.2}$$

provided that

$$|x - x_0| < \delta \ . \tag{2.3}$$

The function f is called *continuous*, if it is continuous at all points of its domain. A function f is of class $C^k(\mathcal{U})$ (k integer ≥ 0) if it is k-times continuously differentiable (that is, it possesses derivatives to k-th order and they are continuous functions).

Example 2.2.1: Functions of different classes C^k

(a) The function $f:\ (0,2) \to \mathbb{R}$ defined as

$$f(x) \ = \ \begin{cases} x & \text{if } 0 < x < 1 \\ 2 - x & \text{if } 1 \leq x < 2 \end{cases}$$

is of class $C^0(\mathcal{U})$, but not of $C^1(\mathcal{U})$, see Figure 2.4.

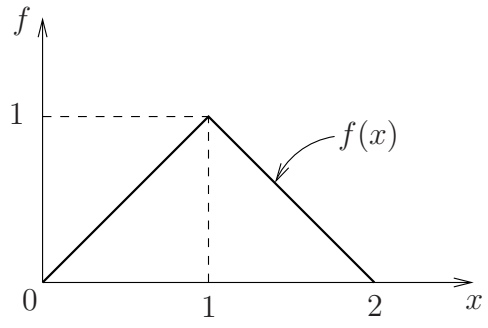


Figure 2.4. *A function of class* $C^0(0,2)$

(b) Any polynomial function $P(x):\ \mathcal{U} \to \mathbb{R}$ is of class $C^\infty(\mathcal{U})$. ◀

The above definition can be easily generalized to certain subsets of $\mathbb{R}^n$: a function $f\ : \mathbb{R}^n \mapsto \mathbb{R}$ is of class $C^k(\mathcal{U})$ if all of its partial derivatives up to k-th order are continuous. Further generalizations to operators will be discussed later.

The "smoothness" (that is, the degree of continuity) of functions plays a significant role in the proper construction of finite element approximations.

2.3 Norms, inner products, and completeness

2.3.1 Norms

By way of background, recall the classical definition of distance (in the Euclidean sense) between two points in $\mathbb{R}^2$: Given any two points $\mathbf{x}_1 = (x_1, y_1)$ and $\mathbf{x}_2 = (x_2, y_2)$ as in Figure 2.5, define the "distance" function $d:\ \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}_0^+$ as

$$d(\mathbf{x}_1, \mathbf{x}_2) \ = \ \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}\ . \tag{2.4}$$

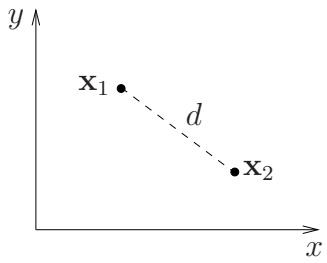


Figure 2.5. *Distance between two points in the classical Euclidean sense*

It is important in the analysis of finite element methods to extend the notion of proximity ("closeness") from points in a Euclidean space to functions. Moreover, we will need to be able to quantify the size ("large" vs. "small") of a function. The appropriate context for these requirements is provided by norms.

Consider a linear space $\{\mathcal{V}, + \,;\, \mathbb{R}, \cdot\}$ and define a mapping $\|\cdot\| : \mathcal{V} \to \mathbb{R}$ such that, for all $u, v \in \mathcal{V}$ and $\alpha \in \mathbb{R}$, the following properties hold:

(i) $\|u + v\| \;\leq\; \|u\| + \|v\|$ (triangle inequality),

(ii) $\|\alpha \cdot u\| \;=\; |a|\|u\|$,

(iii) $\|u\| \;\geq\; 0$ and $\|u\| \;=\; 0 \;\Leftrightarrow\; u \;=\; 0$.

A mapping with the above properties is called a *norm* on $\mathcal{V}$. A linear space $\{\mathcal{V}, +; \mathbb{R}, \cdot\}$ endowed with a norm is called a *normed linear space* (NLS).

Example 2.3.1: Some useful norms

(a) Consider the n-dimensional Euclidean space $\mathbb{R}^n$ and let $\mathbf{x} = [x_1 \;\; x_2 \;\; \ldots \;\; x_n]^T \in \mathbb{R}^n$. Some standard norms in $\mathbb{R}^n$ are defined as follows:

- the 1-norm: $\|\mathbf{x}\|_1 \;=\; \sum_{i=1}^{n} |x_i|$,
- the 2-norm: $\|\mathbf{x}\|_2 \;=\; \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}$,
- the ∞-norm: $\|\mathbf{x}\|_\infty \;=\; \max_{1 \leq i \leq n} |x_i|$.

(b) The L_2-norm of a square-integrable function $u \in \mathcal{U}$ with domain Ω is defined as

$$\|u\|_2 \;=\; \left(\int_\Omega u^2 \, d\Omega\right)^{1/2} .$$

◀

Using any norm on a function space $\mathcal{U}$, we can quantify convergence of a sequence of functions $\{u_n\}$ to u in $\mathcal{U}$ by referring to the *distance function* d between u_n and u, defined as

$$d(u_n, u) \;=\; \|u_n - u\| \;. \tag{2.5}$$

In particular, we say that $u_n \to u \in \mathcal{U}$ if, for any $\epsilon > 0$, there exists an $N(\epsilon)$, such that

$$d(u_n, u) < \epsilon \;, \tag{2.6}$$

for all $n > N$.

Typically, the limit of a convergent sequence $\{u_n\}$ of functions in $\mathcal{U}$ is not known in advance. Indeed, consider the case of a series of approximate function solutions to a partial differential equation having an unknown (and, possibly, unavailable in closed form) exact solution u. A sequence $\{u_n\}$ is called *Cauchy convergent* if, for any $\epsilon > 0$, there exists an $N(\epsilon)$, such that

$$d(u_m, u_n) \;=\; \|u_m - u_n\| \;<\; \epsilon \;, \tag{2.7}$$

for all $m, n > N$. Although it will not be proved here, it is easy to verify that convergence of a sequence $\{u_n\}$ implies Cauchy convergence, but the opposite is not necessary true.

Given any point u in a normed linear space $\{\mathcal{V}, +; \mathbb{R}, \cdot\}$, one may identify the neighborhood $\mathcal{N}_r(u)$ of u with radius $r > 0$ as the set of points v for which

$$d(u, v) \;<\; r \;, \tag{2.8}$$

or, in mathematical notation $\mathcal{N}_r(u) = \{v \in \mathcal{V} \mid d(u, v) < r\}$, see also Figure 2.6. Then, a

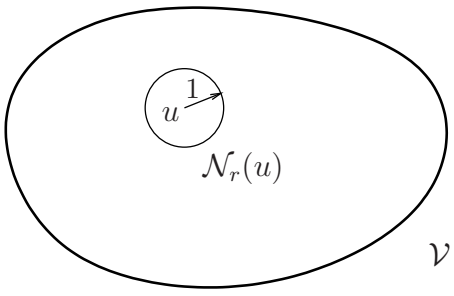


Figure 2.6. The neighborhood $\mathcal{N}_r(u)$ of a point u in $\mathcal{V}$

◀

subset $\mathcal{U}$ of $\mathcal{V}$ is termed *open* if, for each point $u \in \mathcal{U}$, there exists a neighborhood $\mathcal{N}_r(u)$ which is fully contained in $\mathcal{U}$. The complement $\mathcal{U}^c$ of an open set $\mathcal{U}$ (defined as the set of all

points in $\mathcal{V}$ that do not belong to $\mathcal{U}$) is, by definition, a *closed* set. The *closure* of a set $\mathcal{U}$, denoted $\overline{\mathcal{U}}$, is defined as the smallest closed set that contains $\mathcal{U}$.

Example 2.3.2: Open and closed sets in $\mathbb{R}$
Consider the set of real numbers $\mathbb{R}$ equipped with the usual norm (namely, the absolute value).

(a) The set $\mathcal{U} = \{x \in \mathbb{R} \quad | \quad 0 < x < 1\} = (0, 1)$ is open.

(b) The set $\mathcal{V} = \{x \in \mathbb{R} \quad | \quad 0 \leq x \leq 1\} = [0, 1]$ is closed.

(c) The set $\mathcal{W} = \{x \in \mathbb{R} \quad | \quad 0 \leq x < 1\} = [0, 1)$ is neither open nor closed.

(d) The set $\mathbb{R}$ is both open and closed. ◀

2.3.2 Inner products

In addition to notions of size and proximity, we will need to quantify relative orientation (including orthogonality) of functions, just as we do for vectors. To this end, consider a linear space $\{\mathcal{V}, + \ ; \ \mathbb{R}, \cdot\}$ and define a mapping $\langle \cdot \ , \ \cdot \rangle \ : \ \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, such that for all u, v and $w \in \mathcal{V}$ and $\alpha \in \mathbb{R}$, the following properties hold:

(i) $\langle u, v \rangle = \langle v, u \rangle$ (commutativity) ,

(ii) $\langle u + v, w \rangle = \langle u, w \rangle + \langle v, w \rangle$ (distributivity with respect to $+$) ,

(iii) $\langle \alpha \cdot u, v \rangle = \alpha \langle u, v \rangle$ (associativity with respect to $\cdot$) ,

(iv) $\langle u, u \rangle \geq 0$ and $\langle u, u \rangle = 0 \Leftrightarrow u = 0$.

A mapping with the above properties is called an *inner product* on $\mathcal{V} \times \mathcal{V}$. A linear space $\{\mathcal{V}, +; \mathbb{R}, \cdot\}$ endowed with an inner product is called an *inner product space*. If two elements u, v of $\mathcal{V}$ satisfy the condition $\langle u, v \rangle = 0$, then they are *orthogonal* relative to the inner product $\langle \cdot \ , \ \cdot \rangle$.

Example 2.3.3: Inner product spaces

(a) Set $\mathcal{V} = \mathbb{R}^n$ and for any vectors $\mathbf{x} = [x_1 \ x_2 \ \ldots \ x_n]^T$ and $\mathbf{y} = [y_1 \ y_2 \ \ldots \ y_n]^T$ in $\mathcal{V}$, define the mapping

$$\langle \mathbf{x}, \mathbf{y} \rangle \ = \ \mathbf{x}^T \mathbf{y} = \sum_{i=1}^{n} x_i y_i \ .$$

This is the conventional dot-product between vectors in $\mathbb{R}^n$. It is easy to show that the above mapping is an inner product on $\mathcal{V} \times \mathcal{V}$. This inner product-space in called the n-dimensional Euclidean vector space.

(b) The L_2-inner product for functions $u, v \in \mathcal{U}$ with domain Ω is defined as

$$\langle u, v \rangle \;=\; \int_\Omega uv \, d\Omega \; .$$

◀

An inner product on $\mathcal{V} \times \mathcal{V}$ induces an associated norm (called the *natural norm*) on $\mathcal{V}$, defined as

$$\|u\| \;=\; \langle u, u \rangle^{1/2} \; . \tag{2.9}$$

To prove that the function $\langle u, u \rangle^{1/2}$ is indeed a norm, it is sufficient to show that it satisfies the three defining properties of a norm stated in Section 2.3.1. Properties (ii) and (iii) are easily verified using the fact that $\langle \cdot, \cdot \rangle$ is an inner product. Indeed, for (ii)

$$\|\alpha \cdot u\| \;=\; \langle \alpha \cdot u, \alpha \cdot u \rangle^{1/2} \;=\; \left(\alpha^2 \langle u, u \rangle\right)^{1/2} \;=\; |\alpha| \|u\| \; , \tag{2.10}$$

and for (iii)

$$\|u\| \;=\; \langle u, u \rangle^{1/2} \geq 0 \quad , \quad \|u\| \;=\; \langle u, u \rangle^{1/2} \;=\; 0 \Leftrightarrow u \;=\; 0 \; . \tag{2.11}$$

To establish that property (i) (namely, the triangle inequality) holds, we make use of the *Cauchy-Schwartz inequality*, which states that for any $u, v \in \mathcal{V}$

$$|\langle u, v \rangle| \;\leq\; \|u\| \|v\| \; . \tag{2.12}$$

To prove (2.12), first note that it holds trivially as an equality if $u = 0$ or $v = 0$. Next, define a function $F : \mathbb{R} \to \mathbb{R}_0^+$ as

$$F(\lambda) \;=\; \|u + \lambda \cdot v\|^2 \; , \tag{2.13}$$

where u, v are arbitrary (although fixed) non-zero points of $\mathcal{V}$ and λ is a scalar. Making use of the definition of the natural norm and the inner product properties, we have

$$\begin{aligned} F(\lambda) = \langle u + \lambda \cdot v, u + \lambda \cdot v \rangle \;&=\; \langle u, u \rangle \;+\; 2\lambda \langle u, v \rangle \;+\; \lambda^2 \langle v, v \rangle \\ &= \|u\|^2 \;+\; 2\lambda \langle u, v \rangle \;+\; \lambda^2 \|v\|^2 \; . \end{aligned} \tag{2.14}$$

Noting that $F(\lambda) = 0$ has at most one real non-zero root (that is, if and when $u + \lambda \cdot v = 0$), it follows that, since

$$\|v\|^2 \lambda \;=\; -\langle u, v \rangle \pm \sqrt{\langle u, v \rangle^2 - \|u\|^2 \|v\|^2} \; , \tag{2.15}$$

the inequality

$$\langle u, v\rangle^2 \;-\; \|u\|^2\|v\|^2 \;\leq\; 0 \tag{2.16}$$

must hold, thus yielding (2.12).

Using the Cauchy-Schwartz inequality, return to property (i) of a norm and note that

$$\begin{aligned} \|u+v\|^2 \;=\; \langle u+v, u+v\rangle \;&= \langle u,u\rangle \;+\; 2\langle u,v\rangle \;+\; \langle v,v\rangle \\ &= \|u\|^2 \;+2\langle u,v\rangle \;+\; \|v\|^2 \\ &\leq \|u\|^2 \;+\; 2\|u\|\|v\| \;+\; \|v\|^2 \;=\; (\|u\|+\|v\|)^2 \;, \end{aligned} \tag{2.17}$$

which implies that the triangle inequality holds.

With the Cauchy-Schwartz inequality at hand, it is also possible to define a relative orientation angle θ two non-zero points $u, v \in \mathcal{V}$ according to

$$\theta \;=\; \arccos \frac{< u, v >}{\|u\|\|v\|} \;, \tag{2.18}$$

since (2.12) now guarantees that $|\cos\theta| \leq 1$.

2.3.3 Banach and Hilbert spaces

A linear space $\{\mathcal{U}, +; \mathbb{R}, \cdot\}$ for which every Cauchy sequence converges to some "point" in $\mathcal{U}$ is called a *complete* space. Complete normed linear spaces are also referred to as *Banach spaces*. Complete inner product spaces are called *Hilbert spaces*. Clearly, all Hilbert spaces are also Banach spaces (by way of the natural norm of the latter), while the opposite is generally not true. Hilbert spaces form the proper functional context for the mathematical analysis of finite element methods. The basic goal of such mathematical analysis is to establish conditions under which specific finite element approximations lead to a sequence of solutions that converge to the exact solution of the differential equation under investigation.

In the remainder of this section some of the commonly used finite element function spaces are introduced. First, define the L_2-space of functions with domain $\Omega \subset \mathbb{R}^n$ as

$$L_2(\Omega) \;=\; \left\{ u:\; \Omega \to \mathbb{R} \quad | \quad \int_\Omega u^2 d\Omega < \infty \right\} \;. \tag{2.19}$$

The above space contains all *square-integrable* functions defined on Ω.

Next, define the *Sobolev space* $H^m(\Omega)$ of order m (where m is a non-negative integer) as

$$H^m(\Omega) \;=\; \{u:\; \Omega \to \mathbb{R} \quad | \quad D^\alpha u \in L_2(\Omega) \;,\; \alpha \leq m\} \;, \tag{2.20}$$

where

$$D^\alpha u \;=\; \frac{\partial^\alpha u}{\partial x_1^{\alpha_1}\partial x_2^{\alpha_2}...\partial x_n^{\alpha_n}} \qquad , \qquad \alpha = \alpha_1 + \alpha_2 + \ldots + \alpha_n \;, \tag{2.21}$$

is the generic partial derivative of order α, and α is an non-negative integer. Using the above definitions, it is clear that $L_2(\Omega) = H^0(\Omega)$. An inner product is defined for $H^m(\Omega)$ as

$$\langle u, v\rangle_{H^m(\Omega)} \;=\; \int_\Omega \Big\{ \sum_{\alpha=0}^{m} \sum_{\beta=\alpha} D^\beta u \, D^\beta v \Big\} \, d\Omega \;, \tag{2.22}$$

with the corresponding natural norm

$$\|u\|_{H^m(\Omega)} \;=\; \langle u, u\rangle^{1/2}_{H^m(\Omega)} \;=\; \left(\int_\Omega \Big\{ \sum_{\alpha=0}^{m} \sum_{\beta=\alpha} \big(D^\beta u\big)^2 \Big\} \, d\Omega \right)^{1/2} \;=\; \left(\sum_{\alpha=0}^{m} \sum_{\beta=\alpha} \|D^\beta u\|^2_{L_2(\Omega)} \right)^{1/2} . \tag{2.23}$$

Example 2.3.4: Inner product and norm of H^1 in two-dimensions
Assume $\Omega \subset \mathbb{R}^2$ and $m = 1$. Then

$$\langle u, v\rangle_{H^1(\Omega)} = \int_\Omega \left(uv + \frac{\partial u}{\partial x_1}\frac{\partial v}{\partial x_1} + \frac{\partial u}{\partial x_2}\frac{\partial v}{\partial x_2} \right) dx_1 dx_2 \;,$$

and

$$\|u\|_{H^1(\Omega)} = \left[\int_\Omega \left\{ u^2 + \left(\frac{\partial u}{\partial x_1}\right)^2 + \left(\frac{\partial u}{\partial x_2}\right)^2 \right\} dx_1 dx_2 \right]^{1/2} .$$

Clearly, for the above inner product to make sense (or, equivalently, for u to belong to $H^1(\Omega)$), it is necessary that u and both of its first derivatives be square-integrable. ◀

Standard theorems from elementary calculus guarantee that continuous functions are always square-integrable in a domain where they remain bounded. Similarly, piecewise continuous functions are also square integrable, provided that they possess a "small" number of discontinuities. The Dirac-delta function(al), defined on $\mathbb{R}^n$ by the property

$$\int_\Omega \delta(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n) \, dx_1 dx_2 \ldots dx_n \;=\; f(0, 0, \ldots, 0) \;, \tag{2.24}$$

for any continuous function f on Ω, where Ω contains the origin $(0, 0, \ldots, 0)$, is the single example of a function(al) which is not square-integrable and may be encountered in finite element approximations.

Example 2.3.5: A piecewise linear function
Consider the continuous piecewise linear function $u(x)$ in Figure 2.7. Clearly, the function is square-integrable. Its derivative $\dfrac{du}{dx}$ is a Heaviside function, and is also square-integrable. However, the second derivative $\dfrac{d^2u}{dx^2}$, which is a Dirac-delta function is not square-integrable.

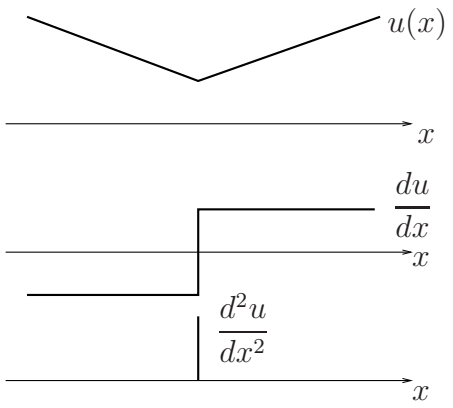


Figure 2.7. A continuous piecewise linear function and its derivatives

◀

Negative Sobolev spaces H^{-m} can also be defined and are of interest in the mathematical analysis of the finite element method. Bypassing the formal definition, one may simply note that a function u defined on Ω belongs to $H^{-1}(\Omega)$ if its anti-derivative belongs to $L_2(\Omega)$.

A formal connection between continuity and integrability of functions can be established by means of *Sobolev's lemma*. The simplest version of this theorem states that given an open set $\Omega \subset \mathbb{R}^n$ with sufficiently smooth boundary, and letting $C_b^k(\Omega)$ be the space of bounded functions of class $C^k(\Omega)$, then

$$H^m(\Omega) \;\subset\; C_b^k(\Omega)\;, \tag{2.25}$$

if, and only if, $m > k + n/2$. For example, setting $m = 2$, $k = 1$ and $n = 1$, one concludes from the preceding theorem that the space of H^2 functions on the real line is embedded in the space of bounded C^1-functions.

2.3.4 Linear operators and bilinear forms in Hilbert spaces

As already argued in Section 2.1, differential operators are a convenient vehicle for the analysis of differential equations. For this reason, we review here some important general properties of linear operators, which may be employed for linear differential equations.

Consider a linear operator $A : \mathcal{U} \mapsto \mathcal{V},\ v = A[u]$, where $\mathcal{U}$, $\mathcal{V}$ are Hilbert spaces, as in Figure 2.8. Some important definitions follow:

A linear operator A is *bounded* if there exists a constant $M > 0$, such that $\|A[u]\|_{\mathcal{V}} \leq M\|u\|_{\mathcal{U}}$, for all $u \in \mathcal{U}$. We say that M is a *bound* to the operator. Next, A is (uniformly) *continuous* if, for any $\epsilon > 0$, there is a $\delta = \delta(\epsilon)$ such that $\|A[u] - A[v]\|_{\mathcal{V}} < \epsilon$ for any $u, v \in \mathcal{U}$ that satisfy $\|u - v\|_{\mathcal{U}} < \delta$. It is easy to show that, in the context of linear operators,

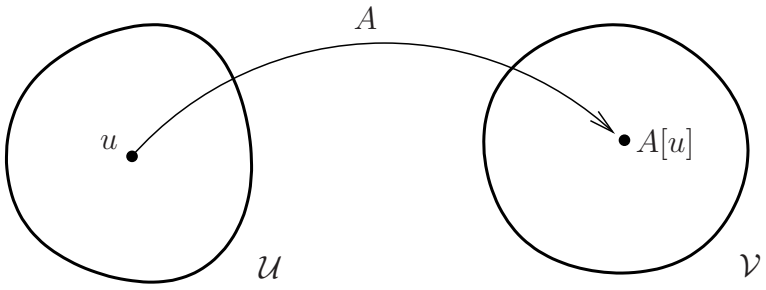


Figure 2.8. *A linear operator mapping $\mathcal{U}$ to $\mathcal{V}$*

boundedness implies (uniform) continuity and vice-versa (that is, the two properties are equivalent).

A linear operator $A : \mathcal{U} \mapsto \mathcal{V} \subset \mathcal{U}$ is *symmetric* relative to a given inner product $\langle \cdot, \cdot \rangle$ defined on $\mathcal{U} \times \mathcal{U}$, if

$$\langle A[u], v \rangle \;=\; \langle u, A[v] \rangle \;, \tag{2.26}$$

for all $u, v \in \mathcal{U}$. Notice that for the symmetry property to be considered, it is essential that the range of the operator A be a subset of its domain.

Example 2.3.6: A symmetric operator

Let $\mathcal{U} = \mathbb{R}^n$ and A be an operator identified with the action of an $n \times n$ symmetric matrix $\mathbf{A}$ on an n-dimensional vector $\mathbf{x}$, so that $A[\mathbf{x}] = \mathbf{A}\mathbf{x}$. Also, define an associated inner product as

$$\langle \mathbf{x}, A[\mathbf{y}] \rangle \;=\; \mathbf{x} \cdot \mathbf{A}\mathbf{y} \;,$$

that is, as the usual dot-product between vectors. Then

$$\begin{aligned} \langle \mathbf{x}, A[\mathbf{y}] \rangle \;&=\; \mathbf{x} \cdot \mathbf{A}\mathbf{y} = \mathbf{x} \cdot \mathbf{A}^T\mathbf{y} \\ &=\; (\mathbf{A}\mathbf{x}) \cdot \mathbf{y} \;=\; \langle A[\mathbf{x}], \mathbf{y} \rangle \end{aligned}$$

implies that A is a symmetric (algebraic) operator. ◀

A symmetric operator A is termed *positive* if $\langle A[u], u \rangle \geq 0$, for all $u \in \mathcal{U}$. The same operator is *strictly positive* if $\langle A[u], u \rangle > 0$, for all non-zero $u \in \mathcal{U}$.

Example 2.3.7: A positive operator

The symmetric algebraic operator of Example 2.3.6 is positive provided

$$\langle A[\mathbf{x}], \mathbf{x} \rangle \;=\; \mathbf{A}\mathbf{x} \cdot \mathbf{x} \;\geq\; 0 \;.$$

In this case, the $n \times n$ symmetric matrix $\mathbf{A}$ is termed *positive-semidefinite*. If the preceding inequality is strict for all non-zero vectors $\mathbf{x}$ in $\mathbb{R}^n$, then the matrix is *positive-definite*. ◀

The *adjoint* A^* of an operator A with reference to the inner product $\langle \cdot, \cdot \rangle$ on $\mathcal{U} \times \mathcal{U}$ is defined by

$$\langle A[u], v \rangle \;=\; \langle u, A^*[v] \rangle \;, \tag{2.27}$$

for all $u, v \in \mathcal{U}$. An operator A is termed *self-adjoint* if $A = A^*$. It is clear that every self-adjoint operator is symmetric, but the converse is not true.

Define $B : \mathcal{U} \times \mathcal{V} \mapsto \mathbb{R}$ as in Figure 2.9, where $\mathcal{U}$ and $\mathcal{V}$ are Hilbert spaces, such that for all $u, u_1, u_2 \in \mathcal{U}$, $v, v_1, v_2 \in \mathcal{V}$ and $\alpha, \beta \in \mathbb{R}$,

(i) $B(\alpha \cdot u_1 + \beta \cdot u_2, v) = \alpha B(u_1, v) + \beta B(u_2, v)$,

(ii) $B(u, \alpha \cdot v_1 + \beta \cdot v_2) = \alpha B(u, v_1) + \beta B(u, v_2)$.

Then, B is called a *bilinear form* on $\mathcal{U} \times \mathcal{V}$. The bilinear form B is *bounded* if there is a constant $M > 0$ (a bound), such that, for all $u \in \mathcal{U}$ and $v \in \mathcal{V}$,

$$|B(u, v)| \leq M \|u\|_{\mathcal{U}} \|v\|_{\mathcal{V}} \;. \tag{2.28}$$

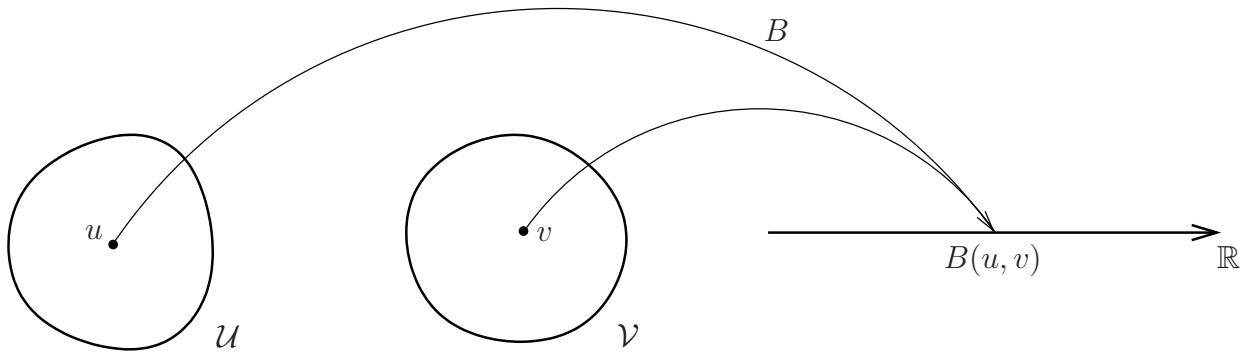


Figure 2.9. *A bilinear form on* $\mathcal{U} \times \mathcal{V}$

Consider a bilinear form $B(u, v)$ and fix $u \in \mathcal{U}$. Then an operator $A_u : \mathcal{V} \mapsto \mathbb{R}$ is defined according to

$$A_u[v] \;=\; B(u, v) \quad ; \quad u \text{ fixed} \;. \tag{2.29}$$

Operator A_u is called the *formal operator* associated with the bilinear form B. Similarly, when $v \in \mathcal{V}$ is fixed in $B(u, v)$, then an operator $A_v : \mathcal{U} \mapsto \mathbb{R}$ is defined as

$$A_v[u] \;=\; B(u, v) \quad ; \quad v \text{ fixed} \;, \tag{2.30}$$

and is called the *formal adjoint* of A_u.

Clearly, both A_u and A_v are linear (since they emanate from a bilinear form) and are often referred to as *linear forms* or linear functionals.

2.4 Background on variational calculus

The solutions to partial differential equations are often associated with extremization of functionals over a properly defined space of admissible functions. This subject will be addressed in detail in Chapter 4. Some preliminary information on variational calculus is presented here as background to forthcoming developments.

Consider a functional $I : \mathcal{U} \mapsto \mathbb{R}$, where $\mathcal{U}$ consists of functions $u = u(x, y, \ldots)$ that can play the role of the dependent variable in a partial differential equation. The *variation* δu of u is an arbitrary function defined on the same domain as u and represents "admissible" changes to the function u. The specific scope of this admissibility is left intentionally ambiguous at this stage, but will be clarified in Chapter 4. For example, if $\Omega \subset \mathbb{R}^n$ is the domain of $u \in \mathcal{U}$ with boundary $\partial\Omega$, where

$$\mathcal{U} \;=\; \big\{ u \in H^1(\Omega) \quad | \quad u = \bar{u} \text{ on } \partial\Omega \big\} \;\; ,$$

then δu may belong to the set

$$\mathcal{U}_0 \;=\; \big\{ u \in H^1(\Omega) \quad | \quad u = 0 \text{ on } \partial\Omega \big\} \;\; .$$

The preceding example illustrates that the variation δu of a function u is essentially restricted only by conditions related to the definition of the function u itself.

As already mentioned, interest will be focused on the determination of functions u^*, which render the functional $I[u]$ stationary (namely, minimum, maximum or a saddle point), as schematically indicated in Figure 2.10.

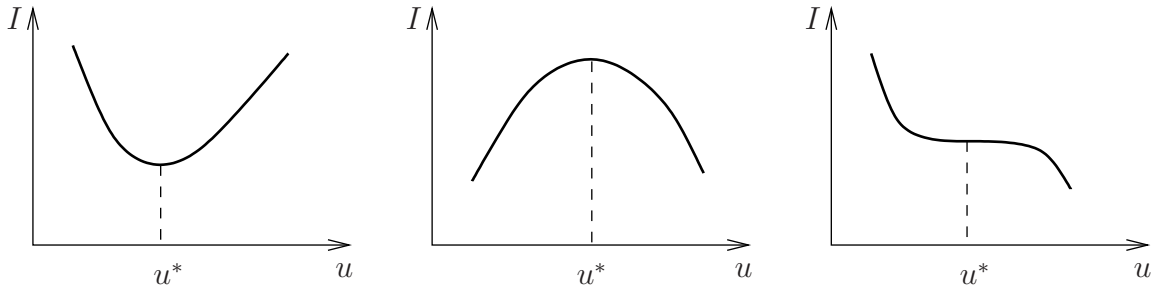


Figure 2.10. *A functional exhibiting a minimum, maximum or saddle point at* $u = u^*$

Define the (first) *variation* $\delta I[u]$ of $I[u]$ as

$$\delta I[u] = \lim_{w \to 0} \frac{I[u + w\delta u] \;-\; I[u]}{w} \;\; , \tag{2.31}$$

and, by induction, the k-th variation as

$$\delta^k I[u] \;=\; \delta(\delta^{k-1} I[u]) \;, \quad k = 2, 3, \ldots \;. \tag{2.32}$$

Alternatively, the variations of $I[u]$ may be determined by first expanding $I[u + \delta u]$ around u and then forming $\delta^k I[u]$, $k = 1, 2, \ldots$, from all terms that involve only the k-th power of δu, according to

$$I[u \;+\; \delta u] \;=\; I[u] \;+\; \delta I[u] \;+\; \frac{1}{2!}\delta^2 I[u] \;+\; \frac{1}{3!}\delta^3 I[u] \;+\; \ldots \;. \tag{2.33}$$

Example 2.4.1: Variations of certain useful functionals

(a) Let I be reduced to a function f defined on $\mathbb{R}^n$ as

$$f(\mathbf{x}) \;=\; \frac{1}{2}\mathbf{x}\cdot\mathbf{A}\mathbf{x} - \mathbf{x}\cdot\mathbf{b} \;,$$

where $\mathbf{A}$ is an $n \times n$ symmetric positive-definite matrix and $\mathbf{b}$ belongs to $\mathbb{R}^n$. Using (2.31), it follows that

$$\delta f(\mathbf{x}) \;=\; \delta\mathbf{x}\cdot\mathbf{A}\mathbf{x} - \delta\mathbf{x}\cdot\mathbf{b} \;=\; \delta\mathbf{x}\cdot(\mathbf{A}\mathbf{x} - \mathbf{b})$$

and

$$\delta^2 f(\mathbf{x}) \;=\; \delta\mathbf{x}\cdot\mathbf{A}\delta\mathbf{x} \;.$$

Therefore, it is seen that extremization (and, more specifically, in this case, minimization) of the above functional yields a system of n linear algebraic equations with n unknowns. Since $\mathbf{A}$ is assumed positive-definite, the system has a unique solution

$$\mathbf{x} \;=\; \mathbf{A}^{-1}\mathbf{b} \;,$$

which coincides with the minimum of $f(\mathbf{x})$. Several iterative methods for the solution of linear algebraic systems effectively exploit this minimization property.

(b) The variations of functional $I[u]$ defined as

$$I[u] \;=\; \int_0^1 u^2\, dx$$

can be determined by directly using (2.31). Thus,

$$\begin{aligned}
\delta I[u] \;&=\; \lim_{\omega\to 0} \frac{\int_0^1 \left[(u \;+\; \omega\delta u)^2 \;-\; u^2\right] dx}{\omega} \\
&=\; \lim_{\omega\to 0} \int_0^1 \left[2u\,\delta u \;+\; \omega(\delta u)^2\right] dx \;=\; \int_0^1 2u\,\delta u\, dx \;,
\end{aligned}$$

$$\begin{aligned}
\delta^2 I[u] \;&=\; \delta(\delta I[u]) \;=\; \lim_{\omega\to 0} \frac{\int_0^1 \left[2(u \;+\; \omega\delta u)\,\delta u \;-\; 2u\,\delta u\right] dx}{\omega} \\
&=\; 2\int_0^1 (\delta u)^2\, dx
\end{aligned}$$

and

$$\delta^k I[u] \;=\; 0\;, \qquad k = 3, 4, \ldots\;.$$

Using the alternative definition (2.33) for the variations of $I[u]$, write

$$\begin{aligned} I[u+\delta u] = \int_0^1 (u+\delta u)^2\,dx \;&=\; \int_0^1 u^2\,dx + 2\int_0^1 u\delta u\,dx + \int_0^1 (\delta u)^2\,dx \\ &=\; I[u] + \delta I[u] + \frac{1}{2!}\delta^2 I[u]\;, \end{aligned}$$

leading again to the expressions for $\delta^k I[u]$ determined above. ◀

Let u^* be a function that extremizes $I[u]$, and write for any variation δu around u^*

$$I[u^*+\delta u] \;=\; I[u^*] \;+\; \delta I[u^*] \;+\; \frac{1}{2!}\delta^2 I[u^*] \;+\; \ldots\;. \tag{2.34}$$

Equation (2.34) implies that necessary and sufficient condition for extremization of I at $u = u^*$ is that

$$\delta I[u^*] \;=\; 0\;. \tag{2.35}$$

Remarks:

- ☛ In the variation of $I[u]$, the independent variables $x, y, \ldots$ that are arguments of u remain "frozen", since the variation is taken over the functions u themselves and not over the variables of their domain.

- ☛ The definition of the first variation of a functional $I[u]$ in (2.31) may be readily extended to the first variation of any operator $A[u]$.

- ☛ Standard operations from differential calculus also apply to variational calculus, e.g., for any two functionals I_1 and I_2 defined on the same function space and any scalar constants α and β,

$$\begin{aligned} \delta(\alpha I_1 + \beta I_2) \;&=\; \alpha\,\delta I_1 + \beta\,\delta I_2\;, \\ \delta(I_1 I_2) \;&=\; \delta I_1\, I_2 + I_1\,\delta I_2\;. \end{aligned}$$

- ☛ Differentiation/integration and variation are operations that generally commute, that is, for $u = u(x)$,

$$\delta\frac{du}{dx} \;=\; \frac{d}{dx}(\delta u)\;,$$

assuming continuity of $\dfrac{du}{dx}$, and

$$\delta \int_\Omega u\, dx \;=\; \int_\Omega \delta u\, dx \;,$$

assuming that the domain of integration Ω is independent of u.

☛ If a functional I depends on functions $u, v, \ldots$, then the variation of I obviously depends on the variations of all $u, v, \ldots$, that is,

$$\delta I[u, v, \ldots] \;=\; \lim_{\omega \to 0} \frac{I[u \;+\; \omega\delta u, v \;+\; \omega\delta v, \ldots] \;-\; I[u, v, \ldots]}{\omega} \;,$$

and

$$\delta^k I[u, v, \ldots] \;=\; \delta(\delta^{k-1} I[u, v, \ldots]) \;, \quad k = 2, 3, \ldots$$

or, alternatively,

$$I[u + \delta u, v + \delta v, \ldots] \;=\; I[u, v, \ldots] \;+\; \delta I[u, v, \ldots] \;+\; \frac{1}{2!}\delta^2 I[u, v, \ldots] \;+\; \ldots \;.$$

☛ If a functional I depends on both u and its derivatives u', u'', ..., then the variation of I also depends on the variation of all u', u'', ..., namely

$$\delta I[u, u', u'', \ldots] \;=\; \lim_{\omega \to 0} \frac{I[u \;+\; \omega\delta u, u' \;+\; \omega\delta u', u'' \;+\; \omega\delta u'', \ldots] \;-\; I[u, u', u'', \ldots]}{\omega} \;.$$

A weaker (that is, more general) definition of the variation of a functional is obtained using the notion of a *directional* (or *Gâteaux*) *differential* of $I[u]$ at point u in the direction v, denoted by $D_v I[u]$ (or $DI(u, v)$). This is defined as

$$D_v I[u] \;=\; \left[\frac{d}{dw} I[u + wv]\right]_{w=0} \;. \tag{2.36}$$

For a large class of functionals, the variation $\delta I[u]$ can be interpreted as the Gâteaux differential of $I[u]$ in the direction δu.

Example 2.4.2: Directional derivative of a simple functional
Consider a functional $I[u]$ defined as

$$I[u] \;=\; \int_\Omega u^2\, d\Omega \;.$$

The directional derivative of I at u in the direction v is given by

$$\begin{aligned} D_v I[u] &= \left[\frac{d}{dw}\int_\Omega (u + wv)^2\, d\Omega\right]_{w=0} \\ &= \left[\frac{d}{dw}\int_\Omega \left[u^2 + w2uv + w^2v^2\right]\, d\Omega\right]_{w=0} \\ &= \left[\int_\Omega \frac{d}{dw}\left[u^2 + w2uv + w^2v^2\right]\, d\Omega\right]_{w=0} \\ &= \left[\int_\Omega \left[2uv + 2wv^2\right]\, d\Omega\right]_{w=0} \\ &= \int_\Omega 2uv\, d\Omega\ . \end{aligned}$$

This result can be compared with that of a previous exercise, where it has been deduced that

$$\delta I[u] = \int_\Omega 2u\delta u\, d\Omega\ .$$

◀

2.5 Exercises

Problem 1

(a) Given $\mathbf{x} \in \mathbb{R}^n$ with components $(x_1, x_2, \ldots, x_n)$, show that the functions $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$, $\|\mathbf{x}\|_2 = (\sum_{i=1}^n x_i^2)^{1/2}$ and $\|\mathbf{x}\|_\infty = \max_{1\le i\le n} |x_i|$ are norms.

(b) Determine the points of $\mathbb{R}^2$ for which $\|\mathbf{x}\|_1 = 1$, $\|\mathbf{x}\|_2 = 1$ and $\|\mathbf{x}\|_\infty = 1$ separately. Sketch on a single plot the geometric curves corresponding to the above points.

Problem 2

Compute the inner product $< u, v >_{L_2(\Omega)}$ for $u = x+1$ and $v = 3x^2+1$, given that $\Omega = [-1, 1]$.

Problem 3

(a) Write the explicit form of the inner product $< u, v >_{H^2(\Omega)}$ and the associated norm on $H^2(\Omega)$ given that $\Omega \subset \mathbb{R}^2$.

(b) Using the result of part (a), find the "distance" in the H^2-norm between functions $u = \sin x + y$ and $v = x$ for $\Omega = \{(x, y) \mid |x| \le \pi,\ |y| \le \pi\}$.

Problem 4

Show the parallelogram law for inner product spaces:

$$\|u + v\|^2 + \|u - v\|^2 = 2\|u\|^2 + 2\|v\|^2\ .$$

Problem 5
Show that the integral I given by

$$I \;=\; \int_a^b \delta^2(x)\, dx \quad , \quad a \;<\; 0 \;<\; b$$

is not well-defined.

Hint: Use the definition of the Dirac-delta function $\delta(x)$ and construct a sequence of integrals I_n converging to I.

Problem 6
Given $\Omega = \{\mathbf{x} \in \mathbb{R}^2 \quad | \quad \|\mathbf{x}\|_2 \leq 1\}$, show that the function $f(\mathbf{x}) = \|\mathbf{x}\|_2^\alpha$ belongs to the Sobolev space $H^1(\Omega)$ for $\alpha > 0$. In addition, determine the range of α so that f also belong to $H^2(\Omega)$.

Problem 7
Consider the operator $A \;:\; \mathbb{R}^3 \mapsto \mathbb{R}^3$ defined as

$$A[(x_1, x_2, x_3)] \;=\; (x_1 \;+\; x_2 \;,\; 2x_1 \;+\; x_3 \;,\; x_2 \;-\; 2x_3) \;.$$

(a) Show that A is linear.

(b) Using the $\|\cdot\|_2$-norm for vectors, show that A is bounded and find an appropriate bound M.

(c) Use the inner product $<\cdot,\cdot>$ defined according to $<\mathbf{x},\mathbf{y}>=\mathbf{x}^T\mathbf{y}$, $\mathbf{x},\mathbf{y} \in \mathbb{R}^3$, to check the operator A for symmetry.

Problem 8
Let an operator $A : C^\infty(\Omega) \mapsto \mathcal{V} \subset C^\infty(\Omega)$ be defined as

$$A[u] \;=\; \frac{d^2}{dx^2}\big[a(x)\frac{d^2u}{dx^2}\big] \;+\; b(x)u \;,$$

where $\Omega = (0, L)$, $a(x) > 0, b(x) > 0, \forall x \in \Omega$. Show that the operator is linear, symmetric and positive, assuming that $u(0) = u(L) = 0$ and $\frac{du}{dx}(0) = \frac{du}{dx}(L) = 0$. Use the L_2-inner product in ascertaining symmetry and positiveness.

Problem 9
Determine the degree of smoothness of the real functions u_1 and u_2 by identifying the classes $C^k(\Omega)$ and $H^m(\Omega)$ to which they belong:

(a) $u_1(x) \;=\; x^n \quad , \quad n \text{ integer} > 0 \quad , \quad \Omega = (0, s) \quad , \quad s < \infty \;.$

(b) The function $u_2(x)$ is defined by means of its first derivative

$$\frac{du_2}{dx}(x) \;=\; \begin{cases} x \;-\; 0.25\;, & 0 < x \leq 1 \\ 0.25 \;-\; x\;, & 1 < x < 2 \end{cases} \quad ,$$

such that $u_2(0) = 0$ and $u_2(2) = -1$, where $\Omega = (0, 2)$.

Problem 10

Show that if u is a real-valued function of class $C^1(\Omega)$, where $\Omega \in \mathbb{R}$, then $\delta\dfrac{du}{dx} = \dfrac{d(\delta u)}{dx}$, that is, the operations of variation and differentiation commute.

Problem 11

Let the functional $I[u, u']$ be defined as

$$I[u, u'] \;=\; \int_0^1 \left(1 \;+\; u^2 \;+\; u'^2\right) dx \;.$$

(a) Compute the variations $\delta I[u, u']$ and $\delta^2 I[u, u']$ using the respective definitions.

(b) What is the value of the differential δI for $u = x^2$ and $\delta u = x$?

2.6 Suggestions for further reading

Sections 2.1-2.3

[1] G. Strang and G.J. Fix. *An Analysis of the Finite Element Method.* Prentice-Hall, Englewood Cliffs, 1973. [*The index of notations (p. 297) offers an excellent, albeit brief, discussion of mathematical preliminaries*].

[2] J.N. Reddy. *Applied Functional Analysis and Variational Methods in Engineering.* McGraw-Hill, New York, 1986. [*This book contains a very comprehensive and readable introduction to Functional Analysis with emphasis to applications in continuum mechanics*].

[3] T.J.R. Hughes. *The Finite Element Method; Linear Static and Dynamic Finite Element Analysis.* Prentice-Hall, Englewood Cliffs, 1987. [*Appendices 1.I and 4.I discuss concisely the mathematical preliminaries to the analysis of the finite element method*].

Section 2.4

[1] O. Bolza. *Lectures on the Calculus of Variations*. Chelsea, New York, 3rd edition, 1973. [*A classic book on calculus of variations that can serve as a reference, but not as a didactic text*].

[2] H. Sagan. *Introduction to the Calculus of Variations*. Dover, New York, 1992. [*A modern text on calculus of variations – Chapter 1 is very readable and pertinent to the present discussion of mathematical concepts*].

Chapter 3

Methods of Weighted Residuals

3.1 Introduction

Consider an open set $\Omega \subset \mathbb{R}^n$ with boundary $\partial\Omega$ that possesses a unique outer unit normal $\mathbf{n}$ at every point, as in Figure 3.1. A differential operator A involving derivatives up to order p is defined on a function space $\mathcal{U}$, and differential operators B_i, $i = 1, \ldots, k$, involving traces $\gamma_j = \dfrac{\partial^j u}{\partial n^j}, 0 \leq j \leq p-1$, are defined on appropriate boundary function spaces. Further, the boundary $\partial\Omega$ is decomposed (arbitrarily at present) into k parts $\partial\Omega_i$, $i = 1, \ldots, k$, such that

$$\overline{\bigcup_{i=1}^{k} \partial\Omega_i} \;=\; \partial\Omega \;.$$

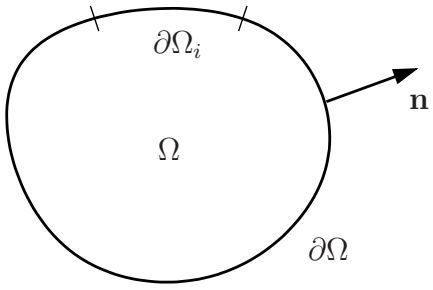


Figure 3.1. *An open and connected domain Ω with smooth boundary written as the union of boundary regions $\partial\Omega_i$*

Given functions f and g_i, $i = 1, \ldots, k$, on Ω and $\partial\Omega_i$, respectively, a mathematical problem

associated with a partial differential equation is described by the system

$$\begin{aligned} A[u] &= f \qquad \text{in } \Omega \ , \\ B_i[u] &= g_i \qquad \text{on } \partial\Omega_i \quad , \quad i = 1, \ldots, k \ . \end{aligned} \tag{3.1}$$

With reference to equations (3.1), define functions w_Ω and w_i, $i = 1, \ldots, k$, on Ω and $\partial\Omega_i$, respectively, such that the scalar quantity R, given by

$$R \ = \ \int_\Omega w_\Omega (A[u] \ - \ f) \, d\Omega \ + \ \sum_{i=1}^{k} \int_{\partial\Omega_i} w_i (B_i[u] \ - \ g_i) \, d\Gamma \ , \tag{3.2}$$

be algebraically consistent (that is, all integrals of the right-hand side have the same units). These functions are called *weighting functions* (or *test functions*).

Equations (3.1) constitute the *strong form* of the differential equation. The scalar equation

$$\int_\Omega w_\Omega (A[u] - f) \, d\Omega \ + \ \sum_{i=1}^{k} \int_{\partial\Omega_i} w_i (B_i[u] - g_i) \, d\Gamma \ = \ 0 \ , \tag{3.3}$$

where functions w_Ω and w_i, $i = 1, \ldots, k$, are *arbitrary* to within consistency of units and sufficient smoothness for all integrals in (3.3) to exist, is the associated general *weighted-residual form* of the differential equation.

By inspection, the strong form (3.1) implies the general weighted-residual form. The converse is also true, conditional upon sufficient smoothness of the involved fields. The following lemma provides the necessary background for the ensuing proof in the context of $\mathbb{R}^n$.

The localization lemma

Let $f : \ \Omega \mapsto \mathbb{R}$ be a continuous function, where $\Omega \subset \mathbb{R}^n$ is an open set. Then,

$$\int_{\Omega_i} f \, d\Omega \ = \ 0 \ , \tag{3.4}$$

for all open $\Omega_i \subset \Omega$, if, and only if, $f = 0$ everywhere in Ω.

In proving the above lemma, one immediately notes that if $f = 0$, then the integral of f will vanish identically over any Ω_i. To prove the converse, assume by contradiction that there exists a point $\mathbf{x}_0$ in Ω where

$$f(\mathbf{x}_0) \ = \ f_0 \ \neq \ 0 \ , \tag{3.5}$$

and without loss of generality, let $f_0 > 0$. It follows that, since f is continuous and Ω is open, there exists an open "sphere" $\mathcal{N}_\delta \subset \Omega$ of radius $\delta > 0$ centered at $\mathbf{x}_0$, that is,

$$\mathcal{N}_\delta \;=\; \{\mathbf{x} \in \mathbb{R}^n \quad | \quad \|\mathbf{x} - \mathbf{x}_0\| < \delta\} \;, \tag{3.6}$$

such that

$$|f(\mathbf{x}) \;-\; f(\mathbf{x}_0)| \;<\; \epsilon \;=\; \frac{f_0}{2} \,, \tag{3.7}$$

for all $\mathbf{x} \in \mathcal{N}_\delta$. Thus, it is seen from (3.7) that

$$f(\mathbf{x}) \;>\; \frac{f_0}{2} \tag{3.8}$$

everywhere in $\mathcal{N}_\delta$, hence

$$\int_{\mathcal{N}} f \, d\Omega \;>\; \frac{1}{2} \int_{\mathcal{N}} f_0 \, d\Omega \;>\; 0 \;, \tag{3.9}$$

which constitutes a contradiction with the original assumption that the integral of f vanishes identically over all open Ω_i.

Returning to the relation between (3.1) and (3.3), note that since the latter holds for arbitrary choices of w_Ω and w_i, one may define functions

$$w_\Omega(\mathbf{x}) \;=\; \begin{cases} 1 & \text{if } \mathbf{x} \in \Omega_i \\ 0 & \text{otherwise} \end{cases} \;, \tag{3.10}$$

for any open $\Omega_i \subset \Omega$, and

$$w_i \;=\; 0 \quad , \quad i = 1, \ldots, k \;. \tag{3.11}$$

Invoking the localization lemma, it is readily concluded that $(3.1)_1$ should hold everywhere in Ω, conditional upon continuity of $A[u]$ and f. Repeating the same process k times (once for each of the boundary conditions) for appropriately defined weighting functions and involving the localization theorem, each one of equations $(3.1)_2$ is recovered on its respective domain.

The equivalence of the strong form and the weighted-residual form plays a fundamental role in the construction of approximate solutions (including finite element solutions) to the underlying problem. Various approximation methods, such as the Galerkin, collocation and least-squares methods, are derived by appropriately restricting the admissible form of the weighting functions and the actual solution.

The above preliminary development applies to linear and non-linear differential operators of any order. A large portion of the forthcoming discussion of weighted-residual methods will involve linear differential equations for which the linear operator A contains derivatives of u up to order $p = 2q$, where q is an integer, whereas linear operators B_i contain only derivatives of order $0, \ldots, 2q - 1$.

3.2 Galerkin methods

Galerkin methods provide a fairly general framework for the numerical solution of differential equations within the context of the weighted-residual formalization. Here, an introduction to Galerkin methods is attempted by means of their application to the solution of a representative boundary-value problem.

Consider domain $\Omega \subset \mathbb{R}^2$ with smooth boundary $\partial\Omega = \overline{\Gamma_u \cup \Gamma_q}$ and $\Gamma_u \cap \Gamma_q = \emptyset$, as in Figure 3.2. Let the strong form of a boundary-value problem be as follows:

$$\begin{aligned} \frac{\partial}{\partial x_1}(k\,\frac{\partial u}{\partial x_1}) \;+\; \frac{\partial}{\partial x_2}(k\,\frac{\partial u}{\partial x_2}) \;&=\; f \qquad \text{in } \Omega\ , \\ -k\,\frac{\partial u}{\partial n} \;&=\; \bar{q} \qquad \text{on } \Gamma_q\ , \\ u \;&=\; \bar{u} \qquad \text{on } \Gamma_u\ , \end{aligned} \tag{3.5}$$

where $u = u(x_1, x_2)$ is the (yet unknown) solution in Ω. Continuous functions $k = k(x_1, x_2)$

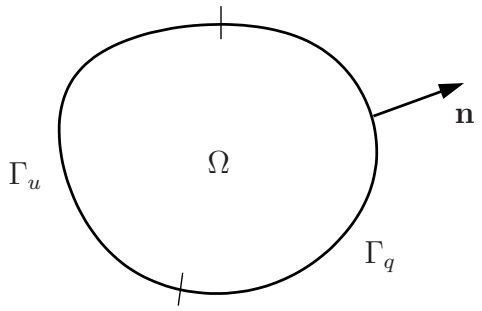


Figure 3.2. *The domain Ω of the Laplace-Poisson equation with Dirichlet boundary Γ_u and Neumann boundary Γ_q*

and $f = f(x_1, x_2)$ defined in Ω, as well as continuous functions $\bar{q} = \bar{q}(x_1, x_2)$ on Γ_q and $\bar{u} = \bar{u}(x_1, x_2)$ on Γ_u are *data* of the problem (that is, they are known in advance). The boundary conditions $(3.5)_2$ and $(3.5)_3$ are termed *Neumann* and *Dirichlet* conditions, respectively.

It is clear from the statement of the strong form that both the domain and the boundary differential operators are linear in u. This is the two-dimensional *Laplace-Poisson equation*, which has applications in the mathematical modeling of numerous systems in structural mechanics, heat conduction, electrostatics, flow in porous media, etc.

Residual functions R_Ω, R_q and R_u are defined according to

$$\begin{aligned} R_\Omega(x_1,x_2) &= \frac{\partial}{\partial x_1}(k\,\frac{\partial u}{\partial x_1}) + \frac{\partial}{\partial x_2}(k\,\frac{\partial u}{\partial x_2}) - f \qquad \text{in } \Omega\ , \\ R_q(x_1,x_2) &= -k\,\frac{\partial u}{\partial n} - \bar{q} \qquad \text{on } \Gamma_q\ , \\ R_u(x_1,x_2) &= u - \bar{u} \qquad \text{on } \Gamma_u\ . \end{aligned} \tag{3.6}$$

Introducing arbitrary functions $w_\Omega = w_\Omega(x_1,x_2)$ in Ω, $w_q = w_q(x_1,x_2)$ on Γ_q and $w_u = w_u(x_1,x_2)$ on Γ_u, the *weighted-residual form* (3.3) is an integro-differential equation, which reads

$$\int_\Omega w_\Omega R_\Omega\,d\Omega + \int_{\Gamma_q} w_q R_q\,d\Gamma + \int_{\Gamma_u} w_u R_u\,d\Gamma = 0\ . \tag{3.7}$$

The weighted-residual form is also referred to as a *weak form* in contrast to the strong form defined in (3.5). As argued earlier, is equivalent to the strong form of the boundary-value problem provided that the weighting functions are arbitrary to within unit consistency and proper definition of the integrals in (3.7).

A series of assumptions are introduced in deriving the Galerkin method. First, assume that boundary condition $(3.5)_3$ is satisfied at the outset, namely that the solution u is sought over a set of candidate functions that already satisfy $(3.5)_3$. Hence, the third integral of the left-hand side of (3.7) vanishes and the choice of function w_u becomes irrelevant.

Observing that the two remaining integral terms in (3.7) are consistent unit-wise, provided that w_Ω and w_q have the same units, introduce the second assumption leading to a so-called *Galerkin formulation*: this is a particular choice of functions w_Ω and w_q according to which

$$\begin{aligned} w_\Omega &= w \qquad \text{in } \Omega\ , \\ w_q &= w \qquad \text{on } \Gamma_q\ . \end{aligned} \tag{3.8}$$

Substitution of the above expressions for the weighting functions into the reduced form of (3.7) yields

$$\int_\Omega w\left[\frac{\partial}{\partial x_1}\left(k\,\frac{\partial u}{\partial x_1}\right) + \frac{\partial}{\partial x_2}\left(k\,\frac{\partial u}{\partial x_2}\right) - f\right] d\Omega - \int_{\Gamma_q} w\left[k\,\frac{\partial u}{\partial n} + \bar{q}\right] d\Gamma = 0\ , \tag{3.9}$$

which, after integration by parts[1] and use of the divergence theorem[2], is rewritten as

$$-\int_\Omega \left[\frac{\partial w}{\partial x_1}\,k\,\frac{\partial u}{\partial x_1} \;+\; \frac{\partial w}{\partial x_2}\,k\,\frac{\partial u}{\partial x_2} \;+\; wf\right] d\Omega \;+\; \int_{\partial\Omega} wk\left[\frac{\partial u}{\partial x_1}n_1 + \frac{\partial u}{\partial x_2}n_2\right] d\Gamma \\ -\int_{\Gamma_q} w\left[k\,\frac{\partial u}{\partial n} + \bar{q}\right] d\Gamma \;=\; 0\,. \quad (3.10)$$

Recall that the projection of the gradient of u in the direction of the outward unit normal $\mathbf{n}$ is given by

$$\frac{\partial u}{\partial n} \;=\; \frac{du}{d\mathbf{x}}\cdot\mathbf{n} \;=\; \frac{\partial u}{\partial x_1}\,n_1 + \frac{\partial u}{\partial x_2}\,n_2\,, \quad (3.11)$$

and, thus, the above weighted-residual equation is also written as

$$-\int_\Omega \left[\frac{\partial w}{\partial x_1}\,k\,\frac{\partial u}{\partial x_1} \;+\; \frac{\partial w}{\partial x_2}\,k\,\frac{\partial u}{\partial x_2} \;+\; wf\right] d\Omega \;+\; \int_{\Gamma_u} wk\,\frac{\partial u}{\partial n}\,d\Gamma \;-\; \int_{\Gamma_q} w\bar{q}\,d\Gamma \;=\; 0\,. \quad (3.12)$$

Here, an additional assumption is introduced, namely

$$w \;=\; 0 \qquad \text{on } \Gamma_u\,. \quad (3.13)$$

This last assumption leads to the weighted residual equation

$$\int_\Omega \left[\frac{\partial w}{\partial x_1}\,k\,\frac{\partial u}{\partial x_1} \;+\; \frac{\partial w}{\partial x_2}\,k\,\frac{\partial u}{\partial x_2} \;+\; wf\right] d\Omega \;+\; \int_{\Gamma_q} w\bar{q}\,d\Gamma \;=\; 0\,, \quad (3.14)$$

which is identified with the Galerkin formulation of the original problem.

Clearly, the purpose of the preceding three assumptions is to simplify the original weighted-residual form (3.7) without sacrificing any essential approximating properties in the resulting equation (3.14). Also, it is important to observe that, owing to the use of the divergence theorem, the highest-order partial derivative in (3.14) is one, as opposed to the strong form, which includes second-order partial derivatives.

[1]The relevant theorem states that if f and g are C^1 functions from Ω to $\mathbb{R}^N$, then

$$\int_\Omega f g_{,i}\,d\Omega \;=\; \int_\Omega (fg)_{,i}\,d\Omega - \int_\Omega f_{,i}\,g\,d\Omega\,.$$

[2]This theorem states that given a closed smooth surface $\partial\Omega$ with interior Ω and a C^1 function f from Ω to $\mathbb{R}^N$, then

$$\int_\Omega f_{,i}\,d\Omega \;=\; \int_{\partial\Omega} f n_i\,d\Gamma\,,$$

where n_i denotes the i-th component of the outer unit normal to $\partial\Omega$.

Alternatively, it is possible to assume that *both* $(3.5)_{2,3}$ are satisfied at the outset and write the weighted residual statement for $w_\Omega = w$ as

$$\int_\Omega w \left[\frac{\partial}{\partial x_1} \left(k \frac{\partial u}{\partial x_1} \right) + \frac{\partial}{\partial x_2} \left(k \frac{\partial u}{\partial x_2} \right) - f \right] d\Omega \;=\; 0 \; . \tag{3.15}$$

Again, integration by parts and use of the divergence theorem transform the above equation into

$$- \int_\Omega \left[\frac{\partial w}{\partial x_1} k \frac{\partial u}{\partial x_1} \;+\; \frac{\partial w}{\partial x_2} k \frac{\partial u}{\partial x_2} \;+\; wf \right] d\Omega \;+\; \int_{\partial\Omega} wk \frac{\partial u}{\partial n} d\Gamma \;=\; 0 \; , \tag{3.16}$$

which, in turn, becomes identical to (3.14) by imposing restriction (3.13) and making explicit use of the assumed condition $(3.5)_2$.

The weighted residual problem associated with equation (3.14) can be expressed operationally as follows: find $u \in \mathcal{U}$, such that, for all $w \in \mathcal{W}$,

$$B(w,u) \;+\; (w,f) \;+\; (w,\bar{q})_{\Gamma_q} \;=\; 0 \; , \tag{3.17}$$

where

$$\mathcal{U} \;=\; \left\{ u \in H^1(\Omega) \quad | \quad u = \bar{u} \text{ on } \Gamma_u \right\} \; , \tag{3.18}$$

$$\mathcal{W} \;=\; \left\{ w \in H^1(\Omega) \quad | \quad w = 0 \text{ on } \Gamma_u \right\} \; . \tag{3.19}$$

In the above, $B(w,u)$ is a (symmetric) bilinear form defined as

$$B(w,u) \;=\; \int_\Omega \left(\frac{\partial w}{\partial x_1} k \frac{\partial u}{\partial x_1} \;+\; \frac{\partial w}{\partial x_2} k \frac{\partial u}{\partial x_2} \right) d\Omega \; , \tag{3.20}$$

whereas (w,f) and $(w,\bar{q})_{\Gamma_q}$ are linear forms defined respectively as

$$(w,f) \;=\; \int_\Omega wf \, d\Omega \tag{3.21}$$

and

$$(w,\bar{q})_{\Gamma_q} \;=\; \int_{\Gamma_q} w\bar{q} \, d\Gamma \; . \tag{3.22}$$

The identification of admissible solution fields $\mathcal{U}$ and weighting function fields $\mathcal{W}$ is dictated by restrictions placed during the derivation of (3.14) and by the requirement that the bilinear form $B(w,u)$ be computable (that is, the integral be well-defined). Clearly, alternative definitions of $\mathcal{U}$ and $\mathcal{W}$ (with regards to smoothness) may also be acceptable.

A finite-dimensional *Galerkin approximation* of (3.14) is obtained by restating the weighted-residual problem as follows: find $u_h \in \mathcal{U}_h$, such that, for all $w_h \in \mathcal{W}_h$,

$$B(w_h, u_h) + (w_h, f) + (w_h, \bar{q})_{\Gamma_q} = 0 \ , \tag{3.23}$$

where $\mathcal{U}_h$ and $\mathcal{W}_h$ are subspaces of $\mathcal{U}$ and $\mathcal{W}$, respectively, so that

$$\begin{aligned} u \doteq u_h \ &= \ \sum_{I=1}^{N} \alpha_I \, \varphi_I(x_1, x_2) \ + \ \varphi_0(x_1, x_2) \ , \\ w \doteq w_h \ &= \ \sum_{I=1}^{N} \beta_I \, \psi_I(x_1, x_2) \ . \end{aligned} \tag{3.24}$$

In the above, $\varphi_I(x_1, x_2)$ and $\psi_I(x_1, x_2)$, $I = 1, 2, \ldots, N$, are given functions (called *interpolation* or *basis functions*), which, for convenience, vanish on Γ_u, and $\varphi_0(x_1, x_2)$ is chosen such that $\varphi_0 = \bar{u}$ on Γ_u so that u_h satisfy boundary condition $(3.5)_3$. Parameters $\alpha_I \in \mathbb{R}$, $I = 1, 2, \ldots, N$, are to be determined by invoking (3.14), while parameters $\beta_I \in \mathbb{R}$, $I = 1, 2, \ldots, N$, are arbitrary.

A *Bubnov*[3]*-Galerkin* approximation is obtained from (3.24) by setting $\psi_I = \varphi_I$ for all $I = 1, 2, \ldots, N$. This is the most popular version of the Galerkin method. Use of $\psi_I \neq \varphi_I$ in the discrete weighting function w_h yields a so-called *Petrov*[4]*-Galerkin* approximation.

Substitution of u_h and w_h, as defined in (3.24), into the weak form (3.14) results in

$$\begin{aligned} \sum_{I=1}^{N} \beta_I \int_\Omega \left[\psi_{I,1} \ \psi_{I,2}\right] k \left(\sum_{J=1}^{N} \begin{bmatrix} \varphi_{J,1} \\ \varphi_{J,2} \end{bmatrix} \alpha_J \ + \ \begin{bmatrix} \varphi_{0,1} \\ \varphi_{0,2} \end{bmatrix} \right) d\Omega \\ + \ \sum_{I=1}^{N} \beta_I \int_\Omega \psi_I f \, d\Omega + \sum_{I=1}^{N} \beta_I \int_{\Gamma_q} \psi_I \, \bar{q} \, d\Gamma \ = \ 0 \ , \end{aligned} \tag{3.25}$$

or, alternatively,

$$\sum_{I=1}^{N} \beta_I \Big(\sum_{J=1}^{N} K_{IJ} \, \alpha_J \ - \ F_I \Big) \ = \ 0 \ , \tag{3.26}$$

where

$$K_{IJ} \ = \ \int_\Omega \left[\psi_{I,1} \ \psi_{I,2}\right] k \begin{bmatrix} \varphi_{J,1} \\ \varphi_{J,2} \end{bmatrix} d\Omega \ , \tag{3.27}$$

and

$$F_I \ = \ - \int_\Omega \psi_I f \, d\Omega \ - \ \int_\Omega \left[\psi_{I,1} \ \psi_{I,2}\right] k \begin{bmatrix} \varphi_{0,1} \\ \varphi_{0,2} \end{bmatrix} d\Omega \ - \ \int_{\Gamma_q} \psi_I \, \bar{q} \, d\Gamma \ . \tag{3.28}$$

[3]Ivan Grigoryevich Bubnov (1872-1919) was a Russian naval architect.

[4]Georgi Ivanovich Petrov (1912-1987) was a Russian aerodynamicist.

Since the parameters β_I are arbitrary, it follows easily from (3.26) that

$$\sum_{J=1}^{N} K_{IJ}\alpha_J \;=\; F_I \quad , \quad I = 1, 2, \ldots, N \;, \tag{3.29}$$

or, in matrix form,

$$\mathbf{K}\boldsymbol{\alpha} \;=\; \mathbf{F} \;, \tag{3.30}$$

where $\mathbf{K}$ is the $N \times N$ *stiffness matrix* with components given by (3.27), $\mathbf{F}$ is the $N \times 1$ *forcing vector* with components as in (3.28), and $\boldsymbol{\alpha}$ is the $N \times 1$ vector of parameters α_I introduced in $(3.24)_1$.

It is important to note that the Galerkin approximation (3.24) transforms the integro-differential equation (3.14) into a system of linear algebraic equations to be solved for $\boldsymbol{\alpha}$.

Remarks:

- It should be noted that, strictly speaking, $\mathcal{U}$ is not a linear space, since it violates the closure property (see Section 2.1). However, it is easy to reformulate equations (3.5) so that they only involve homogeneous Dirichlet boundary conditions, in which case $\mathcal{U}$ is formally a linear space and $\mathcal{U}_h$ a linear subspace of it. Indeed, any linear partial differential equation of the form

  $$A[u] \;=\; f$$

  with non-homogeneous boundary conditions

  $$u \;=\; \bar{u}$$

  on a part of its boundary Γ_u, can be rewritten without loss of generality as

  $$A[v] \;=\; f \;-\; A[u_0]$$

  with homogeneous boundary conditions on Γ_u, where u_0 is any given function in the domain of u, such that $u_0 = \bar{u}$ on Γ_u.

- It can be easily seen from (3.27) that the stiffness matrix $\mathbf{K}$ is symmetric for a Bubnov-Galerkin approximation. For the same type of approximation, it can be shown that, under mild assumptions, $\mathbf{K}$ is also positive-definite (therefore non-singular), so that the system (3.30) possesses a unique solution.

- ☛ Generally, there exists no precisely defined set of assumptions that guarantee the non-singularity of the stiffness matrix $\mathbf{K}$ emanating from a Petrov-Galerkin approximation.
- ☛ The terminology "stiffness" matrix and "forcing" vector originates in structural engineering and is associated with the physical interpretation of these quantities in the context of linear elasticity.

Example 3.2.1: Bubnov-Galerkin approximation for one-dimensional Laplace-Poisson equation

Consider a one-dimensional counterpart of the Laplace-Poisson equation in the form

$$\begin{aligned} \frac{d^2u}{dx^2} &= 1 \qquad \text{in } \Omega = (0,1) \ , \\ -\frac{du}{dx} &= 2 \qquad \text{on } \Gamma_q = \{1\} \ , \\ u &= 0 \qquad \text{on } \Gamma_u = \{0\} \ . \end{aligned}$$

Hence, equation (3.14) takes the form

$$\int_0^1 \left(\frac{dw}{dx}\frac{du}{dx} + w\right) dx + 2w\Big|_{x=1} = 0 \ . \qquad (\dagger)$$

A one-parameter Bubnov-Galerkin approximation can be obtained by setting $N = 1$ in equations (3.24) and choosing

$$\varphi_0(x) = 0$$

and

$$\varphi_1(x) = x \ ,$$

where, of course, $\varphi_1(0) = 0$. Substituting u_h and w_h into (†) gives

$$\int_0^1 (\beta_1\alpha_1 + \beta_1 x)\, dx + 2\beta_1 = 0 \ ,$$

and, since β_1 is an arbitrary parameter, it follows that

$$\alpha_1 = -\frac{5}{2} \ .$$

Thus, the one-parameter Bubnov-Galerkin approximation of the solution to the above differential equation is

$$u_h(x) = -\frac{5}{2}\, x \ .$$

Similarly, a two-parameter Bubnov-Galerkin approximation is obtained by choosing

$$\varphi_0(x) = 0$$

and

$$\varphi_1(x) \;=\; x \qquad , \qquad \varphi_2(x) \;=\; x^2 \;.$$

where, as required, $\varphi_1(0) = \varphi_2(0) = 0$. Again, (†) implies that

$$\int_0^1 \left[(\beta_1 \;+\; 2\beta_2 x)(\alpha_1 \;+\; 2\alpha_2 x) \;+\; (\beta_1 x \;+\; \beta_2 x^2)\right] dx \;+\; 2(\beta_1 \;+\; \beta_2) \;=\; 0 \;,$$

and due to the arbitrariness of β_1 and β_2, one may write

$$\begin{aligned} \int_0^1 \beta_1(\alpha_1 \;+\; 2\alpha_2 x)\, dx \;&=\; -2\beta_1 \;-\; \int_0^1 \beta_1 x\, dx \;, \\ \int_0^1 \beta_2\, 2x(\alpha_1 \;+\; 2\alpha_2 x)\, dx \;&=\; -2\beta_2 \;-\; \int_0^1 \beta_2 x^2\, dx \;, \end{aligned}$$

from where it follows that

$$\begin{aligned} \alpha_1 \;+\; \alpha_2 \;&=\; -\frac{5}{2} \;, \\ \alpha_1 \;+\; \frac{4}{3}\alpha_2 \;&=\; -\frac{7}{3} \;. \end{aligned}$$

Solving the above linear system yields $\alpha_1 = -3$ and $\alpha_2 = \frac{1}{2}$, so that

$$u_h(x) \;=\; -3x \;+\; \frac{1}{2}\, x^2 \;.$$

It can be easily confirmed by direct integration that the exact solution of the differential equation is

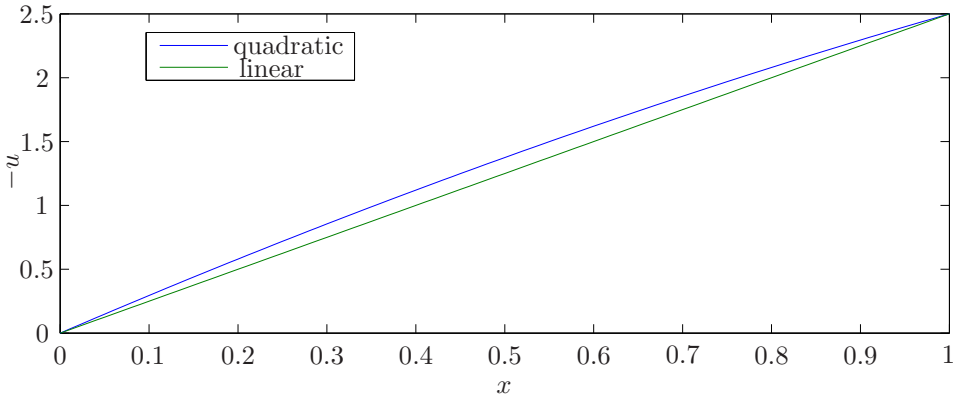


Figure 3.3. Linear and quadratic approximations of the solution to the boundary-value problem in Example 3.2.1

identical to the one obtained by the above two-parameter Bubnov-Galerkin approximation. It can be concluded that in this particular problem, the two-dimensional subspace $\mathcal{U}_h$ of all admissible functions $\mathcal{U}$ contains the exact solution, and, also, that the Bubnov-Galerkin method is capable of recovering it. As will be argued later, the latter is not an "accident", but rather results from an important property of the Bubnov-Galerkin method.

Non-polynomial (*e.g.*, piecewise polynomial or trigonometric) interpolation functions are also acceptable options for ϕ_i. ◀

The Galerkin method is now summarized in the context of the model problem

$$\begin{aligned} A[u] &= f \qquad \text{in } \Omega \ , \\ B[u] &= g \qquad \text{on } \Gamma_q \ , \\ u &= \bar{u} \qquad \text{on } \Gamma_u \ , \end{aligned} \tag{3.31}$$

where A is a linear second-order differential operator on a space of admissible domain functions u, and B is a linear first-order differential operator on the space of the traces of u. In addition, it is assumed that $\partial\Omega = \overline{\Gamma_u \cup \Gamma_q}$ and $\Gamma_u \cap \Gamma_q = \emptyset$. The method is based on the construction of a weighted integral form written as

$$\int_\Omega w_\Omega (A[u] \ - \ f)\, d\Omega \ + \ \int_{\Gamma_q} w_q (B[u] \ - \ g)\, d\Gamma \ = \ 0 \ , \tag{3.32}$$

where the space of admissible solutions u satisfies $(3.31)_3$ at the outset. In addition, w_q is chosen to vanish identically on Γ_u and, depending on unit consistency and the particular form of $(3.31)_2$, is chosen to be equal to w (or $-w$) on Γ_q.

3.3 Collocation methods

Collocation methods are based on the idea that an approximate solution to a boundary- or initial-value problem can be obtained by enforcing the underlying equations at suitably chosen points in the domain and/or on the boundary. Starting from the general weighted-residual form given in (3.3), assume, without loss of generality, that all boundary conditions except those on the region Γ_q are explicitly satisfied by the admissible functions u_h, and obtain the reduced form

$$\int_\Omega w_\Omega (A[u] \ - \ f)\, d\Omega \ + \ \int_{\Gamma_q} w_q (B[u] \ - \ g)\, d\Gamma \ = \ 0 \ , \tag{3.33}$$

for arbitrary functions w_Ω on Ω and w_q on Γ_q. A finite-dimensional admissible field for u_h can be constructed according to

$$u(\mathbf{x}) \doteq u_h(\mathbf{x}) \ = \ \sum_{I=1}^{N} \alpha_I \, \varphi_I(\mathbf{x}) \ + \ \varphi_0(\mathbf{x}) \ , \tag{3.34}$$

with $\varphi_I(\mathbf{x}) = 0$, $I = 1, \ldots, N$, everywhere on $\partial\Omega$ except on Γ_q and with $\varphi_0(\mathbf{x})$ chosen to enforce the boundary conditions in the same region.

3.3.1 Point-collocation method

First, identify n interior points in Ω with coordinates $\mathbf{x}_i$, $i = 1, \ldots, n$, and $N - n$ boundary points on Γ_q with coordinates $\mathbf{x}_i$, $i = n+1, \ldots, N$. These are referred to as *domain* and *boundary collocation points*, respectively, and are shown schematically in Figure 3.4.

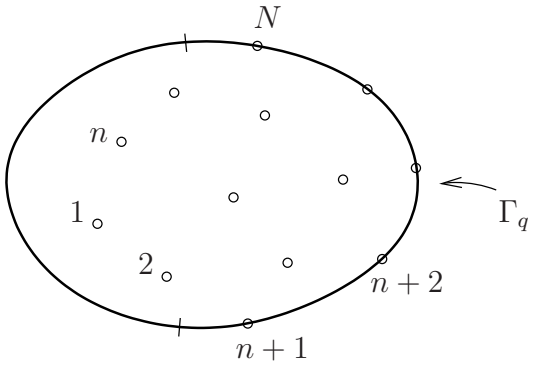


Figure 3.4. *The point-collocation method*

Next, the interior and boundary weighting functions are respectively defined according to

$$w_\Omega(\mathbf{x}) \doteq w_{\Omega h}(\mathbf{x}) = \sum_{i=1}^{n} \beta_i \, \delta(\mathbf{x} - \mathbf{x}_i) \tag{3.35}$$

and

$$w_q(\mathbf{x}) \doteq w_{qh}(\mathbf{x}) = \rho^2 \sum_{i=n+1}^{N} \beta_i \, \delta(\mathbf{x} - \mathbf{x}_i) \ , \tag{3.36}$$

where the scalar parameter ρ^2 is introduced in w_{qh} for unit consistency. Substitution of (3.34-3.36) into the weak form (3.33) yields

$$\sum_{i=1}^{n} \beta_i \left(A[\sum_{I=1}^{N} \alpha_I \, \varphi_I(\mathbf{x}_i) + \varphi_0(\mathbf{x}_i)] - f(\mathbf{x}_i) \right) + \rho^2 \sum_{i=n+1}^{N} \beta_i \left(B[\sum_{I=1}^{N} \alpha_I \, \varphi_I(\mathbf{x}_i) + \varphi_0(\mathbf{x}_i)] - g(\mathbf{x}_i) \right) = 0 \ . \tag{3.37}$$

Recalling that A and B are linear in u, equation (3.37) are rewritten as

$$\sum_{i=1}^{n} \beta_i \left(\sum_{I=1}^{N} \alpha_I A[\varphi_I(\mathbf{x}_i)] + A[\varphi_0(\mathbf{x}_i)] - f(\mathbf{x}_i) \right) + \rho^2 \sum_{i=n+1}^{N} \beta_i \left(\sum_{I=1}^{N} \alpha_I B[\varphi_I(\mathbf{x}_i)] + B[\varphi_0(\mathbf{x}_i)] - g(\mathbf{x}_i) \right) = 0 \ . \tag{3.38}$$

Since the parameters β_i are arbitrary, the above scalar equation results in a system of N linear algebraic equations of the form

$$\sum_{I=1}^{N} K_{iI}\,\alpha_I \;=\; F_i \quad , \quad i = 1, \ldots, N \;, \tag{3.39}$$

where

$$K_{iI} \;=\; \begin{cases} A[\varphi_I(\mathbf{x}_i)] \quad , & 1 \le i \le n \;, \\ \rho^2 B[\varphi_I(\mathbf{x}_i)] \quad , & n+1 \le i \le N \end{cases} \quad , \quad I = 1, \ldots, N \;, \tag{3.40}$$

and

$$F_i \;=\; \begin{cases} -A[\varphi_0(\mathbf{x}_i)] + f(\mathbf{x}_i) \quad , & 1 \le i \le n \;, \\ -\rho^2\big(B[\varphi_0(\mathbf{x}_i)] - g(\mathbf{x}_i)\big) \quad , & n+1 \le i \le N \end{cases} \;. \tag{3.41}$$

These equations may be solved for the parameters α_I, so that the approximate solution u_h is obtained from (3.34).

The particular choice of admissible fields renders the integrals in (3.33) well-defined, since products of Dirac-delta functions (from w_h) and smooth functions (from u_h) are always properly integrable.

Example 3.3.1: Point-collocation method for two-dimensional problem
Consider the partial differential equation

$$\begin{aligned} \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} \;&=\; -1 \qquad \text{in } \Omega = \{(x_1, x_2) \quad | \quad |x_1| \le 1 \;,\; |x_2| \le 1\} \;, \\ \frac{\partial u}{\partial n} \;&=\; 0 \qquad \text{on } \partial\Omega \;. \end{aligned}$$

The domain of the problem is sketched in Figure 3.5. It is immediately concluded that the boundary

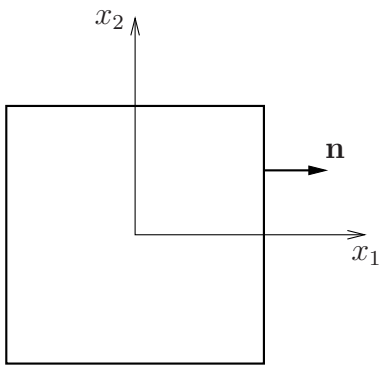


Figure 3.5. *The point collocation method in a square domain*

$\partial\Omega$ does not possess a unique outward unit normal at points $(\pm 1, \pm 1)$. It can be shown, however, that

this difficulty can be surmounted by a limiting process, thus rendering the present method of analysis valid. Further, it is easy to deduce by direct analytical means that the exact solution to this problem is symmetric with respect to the x_1- and x_2-axis, as well as with respect to the axes formed by the equations $x_1 = x_2$ and $x_1 = -x_2$.

The above boundary-value problem is referred to as a *Neumann problem*. It is easily concluded that the solution of this above problem is defined only to within an arbitrary constant, that is, if $u(x_1, x_2)$ is a solution, then so is $u(x_1, x_2) + c$, where c is any constant.

In order to simplify the analysis, use is made of a one-parameter space of admissible solutions which satisfies all boundary conditions. To this end, write u_h as

$$u_h(x_1, x_2) \;=\; \alpha_1(1 - x_1^2)^2(1 - x_2^2)^2 \;.$$

It is easy to show that

$$\frac{\partial^2 u_h}{\partial x_1^2} + \frac{\partial^2 u_h}{\partial x_2^2} \;=\; -4\alpha_1\big[(1 - 3x_1^2)(1 - x_2^2)^2 + (1 - x_1^2)^2(1 - 3x_2^2)\big] \;. \qquad (\dagger)$$

Noting that the solution should be symmetric with respect to axes $x_1 = 0$ and $x_2 = 0$, pick the single interior collocation point to be located at the intersection of these axes, namely at $(0, 0)$. It follows from $(\dagger)$ that

$$K_{11}\,\alpha_1 \;=\; F_1 \;,$$

where $K_{11} = -8$ and $F = -1$, so that $\alpha_1 = \frac{1}{8}$ and the approximate solution is

$$u_h \;=\; \frac{1}{8}(1 - x_1^2)^2(1 - x_2^2)^2 \;.$$

Alternatively, one may choose to start with a one-parameter space of admissible solutions which satisfies the domain equation everywhere, and enforce the boundary conditions at a single point on the boundary. For example, let

$$u_h(x_1, x_2) \;=\; -\frac{1}{4}(x_1^2 + x_2^2) + \alpha_1(x_1^4 + x_2^4 - 6x_1^2x_2^2) \;,$$

and choose to satisfy the boundary condition at point $(1, 0)$ (thus, due to symmetry, also at point $(-1, 0)$). It follows that

$$\frac{\partial u_h}{\partial n}(1, 0) \;=\; \frac{\partial u_h}{\partial x_1}(1, 0) \;=\; -\frac{1}{2} + 4\alpha_1 \;=\; 0 \;,$$

hence $\alpha_1 = \frac{1}{8}$, and

$$u_h(x_1, x_2) \;=\; -\frac{1}{4}(x_1^2 + x_2^2) + \frac{1}{8}(x_1^4 + x_2^4 - 6x_1^2x_2^2) \;.$$

A combined domain and boundary point collocation solution can be obtained by starting with a two-parameter approximation function

$$u_h(x_1, x_2) \;=\; \alpha_1(x_1^2 + x_2^2) + \alpha_2(1 - x_1^2)(1 - x_2^2)$$

and selecting one interior and one boundary collocation point. In particular, taking $(0, 0)$ to be the interior collocation point leads to the algebraic equation

$$\alpha_1 - \alpha_2 \;=\; -1/4 \;.$$

Subsequently, choosing $(1,0)$ as the boundary collocation point yields

$$\alpha_1 - \alpha_2 \;=\; 0 \;.$$

Clearly the system of the preceding two equations is singular, which means here that the two collocations points, in effect, generate conflicting restrictions for the two-parameter approximation function. In such a case, one may choose an alternative boundary collocation point, e.g., $(1, 1/\sqrt{2})$, which results in the equation

$$2\alpha_1 - \alpha_2 \;=\; 0 \;,$$

which, when solved simultaneously with the equation obtained from interior collocation, leads to

$$\alpha_1 \;=\; 1/4 \qquad , \qquad \alpha_2 \;=\; 1/2 \;,$$

hence,

$$u_h(x_1,x_2) \;=\; \frac{1}{4}(x_1^2+x_2^2) + \frac{1}{2}(1-x_1^2)(1-x_2^2) \;.$$

◀

3.3.2 Subdomain-collocation method

A generalization of the point-collocation method is obtained as follows: let Ω_i, $i = 1,\ldots,n$, and $\Gamma_{q,i}$, $i = n+1,\ldots,N$, be mutually disjoint connected subsets of the domain Ω and the boundary Γ_q, respectively, as in Figure 3.6. It follows that

$$\overline{\bigcup_{i=1}^{n} \Omega_i} \;\subset\; \Omega \tag{3.42}$$

and

$$\overline{\bigcup_{i=n+1}^{N} \Gamma_{q,i}} \;\subset\; \Gamma_q \;. \tag{3.43}$$

Recall the weighted residual form (3.33) and define the weighting function on Ω as

$$w_\Omega(\mathbf{x}) \;\doteq\; w_{\Omega h}(\mathbf{x}) \;=\; \sum_{i=1}^{n} \beta_i w_{\Omega,i}(\mathbf{x}) \;, \tag{3.44}$$

with

$$w_{\Omega,i}(\mathbf{x}) \;=\; \begin{cases} 1 & \text{if } \mathbf{x} \in \Omega_i \\ 0 & \text{otherwise} \end{cases} \;. \tag{3.45}$$

Similarly, write on Γ_q

$$w_q(\mathbf{x}) \;\doteq\; w_{qh}(\mathbf{x}) \;=\; \sum_{i=n+1}^{N} \beta_i w_{q,i}(\mathbf{x}) \tag{3.46}$$

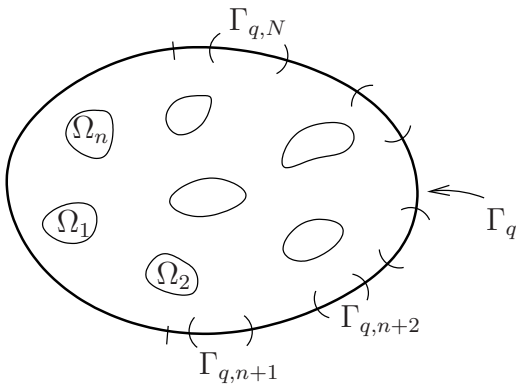


Figure 3.6. *The subdomain-collocation method*

with

$$w_{q,i}(\mathbf{x}) = \begin{cases} \rho^2 & \text{if } \mathbf{x} \in \Gamma_{q,i} \ , \\ 0 & \text{otherwise} \end{cases} \quad . \tag{3.47}$$

Given the above weighting functions, the weighted-residual form (3.33) is rewritten as

$$\sum_{i=1}^{n} \int_{\Omega_i} \beta_i (A[u] \ - \ f) \, d\Omega \ + \ \sum_{i=n+1}^{N} \rho^2 \int_{\Gamma_{q,i}} \beta_i (B[u] \ - \ g) \, d\Gamma \ = \ 0 \ . \tag{3.48}$$

Substitution of u_h from (3.34) into the above weak form yields

$$\sum_{i=1}^{n} \beta_i \int_{\Omega_i} \big(A[\sum_{I=1}^{N} \alpha_I \, \varphi_I(\mathbf{x}) \ + \ \varphi_0(\mathbf{x})] \ - \ f\big) \, d\Omega$$
$$+ \ \rho^2 \sum_{i=n+1}^{N} \beta_i \int_{\Gamma_{q,i}} \big(B[\sum_{I=1}^{N} \alpha_I \varphi_I(\mathbf{x}) \ + \ \varphi_0(\mathbf{x})] \ - \ g\big) \, d\Gamma \ = \ 0 \ . \tag{3.49}$$

Invoking the linearity of A and B in u, the above equation can be also written as

$$\sum_{i=1}^{n} \beta_i \int_{\Omega_i} \big(\sum_{I=1}^{N} \alpha_I A[\varphi_I(\mathbf{x})] \ + \ A[\varphi_0(\mathbf{x})] \ - \ f\big) \, d\Omega$$
$$+ \ \rho^2 \sum_{i=n+1}^{N} \beta_i \int_{\Gamma_{q,i}} \big(\sum_{I=1}^{N} \alpha_I B[\varphi_I(\mathbf{x})] \ + \ B[\varphi_0(\mathbf{x})] \ - \ g\big) \, d\Gamma \ = \ 0 \ , \tag{3.50}$$

from where it can be concluded that, since β_i are arbitrary,

$$\sum_{I=1}^{N} K_{iI} \, \alpha_I \ = \ F_i \quad , \quad i = 1, \ldots, N \ , \tag{3.51}$$

where

$$K_{iI} \;=\; \begin{cases} \displaystyle\int_{\Omega_i} A[\varphi_I(\mathbf{x})]\, d\Omega \quad , & 1 \le i \le n \\ \rho^2 \displaystyle\int_{\Gamma_{q,i}} B[\varphi_I(\mathbf{x})]\, d\Gamma \quad , & n+1 \le i \le N \end{cases} \quad , \quad I = 1, \ldots, N \; , \tag{3.52}$$

and

$$F_i \;=\; \begin{cases} \displaystyle\int_{\Omega_i} \big(-A[\varphi_0(\mathbf{x})] + f(\mathbf{x})\big)\, d\Omega \quad , & 1 \le i \le n \\ \rho^2 \displaystyle\int_{\Gamma_{q,i}} \big(-B[\varphi_0(\mathbf{x})] + g(\mathbf{x})\big)\, d\Gamma \quad , & n+1 \le i \le N \end{cases} \; . \tag{3.53}$$

Remarks:

- The point-collocation method requires very modest computational effort to form the stiffness matrix and the forcing vector, as it involves no integration.
- Collocation methods generally lead to unsymmetric stiffness matrices.
- The choice of collocation points is generally not arbitrary. In fact, for certain classes of differential equations, one may identify collocation points that yield optimal accuracy of the approximate solution.
- The collocation method does not readily lend itself to a general algorithmic implementation, due to the need to select collocation points in problem-specific fashion.

3.4 Least-squares methods

Consider again the model problem (3.33) of the previous section, and assuming that all boundary conditions except for those on Γ_q are satisfied by the choice of the admissible solutions, form the "least-squares" functional $I[u]$, defined as

$$I[u] = \int_\Omega (A[u] \;-\; f)^2 \, d\Omega \;+\; \rho^2 \int_{\Gamma_q} (B[u] - g)^2 \, d\Gamma \; , \tag{3.54}$$

where ρ is a consistency parameter. Clearly, the functional is non-negative and attains an absolute minimum value of zero only at the solution of the corresponding strong form of the problem. In order to find the extrema of the functional defined in (3.54), determine its first variation as

$$\delta I[u] \;=\; 2 \int_\Omega (A[u] \;-\; f)\, \delta(A[u] \;-\; f)\, d\Omega \;+\; 2\rho^2 \int_{\Gamma_q} (B[u] \;-\; g)\, \delta(B[u] \;-\; g)\, d\Gamma \; . \tag{3.55}$$

Since A and B are linear in u, it is easily seen that

$$\begin{aligned} \delta A[u] &= \lim_{\omega\to 0} \frac{A[u + \omega\,\delta u] - A[u]}{\omega} \\ &= \lim_{\omega\to 0} \frac{A[u] + \omega A[\delta u] - A[u]}{\omega} = A[\delta u] \end{aligned} \tag{3.56}$$

and, likewise, $\delta B[u] = B[\delta u]$. Consequently, extremization of $I[u]$ requires that

$$\int_\Omega A[\delta u](A[u] - f)\,d\Omega + \rho^2 \int_{\Gamma_q} B[\delta u](B[u] - g)\,d\Gamma = 0\,. \tag{3.57}$$

At this stage, introduce the finite-dimensional approximation for u_h as in (3.34), and, in addition, write

$$\delta u(\mathbf{x}) \doteq \delta u_h(\mathbf{x}) = \sum_{I=1}^{N} \delta\alpha_I\,\varphi_I(\mathbf{x})\,, \tag{3.58}$$

with $\delta\alpha_I$, $I = 1,\ldots,N$, being arbitrary scalar parameters. Substituting u_h and δu_h in (3.57) results in

$$\begin{aligned} &\int_\Omega A[\sum_{I=1}^{N} \delta\alpha_I\,\varphi_I(\mathbf{x})]\big(A[\sum_{J=1}^{N} \alpha_J\,\varphi_J(\mathbf{x}) + \varphi_0(\mathbf{x})] - f\big)\,d\Omega \\ &\qquad + \rho^2 \int_{\Gamma_q} B[\sum_{I=1}^{N} \delta\alpha_I\,\varphi_I(\mathbf{x})]\big(B[\sum_{J=1}^{N} \alpha_J\,\varphi_J(\mathbf{x}) + \varphi_0(\mathbf{x})] - g\big)\,d\Gamma = 0\,. \end{aligned} \tag{3.59}$$

Again, since A and B are linear in u, it follows that the above equation can be also written as

$$\begin{aligned} &\int_\Omega \sum_{I=1}^{N} \delta\alpha_I A[\varphi_I(\mathbf{x})] \left(\sum_{J=1}^{N} \alpha_J A[\varphi_J(\mathbf{x})] + A[\varphi_0(\mathbf{x})] - f\right) d\Omega \\ &\qquad + \rho^2 \int_{\Gamma_q} \sum_{I=1}^{N} \delta\alpha_I B[\varphi_I(\mathbf{x})] \left(\sum_{J=1}^{N} \alpha_J B[\varphi_J(\mathbf{x})] + B[\varphi_0(\mathbf{x})] - g\right) d\Gamma = 0\,. \end{aligned} \tag{3.60}$$

Invoking the arbitrariness of $\delta\alpha_I$, this gives rise to a system of linear algebraic equations of the form

$$\sum_{J=1}^{N} K_{IJ}\,\alpha_J = F_I \quad , \quad I = 1,\ldots,N\,, \tag{3.61}$$

where

$$K_{IJ} = \int_\Omega A[\varphi_I]A[\varphi_J]\,d\Omega + \rho^2 \int_{\Gamma_q} B[\varphi_I]B[\varphi_J]\,d\Gamma \quad , \quad I,J = 1,\ldots,N \tag{3.62}$$

and

$$F_I \;=\; \int_\Omega A[\varphi_I]\big(f \;-\; A[\varphi_0]\big)\,d\Omega \;+\; \rho^2 \int_{\Gamma_q} B[\varphi_I]\big(g \;-\; B[\varphi_0]\big)\,d\Gamma \quad , \quad I = 1, \ldots, N \;. \tag{3.63}$$

It is important to note that the smoothness requirements for the admissible functions u are governed by the integrals that appear in (3.54). It can be easily deduced that if A is a differential operator of, say, second order (namely, maps functions u to partial derivatives of second order), then for (3.54) to be well-defined, it is necessary that $u \in H^2(\Omega)$. This requirement can be contrasted to the one obtained in the Galerkin approximation of (3.5), where it was concluded that both u and w need only belong to $H^1(\Omega)$.

Remarks:

- The stiffness matrix that emanates from the least-squares functional is symmetric by construction and, may be positive-definite, conditional upon the particular form of the boundary conditions.
- A slightly more general weighted-residual formulation of the least-squares problem based directly on (3.33) is recovered as follows by choosing $w_\Omega = A[w]$ and $w_q = B[w]$. Then, the weak form in (3.57) is reproduced, where w appears in place of δu.

Example 3.4.1: Least-squares method for one-dimensional Laplace-Poisson equation

Consider again the differential equation in Example 3.2.1 and assume a quadratic polynomial solution of the form

$$u_h(x) \;=\; \alpha_1 x + \alpha_2 x^2 \;, \tag{3.64}$$

which clearly satisfies the homogeneous Dirichlet boundary condition at $x = 0$. Taking into account (3.54), the least-squares functional is written as

$$I[u] \;=\; \int_0^1 \left(\frac{d^2u}{dx^2} - 1\right)^2 dx + \rho^2 \left(\frac{du}{dx} + 2\right)^2 \Bigg|_{x=1}$$

or, upon substituting the expression for u_h from (3.64),

$$\begin{aligned} I(\alpha_1, \alpha_2) \;&=\; \int_0^1 (2\alpha_2 - 1)^2\,dx + \rho^2(\alpha_1 + 2\alpha_2)^2 \\ &=\; (2\alpha_2 - 1)^2 + \rho^2(\alpha_1 + 2\alpha_2 + 2)^2 \;, \end{aligned}$$

It is clear, by inspection, that $I(\alpha_1, \alpha_2)$ attains a (global) minimum when

$$\begin{aligned} 2\alpha_2 - 1 \;&=\; 0 \\ \alpha_1 + 2\alpha_2 + 2) \;&=\; 0 \;, \end{aligned}$$

which imply that $\alpha_1 = -3$ and $\alpha_2 = 1/2$. This results in the approximate solution

$$u_h(x) \;=\; -3x + \frac{1}{2}x^2 \;,$$

which is also the exact solution of the boundary-value problem.

3.5 Composite methods

The Galerkin, collocation and least-squares methods can be appropriately combined to yield composite weighted residual methods. The choice of admissible weighting functions defines the degree and form of blending between the above methods. Without attempting to provide an exhaustive presentation, note that a typical composite Galerkin and collocation method can be obtained for the problem (3.33) by defining the admissible solutions as in (3.34) and setting

$$w_\Omega(\mathbf{x}) \;\doteq\; w_{\Omega h}(\mathbf{x}) \;=\; \sum_{I=1}^{m} \beta_I \, \psi_I(\mathbf{x}) \;+\; \sum_{I=m+1}^{n} \beta_I \, \rho_1^2 \delta(\mathbf{x} \;-\; \mathbf{x}_I) \;, \tag{3.65}$$

where ψ_I, $I = 1, \ldots, N$ vanish on $\partial\Omega \setminus \Gamma_q$ and ρ_1 is a scaling factor. In addition, on Γ_q,

$$w_q(\mathbf{x}) \;\doteq\; w_{qh}(\mathbf{x}) \;=\; \sum_{I=1}^{m} \beta_I \, \psi_I(\mathbf{x}) \;+\; \sum_{I=n+1}^{N} \beta_I \, \rho_2^2 \delta(\mathbf{x} \;-\; \mathbf{x}_I) \;, \tag{3.66}$$

where, again, ρ_2 is a scaling factor. This composite method combines m smooth weighting functions, $n - m$ domain collocation points and $N - n$ boundary collocation points.

A typical composite collocation and least-squares method can be similarly obtained by defining the domain and boundary weighting functions according to

$$w_\Omega(\mathbf{x}) \;\doteq\; w_{\Omega h}(\mathbf{x}) \;=\; \sum_{I=1}^{m} \beta_I \, A[\varphi_I(\mathbf{x})] \;+\; \sum_{I=m+1}^{n} \beta_I \, \rho_1^2 \delta(\mathbf{x} \;-\; \mathbf{x}_I) \tag{3.67}$$

and

$$w_q(\mathbf{x}) \;\doteq\; w_{qh}(\mathbf{x}) \;=\; \sum_{I=1}^{m} \beta_I \, B[\varphi_I(\mathbf{x})] \;+\; \sum_{I=n+1}^{N} \beta_I \, \rho_2^2 \delta(\mathbf{x} \;-\; \mathbf{x}_I) \;, \tag{3.68}$$

respectively, where, again, φ_I, $I = 1, \ldots, N$, vanish on $\partial\Omega \setminus \Gamma_q$ and ρ_1, ρ_2 are scaling factors.

Composite weighted residual methods are used for differential equations of mixed type or in cases where high accuracy is desired in localized regions of the domain Ω or boundary $\partial\Omega$.

3.6 An interpretation of finite difference methods

Finite difference methods can be interpreted as special cases of weighted residuals methods. In particular, the difference operators can be viewed as differential operators over appropriately chosen polynomial spaces. As a demonstration of this interpretation, consider the boundary-value problem (1.1), and let grid points x_i, $i = 1, \ldots, N$, be chosen in the interior of the domain $(0, L)$, as in Section 1.2.2. The system of equations

$$\begin{aligned} u_2 - 2u_1 &= \frac{f_1 \Delta x^2}{k} - u_0 \ , \\ u_{l+1} - 2u_l + u_{l-1} &= \frac{f_l \Delta x^2}{k} \quad , \quad l = 2, \ldots, N-1 \ , \\ -2u_N + u_{N-1} &= \frac{f_N \Delta x^2}{k} - u_L \end{aligned} \tag{3.69}$$

is obtained by applying the centered-difference operator

$$\frac{d^2 u}{dx^2} \doteq \frac{u_{l+1} - 2u_l + u_{l-1}}{\Delta x^2} \tag{3.70}$$

at all interior points. Note that in the above equations $(\cdot)_l = (\cdot)(x_l)$.

In order to analyze the above finite-difference approximation, consider the domain-based weighted-residual form

$$\int_0^L w \left(k \frac{d^2 u}{dx^2} - f \right) dx = 0 \ , \tag{3.71}$$

where boundary conditions $(1.1)_{2,3}$ are assumed to hold at the outset. Subsequently, define the approximate solution u_h *within each sub-domain* $(x_l - \frac{\Delta x}{2}, x_l + \frac{\Delta x}{2}]$, $l = 2, \ldots, N-1$, as

$$u_h(x) = \sum_{i=l-1}^{l+1} \alpha_i N_i(x) \ , \tag{3.72}$$

where N_i are polynomials of degree 2 defined as

$$\begin{aligned} N_{l-1}(x) &= \frac{(x - x_l)(x - x_{l+1})}{2\Delta x^2} \ , \\ N_l(x) &= -\frac{(x - x_{l-1})(x - x_{l+1})}{\Delta x^2} \ , \\ N_{l+1}(x) &= \frac{(x - x_{l-1})(x - x_l)}{2\Delta x^2} \ . \end{aligned} \tag{3.73}$$

Figure 3.7 illustrates the three interpolation functions in the representative sub-domain. It can be readily concluded from (3.72) and (3.73) that $u_h(x_i) = \alpha_i$, $i = l-1, l, l+1$, which

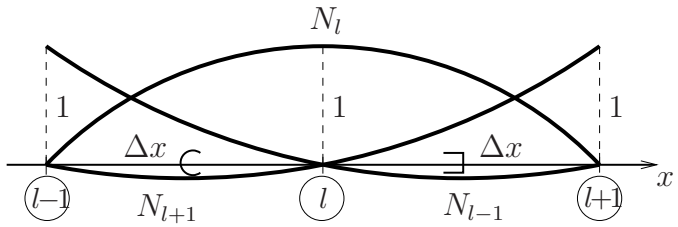


Figure 3.7. *Polynomial interpolation functions used for region* $(x_l - \frac{\Delta x}{2}, x_l + \frac{\Delta x}{2}]$ *in the weighted-residual interpretation of the finite difference method*

means that the parameters $\alpha_{l-1}, \alpha_l, \alpha_{l+1}$ can be interpreted as the values of the dependent variable u at $x = x_{l-1}, x_l, x_{l+1}$, respectively. This means that the approximation u_h in the domain $(x_l - \frac{\Delta x}{2}, x_l + \frac{\Delta x}{2}]$ is

$$u_h(x) \;=\; \sum_{i=l-1}^{l+1} N_i(x) u_i \;. \tag{3.74}$$

Likewise, the function u_h is given in the sub-domains $[0, x_1 + \frac{\Delta x}{2}]$ and $(x_N - \frac{\Delta x}{2}, L]$ by

$$\begin{aligned} u_h(x) \;&=\; N_0(x) u_0 + \sum_{i=1}^{2} N_i(x) u_i \;, \\ u_h(x) \;&=\; \sum_{i=N-1}^{N} N_i(x) u_i + N_{N+1}(x) u_L \;, \end{aligned} \tag{3.75}$$

respectively, so that the boundary conditions are satisfied at both end-points, see Figure 3.8.

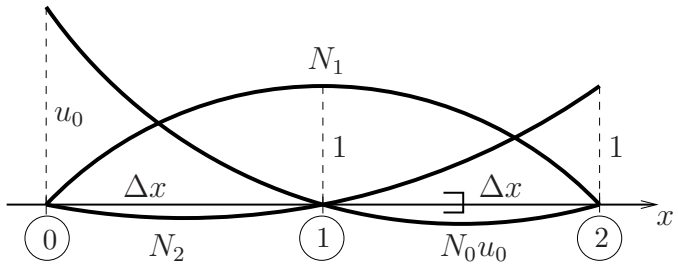


Figure 3.8. *Polynomial interpolation functions used for region* $[0, x_1 + \frac{\Delta x}{2}]$ *in the weighted-residual interpretation of the finite difference method*

The approximation w_h over the full domain $(0, L)$ is taken in the form

$$w_h \;=\; \sum_{l=1}^{N} \beta_l \, \delta(x - x_l) \;, \tag{3.76}$$

so that equation (3.71) is written as

$$\sum_{l=1}^{N} \beta_l \left(k \frac{d^2 u_h}{dx^2}(x_l) - f(x_l) \right) \;=\; 0 \; . \tag{3.77}$$

It follows from (3.73) and (3.74) that in the representative sub-domain $(x_l - \frac{\Delta x}{2}, x_l + \frac{\Delta x}{2}]$

$$\frac{d^2 u_h}{dx^2} \;=\; \frac{u_{l-1}}{\Delta x^2} - 2\frac{u_l}{\Delta x^2} + \frac{u_{l+1}}{\Delta x^2} \; . \tag{3.78}$$

This, in turn, implies that at $x = x_l$

$$\frac{k}{\Delta x^2}(u_{l-1} - 2u_l + u_{l+1}) - f_l \;=\; 0 \; , \tag{3.79}$$

owing to the arbitrariness of parameters β_l, $l = 1, \ldots, N$, in (3.77). Similarly, in sub-domain $[0, x_1 + \frac{\Delta x}{2}]$,

$$\frac{k}{\Delta x^2}(u_0 - 2u_1 + u_2) - f_1 \;=\; 0 \; , \tag{3.80}$$

and, in sub-domain $(x_N - \frac{\Delta x}{2}, L]$,

$$\frac{k}{\Delta x^2}(u_{N-1} - 2u_N + u_L) - f_N \;=\; 0 \; . \tag{3.81}$$

Thus, the finite difference equations are recovered exactly at all interior grid points.

Remark:

- ☛ The choice of admissible fields $\mathcal{U}_h$ and $\mathcal{W}_h$ in the preceding analysis is mathematically appropriate, since the integral in equation (3.71) is always well-defined.

It is instructive at this point to review a finite element solution of the same problem (1.1) starting from a Bubnov-Galerkin weighted-residual formulation. To this end, assume that the interpolation functions in (x_{l-1}, x_{l+1}), $l = 1, 2, \ldots, N$, are continuous piecewise-linear polynomials, namely

$$\begin{aligned}
\varphi_{l-1}(x) \;&=\; \begin{cases} -\dfrac{x}{\Delta x} & , \quad x < 0 \\ 0 & , \quad x \geq 0 \end{cases} \; , \\
\varphi_{l}(x) \;&=\; \begin{cases} 1 + \dfrac{x}{\Delta x} & , \quad x < 0 \\ 1 - \dfrac{x}{\Delta x} & , \quad x \geq 0 \end{cases} \; , \\
\varphi_{l+1}(x) \;&=\; \begin{cases} 0 & , \quad x < 0 \\ \dfrac{x}{\Delta x} & , \quad x \geq 0 \end{cases} \; ,
\end{aligned} \tag{3.82}$$

see Figure 3.9. Note that here the origin of the coordinate system x is shifted to point l from point 0, without any loss of generality. Then, letting

$$u_h \;=\; \sum_{i=l-1}^{l+1} \varphi_i(x) u_i \quad , \quad w_h \;=\; \sum_{i=l-1}^{l+1} \varphi_i(x) w_i \tag{3.83}$$

in (x_{l-1}, x_{l+1}), $l = 1, 2, \ldots, N$, it follows that

$$\frac{du_h}{dx} \;=\; \begin{cases} \dfrac{u_l - u_{l-1}}{\Delta x} & , \quad x < 0 \\ \dfrac{u_{l+1} - u_l}{\Delta x} & , \quad x \geq 0 \end{cases} \tag{3.84}$$

and, likewise,

$$\frac{dw_h}{dx} \;=\; \begin{cases} \dfrac{w_l - w_{l-1}}{\Delta x} & , \quad x < 0 \\ \dfrac{w_{l+1} - w_l}{\Delta x} & , \quad x \geq 0 \end{cases} \quad . \tag{3.85}$$

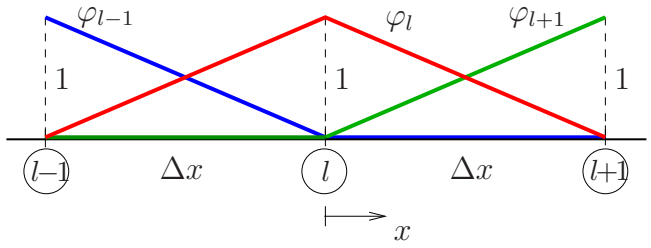


Figure 3.9. *Interpolation functions for a finite element approximation of a one-dimensional two-cell domain*

Neglecting any terms stemming from boundary-like conditions at x_{l-1} and x_{l+1} (which is tantamount to considering Dirichlet boundary conditions at $x = -\Delta x$ and $x = \Delta x$), one may write the weak form form the domain $(-\Delta x, \Delta x)$ as

$$\int_{-\Delta x}^{\Delta x} \left(\frac{dw_h}{dx} k \frac{du_h}{dx} + w_h f \right) dx \;=\; 0 \;, \tag{3.86}$$

which, upon substituting (3.84) and (3.85) into (3.86), becomes

$$\begin{aligned} &\int_{-\Delta x}^{0} \left[\frac{(w_l - w_{l-1}) k (u_l - u_{l-1})}{\Delta x^2} + \left\{ -\frac{x}{\Delta x} w_{l-1} + \left(1 + \frac{x}{\Delta x} \right) w_l \right\} f \right] dx \\ &\quad + \int_{0}^{\Delta x} \left[\frac{(w_{l+1} - w_l) k (u_{l+1} - u_l)}{\Delta x^2} + \left\{ \left((1 - \frac{x}{\Delta x} \right) w_l + \frac{x}{\Delta x} w_{l+1} \right\} f \right] dx \;=\; 0 \;. \end{aligned} \tag{3.87}$$

Setting $w_{l-1} = w_{l+1} = 0$, it follows that

$$w_l \int_{-\Delta x}^{0} \left[\frac{k}{\Delta x^2}(u_l - u_{l-1}) + (1 + \frac{x}{\Delta x}) f \right] dx + w_l \int_{0}^{\Delta x} \left[-\frac{k}{\Delta x^2}(u_{l+1} - u_l) + (1 - \frac{x}{\Delta x}) f \right] dx \;=\; 0 \; . \quad (3.88)$$

Next, define the force f_l according to

$$\Delta x f_l \;=\; \int_{-\Delta x}^{0} (1 + \frac{x}{\Delta x}) f \, dx + \int_{0}^{\Delta x} (1 - \frac{x}{\Delta x}) f \, dx \; . \quad (3.89)$$

Subsequently, recalling that w_l is arbitrary and integrating (3.88) leads again to the equations in (3.79). A corresponding analysis may be performed for the interpolations in the domains $(0, x_2)$ and (x_{N-1}, L), which can be shown to readily recover equations (3.80) and (3.81), respectively. The preceding findings imply that, upon using the definition of force terms in (3.89), the finite element solution of (1.1) with piecewise linear polynomial interpolations coincides with the finite difference solution that uses the classical difference operator (1.2). The latter can be equivalently thought of as a weighted residual method with piecewise quadratic approximations for the dependent variable u and Dirac-delta function approximations for the corresponding weighting functions. This observation is consistent with the findings in Sections 1.2.2 and 1.2.3.

The traditional distinction between the finite difference and the finite element method is summarized by noting that finite differences approximate differential operators by (algebraic) difference operators which apply on admissible fields $\mathcal{U}$, whereas finite elements use the exact differential operators which apply only on subspaces of these admissible fields. The weighted-residual framework allows for a unified interpretation of both methods.

3.7 Exercises

Problem 1

Consider the boundary-value problem

$$\begin{aligned} \frac{d^2u}{dx^2} + u + x &= 0 \quad \text{in } \Omega = (0,1) \; , \\ u &= 1 \quad \text{on } \Gamma_u = \{0\} \; , \\ \frac{du}{dx} &= 0 \quad \text{on } \Gamma_q = \{1\} \; . \end{aligned}$$

Assume a general three-parameter polynomial approximation to the exact solution, in the form

$$u_h(x) \;=\; \alpha_0 + \alpha_1 x + \alpha_2 x^2 \; . \quad (\dagger)$$

(a) Place a restriction on parameters α_i by enforcing *only* the Dirichlet boundary condition, and obtain a *Bubnov-Galerkin* approximation of the solution.

(b) Starting from the general quadratic form of u_h in (†), place a restriction on parameters α_i as in part (a), and determine a *Petrov-Galerkin* approximation of the solution assuming

$$w_h(x) \;=\; \beta_1\psi_1(x) \;+\; \beta_2\psi_2(x)\;,$$

where functions $\psi_1(x)$ and $\psi_2(x)$ are defined as

$$\psi_1(x) \;=\; \begin{cases} 0 & \text{for } 0 \le x \le \frac{1}{2} \\ 1 & \text{for } \frac{1}{2} < x \le 1 \end{cases}\;,$$

$$\psi_2(x) \;=\; \begin{cases} x & \text{for } 0 \le x \le \frac{1}{2} \\ 0 & \text{for } \frac{1}{2} < x \le 1 \end{cases}\;.$$

Clearly justify the admissibility of w_h for the proposed approximation.

(c) Starting again from the general quadratic form of u_h in (†), enforce *all* boundary conditions on u_h and uniquely determine a one-parameter *point-collocation* approximation.

Problem 2

The stiffness matrix $\boldsymbol{K} = [K_{IJ}]$ emanating from a Bubnov-Galerkin approximation of the boundary-value problem

$$\begin{aligned} \frac{\partial}{\partial x_1}\Big(k\frac{\partial u}{\partial x_1}\Big) \;+\; \frac{\partial}{\partial x_2}\Big(k\frac{\partial u}{\partial x_2}\Big) \;&=\; f \quad \text{in } \Omega \subset \mathbb{R}^2\;, \\ u \;&=\; \bar{u} \quad \text{on } \Gamma_u\;, \\ -k\,\frac{\partial u}{\partial n} \;&=\; \bar{q} \quad \text{on } \Gamma_q \end{aligned}$$

has components K_{IJ} given by

$$K_{IJ} \;=\; \int_\Omega \left\{ \begin{array}{cc} \varphi_{I,1} & \varphi_{I,2} \end{array} \right\} k \left\{ \begin{array}{c} \varphi_{J,1} \\ \varphi_{J,2} \end{array} \right\} d\Omega\;,$$

where the approximation for u is of the general form

$$u(x_1,x_2) \;\doteq\; u_h(x_1,x_2) \;=\; \sum_{I=1}^{N} \alpha_I\varphi_I(x_1,x_2) \;+\varphi_0(x_1,x_2)\;,$$

and the functions φ_I, $I = 1,\ldots,N$, are assumed linearly independent. Show that $\boldsymbol{K}$ is positive-definite in $\mathbb{R}^N$, provided $k > 0$ and $\Gamma_u \neq \emptyset$.

Problem 3

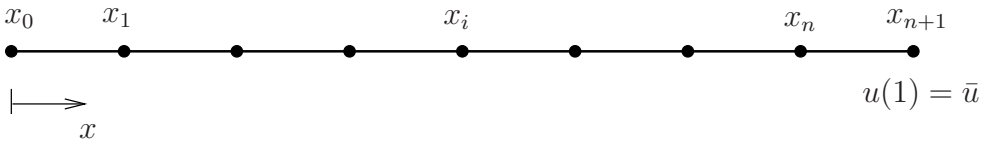
Consider the weak form of the one-dimensional Laplace equation

$$\int_\Omega k w_{,x} u_{,x}\, d\Omega \;+\; \int_\Omega w f\, d\Omega \;+\; w(0)\bar{q} \;=\; 0\;,$$

where $\Omega = (0,1)$ and $k > 0$ is a constant. Assume that the admissible fields $\mathcal{U}$ and $\mathcal{W}$ for u and w, respectively, allow the first derivative of u and w to exhibit finite jumps at points $x_i \in \Omega,\ i = 1, 2, \ldots, n$, and show that

$$\sum_{i=0}^{n} \int_{x_i}^{x_{i+1}} w(k\,u_{,xx} \ - \ f)\,d\Omega \ + \ w(0)[ku_{,x}(0) \ - \ \bar{q}] \\ + \ \sum_{i=1}^{n} w(x_i)k[u_{,x}(x_i^+) \ - \ u_{,x}(x_i^-)] \ = \ 0\ .$$

From the above equation derive the strong form of the problem. What can you conclude regarding smoothness of the exact solution across the x_i's?



Problem 4

Consider the boundary-value problem

$$\frac{\partial^2 u}{\partial x_1^2} \ + \ \frac{\partial^2 u}{\partial x_2^2} \ = \ -1 \quad \text{in} \ \ \Omega = (-1,1) \times (-1,1)\ , \\ u \ + \ \frac{\partial u}{\partial n} \ = \ 0 \quad \text{on} \ \ \partial\Omega\ ,$$

with reference to a fixed Cartesian coordinate system $x_1 - x_2$. The boundary condition on $\partial\Omega$ is referred to as a *Robin* (or *third type*) boundary condition.

(a) Conclude that the (unknown) exact solution is symmetric with respect to lines $x_1 = 0$, $x_2 = 0$, $x_1 - x_2 = 0$ and $x_1 + x_2 = 0$.

(b) Start from the general two-dimensional polynomial field which is complete up to degree 6 and show that using only the aforementioned symmetries of the exact solution, the polynomial approximation is reduced to

$$u_h \ = \ \alpha_0 \ + \ \alpha_1\,(x_1^2 \ + \ x_2^2) \ + \ \alpha_2\,x_1^2 x_2^2 \ + \ \alpha_3\,(x_1^4 \ + \ x_2^4) \\ + \ \alpha_4\,(x_1^4 x_2^2 \ + \ x_1^2 x_2^4) \ + \ \alpha_5\,(x_1^6 \ + \ x_2^6)\ ,$$

where $\alpha_i,\ i = 0, 1, ..., 5$, are arbitrary constants. Subsequently, apply the boundary conditions on u_h and, thus, place restrictions on parameters α_i.

(c) Find two non-trivial approximate solutions to the above problem by means of a one-parameter and a two-parameter interior collocation method using appropriate approximation functions from the family of functions obtained in part (b). Pick collocation points judiciously for both approximations.

Problem 5

Consider the initial-value problem

$$\begin{aligned} \frac{du}{dt} \; + \; u \;\; &= \;\; t \quad \text{in } \Omega = (0,1) \; , \\ u(0) \;\; &= \;\; 1 \; , \end{aligned}$$

and assume a general three-parameter polynomial approximation u_h written as

$$u_h(t) \;=\; \alpha_0 \;+\; \alpha_1 t \;+\; \alpha_2 t^2 \; ,$$

where α_i, $i = 0, 1, 2$, are scalar parameters to be determined.

(a) Place a restriction on u_h by directly enforcing the initial condition, and, subsequently, obtain an approximate solution to the problem using the point-collocation method. Select the collocation points judiciously.

(b) Place the same restriction on u_h as in part (a), and obtain an approximate solution to the problem using a Bubnov-Galerkin method.

Problem 6

Consider the non-linear second-order ordinary differential equation of the form

$$A[u] \;=\; f \quad \text{in } \Omega = (0,1) \; ,$$

where

$$A[u] \;=\; -2u \, \frac{d^2u}{dx^2} \;+\; \Big(\frac{du}{dx}\Big)^2$$

and

$$f \;=\; 4 \; ,$$

with boundary conditions $u(0) = 1$ and $u(1) = 0$. Find each of the approximate polynomial solutions u_h, as instructed in the problem statement, and compute the *residual* error norm $\mathcal{E}$ defined as

$$\mathcal{E}(u_h) \;=\; \|A[u_h] \;-\; f\|_{L_2} \; .$$

Compare the approximate solutions by means of $\mathcal{E}$. Comment on the results of your analysis.

Problem 7

Consider the partial differential equation

$$\frac{\partial^2 u}{\partial x^2} \;-\; \frac{\partial u}{\partial t} \;=\; 0 \quad \text{for } (x,t) \in (0,1) \times (0,T) \; , \tag{‡}$$

subject to boundary conditions

$$u(0,t) \;=\; 0 \quad \text{for } t \in (0,T) \; , \tag{††}$$

$$\frac{\partial u}{\partial x}(1,t) \;=\; 0 \quad \text{for } t \in (0,T)\ , \tag{‡‡}$$

and initial condition

$$u(x,0) \;=\; 1 \quad \text{for } x \in (0,1)\ , \tag{†‡}$$

where $T > 0$. Let a family of approximations $u_h(x,t)$ be written as

$$u_h(x,t) \;=\; \{\alpha_0 \;+\; \alpha_1 x \;+\; \alpha_2 x^2\}\,\theta(t)\ , \tag{‡†}$$

where $\theta(t)$ is a (yet unknown) function of time, and α_0, α_1 and α_2 are scalar parameters to be determined.

(a) Obtain a reduced form of $u_h(x,t)$ by enforcing boundary conditions (††) and (‡‡).

(b) Determine an initial condition $\theta(0)$ for function $\theta(t)$ by forcing the reduced form of $u_h(x,t)$ obtained in part (a) to satisfy equation (†‡) in the sense of the least-squares method.

(c) Arrive at a first-order ordinary differential equation for function $\theta(t)$ by applying to differential equation (‡) a Petrov-Galerkin method *in the spatial domain*. Use the approximation function $u_h(x,t)$ of part (a) and a weighting function $w_h(x,t)$ given by

$$w_h(x,t) \;=\; x\ .$$

Find a closed-form expression for $\theta(t)$ by solving the differential equation analytically and using the initial condition obtained in part (b).

The solution procedure outlined above, where a partial differential equation is reduced into an ordinary differential equation by an approximation of the form (‡†), is referred to as the *Kantorovich method*.

Problem 8

Consider the differential equation

$$\frac{\partial^2 u}{\partial x_1^2} \;+\; 2\,\frac{\partial^2 u}{\partial x_2^2} \;=\; -2$$

in the square domain $\Omega = \{(x_1,x_2) \quad | \quad |x_1| < 1\ ,\ |x_2| < 1\}$, with homogeneous Dirichlet boundary condition $u = 0$ everywhere on $\partial\Omega$.

(a) Reformulate the above boundary-value problem in terms of a new dependent variable v defined as

$$v \;=\; u \;+\; \frac{1}{2}x_1^2 \;+\; \frac{1}{4}x_2^2\ .$$

Verify that the resulting partial differential equation in v is homogeneous, while the boundary condition is non-homogeneous and expressed as

$$v \;=\; \bar{v} \;=\; \frac{1}{2}x_1^2 \;+\; \frac{1}{4}x_2^2 \quad \text{on } \partial\Omega\ .$$

(b) Introduce a one-parameter family of approximate solutions v_h of the form

$$v_h \;=\; \alpha\,\varphi \;,$$

where

$$\varphi \;=\; x_1^2 \;-\; \frac{1}{2}x_2^2 \;,$$

and show that v_h satisfies the homogeneous partial differential equation obtained in part (a). Subsequently, determine the scalar parameter α using a weighted-residual method on $\partial\Omega$ by requiring that

$$\int_{\partial\Omega} w_u\,(v_h \;-\; \bar{v})\,d\Gamma \;=\; 0 \;,$$

where $w_u = \beta$, and β is an arbitrary scalar.

Problem 9

Consider the boundary-value problem

$$\begin{aligned} \frac{d^2u}{dx^2} + u &= 0 \quad \text{in } \Omega = (0,\pi) \;, \\ u(0) &= 0 \;, \\ \frac{du}{dx}(\pi) &= -1 \;, \end{aligned}$$

and assume a three-parameter polynomial approximation u_h written as

$$u_h(x) \;=\; \alpha_0 + \alpha_1 x + \alpha_2 x^2 \;,$$

where α_0, α_1, and α_2 are scalar parameters to be determined.

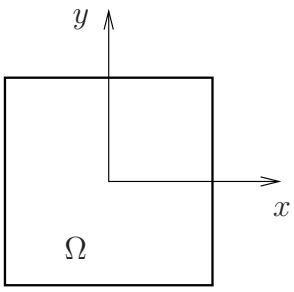
(a) Obtain a reduced form of $u_h(x)$ by directly enforcing the Dirichlet boundary condition.

(b) Starting from the reduced form of $u_h(x)$ obtained in part (a), use a composite Galerkin/collocation method to find an approximate solution to the above boundary-value problem. In particular, define the weighting function w_h to be

$$w_h(x) \;=\; \beta_1 x \;+\; \beta_2\delta(x-\pi/2) \;,$$

where β_1 and β_2 are arbitrary scalar parameters, and $\delta(x)$ denotes the Dirac-delta function.

(c) Calculate the error of the approximate solution obtained in part (b), in the sense of the L_2-norm over the domain Ω. In particular, employ the trapezoidal rule in the subdomains $[0,\pi/2)$ and $[\pi/2,\pi)$ to estimate this error.



Problem 10

Consider the partial differential equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \;=\; 1 \;,$$

in the domain $\Omega = \{(x,y) \;\mid\; -1 < x < 1, -1 < y < 1\}$ shown in the following figure, and subject to boundary condition $u = 0$ on $\partial\Omega$.

Assume that an approximate solution of the resulting boundary-value problem is sought in the form

$$u_h(x,y) \;=\; \alpha_1(1-x^2)(1-y^2) + \alpha_2(1-x)^4 + \alpha_3 xy + \alpha_4 \;,$$

where α_1, α_2, α_3 and α_4 are parameters to be determined.

(a) Without solving any equations, argue convincingly that the term $\alpha_2(1-x)^4$ should be dropped from the approximate solution.

(b) Enforce the boundary condition on the reduced form of u_h to deduce a one-parameter approximation.

(c) Starting from the one-parameter approximation of part (b), determine the solution u_h using a domain least-squares weighted residual method.

Problem 11

Consider the initial-value problem

$$\begin{aligned} \frac{du}{dt} \;-\; u \;&=\; 0 \quad \text{in } \Omega = (0,T) \;, \\ u(0) \;&=\; 1 \;, \end{aligned}$$

where T is a given positive number, and let a two-parameter polynomial approximation u_h be expressed as

$$u_h(t) \;=\; \alpha_0 \;+\; \alpha_1 t \;.$$

(a) Obtain a reduced form of $u_h(t)$ by directly enforcing the initial condition.

(b) Determine u_h as a function of t and T using a domain least-squares method in $(0,T)$.

(c) Find the limit of the approximate solution $u_h(t)$ obtained in part (b), as T approaches zero, i.e., as the domain $(0,T)$ of the analysis becomes arbitrarily small. How does the approximate solution compare with the exact solution $u = e^t$ in this limiting case?

Problem 12
Consider the boundary-value problem

$$\begin{aligned} \frac{d^2u}{dx^2} &= x \quad \text{in } \Omega = (0,1) \ , \\ u(0) &= 1 \ , \\ u'(1) &= 0 \ , \end{aligned}$$

and assume a four-parameter polynomial approximation u_h written as

$$u_h(x) \ = \ \alpha_0 \ + \ \alpha_1 x \ + \ \alpha_2 x^2 \ + \ \alpha_3 x^3 \ ,$$

where α_0, α_1, α_2 and α_3 are scalar parameters to be determined.

(a) Obtain a reduced form of $u_h(x)$ by directly enforcing both boundary conditions.

(b) Starting from the reduced form of $u_h(x)$ in part (a), use a composite Galerkin/collocation method to find an approximate solution to the above boundary-value problem. In particular, define the weighting function w_h to be

$$w_h(x) \ = \ \beta_1 x \ + \ \beta_2 \delta(x - \bar{x}) \ ,$$

where β_1 and β_2 are arbitrary scalar parameters, and $\delta(x)$ denotes the Dirac-delta function. Notice that collocation point $\bar{x} \in (0,1)$ is chosen to be any point inside the domain of the analysis. Express the approximate solution as a function of x and $\bar{x}$. What do you notice?

(c) Calculate the residual error of the approximate solution obtained in part (b), in the sense of the L_2-norm over the domain Ω.

3.8 Suggestions for further reading

Sections 3.1-3.5

[1] B.A. Finlayson and L.E. Scriven. The method of weighted residuals – a review. *Appl. Mech. Rev.*, 19:735–748, 1966. [*This is an excellent review of weighted residual methods, including a discussion of their relation to variational methods*].

[2] G.F. Carey and J.T. Oden. *Finite Elements: a Second Course*, volume II. Prentice-Hall, Englewood Cliffs, 1983. [*This volume discusses the Galerkin method in Chapter 1 and the other weighted residual methods in Chapter 4*].

[3] O.D. Kellogg. *Foundations of Potential Theory.* Dover, New York, 1953. [*Chapter IV of this book contains an excellent discussion of the divergence theorem for domains with boundaries that possess corners*].

Section 3.4

[1] P.P. Lynn and S.K. Arya. Use of the least-squares criterion in the finite element formulation. *Int. J. Num. Meth. Engr.*, 6:75–88, 1973. [*This article uses the least-squares method for the solution of the two-dimensional Laplace-Poisson equation, in conjunction with the finite element method for the construction of the admissible fields*].

Section 3.6

[1] O.C. Zienkiewicz and K. Morgan. *Finite Elements and Approximation.* Wiley, New York, 1983. [*The relation between finite element and finite difference methods is addressed in Section 3.10*].

[2] K.W. Morton. Finite Difference and Finite Element Methods. *Comp. Phys. Comm.*, 12:99-108, (1976). [*This article presents a comparison between finite difference and finite element methods from a finite difference viewpoint*].

Chapter 4

Variational Methods

4.1 Introduction to variational principles

Certain classes of partial differential equations possess a variational structure. This means that their solutions u can be interpreted as extremal points over a properly defined function space $\mathcal{U}$, with reference to given functionals $I[u]$. As a result, determining solutions of such equations is tantamount for finding the extrema of the corresponding functionals, which, in some cases, may be an easier problem to tackle.

By way of introduction to variational methods, consider a functional $I[u]$ defined as

$$I[u] \;=\; \int_\Omega \left[\frac{k}{2} \left(\frac{\partial u}{\partial x_1} \right)^2 + \frac{k}{2} \left(\frac{\partial u}{\partial x_2} \right)^2 + f u \right] \, d\Omega \;, \tag{4.1}$$

where $k = k(x_1, x_2) > 0$ and $f = f(x_1, x_2)$ are continuous functions in Ω. In addition, assume that the domain Ω possesses a smooth boundary $\partial\Omega$ with uniquely defined outward unit normal $\mathbf{n}$.

The functional $I[u]$ attains an extremum if, and only if, its first variation vanishes, namely

$$\begin{aligned} \delta I[u] \;&=\; \int_\Omega \left[k \, \frac{\partial u}{\partial x_1} \delta \left(\frac{\partial u}{\partial x_1} \right) + k \, \frac{\partial u}{\partial x_2} \delta \left(\frac{\partial u}{\partial x_2} \right) + f \delta u \right] \, d\Omega \\ &=\; \int_\Omega \left[\frac{\partial u}{\partial x_1} k \, \frac{\partial \delta u}{\partial x_1} + \frac{\partial u}{\partial x_2} k \, \frac{\partial \delta u}{\partial x_2} + f \delta u \right] \, d\Omega \;=\; 0 \;, \end{aligned} \tag{4.2}$$

where u is assumed continuously differentiable and the order in which differential and variation operators are applied is switched, as argued in Section 2.4. Following the developments

of Section 3.2, integration by parts and application of the divergence theorem on (4.2) yields

$$\delta I[u] \;=\; \int_{\partial\Omega}\left[k\left(\frac{\partial u}{\partial x_1}n_1+\frac{\partial u}{\partial x_2}n_2\right)\delta u\right]\,d\Gamma-\int_{\Omega}\left[\frac{\partial}{\partial x_1}(k\frac{\partial u}{\partial x_1})+\frac{\partial}{\partial x_2}(k\frac{\partial u}{\partial x_2})-f\right]\delta u\,d\Omega \;=\; 0\,. \tag{4.3}$$

Recalling again that $\dfrac{\partial u}{\partial x_1}n_1+\dfrac{\partial u}{\partial x_2}n_2=\dfrac{\partial u}{\partial n}$, equation (4.3) assumes the form

$$\delta I[u] \;=\; \int_{\partial\Omega}k\frac{\partial u}{\partial n}\delta u\,d\Gamma-\int_{\Omega}\left[\frac{\partial}{\partial x_1}(k\frac{\partial u}{\partial x_1})+\frac{\partial}{\partial x_2}(k\frac{\partial u}{\partial x_2})-f\right]\delta u\,d\Omega \;=\; 0\,. \tag{4.4}$$

Owing to the arbitrariness of δu in Ω, the localization theorem of Section 3.1 implies that

$$\frac{\partial}{\partial x_1}(k\frac{\partial u}{\partial x_1})+\frac{\partial}{\partial x_2}(k\frac{\partial u}{\partial x_2}) \;=\; f \quad \text{in } \Omega\,, \tag{4.5}$$

while

$$k\frac{\partial u}{\partial n}\delta u \;=\; 0 \quad \text{on } \partial\Omega\,, \tag{4.6}$$

conditional upon sufficient smoothness of the respective fields. The first of the above two equations is identical to the Laplace-Poisson equation $(3.5)_1$, while the second equation presents three distinct alternatives:

(i) Set

$$\delta u \;=\; 0 \quad \text{on } \partial\Omega\,. \tag{4.7}$$

This condition implies that δu is not arbitrary on $\partial\Omega$, but rather the dependent variable u is prescribed throughout the boundary. Therefore, the space of admissible fields u is defined as

$$\mathcal{U} \;=\; \left\{u\in H^1(\Omega) \quad | \quad u=\bar{u} \text{ on } \partial\Omega\right\}\,, \tag{4.8}$$

where $\bar{u}$ is prescribed independently of the functional $I[u]$, in the sense that the functional contains no information regarding the actual value of u on $\partial\Omega$. Boundary conditions such as $u=\bar{u}$, which appear in the space of admissible fields, are referred to as *essential* (or *geometrical*). In this case, the admissible variations belong to the space $\mathcal{U}_0$ given by

$$\mathcal{U}_0 \;=\; \left\{u\in H^1(\Omega) \quad | \quad u=0 \text{ on } \partial\Omega\right\}\,. \tag{4.9}$$

(ii) Set

$$k\frac{\partial u}{\partial n} \;=\; 0 \quad \text{on } \partial\Omega\,. \tag{4.10}$$

In this case, δu is arbitrary on $\partial\Omega$, so that the boundary condition applies on the extremal function u, and is exactly derivable from the functional. Boundary conditions that directly apply to the extremal function (and its derivatives) are referred to as *natural* (or *suppressible*). No boundary restrictions are imposed on $\mathcal{U}$ in the present case and the space of admissible variations coincides with $\mathcal{U}$, that is,

$$\mathcal{U} \;=\; \mathcal{U}_0 \;=\; \left\{u \in H^1(\Omega)\right\} \;. \tag{4.11}$$

(iii) Admit a decomposition of boundary $\partial\Omega$ into parts Γ_u and Γ_q, such that $\partial\Omega = \overline{\Gamma_u \cup \Gamma_q}$. Subsequently, set

$$\begin{aligned} \delta u \;&=0 \quad \text{on } \Gamma_u \;, \\ k\frac{\partial u}{\partial n} \;&=0 \quad \text{on } \Gamma_q \;. \end{aligned} \tag{4.12}$$

Here, essential and natural boundary conditions are enforced on mutually disjoint portions of the boundary. In this case, the problem is said to involve *mixed* boundary conditions, and the space of admissible fields is defined as

$$\mathcal{U} \;=\; \left\{u \in H^1(\Omega) \quad | \quad u = \bar{u} \text{ on } \Gamma_u\right\} \;. \tag{4.13}$$

Here, the space of admissible variations $\mathcal{U}_0$ is defined as

$$\mathcal{U}_0 \;=\; \left\{u \in H^1(\Omega) \quad | \quad u = 0 \text{ on } \Gamma_u\right\} \;. \tag{4.14}$$

It can be concluded from the above, with reference to (4.4) that essential boundary conditions appear on variations of u and, possibly, its derivatives (and therefore place restrictions on the space of admissible fields), while natural boundary conditions appear directly on derivatives of the extremal function u. Equation (4.4) reveals that extremization of the functional in (4.1) yields a function u which satisfies the differential equation $(3.5)_1$ and boundary conditions selected in conjunction to the space of admissible fields $\mathcal{U}$.

It can be easily seen that the option of non-homogeneous natural boundary conditions of the form

$$-\,k\frac{\partial u}{\partial n} \;=\; \bar{q} \quad \text{on } \Gamma_q \tag{4.15}$$

can be accommodated if the original functional is amended so that it reads

$$\bar{I}[u] \;=\; \int_\Omega \left[\frac{k}{2}\left(\frac{\partial u}{\partial x_1}\right)^2 \;+\; \frac{k}{2}\left(\frac{\partial u}{\partial x_2}\right)^2 + fu\right] d\Omega \;+\; \int_{\Gamma_q} \bar{q}u \, d\Gamma \;, \tag{4.16}$$

where $\bar{q} = \bar{q}(x_1, x_2)$ is a continuous function on Γ_q. In this case, the vanishing of the first variation of $\bar{I}[u]$ implies that

$$\int_\Omega \left[\frac{\partial u}{\partial x_1} k \frac{\partial \delta u}{\partial x_1} \; + \; \frac{\partial u}{\partial x_2} k \frac{\partial \delta u}{\partial x_2} \; + \; f \delta u\right] d\Omega \; + \; \int_{\Gamma_q} \bar{q} \delta u \, d\Gamma \; = \; 0 \tag{4.17}$$

so that, assuming mixed boundary conditions and recalling (4.4),

$$\int_{\Gamma_q} (k \frac{\partial u}{\partial n} + \bar{q}) \delta u \, d\Gamma - \int_\Omega \left[\frac{\partial}{\partial x_1}(k \frac{\partial u}{\partial x_1}) + \frac{\partial}{\partial x_2}(k \frac{\partial u}{\partial x_2}) - f\right] \delta u \, d\Omega \; = \; 0 \; . \tag{4.18}$$

Equation (4.17) is termed the *variational form* of boundary-value problem (3.5). This is another version of a weak form associated with the strong form defined in (3.5). Comparing the above equation to (3.14), it is obvious that they are identical provided that the space of admissible field $\mathcal{W}$ for w in (3.14) is identical to that of δu in (4.17).

The variational form of problem (3.5) can be stated operationally as follows: find $u \in \mathcal{U}$, such that for all $\delta u \in \mathcal{U}_0$

$$B(\delta u, u) \; + \; (\delta u, f) \; + \; (\delta u, \bar{q})_{\Gamma_q} \; = \; 0 \; , \tag{4.19}$$

where the bilinear form $B(\delta u, u)$ is defined as

$$B(\delta u, u) \; = \; \int_\Omega \left(\frac{\partial \delta u}{\partial x_1} \, k \, \frac{\partial u}{\partial x_1} \; + \; \frac{\partial \delta u}{\partial x_2} \, k \, \frac{\partial u}{\partial x_2} \right) d\Omega \; , \tag{4.20}$$

and the linear forms $(\delta u, f)$ and $(\delta u, \bar{q})_{\Gamma_q}$ are defined, respectively, as

$$(\delta u, f) \; = \; \int_\Omega \delta u f \, d\Omega \tag{4.21}$$

and

$$(\delta u, \bar{q})_{\Gamma_q} \; = \; \int_{\Gamma_q} \delta u \bar{q} \, d\Gamma \; . \tag{4.22}$$

The correspondence of the above operational form with that of Section 3.2 is noted for the purpose of the forthcoming comparison between the Galerkin method and the variational method, when applied to problem (3.5).

The nature of the extremum point u (that is, whether it renders the functional $\bar{I}[u]$ minimum, maximum or merely stationary) can be determined by means of the second variation of $\bar{I}[u]$. Specifically, write

$$\delta^2 \bar{I}[u] \; = \; \delta \left(\delta \bar{I}[u] \right) \; = \; \int_\Omega \left[k \left(\frac{\partial \delta u}{\partial x_1} \right)^2 + k \left(\frac{\partial \delta u}{\partial x_2} \right)^2 \right] d\Omega \; , \tag{4.23}$$

and note that $\delta^2\bar{I}[u] > 0$, for all $\delta u \neq 0$, provided $\Gamma_u \neq \emptyset$. This is true because, if δu is assumed to be constant throughout the domain, it has to vanish everywhere, by definition of $\mathcal{U}_0$. It turns out that the conditions $\delta\bar{I}[u] = 0$ and $\delta^2\bar{I}[u] > 0$ are sufficient for any $I[u]$ to attain a local minimum at u, provided that $\delta^2\bar{I}[u]$ is also bounded from below at u, namely that

$$\delta^2\bar{I}[u] \ \geq c\|\delta u\|^2 \ , \tag{4.24}$$

where c is a positive constant. It turns out that (4.24) holds true, although discussion of the proof is postponed until Chapter 7.

In the preceding case, in addition to the variational form (4.17), there exists a *variational principle* associated with the solution u of problem (3.5). This can be stated as follows: find $u \in \mathcal{U}$, such that

$$\bar{I}[u] \ \leq \ \bar{I}[v] \ , \tag{4.25}$$

for all $v \in \mathcal{U}$, where $\bar{I}[u]$ is defined in (4.16).

Remark:

- ☛ Directional derivatives can be used in deriving the variational equation (4.17) from functional $\bar{I}[u]$. Indeed, write

$$D_v\bar{I}[u] \ = \ 0 \ \Rightarrow \Big[\frac{d}{d\omega}\bar{I}[u+\omega v]\Big]_{\omega=0} \ = \ 0 \ \Rightarrow$$
$$\int_\Omega \big[\frac{\partial u}{\partial x_1}k\,\frac{\partial v}{\partial x_1} \ + \ \frac{\partial u}{\partial x_2}k\,\frac{\partial v}{\partial x_2} \ + \ fv\big]\,d\Omega \ + \ \int_{\Gamma_q}\bar{q}v\,d\Gamma \ = \ 0 \ , \tag{4.26}$$

  where $v \in \mathcal{U}_0$.

4.2 Variational forms and variational principles

The analysis in Section 4.1, as applied to the model problem (3.5), provides an attractive perspective to the solution of certain partial differential equations: the solution is identified with a "point" u, which minimizes an appropriately constructed functional $I[u]$ over an admissible function space $\mathcal{U}$. Variational forms can be made fully equivalent to respective strong forms, as evidenced in the discussion of the weighted residual methods, under certain smoothness assumptions. However, the equivalence between variational forms and variational principles is not guaranteed: indeed, there exists no general method of constructing functionals $I[u]$, whose extremization recovers a desired variational form. In this sense, only

certain partial differential equations are amenable to analysis and solution by variational methods.

Vainberg's theorem provides the necessary and sufficient condition for the equivalence of a variational form to a functional extremization problem. If such an equivalence holds, the functional is typically referred to as a *potential*.

Theorem (Vainberg)
Consider a variational form

$$G(u,\delta u) \;=\; B(u,\delta u) \;+\; (f,\delta u) \;+\; (\bar{q},\delta u)_{\Gamma_q} \;=\; 0 \;, \tag{4.27}$$

where $u \in \mathcal{U}$, $\delta u \in \mathcal{U}_0$, and f and $\bar{q}$ are independent of u. Assume that the variation of G exists in a neighborhood $\mathcal{N}$ of u, and it is continuous in u at every point of $\mathcal{N}$. Then, the necessary and sufficient condition for the above weak form to be derivable from a potential in $\mathcal{N}$ is that

$$B(\delta u_1,\delta u_2) \;=\; B(\delta u_2,\delta u_1) \;, \tag{4.28}$$

namely that B be symmetric for all δu_1, $\delta u_2 \in \mathcal{U}_0$ and all $u \in \mathcal{N}$. Moreover, the functional $I[u]$ is defined as

$$I[u] \;=\; \frac{1}{2}B(u,u) + (f,u) + (\bar{q},u)_{\Gamma_q} \;. \tag{4.29}$$

Remarks:

- ☛ Apart from some technicalities, Vainberg's theorem can be proved following the above general procedure, even when B is non-linear in u.
- ☛ Checking the satisfaction of condition (4.28) is typically an easy task.

Example 4.2.1: Application of Vainberg's theorem
Recall the variational form (4.19), which is associated with boundary-value problem (3.5). In this case, recalling (4.20),

$$B(u,\delta u) \;=\; \int_\Omega \left(\frac{\partial \delta u}{\partial x_1}\, k\, \frac{\partial u}{\partial x_1} \;+\; \frac{\partial \delta u}{\partial x_2}\, k\, \frac{\partial u}{\partial x_2} \right) d\Omega \;,$$

$$(f,\delta u) \;=\; \int_\Omega f\, \delta u\, d\Omega \;,$$

and

$$(\bar{q},\delta u)_{\Gamma_q} \;=\; \int_{\Gamma_q} \bar{q}\, \delta u\, d\Gamma \;.$$

Using Vainberg's theorem, it can be immediately concluded that, since B is symmetric, there exists a potential $I[u]$, which, according to (4.29), is given by

$$\begin{aligned} I[u] &= \frac{1}{2}B(u,u) + (f,u) + (\bar{q},u)_{\Gamma_q} \\ &= \int_\Omega \left[\frac{k}{2}\left(\frac{\partial u}{\partial x_1}\right)^2 + \frac{k}{2}\left(\frac{\partial u}{\partial x_2}\right)^2 + fu \right] d\Omega + \int_{\Gamma_q} u\bar{q}\, d\Gamma \ , \end{aligned}$$

and whose extremization yields the above variational form. ◀

4.3 Rayleigh-Ritz method

The *Rayleigh-Ritz method* yields approximate solutions to partial differential equations, whose variational form is derivable from a potential $I[u]$. The central idea of the Rayleigh-Ritz method is to extremize $I[u]$ over a properly constructed subspace $\mathcal{U}_h$ of the space of admissible fields $\mathcal{U}$. To this end, write

$$u \doteq u_h = \sum_{I=1}^{N} \alpha_I \varphi_I + \varphi_0 \ , \tag{4.30}$$

where φ_I, $I = 1, \ldots, N$, is a specified family of interpolation functions that vanish where essential boundary conditions are enforced. In addition, function φ_0 is defined so that u_h satisfy identically the essential boundary conditions. Consequently, an N-dimensional subspace $\mathcal{U}_h$ is completely defined by (4.30). Extremization of $I[u]$ over $\mathcal{U}_h$ yields

$$\delta I[u_h] = \delta I[\sum_{I=1}^{N} \alpha_I \varphi_I + \varphi_0] = 0 \ . \tag{4.31}$$

Instead of directly obtaining the variational form of the problem by determining the explicit form of $\delta I[u_h]$ as a function of u_h, one may rewrite the extremization statement in terms of a real function $\tilde{I}$ of parameters α_I, $I = 1, \ldots, N$, namely

$$\delta \tilde{I}(\alpha_1, \ldots, \alpha_N) = 0 \ . \tag{4.32}$$

It follows from (4.32) that $\tilde{I}(\alpha_1, \ldots, \alpha_N)$ is a scalar function of α_1, α_2, ..., α_N, which attains an extremum over $\mathcal{U}_h$ if, and only if,

$$\frac{\partial \tilde{I}}{\partial \alpha_1}\delta\alpha_1 + \frac{\partial \tilde{I}}{\partial \alpha_2}\delta\alpha_2 + \ldots + \frac{\partial \tilde{I}}{\partial \alpha_N}\delta\alpha_N = 0 \ . \tag{4.33}$$

Since the variations $\delta\alpha_I$, $I = 1, \ldots, N$, are arbitrary, it may be immediately concluded that

$$\frac{\partial \tilde{I}}{\partial \alpha_I} \;=\; 0 \qquad , \qquad I = 1, \ldots, N \; . \tag{4.34}$$

Equations (4.34) may be solved for parameters α_I, so that an approximate solution to the variational problem is expressed by means of (4.30).

Example 4.3.1: The Rayleigh-Ritz method for an ordinary differential equation
Consider the functional $I[u]$ defined in the domain $(0, 1)$ as

$$I[u] \;=\; \int_0^1 \left[\frac{1}{2} \left(\frac{du}{dx} \right)^2 \;+\; u \right] dx \;+\; 2u \Big|_{x=1} \; ,$$

and the associated essential boundary condition $u(0) = 0$. The above functional is associated with the one-dimensional version of the Laplace-Poisson equation discussed in Section 3.2. In particular, it can be readily established that extremization of $I[u]$ recovers the solution to a boundary-value problem of the form

$$\begin{aligned} \frac{d^2u}{dx^2} \;&=\; 1 \qquad \text{in } \Omega = (0,1) \; , \\ -\frac{du}{dx} \;&=\; 2 \qquad \text{on } \Gamma_q = \{1\} \; , \\ u \;&=\; 0 \qquad \text{on } \Gamma_u = \{0\} \; . \end{aligned}$$

In order to obtain a Rayleigh-Ritz approximation to the solution of the preceding boundary-value problem, write u_h as

$$u_h(x) \;=\; u_N(x) \;=\; \sum_{I=1}^{N} \alpha_I \varphi_I(x) \;+\; \varphi_0(x) \; ,$$

and set, for simplicity, $\varphi_0 = 0$, so that the homogeneous essential boundary condition at $x = 0$ be satisfied. A one-parameter Rayleigh-Ritz approximation can be determined by choosing $\varphi_1(x) = x$. Then,

$$\begin{aligned} I[u_1] \;&=\; \int_0^1 \left[\frac{1}{2}\alpha_1^2 \;+\; \alpha_1 x \right] dx \;+\; 2\alpha_1 \\ &=\; \frac{1}{2}\alpha_1^2 \;+\; \frac{5}{2}\alpha_1 \; . \end{aligned}$$

Setting the first variation of $I[u_1]$ to zero, it follows that

$$\alpha_1 \;+\; \frac{5}{2} \;=\; 0 \; ,$$

from where it is concluded that $\alpha_1 = -\frac{5}{2}$, and

$$u_1(x) \;=\; -\frac{5}{2}x \; .$$

Similarly, one may consider a two-parameter polynomial Rayleigh-Ritz approximation by choosing $\varphi_1(x) = x$ and $\varphi_2(x) = x^2$. In this case, $I[u]$ takes the form

$$\begin{aligned} I[u_2] &= \int_0^1 \left[\frac{1}{2}\left(\alpha_1 + 2\alpha_2 x\right)^2 + \left(\alpha_1 x + \alpha_2 x^2\right)\right] dx + 2(\alpha_1 + \alpha_2) \\ &= \frac{1}{2}\alpha_1^2 + \alpha_1\alpha_2 + \frac{2}{3}\alpha_2^2 + \frac{5}{2}\alpha_1 + \frac{7}{3}\alpha_2 \, . \end{aligned}$$

Setting the first variation of $I[u_2]$ to zero, results in the system of equations

$$\begin{aligned} \alpha_1 + \alpha_2 &= -\frac{5}{2} \, , \\ \alpha_1 + \frac{4}{3}\alpha_2 &= -\frac{7}{3} \, , \end{aligned}$$

whose solution gives $\alpha_1 = -3$ and $\alpha_2 = \frac{1}{2}$, hence

$$u_2(x) = -3x + \frac{1}{2}x^2 \, .$$

The approximate solution $u_2(x)$ coincides with the exact solution of the boundary-value problem. Furthermore, $u_1(x)$ and $u_2(x)$ coincide with the respective solutions obtained in Section 3.2 using the Bubnov-Galerkin method with the same interpolation functions.

A different approximate solution $\tilde{u}_2$ can be obtained using the Rayleigh-Ritz method in connection with *piece-wise* linear polynomial interpolation functions of the form

$$\varphi_1(x) = \begin{cases} 2x & \text{if } 0 \le x \le 0.5 \\ 2(1 - x) & \text{if } 0.5 < x \le 1 \end{cases}$$

and

$$\varphi_2(x) = \begin{cases} 0 & \text{if } 0 \le x \le 0.5 \\ 2(x - \frac{1}{2}) & \text{if } 0.5 < x \le 1 \end{cases} \, ,$$

where functions φ_1 and φ_2 are depicted in Figure 4.1. Then, $I[\tilde{u}_2]$ is written as

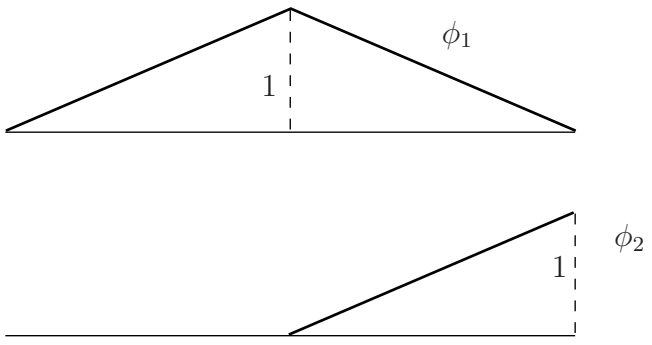


Figure 4.1. *Piecewise linear interpolations functions in one dimension*

$$
\begin{aligned}
I[\tilde{u}_2] \;&=\; \int_0^{0.5} [\frac{1}{2}\,(2\alpha_1)^2 \;+\; 2\alpha_1 x]\,dx \\
&\quad +\; \int_{0.5}^{1} [\frac{1}{2}\,(-2\alpha_1 \;+\; 2\alpha_2)^2 \;+\; 2\alpha_1(1 \;-\; x) \;+\; 2\alpha_2(x \;-\; \frac{1}{2})]\,dx \;+\; 2\alpha_2 \\
&=\; 2\alpha_1^2 \;-\; 2\alpha_1\alpha_2 \;+\; \alpha_2^2 \;+\; \frac{1}{2}\alpha_1 \;+\; \frac{9}{4}\alpha_2\;.
\end{aligned}
$$

Again, setting the variation of $I[\tilde{u}_2]$ to zero yields

$$
\begin{aligned}
4\alpha_1 \;-\; 2\alpha_2 \;&=\; -\frac{1}{2}\;, \\
-2\alpha_1 \;+\; 2\alpha_2 \;&=\; -\frac{9}{4}\;,
\end{aligned}
$$

so that $\alpha_1 = -\frac{11}{8}$ and $\alpha_2 = -\frac{5}{2}$, and

$$
\tilde{u}_2(x) \;=\; \begin{cases} -\frac{11}{4}x & \text{if } 0 \le x \le 0.5 \\ -\frac{1}{4}(1 \;+\; 9x) & \text{if } 0.5 < x \le 1 \end{cases}\quad .
$$

Solutions u_1, u_2 and $\tilde{u}_2$ are plotted in Figure 4.2. ◀

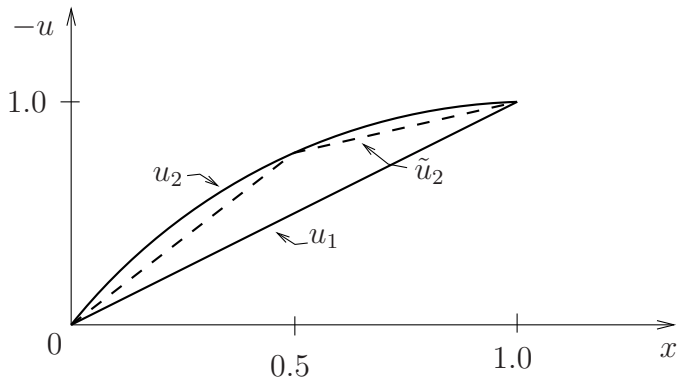


Figure 4.2. *Comparison of exact and approximate solutions*

The Rayleigh-Ritz method is related to the Bubnov-Galerkin method, in the sense that, whenever the former is applicable, it yields identical approximate solutions with the latter, when using the same interpolation functions. However, it should be understood that, even in these cases, the methods are fundamentally different in that the former is a variational method, whereas the latter is not.

4.4 Exercises

Problem 1

Consider the fourth-order ordinary differential equation

$$\frac{d^4u}{dx^4} \;=\; f \quad \text{in} \;\; \Omega = (0,1) \;,$$

where f is a function of x. No boundary conditions are prescribed at this stage on $\partial\Omega = \big\{\{0\},\{1\}\big\}$.

(a) Multiply the differential equation by a function v and subsequently integrate over the domain Ω (note that other than the standard integrability requirement, no restrictions are placed on v, since no boundary conditions have been specified).

(b) Perform two successive integrations by parts on the above integral to obtain

$$\begin{aligned}\int_\Omega \big(\frac{d^4u}{dx^4} \;-\; f\big) v \, dx \;&=\; D_v\Big\{ \int_\Omega \Big(\frac{1}{2}\big(\frac{d^2u}{dx^2}\big)^2 \;-\; fu \Big)\, dx \Big\} \;+ \\ &\frac{d^3u}{dx^3}(1)\, v(1) \;-\frac{d^3u}{dx^3}(0)\, v(0) \;-\; \frac{d^2u}{dx^2}(1)\, \frac{dv}{dx}(1) \;+\; \frac{d^2u}{dx^2}(0)\, \frac{dv}{dx}(0)\,.\end{aligned}$$

(c) Conclude from part (b) that stationarity of the functional $I[u]$, defined as

$$I[u] \;=\; \int_\Omega \Big(\frac{1}{2}\big(\frac{d^2u}{dx^2}\big)^2 \;-\; fu \Big)\, dx \;,$$

implies that the given differential equation is satisfied in Ω and, moreover,

$$\begin{aligned}\frac{d^3u}{dx^3}(1)\, v(1) \;&=\; 0 \;, \\ \frac{d^3u}{dx^3}(0)\, v(0) \;&=\; 0 \;, \\ \frac{d^2u}{dx^2}(1)\, \frac{dv}{dx}(1) \;&=\; 0 \;, \\ \frac{d^2u}{dx^2}(0)\, \frac{dv}{dx}(0) \;&=\; 0 \;.\end{aligned}$$

(d) Identify all possible essential and natural boundary conditions on $\partial\Omega$. Note that essential boundary conditions appear on the variations v (and therefore restrict the admissible fields), while natural boundary conditions appear directly on derivatives of the extremal function u.

(e) Consider an expanded functional $I_1[u]$ given by

$$I_1[u] \;=\; \int_\Omega \Big(\frac{1}{2}\big(\frac{d^2u}{dx^2}\big)^2 \;-\; fu \Big)\, dx \;+\; q_1(0)u(0) \;+\; q_1(1)\frac{du}{dx}(1) \;,$$

where q_1 is defined on $\partial\Omega$, and derive the boundary equations associated with stationarity of $I_1[u]$. Again, identify all possible essential and natural boundary conditions on $\partial\Omega$. Can the functional be further amended so that it read

$$I_2[u] \;=\; \int_\Omega \Big(\frac{1}{2}\big(\frac{d^2u}{dx^2}\big)^2 \;-\; fu \Big)\, dx \;+\; q_1(0)u(0) \;+\; q_1(1)\frac{du}{dx}(1) \;+\; q_2(1)\frac{d^2u}{dx^2}(1) \;,$$

where q_2 is defined at $x = 1$? Clearly explain your answer.

Problem 2

Consider the initial-value problem

$$\begin{aligned} \frac{\partial}{\partial x}\big(k\frac{\partial u}{\partial x}\big) \; - \; f \;\; &= \;\; l\,\frac{\partial u}{\partial t} \quad \text{in} \;\; \Omega \times (0,T) \;, \\ u \;\; &= \;\; \bar{u}(x,t) \quad \text{on} \;\; \partial\Omega \times (0,T) \;, \\ u(x,0) \;\; &= \;\; u_0(x) \quad \text{in} \;\; \Omega \;, \end{aligned}$$

for the determination of $u = u(x,t)$, where $\Omega \subset \mathbb{R}$, and k, l and f are given non-vanishing functions of x. Starting from the above strong form, obtain the variational form of the problem and use Vainberg's theorem to show that there exists no variational theorem associated with the weak form.

Problem 3

Consider the boundary-value problem

$$\begin{aligned} -\frac{d}{dx}\big((1+x)\frac{du}{dx}\big) \;\; &= \;\; 0 \quad \text{in} \;\; (0,1) \;, \\ u(0) \;\; &= \;\; 0 \;, \\ u(1) \;\; &= \;\; 1 \;. \end{aligned}$$

Construct the variational form of this problem and use Vainberg's theorem to verify that there exists a variational principle associated with the problem. Also, construct the relevant functional $I[u]$ and specify the space over which it attains a minimum. Obtain a Rayleigh-Ritz approximation to the solution of the above problem, assuming a two-parameter polynomial approximation. Submit a plot of the exact and the approximate solution.

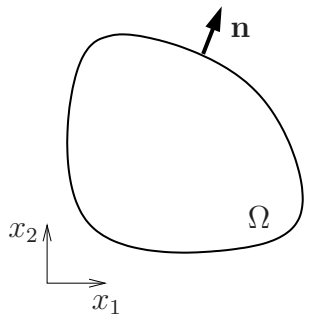
Problem 4

Consider the boundary-value problem

$$\begin{aligned} \left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}\right) \;\; &= \;\; f \qquad \text{in } \Omega = \{(x_1,x_2) \;\mid\; 0 < x_1, x_2 < 1\} \;, \\ \frac{\partial u}{\partial n} \;\; &= \;\; 0 \qquad \text{on } \Gamma_q = \{(x_1,x_2) \;\mid\; x_1 = 0 \text{ or } x_2 = 0\} \;, \\ u \;\; &= \;\; 0 \qquad \text{on } \Gamma_u = \{(x_1,x_2) \;\mid\; x_1 = 1 \text{ or } x_2 = 1\} \;, \end{aligned}$$

where $f = f(x_1,x_2)$. In particular, minimize an appropriate functional $I[u]$ over a properly defined functional space $\mathcal{U}$ (treat boundary conditions on $x_1 = 0$ and $x_2 = 0$ as natural) using a one-parameter polynomial approximation

$$u_1(x_1,x_2) \;\; = \;\; \alpha_1\,\varphi_1 \;,$$



where $\varphi_1(x_1, x_2) = (1-x_1)(1-x_2)$. Also, suggest a two-parameter polynomial approximation

$$u_2(x, y) \;=\; \alpha_1\,\varphi_1 \;+\; \alpha_2\,\varphi_2$$

by specifying φ_2 to be the next to φ_1 in the hierarchy of admissible polynomials in $\mathcal{U}$. In this part, do not solve for the coefficients α_1 and α_2.

Problem 5

Consider the homogeneous boundary-value problem

$$\frac{\partial^4 u}{\partial x_1^4} \;+\; 2\frac{\partial^4 u}{\partial x_1^2 \partial x_2^2} \;+\; \frac{\partial^4 u}{\partial x_2^4} \;=\; f \quad \text{in} \;\; \Omega \subset \mathbb{R}^2 \;,$$

$$u \;=\; 0 \;\;,\;\; \frac{\partial u}{\partial n} \;=\; 0 \quad \text{on} \;\; \partial\Omega \;,$$

where $f = f(x_1, x_2)$.

Show that the solution u of the above problem extremizes the functional $I[u]$ defined as

$$I[u] \;=\; \int_\Omega \Big[\frac{1}{2}\Big\{\big(\frac{\partial^2 u}{\partial x_1^2}\big)^2 \;+\; 2\,\frac{\partial^2 u}{\partial x_1^2}\frac{\partial^2 u}{\partial x_2^2} \;+\; \big(\frac{\partial^2 u}{\partial x_2^2}\big)^2\Big\} \;-\; f u\Big]\, d\Omega \;,$$

over the set of all admissible functions $\mathcal{U}$. Verify that the same conclusion can be reached for the functional $I_1[u]$ defined as

$$I_1[u] \;=\; \int_\Omega \Big[\frac{1}{2}\Big\{\big(\frac{\partial^2 u}{\partial x_1^2}\big)^2 \;+\; 2\,\big(\frac{\partial^2 u}{\partial x_1 \partial x_2}\big)^2 \;+\; \big(\frac{\partial^2 u}{\partial x_2^2}\big)^2\Big\} \;-\; f u\Big]\, d\Omega \;.$$

This problem confirms that the functional associated with a differential equation is not necessarily unique.

Problem 6

Consider functional $I[u]$ defined as

$$I[u] \;=\; \int_0^1 \left[\frac{1}{2}(x+1)\left(\frac{du}{dx}\right)^2 - u\right] dx \;,$$

on the admissible field $\mathcal{U} = \big\{u \in H^1(0,1) \;\mid\; u(0) = 0,\; u(1) = 1\big\}$.

(a) Obtain the strong form of the boundary-value problem, whose variational form is associated with the extremization of $I[u]$ over $\mathcal{U}$.

(b) Argue that $I[u]$ attains a minimum at the exact solution $u = u(x)$ of the boundary-value problem identified in part (a).

(c) Find a *Rayleigh-Ritz* solution to the problem assuming the simplest possible (non-trivial) one-parameter polynomial approximation u_h.

Problem 7

Consider the boundary-value problem

$$\begin{aligned} \frac{d^2}{dr^2}\Big(r\frac{d^2u}{dr^2}\Big) \;-\; \frac{d}{dr}\Big(\frac{1}{r}\frac{du}{dr}\Big) \;&=\; f \quad \text{in } \Omega \;=\; (a,b)\;, \\ u \;&=\; 0 \quad \text{on } \partial\Omega \;=\; \big\{\{a\},\{b\}\big\}\;, \\ \frac{du}{dr} \;&=\; 0 \quad \text{on } \partial\Omega \;=\; \big\{\{a\},\{b\}\big\}\;, \end{aligned}$$

where $0 < a < b$.

(a) Starting from the integral equation

$$\int_a^b \delta u\,\Big[\frac{d^2}{dr^2}\Big(r\frac{d^2u}{dr^2}\Big) \;-\; \frac{d}{dr}\Big(\frac{1}{r}\frac{du}{dr}\Big) \;-\; f\Big]\,dr \;=\; 0\;,$$

derive the weak (variational) form of the above boundary-value problem and identify explicitly the spaces of admissible functions u and δu. Assume that f is of class C^∞.

(b) Verify that Vainberg's theorem is applicable to the problem and use it to construct a functional $I[u]$, whose extremization over all admissible functions recovers the variational form of part (a).

(c) Suggest the general form of a two-parameter polynomial function that can be used in a Rayleigh-Ritz approximate solution to the problem (you do not have to actually solve the Rayleigh-Ritz problem).

4.5 Suggestions for further reading

Sections 4.1

[1] K. Washizu. *Variational Methods in Elasticity & Plasticity.* Pergamon Press, Oxford, 1982. [*This is a classic book on variational methods with emphasis on structural and solid mechanics*].

[2] H. Sagan. *Introduction to the Calculus of Variations.* Dover, New York, 1992. [*This book contains a complete discussion of the theory of first and second variation*].

Section 4.2

[1] M.M. Vainberg. *Variational Methods for the Study of Nonlinear Operators.* Holden-Day, San Francisco, 1964. [*This book contains many important mathematical results, including a non-linear version of Vainberg's theorem*].

Section 4.3

[1] B.A. Finlayson and L.E. Scriven. The method of weighted residuals – a review. *Appl. Mech. Rev.*, 19:735–748, 1966. [*This review article contains on page 741 an exceptionally clear discussion of the relationship between Rayleigh-Ritz and Galerkin methods*].

Chapter 5

Construction of Finite Element Subspaces

5.1 Introduction

The finite element method provides a general procedure for the construction of admissible spaces $\mathcal{U}_h$ and, if necessary, $\mathcal{W}_h$, in connection with the weighted-residual and variational methods discussed in the previous two chapters.

By way of background, define the *support* of a real-valued function $f(\mathbf{x})$ in its domain $\Omega \subset \mathbb{R}^n$ as the closure of the set of all points $\mathbf{x}$ in the domain for which $f(\mathbf{x}) \neq 0$, namely

$$\operatorname{supp} f \;=\; \overline{\{\mathbf{x} \in \Omega \;\mid\; f(\mathbf{x}) \neq 0\}}\,. \tag{5.1}$$

With reference to the general form of the approximation functions u_h and w_h given by equations (3.24), respectively, one may establish a distinction between *global* and *local* approximation methods. Local approximation methods are those for which $\operatorname{supp} \varphi_I$ is "small" compared to the size of the domain of approximation, whereas global methods employ interpolation functions with relatively "large" support.

Global and local approximation methods present both advantages and disadvantages. Global methods are often capable of providing excellent estimates of a solution with relatively small computational effort, especially when the analyst has a good understanding of the expected solution characteristics. However, a proper choice of global interpolation functions may not always be readily available, as in the case of complicated domains, where satisfaction of any boundary conditions could be a difficult, if not an insurmountable task. In addition,

global methods rarely lend themselves to a straightforward algorithmic implementation, and even when they do, they almost invariably yield dense linear systems of the form (3.30), which may require substantial computational effort to solve.

Local methods are more suitable for algorithmic implementation than global methods, as they can easily satisfy Dirichlet (or essential) boundary conditions, and they typically yield "banded" linear algebraic systems. Moreover, these methods are flexible in allowing local refinements in the approximation, when warranted by the analysis. However, local methods can be surprisingly expensive, even for simple problems, when the desired degree of accuracy is high. The so-called *global-local* approximation methods combine both global and local interpolation functions in order to exploit the positive characteristics of both methods.

Interpolation functions that appear in equations (3.24) need to satisfy certain general admissibility criteria. These criteria are motivated by the requirement that the resulting finite-dimensional solution spaces be well-defined and capable of accurately and uniformly approximating the exact solutions. In particular, all families of interpolation functions $\{\varphi_1, \ldots, \varphi_N\}$ should have the following properties:

(a) For any $\mathbf{x} \in \Omega$, there exists an I with $1 \leq I \leq N$, such that $\varphi_I(\mathbf{x}) \neq 0$. In other words, the interpolation functions should "cover" the whole domain of analysis. Indeed, if the above property is not satisfied, it follows that there exist interior points of Ω where the exact solution is not at all approximated.

(b) The interpolation functions should satisfy the Dirichlet (or essential) boundary conditions, if required by the underlying weak form, as discussed in Chapters 3 and 4.

(c) The interpolation functions should be *linearly independent* in the domain of analysis. To further elaborate on this point, let $\mathcal{U}_h$ be the space of admissible solutions and consider the interpolation functions $\varphi_I \in \mathcal{U}_h$, $I = 1, 2, \ldots, N$. Linear independence of these interpolation functions means that

$$\sum_{I=1}^{N} \alpha_I \, \varphi_I \;=\; 0 \quad \Leftrightarrow \quad \alpha_I \;=\; 0 \quad , \quad I = 1, \ldots, N \; . \tag{5.2}$$

An alternative statement of linear independence of the functions $\{\varphi_1, \ldots, \varphi_N\}$ is that given any $u_h \in \mathcal{U}_h$, there exists a unique set of parameters $\{\alpha_1, \ldots, \alpha_N\}$, such that

$$u_h \;=\; \sum_{I=1}^{N} \alpha_I \, \varphi_I \; . \tag{5.3}$$

If property (c) holds, then functions $\{\varphi_1, \ldots, \varphi_N\}$ are said to form a *basis* of $\mathcal{U}_h$, and $\mathcal{U}_h$ is an N-dimensional space. Linear independence of the interpolation functions is essential for the derivation of approximate solutions. Indeed, if the parameters $\{\alpha_1, \ldots, \alpha_N\}$ are not uniquely defined for any given $u_h \in \mathcal{U}_h$, then the linear algebraic system resulting from the approximation (see, e.g., equation (3.30)) does not possess a unique solution and, consequently, the discrete problem is ill-posed.

(d) The interpolation functions must satisfy the integrability requirements emanating from the associated weak forms, as discussed in Chapters 3 and 4. Otherwise, the integrals comprising the weak form used for the approximation are not well-defined.

(e) The family of interpolation functions should possess sufficient "approximating power". One of the most important features of Hilbert spaces is that they provide a suitable framework for examining the issue of how (and in what sense) a function $u_h \in \mathcal{U}_h \subset \mathcal{U}$, defined as

$$u_h \;=\; \sum_{I=1}^{N} \alpha_I \, \varphi_I \tag{5.4}$$

approximates a function $u \in \mathcal{U}$ as N increases towards infinity. In order to address the above point, consider a set of functions $\{\varphi_1, \varphi_2, \ldots, \varphi_N, \ldots\}$, which are linearly independent in $\mathcal{U}$ and, thus, form a countably infinite basis.[1] These functions are termed *orthonormal* in $\mathcal{U}$ if

$$< \varphi_I, \varphi_J > = \begin{cases} 1 & \text{if } I = J \\ 0 & \text{if } I \neq J \end{cases} \; . \tag{5.5}$$

Any countably infinite basis can be orthonormalized by means of a *Gram-Schmidt orthogonalization* procedure, as follows: starting with the first function φ_1, let $a_1 = 1/\|\varphi_1\|$, where $\|\cdot\|$ is the natural norm, and define

$$\psi_1 \;=\; a_1 \varphi_1 \;=\; \frac{\varphi_1}{\|\varphi_1\|} \, , \tag{5.6}$$

so that, clearly,

$$< \psi_1, \psi_1 > = \; \|\psi_1\|^2 \;=\; 1 \; . \tag{5.7}$$

Then, let

$$\psi_2 \;=\; a_2 [\varphi_2 - \; < \varphi_2, \psi_1 > \psi_1] \; , \tag{5.8}$$

[1] Hilbert spaces can be shown to always possess such a basis.

where a_2 is a scalar parameter to be determined. In geometric terms, ψ_2 is obtained by subtracting the projection of φ_2 on the function ψ_1 from φ_2 and then scaling the outcome by a_2. It is immediately seen from (5.8) that

$$\begin{aligned} < \psi_1, \psi_2 > &= < \psi_1, a_2\varphi_2 \; - \; a_2 < \varphi_2, \psi_1 > \psi_1 > \\ &= a_2[< \psi_1, \varphi_2 > \; - \; < \psi_1, \psi_1 >< \psi_1, \varphi_2 >] \; = \; 0 \; . \end{aligned} \tag{5.9}$$

The scalar parameter a_2 is determined so that $\|\psi_2\| = 1$, namely

$$a_2 \; = \; \frac{1}{\|\varphi_2 \; - \; < \varphi_2, \psi_1 > \psi_1\|} \; . \tag{5.10}$$

In general, the function φ_{K+1}, $K = 1, 2, \ldots$, gives rise to ψ_{K+1} defined as

$$\psi_{K+1} \; = \; a_{K+1}[\varphi_{K+1} \; - \; \sum_{I=1}^{K} < \varphi_{K+1}, \psi_I > \psi_I] \; , \tag{5.11}$$

where

$$a_{K+1} \; = \; \frac{1}{\|\varphi_{K+1} \; - \; \sum_{I=1}^{K} < \varphi_{K+1}, \psi_I > \psi_I\|} \; . \tag{5.12}$$

To establish that $\{\psi_1, \psi_2, \ldots, \psi_N, \ldots\}$ are orthonormal, it suffices to show by induction that if $\{\psi_1, \psi_2, \ldots, \psi_K\}$ are orthonormal, then ψ_{K+1} is orthonormal with respect to each of the first K members of the sequence. Indeed, using (5.11) it is seen that

$$\begin{aligned} < \psi_{K+1}, \psi_K > &= < a_{K+1}\varphi_{K+1} \; - \; a_{K+1} \sum_{I=1}^{K} < \varphi_{K+1}, \psi_I > \psi_I, \psi_K > \\ &= < a_{K+1}\varphi_{K+1}, \psi_K > \; - \; \sum_{I=1}^{K-1} a_{K+1} < \varphi_{K+1}, \psi_I >< \psi_I, \psi_K > \\ &\quad - a_{K+1} < \varphi_{K+1}, \psi_K >< \psi_K, \psi_K > \; = \; 0 \end{aligned} \tag{5.13}$$

and, for $N < K$,

$$\begin{aligned} < \psi_{K+1}, \psi_N > = & < a_{K+1}\varphi_{K+1} - a_{K+1} \sum_{I=1}^{K} < \varphi_{K+1}, \psi_I > \psi_I, \psi_N > \\ = & < a_{K+1}\varphi_{K+1}, \psi_N > - \sum_{I=1}^{N-1} a_{K+1} < \varphi_{K+1}, \psi_I >< \psi_I, \psi_N > \\ & - \sum_{I=N+1}^{K} a_{K+1} < \varphi_{K+1}, \psi_I >< \psi_I, \psi_N > \\ & - a_{K+1} < \varphi_{K+1}, \psi_N >< \psi_N, \psi_N > = 0 , \end{aligned} \tag{5.14}$$

which establishes the desired result.

Since $\{\psi_1, \psi_2, \ldots, \psi_N, \ldots\}$ is an orthonormal basis in $\mathcal{U}$, one may uniquely write any $u \in \mathcal{U}$ in the form

$$u = \sum_{I=1}^{\infty} \alpha_I \psi_I , \tag{5.15}$$

which may be interpreted as meaning that given any $\epsilon > 0$, there exists a positive integer $N = N(\epsilon)$ and scalars α_I, such that

$$\| u - \sum_{I=1}^{n} \alpha_I \psi_I \| < \epsilon , \tag{5.16}$$

for all $n > N$. The terms α_I in (5.15) are known as the *Fourier coefficients* of u with respect to the given basis and can be easily determined by exploiting the orthonormality of ψ_I and noting that

$$\begin{aligned} < u, \psi_J > = & < \sum_{I=1}^{\infty} \alpha_I \psi_I, \psi_J > \\ = & \sum_{I=1}^{\infty} \alpha_I < \psi_I, \psi_J > = \alpha_J . \end{aligned} \tag{5.17}$$

This means that α_J is the projection of u on the unit function ψ_J, as in Figure 5.1. Therefore, one obtains the *Fourier representation* of u with respect to the given orthonormal basis as

$$u = \sum_{I=1}^{\infty} < u, \psi_I > \psi_I . \tag{5.18}$$

It is noted that the natural norm of u satisfies *Parseval's identity*, that is,

$$\| u \|^2 = \sum_{I=1}^{\infty} |\alpha_I|^2 , \tag{5.19}$$

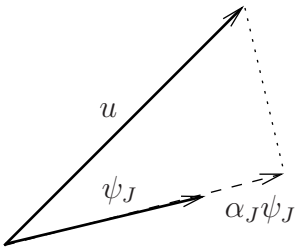


Figure 5.1. *Geometric interpretation of Fourier coefficients*

where α_I are obtained from (5.17). Indeed,

$$\begin{aligned} \|u\|^2 \; = < u, u > \; = \; & < \sum_{I=1}^{\infty} < u, \psi_I > \psi_I, \sum_{J=1}^{\infty} < u, \psi_J > \psi_J > \\ & = \sum_{I=1}^{\infty} < u, \psi_I > \sum_{J=1}^{\infty} < u, \psi_J >< \psi_I, \psi_J > \\ & = \sum_{I=1}^{\infty} < u, \psi_I >^2 \; = \; \sum_{I=1}^{\infty} |\alpha_I|^2 \; . \end{aligned} \tag{5.20}$$

Parseval's identity implies that the series $\sum_{I=1}^{\infty} |\alpha_I|^2$ converges for any given u. Therefore, the Fourier coefficients satisfy the property $\lim_{I \to \infty} \alpha_I = 0$.

The interpolation functions φ_I, $I = 1, 2, \ldots, N$, used in the finite element method should satisfy the *completeness* property in the space of admissible functions $\mathcal{U}$. This means that they should be a finite subset of a countably infinite basis of $\mathcal{U}$ which asymptotically replicates, as N increases, the original countably infinite basis. In light of Parseval's identity (5.19), this implies that the series $\sum_{I=1}^{N} |\alpha_I|^2$ converges with N, where α_I is the Fourier coefficient associated with interpolation function φ_I for any $u_h \in \mathcal{U}_h$.

In order to motivate the choice of φ_I's, recall the Weierstrass approximation theorem of elementary real analysis:

Weierstrass Approximation Theorem (1885)

Given a continuous function f in $[a, b] \subset \mathbb{R}$ and any scalar $\epsilon > 0$, there exists a polynomial P_N of degree N, such that

$$|f(x) \; - \; P_N(x)| \; < \; \epsilon \; , \tag{5.21}$$

for all $x \in [a, b]$.

The above theorem states that any continuous function f on a closed subset of $\mathbb{R}$ can be uniformly approximated by a polynomial function to within any desired level of accuracy. Using this theorem, one may conclude that the exact solution u to a given problem defined on $\mathbb{R}$ can be potentially approximated by a polynomial u_h of degree N, so that

$$\lim_{N\to\infty} \|u \; - \; u_h\| \; = \; 0 \tag{5.22}$$

or, equivalently, in view of (5.15), that

$$u(x) \; = \; \sum_{I=0}^{\infty} \alpha_I x^I \; . \tag{5.23}$$

In addition to polynomials in $\mathbb{R}$ (that is, the sequence of functions 1, x, x^2, ...), the completeness property is also satisfied by trigonometric functions, as evidenced by the classical Fourier representation of a continuous real function u in the form

$$u(x) \; = \; \sum_{k=0}^{\infty} (\alpha_k \sin kx \; + \; \beta_k \cos kx) \; . \tag{5.24}$$

The above theorem can also be extended to polynomials defined in closed and bounded subsets of $\mathbb{R}^n$.

The interpolation functions are required to be complete, so that any smooth solution u be representable to within specified error by means of u_h. It should be noted that the preceding theorem does *not* guarantee that a numerical method, which involves complete interpolation functions, will necessarily provide a uniformly accurate approximate solution.

In certain occasions, properties (a), (b) and (e) of the interpolation functions are relaxed, in order to accommodate special requirements of the approximation.

5.2 Finite element spaces

The finite element method provides a systematic procedure for constructing local piecewise polynomial interpolation functions, in accordance with the guidelines of the previous section. In order to initiate the discussion of the finite element method, introduce the notion of the finite element discretization: given the open domain Ω of analysis, admit the existence of

open finite element sub-domains Ω^e, such that

$$\bar{\Omega} \;=\; \overline{\bigcup_e \Omega^e}\;, \tag{5.25}$$

as shown schematically in Figure 5.2. Similarly, the boundary $\partial\Omega$ is decomposed into

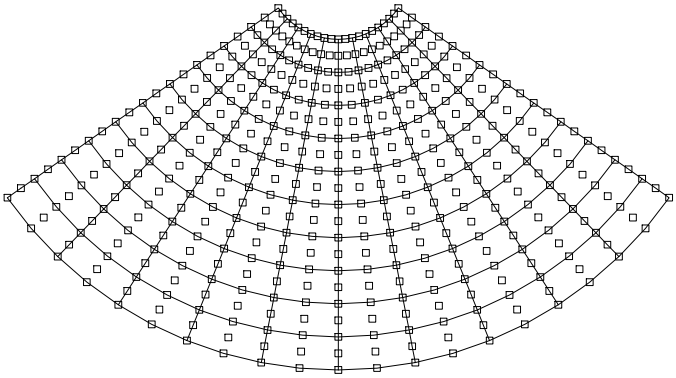


Figure 5.2. *A finite element mesh*

sub-domains $\partial\Omega^e$ consistently with (5.25), so that

$$\partial\Omega \;\subseteq\; \overline{\bigcup_e \partial\Omega^e}\;. \tag{5.26}$$

Also, admit the existence of points $I \in \bar{\Omega}$, associated with sub-domains Ω^e. Points I have position vectors $\mathbf{x}_I$ with reference to a fixed coordinate system, and are referred to as the *nodal points* (or simply *nodes*). The collection of finite element sub-domains and nodal points within $\bar{\Omega}$ constitutes a *finite element mesh*. The geometry of each Ω^e is completely defined by the nodal points that lie on $\partial\Omega^e$ and in Ω^e.

Continuous piecewise polynomial interpolation functions φ_I are defined for each interior finite element node I, so that, by convention,

$$\varphi_I(\mathbf{x}_J) \;=\; \begin{cases} 1 & \text{if } I = J \\ 0 & \text{otherwise} \end{cases}\;, \tag{5.27}$$

where $\mathbf{x}_J$ is the position vector of node J. Similarly, one may define local interpolation functions for exterior boundary nodes that do not lie on the portion of the boundary where Dirichlet (or essential) conditions are enforced. The latter are satisfied locally by approximation functions which vanish at all other boundary and interior nodes. Moreover, the support of φ_I is restricted to the element domains in the immediate neighborhood of node I, as shown in Figure 5.3.

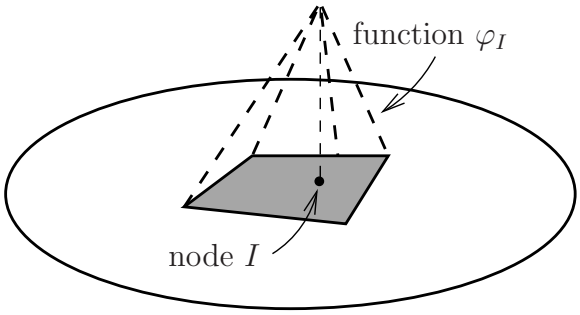


Figure 5.3. *A finite element-based interpolation function*

At this stage, it is possible to formally define a *finite element* as a mathematical object which consists of three basic ingredients:

(i) a finite element sub-domain Ω^e,

(ii) a linear space of interpolation functions, or more specifically, the restriction of the interpolation functions to Ω^e, and

(iii) a set of "degrees of freedom", namely those parameters α_I that are associated with non-vanishing interpolation functions in Ω^e.

Given the above general description of the finite element interpolation functions, one may proceed in establishing their admissibility in connection with the properties outlined in the preceding section.

Property (a) is generally satisfied by construction of the interpolation functions. Indeed, given any interior point P of Ω, there exist neighboring nodal points whose interpolation functions are non-zero at P. However, it is conceivable that (5.25) holds only approximately, that is, subdomains Ω^e only partially cover the domain Ω, as seen in Figure 5.4. In this case, property (a) may be violated in certain small regions on the domain, thus inducing an error in the approximation.

Property (b) is directly satisfied by fixing the degrees-of-freedom associated with the portion of the exterior boundary where Dirichlet (or essential) conditions are enforced. Again, an error in the approximation is introduced when the actual exterior boundary is not represented exactly by the finite element domain discretization, see Figure 5.5.

In order to show that property (c) is satisfied, assume, by contradiction, that for all $\mathbf{x} \in \Omega$, $u_h = 0$, while not all scalar parameters α_I are zero. Owing to (5.27), one may immediately

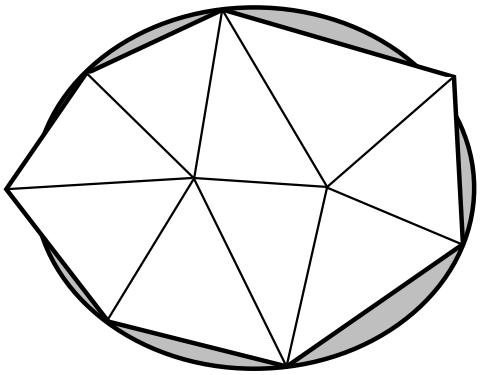


Figure 5.4. *Finite element vs. exact domain*

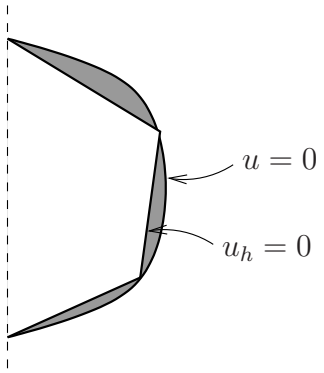


Figure 5.5. *Error in the enforcement of Dirichlet boundary conditions due to the difference between the exact and the finite element domain*

conclude that at any node J

$$u_h(\mathbf{x}_J) \;=\; \sum_{I=1}^{N} \alpha_I \, \varphi_I(\mathbf{x}_J) \;=\; \alpha_J \;, \tag{5.28}$$

hence $\alpha_J = 0$. Since the nodal point J is chosen arbitrary, it follows that all α_J vanish, which constitutes a contradiction. Therefore, the proposed interpolation functions are linearly independent.

As already seen in Chapters 3 and 4, property (d) dictates that the admissible fields $\mathcal{U}_h$ and, if applicable, $\mathcal{W}_h$ must render the associated weak forms well-defined. In the finite element literature, this property is frequently referred to as the *compatibility condition*. The terminology stems from certain second-order differential equations of structural mechanics (e.g., the displacement-based equations of motion for linearly elastic solids), where integrability of the weak forms amounts to the requirement that the assumed displacement fields u_h

belong to H^1. This, in turn, implies that the displacements should be "compatible", namely the displacements of individual finite elements domains should not exhibit overlaps or voids, as in Figure 5.6.

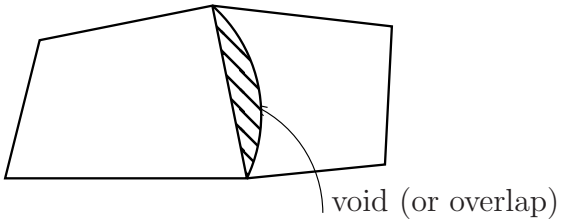


Figure 5.6. *A potential violation of the integrability (compatibility) requirement*

Property (e) and its implications within the context of the finite element method deserve special attention, and are discussed separately in the following section.

5.3 Completeness property

The completeness property requires that piecewise polynomial fields $\mathcal{U}_h$ contain "points" u_h that may uniformly approximate the exact solution u of a differential equation to within desirable accuracy. This approximation may be achieved by enriching $\mathcal{U}_h$ in various ways:

(a) Refinement of the domain discretization, while keeping the order of the polynomial interpolation fixed (*h-refinement*).

(b) Increase in the order of polynomial interpolation within a fixed domain discretization (*p-refinement*).

(c) Combined refinement of the domain discretization and increase of polynomial order of interpolation (*hp-refinement*).

(d) Repositioning of a domain discretization with fixed order of polynomial interpolation and element topology to enhance the accuracy of the approximation in a selective manner (*r-refinement*).

It can be shown that in order to assess completeness of a given finite element interpolation, one must be able to conclude that the error in the approximation of the highest derivative of u in the weak form is at most of order $o(h)$, where h is a measure of the "fineness" of the

approximation. To see this point, consider a smooth real function u, and fix a point $x = \bar{x}$ in its domain, as in Figure 5.7. With reference to Taylor's theorem, write

$$u(\bar{x}+h) \;=\; u(\bar{x}) \;+\; hu'(\bar{x}) \;+\; \frac{1}{2!}h^2u''(\bar{x}) \;+\; \ldots \;+\; \frac{1}{q!}h^q u^{(q)}(\bar{x}) \;+\; o(h^{q+1}) \;, \tag{5.29}$$

in the domain $(\bar{x}, \bar{x}+h)$, for any given $h > 0$. Assuming that $\mathcal{U}_h$ contains all polynomials in h that are complete to degree q, it follows that there exists a $u_h \in \mathcal{U}_h$ so that at $\bar{x}+h$

$$u \;=\; u_h \;+\; o(h^{q+1}) \;. \tag{5.30}$$

Letting p be the order of the highest derivative of u in any weak form, it follows from (5.30)

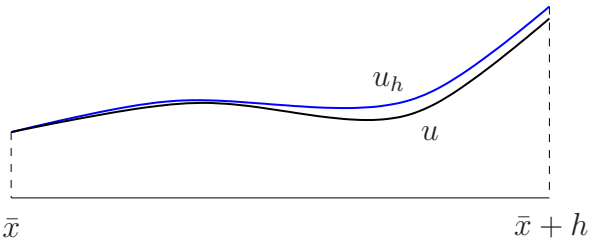


Figure 5.7. *A function u and its approximation u_h in the domain $(\bar{x}, \bar{x}+h)$*

that

$$\frac{d^p u}{dx^p} \;=\; \frac{d^p u_h}{dx^p} \;+\; o(h^{q-p+1}) \;. \tag{5.31}$$

Thus, for $\mathcal{U}_h$ to be a complete field, it suffices to establish that

$$q - p + 1 \;\geq\; 1 \;, \tag{5.32}$$

or, equivalently,

$$q \;\geq\; p \;. \tag{5.33}$$

Indeed, in this case $\dfrac{d^p u_h}{dx^p}$ converges to $\dfrac{d^p u}{dx^p}$ as $h \to 0$ (that is, under h-refinement). Hence, in order to guarantee completeness, any approximation to u must contain all polynomial terms of degree at least p. The same argument can be easily made for functions of several variables.

In the context of weighted residual methods, completeness guarantees that weak forms are computed to full resolution as the approximation becomes finer in the sense that $h \to 0$ (h-refinement) or $q \to \infty$ (p-refinement). Indeed, consider a weak form

$$B(w,u) \;+\; (w,f) \;+\; (w,\bar{q})_{\Gamma_q} \;=\; 0 \;, \tag{5.34}$$

associated with a linear partial differential equation and let both u_h and w_h be refined in the same fashion (that is, using h- or p-refinement). It is easily seen that

$$B(w,u) \;=\; B(w_h,u_h) \;+\; B(w-w_h,u-u_h) \;+\; B(w-w_h,u_h) \;+\; B(w_h,u-u_h)\;. \quad (5.35)$$

Owing to (5.33), the last three terms on the right-hand side of the above identity are of order at least $o(h^{q-p+1})$ before integration. Taking the limit of the above identity as h approaches zero, it is desired that

$$B(w,u) \;=\; \lim_{h\to 0} B(w_h,u_h)\;. \quad (5.36)$$

under h-refinement. This holds true if condition (5.33) is satisfied. Likewise, taking the limit as $q \to \infty$, it is desired that

$$B(w,u) \;=\; \lim_{q\to\infty} B(w_h,u_h)\;. \quad (5.37)$$

under p-refinement. Equation (5.31) implies that this is also true as q approaches infinity.

Similar conclusions can be reached for the linear forms (w,f) and $(w,\bar{q})_{\Gamma_q}$.

Example 5.3.1: Completeness of approximations

(a) Consider the differential equation

$$k\frac{d^2u}{dx^2} \;=\; f \quad \text{in } (0,1)\;,$$

where $p=1$ when using the Galerkin method (see Example 3.2.1). Then, (5.33) implies that all polynomial approximations of u should be complete up to linear terms in x, namely should contain independent monomials $\{1,x\}$.

(b) Consider the differential equation

$$\frac{\partial^4 u}{\partial x_1^4} \;+\; 2\frac{\partial^4 u}{\partial x_1^2\partial x_2^2} \;+\; \frac{\partial^4 u}{\partial x_2^4} \;=\; f \quad \text{in } \Omega\subset\mathbb{R}^2\;,$$

where it is been shown (see Chapter 4, Problem 5) that a variational form is derivable such that $p=2$. Then, the monomial terms that should be independently present in any complete approximation are $\{1,x_1,x_2,x_1^2,x_1x_2,x_2^2\}$. ◀

Obviously, setting $q=p$ as in the preceding examples satisfies only the minimum requirement for completeness. Generally, the higher the order q relative to p, the richer the space of admissible functions $\mathcal{U}_h$. Thus, an increase in the order of completeness beyond the minimum requirements set by (5.33) yields more accurate approximations of the exact solution to a given problem.

A polynomial approximation in $\mathbb{R}^n$ is said to be *polynomially complete up to order* q, if it contains *independently* all monomials $x_1^{q_1} x_2^{q_2} \ldots x_n^{q_n}$, where $q_1 + q_2 + \ldots + q_n \leq q$. In $\mathbb{R}$, the above implies that terms $\{1, x, \ldots, x^q\}$ should be independently represented. In $\mathbb{R}^2$, completeness up to order q can be conveniently visualized by means of a *Pascal triangle*, as shown in Figure 5.8. In this case, the number of independent monomials is $\frac{(q+1)(q+2)}{2}$.

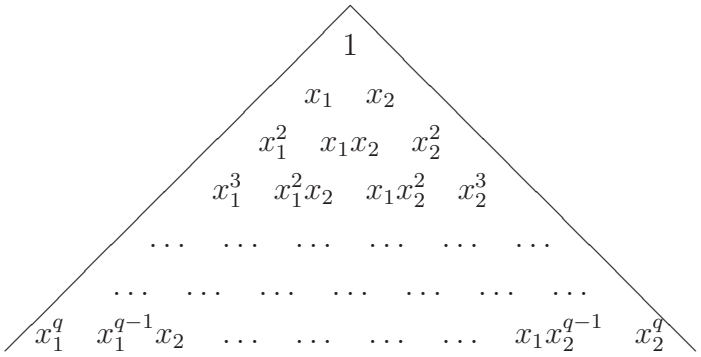


Figure 5.8. *Pascal triangle*

An alternative (and somewhat stronger) formalization of the completeness property can be obtained by noting that application of a weighted residual method to linear differential equation

$$A[u] \;=\; f \tag{5.38}$$

within *fixed* spaces $\mathcal{U}_h$ and, if necessary, $\mathcal{W}_h$, yields an approximate solution u_h typically obtained by solving a system of linear algebraic equations of the form (3.30). Therefore, for fixed h, one may define a *discrete* operator A_h associated with the operator A, so that

$$A_h[u] \;=\; A[u_h] \;, \tag{5.39}$$

for any $u_h \in \mathcal{U}_h$. Subsequently, the domain of the discrete operator can be appropriately extended, so that it encompasses the whole space $\mathcal{U}$. Then, completeness implies that

$$A_h[u] \;=\; f \;+\; o(h^\alpha) \quad ; \quad \alpha > 0 \;. \tag{5.40}$$

Assuming sufficient smoothness of u, equation (5.40) implies that the discrete operator A_h converges to the continuous operator A as h approaches zero.

5.4 Basic finite element shapes in one, two and three dimensions

The geometric shape of a finite element domain Ω^e is fully determined by two sets of data:

(i) The position of the nodal points,

(ii) A domain interpolation procedure, which may coincide with the interpolation employed for the dependent variables of the problem.

Thus, the position vector $\mathbf{x}$ of a point in Ω^e can be written as a function of the position vectors $\mathbf{x}_I$ of nodes I and the given domain interpolation functions.

5.4.1 One dimension

One-dimensional finite element domains are line segments, straight or curved, as in Figure 5.9.

Figure 5.9. *Finite element domains in one dimension*

5.4.2 Two dimensions

Two-dimensional finite element domains are typically triangular or quadrilateral, with straight or curved edges, as in Figure 5.10. Elements with more complicated geometric shapes are rarely used in practice.

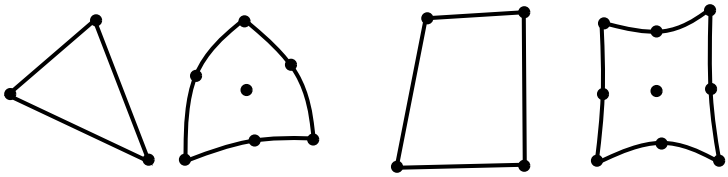


Figure 5.10. *Finite element domains in two dimensions*

5.4.3 Three dimensions

The most useful three-dimensional finite element domains are tetrahedral (tets), pentahedral (pies) and hexahedral (bricks), with straight or curved edges and flat or non-flat faces, see Figure 5.11. Again, elements with more complicated geometric shapes are generally avoided.

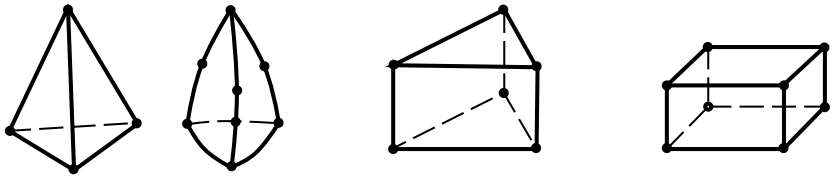


Figure 5.11. *Finite element domains in three dimensions*

5.4.4 Higher dimensions

Elements in four or higher dimensions are very seldom used and will not be discussed here.

5.5 Polynomial element interpolation functions

Element interpolation functions are generally used for two purposes, namely to generate an approximation for the dependent variable and to parametrize the element domain. The second use of these functions justifies their frequent identification as *shape* functions. In what follows, polynomial element interpolation functions are visited in connection with the construction of finite element approximations in one, two and three dimensions.

5.5.1 Interpolations in one dimension

First, consider the case of continuous piecewise polynomial interpolation functions. These functions are admissible for the Galerkin-based finite element approximations associated with the solution of the one-dimensional counterpart of the Laplace-Poisson equation discussed in earlier sections. Furthermore, assume that the order of the highest derivative in the weak form is $p = 1$, so that the completeness requirement necessitates the construction of a polynomial approximation which is complete to degree $q \geq 1$.

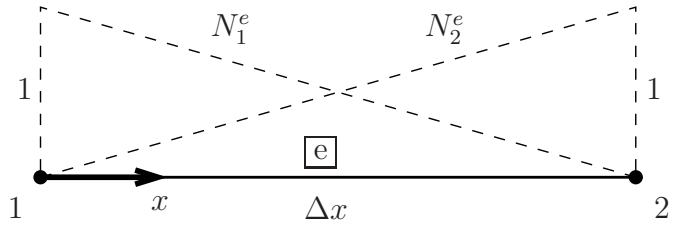


Figure 5.12. *Linear element interpolation in one dimension*

The simplest finite element which satisfies the above integrability and completeness requirements is the 2-node element of length Δx, as in Figure 5.12. This element is extracted from the one-dimensional finite element mesh with domain Ω and consisting of N nodes and $N-1$ equally-sized elements with piecewise linear interpolation, see Figure 5.13. Associated with this element (identified here as element e), there is a local node numbering system and a coordinate system x (here having its origin at node 1). The interpolation u_h of the dependent variable u in the element domain Ω^e takes the form

$$u_h(x) \;=\; N_1^e(x)\, u_1^e + N_2^e(x)\, u_2^e \;, \tag{5.41}$$

where the element interpolation functions N_1^e and N_2^e are defined as

$$N_1^e(x) \;=\; 1 - \frac{x}{\Delta x} \qquad , \qquad N_2^e(x) \;=\; \frac{x}{\Delta x} \;. \tag{5.42}$$

As easily concluded from Figures 5.12 and 5.13, these element interpolation function are the

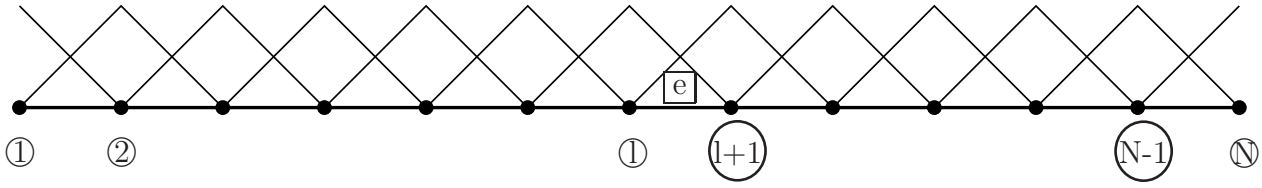


Figure 5.13. *One-dimensional finite element mesh with piecewise linear interpolation*

restrictions of the global interpolation functions φ_l and φ_{l+1} to the domain Ω^e of element e. It is immediately see from (5.42) that $N_1^e(0) = 1$ and $N_1^e(\Delta x) = 0$, while $N_2^e(0) = 0$ and $N_2^e(\Delta x) = 1$. Also, u_1^e and u_2^e in (5.41) denote the element "degrees of freedom", which, given the form of the element interpolation functions, can be directly identified with the ordinates of the dependent variable at nodes 1 and 2 (numbered locally as shown in Figure 5.12), respectively.

Clearly, the above finite element approximation is complete in 1 and x (that is, $q = 1$). In addition, it satisfies the compatibility requirement by construction, since the dependent variable is continuous in Ω^e, as well as at all interelement boundaries. The last conclusion can be reached by noting that the nodal degrees of freedom are shared when the nodes themselves are shared between contiguous elements.

A complete quadratic interpolation can be obtained by constructing 3-node elements, as in Figure 5.14. Here, the dependent variable is given by

$$u_h(x) \;=\; N_1^e(x)\, u_1^e + N_2^e(x)\, u_2^e + N_3^e(x)\, u_3^e \;, \tag{5.43}$$

where

$$N_1^e(x) \;=\; \frac{(x-\Delta x)(x-2\Delta x)}{2\Delta x^2} \quad , \quad N_2^e(x) \;=\; \frac{x(x-\Delta x)}{2\Delta x^2} \quad , \quad N_3^e(x) \;=\; -\frac{x(x-2\Delta x)}{\Delta x^2} \;. \tag{5.44}$$

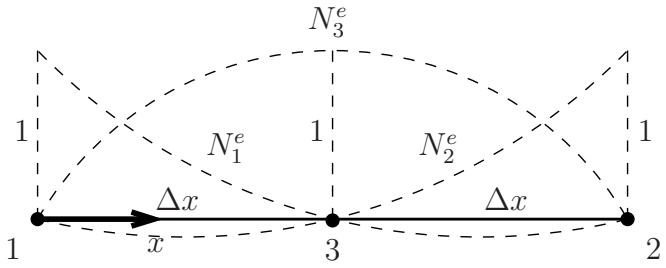


Figure 5.14. *Standard quadratic element interpolations in one dimension*

Again, compatibility and completeness (to degree $q = 2$) are satisfied by the interpolation in (5.43).

Generally, for an element with $q+1$ nodes having coordinates x_i, $i = 1, \ldots, q+1$, one may obtain a *Lagrangian interpolation* of the form

$$u_h(x) \;=\; \sum_{i=1}^{q+1} N_i^e(x) u_i^e \;. \tag{5.45}$$

The generic element interpolation function N_i^e is a polynomial of degree q written as

$$N_i^e(x) \;=\; a_0 \;+\; a_1 x \;+\; \ldots \;+\; a_q x^q \;, \tag{5.46}$$

where

$$N_i^e(x_j) \;=\; \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \;. \tag{5.47}$$

Conditions (5.47) give rise to a system of $q+1$ equations for the $q+1$ parameters c_0 to c_q. Interestingly, a direct solution of this system is not necessary to determine the explicit functional form of N_i^e. Indeed, it can be immediately verified that

$$N_i^e(x) \;=\; l_i(x) \;=\; \frac{(x \;-\; x_1)\ldots(x \;-\; x_{i-1})(x \;-\; x_{i+1})\ldots(x \;-\; x_{q+1})}{(x_i \;-\; x_1)\ldots(x_i \;-\; x_{i-1})(x_i \;-\; x_{i+1})\ldots(x_i \;-\; x_{q+1})}\;, \tag{5.48}$$

for $i = 1, 2, \ldots, q+1$.

The above general procedure by way of which the degree of polynomial completeness is progressively increased by adding nodes and associated degrees of freedom is referred to as *standard* interpolation. An alternative to this procedure is provided by the so-called *hierarchical* interpolation. To illustrate an application of hierarchical interpolation, consider the 2-node element discussed earlier in this section, and modify (5.41) so that

$$u_h(x) \;=\; N_1^e(x)\,u_1^e + N_2^e(x)\,u_2^e + \tilde{N}_3^e(x)\,\alpha^e\;, \tag{5.49}$$

where both the function $\tilde{N}_3^e$ and the degree of freedom α^e are to be determined. Clearly, $\tilde{N}_3^e$ should be a quadratic function of x, since a complete linear interpolation is already guaranteed by the original form of u_h in (5.41). Therefore,

$$\tilde{N}_3^e(x) \;=\; a_0 \;+\; a_1 x \;+\; a_2 x^2\;, \tag{5.50}$$

where a_0, a_1 and a_2 are parameters to be determined. In order to satisfy compatibility (which here means continuity of u_h at interelement boundaries), it is sufficient to assume that $\tilde{N}_3^e(0) = 0$ and $\tilde{N}_3^e(\Delta x) = 0$. These conditions imply that

$$\tilde{N}_3^e(x) \;=\; \frac{ax}{\Delta x}\,(1 \;-\; \frac{x}{\Delta x})\;, \tag{5.51}$$

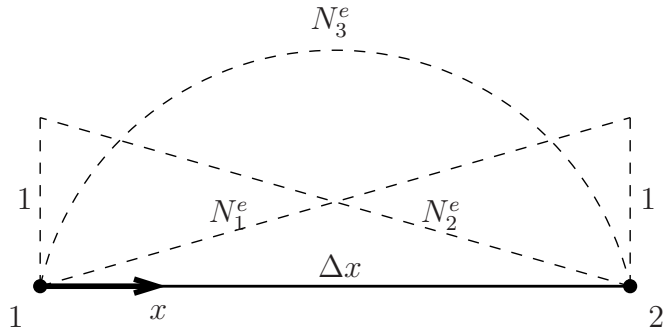


Figure 5.15. *Hierarchical quadratic element interpolations in one dimension*

where a can be any non-zero constant. The three interpolation functions obtained by the above hierarchical procedure are depicted in Figure 5.15. In contrast to the standard interpolation, here the degree of freedom α^e is not associated with a finite element node. A simple algebraic interpretation of α^e can be obtained as follows: let the element interpolation function $\tilde{N}_3^e(x)$ take the specific form

$$\tilde{N}_3^e(x) \;=\; \frac{4x}{\Delta x}\,(1 \;-\; \frac{x}{\Delta x})\;. \tag{5.52}$$

Then, it can be trivially concluded that α^e quantifies the deviation from linearity of u_h at the mid-point of Ω^e, namely at $x = \Delta x/2$.

Remark:

- ☛ By construction, the degree of freedom α^e is not shared between contiguous elements. Consequently, it is possible to determine its value *locally* (that is, at the element level), as a function of the other element degrees of freedom. As a result, α^e does not need to enter the global system of equations. In the structural mechanics literature, the process of locally eliminating hierarchical degrees of freedom at the element level is referred to as *static condensation*.

Finite element approximations that maintain continuity of the first derivative of the dependent variable are necessary for the solution of certain higher-order partial differential equations. As a representative example, consider the fourth-order differential equation $\dfrac{d^4u}{dx^4} = f$, which, after application of the Bubnov-Galerkin method gives rise to a weak form that involves second-order derivatives of both the dependent variable and the weighting function. Here, continuity of the first derivative of u is sufficient to guarantee well-posedness of the weak form. In addition, the completeness requirement is met by ensuring that the approximation in each element is polynomially complete to degree $q \geq 2$.

A simple element which satisfies the above requirements is the 2-node element of Figure 5.16, in which each node is associated with two degrees of freedom, identified as the ordinates of the dependent variable u and its first derivative $\theta = \dfrac{du}{dx}$, respectively.

Given that there are four degrees of freedom in each element, a cubic polynomial interpolation of the form

$$u_h(x) \;=\; c_0 \;+\; c_1 x \;+\; c_2 x^2 \;+\; c_3 x^3 \tag{5.53}$$

can be determined uniquely under the conditions

$$u_h(0) \;=\; u_1^e \quad , \quad \frac{du_h}{dx}(0) \;=\; \theta_1^e \quad , \quad u_h(\Delta x) \;=\; u_2^e \quad , \quad \frac{du_h}{dx}(\Delta x) \;=\; \theta_2^e\;. \tag{5.54}$$

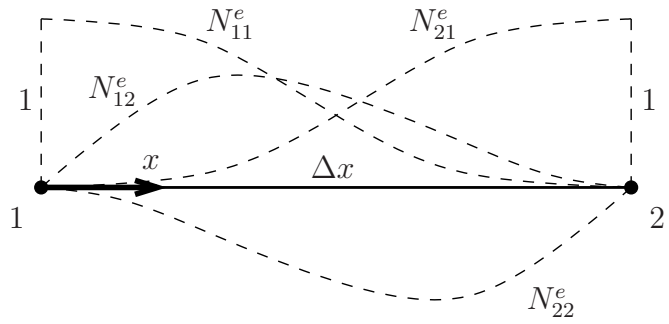


Figure 5.16. *Hermitian interpolation functions in one dimension*

Solving the above equations for the four parameters c_0 to c_3 yields a standard *Hermitian interpolation*, in which

$$u_h(x) \;=\; \sum_{i=1}^{2} N^e_{i1}(x)\, u^e_i \;+\; \sum_{i=1}^{2} N^e_{i2}(x)\, \theta^e_i \;, \tag{5.55}$$

where

$$\begin{aligned} N^e_{11} \;&=\; 1 \;-\; 3\,\big(\frac{x}{\Delta x}\big)^2 \;+\; 2\,\big(\frac{x}{\Delta x}\big)^3 \qquad , \qquad N^e_{21} \;=\; 3\,\big(\frac{x}{\Delta x}\big)^2 \;-\; 2\,\big(\frac{x}{\Delta x}\big)^3 \\ N^e_{12} \;&=\; \Delta x\Big[\frac{x}{\Delta x} \;-\; 2\,\big(\frac{x}{\Delta x}\big)^2 \;+\; \big(\frac{x}{\Delta x}\big)^3\Big] \qquad , \qquad N^e_{22} \;=\; \Delta x\Big[-\,\big(\frac{x}{\Delta x}\big)^2 \;+\; \big(\frac{x}{\Delta x}\big)^3\Big] \;. \end{aligned} \tag{5.56}$$

Generally, a Hermitian interpolation can be introduced for a $(q+1)$-node element, where each node i is associated with coordinate x_i and with degrees of freedom u^e_i and θ^e_i. It follows that u_h is a polynomial of degree $2q+1$ in the form

$$u_h(x) \;=\; \sum_{i=1}^{q+1} N^e_{i1}(x)\, u^e_i \;+\; \sum_{i=1}^{q+1} N^e_{i2}(x)\, \theta^e_i \;. \tag{5.57}$$

The element interpolation functions N^e_{i1} in the above equation satisfy

$$N^e_{i1}(x_j) \;=\; \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \qquad , \qquad \frac{dN^e_{i1}}{dx}(x_j) \;=\; 0 \;. \tag{5.58}$$

Similarly, the functions N^e_{i2} satisfy the conditions

$$N^e_{i2}(x_j) \;=\; 0 \qquad , \qquad \frac{dN^e_{i2}}{dx}(x_j) \;=\; \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \;. \tag{5.59}$$

It can be easily verified that the above *Hermitian polynomials* are defined as

$$N^e_{i1}(x) \;=\; \big[1 \;-\; 2l'_i(x_i)\,(x \;-\; x_i)\big]\, l^2_i(x) \qquad , \qquad N^e_{i2}(x) \;=\; (x \;-\; x_i)\, l^2_i(x) \;, \tag{5.60}$$

where $l_i(x)$ denotes the Lagrangian polynomial of degree q defined in (5.48). The above interpolation satisfies continuity of the dependent variable and its first derivative across interelement boundaries. In addition, it guarantees polynomial completeness up to degree $q = 3$.

Higher-order accurate elements can be also constructed starting from the 2-node element and adding hierarchical degrees of freedom. For example, one may assume a quartic interpolation of the form

$$u_h(x) \;=\; \sum_{i=1}^{2} N^e_{i1}(x)\, u^e_i \;+\; \sum_{i=1}^{2} N^e_{i2}(x)\, \theta^e_i \;+\; \tilde{N}^e_5(x)\, \alpha^e, \tag{5.61}$$

where the interpolation function $\tilde{N}^e_5$ is written as

$$\tilde{N}^e_5(x) \;=\; a_0 \;+\; a_1 x \;+\; a_2 x^2 \;+\; a_3 x^3 \;+\; a_4 x^4 \;. \tag{5.62}$$

Given the conditions $\tilde{N}^e_5(0) = \tilde{N}^e_5(\Delta x) = 0$ and $\dfrac{d\tilde{N}^e_5}{dx}(0) = \dfrac{d\tilde{N}^e_5}{dx}(\Delta x) = 0$, it follows that

$$\tilde{N}^e_5(x) \;=\; a\Big[\big(\frac{x}{\Delta x}\big)^2 \;-\; 2\,\big(\frac{x}{\Delta x}\big)^3 \;+\; \big(\frac{x}{\Delta x}\big)^4\Big] \;. \tag{5.63}$$

Finite element approximations which enforce continuity of higher-order derivatives are conceptually simple. The idea is to introduce degrees of freedom identified with the dependent variable and its derivatives up to the highest order in which continuity is desired. However, such elements are rarely used in practice and will not be discussed here in more detail.

5.5.2 Interpolations in two dimensions

First, consider finite element interpolations in two dimensions, where continuity of the dependent variable across interelement boundaries is initially assumed to be sufficient to satisfy the compatibility requirement, while polynomial completeness is necessary only to degree $p = 1$. It can be easily verified that the above requirements lead to a proper finite element approximation of the Laplace-Poisson equation discussed in connection with the Galerkin method in Section 3.2.

The simplest two-dimensional element is the 3-node straight-edge triangle Ω^e with one degree-of-freedom per node, as seen in Figure 5.17. For this element, assume a linear poly-

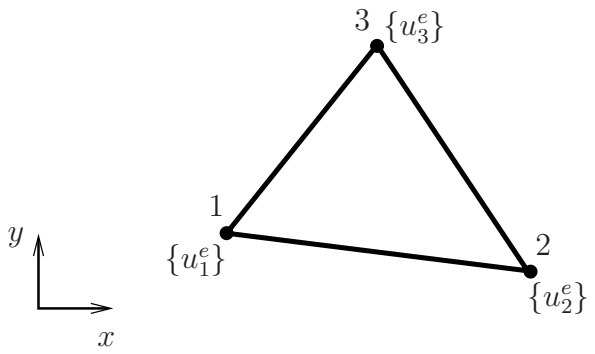


Figure 5.17. *A 3-node triangular element*

nomial interpolation u_h of the dependent variable u in the form

$$u_h(x,y) \;=\; \sum_{i=1}^{3} N_i^e(x,y)u_i^e \;=\; c_0 \;+\; c_1 x \;+\; c_2 y \;, \tag{5.64}$$

with reference to a fixed Cartesian coordinate system (x, y). Upon identifying the degrees of freedom at each node $i = 1, 2, 3$ having coordinates (x_i, y_i) with the ordinate u_i^e of the dependent variable at that node, one obtains a system of three linear algebraic equations with unknowns c_0, c_1 and c_2, in the form

$$\begin{aligned} u_1^e &= c_0 \;+\; c_1 x_1 \;+\; c_2 y_1 \;, \\ u_2^e &= c_0 \;+\; c_1 x_2 \;+\; c_2 y_2 \;, \\ u_3^e &= c_0 \;+\; c_1 x_3 \;+\; c_2 y_3 \;. \end{aligned} \tag{5.65}$$

Assuming that the solution of the above system is unique, one may write

$$\begin{aligned} c_0 &= \frac{1}{2A^e}\left[u_1^e(x_2y_3 \;-\; x_3y_2) \;+\; u_2^e(x_3y_1 - x_1y_3) \;+\; u_3^e(x_1y_2 - x_2y_1)\right] \;, \\ c_1 &= \frac{1}{2A^e}\left[u_1^e(y_2 \;-\; y_3) \;+\; u_2^e(y_3 \;-\; y_1) \;+\; u_3^e(y_1 \;-\; y_2)\right] \;, \\ c_2 &= \frac{1}{2A^e}\left[u_1^e(x_3 \;-\; x_2) \;+\; u_2^e(x_1 \;-\; x_3) \;+\; u_3^e(x_2 \;-\; x_1)\right] \;, \end{aligned} \tag{5.66}$$

where

$$A^e \;=\; \frac{1}{2}\det\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix} . \tag{5.67}$$

It is interesting to note that A^e represents the (signed) area of the triangle Ω^e. Therefore, the system (5.65) is uniquely solvable if, and only if, the nodes 1,2,3 do not lie on the same

line. In addition, it can be easily concluded that the area A^e of a non-degenerate triangle is positive if, and only if, the nodes are numbered in a counter-clockwise manner, as in Figure 5.17.

Explicit polynomial expressions for the element interpolation functions are obtained from (5.64) and (5.66) in the form

$$\begin{aligned} N_1^e(x,y) &= \frac{1}{2A^e}\Big[(x_2y_3 - x_3y_2) + (y_2 - y_3)x + (x_3 - x_2)y\Big] \\ N_2^e(x,y) &= \frac{1}{2A^e}\Big[(x_3y_1 - x_1y_3) + (y_3 - y_1)x + (x_1 - x_3)y\Big] \\ N_3^e(x,y) &= \frac{1}{2A^e}\Big[(x_1y_2 - x_2y_1) + (y_1 - y_2)x + (x_2 - x_1)y\Big] \end{aligned} \quad . \tag{5.68}$$

It can be noted from $(5.68)_1$ that $N_1^e(x,y) = 0$ coincides with the equation of the straight line passing through nodes 2 and 3. This observation is sufficient to guarantee continuity of u_h across interelement boundaries. Indeed, since N_1^e vanishes identically along 2-3, the interpolation u_h, which varies linearly along this line, is fully determined as a function of the degrees-of-freedom u_2^e and u_3^e. These degrees-of-freedom, in turn, are shared between the elements with common edge 2-3, which establishes the continuity of u_h as the edge 2-3 is crossed between these two elements. Obviously, entirely analogous arguments apply to edges 3-1 and 1-2. Furthermore, completeness to degree $q = 1$ is satisfied, since any linear polynomial function of x and y can be uniquely represented by three parameters, such as u_i^e, $i = 1, 2, 3$, and can be spanned over Ω^e by the interpolation functions in (5.68).

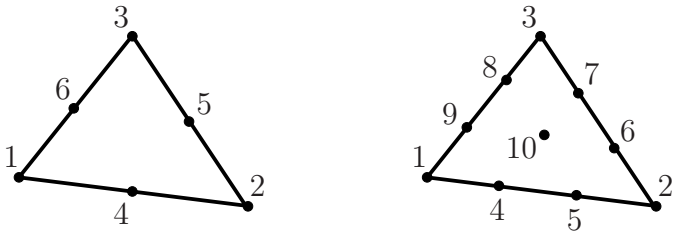


Figure 5.18. *Higher-order triangular elements (left: 6-node element, right: 10-node element)*

Triangular elements with polynomial order of completeness $q \geq 1$ can be constructed by adding nodes accompanied by degrees-of-freedom to the straight-edge triangle. Examples of 6- and 10-node triangular elements which are polynomially complete to degree $q = 2$ and 3 with reference to the Pascal triangle of Figure 5.8 are illustrated in Figure 5.18. It should be noted that the nodes are generally placed with geometric regularity. Thus, for the 6-node triangle, the nodes are located at the corners and the mid-edges of the triangular domain.

Again, the element interpolation functions can be determined by the procedure followed earlier for the 3-node triangle. Similarly, continuity of the dependent variable in these elements can be proved by arguments identical to those used for the 3-node triangle. Elements featuring irregular placement of the nodes, such as the 4-node element in Figure 5.19 are typically not desirable, as they produce a biased interpolation of the dependent variable without appreciably contributing towards increasing the polynomial degree of completeness. Such elements are sometimes used as "transitional" interfaces intended to properly connect meshes of different types of elements (e.g., a mesh consisting of 3-node triangles with another consisting of 6-node triangles).

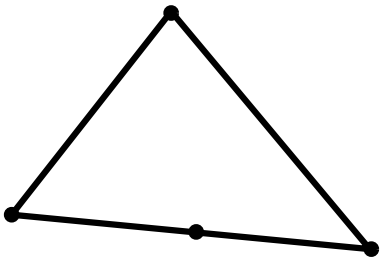


Figure 5.19. *A transitional 4-node triangular element*

In the study of triangular elements, it is analytically advantageous to introduce an alternative coordinate representation and use it instead of the standard Cartesian representation introduced earlier in this section. To this end, note that an arbitrary interior point of Ω^e with Cartesian coordinates (x, y) divides the element domain into three triangular sub-regions with areas A_1^e, A_2^e and A_3^e, as shown in Figure 5.20. Noting that

$$A_1^e \;=\; \frac{1}{2}\det\begin{bmatrix} 1 & x & y \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix} \;, \tag{5.69}$$

with similar expressions for A_2^e and A_3^e, define the so-called *area coordinates* of the point (x, y) as

$$L_i \;=\; \frac{A_i^e}{A^e} \quad , \quad i = 1, 2, 3 \;. \tag{5.70}$$

Clearly, only two of the three *area coordinates* are independent since it is readily seen from (5.70) that $L_1 + L_2 + L_3 = 1$. Interestingly, comparing (5.68) to (5.70) it is immediately apparent that $N_i^e = L_i$, $i = 1, 2, 3$. Generally, the area coordinates can vastly simplify the calculation of element interpolation functions in straight-edge triangular elements. With

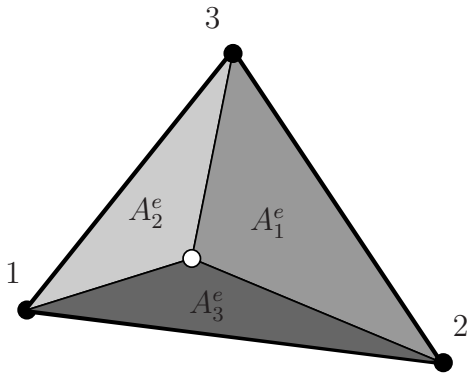


Figure 5.20. *Area coordinates in a triangular domain*

reference to the 6-node element depicted in Figure 5.18, note that the area coordinate representation of node 1 is $(1, 0, 0)$, while that of node 4 is $(\frac{1}{2}, \frac{1}{2}, 0)$. The representation of the edge 2-3 is $L_1 = 0$ (or, equivalently, $L_2 + L_3 = 1$), while that of the line connecting nodes 5 and 6 is $L_3 = \frac{1}{2}$. Given the above, the six element interpolation functions of this element can be expressed in terms of the area coordinates as

$$N_1^e = 2L_1(L_1 - \frac{1}{2}) \quad , \quad N_2^e = 2L_2(L_2 - \frac{1}{2}) \quad , \quad N_3^e = 2L_3(L_3 - \frac{1}{2})$$
$$N_4^e = 4L_1L_2 \quad , \quad N_5^e = 4L_2L_3 \quad , \quad N_6^e = 4L_3L_1 \; . \tag{5.71}$$

The element interpolation functions of other higher-order triangular elements (e.g., the 10-node element in Figure 5.18) may be readily obtained in terms of area coordinates using the procedure outlined above.

An important formula for the integration of polynomial functions of the area coordinates over the region of a straight-edge triangle Ω with area A can be established in the form

$$\int_\Omega L_1^\alpha L_2^\beta L_3^\gamma \, dA = \frac{\alpha! \, \beta! \, \gamma!}{(\alpha + \beta + \gamma + 2)!} \, 2A \; , \tag{5.72}$$

where α, β and γ are any non-negative integers. This formula permits the exact evaluation of integrals associated with the weak form of differential equations, provided that the integrals involve polynomial terms in the area coordinates L_i.

Quadrilateral elements are also used widely in finite element practice. First, attention is focused on the special case of rectangular elements for $p = 1$. The simplest possible such element is the 4-node rectangle of Figure 5.21. Here, it is assumed that the dependent

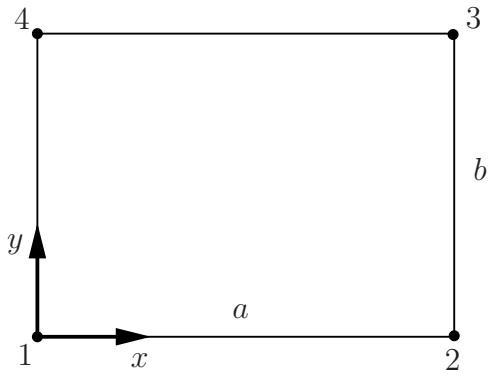


Figure 5.21. *Four-node rectangular element*

variable is interpolated as

$$u_h \;=\; \sum_{i=1}^{4} N_i^e u_i^e \;=\; c_0 + c_1 x + c_2 y + c_3 xy \;, \tag{5.73}$$

where u_i^e, $i = 1 - 4$, are the nodal degrees of freedom (corresponding to the ordinates of the dependent variable at the nodes) and $c_0 - c_3$ are constants. Following the process outlined earlier, one may determine these constants by requiring that

$$\begin{aligned}
u_1^e &= c_0 + c_1 x_1^e + c_2 y_1^e + c_3 x_1^e y_1^e \;, \\
u_2^e &= c_0 + c_1 x_2^e + c_2 y_2^e + c_3 x_2^e y_2^e \;, \\
u_3^e &= c_0 + c_1 x_3^e + c_2 y_3^e + c_3 x_3^e y_3^e \;, \\
u_4^e &= c_0 + c_1 x_4^e + c_2 y_4^e + c_3 x_4^e y_4^e \;.
\end{aligned} \tag{5.74}$$

As before, the solution of the preceding linear system yields expressions for $c_0 - c_3$, which, in turn, can be used in connection with (5.73) to establish expressions for N_i^e, $i = 1 - 4$. However, it is rather simple to deduce these expressions directly by exploiting the fundamental property of the shape functions, namely that they vanish at all nodes except for one where they attain unit value. Indeed, in the case of the 4-node rectangle of Figure 5.21, these functions are given by

$$\begin{aligned}
N_1^e &= \frac{1}{ab}(x-a)(y-b) \;, \\
N_2^e &= -\frac{1}{ab}x(y-b) \;, \\
N_3^e &= \frac{1}{ab}xy \;, \\
N_4^e &= -\frac{1}{ab}(x-a)y \;.
\end{aligned} \tag{5.75}$$

Figure 5.22 depicts a typical interpolation function for the 4-node rectangle.

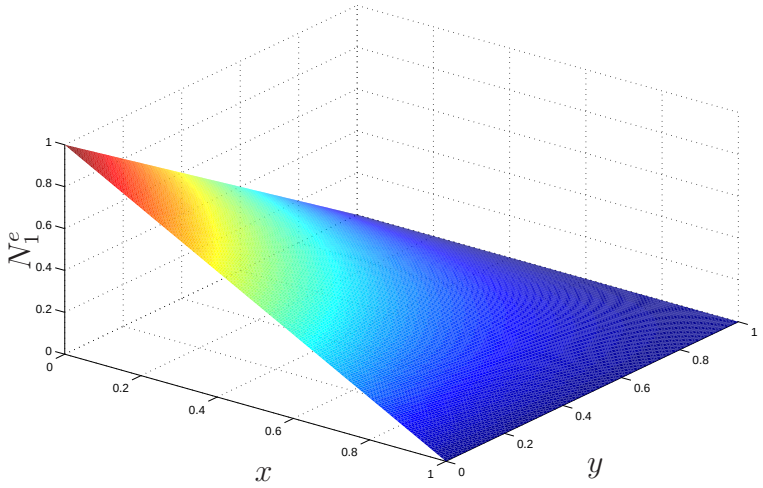


Figure 5.22. *Interpolation function N_1^e for $a = b = 1$ (a hyperbolic paraboloid)*

The completeness property of this element is readily apparent, as one may represent any polynomial with terms $\{1, x, y, xy\}$[2]. Integrability is also guaranteed; indeed, taking any element edge, say, for example, edge 1-2, it is clear that $N_3^e = N_4^e = 0$. Hence, along this edge u_h is a linear function fully determined by the values of u_1^e and u_2^e, which, it turn, are shared with the neighboring element on the other side of edge 1-2.

Higher-order rectangular elements can be divided into two families based on the methodology used to generate them: these are the *serendipity* and the *Lagrangian* elements. The 4-node rectangle is common to both families. The next three elements of the serendipity family are the 8-, 12- and 17-node elements, see Figure 5.23. These elements are polynomially complete to degree $q = 2$, 3, and 4, respectively. The 8-node rectangle may represent any

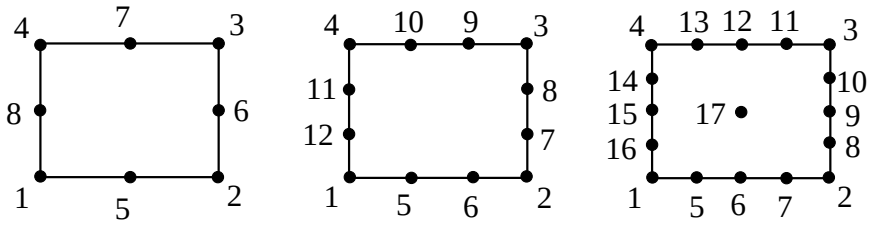


Figure 5.23. *Three members of the serendipity family of rectangular elements*

[2]Note that the degree of completeness is still only $q = 1$ despite the presence of the bilinear term xy.

polynomial with terms $\{1, x, y, x^2, xy, y^2, x^2y, xy^2\}$. This can be either assumed at the outset (following the approach used earlier for the 4-node rectangle) and confirmed by enforcing the restrictions $u_h(x_i, y_i) = u_i^e$, $i = 1 - 8$, or by directly "guessing" the mathematical form of the interpolation functions using their fundamental property. Regrettably, this guessing becomes more difficult for the 12- and the 17-node elements, which explains the characterization of this family as "serendipity". It can be shown that for a rectangular element of the serendipity family with $m+1$ nodes per edge, the represented monomials in Pascal's triangle are as shown in Figure 5.24 before accounting for any interior nodes, such as node 17 in the 17-node element.

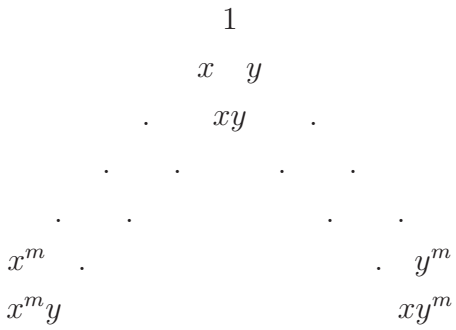


Figure 5.24. *Pascal's triangle for two-dimensional serendipity elements (before accounting for any interior nodes)*

The Lagrangian family of rectangular elements is comprised of the 4-node element discussed earlier, followed by the 9-, 16- and 25-node element, see Figure 5.25. The 9-node rect-

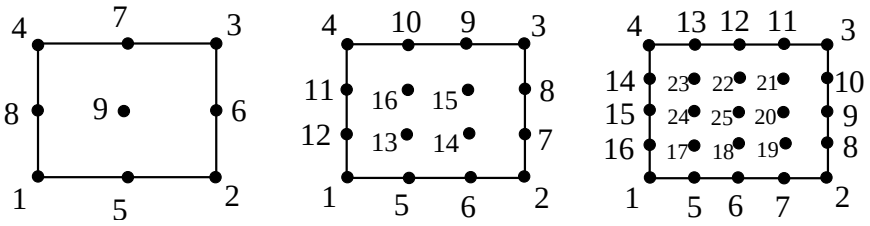


Figure 5.25. *Three members of the Lagrangian family of rectangular elements*

angle is capable of representing any polynomial with terms $\{1, x, y, x^2, xy, y^2, x^2y, xy^2, x^2y^2\}$. In contrast to the serendipity elements, the interpolation functions of the Lagrangian elements can be determined trivially as products of one-dimensional Lagrange interpolation functions. As an example, consider the interpolation function N_{18}^e associated with node 18

of the 25-node element of Figure 5.25. This can be written as

$$N^e_{18} \;=\; l_3(x)l_2(y)\;, \tag{5.76}$$

where

$$\begin{aligned} l_3(x) &= \frac{(x-x_{16})(x-x_{17})(x-x_{19})(x-x_8)}{(x_{18}-x_{16})(x_{18}-x_{17})(x_{18}-x_{19})(x_{18}-x_8)}\;, \\ l_2(y) &= \frac{(y-y_6)(y-y_{25})(y-x_{22})(y-y_{12})}{(y_{18}-y_6)(y_{18}-y_{25})(y_{18}-y_{22})(y_{18}-y_{12})}\;. \end{aligned} \tag{5.77}$$

Again, it is straightforward to see that for a rectangular element of the Lagrangian family with $m+1$ nodes per edge, the represented monomials in Pascal's triangle are as shown in Figure 5.26.

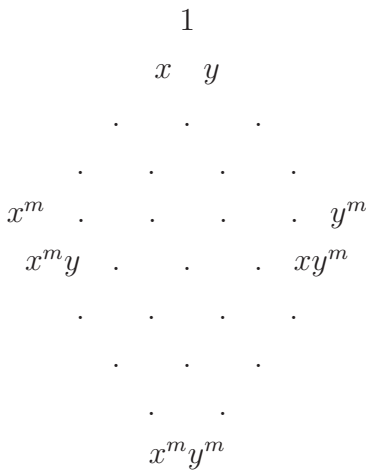


Figure 5.26. *Pascal's triangle for two-dimensional Lagrangian elements*

All serendipity and Lagrangian rectangular elements are invariant under 90^o rotations, meaning that they represent the same monomial terms in x and y. This is clear from the symmetry in x and y of the represented monomials in the associated Pascal triangles of Figure 5.24 and Figure 5.26.

General quadrilaterals, such as the 4-node element in Figure 5.27, present a difficulty. In particular, it is easy to see that if one assumes at the outset a bilinear interpolation as in equation (5.73), then the value of u_h along a given edge generally depends not only on the nodal values at the two end-points of the edge, but also on the other two nodal values, which immediately implies violation of the interelement continuity of u_h. If, on the

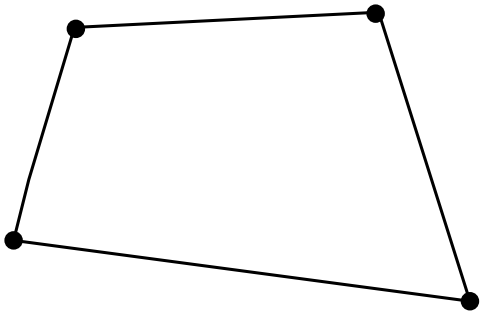


Figure 5.27. *A general quadrilateral finite element domain*

other hand, one constructs a set of interpolation functions that satisfy the fundamental property, then it is easily seen that these functions are not complete to the minimum degree $q = 1$. One simple way to circumvent this difficulty is to construct a composite 4-node rectangle consisting of two connected triangles or a composite 5-node triangle consisting of four connected triangles, as in Figure 5.28. In both cases, the interpolation in each triangle is linearly complete and continuity of the dependent variable is guaranteed at all interelement boundaries. In Section 5.6, the question of general quadrilateral elements will be revisited within the context of the so-called isoparametric mapping.

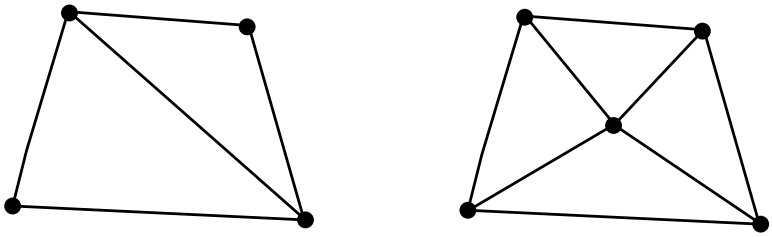


Figure 5.28. *Rectangular finite elements made of two or four joined triangular elements*

The construction of two-dimensional finite elements with $p = 2$ is substantially more complicated than the respective one-dimensional case. To illustrate this point, consider a simple cubically complete interpolation of the dependent variable u_h as

$$u_h \;=\; c_0 + c_1 x + c_2 y + c_3 x^2 + c_4 xy + c_5 y^2 + c_6 x^3 + c_7 x^2 y + c_8 x y^2 + c_9 y^3 \;. \qquad (5.78)$$

One may choose to associate this interpolation with the 3-node triangular element in Figure 5.29. Here, there are three degrees of freedom per node, namely the dependent variable u_h and its two partial derivatives $\dfrac{\partial u_h}{\partial x}$ and $\dfrac{\partial u_h}{\partial y}$. Given that there are 10 unknown coefficients c_i, $i = 0-9$, and only 9 degrees of freedom, one has to either add an extra degree of freedom

or restrict the interpolation. The former may be accomplished by adding a fourth node at the centroid of the triangle and assigning the degree of freedom to be equal to the ordinate of the dependent variable at that point. The latter may be effected by requiring the monomials x^2y and xy^2 to have the same coefficient, namely, $c_7 = c_8$.

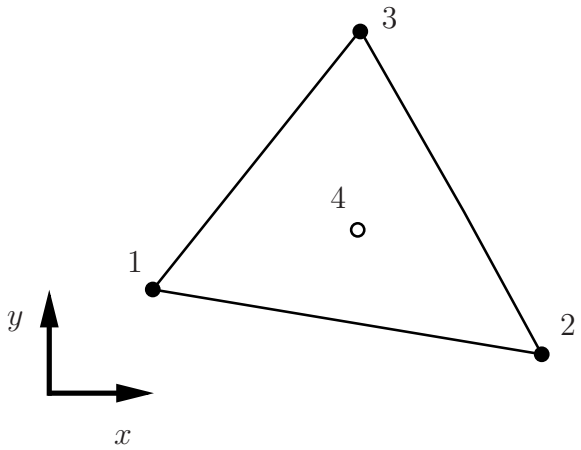


Figure 5.29. *A simple potential 3- or 4-node triangular element for the case* $p = 2$ *(*$u, \dfrac{\partial u}{\partial s}, \dfrac{\partial u}{\partial n}$ *dofs at nodes* $1, 2, 3$ *and, possibly,* u *dof at node* 4*)*

In either case, consider a typical edge, say 1-2, of this element and, without any loss of generality, recast the degrees of freedom associated with this edge relative to the tangential and normal coordinates (s, n), as shown in Figure 5.30. It is clear from the original interpolation assumption that u_h varies cubically in edge 1-2. Hence, given that both u_h and

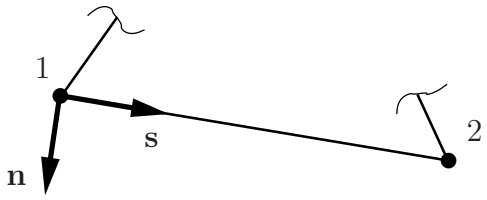


Figure 5.30. *Illustration of violation of the integrability requirement for the 9- or 10-dof triangle for the case* $p = 2$

$\dfrac{\partial u_h}{\partial s}$ are specified on this edge, it follows that u_h, as well as $\dfrac{\partial u_h}{\partial s}$ are continuous across 1-2. However, this is not the case for the normal derivative $\dfrac{\partial u_h}{\partial n}$, which varies quadratically along 1-2, but cannot be determined uniquely from the two normal derivative degrees of freedom on the edge. This implies that $\dfrac{\partial u_h}{\partial n}$ is discontinuous across 1-2, therefore this simple element

violates the integrability requirement for the case $p = 2$. Hence, a direct extension of the one-dimensional Hermitian interpolation-based elements to the two-dimensional case is not permissible. To remedy this problem, one may resort to elements that have mid-edge degrees of freedom, such as the 6-node triangle in Figure 5.31. This element has the previously noted three degrees of freedom at the vertices, as well as a normal derivative degree of freedom at each of the mid-edges.

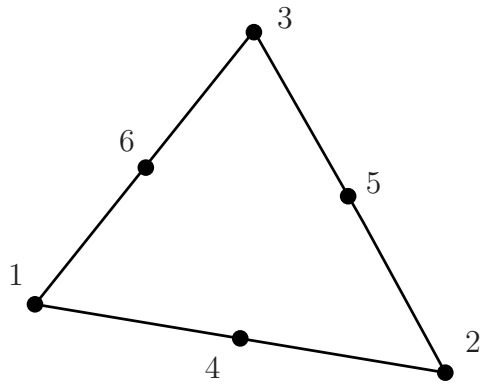


Figure 5.31. *A 12-dof triangular element for the case* $p = 2$ *(*$u, \dfrac{\partial u}{\partial s}, \dfrac{\partial u}{\partial n}$ *dofs at nodes* $1, 2, 3$ *and* $\dfrac{\partial u}{\partial n}$ *at nodes* $4, 5, 6$*)*

The mid-edge nodes of the previous 12-dof element are somewhat undesirable from a data management viewpoint (they have different number of degrees of freedom than vertex nodes), as well as because of the special care needed in order to specify a unique normal to a given edge (otherwise, the shared degree of freedom would be inconsistently interpreted by the two neighboring elements that share it). More importantly, it turns out that this element requires algebraically complex rational polynomial interpolations for the mid-edge degrees of freedom.

Composite triangles, such as the celebrated *Clough-Tocher element*, were developed to circumvent the need for rational polynomial interpolation functions. This element is comprised of three joined triangles, each employing a complete cubic interpolation of u_h, see Figure 5.32. This means that, at the outset, the element has $3 \times 10 = 30$ degrees of freedom. Taking into account that the values of u_h and its two first derivatives $\dfrac{\partial u_h}{\partial x}$ and $\dfrac{\partial u_h}{\partial y}$ are shared at each of the four vertices (three exterior and one interior), the total number of degrees of freedom is immediately reduced to 30-(3×3 + 3×2)=15. At this stage, the normal derivative is not continuous across the internal edges, hence u_h is not internally C^1-continuous. To fix

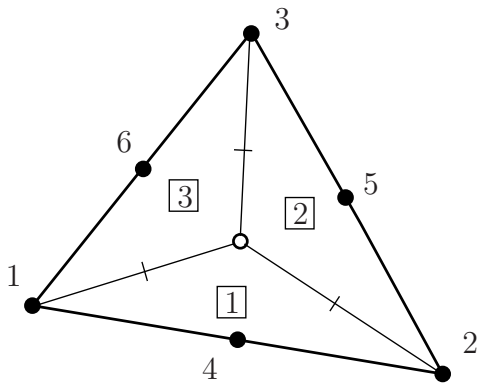


Figure 5.32. *Clough-Tocher triangular element for the case* $p = 2$ *(*$u, \dfrac{\partial u}{\partial s}, \dfrac{\partial u}{\partial n}$ *dofs at nodes* $1, 2, 3$ *and* $\dfrac{\partial u}{\partial n}$ *at nodes* $4, 5, 6$*)*

this problem, Clough and Tocher required that the normal derivative be matched at the mid-point of each internal edge, which further reduces the number of degrees of freedom from 15 to 12. These degrees of freedom are u_h, $\dfrac{\partial u_h}{\partial x}$ and $\dfrac{\partial u_h}{\partial y}$ at the corner nodes and $\dfrac{\partial u_h}{\partial n}$ at the mid-edges. In addition, the mid-edge degrees of freedom may be suppressed by requiring that $\dfrac{\partial u_h}{\partial n}$ at the mid-edges be averaged over the two corresponding corner values, thus leading to a 9 degree-of-freedom element. In either case, the Clough-Tocher element possesses piecewise cubic polynomial interpolation of the dependent variable in each triangular subdomain and satisfies both the integrability and the completeness requirement.

A composite compatible quadrilateral element may be constructed from 4 triangular elements having piecewise cubic interpolations following the procedure used in the derivation of the preceding Clough-Tocher triangular element.

There are numerous triangular and quadrilateral elements for the case $p = 2$. However, their use has gradually diminished in finite element practice. For this reason, they will not be discussed here.

5.5.3 Interpolations in three dimensions

In this section, three dimensional polynomial interpolations are considered in connection with tetrahedral, pentahedral and hexahedral elements.

The simplest three-dimensional element is the 4-node tetrahedron with one node at each vertex, see Figure 5.33. This element has one degree-of-freedom at each node and the

dependent variable is interpolated as

$$u_h \;=\; \sum_{i=1}^{4} N_i^e u_i^e \;=\; c_0 + c_1 x + c_2 y + c_3 z \;, \tag{5.79}$$

where u_i^e, $i = 1 - 4$, are the nodal degrees of freedom and $c_0 - c_3$ are constants. Recalling again that $u_i^e = u_h(x_i^e, y_i^e, z_i^e)$, that is, the degrees of freedom take the values of the ordinates of the depended variable at nodes i with coordinates (x_i^e, y_i^e, z_i^e), it follows that the constants $c_0 - c_3$ can be determined by solving the system of equations

$$\begin{aligned}
u_1^e &= c_0 + c_1 x_1^e + c_2 y_1^e + c_3 z_1^e \;, \\
u_2^e &= c_0 + c_1 x_2^e + c_2 y_2^e + c_3 z_2^e \;, \\
u_3^e &= c_0 + c_1 x_3^e + c_2 y_3^e + c_3 z_3^e \;, \\
u_4^e &= c_0 + c_1 x_4^e + c_2 y_4^e + c_3 z_4^e \;.
\end{aligned} \tag{5.80}$$

Clearly, this element is polynomially complete to degree $q = 1$. In addition, it is easy to show that this element is suitable for approximating weak forms in which $p = 1$, that is, it satisfies the integrability condition for this class of weak forms.

Higher-order tetrahedral elements are possible and, in fact, often used in engineering practice. The next element in this hierarchy is the 10-node tetrahedron with nodes added to each of the six mid-edges. This element is polynomially complete to degree $q = 2$ and can exactly represent any polynomial function consisting of the monomial terms $\{1, x, y, z, x^2, y^2, z^2, xy, yz, zx\}$, see Figure 5.34.

The task of deducing analytical representations of the element interpolation functions N_i^e for tetrahedra is vastly simplified by the introduction of *volume coordinates*, in complete analogy to the area coordinates employed for triangular elements in two dimensions.

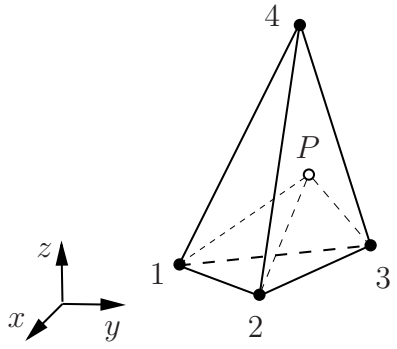


Figure 5.33. *The 4-node tetrahedral element*

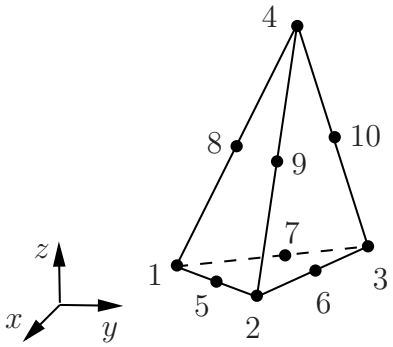


Figure 5.34. *The 10-node tetrahedral element*

With reference to the 4-node tetrahedral element of Figure 5.33, one may define the volume coordinate L_i of a typical point P of the tetrahedron as

$$L_i \;=\; \frac{V_i}{V} \quad , \quad i = 1 - 4 \; , \tag{5.81}$$

where V_i is the volume of the tetrahedron formed by the point P and the face opposite to node i, while V is the volume of the full tetrahedron. It is readily obvious that $L_1 + L_2 + L_3 + L_4 = 1$, hence only three of the volume coordinates are independently specified. Also, with reference to the 4-node tetrahedron, it follows that $N_i^e = L_i$, $i = 1 - 4$. Element interpolation functions for higher-order tetrahedra can be derived with great ease using volume coordinates. Furthermore, when evaluating integral terms over a tetrahedral region Ω of volume V, one may employ a convenient formula, according to which

$$\int_\Omega L_1^\alpha L_2^\beta L_3^\gamma L_4^\delta \, dV \;=\; \frac{\alpha!\,\beta!\,\gamma!\,\delta!}{(\alpha \,+\, \beta \,+\, \gamma \,+\, \delta \,+\, 3)!}\, 6V \; , \tag{5.82}$$

where α, β, γ, and δ are non-negative integers.

The first two pentahedral elements of interest are the 6-node and the 15-node pentahedron, shown in Figure 5.35. The former is complete up to polynomial degree $q = 1$ and its interpolation functions are capable of representing the monomial terms $\{1, x, y, z, xz, yz\}$. The latter is complete up to polynomial degree $q = 2$ and its interpolation functions may independently reproduce the monomials $\{1, x, y, x^2, xy, y^2, z, xz, yz, x^2z, xyz, y^2z, z^2, xz^2, yz^2\}$. It is worth noting that the interpolation functions of a pentahedral element are products of the triangle-based functions of the top and bottom (triangular) faces and the rectangle-based functions of the lateral (rectangular) faces.

Hexahedral elements are widely used in three-dimensional finite element analyses. The simplest such element is the 8-node hexahedron with nodes at each of its vertices, see Figure 5.36. This element is polynomially complete up to degree $q = 1$ and its interpolation

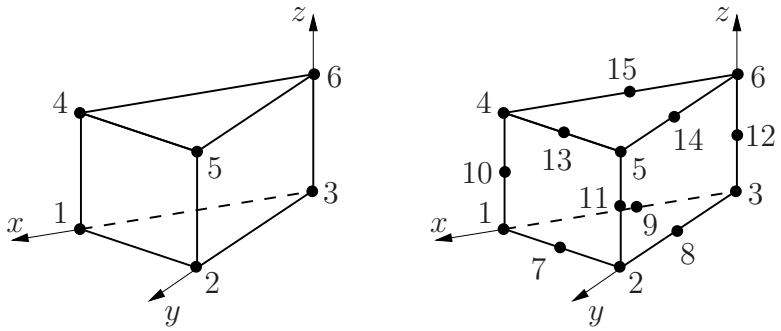


Figure 5.35. *The 6- and 15-node pentahedral elements*

functions are capable of representing any polynomial consisting of $\{1, x, y, z, xy, yz, zx, xyz\}$. The element interpolation functions of the *orthogonal* 8-node hexahedron of Figure 5.36, can be written by inspection as

$$\begin{aligned}
N_1^e &= -\frac{1}{abc}(x-a)(y-b)(z-c) \ , \\
N_2^e &= \frac{1}{abc}(x-a)y(z-c) \ , \\
N_3^e &= -\frac{1}{abc}(x-a)yz \ , \\
N_4^e &= \frac{1}{abc}(x-a)(y-b)z \ , \\
N_5^e &= \frac{1}{abc}x(y-b)(z-c) \ , \\
N_6^e &= -\frac{1}{abc}xy(z-c) \ , \\
N_7^e &= \frac{1}{abc}xyz \ , \\
N_8^e &= -\frac{1}{abc}x(y-b)z \ .
\end{aligned} \tag{5.83}$$

The next two useful hexahedral elements are the 20- and the 27-node element, see Figure 5.37. These can be viewed as the three-dimensional members of the serendipity and Lagrangian family for the case of polynomial completeness of order $q = 2$. The interpolation functions of the 20-node hexahedron can independently represent the monomials

$$\{1, x, y, z, x^2, y^2, z^2, xy, yz, zx, xyz, xy^2, xz^2, yz^2, yx^2, zx^2, zy^2, x^2yz, y^2zx, z^2xy\} \ ,$$

while the interpolation functions of the 27-node hexahedron can additionally represent the monomials

$$\{x^2y^2, y^2z^2, z^2x^2, x^2y^2z, y^2z^2x, z^2x^2y, x^2y^2z^2\} \ .$$

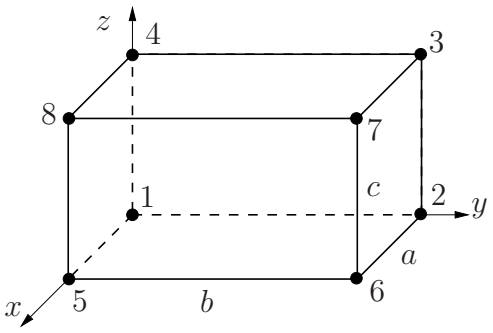


Figure 5.36. *The 8-node hexahedral element*

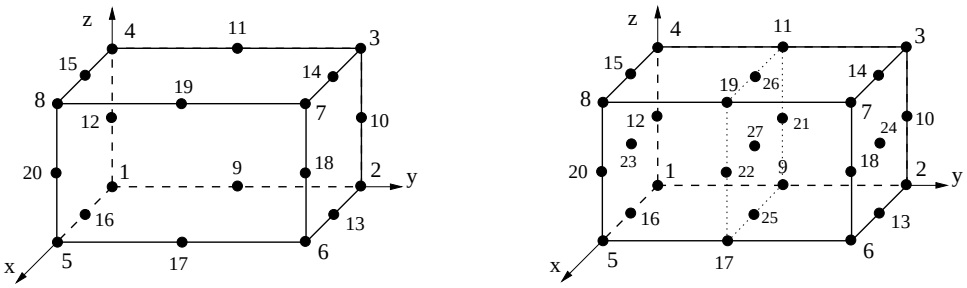


Figure 5.37. *The 20- and 27-node hexahedral elements*

There exist various three-dimensional elements for the case $p = 2$. However, they will not be discussed here owing to their limited usefulness.

5.6 The concept of isoparametric mapping

In finite element practice, one often distinguishes between analyses conducted on *structured* or *unstructured* meshes. The former are applicable to domains that are very regular, such as rectangles, cubes, etc, and which can be subdivided into equal-sized elements, themselves having a regular shape. The latter is encountered in the discretization of complex two- and three-dimensional domains, where it is frequently essential to use elements with "irregular" shapes, such as arbitrary straight-edge quadrilaterals, curved-edge triangles and quadrilaterals, etc. For these cases, it becomes extremely important to establish a general methodology for constructing irregular-shaped elements which satisfy the appropriate completeness and integrability requirements.

The concept of isoparametric mapping offers precisely the means for constructing irregular-shaped elements that inherit the well-established completeness and integrability properties of their regular-shaped counterparts. The original conception of this mapping is due to Ian Taig[3], which was later expanded and formalized by Bruce Irons[4] The main idea of the isoparametric mapping is to construct the irregularly-shaped element in the *physical* domain (namely, the domain of interest) as a mapping from a *parent* (or *natural*) domain, in which this same element has a regular shape. This mapping can be expressed in three-dimensions as

$$x \;=\; \hat{x}(\xi,\eta,\zeta) \quad , \quad y \;=\; \hat{y}(\xi,\eta,\zeta) \quad , \quad z \;=\; \hat{z}(\xi,\eta,\zeta)\;, \tag{5.84}$$

where (ξ,η,ζ) and (x,y,z) are coordinates in the natural and physical domain, respectively. The mapping of equations (5.84) can be equivalently (and more succinctly) represented in vector form as

$$\mathbf{x} \;=\; \boldsymbol{\phi}(\boldsymbol{\xi})\;, \tag{5.85}$$

where (x,y,z) and (ξ,η,ζ) are Cartesian components of $\mathbf{x}$ and $\boldsymbol{\xi}$, respectively. Here, $\boldsymbol{\phi}$ maps the regular-shaped domain $\Omega^e_\square$ to the irregular-shaped domain Ω^e, see Figure 5.38. By way of background, the mapping $\boldsymbol{\phi}$ is termed *one-to-one* (or *injective*) if for any two distinct points $\boldsymbol{\xi}_1 \neq \boldsymbol{\xi}_2$ in $\Omega^e_\square$, their images $\mathbf{x}_1$ and $\mathbf{x}_2$ under $\boldsymbol{\phi}$ satisfy $\mathbf{x}_1 \neq \mathbf{x}_2$. Further, the mapping $\boldsymbol{\phi}$ is termed *onto* (or *surjective*) if $\boldsymbol{\phi}(\Omega^e_\square) = \Omega^e$, or, said equivalently, any point $\mathbf{x} \in \Omega^e$ is the image of some point $\boldsymbol{\xi} \in \Omega^e_\square$.

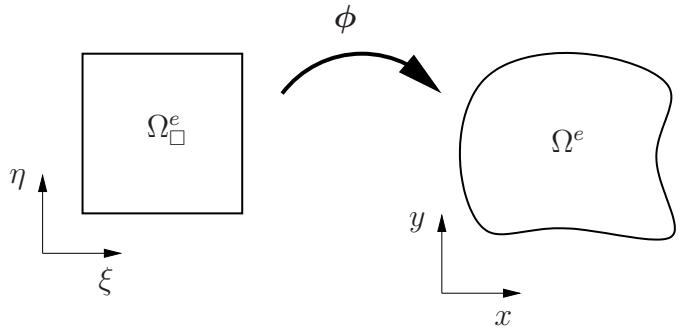


Figure 5.38. *Schematic of a parametric mapping from $\Omega^e_\square$ to Ω^e*

In order to define what constitutes an isoparametric mapping, let the dependent variable u

[3]Ian Taig (1927-?) was a British aerospace engineer.

[4]Bruce Irons (1924-1983) was a British physicist, mathematician, and engineer.

be approximated in the element Ω^e of interest as

$$u_h^e \;=\; \sum_{i=1}^{n} N_i^e u_i^e \;, \tag{5.86}$$

where u_i^e, $i = 1, 2, \ldots, n$, are the element degrees of freedom. Likewise, suppose that the geometry of the element Ω^e is defined by the equations

$$\mathbf{x} \;=\; \sum_{j=1}^{m} N_j^e \mathbf{x}_j^e \;, \tag{5.87}$$

where $\mathbf{x}_j^e$, $j = 1, 2, \ldots, m$, are the coordinates of element nodes. It is important to stress that in the preceding equations, the interpolation functions N_i^e and N_j^e are identical for $i = j$ and they are defined on $\Omega_\square^e$, namely they are functions of the natural coordinates (ξ, η, ζ).

With reference to equations (5.86) and (5.87), a finite element is termed *isoparametric* if $n = m$. Otherwise, it is called *subparametric* if $n > m$ or *superparametric* if $n < m$. From the foregoing definition, it follows that in isoparametric elements the exact same functions are employed to define the element geometry and the interpolation of the dependent variable. As stated earlier in this section, this justifies the term *shape functions* which is frequently used in finite element literature as an alternative to "element interpolation functions". The implications of the isoparametric assumption will become apparent in the ensuing developments.

One of the key questions associated with all parametric finite elements is whether the mapping $\boldsymbol{\phi}$, expressed here through equations $(5.90)_2$, is invertible. Said differently, the relevant question is whether one may uniquely associate points $(\xi, \eta, \zeta) \in \Omega_\square^e$ with points $(x, y, z) \in \Omega^e$ and vice-versa. This question is addressed by the *inverse function theorem*, which, when adapted to the context of this problem may be stated as follows: Consider a mapping $\boldsymbol{\phi} : \Omega_\square^e \mapsto \Omega^e$ of class C^r, such that $\boldsymbol{\xi} \in \Omega_\square^e$ is mapped to $\mathbf{x} = \boldsymbol{\phi}(\boldsymbol{\xi}) \in \Omega^e$, where $\Omega_\square^e$ and Ω^e are open sets. If $J = \det \dfrac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\xi}} \neq 0$ at a point $\bar{\boldsymbol{\xi}} \in \Omega_\square^e$, then there is an open neighborhood around $\bar{\boldsymbol{\xi}}$, such that $\boldsymbol{\phi}$ is one-to-one and onto an open subset of Ω^e containing the point $\bar{\mathbf{x}} = \boldsymbol{\phi}(\bar{\boldsymbol{\xi}})$ and the inverse function $\boldsymbol{\phi}^{-1}$ exists and is of class C^r. The derivative $\mathbf{J} = \dfrac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\xi}}$ can be written in matrix form as

$$[\mathbf{J}] \;=\; \left[\begin{array}{ccc} \dfrac{\partial \boldsymbol{\phi}}{\partial \xi} & \dfrac{\partial \boldsymbol{\phi}}{\partial \eta} & \dfrac{\partial \boldsymbol{\phi}}{\partial \zeta} \end{array}\right] \;=\; \left[\begin{array}{ccc} \dfrac{\partial \hat{x}}{\partial \xi} & \dfrac{\partial \hat{x}}{\partial \eta} & \dfrac{\partial \hat{x}}{\partial \zeta} \\ \dfrac{\partial \hat{y}}{\partial \xi} & \dfrac{\partial \hat{y}}{\partial \eta} & \dfrac{\partial \hat{y}}{\partial \zeta} \\ \dfrac{\partial \hat{z}}{\partial \xi} & \dfrac{\partial \hat{z}}{\partial \eta} & \dfrac{\partial \hat{z}}{\partial \zeta} \end{array}\right] , \tag{5.88}$$

and is referred to as the *Jacobian matrix* of the isoparametric transformation. The inverse function theorem guarantees that every interior point $(x, y) \in \Omega^e$ is uniquely associated with a single point $(\xi, \eta) \in \Omega^e_\square$ provided that the determinant J is non-zero everywhere in $\Omega^e_\square$.[5]

By way of a concrete example, consider in detail the isoparametric 4-node quadrilateral element of Figure 5.39. The element interpolation functions in the parent domain are given by

$$\begin{aligned} N^e_1(\xi,\eta) &= \frac{1}{4}(1-\xi)(1-\eta) \ , \\ N^e_2(\xi,\eta) &= \frac{1}{4}(1+\xi)(1-\eta) \ , \\ N^e_3(\xi,\eta) &= \frac{1}{4}(1+\xi)(1+\eta) \ , \\ N^e_4(\xi,\eta) &= \frac{1}{4}(1-\xi)(1+\eta) \ . \end{aligned} \tag{5.89}$$

Given that the element is isoparametric, it follows that

$$u^e_h = \sum_{i=1}^{4} N^e_i u^e_i \qquad , \qquad \mathbf{x} = \sum_{i=1}^{4} N^e_i \mathbf{x}^e_i \ , \tag{5.90}$$

where $\mathbf{x}^e_i$ are the vectors with coordinates (x^e_i, y^e_i) pointing to the positions of the four nodes $1, 2, 3, 4$ in the physical domain.

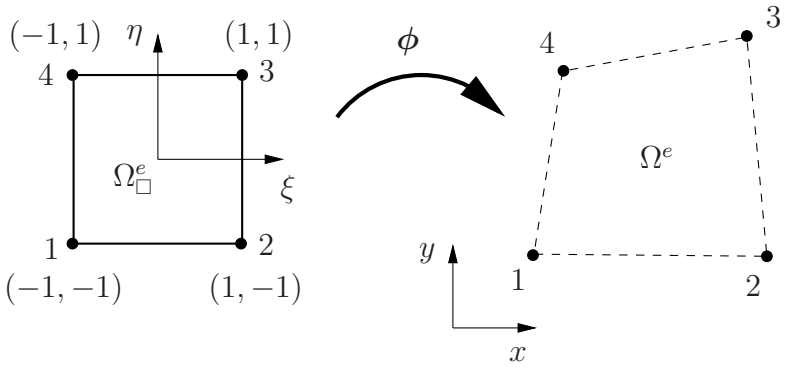


Figure 5.39. *The 4-node isoparametric quadrilateral*

First, verify that the edges of the element in the physical domain are straight. To this end, consider a typical edge, say 1-2: clearly, this edge corresponds in the parent domain

[5] Note that here ϕ is of class C^∞.

to $\xi \in (-1,1)$ and $\eta = -1$. In view of (5.89) and $(5.90)_2$, this means that the equations describing the edge 1-2 are:

$$\begin{aligned} x &= \frac{1}{2}(1-\xi)x_1^e + \frac{1}{2}(1+\xi)x_2^e = \frac{1}{2}(x_1^e + x_2^e) + \frac{1}{2}\xi(x_2^e - x_1^e) \ , \\ y &= \frac{1}{2}(1-\xi)y_1^e + \frac{1}{2}(1+\xi)y_2^e = \frac{1}{2}(y_1^e + y_2^e) + \frac{1}{2}\xi(y_2^e - y_1^e) \ . \end{aligned} \tag{5.91}$$

The above are parametric equations of a straight line passing through points (x_1^e, y_1^e) and (x_2^e, y_2^e), namely through nodes 1 and 2, which proves the original assertion. Hence, the mapped domain Ω^e is a quadrilateral with straight edges.

Next, establish the completeness and integrability properties of this element. Starting with the former, note that for completeness to polynomial degree $q = 1$, the interpolation of equation $(5.90)_1$ needs to be able to exactly represent any polynomial of the form

$$u_h = c_0 + c_1 x + c_2 y \ . \tag{5.92}$$

However, equation $(5.90)_1$ implies that, if the four degrees of freedom u_i^e coincide with the nodal values of u_h, then setting u_h at the four nodes according to $(5.92)_1$ yields

$$\begin{aligned} u_h &= \sum_{i=1}^4 N_i^e u_i^e = \sum_{i=1}^4 N_i^e u_h(x_i^e, y_i^e) = \sum_{i=1}^4 N_i^e (c_0 + c_1 x_i^e + c_2 y_i^e) \\ &= (\sum_{i=1}^4 N_i^e) c_0 + (\sum_{i=1}^4 N_i^e x_i^e) c_1 + (\sum_{i=1}^4 N_i^e y_i^e) c_2 = (\sum_{i=1}^4 N_i^e) c_0 + c_1 x + c_2 y \ , \end{aligned} \tag{5.93}$$

where equation $(5.90)_2$ is used. In view of equation (5.92), completeness of the 4-node isoparametric quadrilateral is guaranteed as long as $\sum_{i=1}^4 N_i^e = 1$, which can be easily verified from equations (5.89).

Integrability for the case $p = 1$ can be established as follows: consider a typical element edge, say 1-2, along which

$$u_h(\xi, -1) = \frac{1}{2}(1-\xi)u_1^e + \frac{1}{2}(1+\xi)u_2^e \ , \tag{5.94}$$

as readily seen from equations (5.89) and $(5.90)_1$. The preceding expression confirms that the value of u_h along edge 1-2 is a linear function of the variable ξ and depends solely on the nodal values of u_h at nodes 1 and 2. Since the values of u_h at these nodes are shared between the two contiguous elements along the edge 1-2, this implies continuity of u_h across 1-2, which is a sufficient condition for integrability.

Reducing equation (5.88) to the two dimensional case, the Jacobian determinant J is given by

$$J \;=\; \frac{\partial \hat{x}}{\partial \xi}\frac{\partial \hat{y}}{\partial \eta} - \frac{\partial \hat{y}}{\partial \xi}\frac{\partial \hat{x}}{\partial \eta}\;, \tag{5.95}$$

which, taking into account equation $(5.90)_2$, leads, after some elementary algebra, to

$$\begin{aligned} J \;=\; \frac{1}{8}\big[&(x_1^e y_2^e - x_2^e y_1^e + x_2^e y_3^e - x_3^e y_2^e + x_3^e y_4^e - x_4^e y_3^e + x_4^e y_1^e - x_1^e y_4^e) \\ &+ \xi(x_1^e y_4^e - x_4^e y_1^e + x_2^e y_3^e - x_3^e y_2^e + x_3^e y_1^e - x_1^e y_3^e + x_4^e y_2^e - x_2^e y_4^e) \\ &+ \eta(x_1^e y_3^e - x_3^e y_1^e + x_2^e y_1^e - x_1^e y_2^e + x_3^e y_4^e - x_4^e y_3^e + x_4^e y_2^e - x_2^e y_4^e)\big]\;. \end{aligned} \tag{5.96}$$

It is instructive to observe here that since J is linear in ξ and η, then $J > 0$ everywhere in

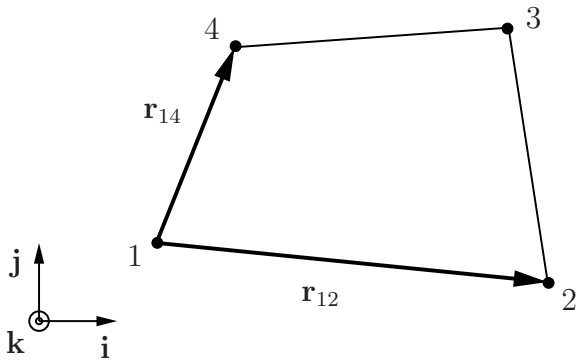


Figure 5.40. *Geometric interpretation of one-to-one isoparametric mapping in the 4-node quadrilateral*

the interior of the domain $\Omega^e_\square$ provided $J > 0$ at all four nodal points. Now, consider, say, node 1, with natural coordinates $(-1,-1)$ and conclude from equation (5.96) that at this node

$$J(-1,-1) \;=\; \frac{1}{4}\big[(x_2^e - x_1^e)(y_4^e - y_1^e) - (x_4^e - x_1^e)(y_2^e - y_1^e)\big]\;. \tag{5.97}$$

It follows from the above equation that $J > 0$ if the physical domain Ω^e is convex at node 1. This is because, with reference to Figure 5.40, one may interpret the Jacobian determinant at node 1 according to

$$4J\mathbf{k} \;=\; \mathbf{r}_{12} \times \mathbf{r}_{14}\;, \tag{5.98}$$

where $\mathbf{r}_{ij}$ denotes the vector connecting nodes i and j and $\mathbf{k}$ is the unit vector normal to the plane of the element, such that $(\mathbf{r}_{12}, \mathbf{r}_{14}, \mathbf{k})$ form a right-handed triad. An analogous conclusion can be drawn for the other three nodes. Hence, invertibility of the isoparametric

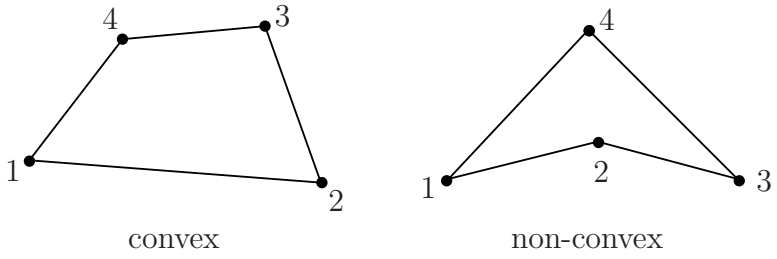


Figure 5.41. *Convex and non-convex 4-node quadrilateral element domains*

mapping for the 4-node quadrilateral is guaranteed as long as the element domain Ω^e is convex, see Figure 5.41.

It is easy to see that the isoparametric mapping in Figure 5.39 is orientation-preserving, in the sense that the nodal sequencing (say 1-2-3-4, if following a counter-clockwise convention) is preserved under the mapping $\boldsymbol{\phi}$. While this orientation preservation property is not essential, it is typically adopted in finite element practice. Reversal of the node sequencing, say from (1-2-3-4) to (1-4-3-2) when following a counterclockwise convention implies that $J < 0$. This can be immediately seen using the foregoing interpretation of the Jacobian determinant at nodal points.

An additional property of the Jacobian determinant J, which becomes important when evaluating integrals emanating from weak forms, is now deduced from the preceding analysis of the 4-node quadrilateral. To this end, start from $(5.84)_{1,2}$ and note that

$$dx \;=\; \frac{\partial \hat{x}}{\partial \xi} d\xi + \frac{\partial \hat{x}}{\partial \eta} d\eta \qquad , \qquad dy \;=\; \frac{\partial \hat{y}}{\partial \xi} d\xi + \frac{\partial \hat{y}}{\partial \eta} d\eta \; . \tag{5.99}$$

Hence, with reference to Figure 5.42, the infinitesimal vector area $d\mathbf{A}$ is written as

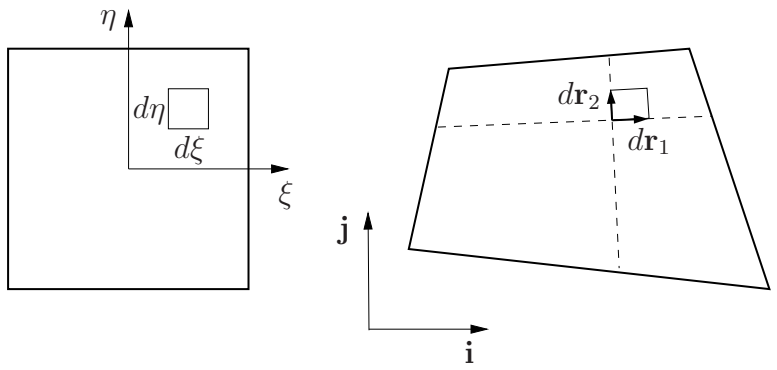


Figure 5.42. *Relation between area elements in the natural and physical domain*

$$d\mathbf{A} \;=\; d\mathbf{r}_1 \times d\mathbf{r}_2 \; , \tag{5.100}$$

where $d\mathbf{r}_1$ and $d\mathbf{r}_2$ are the infinitesimal vectors along lines of constant η and ξ, respectively. This and equations (5.99) imply that

$$d\mathbf{A} \;=\; (\frac{\partial \hat{x}}{\partial \xi} d\xi \mathbf{i} + \frac{\partial \hat{y}}{\partial \xi} d\xi \mathbf{j}) \times (\frac{\partial \hat{x}}{\partial \eta} d\eta \mathbf{i} + \frac{\partial \hat{y}}{\partial \eta} d\eta \mathbf{j}) \;=\; J d\xi d\eta \mathbf{k} \;, \tag{5.101}$$

where $(\mathbf{i}, \mathbf{j})$ are unit vectors along the x- and y-axis, respectively. It follows from the above equation that the infinitesimal area element dA in the physical domain is related to the infinitesimal area element $d\xi d\eta$ in the natural domain as

$$dA \;=\; J d\xi d\eta \;. \tag{5.102}$$

Note that the above argument is not specific to the isoparametric mapping of the 4-node quadrilateral element, hence it applies to all planar isoparametric elements. Furthermore, this argument can be easily extended to three-dimensional elements (where $dV = J d\xi d\eta d\zeta$) or restricted to one-dimensional elements (where, say, $dx = J d\xi$) of the isoparametric type.

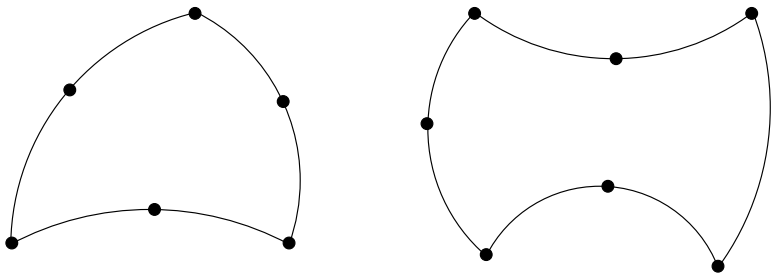


Figure 5.43. *Isoparametric 6-node triangle and 8-node quadrilateral*

The isoparametric approach can be applied to triangles and quadrilaterals without appreciable complication over what has been described for the 4-node quadrilateral. For example, higher-order planar isoparametric elements can be constructed based on the 6-node triangle and the 8-node serendipity rectangle, see Figure 5.43. Both elements may have curved boundaries, which is a desirable feature when modeling arbitrary domains.

Three-dimensional isoparametric elements are also possible and, in fact, quite popular. The simplest such hexahedral element is the 8-node isoparametric brick of Figure 5.44. The geometry of this element is defined in terms of the position vectors $\mathbf{x}_i^e$ of its eight vertex nodes and the corresponding interpolation functions N_i^e in the natural domain. The latter

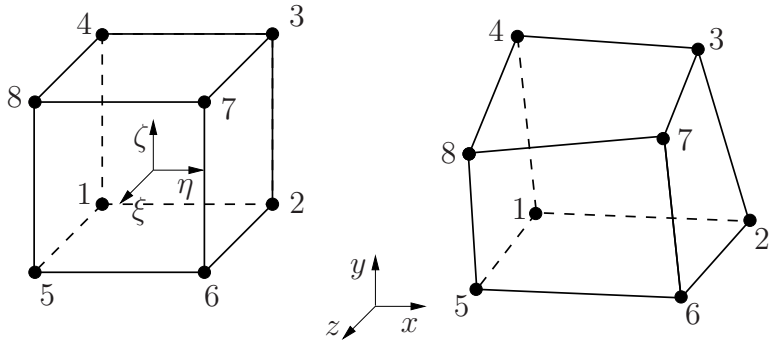


Figure 5.44. *Isoparametric 8-node hexahedral element*

can be written relative to the coordinate system shown in Figure 5.44 as

$$\begin{aligned}
N_1^e &= \frac{1}{4}(1-\xi)(1-\eta)(1-\zeta) \quad , \quad N_2^e = \frac{1}{4}(1-\xi)(1+\eta)(1-\zeta) \ , \\
N_3^e &= \frac{1}{4}(1-\xi)(1+\eta)(1+\zeta) \quad , \quad N_4^e = \frac{1}{4}(1-\xi)(1-\eta)(1+\zeta) \ , \\
N_5^e &= \frac{1}{4}(1+\xi)(1-\eta)(1-\zeta) \quad , \quad N_6^e = \frac{1}{4}(1+\xi)(1+\eta)(1-\zeta) \ , \\
N_7^e &= \frac{1}{4}(1+\xi)(1+\eta)(1+\zeta) \quad , \quad N_8^e = \frac{1}{4}(1+\xi)(1-\eta)(1+\zeta) \ .
\end{aligned} \tag{5.103}$$

All element edges in the 8-node isoparametric brick are straight. Indeed, a typical edge, say 8-7 with coordinates $(1, \eta, 1)$, is described by the equations

$$\begin{aligned}
x &= \frac{1}{2}(x_7^e + x_8^e) + \frac{1}{2}(x_7^e - x_8^e)\eta \ , \\
y &= \frac{1}{2}(y_7^e + y_8^e) + \frac{1}{2}(y_7^e - y_8^e)\eta \ , \\
z &= \frac{1}{2}(z_7^e + z_8^e) + \frac{1}{2}(z_7^e - z_8^e)\eta \ ,
\end{aligned} \tag{5.104}$$

which are precisely the parametric equations of a straight line passing through the nodal points 7 and 8 with coordinates (x_7^e, y_7^e, z_7^e) and (x_8^e, y_8^e, z_8^e), respectively. However, element faces are not necessarily flat. To argue this point, take a typical face, say 8-7-4-3 with coordinates $(\xi, \eta, 1)$ and note that it is defined by the equations

$$\begin{aligned}
x &= \frac{1}{4}(1-\xi)(1+\eta)x_3^e + \frac{1}{4}(1-\xi)(1-\eta)x_4^e + \frac{1}{4}(1+\xi)(1+\eta)x_7^e + \frac{1}{4}(1+\xi)(1-\eta)x_8^e \ , \\
y &= \frac{1}{4}(1-\xi)(1+\eta)y_3^e + \frac{1}{4}(1-\xi)(1-\eta)y_4^e + \frac{1}{4}(1+\xi)(1+\eta)y_7^e + \frac{1}{4}(1+\xi)(1-\eta)y_8^e \ , \\
x &= \frac{1}{4}(1-\xi)(1+\eta)z_3^e + \frac{1}{4}(1-\xi)(1-\eta)z_4^e + \frac{1}{4}(1+\xi)(1+\eta)z_7^e + \frac{1}{4}(1+\xi)(1-\eta)z_8^e \ ,
\end{aligned} \tag{5.105}$$

which contain a bilinear term $\xi\eta$ responsible for the non-flatness of the resulting surface.[6]

As in the case of planar elements, it is straightforward to formulate higher-order three-dimensional isoparametric elements based, e.g., on the 10-node tetrahedron or the 20-node brick. In general, these higher-order elements have both curved edges and non-flat faces.

5.7 Exercises

Problem 1

Write the shape functions for a complete and integrable ($p = 1$) cubic one-dimensional element in the domain $(0, 1)$ using standard and hierarchical interpolation.

Problem 2

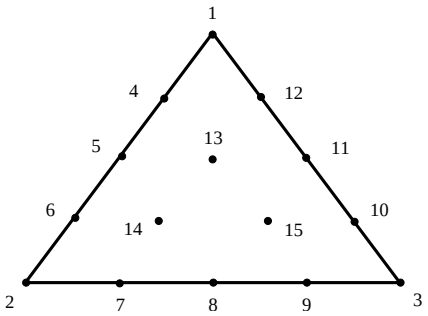
Suppose that the differential equation $\dfrac{d^2u}{dx^2} = 1$ is solved in a given domain using the Bubnov-Galerkin method with two-node hierarchically interpolated quadratic elements subject to some boundary conditions. Consider a typical element e of this approximation, as in Figure 5.15 of the notes, and let the hierarchical interpolation be as in equation (5.52). Solve the differential equation *locally* in the domain of this element and express the hierarchical degree of freedom α^e as a function of the element length Δx. Why isn't α^e also a function of the nodal degrees of freedom u_1^e and u_2^e in this case?

Problem 3

Write the shape functions for a 10-node triangular element using area coordinates.

Problem 4

Write the interpolation functions N_i^e, $I = 1 - 15$, for the 15-node triangle using area coordinates. Also, compute the integral $\int_{\Omega^e} N_1 N_6 \, dA$ over the region Ω_e of an equilateral triangle with unit side. What is the degree of polynomial completeness of this element?



[6] Another way of arguing the same point is to simply note that the element edge needs to pass through 4 points which do not necessarily lie on the same plane.

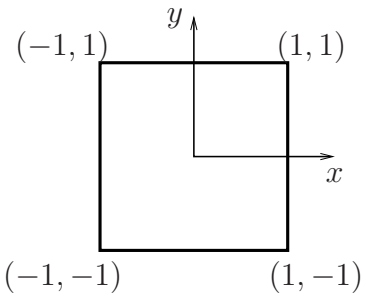
Problem 5

Find the shape functions for the following elements of square domain as in the adjacent figure:

(a) the 8- and 12-node members of the serendipity family,

(b) the 9- and 16-node members of the Lagrangian family,

(c) the hierarchically interpolated 4-node quadratic element corresponding to the 9-node member of the Lagrangian family.



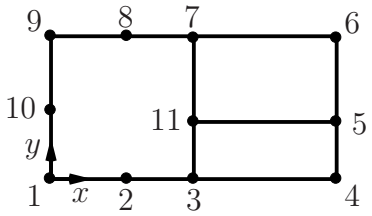
Problem 6

Consider a non-rectangular 4-node quadrilateral element with standard polynomial shape functions. Give an example showing that in this element the values of the dependent variable along an edge generally depend not only on its values at the two nodes defining the edge, but also on its values at one or both of the other nodes. This observation verifies that the arbitrarily-shaped 4-node quadrilateral typically yields incompatible polynomial finite element approximations for problems with $p = 1$.

Hint: to avoid lengthy calculations, consider a simple non-rectangular element shape for your analysis.

Problem 7

A finite element analysis is performed with rectangular serendipity elements. At a certain location in the domain, the mesh pattern of the figure is desired. Which restriction should be imposed on the nodal values of the dependent variable u_h, so as to maintain compatibility of the admissible field $\mathcal{U}$ for problems with $p = 1$? State precisely the required condition.



Problem 8

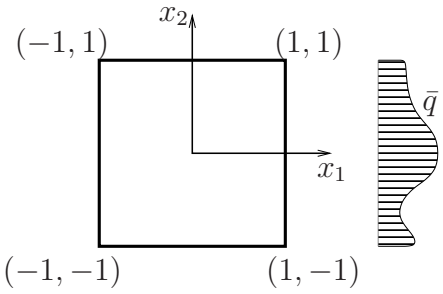
Consider the boundary-value problem of *Problem 2* in Chapter 3, which is to be solved for $k = 1$ and $f = 0$ by a Bubnov-Galerkin approximation. The domain Ω is discretized using finite elements with rectangular domains Ω^e, such that in each element $u_h(\Omega^e) = \sum_i N_i^e u_i^e$ and $w_h(\Omega^e) = \sum_i N_i^e w_i^e$. Subsequently, the weak form of the problem is written for a typical element as

$$\sum_i w_i^e \Big(\sum_j K_{ij}^e u_j^e \; - \; F_i^e \Big) \; = \; 0 \; ,$$

where $[K_{ij}^e]$ is the element stiffness matrix and $[F_i^e]$ is the forcing vector due to the non-vanishing Neumann boundary conditions on $\partial\Omega^e$.

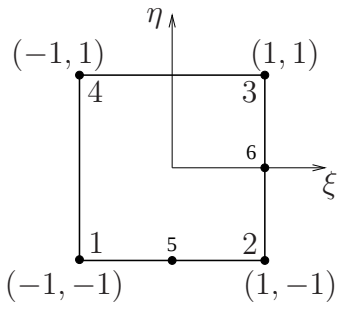
(a) Derive general formulae for $[K^e_{ij}]$ and $[F^e_i]$ for an element with domain Ω^e, in terms of the element interpolation functions N^e_i.

(b) For the element depicted in the adjacent figure, assume that only the edge with equation $x_1 = 1$ possesses non-zero Neumann boundary conditions and determine specific expressions for $[F^e_i]$, provided that the interpolation functions in Ω_e are based on (a) 4-node rectangular, (b) 8-node serendipity, or (c) 9-node Lagrangian elements. In the above analysis, let $\bar{q} = q_0$, where q_0 is a constant, and $\bar{q} = \frac{1}{2}(1-x_2)q_1 + \frac{1}{2}(1+x_2)q_2$, where q_1 and q_2 are constants.



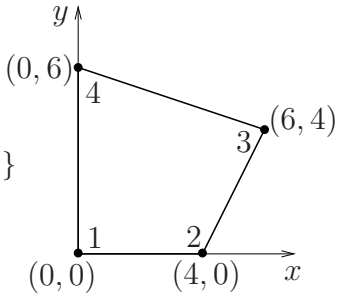
Problem 9

To illustrate a general procedure for the determination of shape functions in rectangular elements, consider the 6-node element in the adjacent figure. First, write the standard bi-linear shape functions for the 4-node element (that is, ignore, for a moment, the presence of nodes 5 and 6). Subsequently, ignoring only the presence of node 6, write the shape function for node 5 and use it to selectively modify the shape functions for nodes 1 through 4, so as to satisfy all relevant requirements for all five shape functions. Then, repeat this procedure to determine the shape functions of the full 6-node element.



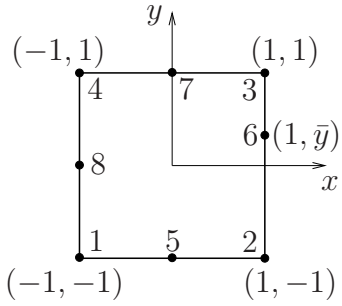
Problem 10

Determine the isoparametric transformation equations and the Jacobian matrix for the element shown in the adjacent figure. Use the standard square parent element $\Omega^e_\square = \{(\xi,\eta) \in (-1,1) \times (-1,1)\}$ for the isoparametric transformation.



Problem 11

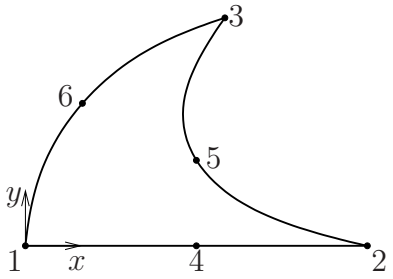
Consider the 8-node isoparametric element of the adjacent figure. Assuming that the location of nodes 1-5 and 7-8 is fixed, determine the extent to which node 6 can move vertically away from the mid-edge point without rendering the mapping from the parent to the actual element singular.



Problem 12

The 6-node triangular element in the figure is mapped isoparametrically from a straight-edge triangle. The table below contains the Cartesian coordinates for the six nodes in the physical domain:

Node	x	y
1	0.0	0.0
2	6.0	0.0
3	3.5	4.0
4	3.0	0.0
5	3.0	1.5
6	1.0	2.5



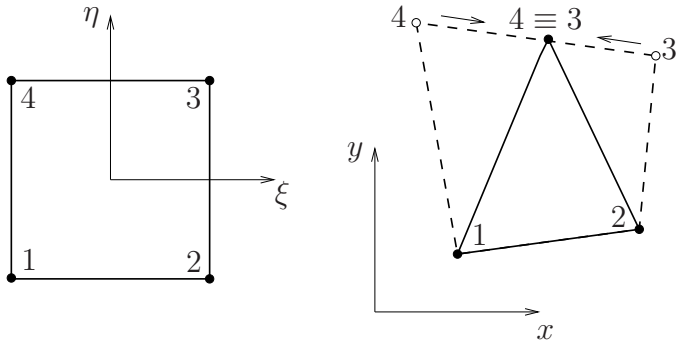
(a) Write the equation of the edge 2-5-3 in terms of the area coordinates.

(b) Determine the Cartesian coordinates (x, y) of the point in the physical domain with area coordinates $(1/4, 1/2, 1/4)$.

(c) Determine the area coordinates (L_1, L_2, L_3) of the point in the natural domain with Cartesian coordinates $(2, 2)$.

Problem 13

A 3-node triangular element is constructed from a 4-node isoparametric quadrilateral element by collapsing two neighboring nodes to a single point, as shown in the following figure:



Using a formal procedure, show that the field $u_h = \sum_{I=1}^{4} N_I^e u_I^e$ associated with the above degenerate quadrilateral element is identical to the (linear) field of a standard 3-node triangular element. What can you conclude about the isoparametric transformation at node $3 \equiv 4$?

Hint: Let $(\bar{\xi}, \bar{\eta})$ be the natural coordinates associated with an arbitrary interior point of the above degenerate triangular element. Show that at this point the gradient of $u_h(x, y)$ is independent of $(\bar{\xi}, \bar{\eta})$.

Chapter 6

Computer Implementation of Finite Element Methods

The computer implementation of finite element methods entails various practical aspects that have a well-developed state-of-the-art and merit special attention. The detailed exposition of all these implementational aspects is beyond the scope of these notes. However, some of these aspects are discussed below.

6.1 Numerical integration of element matrices

All finite element methods, with the exception of those which are derived from point-collocation, are based on weak forms that are expressed as integrals the domain Ω and its boundary $\partial\Omega$ (or parts of it), see, e.g., equations (3.14) and (3.57). These integrals are ultimately evaluated as sums over integrals at the single element level, that is, over the typical element domain Ω^e and its boundary $\partial\Omega^e$ (or parts of it). Therefore, it is important to be able to evaluate such element-wise integrals either exactly or by approximate numerical techniques. The latter case is the subject of the remainder of this section.

By way of background, consider a one-dimensional integral

$$I \;=\; \int_a^b f(x)\,dx \;, \tag{6.1}$$

where f is a real-valued function and a, b are constant integration limits. The domain (a, b) of integration can be readily mapped into the domain $(-1, 1)$ relative to a new coordinate ξ

by merely setting

$$x \;=\; \frac{1}{2}(a+b) + \frac{1}{2}(b-a)\xi \;. \tag{6.2}$$

By the inverse function theorem (see Section 5.6), this transformation is one-to-one and onto as long as $\frac{dx}{d\xi} = \frac{1}{2}(b-a) \neq 0$. The transformation also establishes symmetry of the domain relative to the origin. Taking into account the preceding transformation, one may write the original integral as

$$I \;=\; \int_a^b f(x)\,dx \;=\; \int_{-1}^{1} f\big(\frac{1}{2}(a+b) + \frac{1}{2}(b-a)\xi\big)\frac{1}{2}(b-a)\,d\xi \;=\; \int_{-1}^{1} g(\xi)\,d\xi \;. \tag{6.3}$$

Now, the integral is evaluated numerically as

$$I \;=\; \int_{-1}^{1} g(\xi)\,d\xi \;\doteq\; \sum_{k=1}^{L} w_k g(\xi_k) \;. \tag{6.4}$$

Here, ξ_k, $k = 1, 2, \ldots, L$, are the *sampling points* and w_k, $k = 1, 2, \ldots, L$, are the corresponding *weights*. Setting $g(\xi) = 1$, it follows from $(6.4)_2$ that

$$\sum_{k=1}^{L} w_k \;=\; 2 \;. \tag{6.5}$$

Equation $(6.4)_2$ encompasses virtually all the numerical integration methods used in one-dimensional finite elements.

Example 6.1.1: Some classical integrators

(a) The classical *trapezoidal rule* can be expressed, in view of equation $(6.4)_2$, as

$$\int_{-1}^{1} g(\xi)\,d\xi \;\doteq\; \sum_{k=1}^{2} w_k g(\xi_k) \;,$$

where $w_1 = w_2 = 1$, $\xi_1 = -1$ and $\xi_2 = 1$. The trapezoidal rule integrates exactly all polynomials up to degree $q = 1$.

(b) *Simpson's rule* is written as

$$\int_{-1}^{1} g(\xi)\,d\xi \;\doteq\; \sum_{k=1}^{3} w_k g(\xi_k) \;,$$

where $w_1 = w_3 = \frac{1}{3}$, $w_2 = \frac{4}{3}$, and $\xi_1 = -1$, $\xi_2 = 0$, and $\xi_3 = 1$. Simpson's rule integrates exactly all polynomials up to degree $q = 3$. Notice that Simpson's rule attains accuracy of two additional orders of magnitude as compared to the trapezoidal rule despite only adding one extra function evaluation. This is due to the optimal placement $\xi_2 = 0$ of the interior sampling point. ◀

The trapezoidal and Simpson's rules are special cases of the *Newton-Cotes closed* numerical integration formulae. Given that function evaluations are computationally expensive and need to be repeated for all element-based integrals, one may reasonably ask if the Newton-Cotes formulae are optimal, that is, whether they furnish the maximum possible polynomial degree of accuracy for the given cost. It turns out that the answer to this question is, in fact, negative. This point can be argued as follows: recalling equation $(6.4)_2$, it is clear that the right-hand side contains L weights and L coordinates of the sampling points to be determined, hence a total of $2L$ "tunable" parameters. Suppose that one wishes to exactly integrate with such a formula a polynomial of degree q, expressed as

$$P(\xi) \;=\; a_0 + a_1\xi + \ldots + a_q\xi^q \tag{6.6}$$

over the canonical domain $(-1, 1)$. This implies that

$$\int_{-1}^{1} P(\xi)\, d\xi \;=\; \sum_{k=1}^{L} w_k g(\xi_k) \;, \tag{6.7}$$

or

$$\int_{-1}^{1} (a_0 + a_1\xi + \ldots + a_q\xi^q)\, d\xi \;=\; \sum_{k=1}^{L} w_k (a_0 + a_1\xi_k + \ldots + a_q\xi_k^q) \;. \tag{6.8}$$

Since the constant coefficients a_k, $k = 0, 1, \ldots, q$, are independent of each other and arbitrary, it follows from the above equation that

$$\left[\frac{\xi^{i+1}}{i+1}\right]_{-1}^{1} \;=\; \sum_{k=1}^{L} w_k \xi_k^i \;=\; \begin{cases} \dfrac{2}{i+1} \;, & i \text{ even} \\ 0 \;, & i \text{ odd} \end{cases} \quad, \quad i = 0, 1, \ldots, q \;. \tag{6.9}$$

The $q+1$ equations in $(6.9)_2$ contain $2L$ unknowns, which means that a unique solution can be expected if, and only if, $q = 2L - 1$. It turns out that the system $(6.9)_2$ possesses such a unique solution, which yields the *Gaussian quadrature* rules. These are optimal in terms of accuracy and, for this reason, they are used extensively in finite element computations.

Example 6.1.2: Three cases of Gaussian quadrature in one-dimension

(a) When $L = 1$, it is clear that, owing to symmetry, $\xi_1 = 0$ and $w_1 = 2$ (alternatively, the values of ξ_1 and w_1 may be derived from $(6.9)_2$ for $i = 0$ and $i = 1$). This is the well-known *mid-point rule*, which is exact for integration of polynomials of degree up to $q = 1$.

(b) When $L = 2$, symmetry dictates that $\xi_1 = -\xi_2$ and $w_1 = w_2 = 1$. The value of ξ_1 can be deduced from (6.9) taking into account that this rule should be exact for the integration of polynomials of degree up to $q = 3$. Indeed, taking $i = 2$ in $(6.9)_2$ leads to $-\xi_1 = \xi_2 = \dfrac{1}{\sqrt{3}}$.

(c) When $L = 3$, symmetry necessitates that $w_1 = w_3$ and, also, $\xi_1 = -\xi_3$, and $\xi_2 = 0$. Appealing again to equation $(6.9)_2$, now for $i = 2$ and $i = 4$, one finds that $w_1 = w_3 = \frac{5}{9}$, $w_2 = \frac{8}{9}$, and $-\xi_1 = \xi_3 = \sqrt{\frac{3}{5}}$. In this case, polynomial integration is exact up to degree $q = 5$.

Similar results may be obtained for higher-order accurate Gaussian quadrature formulae. ◀

Gaussian quadrature is a prime example of the *Newton-Cotes open* numerical integration formulae.

The preceding formulae are readily applicable to integration over multi-dimensional domains which are products of one-dimensional domains. These include rectangles in two dimensions and orthogonal parallelepipeds in three dimensions. This observation is particularly relevant to general isoparametric quadrilateral and hexahedral elements which are mapped to the physical domain from squares and cubes. Taking, for example, the case of an isoparametric quadrilateral, write a typical domain integral as

$$I \;=\; \int_{\Omega^e} f(x,y)\,dxdy \;=\; \int_{-1}^{1}\int_{-1}^{1} f\big(\hat{x}(\xi,\eta),\hat{y}(\xi,\eta)\big)\,J\,d\xi d\eta \;=\; \int_{-1}^{1}\int_{-1}^{1} g(\xi,\eta)\,d\xi d\eta\;, \quad (6.10)$$

where use is made of $(5.84)_{1,2}$ and (5.102). Since the coordinates ξ and η are independent, one may write

$$\begin{aligned}\int_{-1}^{1}\int_{-1}^{1} g(\xi,\eta)\,d\xi d\eta \;\doteq\; \int_{-1}^{1}\Big\{\sum_{k=1}^{L} w_k g(\xi_k,\eta)\Big\}\,d\eta \;\doteq\; \sum_{l=1}^{L} w_l \sum_{k=1}^{L} w_k g(\xi_k,\eta_l)\\ = \sum_{k=1}^{L}\sum_{l=1}^{L} w_k w_l g(\xi_k,\eta_l)\;. \quad (6.11)\end{aligned}$$

Generally, multi-dimensional integrals over product domains can be evaluated numerically using multiple summations (one per dimension). Figure 6.1 illustrates certain two-dimensional integration rules over square domains and the degree of polynomials of the form $\xi^{q_1}\eta^{q_2}$ that are integrated exactly by each of the rules.

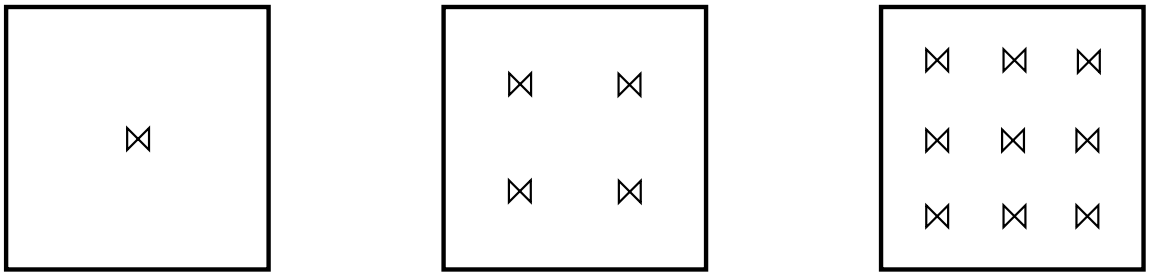


Figure 6.1. *Two-dimensional Gauss quadrature rules for* $q_1, q_2 \le 1$ *(left),* $q_1, q_2 \le 3$ *(center), and* $q_1, q_2 \le 5$ *(right)*

Remark:

☛ It can be shown that the locations of the Gauss points in the domain $(-1, 1)$ are roots of the *Legendre polynomials* P_k. These are defined by the recurrence formula

$$P_{k+1}(\xi) \;=\; \frac{(2k+1)\xi P_k(\xi) - kP_{k-1}(\xi)}{k+1} \quad , \quad k = 1, 2, \ldots \;,$$

with $P_0(\xi) = 1$ and $P_1(\xi) = \xi$. The Legendre polynomials satisfy the orthogonality property

$$\int_{-1}^{1} P_i(\xi) P_j(\xi)\, d\xi \;=\; \begin{cases} \dfrac{2}{i+1} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} .$$

This property plays an essential role in establishing the aforementioned connection between Legendre polynomials and Gauss points.

Integration over triangular and tetrahedral domains can be performed either by using the exact formulae presented in Chapter 5 for polynomial functions of the area or volume coordinates or by approximate formulae of the form

$$I \;=\; \int_{\Omega} g(L_1, L_2, L_3)\, dA \;\doteq\; A \sum_{k=1}^{L} w_k g(L_{1k}, L_{2k}, L_{3k}) \tag{6.12}$$

for a straight-edge triangular region Ω of area A, or

$$I \;=\; \int_{\Omega} g(L_1, L_2, L_3, L_4)\, dV \;\doteq\; V \sum_{k=1}^{L} w_k g(L_{1k}, L_{2k}, L_{3k}, L_{4k}) \tag{6.13}$$

for a straight-edge, flat face tetrahedral region Ω of volume V. Figure 6.2 depicts three simple integration rules for triangles.

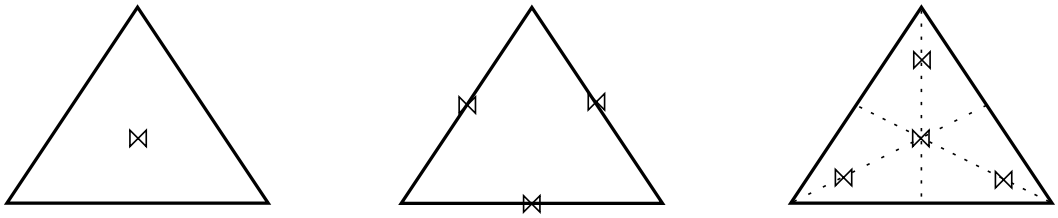


Figure 6.2. *Integration rules in triangular domains for $q \leq 1$ (left), $q \leq 2$ (center), and $q \leq 3$ (right). At left, the integration point is located at the barycenter of the triangle and the weight is $w_1 = 1$; at center, the integration points are located at the mid-edges and the weights are $w_1 = w_2 = w_3 = 1/3$; at right, one integration point is located at the barycenter and has weight $w_1 = -27/48$, while the other three are at points with coordinates $(0.6, 0.2, 0.2)$, $(0.2, 0.6, 0.2)$, and $(0.2, 0.2, 0.6)$, with associated weights $w_2 = w_3 = w_4 = 25/48$.*

6.2 Assembly of global element arrays

The purpose of this section is to establish a procedure by which one may start with weak forms at the element level, assemble all the element-wide information and derive global equations, whose solution yields the finite element approximation of interest.

Weak forms emanating from Galerkin, least squares, collocation or variational approaches can be written without any restrictions in any subdomain of the original domain Ω over which a differential equation is assumed to hold. Indeed, if a differential equation holds over a domain Ω, then it also holds over any subset of Ω. Further, assuming that the finite element approximation and weighting functions are smooth, the use of integration by parts and the divergence theorem is allowable. Therefore, weak forms such as (3.14) can be written over the domain Ω^e of a given element, that is,

$$\int_{\Omega^e} \left[\frac{\partial w_h}{\partial x_1} k \frac{\partial u_h}{\partial x_1} + \frac{\partial w_h}{\partial x_2} k \frac{\partial u_h}{\partial x_2} + w_h f \right] d\Omega + \int_{\partial\Omega^e \cap \Gamma_q} w_h \bar{q}\, d\Gamma + \int_{\partial\Omega^e \setminus \partial\Omega} w_h q\, d\Gamma = 0 \,. \tag{6.14}$$

The two boundary integral terms in (6.14) constitute a small, yet important, departure from those in (3.14). The first boundary term applies to the part of the element boundary $\partial\Omega^e$, if any, that happens to lie on the exterior Neumann boundary Γ_q of the domain Ω. The second boundary term in (6.14) refers to the interior part of the element boundary (that is, the portion of the element boundary $\partial\Omega^e$ that is shared with other elements), which is subject to a (yet unknown) Neumann boundary condition specifying the flux q across two

neighboring elements.

Starting from equation (6.14), one may write corresponding element-wide weak forms for all finite elements and add them together. This leads to

$$\sum_e \int_{\Omega^e} \left[\frac{\partial w_h}{\partial x_1} k \frac{\partial u_h}{\partial x_1} + \frac{\partial w_h}{\partial x_2} k \frac{\partial u_h}{\partial x_2} + w_h f \right] d\Omega + \sum_e \int_{\partial\Omega^e \cap \Gamma_q} w_h \bar{q} \, d\Gamma + \sum_{e,e' \text{neighbors}} \int_{C_{e,e'}} w_h [\![q]\!] \, d\Gamma = 0 , \quad (6.15)$$

where $C_{e,e'}$ denotes an edge shared between two contiguous elements e and e' and $[\![q]\!]$ denotes the jump of the (yet unknown) normal flux q from element e to element e' across $C_{e,e'}$. Equation (6.15) may be readily rewritten as

$$\int_{\Omega} \left[\frac{\partial w_h}{\partial x_1} k \frac{\partial u_h}{\partial x_1} + \frac{\partial w_h}{\partial x_2} k \frac{\partial u_h}{\partial x_2} + w_h f \right] d\Omega + \int_{\Gamma_q} w_h \bar{q} \, d\Gamma + \sum_{e,e' \text{neighbors}} \int_{C_{e,e'}} w_h [\![q]\!] \, d\Gamma = 0 , \quad (6.16)$$

if one assumes that $\sum_e \int_{\Omega^e} d\Omega = \Omega$ and $\sum_e \int_{\partial\Omega^e \cap \Gamma_q} d\Gamma = \Gamma_q$. Note that, in general, the preceding two conditions are satisfied only in an approximate sense, due to the error in domain and boundary discretization. Either way, it is readily apparent that the weak form in (6.16) differs from the original form in (3.14) because of the introduction of the finite element fields (w_h, u_h) and the jump conditions in the last term of the right-hand side.

The finite element *assembly operation* entails the summation of all element-wide discrete weak forms to form the global discrete weak form from which one may derive a system of algebraic equations, whose solution provides the scalar coefficients that define u_h, see, e.g., equation $(3.24)_1$. Hence, for each element e with n degrees of freedom, one may start from (6.14) and derive an equation of the form

$$\sum_{i=1}^{n} \beta_i \Big(\sum_{j=1}^{n} K_{ij}^e \, u_j^e - F_i^e - F_i^{\text{int},e} \Big) = 0 , \quad (6.17)$$

where K_{ij}^e are the components of the element stiffness matrix and F_i^e are the components of the forcing vector contributed in the domain Ω^e and on the boundary $\partial\Omega^e \cap \Gamma_q$. Lastly, $F_i^{\text{int},e}$ are the components of the forcing vector due to the boundary term on $\partial\Omega^e \setminus \partial\Omega$. This term is unknown at the outset, as the element interior boundary fluxes are not specified in

the original boundary-value problem.[1] Recalling the arbitrariness of the weighting function coefficients β_i, it follows immediately that for each element

$$\sum_{j=1}^{n} K_{ij}^e \, u_j^e - F_i^e - F_i^{\text{int},e} \;=\; 0 \quad , \quad i = 1, 2, \ldots, n \; . \tag{6.18}$$

The assembly operation now amounts to combining all element-wide equations of the form (6.18) into a global system that applies to the whole domain Ω. Symbolically, this may be represented by way of an *assembly operator* $\mathop{\mathbf{A}}_{e}$, such that

$$\mathop{\mathbf{A}}_{e} \left[\sum_{j=1}^{n} K_{ij}^e \, u_j^e \;-\; F_i^e - F_i^{\text{int},e} \right] \;=\; 0 \; . \tag{6.19}$$

This operator aggregates the contributions to the N global degrees of freedom from all elements, thus producing the global set of linear algebraic equations

$$\sum_{J=1}^{N} K_{IJ} u_J \;=\; F_I \quad , \quad I = 1, 2, \ldots, N \; . \tag{6.20}$$

Now, the components K_{IJ} of the global stiffness matrix and the components F_I of the global forcing vector are expressed as

$$[K_{IJ}] \;=\; \mathop{\mathbf{A}}_{e} [K_{ij}^e] \quad , \quad [F_I] \;=\; \mathop{\mathbf{A}}_{e} [F_i^e] \; . \tag{6.21}$$

Note that the assembled contributions of the interior boundary fluxes are neglected, namely

$$[F_I^{\text{int}}] \;=\; \mathop{\mathbf{A}}_{e} [F_i^{\text{int},e}] \;\doteq\; 0 \; . \tag{6.22}$$

This assumption is tantamount to outright neglecting the last boundary integral term in (6.16), and is made in finite element methods by necessity, although it clearly induces an error. Indeed, the interior boundary fluxes are unknown at the outset, so that including the corresponding forces in the assembled state equations is not an option. On the other hand, omission of these interelement jump terms is justified by the fact that the exact solution of the differential equation guarantees flux continuity across any surface, hence in an asymptotic sense (that is, as the approximation becomes more accurate), the force contributions of these jumps tend to vanish.

[1] This is precisely why one cannot, in general, solve the original boundary-value problem on a direct element-by-element basis.

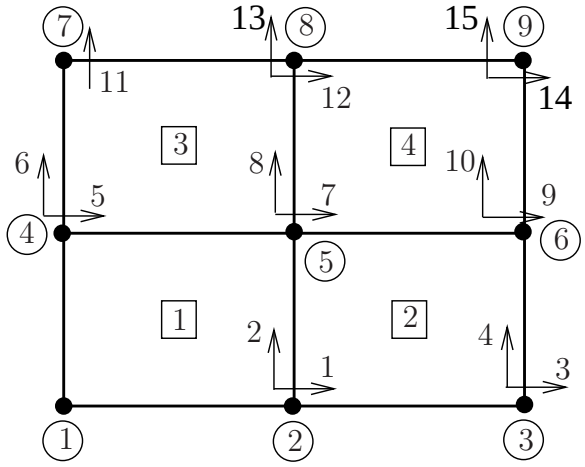


Figure 6.3. *Finite element mesh depicting global node and element numbering, as well as global degree of freedom assignments (both degrees of freedom are fixed at node 1 and the second degree of freedom is fixed at node 7)*

It is instructive here to illustrate the action of the assembly operator $\underset{e}{\mathbf{A}}$ by means of an example. To this end, consider a simple finite element mesh of 4-node rectangular elements with two-degrees of freedom per node, see Figure 6.3. All *active* degrees of freedom (that is, those that are not fixed) are numbered in increasing order of the global node numbers. In this manner, one forms the ID array, defined as

$$[\mathrm{ID}] \;=\; \begin{bmatrix} 0 & 1 & 3 & 5 & 7 & 9 & 0 & 12 & 14 \\ 0 & 2 & 4 & 6 & 8 & 10 & 11 & 13 & 15 \end{bmatrix} , \tag{6.23}$$

by looping over all nodes. The dimension of this array is `ndf`×`numnp`, where `ndf` denotes the number of degrees of freedom per node before any boundary conditions are imposed and `numnp` denotes the total number of nodes in the mesh (here, `ndf`=2 and `numnp`=9). Taking into account the local nodal numbering convention for 4-node elements (see Figure 5.40), one may now generate the IX array as

$$[\mathrm{IX}] \;=\; \begin{bmatrix} 1 & 2 & 4 & 5 \\ 2 & 3 & 5 & 6 \\ 5 & 6 & 8 & 9 \\ 4 & 5 & 7 & 8 \end{bmatrix} \tag{6.24}$$

This array contains the mapping between the local and the global numbers for all nodes in each element and its dimension is `nen`×`numel`, where `nen` is the number of nodes per element

and `numel` is the total number of elements in the mesh (here, `nen`=4 and `numel`=4). Using the ID and IX arrays, it is now straightforward to deduce the LM array as

$$[\mathrm{LM}] \;=\; \begin{bmatrix} 0 & 1 & 5 & 7 \\ 0 & 2 & 6 & 8 \\ 1 & 3 & 7 & 9 \\ 2 & 4 & 8 & 10 \\ 7 & 9 & 12 & 14 \\ 8 & 10 & 13 & 15 \\ 5 & 7 & 0 & 12 \\ 6 & 8 & 11 & 13 \end{bmatrix} \tag{6.25}$$

by looping over all elements. The dimension of this array is (`ndf*nen`)×`numel`. Each column of the LM array contains the list of globally numbered degrees of freedom in the corresponding order to that of the local degrees of freedom of the element. As a result, the task of assembling an element-wide array, say $[K^4_{ij}]$ into the global stiffness array is reduced to identifying the correspondence between local and global degrees of freedom for each component of $[K^4_{ij}]$ by direct reference to the LM array. For instance, the entry K^4_{12} is added to the global stiffness (of dimension 15×15) in the 7th row/8th column entry, as dictated by the first two entries of the 4th column of the LM array in (6.25). Likewise, the entry F^3_5 is added to the global forcing vector (of dimension 15×1) in the 12th row, as dictated by the 5th row of the 3rd column of the LM array. The preceding matrix-based data structure, due to E.L. Wilson,[2] shows that the task of assembling global arrays amounts to a simple reindexing of the local arrays with the aid of the LM array.

A slight generalization of the preceding assembly procedure, which allows for the efficient processing on non-homogeneous Dirichlet boundary conditions, involves the numbering all degrees of freedom (including those which are restrained by Dirichlet boundary conditions) and generating an expanded stiffness matrix (in the preceding example, this would be of dimension 18×18). Then, any rows and columns associated with Dirichlet boundary conditions would be reduced and any corresponding non-homogeneous prescribed boundary values would give rise to force terms appropriately added to the global forcing vector. To explain this reduction process, express the expanded linear algebraic system $[K][\alpha] = [F]$ resulting

[2] Edward L. Wilson (1931–) is an American civil engineer.

from the finite element approximation of a given problem in the form

$$\underbrace{\begin{bmatrix} [K_{uu}] & [K_{u\bar{u}}] \\ [K_{\bar{u}u}] & [K_{\bar{u}\bar{u}}] \end{bmatrix}}_{[K]} \underbrace{\begin{bmatrix} [u] \\ [\bar{u}] \end{bmatrix}}_{[\alpha]} = \underbrace{\begin{bmatrix} [F_u] \\ [F_{\bar{u}}] \end{bmatrix}}_{[F]} , \tag{6.26}$$

where the column-vectors $[u]$ and $[\bar{u}]$ represent free and fixed global degrees of freedom, and the global stiffness matrix $[K]$ and forcing vector $[F]$ are partitioned accordingly. The reduced system of finite element equations then takes the form

$$[K_{uu}][u] \;=\; [F_u] - [K_{u\bar{u}}][\bar{u}] \;, \tag{6.27}$$

which may be solved for $[u]$, provided that $[K_{uu}]$ is invertible. Subsequently, the "reactions" $[F_{\bar{u}}]$ (that is, the forces required to impose the prescribed values $[\bar{u}]$ for the fixed degrees of freedom) may be calculated from the second set of matrix equations in (6.26), with the aid of (6.27), as

$$\begin{aligned} [F_{\bar{u}}] \;&=\; [K_{\bar{u}u}][u] + [K_{\bar{u}\bar{u}}][\bar{u}] \\ &=\; [K_{\bar{u}u}] \left([K_{uu}^{-1}][F_u] - [K_{uu}^{-1}][K_{u\bar{u}}][\bar{u}] \right) + [K_{\bar{u}\bar{u}}][\bar{u}] \\ &=\; \left([K_{\bar{u}\bar{u}}] - [K_{\bar{u}u}][K_{uu}^{-1}][K_{u\bar{u}}] \right) [\bar{u}] + [K_{\bar{u}u}][K_{uu}^{-1}][F_u] \;. \end{aligned} \tag{6.28}$$

The matrix inside the parenthesis on the right-hand side of $(6.28)_3$ is known in linear algebra as the *Schur complement* of the submatrix $[K_{uu}]$ relative to the matrix $[K]$.

6.3 Algebraic equation solving in finite element methods

The system of linear algebraic equations (6.20) obtained upon assembling the local arrays into their global counterparts can be solved using a number of different and well-established methods. These include

(1) Iterative methods, such as Jacobi, Gauss-Seidel, steepest descent, conjugate gradient, and multigrid.

(2) Direct methods, such as Gauss elimination (or LU decomposition), Choleski, frontal.

Generally, direct methods are used for relatively small systems (say, for $N \leq 10^6$), owing to their relatively high expense. Iterative methods are typically more efficient for larger systems.

One of the most important properties of finite element methods is that they lead to stiffness matrices that are *banded*, which means that the non-zero terms are situated around the major diagonal. This is an immediate implication of the local nature of the finite element approximation. Indeed, the fact that the interpolation functions have small support (see the discussion in Section 5.1) implies that coupling between degrees of freedom is restricted only to neighboring elements. Figure 6.4 illustrates a typical example of the profile of a finite element stiffness matrix:

$$[K] = \begin{bmatrix} * & * & & & & & & \\ * & * & * & * & * & & & \\ & * & * & * & * & & & \\ & * & * & * & * & & & \\ & * & * & * & * & * & & \\ & & & & * & * & * \\ & & & & & * & * \end{bmatrix}$$

Figure 6.4. *Profile of a typical finite element stiffness matrix ($*$ denotes a non-zero entry or a zero entry having at least one non-zero entry below and above it in the column to which it belongs)*

It is important to note that the profile of the finite element stiffness matrix is generally symmetric, even if the entries themselves are not. This is because the coupling between two degrees of freedom is generally bilateral, namely one degree of freedom affects the other and vice-versa. In the model stiffness matrix of Figure 6.4, one may define the *half-bandwidth* b as the maximum height of a column above the major diagonal for which the highest entry is non-zero. By this definition, for the system in Figure 6.4, it is clear that $b = 3$.

The significance of the bandedness of finite element stiffnesses becomes apparent when one considers the cost of solving a system of N linear algebraic equations. In particular, the solution of N such equations using standard Gauss elimination without accounting for bandedness requires approximately $\frac{2}{3}N^3$ flops,[3] when N is large[4]. If the stiffness matrix

[3] A *flop* is a floating-point operation, namely an addition, subtraction, multiplication or division.

[4] The precise number of flops for Gauss elimination for a linear system of N equations with N unknowns is $\frac{N^3}{3} + \frac{N^2}{2} - \frac{5N}{6}$ additions/subtractions and $\frac{N^3}{3} + N^2 - \frac{N}{3}$ multiplications/divisions.

is banded, with half-bandwidth $b \ll N$, the cost is reduced to approximately $2Nb^2$. To appreciate the difference, consider a moderate-size system of $N = 10^5$ equations with half-bandwidth $b = 10^3$. Standard Gauss elimination on a single processor that delivers 500 MFLOPS[5] takes approximately

$$\frac{\frac{2}{3}(10^5)^3}{500 \times 10^6} \;=\; \frac{2}{15}10^7 \text{ sec} \;\doteq\; 15 \text{ days} \,. \tag{6.29}$$

On the other hand, using a banded Gauss elimination solver requires

$$\frac{2(10^5)(10^3)^2}{500 \times 10^6} \;=\; \frac{2}{5}10^3 \text{ sec} \;\doteq\; 7 \text{ minutes} \,. \tag{6.30}$$

In addition to the substantial savings in solution time, the banded structure of the stiffness matrix allows for compact storage of its components, which reduces the associated memory requirements. For instance, taking again by way of example the matrix $[K]$ in Figure 6.4, it is possible to store all of its non-zero components in a 25×1 array, say $[A]$ and recreate its exact form by defining a 7×1 array $[B]$ as

$$[B] \;=\; \begin{bmatrix} 1 \\ 4 \\ 9 \\ 14 \\ 19 \\ 22 \\ 25 \end{bmatrix} \,, \tag{6.31}$$

provided that the profile of $[K]$ is symmetric with respect to the major diagonal. Note that the entries of $[B]$ are equal to the index in $[A]$ of each successive diagonal entry of $[K]$. Compact storage becomes even more efficient if the matrix $[K]$ is symmetric, in which case only the diagonal and upper (or, equivalently, lower) triangular entries need to be stored.

6.4 Finite element modeling: mesh design and generation

Finite element modeling is a relatively complex undertaking. It requires:

[5] MFLOPS stands for Millions of Floating-point Operations per Second.

- Complete and unambiguous understanding of the boundary/initial-value problem (question: what are the relevant differential equations and boundary and/or initial conditions?)

- Familiarity with the nature of the solutions to this class of problems (question: is the finite element solution consistent with physically-motivated expectations of it?).

- Experience in geometric modeling (question: how does one create a finite element mesh that accurately represents the domain of interest?)

- Deep knowledge of the technical aspects of the finite element method (questions: how does one impose Dirichlet boundary conditions, input equivalent nodal forces, choose element types and number of element integration points, etc.?)

Creating sophisticated finite element models typically involves two well-tested steps. These are:

- Simplification of the model to the highest possible degree without loss of any of its salient features.

- Decomposition of the reduced model into simpler submodels, meshing of the submodels, and "tying" of these back into the full model.

Two aspects of mesh modeling that merit special attention are symmetry and optimal node numbering.

6.4.1 Symmetry

If the differential equation, boundary conditions, and domain are all symmetric with respect to certain axes or planes, the finite element analyst can exploit the resulting symmetry(ies) of the solution to simplify the task of modeling. The most important step here is to apply the appropriate boundary conditions on the symmetry axis or plane.

Some typical examples of symmetry are illustrated in Figure 6.5 below.

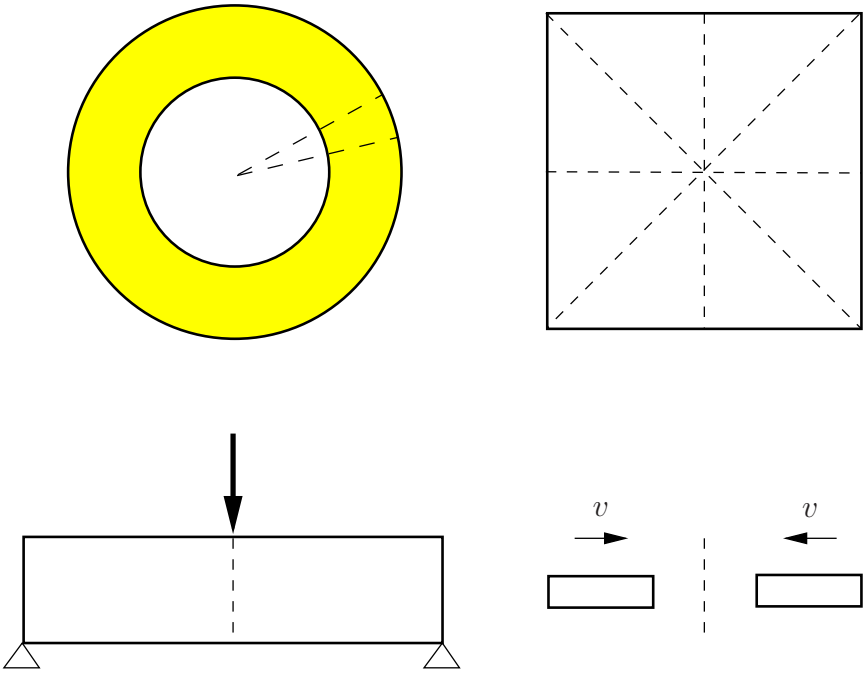


Figure 6.5. *Representative examples of symmetries in the domains of differential equations (corresponding symmetries in the boundary conditions, loading, and equations themselves are assumed)*

Judicious use of symmetry may drastically reduce the cost of a finite element analysis without sacrificing the reliability of the solution.

6.4.2 Optimal node numbering

The manner in which finite element nodes are globally numbered may play an important role in the shape and size of the profile of the resulting finite element stiffness matrix. This point is illustrated by means of a simple example in Figure 6.6, where the nodes of the same mesh are numbered in two distinct and regular ways, such as row-wise or column-wise. Assuming that each node has two active degrees of freedom, row-wise numbering leads to a half-bandwidth $b_A = 17$, while column-wise numbering leads to $b_B = 7$.

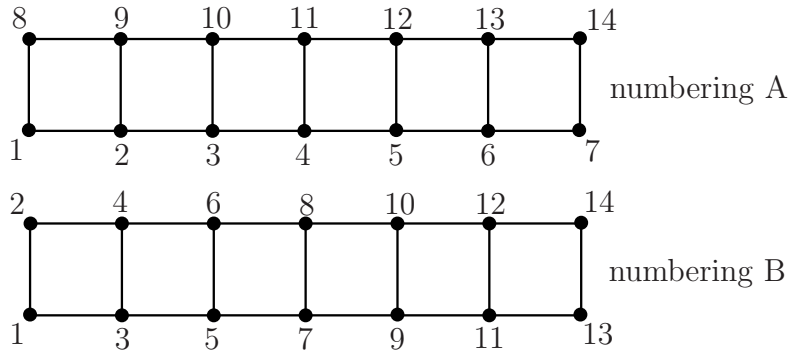


Figure 6.6. *Two possible ways of node numbering in a finite element mesh*

Generally, global node numbering should be done in a manner that minimizes the half-bandwidth of the stiffness matrix. This becomes quite challenging when creating parts of a mesh separately and then tying all of them together. Several algorithms have been devised to perform optimal (or, more often, nearly optimal) node numbering. Most commercial element codes have a built-in node numbering algorithm and the user does not have to occupy him/herself with this task.

6.5 Computer program organization

All commercial and (many) stand-along research/education finite element codes contain three basic modules: input (pre-processing), solution and output (post-processing).

The input module concerns primarily the generation of the finite element mesh and the application of boundary conditions. In this module, one may also specify the physics of the problem together with the values of any required constants, as well as the element type and other related parameters. The solution module concerns the determination of the element arrays, the assembly of the global arrays, and the solution of the resulting algebraic systems (linear or non-linear). The output module handles the computation of any quantities of interest at the mesh or individual element level and the visualization of the solution.

Some of the desirable features of finite element codes are:

- General-purpose, namely employing a wide range of finite element methods to solve diverse problems (e.g., time-independent/dependent, linear/non-linear, multi-physics, etc.)

- Full non-linearity, namely designed at the outset to treat all problems as non-linear and handling linear problems as a trivial special case.

- Modularity, namely able to incorporate new elements written by (advanced) users and finite element programmers without requiring that they know (or have access) to all parts of the program.

In recent years, there is a trend toward integration of computer-aided design software tools into finite element codes to provide "one-stop shopping" for engineering analysis/design needs. In concert with this trend, there is an effort to limit the discretion of the user in intervening in the code through the input files, thereby protecting the integrity of the analysis, albeit at the expense of sometimes frustrating the experienced users.

6.6 Suggestions for further reading

Section 6.1

[1] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions; With Formulas, Graphs, and Mathematical Tables.* Dover, New York, 1970. [*This handbook includes a set of integration rules for special polygonal domains*].

Section 6.3

[1] M. Hoit and E.L. Wilson. An equation numbering algorithm based on minimum front criteria. *Comp. Struct.*, 16:225–239, 1983.

[2] S.W. Sloan and M.F. Randolph. Automatic element reordering for finite element analysis with frontal solution schemes. *Int. J. Num. Meth. Engr.*, 19:1153–1181, 1983.

[*These articles describe two alternative algorithms for optimal node number in finite element meshes*].

6.7 Exercises

Problem 1

A function $f(x, y) : \mathbb{R}^2 \mapsto \mathbb{R}$ varies linearly within a 3-node triangular element with domain Ω_e

(that is, $f(x,y) = a + bx + cy$ in the element). Show that the average value of this function over the element domain, defined as

$$f_{av} \;=\; \frac{1}{A}\int_{\Omega_e} f(x,y)\, d\Omega \;,$$

where A is the area of the element, is equal to the average of the three nodal values of f.

Hint: evaluate the above integral using area coordinates.

Problem 2

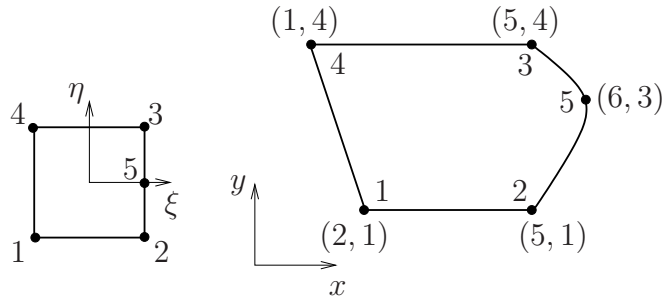
A 5-node quadrilateral element is constructed by mapping isoparametrically the square domain $\Omega^e_\square$ of the natural space (ξ,η) into the partially curved domain Ω_e of the physical space (x,y), as in the figure below.

(a) Determine the shape functions of the element in the natural domain.

(b) Show that the isoparametric mapping is one-to-one for all points (ξ,η) of domain $\Omega^\square_e$.

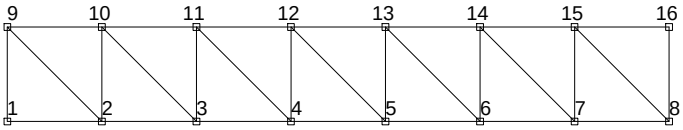
(c) Determine the minimum number of Gauss points per direction required to compute the area of the element exactly in the physical domain.

(d) Evaluate the exact area of the element in the physical domain using Gaussian quadrature.

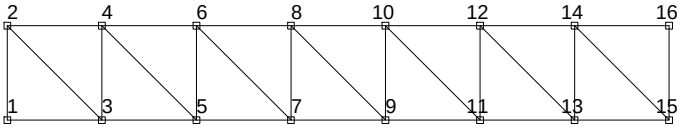


Problem 3

The two finite element meshes shown below are used in the solution of a planar problem for which every node has two active degrees of freedom numbered in increasing order with the global node numbers. For each mesh, indicate which entries of the stiffness matrix are generally non-zero by blacking them out in a 32×32 grid (you may use the grid on engineering paper). What is the half-bandwidth of each mesh for the given problem?



Mesh A



Mesh B

Chapter 7

Elliptic Differential Equations

The finite element method was originally conceived for elliptic partial differential equations. For such equations, Bubnov-Galerkin based finite element formulations can be shown to possess highly desirable properties of convergence, as will be established later in this chapter.

7.1 The Laplace equation in two dimensions

The Laplace equation is a classic example of an elliptic partial differential equation, see the discussion in Section 1.3. Galerkin-based and variational weak forms for the Laplace equation have been derived and discussed in detail earlier, see Sections 3.2 and 4.1.

7.2 Linear elastostatics

Consider a deformable solid body that occupies the region $\Omega \subset \mathbb{R}^3$ in its reference state at time $t = 0$, see Figure 7.1. Also, let the boundary $\partial\Omega$ of the region Ω be smooth with outward unit normal $\mathbf{n}$, and be decomposed into two regions $\Gamma_u \neq \emptyset$ and Γ_q, such that $\partial\Omega = \overline{\Gamma_u \cup \Gamma_q}$. Further, assume that there is a vector function $\mathbf{u} : \overline{\Omega} \times \mathbb{R}^+ \mapsto \mathbb{R}^3$, such that the position vector $\mathbf{y}$ of a material point X at time t is related to the position vector $\mathbf{x}$ of the same material point at time $t = 0$ by

$$\mathbf{y}(\mathbf{x}, t) \;=\; \mathbf{x} + \mathbf{u}(\mathbf{x}, t) \; . \tag{7.1}$$

The vector function $\mathbf{u}$ is referred to as the *displacement* field. The body is assumed to be made of a linearly elastic material and is being deformed due to body force $\mathbf{f}$ per unit

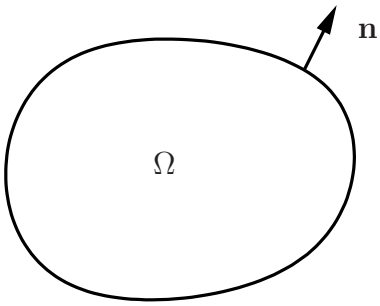


Figure 7.1. *The domain Ω of the linear elastostatics problem*

volume, prescribed surface tractions $\bar{\mathbf{t}}$ on Γ_q, and prescribed displacements $\bar{\mathbf{u}}$ on Γ_u. All data functions $\mathbf{f}$, $\bar{\mathbf{t}}$ and $\mathbf{u}$ are assumed continuous in their respective domains.

The strong form of the equations of linear elastostatics are written as

$$\begin{aligned} \nabla \cdot \boldsymbol{\sigma} + \mathbf{f} &= \mathbf{0} \qquad \text{in } \Omega \ , \\ \boldsymbol{\sigma}\mathbf{n} &= \bar{\mathbf{t}} \qquad \text{on } \Gamma_q \ , \\ \mathbf{u} &= \bar{\mathbf{u}} \qquad \text{on } \Gamma_u \ , \end{aligned} \tag{7.2}$$

where $\boldsymbol{\sigma}$ is the stress tensor and $\nabla \cdot \boldsymbol{\sigma}$ denotes the divergence of $\boldsymbol{\sigma}$. Here, $(7.2)_1$ are the equations of equilibrium for the body, while $(7.2)_{2,3}$ are the Neumann and Dirichlet boundary conditions,[1] respectively. If the stress tensor has Cartesian component representation

$$[\boldsymbol{\sigma}] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} \ , \tag{7.3}$$

then its divergence can be expressed as

$$[\nabla \cdot \boldsymbol{\sigma}] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} \begin{bmatrix} \dfrac{\partial}{\partial x_1} \\ \dfrac{\partial}{\partial x_2} \\ \dfrac{\partial}{\partial x_3} \end{bmatrix} = \begin{bmatrix} \sigma_{11,1} + \sigma_{12,2} + \sigma_{13,3} \\ \sigma_{21,1} + \sigma_{22,2} + \sigma_{23,3} \\ \sigma_{31,1} + \sigma_{32,2} + \sigma_{33,3} \end{bmatrix} \ , \tag{7.4}$$

where x_j, $j = 1, 2, 3$, are the Cartesian components of $\mathbf{x}$, and $\sigma_{ij,k} = \dfrac{\partial \sigma_{ij}}{\partial x_k}$.

[1] In general, different boundary condition may apply to the same boundary point at different directions.

Assuming that the elastic material is isotropic and homogeneous, one may express the stress tensor as

$$\boldsymbol{\sigma} \;=\; \lambda(\mathrm{tr}\,\boldsymbol{\epsilon})\mathbf{I} + 2\mu\boldsymbol{\epsilon}\;, \tag{7.5}$$

in terms of the *Lamé constants* λ and μ, the identity tensor $\mathbf{I}$, and the infinitesimal strain tensor $\boldsymbol{\epsilon}$. The Lamé constants are taken to satisfy the conditions

$$\lambda + \frac{2}{3}\mu \;>\; 0 \qquad , \qquad \mu \;>\; 0\;, \tag{7.6}$$

which will be justified later in this section. An alternative set of constants, the *Young's modulus* E and *Poisson's ratio* ν may be used in (7.5), where $E = \frac{\mu(3\lambda+2\mu)}{\lambda+\mu}$ and $\nu = \frac{\lambda}{2(\lambda+\mu)}$, which, in view of (7.6) implies that $E > 0$ and $-1 < \nu < 1$.

The infinitesimal strain tensor which appears in (7.5) is defined as

$$\boldsymbol{\epsilon} \;=\; \frac{1}{2}[\nabla\mathbf{u} + (\nabla\mathbf{u})^T] \;=\; \nabla_s\mathbf{u}\;, \tag{7.7}$$

where $\nabla\mathbf{u}$ is the gradient of $\mathbf{u}$ expressed in Cartesian component form as

$$[\nabla\mathbf{u}] \;=\; \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \begin{bmatrix} \dfrac{\partial}{\partial x_1} & \dfrac{\partial}{\partial x_2} & \dfrac{\partial}{\partial x_3} \end{bmatrix} \;=\; \begin{bmatrix} u_{1,1} & u_{1,2} & u_{1,3} \\ u_{2,1} & u_{2,2} & u_{2,3} \\ u_{3,1} & u_{3,2} & u_{3,3} \end{bmatrix}\;, \tag{7.8}$$

where u_i, $i = 1, 2, 3$ are the Cartesian components of the displacement $\mathbf{u}$ and $u_{i,j} = \dfrac{\partial u_i}{\partial x_j}$. Taking into account (7.7) and (7.8), it follows that the Cartesian components of the strain tensor are

$$[\boldsymbol{\epsilon}] \;=\; \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{bmatrix} \;=\; \begin{bmatrix} u_{1,1} & \frac{1}{2}(u_{1,2}+u_{2,1}) & \frac{1}{2}(u_{1,3}+u_{3,1}) \\ \frac{1}{2}(u_{1,2}+u_{2,1}) & u_{2,2} & \frac{1}{2}(u_{2,3}+u_{3,2}) \\ \frac{1}{2}(u_{1,3}+u_{3,1}) & \frac{1}{2}(u_{2,3}+u_{3,2}) & u_{3,3} \end{bmatrix}\;. \tag{7.9}$$

Clearly, the strain tensor is symmetric and, in view of equation (7.5), so is the stress tensor. The trace of the strain, $\mathrm{tr}\,\boldsymbol{\epsilon}$, which appears in (7.5), is equal to $u_{1,1} + u_{2,2} + u_{3,3}$.

The strong form of the linear elastostatics problem can be summarized as follows: given $\mathbf{f}$ in Ω, $\bar{\mathbf{t}}$ on Γ_q, and $\bar{\mathbf{u}}$ on Γ_u, find $\mathbf{u}$ in Ω, such that equations (7.2) are satisfied. The precise sense in which equations (7.2) are elliptic will be discussed later in this section.

A Galerkin-based weak form for linear elastostatics can be deduced in analogy with earlier developments in Section 3.2, by first assuming that: (a) the Dirichlet boundary conditions are satisfied at the outset, (b) the weighting functions $\mathbf{w}_\Omega$ and $\mathbf{w}_q$ are chosen to be identical,

that is, $\mathbf{w}_\Omega = \mathbf{w}_q = \mathbf{w}$, and (c) that $\mathbf{w} = \mathbf{0}$ on Γ_u. Taking into account the preceding conditions, the weak form may be written as

$$\int_\Omega \mathbf{w} \cdot (-\nabla \cdot \boldsymbol{\sigma} - \mathbf{f})\, d\Omega + \int_{\Gamma_q} \mathbf{w} \cdot (\boldsymbol{\sigma}\mathbf{n} - \bar{\mathbf{t}})\, d\Gamma \;=\; 0\ . \tag{7.10}$$

Concentrating on the first term of the left-hand side of equation (7.10), use the Einsteinian summation convention[2] to write

$$\begin{aligned}
\int_\Omega \mathbf{w} \cdot (\nabla \cdot \boldsymbol{\sigma})\, d\Omega \;&=\; \int_\Omega w_i \sigma_{ij,j}\, d\Omega \\
&=\; \int_\Omega (w_i \sigma_{ij})_{,j}\, d\Omega - \int_\Omega w_{i,j}\sigma_{ij}\, d\Omega \\
&=\; \int_{\partial\Omega} w_i \sigma_{ij} n_j\, d\Gamma - \int_\Omega w_{i,j}\sigma_{ij}\, d\Omega \\
&=\; \int_{\Gamma_q} w_i \sigma_{ij} n_j\, d\Gamma - \int_\Omega w_{i,j}\sigma_{ij}\, d\Omega\ ,
\end{aligned} \tag{7.11}$$

where use is made of integration by parts, the divergence theorem, and the fact that $\mathbf{w} = \mathbf{0}$ on Γ_u. Further, it is easily seen that

$$w_{i,j}\sigma_{ij} \;=\; \left[\frac{1}{2}(w_{i,j} + w_{j,i}) + \frac{1}{2}(w_{i,j} - w_{j,i})\right]\sigma_{ij} \;=\; \frac{1}{2}(w_{i,j} + w_{j,i})\sigma_{ij}\ . \tag{7.12}$$

Taking into account equations (7.11) and (7.12), it follows that

$$\int_\Omega \mathbf{w} \cdot (\nabla \cdot \boldsymbol{\sigma})\, d\Omega \;=\; \int_{\Gamma_q} \mathbf{w} \cdot (\boldsymbol{\sigma}\mathbf{n})\, d\Gamma \;- \int_\Omega \nabla_s \mathbf{w} : \boldsymbol{\sigma}\, d\Omega\ , \tag{7.13}$$

where $\nabla_s \mathbf{w} : \boldsymbol{\sigma}$ denotes the contraction of the tensors $\nabla_s \mathbf{w}$ and $\boldsymbol{\sigma}$, expressed in component form as $\nabla_s \mathbf{w} : \boldsymbol{\sigma} = w_{i,j}\sigma_{ij}$. With the aid of (7.13), the weak form in (7.10) can be rewritten as

$$\int_\Omega \nabla_s \mathbf{w} : \boldsymbol{\sigma}\, d\Omega \;=\; \int_\Omega \mathbf{w} \cdot \mathbf{f}\, d\Omega + \int_{\Gamma_q} \mathbf{w} \cdot \bar{\mathbf{t}}\, d\Gamma\ . \tag{7.14}$$

Given $\mathbf{f}$, $\bar{\mathbf{t}}$, $\bar{\mathbf{u}}$, and the stress-strain law (7.5), the weak form of the problem of linear elastostatics amounts to finding $\mathbf{u} \in \mathcal{U}$, such that equation (7.14) holds for all admissible $\mathbf{w} \in \mathcal{W}$ and $\boldsymbol{\sigma}$ is related to $\mathbf{u}$ through (7.5) and (7.7). Here, the admissible spaces $\mathcal{U}$ and $\mathcal{W}$ are defined as

$$\begin{aligned}
\mathcal{U} \;&=\; \left\{\mathbf{u} \in \mathbf{H}^1(\Omega) \quad | \quad \mathbf{u} = \bar{\mathbf{u}} \text{ on } \Gamma_u\right\}\ , \\
\mathcal{W} \;&=\; \left\{\mathbf{w} \in \mathbf{H}^1(\Omega) \quad | \quad \mathbf{w} = \mathbf{0} \text{ on } \Gamma_u\right\}\ ,
\end{aligned} \tag{7.15}$$

[2]According to this convention, all indices that appear in a product term twice (*dummy* indices) are summed from 1 to 3, while all indices that appear once (*free* indices) are assumed to take value 1,2 or 3. In this convention, no indices are allowed to appear in a product term more than twice.

where $\mathbf{H}^1(\Omega)$ is the Sobolev space of all vector functions with square-integrable first derivatives. Adopting the terminology of *virtual displacements*, equation (7.13) can be viewed as a statement of the *theorem of virtual work*, according to which the work done by the actual internal forces (that is, the stress $\boldsymbol{\sigma}$) over the virtual strains $\nabla_s \mathbf{w}$ is equal to the work done by the actual external forces (that is, the body force $\mathbf{f}$ and surface traction $\bar{\mathbf{t}}$) over the virtual displacement $\mathbf{w}$.

The weak form (7.14) can be written operationally as

$$B(\mathbf{w},\mathbf{u}) \;=\; (\mathbf{w},\mathbf{f}) + (\mathbf{w},\bar{\mathbf{t}})_{\Gamma_q} \;, \tag{7.16}$$

where

$$\begin{aligned} B(\mathbf{w},\mathbf{u}) \;&=\; \int_\Omega \nabla_s \mathbf{w} : \boldsymbol{\sigma}\, d\Omega \;, \\ (\mathbf{w},\mathbf{f}) \;&=\; \int_\Omega \mathbf{w}\cdot\mathbf{f}\, d\Omega \;, \\ (\mathbf{w},\bar{\mathbf{t}})_{\Gamma_q} \;&=\; \int_{\Gamma_q} \mathbf{w}\cdot\bar{\mathbf{t}}\, d\Gamma \;. \end{aligned} \tag{7.17}$$

The preceding bilinear form $B(\cdot,\cdot)$ is symmetric. To see this in a transparent manner, one may express the components of tensorial quantities such as $\nabla_s \mathbf{w}$ and $\boldsymbol{\sigma}$ in vector form. In particular, one may start with the 3×3 symmetric matrix of components of the infinitesimal strain tensor in $(7.9)_1$, and rewrite them with a slight change in notation as

$$\langle \boldsymbol{\epsilon} \rangle \;=\; \left[\begin{array}{cccccc} \epsilon_{11} & \epsilon_{22} & \epsilon_{33} & 2\epsilon_{12} & 2\epsilon_{23} & 2\epsilon_{31} \end{array}\right]^T \;, \tag{7.18}$$

where $\langle\cdot\rangle$ is reserved for representation of tensors using column vector form. Likewise, the 3×3 symmetric matrix of components of the stress tensor in (7.3) can be written in column-vector form as

$$\langle \boldsymbol{\sigma} \rangle \;=\; \left[\begin{array}{cccccc} \sigma_{11} & \sigma_{22} & \sigma_{33} & \sigma_{12} & \sigma_{23} & \sigma_{31} \end{array}\right]^T \;. \tag{7.19}$$

Note that the factor "2" in the last three rows of the strain vector in (7.18) is included to ensure that the contraction $\boldsymbol{\sigma} : \boldsymbol{\epsilon} = \sigma_{ij}\epsilon_{ij}$ is defined consistently when employing the vector convention, that is $\boldsymbol{\sigma} : \boldsymbol{\epsilon} = \langle\boldsymbol{\sigma}\rangle^T\langle\boldsymbol{\epsilon}\rangle$. The preceding vector notation is employed in the remainder of this section.

The stress-strain law (7.5) can be written using the vector convention as

$$\langle \boldsymbol{\sigma} \rangle \;=\; [\mathbf{D}]\langle \boldsymbol{\epsilon} \rangle \;, \tag{7.20}$$

where $[\mathbf{D}]$ is a 6×6 elasticity matrix, such that

$$\begin{bmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{33} \\ \sigma_{12} \\ \sigma_{23} \\ \sigma_{31} \end{bmatrix} = \begin{bmatrix} \lambda+2\mu & \lambda & \lambda & 0 & 0 & 0 \\ \lambda & \lambda+2\mu & \lambda & 0 & 0 & 0 \\ \lambda & \lambda & \lambda+2\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu \end{bmatrix} \begin{bmatrix} \epsilon_{11} \\ \epsilon_{22} \\ \epsilon_{33} \\ 2\epsilon_{12} \\ 2\epsilon_{23} \\ 2\epsilon_{31} \end{bmatrix} . \tag{7.21}$$

In the special case of plane strain on the $(1,2)$-plane, where $\epsilon_{33} = \epsilon_{13} = \epsilon_{23} = 0$, the preceding system reduces to

$$\begin{bmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{12} \end{bmatrix} = \begin{bmatrix} \lambda+2\mu & \lambda & 0 \\ \lambda & \lambda+2\mu & 0 \\ 0 & 0 & \mu \end{bmatrix} \begin{bmatrix} \epsilon_{11} \\ \epsilon_{22} \\ 2\epsilon_{12} \end{bmatrix} , \tag{7.22}$$

while $\sigma_{33} = \lambda(\epsilon_{11} + \epsilon_{22})$. Lastly, in the special case of plane stress on the $(1,2)$-plane, where $\sigma_{33} = \sigma_{13} = \sigma_{23} = 0$, one may write the stress-strain relations in reduced matrix form as

$$\begin{bmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{12} \end{bmatrix} = \begin{bmatrix} \dfrac{4\mu(\lambda+\mu)}{\lambda+2\mu} & \dfrac{2\lambda\mu}{\lambda+2\mu} & 0 \\ \dfrac{2\lambda\mu}{\lambda+2\mu} & \dfrac{4\mu(\lambda+\mu)}{\lambda+2\mu} & 0 \\ 0 & 0 & \mu \end{bmatrix} \begin{bmatrix} \epsilon_{11} \\ \epsilon_{22} \\ 2\epsilon_{12} \end{bmatrix} , \tag{7.23}$$

while $\epsilon_{33} = -\frac{\lambda}{\lambda+2\mu}(\epsilon_{11} + \epsilon_{22})$.

Since the matrix $[\mathbf{D}]$ is always symmetric, it follows that the integrand of the bilinear form in (7.17) can be written with the aid of (7.20) as

$$\nabla_s \mathbf{w} : \boldsymbol{\sigma} \;=\; \langle \boldsymbol{\epsilon}(\mathbf{w}) \rangle^T [\mathbf{D}] \langle \boldsymbol{\epsilon}(\mathbf{u}) \rangle \;=\; \langle \boldsymbol{\epsilon}(\mathbf{u}) \rangle^T [\mathbf{D}] \langle \boldsymbol{\epsilon}(\mathbf{w}) \rangle \;, \tag{7.24}$$

which shows that the bilinear form in $(7.17)_1$ is indeed symmetric. Using the vector forms, one may rewrite (7.14) with matrices as

$$\int_\Omega \langle \boldsymbol{\epsilon}(\mathbf{w}) \rangle^T [\mathbf{D}] \langle \boldsymbol{\epsilon}(\mathbf{u}) \rangle \, d\Omega \;=\; \int_\Omega [\mathbf{w}]^T [\mathbf{f}] \, d\Omega + \int_{\Gamma_q} [\mathbf{w}]^T [\bar{\mathbf{t}}] \, d\Gamma \;. \tag{7.25}$$

Generally, the matrix representation of weak forms, such as (7.25), is more practical than the corresponding tensorial representation for the purpose of computer implementation.

The symmetry of the bilinear form in $(7.17)_1$ implies that Vainberg's theorem of Section 4.2 is applicable to the weak form (7.16) or, equivalently (7.25), therefore there exists a functional $I[\mathbf{u}]$, given by

$$\begin{aligned} I[\mathbf{u}] &= \frac{1}{2}B(\mathbf{u},\mathbf{u}) - (\mathbf{u},\mathbf{f}) - (\mathbf{u},\bar{\mathbf{t}})_{\Gamma_q} \\ &= \frac{1}{2}\int_\Omega \langle\boldsymbol{\epsilon}(\mathbf{u})\rangle^T[\mathbf{D}]\langle\boldsymbol{\epsilon}(\mathbf{u})\rangle\, d\Omega - \int_\Omega [\mathbf{u}]^T[\mathbf{f}]\, d\Omega - \int_{\Gamma_q}[\mathbf{u}]^T[\bar{\mathbf{t}}]\, d\Gamma\ , \end{aligned} \tag{7.26}$$

whose extremization yields the weak form for the problem of linear elastostatics. The first term on the right-hand side of $(7.26)_{1,2}$ is the strain energy, while the second and third terms represent together the energy associated with the applied forces. The functional $I[\mathbf{u}]$ in (7.26) is referred to as the *total potential energy* of the solid body occupying the region Ω.

The *Minimum Total Potential Energy theorem* states that among all displacements $\mathbf{u} \in \mathcal{U}$, the actual solution $\mathbf{u}$ renders the total potential energy an absolute minimum. To prove this theorem, note that the extremization of $I[\mathbf{u}]$ yields the condition

$$\delta I[\mathbf{u}] = \int_\Omega \langle\boldsymbol{\epsilon}(\delta\mathbf{u})\rangle^T[\mathbf{D}]\langle\boldsymbol{\epsilon}(\mathbf{u})\rangle\, d\Omega - \int_\Omega [\delta\mathbf{u}]^T[\mathbf{f}]\, d\Omega - \int_{\Gamma_q}[\delta\mathbf{u}]^T[\bar{\mathbf{t}}]\, d\Gamma = 0\ , \tag{7.27}$$

which coincides with the weak form (7.25) when setting $\delta\mathbf{u} = \mathbf{w}$. Furthermore, given any $\delta\mathbf{u} \in \mathcal{W}$, it is easy to conclude with the aid of (7.27) that

$$I[\mathbf{u}+\delta\mathbf{u}] - I[\mathbf{u}] = \int_\Omega \langle\boldsymbol{\epsilon}(\delta\mathbf{u})\rangle^T[\mathbf{D}]\langle\boldsymbol{\epsilon}(\delta\mathbf{u})\rangle\, d\Omega\ . \tag{7.28}$$

The right-hand side of (7.28) is necessarily positive for $\delta\mathbf{u} \neq \mathbf{0}$, given that $[\mathbf{D}]$ is a positive-definite matrix, since its eigenvalues $\lambda_1 = 3\lambda + 2\mu$, are $\lambda_{2,3} = 2\mu$, and $\lambda_{4,5,6} = \mu$ are all positive in light of (7.6). Hence, $I[\mathbf{u}] \leq I[\mathbf{v}]$, for all $\mathbf{v} \in \mathcal{U}$, therefore $\mathbf{u}$ minimizes I over all admissible displacements.

7.2.1 A Galerkin approximation to the weak form

The discrete counterpart of (7.25) can be written as

$$\int_\Omega \langle\boldsymbol{\epsilon}(\mathbf{w}_h)\rangle^T[\mathbf{D}]\langle\boldsymbol{\epsilon}(\mathbf{u}_h)\rangle\, d\Omega = \int_\Omega [\mathbf{w}_h]^T[\mathbf{f}]\, d\Omega + \int_{\Gamma_q}[\mathbf{w}_h]^T[\bar{\mathbf{t}}]\, d\Gamma\ , \tag{7.29}$$

where $\mathbf{u}_h \in \mathcal{U}_h \subset \mathcal{U}$ and where $\mathbf{w}_h \in \mathcal{W}_h \subset \mathcal{W}$. Within a given finite element e with domain Ω^e, one may write

$$\mathbf{u}_h = \sum_{i=1}^{\texttt{nen}} N_i^e \mathbf{u}_i^e \quad , \quad \mathbf{w}_h = \sum_{i=1}^{\texttt{nen}} N_i^e \mathbf{w}_i^e\ , \tag{7.30}$$

where `nen` is the total number of element nodes, $\mathbf{u}_i^e$ are the `ndf` degrees of freedom of node i, and $\mathbf{w}_i^e$ are the `ndf` values of the weighting function at node i (with `ndf` being nominally equal to 3 in the three-dimensional case). Equations (7.30) can be written compactly as

$$\mathbf{u}_h \;=\; [\mathbf{N}^e][\mathbf{u}^e] \qquad , \qquad \mathbf{w}_h \;=\; [\mathbf{N}^e][\mathbf{w}^e] \;, \tag{7.31}$$

in terms of the `ndf`$*$`nen` vectors

$$[\mathbf{u}^e] \;=\; \begin{bmatrix} \mathbf{u}_1^e \\ \mathbf{u}_2^e \\ \cdot \\ \cdot \\ \mathbf{u}_{nen}^e \end{bmatrix} \quad , \quad [\mathbf{w}^e] \;=\; \begin{bmatrix} \mathbf{w}_1^e \\ \mathbf{w}_2^e \\ \cdot \\ \cdot \\ \mathbf{w}_{nen}^e \end{bmatrix} \;, \tag{7.32}$$

and the `ndf`$\times$`ndf`$*$`nen` matrix

$$[\mathbf{N}^e] \;=\; \begin{bmatrix} N_1^e \mathbf{I}_{ndf} & N_2^e \mathbf{I}_{ndf} & \cdot & \cdot & N_{nen}^e \mathbf{I}_{ndf} \end{bmatrix} \;, \tag{7.33}$$

and $\mathbf{I}_{ndf}$ is the `ndf`$\times$`ndf` identity matrix.

The strain tensor, expressed in vector form, can be also written as

$$\begin{bmatrix} \epsilon_{11} \\ \epsilon_{22} \\ \epsilon_{33} \\ 2\epsilon_{12} \\ 2\epsilon_{23} \\ 2\epsilon_{31} \end{bmatrix} \;=\; \begin{bmatrix} \dfrac{\partial}{\partial x_1} & 0 & 0 \\ 0 & \dfrac{\partial}{\partial x_2} & 0 \\ 0 & 0 & \dfrac{\partial}{\partial x_3} \\ \dfrac{\partial}{\partial x_2} & \dfrac{\partial}{\partial x_1} & 0 \\ 0 & \dfrac{\partial}{\partial x_3} & \dfrac{\partial}{\partial x_2} \\ \dfrac{\partial}{\partial x_3} & 0 & \dfrac{\partial}{\partial x_1} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \;. \tag{7.34}$$

This implies, with the aid of (7.30) that the strains in Ω^e are expressed as

$$\langle \boldsymbol{\epsilon}(\mathbf{u}_h) \rangle \;=\; \sum_{i=1}^{\texttt{nen}} [\mathbf{B}_i^e][\mathbf{u}_i^e] \quad , \quad \langle \boldsymbol{\epsilon}(\mathbf{w}_h) \rangle \;=\; \sum_{i=1}^{\texttt{nen}} [\mathbf{B}_i^e][\mathbf{w}_i^e] \;, \tag{7.35}$$

in terms of the 6×`ndf` strain-displacement matrix

$$[\mathbf{B}_i^e] \;=\; \begin{bmatrix} \dfrac{\partial N_i^e}{\partial x_1} & 0 & 0 \\ 0 & \dfrac{\partial N_i^e}{\partial x_2} & 0 \\ 0 & 0 & \dfrac{\partial N_i^e}{\partial x_3} \\ \dfrac{\partial N_i^e}{\partial x_2} & \dfrac{\partial N_i^e}{\partial x_1} & 0 \\ 0 & \dfrac{\partial N_i^e}{\partial x_3} & \dfrac{\partial N_i^e}{\partial x_2} \\ \dfrac{\partial N_i^e}{\partial x_3} & 0 & \dfrac{\partial N_i^e}{\partial x_1} \end{bmatrix} . \tag{7.36}$$

Again, resorting to compact notation, equations (7.35) can be recast in the form

$$\langle \boldsymbol{\epsilon}(\mathbf{u}_h)\rangle \;=\; [\mathbf{B}^e][\mathbf{u}^e] \quad , \quad \langle \boldsymbol{\epsilon}(\mathbf{w}_h)\rangle \;=\; [\mathbf{B}^e][\mathbf{w}^e] \; , \tag{7.37}$$

where $[\mathbf{B}^e]$ is a 6×`ndf`∗`nen` matrix defined as

$$[\mathbf{B}^e] \;=\; \begin{bmatrix} \mathbf{B}_1^e & \mathbf{B}_2^e & \cdot & \cdot & \mathbf{B}_{nen}^e \end{bmatrix} . \tag{7.38}$$

The weak form (7.29) can be applied to element e, so that

$$\int_{\Omega^e} \langle \boldsymbol{\epsilon}(\mathbf{w}_h)\rangle^T [\mathbf{D}] \langle \boldsymbol{\epsilon}(\mathbf{u}_h)\rangle \, d\Omega \;=\; \int_{\Omega^e} [\mathbf{w}_h]^T [\mathbf{f}] \, d\Omega + \int_{\partial\Omega^e \cap \Gamma_q} [\mathbf{w}_h]^T [\bar{\mathbf{t}}] \, d\Gamma + \int_{\partial\Omega^e \setminus \partial\Omega} [\mathbf{w}_h]^T [\mathbf{t}] \, d\Gamma \; . \tag{7.39}$$

Appealing to equations (7.31) and (7.37), the preceding weak form is written as

$$\begin{aligned} &\int_{\Omega^e} ([\mathbf{B}^e][\mathbf{w}^e])^T [\mathbf{D}] ([\mathbf{B}^e][\mathbf{u}^e]) \, d\Omega \\ &\qquad = \int_{\Omega^e} ([\mathbf{N}^e][\mathbf{w}^e])^T [\mathbf{f}] \, d\Omega + \int_{\partial\Omega^e \cap \Gamma_q} ([\mathbf{N}^e][\mathbf{w}^e])^T [\bar{\mathbf{t}}] \, d\Gamma + \int_{\partial\Omega^e \setminus \partial\Omega} ([\mathbf{N}^e][\mathbf{w}^e])^T [\mathbf{t}] \, d\Gamma \end{aligned} \tag{7.40}$$

or

$$\begin{aligned} [\mathbf{w}^e]^T \Bigg[&\left\{ \int_{\Omega^e} [\mathbf{B}^e]^T [\mathbf{D}] [\mathbf{B}^e] \, d\Omega \right\} [\mathbf{u}^e] \\ &- \int_{\Omega^e} [\mathbf{N}^e]^T [\mathbf{f}] \, d\Omega - \int_{\partial\Omega^e \cap \Gamma_q} [\mathbf{N}^e]^T [\bar{\mathbf{t}}] \, d\Gamma - \int_{\partial\Omega^e \setminus \partial\Omega} [\mathbf{N}^e]^T [\mathbf{t}] \, d\Gamma \Bigg] \;=\; 0 \; . \end{aligned} \tag{7.41}$$

Given the arbitrariness of $[\mathbf{w}^e]$, equation (7.41) leads to the linear system

$$[\mathbf{K}^e][\mathbf{u}^e] \;=\; [\mathbf{F}^e] + [\mathbf{F}^{\text{int},e}] \; , \tag{7.42}$$

where

$$
\begin{aligned}
[\mathbf{K}^e] &= \int_{\Omega^e} [\mathbf{B}^e]^T [\mathbf{D}] [\mathbf{B}^e] \, d\Omega \ , \\
[\mathbf{F}^e] &= \int_{\Omega^e} [\mathbf{N}^e]^T [\mathbf{f}] \, d\Omega + \int_{\partial\Omega^e \cap \Gamma_q} [\mathbf{N}^e]^T [\bar{\mathbf{t}}] \, d\Gamma \ , \\
[\mathbf{F}^{\text{int},e}] &= \int_{\partial\Omega^e \setminus \partial\Omega} [\mathbf{N}^e]^T [\mathbf{t}] \, d\Gamma \ .
\end{aligned}
\tag{7.43}
$$

As already discussed in Section 6.2, the forcing vector $[\mathbf{F}^{\text{int},e}]$ due to interelement tractions is unknown at the outset and its contribution to the global forcing vector will be neglected upon assembly.

Example 7.2.1: 4-node isoparametric quadrilateral in plane strain
The element interpolation functions N_i^e, $i = 1-4$, for this element are given in equation (5.89) relative to the natural coordinates (ξ, η). The element interpolation array $[\mathbf{N}^e]$ is now given by

$$
[\mathbf{N}^e] = \begin{bmatrix} N_1^e \mathbf{I}_2 & N_2^e \mathbf{I}_2 & N_3^e \mathbf{I}_2 & N_4^e \mathbf{I}_2 \end{bmatrix} \ ,
$$

and is of dimension 2×8 (note that here `nen` $= 4$ and `ndf` $=2$). Moreover, the strain-displacement array $[\mathbf{B}_i^e]$ is given by

$$
[\mathbf{B}_i^e] = \begin{bmatrix} \dfrac{\partial N_i^e}{\partial x_1} & 0 \\ 0 & \dfrac{\partial N_i^e}{\partial x_2} \\ \dfrac{\partial N_i^e}{\partial x_2} & \dfrac{\partial N_i^e}{\partial x_1} \end{bmatrix} \ ,
$$

hence

$$
[\mathbf{B}^e] = \begin{bmatrix} \dfrac{\partial N_1^e}{\partial x_1} & 0 & \dfrac{\partial N_2^e}{\partial x_1} & 0 & \dfrac{\partial N_3^e}{\partial x_1} & 0 & \dfrac{\partial N_4^e}{\partial x_1} & 0 \\ 0 & \dfrac{\partial N_1^e}{\partial x_2} & 0 & \dfrac{\partial N_2^e}{\partial x_2} & 0 & \dfrac{\partial N_3^e}{\partial x_2} & 0 & \dfrac{\partial N_4^e}{\partial x_2} \\ \dfrac{\partial N_1^e}{\partial x_2} & \dfrac{\partial N_1^e}{\partial x_1} & \dfrac{\partial N_2^e}{\partial x_2} & \dfrac{\partial N_2^e}{\partial x_1} & \dfrac{\partial N_3^e}{\partial x_2} & \dfrac{\partial N_3^e}{\partial x_1} & \dfrac{\partial N_4^e}{\partial x_2} & \dfrac{\partial N_4^e}{\partial x_1} \end{bmatrix} \ .
$$

Given that the elasticity matrix $[\mathbf{D}]$ for plane strain is of dimension 3×3, as in (7.22), it follows from the above representation of $[\mathbf{B}^e]$ that the element stiffness matrix $[\mathbf{K}^e]$ in $(7.43)_1$ is of dimension 8×8. ◀

7.2.2 On the order of numerical integration

The stiffness matrix and forcing vector in equations $(7.43)_{1,2}$ require the evaluation of domain and boundary integrals. In the case of general isoparametric elements, the stiffness matrix

typically involves rational polynomials of the natural coordinates (ξ, η, ζ). To see this, recall that the matrix $[\mathbf{B}^e]$ contains derivatives of the element interpolation functions $N_i^e(\xi, \eta, \zeta)$ with respect to the physical coordinates (x_1, x_2, x_3). Appealing to the chain rule, a typical such derivative $\dfrac{\partial N_i^e}{\partial x_j}$ can be written as

$$\frac{\partial N_i^e}{\partial x_j} = \frac{\partial N_i^e}{\partial \xi}\frac{\partial \hat{\xi}}{\partial x_j} + \frac{\partial N_i^e}{\partial \eta}\frac{\partial \hat{\eta}}{\partial x_j} + \frac{\partial N_i^e}{\partial \zeta}\frac{\partial \hat{\zeta}}{\partial x_j} \ . \tag{7.44}$$

Here, the functions $\hat{\xi}$, $\hat{\eta}$ and $\hat{\zeta}$ constitute the inverse of the functions $\hat{x}$, $\hat{y}$ and $\hat{z}$ in (5.84).[3] While terms of the type $\dfrac{\partial N_i^e}{\partial \xi}$ in equation (7.44) are clearly polynomial in (ξ, η, ζ), this is not the case with terms of the type $\dfrac{\partial \hat{\xi}}{\partial x_j}$, which are, in fact, inverse polynomial in (ξ, η, ζ). To find an analytical expression for derivatives of the type $\dfrac{\partial N_i^e}{\partial x_j}$, write

$$\begin{aligned}
\frac{\partial N_i^e}{\partial \xi} &= \frac{\partial N_i^e}{\partial x_1}\frac{\partial \hat{x}_1}{\partial \xi} + \frac{\partial N_i^e}{\partial x_2}\frac{\partial \hat{x}_2}{\partial \xi} + \frac{\partial N_i^e}{\partial x_3}\frac{\partial \hat{x}_3}{\partial \xi} \ , \\
\frac{\partial N_i^e}{\partial \eta} &= \frac{\partial N_i^e}{\partial x_1}\frac{\partial \hat{x}_1}{\partial \eta} + \frac{\partial N_i^e}{\partial x_2}\frac{\partial \hat{x}_2}{\partial \eta} + \frac{\partial N_i^e}{\partial x_3}\frac{\partial \hat{x}_3}{\partial \eta} \ , \\
\frac{\partial N_i^e}{\partial \zeta} &= \frac{\partial N_i^e}{\partial x_1}\frac{\partial \hat{x}_1}{\partial \zeta} + \frac{\partial N_i^e}{\partial x_2}\frac{\partial \hat{x}_2}{\partial \zeta} + \frac{\partial N_i^e}{\partial x_3}\frac{\partial \hat{x}_3}{\partial \zeta} \ ,
\end{aligned}$$

or, in matrix form

$$\begin{bmatrix} \dfrac{\partial N_i^e}{\partial \xi} \\ \dfrac{\partial N_i^e}{\partial \eta} \\ \dfrac{\partial N_i^e}{\partial \zeta} \end{bmatrix} = \underbrace{\begin{bmatrix} \dfrac{\partial \hat{x}_1}{\partial \xi} & \dfrac{\partial \hat{x}_2}{\partial \xi} & \dfrac{\partial \hat{x}_3}{\partial \xi} \\ \dfrac{\partial \hat{x}_1}{\partial \eta} & \dfrac{\partial \hat{x}_2}{\partial \eta} & \dfrac{\partial \hat{x}_3}{\partial \eta} \\ \dfrac{\partial \hat{x}_1}{\partial \zeta} & \dfrac{\partial \hat{x}_2}{\partial \zeta} & \dfrac{\partial \hat{x}_3}{\partial \zeta} \end{bmatrix}}_{[\mathbf{J}^e]^T} \begin{bmatrix} \dfrac{\partial N_i^e}{\partial x_1} \\ \dfrac{\partial N_i^e}{\partial x_2} \\ \dfrac{\partial N_i^e}{\partial x_3} \end{bmatrix} \ . \tag{7.45}$$

Equation (7.45) demonstrates that the computation of partial derivatives of the type $\dfrac{\partial N_i^e}{\partial x_j}$ requires inversion of the 3×3 Jacobian matrix $[\mathbf{J}^e]^T$. This inverse is equal to $\dfrac{1}{J^e}\text{adj}[\mathbf{J}^e]^T$, where $\text{adj}[\mathbf{J}^e]$ is the adjugate of $[\mathbf{J}^e]$ and $J^e = \det \mathbf{J}^e$. Given that J^e is the product of

[3] With a slight abuse of notation, the functions $\hat{x}$, $\hat{y}$ and $\hat{z}$ of (5.84) are substituted here by functions $\hat{x}_1$, $\hat{x}_2$ and $\hat{x}_3$.

polynomials in (ξ, η, ζ), the presence of the determinant J^e in the denominator of $[\mathbf{J}^e]^{-T}$ establishes the rational polynomial form of $\dfrac{\partial N_i^e}{\partial x_j}$ (hence, also of $[\mathbf{B}^e]$ and $[\mathbf{K}^e]$).

Exact integration of $[\mathbf{K}^e]$ is possible, yet, for general isoparametric mappings, is typically cumbersome and ill-posed, which justifies the use of numerical integration using Gaussian quadrature. Two criteria exist for the choice of the order of the numerical integration, as follows:

(a) *Minimum order of integration for completeness*

Recalling the definition of the bilinear form $B(\cdot,\cdot)$ in equation $(7.17)_1$, note that the highest derivative it involves is of order $p = 1$. Therefore, as argued earlier, completeness requires that the finite element fields $\mathbf{u}_h$ be capable of representing any polynomial up to degree $q \geq 1$. Therefore, at the minimum, completeness requires that $\mathbf{u}_h$ (and also $\mathbf{w}_h$, in the Bubnov-Galerkin approximation) be capable of representing a linear distribution of the displacement, hence a constant distribution of the strain $\boldsymbol{\epsilon}$. In this case, and assuming that the elasticity matrix $[\mathbf{D}]$ is also constant within the element, the bilinear form becomes

$$B(\mathbf{w}_h, \mathbf{u}_h) \;=\; \int_{\Omega^e} \langle \bar{\boldsymbol{\epsilon}}(\mathbf{w}) \rangle^T [\mathbf{D}] \langle \bar{\boldsymbol{\epsilon}}(\mathbf{u}) \rangle \, d\Omega \;=\; \langle \bar{\boldsymbol{\epsilon}}(\mathbf{w}) \rangle^T [\mathbf{D}] \langle \bar{\boldsymbol{\epsilon}}(\mathbf{u}) \rangle \int_{\Omega^e} d\Omega \;, \tag{7.46}$$

where $\langle \bar{\boldsymbol{\epsilon}}(\mathbf{w}) \rangle$ and $\langle \bar{\boldsymbol{\epsilon}}(\mathbf{u}) \rangle$ are constant vectors. It follows that the minimum order of integration for completeness is such that the integral $\int_{\Omega^e} d\Omega$ be evaluated exactly. Recalling that

$$\int_{\Omega^e} d\Omega \;=\; \int_{\Omega^e_\square} J^e \, d\xi d\eta d\zeta \;, \tag{7.47}$$

this implies that the minimum order of integration for completeness is the order required to integrate exactly the Jacobian determinant J^e.

As an example, consider the 4-node quadrilateral element in plane strain, for which it has been established in Section 5.6 that the Jacobian determinant is linear in (ξ, η). It follows immediately that in this case the minimum order of Gaussian integration for completeness is 1×1 (namely, one-point Gaussian integration).

(b) *Minimum order of integration for stability*

The numerical integration of the element arrays should preserve the spectral properties of the original problem. This effectively means that the numerical integration should not introduce artificial zero eigenvalues in the element stiffness matrix.

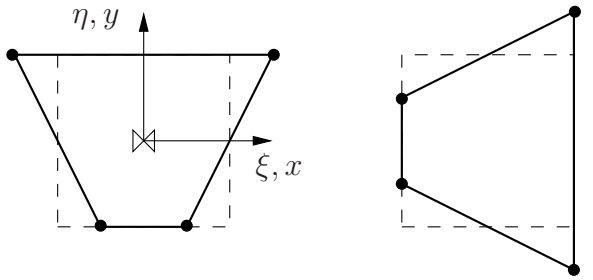


Figure 7.2. *Zero-energy modes for the 4-node quadrilateral with* 1×1 *Gaussian quadrature*

To illustrate the above point, consider again the 4-node isoparametric quadrilateral element in plane strain with the previously deduced minimum order of integration for completeness, namely 1×1 Gaussian quadrature. The deformation modes shown in Figure 7.2 are associated in the exact elastostatics problem with positive strain energy. Indeed, letting for simplicity the natural and physical domains and coordinates coincide (which implies that $J^e = 1$), the displacement and strain vectors associated with one of these modes is

$$[\mathbf{u}_h] \;=\; \begin{bmatrix} \alpha\xi\eta \\ 0 \end{bmatrix} \quad , \quad \langle \boldsymbol{\epsilon}(\mathbf{u}_h) \rangle \;=\; \begin{bmatrix} \alpha\eta \\ 0 \\ \alpha\xi \end{bmatrix} \;, \tag{7.48}$$

where $\alpha(>0)$ is a constant. The strain energy of this deformation mode is

$$\begin{aligned} W \;&=\; \frac{1}{2}\int_{\Omega^e} \langle \boldsymbol{\epsilon}(\mathbf{u}_h) \rangle^T [\mathbf{D}] \langle \boldsymbol{\epsilon}(\mathbf{u}_h) \rangle \, d\Omega \\ &=\; \frac{\alpha^2}{2}\int_{\Omega^e} \begin{bmatrix} \eta & 0 & \xi \end{bmatrix} \begin{bmatrix} \lambda+2\mu & \lambda & 0 \\ \lambda & \lambda+2\mu & 0 \\ 0 & 0 & \mu \end{bmatrix} \begin{bmatrix} \eta \\ 0 \\ \xi \end{bmatrix} d\Omega \;>\; 0 \;, \end{aligned} \tag{7.49}$$

since $[\mathbf{D}]$ is positive-definite. However, when using 1×1 Gauss quadrature, it is readily seen from (7.49) that the strain energy of this mode is zero, as the single Gauss point is located at $\xi = \eta = 0$. Deformation modes which are artificially associated with zero strain energy due to low order of numerical integration of the element stiffness matrix are referred to as *zero-energy modes*. The zero energy mode of equation (7.48) disappears upon using 2×2 Gaussian quadrature, since, in this case, its strain energy

is approximated by

$$W \doteq \frac{\alpha^2}{2}\sum_{l=1}^{2}\sum_{m=1}^{2}\begin{bmatrix}\eta_m & 0 & \xi_l\end{bmatrix}\begin{bmatrix}\lambda+2\mu & \lambda & 0\\ \lambda & \lambda+2\mu & 0\\ 0 & 0 & \mu\end{bmatrix}\begin{bmatrix}\eta_m\\ 0\\ \xi_l\end{bmatrix} > 0\,, \qquad (7.50)$$

where $(\xi_l,\eta_m)=(\pm\frac{1}{\sqrt{3}},\pm\frac{1}{\sqrt{3}})$, $l,m=1,2$. Hence, in this case the minimum order of Gaussian integration for stability is 2×2.

A similar occurrence of zero energy modes can be detected in 8-node isoparametric quadrilateral elements with 2×2 Gaussian quadrature. Here, the deformation mode

$$[\mathbf{u}_h] = \begin{bmatrix}\alpha\xi(\eta^2-\frac{1}{3})\\ -\alpha\eta(\xi^2-\frac{1}{3})\end{bmatrix}\,, \quad \langle\boldsymbol{\epsilon}(\mathbf{u}_h)\rangle = \begin{bmatrix}\alpha(\eta^2-\frac{1}{3})\\ -\alpha(\xi^2-\frac{1}{3})\\ 0\end{bmatrix}\,, \qquad (7.51)$$

with $\alpha>0$, is obviously associated with positive strain energy, see Figure 7.3. However, using 2×2 Gaussian quadrature the strain energy of this mode is approximated as

$$W \doteq \frac{\alpha^2}{2}\sum_{l=1}^{2}\sum_{m=1}^{2}\begin{bmatrix}\eta_m^2-\frac{1}{3} & -(\xi_l^2-\frac{1}{3}) & 0\end{bmatrix}\begin{bmatrix}\lambda+2\mu & \lambda & 0\\ \lambda & \lambda+2\mu & 0\\ 0 & 0 & \mu\end{bmatrix}\begin{bmatrix}\eta_m^2-\frac{1}{3}\\ -(\xi_l^2-\frac{1}{3})\\ 0\end{bmatrix} = 0\,, \qquad (7.52)$$

which means that the mode of equation (7.51) is reduced to zero energy under 2×2 Gaussian quadrature. In this problem, it is evident that the minimum order of integration for stability is 3×3.

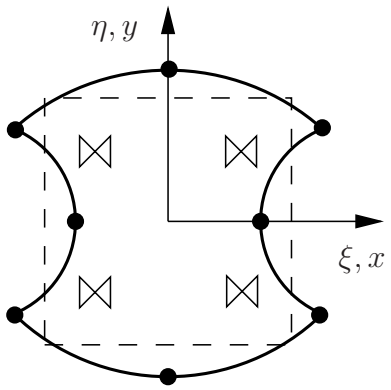


Figure 7.3. *Zero-energy modes for the 8-node quadrilateral with* 2×2 *Gaussian quadrature*

Zero energy modes are easily detected by an eigenvalue analysis of the element stiffness matrix $[\mathbf{K}^e]$. Without applying any boundary conditions, $[\mathbf{K}^e]$ should have as many zero eigenvalues as there are rigid-body modes, namely 3 (corresponding to two translations and one rotation) in two dimensions and 6 (corresponding to three translations and three rotations) in three dimensions. Any *additional* null eigenvectors are zero-energy modes due to lower than required order of integration.

In some cases, it is possible to detect the existence of zero energy modes by a simple counting procedure. For instance, referring again to the general 4-node isoparametric quadrilateral with 1×1 Gaussian quadrature, write its integrated stiffness directly as

$$[\mathbf{K}^e] \doteq 4J^e(0,0)[\mathbf{B}^e(0,0)]^T[\mathbf{D}][\mathbf{B}^e(0,0)] \; . \tag{7.53}$$

Since the dimension of this matrix is 8×8, its maximum rank is 3 (which is the rank of $[\mathbf{D}]$), and two-dimensional motions include only three rigid-body modes, it follows that this matrix has at least two zero-energy modes. Likewise, for the 8-node isoparametric quadrilateral with 2×2 Gaussian quadrature, the stiffness matrix is given by

$$[\mathbf{K}^e] \doteq \sum_{l=1}^{2}\sum_{m=1}^{2} J^e(\xi_l,\eta_m)[\mathbf{B}^e(\xi_l,\eta_m)]^T[\mathbf{D}][\mathbf{B}^e(\xi_l,\eta_m)] \; . \tag{7.54}$$

Since the maximum rank of this 16×16 matrix is $4 \times 3 = 12$ and there exist exactly three rigid-body modes, it follows that there is at least one zero-energy mode, which is precisely the one depicted in Figure 7.3.

Zero-energy modes are often suppressed by the actual deformation of the body, that is, they are rendered *non-communicable*. However, it is generally important to integrate the stiffness matrix by the minimum order of integration for stability to eliminate such modes altogether.

7.2.3 The patch test

A simple test of completeness for finite element approximations was originally proposed in 1965 by B. Irons. The idea is that any patch of elements should be unconditionally able to reproduce a constant field of the p-th derivative of the dependent variable when subjected to appropriate boundary conditions, where p is the order of the highest derivative of this variable in the weak form. Irons argued somewhat heuristically that the above requirement is necessary to guarantee that the error in approximating the p-th derivative of the dependent

variable is at most of order $o(h)$, where h is a measure of mesh size. Under mesh refinement (that is, as h approaches zero), this produces a sequence of solutions that converge to the exact solution. In the engineering literature, this test is referred to as the *patch test*. Since its inception, the patch test has been subjected to the scrutiny of engineers and applied mathematicians alike. Some have attempted to mathematically formalize and validate it while others have sought to discredit and dismiss it. Today, satisfaction of the patch test is widely considered as a good indicator of convergence of finite element approximations.

By way of background, consider an elliptic linear differential equation described operationally as $A[u] = f$ and let p be the order of the highest derivative in the weak counterpart of this equation. Three separate forms of the patch test that feature an increasing degree of severity are identified as follows:

Form A (full nodal restraint)

The values of all dependent variables in the finite element approximation are prescribed at every node according to a specified global polynomial field u_h of degree less than or

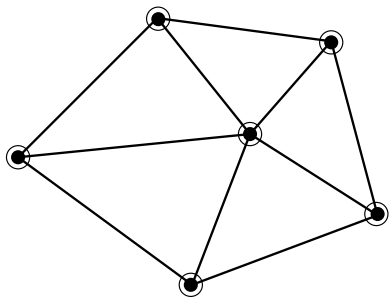


Figure 7.4. *Schematic of the patch test (Form A)*

equal to p which satisfies $A[u_h] = f$, see Figure 7.4. Since all the degrees of freedom are prescribed, the finite element solution of this problem consists of merely evaluating the "forces" corresponding to u_h. The test is designed to provide a comparison of $A[u_h]$ (which is directly available, since A and u_h are given) with $A_h[u_h]$, where A_h is the finite element counterpart of A[4]. Operators A and A_h should be identical, when applied to the given polynomial u_h of degree less or equal to p, see Section 5.3.

[4]This means that, in general, the discrete operator A_h emanating from a weighted-residual approximation satisfies $A_h[u_h] = f$ in a weak sense.

Form B (full boundary restraint)

The values of all dependent variables are prescribed at the boundary of the domain, according to an arbitrarily chosen global polynomial field u_h of degree less or equal to p, which satisfies the homogeneous counterpart of $A[u] = f$, see Figure 7.5. In this

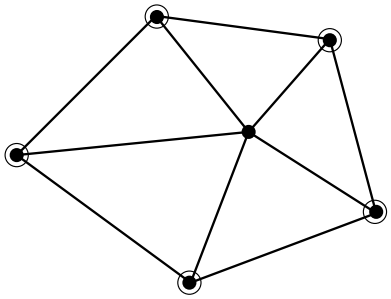


Figure 7.5. *Schematic of the patch test (Form B)*

test, all interior degrees of freedom are to be determined. Subsequently, the solution to the discrete problem is compared to u_h. The finite element solution should coincide with u_h throughout the domain. The above test is designed to check that the inverse operators A^{-1} and A_h^{-1} coincide when applied on a "force" field resulting from the polynomial field u_h prescribed on the boundary.

Form C (minimum restraint)

In this test, an arbitrary finite element patch is restrained by the minimum boundary conditions required to make the problem well-posed by suppressing all global singularities of the boundary-value problem, and is subjected to Neumann boundary conditions that, whenever possible, yield an exact polynomial solution of degree up to p, see Figure 7.6. The finite element solution is also expected to yield the exact answer. This test can detect potential singularities in the stiffness matrix, and provides a measure of the overall robustness of the finite element approximation.

The above forms of the patch test are employed routinely when examining the completeness of a given finite element formulation. They also form a systematic set of tests for a finite element implementation.

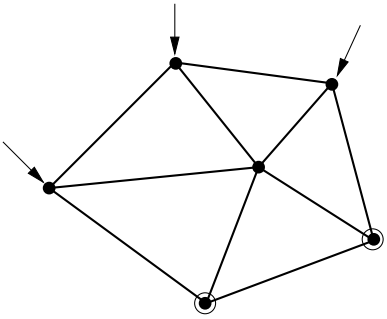


Figure 7.6. *Schematic of the patch test (Form C)*

7.3 Best approximation property of the finite element method

Consider a weak form according to which one needs to find $u \in \mathcal{U}$, such that

$$B(w,u) \;=\; (w,f)\;, \tag{7.55}$$

for all $w \in \mathcal{W}$, where $B(\cdot,\cdot)$ is a symmetric bilinear form and $(\cdot,f)$ is a linear form. The bilinear form B is termed V*-elliptic* (or *bounded from below*) if there exists a constant $\alpha > 0$, such that

$$B(u,u) \;\geq\; \alpha\|u\|^2\;, \tag{7.56}$$

where $\|\cdot\|$ is the norm associated with the inner product of $\mathcal{U}$. In a discrete setting, the above condition translates to

$$[\mathbf{u}]^T[\mathbf{K}][\mathbf{u}] \;\geq\; \alpha[\mathbf{u}]^T[\mathbf{u}]\;, \tag{7.57}$$

where $[\mathbf{u}]$ is any vector in $\mathbb{R}^n$ and $[\mathbf{K}]$ is a matrix in $\mathbb{R}^n \times \mathbb{R}^n$. In the latter case, it is immediately evident that boundedness from below implies positive-definiteness of $[\mathbf{K}]$.

The existence and uniqueness of the solution to (7.55) is guaranteed by the *Lax-Milgram theorem*. This states that the solution to the problem (7.55) exists and is unique if the bilinear form $B(\cdot,\cdot)$ is continuous and V-elliptic and the linear form $(\cdot,f)$ is continuous on the Hilbert space $\mathcal{U}$.

In the case of the two-dimensional Laplace-Poisson equation (3.5) in a domain Ω, where $\mathcal{U} = \{u \in H^1(\Omega) \mid u = 0 \text{ on } \Gamma_u\}$ and $\Gamma_u \neq \emptyset$, V-ellipticity of B translates to the existence of a positive α, such that

$$\int_\Omega \left(\frac{\partial u}{\partial x_1} k \frac{\partial u}{\partial x_1} + \frac{\partial u}{\partial x_2} k \frac{\partial u}{\partial x_2}\right) d\Omega \;\geq\; \alpha \int_\Omega \left[u^2 + \left(\frac{\partial u}{\partial x_1}\right)^2 + \left(\frac{\partial u}{\partial x_2}\right)^2\right] d\Omega\;. \tag{7.58}$$

The preceding result can be proved by appealing to the celebrated *Poincaré inequality*, according to which there exists a constant $c > 0$, such that

$$\int_\Omega u^2 \, d\Omega \;\leq\; c \int_\Omega \left[\left(\frac{\partial u}{\partial x_1} \right)^2 + \left(\frac{\partial u}{\partial x_2} \right)^2 \right] d\Omega \;, \tag{7.59}$$

for all $u \in \mathcal{U}$. This important result holds for regular domains Ω and effectively stipulates that the L_2-norm of a function $u \in \mathcal{U}$ is bounded from above by the L_2-norm of its derivatives. A proof of Poincaré's inequality is outside the scope of these notes.

Taking into account (7.59), and assuming without loss of generality that $k > 0$ is constant, if follows that

$$\frac{c+1}{k} \int_\Omega \left[k \left(\frac{\partial u}{\partial x_1} \right)^2 + k \left(\frac{\partial u}{\partial x_2} \right)^2 \right] d\Omega \;\geq\; \int_\Omega \left[u^2 + \left(\frac{\partial u}{\partial x_1} \right)^2 + \left(\frac{\partial u}{\partial x_2} \right)^2 \right] d\Omega \;, \tag{7.60}$$

hence $\alpha = \dfrac{k}{c+1}$.

In the case of linear elastostatics in a domain Ω, where now the space of admissible displacements is defined as $\mathcal{U} = \{\mathbf{u} \in \mathbf{H}^1(\Omega) \mid \mathbf{u} = \mathbf{0} \text{ on } \Gamma_u\}$ and, again, $\Gamma_u \neq \emptyset$. Here, V-ellipticity can be directly proved by means of *Korn's inequality*, which states that there exists a constant $c > 0$, such that

$$\int_\Omega \langle \boldsymbol{\epsilon} \rangle^T \langle \boldsymbol{\epsilon} \rangle \, d\Omega \;\geq\; c \|\mathbf{u}\|^2_{\mathbf{H}^1(\Omega)} \;, \tag{7.61}$$

assuming that $\lambda > 0$ and $\mu > 0$. Again, the proof of this important inequality is quite technical and, therefore, omitted here.

When the bilinear form $B(\cdot, \cdot)$ is V-elliptic, it is easy to see that it induces an inner product $\langle \cdot, \cdot \rangle_E$ on $\mathcal{U} \times \mathcal{U}$, that is $\langle u, v \rangle_E = B(u, v)$. Indeed, $B(\cdot, \cdot)$ is bilinear, symmetric, and

$$B(u, u) \;\geq\; \alpha \|u\|^2 \;\geq\; 0 \tag{7.62}$$

and $B(u, u) = 0 \Leftrightarrow \|u\| = 0 \Leftrightarrow u = 0$. The natural norm $\| \cdot \|_E$ associated with this inner product is defined by

$$\|u\|_E \;=\; \langle u, u \rangle_E^{1/2} \;=\; [B(u, u)]^{1/2} \;, \tag{7.63}$$

for any $u \in \mathcal{U} = \{u \in H^1(\Omega) \mid u = 0 \text{ on } \Gamma_u\}$. This is widely referred to as the *energy norm*, due to its physical interpretation of its square as being equal to twice the strain energy in the case of linear elastostatics, where, according to $(7.17)_1$ and (7.24),

$$B(\mathbf{u}, \mathbf{u}) \;=\; \|\mathbf{u}\|^2_E \;=\; \int_\Omega \langle \boldsymbol{\epsilon}(\mathbf{u}) \rangle^T [\mathbf{D}] \langle \boldsymbol{\epsilon}(\mathbf{u}) \rangle \, d\Omega \;. \tag{7.64}$$

Returning to (7.55), write its discrete Bubnov-Galerkin counterpart as: find $u_h \in \mathcal{U}_h \subset \mathcal{U}$, such that

$$B(w_h, u_h) \;=\; (w_h, f) \;, \tag{7.65}$$

for all $w_h \in \mathcal{W}_h \subset \mathcal{W}$. Since $\mathcal{W}_h \subset \mathcal{W}$, one may start again from (7.55) and also write

$$B(w_h, u) \;=\; (w_h, f) \;, \tag{7.66}$$

for all $w_h \in \mathcal{W}_h$. Subtracting (7.65) from (7.66), it follows that

$$B(w_h, u - u_h) \;=\; 0 \;, \tag{7.67}$$

for all $w_h \in \mathcal{W}_h$. This is a fundamental orthogonality condition, which states that the error $u - u_h$ is orthogonal to all the weighting functions $w_h \in \mathcal{W}_h$ with respect to the inner product induced by $B(\cdot, \cdot)$.

Now, given the exact solution u to (7.55), proceed to determine the function $\tilde{u} \in \mathcal{U}_h$ which minimizes the energy norm of the difference $u - \tilde{u}$ over all elements of $\mathcal{U}_h$. Clearly, for such a function $\tilde{u}$,

$$\delta B(u - \tilde{u}, u - \tilde{u}) \;=\; 0 \tag{7.68}$$

for any $\delta\tilde{u} \in \mathcal{W}_h$. Writing $\delta u = w_h$ and exploiting the linearity and symmetry of B, the preceding equation may be expressed as

$$B(w_h, u - \tilde{u}) \;=\; 0 \;, \tag{7.69}$$

for all $w_h \in \mathcal{W}_h$. Comparing (7.69) to (7.67), it is immediately seen that $\tilde{u} = u_h$. Since the second variation of the energy norm of $u - \tilde{u}$ is

$$\delta^2 B(u - \tilde{u}, u - \tilde{u}) \;=\; 2B(\delta\tilde{u}, \delta\tilde{u}) \;\geq\; \alpha \|\delta\tilde{u}\|^2 \;\geq\; 0 \;, \tag{7.70}$$

owing to the V-ellipticity of B, it is concluded that the finite element solution u_h minimizes the error in the energy norm over all functions in $\mathcal{U}_h$ and, in this sense, it constitutes the *best approximation* to the exact solution u, see Figure 7.7 for a geometric interpretation.

An important corollary of the orthogonality condition (7.67) is noted here: let u be the solution to an elliptic problem of the type (7.55) and u_h be the Bubnov-Galerkin approximation to this solution. It follows that

$$B(u, u) \;=\; B(u_h + e, u_h + e) \;=\; B(u_h, u_h) + B(e, e) + 2B(u_h, e) \;, \tag{7.71}$$

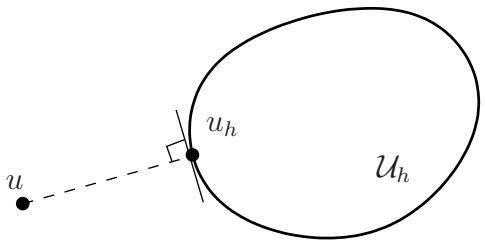


Figure 7.7. *Geometric interpretation of the best approximation property as a closest-point projection from u to $\mathcal{U}_h$ in the sense of the energy norm*

where $e = u - u_h$ is the error in the approximation. Assuming (without loss of generality) that $\mathcal{U}_h = \mathcal{W}_h$, the orthogonality condition (7.67) implies that $B(u_h, e) = 0$, hence

$$B(u,u) \;=\; B(u_h,u_h) + B(e,e) \;\geq\; B(u_h,u_h) \;, \tag{7.72}$$

since $B(e,e) \geq 0$ due to the V-ellipticity of B. The inequality (7.72) shows that the energy of the exact solution is underestimated by the finite element approximation. This is an important property that holds true in all Bubnov-Galerkin formulations of elliptic problems.

7.4 Error sources and estimates

Any finite element solution contains errors due to several sources. These include:

(a) Error in the discretization of the domain (*first fundamental error*)
Such errors are associated with the fact that

$$\Omega_h \;\doteq\; \Omega \qquad , \qquad \partial\Omega_h \;\doteq\; \partial\Omega \;. \tag{7.73}$$

There exist some formal estimates for such errors, which will not be discussed here. The first fundamental error can be controlled by using finer meshes and/or higher-order elements.

(b) Error due to inexact numerical integration
These errors occur when integrating non-polynomial quantities using Gaussian quadrature or other inexact formulae, such that, e.g.,

$$\int_{\square} f(\xi,\eta,\zeta)\, d\xi d\eta d\zeta \;\doteq\; \sum_{k=1}^{L}\sum_{l=1}^{L}\sum_{m=1}^{L} w_k w_l w_m f(\xi_k,\eta_l,\zeta_m) \;, \tag{7.74}$$

where (ξ_k, η_l, ζ_m) are the sampling points and w_k, w_l, and w_m, are the associated weights.

The estimation of such errors is quite easy, given that the integration rules are polynomially accurate to a known degree, as discussed in Section 6.1. The error can be controlled by increasing the order of numerical integration.

(c) Error in the solution of linear algebraic systems

Such errors are associated with the spectral properties of the global finite element stiffness matrix $[\mathbf{K}]$. The accuracy of a direct or iterative solution is generally dependent on the conditioning of $[\mathbf{K}]$, which, in turn, is defined by the *condition number* κ as

$$\kappa \;=\; \|[\mathbf{K}]\|\|[\mathbf{K}^{-1}]\| \;, \tag{7.75}$$

where $\|\cdot\|$ is any matrix norm. If, in particular, the matrix norm is taken to be the *spectral norm*, defined as

$$\|[\mathbf{K}]\| \;=\; \max\{\sqrt{\rho} \quad | \quad \rho : \text{ eigenvalue of } [\mathbf{K}]^T[\mathbf{K}]\} \;, \tag{7.76}$$

then the condition number of equation (7.75) for a symmetric $[\mathbf{K}]$ takes the particular form

$$\kappa \;=\; \left|\frac{\rho_{max}}{\rho_{min}}\right| \;, \tag{7.77}$$

where ρ_{max}, ρ_{min} are the maximum and minimum eigenvalues of $[\mathbf{K}]$, respectively. The higher the condition number, the less accurate the solution of the linear algebraic system.

(d) Other floating-point related errors

These are related to round-off in cases other than the solution of algebraic systems.

(e) Errors in the finite element approximation

These errors are due to the fact that the finite element solution is sought over a subset $\mathcal{U}_h$ of the space of admissible functions $\mathcal{U}$, and, in general, the exact solution u lies in $\mathcal{U} \setminus \mathcal{U}_h$.

A simple error estimate of this class may be obtained by starting with the orthogonality

condition (7.67) and writing

$$\begin{aligned} B(u-u_h, u-u_h) &= B(u-u_h, u-u_h) + B(u-u_h, w_h) \\ &= B(u-u_h, u-u_h+w_h) \\ &= B(u-u_h, u-v) \ , \end{aligned} \tag{7.78}$$

where v is an arbitrary element of $\mathcal{U}_h$ written as $v = u_h - w_h$. Recalling that B is assumed continuous in both of its arguments, it follows that there is a constant $M > 0$, such that

$$B(u-u_h, u-v) \ \leq \ M\|u-u_h\|\|u-v\| \ , \tag{7.79}$$

for all $u \in \mathcal{U}$, and u_h, $v \in \mathcal{U}_h$. Furthermore, taking into account (7.78) and (7.79), the V-ellipticity condition (7.56) leads to

$$\alpha\|u-u_h\|^2 \ \leq \ M\|u-u_h\|\|u-v\| \ , \tag{7.80}$$

which, in turn, implies that

$$\|u-u_h\| \ \leq \ \frac{M}{\alpha}\|u-v\| \ , \tag{7.81}$$

for all $v \in \mathcal{U}_h$. The inequality (7.81) is referred to as *Céa's lemma*. This states that the finite element solution u_h yields to within the mesh-independent constant $\frac{M}{\alpha}$ the best approximation of the exact solution u in the sense of the energy norm over any potential solution v in $\mathcal{U}_h$.

A corresponding inequality may be written using the energy norm. To this end, start by recalling the definition of the energy norm in (7.63), which leads to recasting (7.78) in the form

$$\|u-u_h\|_E^2 \ = \ B(u-u_h, u-v) \ = \ \langle u-u_h, u-v\rangle_E \ . \tag{7.82}$$

It follows from (7.82) and the Cauchy-Schwartz inequality (2.12) that Céa's lemma may be expressed in the energy norm as

$$\|u-u_h\|_E \ \leq \ \|u-v\|_E \ . \tag{7.83}$$

This is yet another statement of the previously discussed best approximation property.

The error estimates (7.81) and (7.83) bound the error from above by the difference between the exact solution u and any other element v of $\mathcal{U}_h$. Although interesting in

their own right, these results are of limited practical significance, as they involve the (unknown) exact solution on both sides of the inequality. However, these results may be used as starting points to deduce practical error estimates. One such error estimate applicable to the case of h-adaptivity can be derived for elliptic problems in the form

$$\|u - u_h\|_E \;\leq\; C_1 h^{q-p+1} |u|_{q+1} \;, \tag{7.84}$$

where

$$|u|^2_{q+1} \;=\; \sum_{\alpha_1+\alpha_2+\ldots+\alpha_n=q+1} \int_\Omega \left| \frac{\partial^{q+1} u}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \ldots \partial x_n^{\alpha_n}} \right|^2 d\Omega \;. \tag{7.85}$$

Here, h is a measure of the mesh size, p is the order of highest derivative in the weak form, q is the polynomial degree of completeness of $\mathcal{U}_h$, and $C_1 > 0$ is a positive constant that is independent of h. Also, the term $|u|_{q+1}$ is a measure of smoothness of the exact solution. The error estimate (7.85) implies that the corresponding finite element method convergences as h^{q-p+1}, that is, at a rate $q - p + 1$ in h.

In the case of p-adaptivity, a typical error estimate is of the form

$$\|u - u_h\|_E \;\leq\; C_2 q^{-r} \sum_{i=q+1}^{r+1} |u|_i \;, \tag{7.86}$$

where $r \geq q$ and $C_2 (> 0)$ is a constant independent of q. A finite element method that is subject to the error estimate (7.86) converges as q^{-r}.

The error estimates (7.84) and (7.86) are of practical use because they establish the rate of convergence of the finite element approximation under mesh refinement (that is, when $h \mapsto 0$), or under increase of the degree of polynomial completeness (that is, when $q \mapsto \infty$), respectively. Although, again, they contain on the right-hand side a term that depends on the exact solution, this does not limit their usefulness, because knowledge of the exact solution is not needed to establish the rate of convergence.

7.5 Application to incompressible elastostatics and Stokes' flow

The presence of constraints introduces challenges in the finite element formulation and solution of partial differential equations. A classical example is encountered when assuming

that a linearly elastic material is incompressible. Preliminary to the introduction of the incompressibility constraint, decompose the strain and stress tensor additively as

$$\boldsymbol{\epsilon} \;=\; \mathbf{e} + \frac{1}{3}(\mathrm{tr}\,\boldsymbol{\epsilon})\mathbf{I} \qquad , \qquad \boldsymbol{\sigma} \;=\; \mathbf{s} + \frac{1}{3}(\mathrm{tr}\,\boldsymbol{\sigma})\mathbf{I}\;, \tag{7.87}$$

where $\mathbf{e}$ and $\mathbf{s}$ are the *deviatoric* strain and stress, respectively. It follows from (7.87) that the deviatoric strain and stress are traceless, namely that $\mathrm{tr}\,\mathbf{e} = 0$ and $\mathrm{tr}\,\mathbf{s} = 0$. Also, the *volumetric* strain θ and the *pressure* p are defined as

$$\theta \;=\; \mathrm{tr}\,\boldsymbol{\epsilon} \;=\; \nabla\cdot\mathbf{u} \qquad , \qquad p \;=\; \frac{1}{3}\,\mathrm{tr}\,\boldsymbol{\sigma}\;. \tag{7.88}$$

As its name indicates, the volumetric strain θ measures the change of volume undergone by the material under the influence of the stresses. Indeed, denoting by dV an infinitesimal material volume element before the deformation and dv the same material volume element after the deformation, it is clear from the definition of strain in (7.7) that

$$dv \;=\; (1+\epsilon_{11})(1+\epsilon_{22})(1+\epsilon_{33})dV \tag{7.89}$$

or, upon ignoring higher-order terms,

$$\frac{dv}{dV} \;\doteq\; 1+\epsilon_{11}+\epsilon_{22}+\epsilon_{33} \;=\; 1+\theta\;. \tag{7.90}$$

Taking into account (7.87) and (7.88), the isotropic stress-strain relation (7.5) can be rewritten as

$$\mathbf{s} \;=\; 2\mu\mathbf{e} \qquad , \qquad p \;=(\lambda+\frac{2}{3}\mu)\theta \;=\; K\theta\;, \tag{7.91}$$

where K is the *bulk modulus*. As seen from (7.91), the original isotropic stress-strain relation (7.5) can be decomposed into two stress-strain relations which associate the deviatoric and volumetric stresses to the corresponding strains.

In examining the problem of incompressible elasticity, one may distinguish between the *nearly incompressible* and the *exact incompressible* cases. Noting that the bulk modulus is related to the Young modulus E and the Poisson's ratio ν as $K = \dfrac{E}{3(1-2\nu)}$, the former case corresponds to ν approaching (but not reaching) the limiting value $\nu = 0.5$, while the latter to $\nu = 0.5$. In the former case, the constitutive equation $(7.91)_2$ applies and the pressure is computed from it. In the latter case, $(7.91)_2$ ceases to apply and the pressure becomes indeterminate from it as $K \to \infty$ and $\theta \to 0$. In this case, it turns out that the pressure becomes a Lagrange multiplier which may be determined by enforcing the constraint of incompressibility $\theta = 0$.

The strong form of the exact incompressible problem of linear elastostatics is defined as

$$\begin{aligned} \nabla \cdot \boldsymbol{\sigma} + \mathbf{f} &= \mathbf{0} \quad \text{in } \Omega \ , \\ \boldsymbol{\sigma}\mathbf{n} &= \bar{\mathbf{t}} \quad \text{on } \Gamma_q \ , \\ \mathbf{u} &= \bar{\mathbf{u}} \quad \text{on } \Gamma_u \ , \\ \nabla \cdot \mathbf{u} &= 0 \quad \text{in } \Omega \ , \end{aligned} \tag{7.92}$$

where now the stress tensor $\boldsymbol{\sigma}$ is given by

$$\boldsymbol{\sigma} \ = \ p\mathbf{I} + 2\mu\boldsymbol{\epsilon} \ = \ p\mathbf{I} + 2\mu\mathbf{e} \ . \tag{7.93}$$

In this strong form, the unknown quantities are the displacement $\mathbf{u}$ and the pressure p. This is in contrast to the strong form in (7.2), where the only unknown is the displacement $\mathbf{u}$, while the pressure p is determined by the constitutive equation $(7.91)_2$.

The strong form (7.92) is identical to the one governing the problem of steady incompressible creeping Newtonian viscous flow (also referred to frequently as *Stokes' flow*). In this case, $\mathbf{u}$ represents the velocity and μ the dynamic viscosity of the fluid.

The weak form of the preceding boundary-value problem can be obtained by starting from the general weighted-residual statement

$$\int_\Omega \mathbf{w} \cdot (-\nabla \cdot \boldsymbol{\sigma} - \mathbf{f}) \, d\Omega + \int_{\Gamma_q} \mathbf{w} \cdot (\boldsymbol{\sigma}\mathbf{n} - \bar{\mathbf{t}}) \, d\Gamma + \int_\Omega q \nabla \cdot \mathbf{u} \, d\Omega \ = \ 0 \ , \tag{7.94}$$

where the weighting functions $(\mathbf{w}, q)$ belong to $\mathcal{W} \times \mathcal{Q}$. Note that the last term on the left-hand side of (7.94) can be merely added to the original weak form because the scalar weighting function q is arbitrary and independent of the vector weighting function $\mathbf{w}$. Upon following the standard process of employing integration by parts and the divergence theorem as in Section 7.5, equation (7.94), transforms into

$$\int_\Omega \nabla_s \mathbf{w} : \boldsymbol{\sigma} \, d\Omega + \int_\Omega q \nabla \cdot \mathbf{u} \, d\Omega \ = \ \int_\Omega \mathbf{w} \cdot \mathbf{f} \, d\Omega + \int_{\Gamma_q} \mathbf{w} \cdot \bar{\mathbf{t}} \, d\Gamma \ . \tag{7.95}$$

Recalling (7.93), one may write

$$\nabla_s \mathbf{w} : \boldsymbol{\sigma} \ = \ \boldsymbol{\epsilon}(\mathbf{w}) : (2\mu\boldsymbol{\epsilon} + p\mathbf{I}) \ = \ \boldsymbol{\epsilon}(\mathbf{w}) : 2\mu\boldsymbol{\epsilon}(\mathbf{u}) + p\nabla \cdot \mathbf{w} \ = \ \boldsymbol{\epsilon}(\mathbf{w}) : 2\mu\boldsymbol{\epsilon}(\mathbf{u}) + p\nabla \cdot \mathbf{w} \ . \tag{7.96}$$

Hence, the weak form (7.95) can be expressed equivalently as

$$\int_\Omega \nabla_s \mathbf{w} : 2\mu\nabla_s \mathbf{u} \, d\Omega + \int_\Omega p\nabla \cdot \mathbf{w} \, d\Omega + \int_\Omega q \nabla \cdot \mathbf{u} \, d\Omega \ = \ \int_\Omega \mathbf{w} \cdot \mathbf{f} \, d\Omega + \int_{\Gamma_q} \mathbf{w} \cdot \bar{\mathbf{t}} \, d\Gamma \ , \tag{7.97}$$

or, resorting to vector representation,

$$\int_\Omega \langle \boldsymbol{\epsilon}(\mathbf{w})\rangle^T 2\mu \langle \boldsymbol{\epsilon}(\mathbf{u})\rangle \, d\Omega + \int_\Omega p\nabla\cdot\mathbf{w}\, d\Omega + \int_\Omega q\nabla\cdot\mathbf{u}\, d\Omega \;=\; \int_\Omega [\mathbf{w}]^T[\mathbf{f}]\, d\Omega + \int_{\Gamma_q} [\mathbf{w}]^T[\bar{\mathbf{t}}]\, d\Gamma \;. \quad (7.98)$$

The weighted-residual problem amounts to finding $(\mathbf{u}, p) \in \mathcal{U}\times\mathcal{P}$, such that (7.98) hold for all $(\mathbf{w}, q) \in \mathcal{W}\times\mathcal{Q}$. The spaces of admissible displacements and associated weighting functions are

$$\begin{aligned} \mathcal{U} &= \left\{ \mathbf{u}\in H^1(\Omega) \quad | \quad \mathbf{u} = \bar{\mathbf{u}} \text{ on } \Gamma_u \right\} \;, \\ \mathcal{W} &= \left\{ \mathbf{w}\in H^1(\Omega) \quad | \quad \mathbf{w} = \mathbf{0} \text{ on } \Gamma_u \right\} \;, \end{aligned}$$

whereas the spaces of admissible pressures and associated weighting functions are

$$\mathcal{P} \;=\; \mathcal{Q} \;=\; \left\{ p\in H^0(\Omega) \right\} \;. \quad (7.99)$$

The second and third terms on the left-hand side of (7.97) (or, equivalently, (7.98)) justify the characterization of the pressure p as a Lagrange multiplier enforcing the constraint of incompressibility.

In the special case when all boundary conditions are of Dirichlet type, the pressure field p is indeterminate to within an additive constant. This is because, if $\overline{p}$ is a constant over the domain Ω, then integration by parts and the divergence theorem imply that

$$\int_\Omega (p+\overline{p})\nabla\cdot\mathbf{w}\, d\Omega \;=\; \int_{\partial\Omega} (p+\overline{p})\mathbf{w}\cdot\mathbf{n}\, d\Gamma - \int_\Omega \nabla(p+\overline{p})\cdot\mathbf{w}\, d\Omega \;=\; -\int_\Omega \nabla p\cdot\mathbf{w}\, d\Omega \;=\; \int_\Omega p\nabla\cdot\mathbf{w}\, d\Omega \,, \quad (7.100)$$

since now $\mathbf{w} = \mathbf{0}$ on $\partial\Omega$. To eliminate this indeterminacy in the Dirichlet problem, one may redefine $\mathcal{P}$ as

$$\mathcal{P} \;=\; \mathcal{Q} \;=\; \left\{ p\in H^0(\Omega) \quad | \quad \int_\Omega p\, d\Omega = 0 \right\} \;. \quad (7.101)$$

Equation (7.98) can be written in operational form as

$$\begin{aligned} B(\mathbf{w},\mathbf{u}) + C(\mathbf{w},p) &= (\mathbf{w},\mathbf{f}) \\ C(\mathbf{u},q) &= 0 \;, \end{aligned} \quad (7.102)$$

where

$$B(\mathbf{w},\mathbf{u}) \;=\; \int_\Omega \langle \epsilon(\mathbf{w})\rangle^T 2\mu \langle \epsilon(\mathbf{u})\rangle \, d\Omega \quad (7.103)$$

and

$$C(\mathbf{w},p) \;=\; \int_\Omega p\nabla\cdot\mathbf{w}\, d\Omega \;. \quad (7.104)$$

Note that the spaces $\mathcal{U}$ and $\mathcal{W}$ are identical to those of the unconstrained elastostatics problem in Section 7.2. This observation has important ramifications in the finite element approximation of the incompressible elastostatics problem. Alternatively, one may choose to incorporate the constraint $(7.92)_4$ directly into the space of admissible displacements $\mathcal{U}_c$, namely define

$$\mathcal{U}_c \;=\; \left\{ \mathbf{u} \in H^1(\Omega) \quad | \quad \mathbf{u} = \bar{\mathbf{u}} \text{ on } \Gamma_u \quad , \quad \nabla \cdot \mathbf{u} = 0 \text{ in } \Omega \right\} \; . \tag{7.105}$$

It turns out that constructing discrete counterparts of $\mathcal{U}_c$ for the purpose of obtaining finite element solutions leads to so-called *primal* approximations methods, which are generally quite cumbersome, hence rarely used in practice. The alternative of employing the Lagrange multiplier formulation in connection with the weak form (7.98) leads to *dual* methods, which are generally simpler to implement.

The weak form (7.102) constitutes the basis for a finite element approximation of the constrained problem. Indeed, such an approximation amounts to defining discrete admissible fields $\mathcal{U}_h \subset \mathcal{U}$, $\mathcal{W}_h \subset \mathcal{W}$ and $\mathcal{P}_h \subset \mathcal{P}$ and seeking a solution to (7.102) within these fields. The discrete problem leads to a global system of algebraic equations of the form

$$\left[\begin{array}{cc} [K_{uu}] & [K_{up}] \\ [K_{pu}]^T & [0] \end{array} \right] \left[\begin{array}{c} [u] \\ [p] \end{array} \right] \;=\; \left[\begin{array}{c} [F] \\ [0] \end{array} \right] \; , \tag{7.106}$$

where $[u]$ and $[p]$ are the displacement and pressure degrees of freedom. Recalling the structure of (7.102), it is immediately seen that the global stiffness matrix is symmetric for the Bubnov-Galerkin case. Further, it is seen that the global stiffness matrix contains zeros on its major diagonal, which implies that pivoting may be required when solving the system using Gauss elimination. Finally, it is clear that the constrained problem requires the solution of additional equations (those corresponding to the pressure degrees of freedom) as compared to the unconstrained problem.

The choice of finite element subspaces for the approximation of constrained problems within the Lagrange multiplier formulation is not as straightforward as in the unconstrained problem. The following example illustrates a fundamental difficulty: consider the deformation of an incompressible isotropic linearly elastic solid in plane strain and assume that it is modeled using 3-node triangular elements with linear displacement $\mathbf{u}_h$ and constant pressure p_h in each element, see Figure 7.8. The constraint of incompressibility can be expressed at the element level as

$$\int_{\Omega^e} q_h \nabla \cdot \mathbf{u}_h \, d\Omega \;=\; 0 \; , \tag{7.107}$$

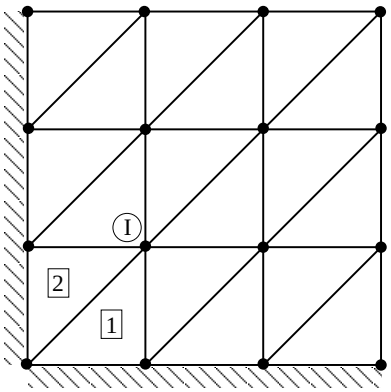


Figure 7.8. *Illustration of volumetric locking in plane strain when using 3-node triangular elements*

or, since the pressure (hence also the weighting function q_h) is assumed piecewise constant,

$$\int_{\Omega^e} \nabla \cdot \mathbf{u}_h \, d\Omega \;=\; 0 \; . \tag{7.108}$$

Given that $\nabla \cdot \mathbf{u}_h$ represents change of volume (here area, since the problem is two-dimensional), it is readily concluded that the total area of each element e should remain constant. Referring to Figure 7.8, area conservation for element 1 implies that node I should only move horizontally. At the same time, area conservation of element 2 implies that node I should only move vertically. The preceding conditions can be satisfied simultaneously only if I stays fixed. The same analysis can be applied successively to the rest of the nodes, thus leading to the conclusion that the whole mesh is locked in place regardless of the external loading! This condition is referred to as *volumetric locking* and is a byproduct of a poor choice of admissible displacements and pressures.

Fortunately, there exist choices of admissible displacement and pressure fields that bypass the problem of volumetric locking and yield convergent finite element approximations. Moreover, there exists a well-established mathematical theory for assessing whether a given formulation is free of volumetric locking. The simplest two-dimensional element that is known to produce convergent solutions to the incompressible elastostatics/Stokes' flow problem is a 4-node quadrilateral with the usual bilinear displacement interpolation in the natural coordinates and constant elementwise pressure, see Figure 7.9.

The nearly incompressible case can be viewed as a *penalty regularization* of the exact incompressible case, in the sense that the constraint is enforced approximately and with increasing accuracy as the value of a *penalty parameter*, here the bulk modulus K, increases

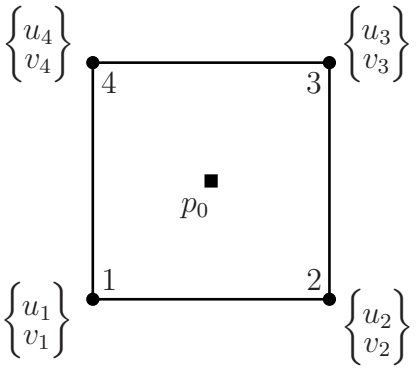


Figure 7.9. *The simplest convergent planar element for incompressible elastostatics/Stokes' flow*

to infinity. To understand the nearly incompressible case, recall that the total potential energy $I[\mathbf{u}]$ of equation (7.26) attains an absolute minimum at the equilibrium point, and write the strain energy with the aid of (7.91) as

$$W[\mathbf{u}] \;=\; \frac{1}{2}\int_\Omega \boldsymbol{\epsilon} : \boldsymbol{\sigma}\, d\Omega \;=\; \frac{1}{2}\int_\Omega \left[2\mu \mathbf{e} : \mathbf{e} + K\theta^2\right] d\Omega \;. \tag{7.109}$$

Clearly, as K increases toward infinity, θ needs to converge to zero for $I[\mathbf{u}]$ to attain an absolute minimum. Otherwise, $I[\mathbf{u}]$ would also explode to infinity, hence violating the Minimum Potential Energy Theorem. The near incompressible treatment is conceptually and implementationally simpler than the exact incompressible treatment, as it involves only displacement degrees of freedom. Its drawbacks are that it satisfies the incompressibility constraint in an approximate fashion and it may lead to poor conditioning of the stiffness matrix with increasing values of K.

7.6 Suggestions for further reading

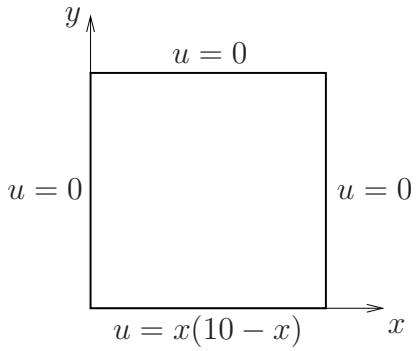
Section 7.2

[1] R.L. Taylor, J.C. Simo, O.C. Zienkiewicz, and A.C.H. Chan. The patch test – a condition for assessing FEM convergence. *Int. J. Num. Meth. Engr.*, 22:39–62, 1986. [*This article contains a detailed account of the patch test and its significance in finite element technology*]

7.7 Exercises

Problem 1

Write a finite element code to find the steady-state temperature distribution on a square rigid-body of side length $a = 10.0$, for which the boundary conditions are shown in the following figure. Assume that there is no heat supply per unit area.



Specifically, solve the problem using four uniform meshes with a total number of 4, 16, 64 and 256 4-node quadrilateral isoparametric elements using exact integration for all element arrays. Submit a copy of the code including any input files. Plot contours of the computed temperature distribution throughout the body. The exact solution of the problem can be found in series form as

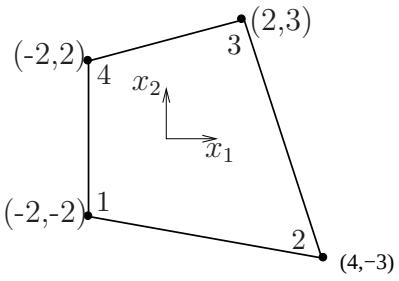
$$u(x,y) \;=\; \sum_{n=1}^{\infty} \frac{0.2}{\sinh n\pi} \Big\{ \int_0^{10} z(10-z) \sin\frac{n\pi z}{10}\, dz \Big\} \sin\frac{n\pi x}{10} \sinh\frac{n\pi(10-y)}{10} \;.$$

Compute the exact solution at the center of the body ($x = 5, y = 5$) and plot the error at this point as a function of the size h of the elements (use both decimal and log-log scale). Comment on the rate of convergence of the finite element solution.

Problem 2

Consider a 4-node isoparametric quadrilateral element in plane strain, with reference configuration as in the following figure. After a finite element analysis is conducted, the nodal displacements (u_1, u_2) of the element are found to be:

Node	u_1	u_2
1	0.005	-0.003
2	0.	0.002
3	0.004	0.
4	-0.005	0.001



Compute the normal strains $\epsilon_{11} = u_{1,1}$, $\epsilon_{22} = u_{2,2}$ and the engineering shear strain $\gamma_{12} = u_{1,2} + u_{2,1}$ at point P with coordinates $(x_1, x_2) = (1, 1)$. Also, compute all components of the stress tensor at point P, assuming that the material is isotropic linear elastic with $\lambda = 6 \times 10^5$ and $\mu = 4 \times 10^5$.

Problem 3

A 4-node isoparametric quadrilateral element Ω^e is used in the analysis of a linear elastic body in plane strain. Assuming that the stress field $\boldsymbol{\sigma}$ is constant over the element, determine the number of Gauss points required to *exactly* compute the integral

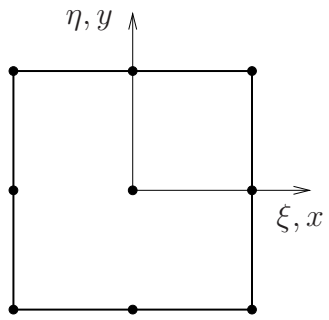
$$\boldsymbol{R}^e \;=\; \int_{\Omega^e} \boldsymbol{B}^{e\,T} \boldsymbol{\sigma}\, d\Omega$$

which emanates from the stress-divergence term

$$\boldsymbol{w}^{e\,T} \boldsymbol{R}^e = \int_{\Omega^e} \boldsymbol{\epsilon}^T(\boldsymbol{w}_h)\, \boldsymbol{\sigma}\, d\Omega \;.$$

Problem 4

Consider a 9-node isoparametric rectangular element, in which the coordinate systems of the natural and physical space coincide, as in the adjacent figure. Assuming that the element is used in modeling a linear elastic solid, whose material parameters remain constant within the element, determine the number of Gauss points per direction required to *exactly* integrate the element stiffness matrix. You do *not* have to actually compute the stiffness matrix.



Problem 5

Perform a spectral analysis of the element stiffness $\mathbf{K}^e$ for a 4-node rectangular element in plane strain with $\lambda = 20.0$ and $\mu = 10.0$ using the 2×2 Gaussian integration rule. You may assume that the element has dimensions 10×6 in the prescribed length unit. Subsequently, repeat your analysis using a 1×1 Gaussian integration rule. Do the eigenvalues change? Comment on the results.

In each case, plot the resulting eigenvectors versus the undeformed mesh and identify them according to the fundamental deformation mode that they represent. You may use MATLAB (or another programming language) for your calculations.

Problem 6

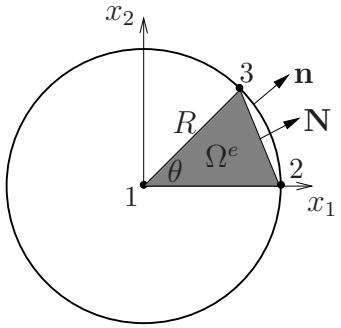
A long cylindrical body of radius R is subjected to boundary traction $\bar{\mathbf{t}} = p_0\, \mathbf{n}$, where p_0 is a constant and $\mathbf{n}$ is the outward unit normal to the boundary.

(a) Let the body be discretized using 3-node triangular elements. With reference to the following figure, compute the equivalent nodal forces on nodes 2 and 3 of a representative element Ω^e due to the prescribed traction, *assuming that p_0 is applied along the normal to the finite element boundary* with constant outward unit normal $\mathbf{N}$.

(b) Determine the exact resultant traction vector $\mathbf{F}$, defined as

$$\mathbf{F} \;=\; \int \bar{\mathbf{t}}\, ds \;,$$

which applies on *the actual circular boundary* of length $s = R\theta$. How does $\mathbf{F}$ compare to the sum of the equivalent nodal forces computed in part (a)?



Chapter 8

Parabolic Differential Equations

Parabolic partial differential equations involve time (or a time-like quantity) as an independent variable. Therefore, the resulting initial/boundary-value problems include two types of independent variables, namely, spatial variables (e.g., x_i, $i = 1, 2, 3$) and a temporal variable (t). In the context of the finite element method, there are two general approaches in dealing with the two types of variables. These are:

(a) Discretize the spatial variables independently from the temporal variable.

In this approach, the spatial discretization typically occurs first and yields a system of ordinary differential equations in time. These equations are subsequently integrated in time by means of some standard numerical integration method. This approach is referred to as *semi-discretization* and is used widely in engineering practice due to its conceptual simplicity and computational efficiency.

(b) Discretize spatial and temporal variables together.

Here, all independent variables are treated simultaneously, although the discretization is generally different for spatial and temporal variables. This approach yields *space-time finite elements*. Such elements are typically used for special problems, as they tend to be more complicated and expensive than those resulting from semi-discretization.

Figure 8.1 includes a schematic depiction of the two approaches.

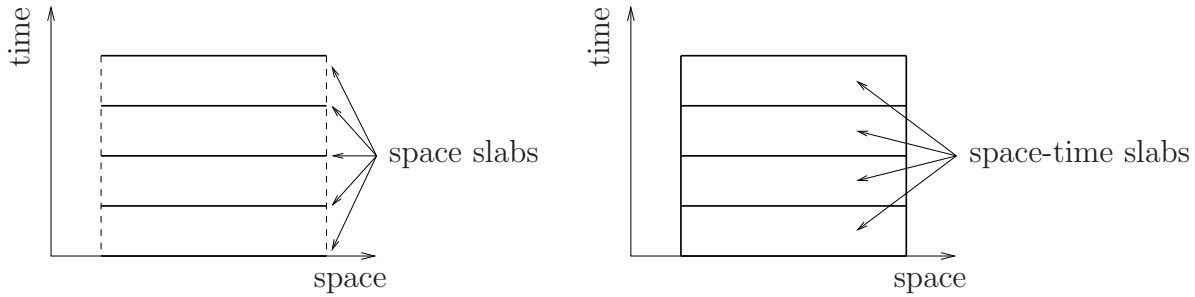


Figure 8.1. *Schematic depiction of semi-discretization (left) and space-time discretization (right)*

8.1 Standard semi-discretization methods

Consider the time-dependent version of the Laplace-Poisson equation in two dimensions. The initial/boundary-value problem takes the form

$$\begin{aligned} \frac{\partial}{\partial x_1}(k\,\frac{\partial u}{\partial x_1}) \;+\; \frac{\partial}{\partial x_2}(k\,\frac{\partial u}{\partial x_2}) - f \;&=\; \rho c \frac{\partial u}{\partial t} \quad \text{in } \Omega \times I \;, \\ -k\,\frac{\partial u}{\partial n} \;&=\; \bar{q} \quad \text{on } \Gamma_q \times I \;, \\ u \;&=\; \bar{u} \quad \text{on } \Gamma_u \times I \;, \\ u(x_1, x_2, 0) \;&=\; u_0(x_1, x_2) \quad \text{in } \Omega \;, \end{aligned} \tag{8.1}$$

where $u = u(x_1, x_2, t)$ is the (yet unknown) solution and $I = (0, T]$ is the time domain of the analysis, with T being a given end-time. Continuous functions $k = k(x_1, x_2)$, $\rho = \rho(x_1, x_2)$, $c = c(x_1, x_2)$ and $u_0 = u_0(x_1, x_2)$ are defined in Ω and a continuous time-dependent function $f = f(x_1, x_2, t)$ is defined in $\Omega \times I$. Further, continuous time-dependent functions $\bar{q} = \bar{q}(x_1, x_2, t)$ and $\bar{u} = \bar{u}(x_1, x_2, t)$ are defined on $\Gamma_u \times I$ and $\Gamma_q \times I$, respectively. Equations $(8.1)_2$ and $(8.1)_3$ are the *time-dependent Neumann* and *time-dependent Dirichlet* conditions, respectively. Finally, equation $(8.1)_4$ is the *initial condition* specified here at time $t = 0$. The strong form of the initial/boundary-value problem is stated as follows: given functions k, ρ, c, u_0, f, $\bar{q}$ and $\bar{u}$, find a function u that satisfies equations (8.1).

A Galerkin-based weighted-residual form of the above problem can be deduced from the counterpart of (3.7) for the initial/boundary-value problem (8.1) by assuming that: (i) the time-dependent Dirichlet boundary conditions are satisfied *a priori* by the choice of the space of admissible solutions $\mathcal{U}$, hence the weighting function w_u vanishes, that is, $w_u = 0$ on $\Gamma_u \times I$, (ii) the remaining weighting functions satisfy $w_\Omega = w$ in $\Omega \times I$, $w_q = w$ on $\Gamma_q \times I$,

(iii) $w = 0$ on $\Gamma_u \times I$, and (iv) the initial condition is satisfied *a priori* in Ω, hence it also enters the space of admissible solutions $\mathcal{U}$.

Taking into account the preceding assumptions, one may write a weighted-residual statement of the form

$$\int_{\Omega\times I} w \left[-\rho c\frac{\partial u}{\partial t} + \frac{\partial}{\partial x_1}\left(k\,\frac{\partial u}{\partial x_1}\right) + \frac{\partial}{\partial x_2}\left(k\,\frac{\partial u}{\partial x_2}\right) - f\right] d(\Omega\times I) - \int_{\Gamma_q\times I} w\left[k\,\frac{\partial u}{\partial n} + \bar{q}\right] d(\Gamma\times I) \;=\; 0\,. \quad (8.2)$$

Clearly, equation (8.2) involves a space-time integral. Since the spatial and temporal dimensions are independent of each other, the corresponding integrals may be readily decoupled, so that (8.2) is rewritten as

$$\int_I\int_\Omega w \left[-\rho c\frac{\partial u}{\partial t} + \frac{\partial}{\partial x_1}\left(k\,\frac{\partial u}{\partial x_1}\right) + \frac{\partial}{\partial x_2}\left(k\,\frac{\partial u}{\partial x_2}\right) - f\right] d\Omega dt - \int_I\int_{\Gamma_q} w\left[k\,\frac{\partial u}{\partial n} + \bar{q}\right] d\Gamma dt \;=\; 0\,. \quad (8.3)$$

One may taking advantage of the decoupling of space from time in the semi-discretization method, and "freeze" time in order to first operate on the space integrals, that is, on the integro-differential equation

$$\int_\Omega w \left[-\rho c\frac{\partial u}{\partial t} + \frac{\partial}{\partial x_1}\left(k\,\frac{\partial u}{\partial x_1}\right) + \frac{\partial}{\partial x_2}\left(k\,\frac{\partial u}{\partial x_2}\right) - f\right] d\Omega - \int_{\Gamma_q} w\left[k\,\frac{\partial u}{\partial n} + \bar{q}\right] d\Gamma \;=\; 0\,, \quad (8.4)$$

which, upon using integration by parts, the divergence theorem, and assumption (iii) takes the form

$$\int_\Omega w\rho c\frac{\partial u}{\partial t}\,d\Omega + \int_\Omega \left[\frac{\partial w}{\partial x_1}\,k\,\frac{\partial u}{\partial x_1} \;+\; \frac{\partial w}{\partial x_2}\,k\,\frac{\partial u}{\partial x_2} \;+\; wf\right] d\Omega \;+\; \int_{\Gamma_q} w\bar{q}\,d\Gamma \;=\; 0\,. \quad (8.5)$$

The Galerkin weighted-residual form can be now stated as follows: given k, ρ, c, f, and $\bar{q}$, find a function $u \in \mathcal{U}$, such that

$$\int_I \left[\int_\Omega w\rho c\frac{\partial u}{\partial t}\,d\Omega + \int_\Omega \left[\frac{\partial w}{\partial x_1}\,k\,\frac{\partial u}{\partial x_1} \;+\; \frac{\partial w}{\partial x_2}\,k\,\frac{\partial u}{\partial x_2} \;+\; wf\right] d\Omega \;+\; \int_{\Gamma_q} w\bar{q}\,d\Gamma\right] dt \;=\; 0\,, \quad (8.6)$$

for all $w \in \mathcal{W}$. Here, the space of admissible solutions $\mathcal{U}$ and the space of weighting functions $\mathcal{W}$ are defined respectively as

$$\mathcal{U} \;=\; \big\{ u \in H^1(\Omega \times I) \quad | \quad u = \bar{u} \text{ on } \Gamma_u \times I \quad , \quad u(x_1, x_2, 0) = u_0 \big\} \;, \tag{8.7}$$

and

$$\mathcal{W} \;=\; \big\{ w \in H^1(\Omega \times I) \quad | \quad w = 0 \text{ on } \Gamma_u \times I \quad , \quad w(x_1, x_2, 0) = 0 \big\} \;. \tag{8.8}$$

A Bubnov-Galerkin approximation of the weak form (8.6) can be effected by writing

$$\begin{aligned} u \;\doteq u_h \;&=\; \sum_{I=1}^{N} \varphi_I(x_1, x_2)\, u_I(t) \;+\; u_b(x_1, x_2, t) \;, \\ w \;\doteq w_h \;&=\; \sum_{I=1}^{N} \varphi_I(x_1, x_2)\, w_I(t) \;, \end{aligned} \tag{8.9}$$

where $\varphi_I = 0$ on Γ_u and $u_I(0) = w_I(0) = 0$. Also, the function $u_b(x_1, x_2, t)$ is chosen to satisfy the time-dependent Dirichlet boundary condition $(8.1)_3$ and the initial condition $(8.1)_4$. It is clear from (8.9) that this approximation induces a separation of spatial and temporal variables, which plays an essential role in the ensuing developments.

Substitution of u_h and w_h into the weak form (8.6) leads to

$$\begin{aligned} \int_I \Bigg[\sum_{I=1}^{N} w_I \int_\Omega \varphi_I \rho c \Big(\sum_{J=1}^{N} \varphi_J \dot{u}_J + \dot{u}_b \Big)\, d\Omega \\ + \sum_{I=1}^{N} w_I \int_\Omega \{\varphi_{I,1} \;\; \varphi_{I,2}\} k \left(\sum_{J=1}^{N} \left\{ \begin{array}{c} \varphi_{J,1} \\ \varphi_{J,2} \end{array} \right\} u_J + \left\{ \begin{array}{c} u_{b,1} \\ u_{b,2} \end{array} \right\} \right) d\Omega \\ + \sum_{I=1}^{N} w_I \int_\Omega \varphi_I f \, d\Omega + \sum_{I=1}^{N} w_I \int_{\Gamma_q} \varphi_I \bar{q} \, d\Gamma \Bigg] \, dt \;=\; 0 \;, \end{aligned} \tag{8.10}$$

where $\overline{(\cdot)} = \frac{d(\cdot)}{dt}$. This equation may be rewritten as

$$\int_I \left[\sum_{I=1}^{N} w_I \Big\{ \sum_{J=1}^{N} (M_{IJ} \dot{u}_J + K_{IJ} u_J) - F_I \Big\} \right] \;=\; 0 \;, \tag{8.11}$$

where

$$M_{IJ} \;=\; \int_\Omega \varphi_I \rho c \varphi_J \, d\Omega \;, \tag{8.12}$$

$$K_{IJ} \;=\; \int_\Omega \{\varphi_{I,1} \;\; \varphi_{I,2}\} \, k \left\{ \begin{array}{c} \varphi_{J,1} \\ \varphi_{J,2} \end{array} \right\} d\Omega \;, \tag{8.13}$$

and

$$F_I \;=\; -\int_\Omega \varphi_I \rho c \dot{u}_b \, d\Omega - \int_\Omega \{\varphi_{I,1}\ \varphi_{I,2}\}\, k \left\{ \begin{array}{c} u_{b,1} \\ u_{b,2} \end{array} \right\} d\Omega \;-\int_\Omega \varphi_I f \, d\Omega \;-\; \int_{\Gamma_q} \varphi_I \, \bar{q} \, d\Gamma \;. \quad (8.14)$$

The arrays [**M**], [**K**] and [**F**], whose components are given above, are termed the *mass* (or *capacitance*) matrix, the *stiffness* matrix and the *forcing* vector, respectively. Equations (8.12) and (8.13) clearly demonstrate that [**M**] and [**K**] are symmetric. In addition, it is easy to establish that [**M**] is positive-definite provided $\rho c > 0$, while [**K**] is positive-semidefinite provided $k > 0$, as already argued for the steady problem.

In conclusion, one arrives at the semi-discrete form (8.11), which may be also written in matrix form as

$$\int_I [\mathbf{w}]^T \big([\mathbf{M}][\dot{\mathbf{u}}] + [\mathbf{K}][\mathbf{u}] - [\mathbf{F}] \big) \, dt \;=\; 0 \;, \quad (8.15)$$

where $[\mathbf{u}] = [u_1(t)\ u_2(t)\ \ldots\ u_N(t)]^T$ and $[\mathbf{w}] = [w_1(t)\ w_2(t)\ \ldots\ w_N(t)]^T$. Equation (8.15) is now an integro-differential equation in time only, as all the spatial derivatives and integrals have been evaluated and "stored" in the arrays [**M**], [**K**] and [**F**].

In the semi-discretization method, once the spatial problem has been discretized, one may proceed to the temporal problem. Here, there are two distinct options:

(a) Discretize [**u**] and [**w**] in time according to some polynomial series, that is,

$$[\mathbf{u}] \;\doteq\; [\hat{\mathbf{u}}] \;=\; \sum_{n=1}^{M} [\boldsymbol{\alpha}_n] t^n \qquad , \qquad [\mathbf{w}] \;\doteq\; [\hat{\mathbf{w}}] \;=\; \sum_{n=1}^{M} [\boldsymbol{\beta}_n] t^n \;, \quad (8.16)$$

where $[\boldsymbol{\alpha}_n]$ is a vector to be determined and $[\boldsymbol{\beta}_n]$ is an arbitrary vector. The preceding polynomial approximations in time satisfy the requirements for the initial values of **u** and **w**, as stipulated in the admissible space $\mathcal{U}$ and $\mathcal{W}$ in (8.7) and (8.8), respectively. These approximate functions are then substituted into the semi-discrete form (8.15) and the resulting system is solved for the values of $[\boldsymbol{\alpha}_n]$. This is essentially a Bubnov-Galerkin approximation in time.

(b) Apply a standard discrete time integrator directly on the semi-discrete form (8.15). This amounts to choosing **w** to consist of Dirac-delta functions at discrete times t_1, t_2, ..., t_n, t_{n+1}, ..., T, which would imply that the system of ordinary differential equations

$$[\mathbf{M}][\dot{\mathbf{u}}] + [\mathbf{K}][\mathbf{u}] \;=\; [\mathbf{F}] \quad (8.17)$$

is to be exactly satisfied at these times.

In the remainder of this section, the second option is pursued. In addition, without loss of generality, the initial conditions on the dependent variable u are subsumed into the vector $[\mathbf{u}]$ (as opposed to being delegated to the function $u_b(x_1, x_2, t)$), hence the initial condition for the solution vector $[\mathbf{u}]$ is $[\mathbf{u}(0)]$.

To find the solution to the system of ordinary differential equations (8.17), start by recall that the general solution of the homogeneous counterpart of (8.17), that is, when $[\mathbf{F}] = [\mathbf{0}]$, is of the form

$$[\mathbf{u}(t)] \;=\; \sum_{I=1}^{N} c_I e^{-\lambda_I t} [\mathbf{z}_I] \;, \tag{8.18}$$

where the pairs $(\lambda_I, [\mathbf{z}_I])$, $I = 1, 2, \ldots, N$, and the constants c_I are to be determined. Upon substituting a typical such pair $(\lambda, [\mathbf{z}])$ into the homogeneous equation, one gets

$$e^{-\lambda t}(-\lambda[\mathbf{M}] + [\mathbf{K}])[\mathbf{z}] \;=\; [\mathbf{0}] \;, \tag{8.19}$$

hence,

$$\lambda[\mathbf{M}][\mathbf{z}] \;=\; [\mathbf{K}][\mathbf{z}] \;. \tag{8.20}$$

Equation (8.20) corresponds to the general symmetric linear eigenvalue problem, which can be solved for the eigenpairs $(\lambda_I, [\mathbf{z}_I])$, $I = 1, 2, \ldots, N$. For notational simplicity, define the $N \times N$ arrays

$$[\mathbf{\Lambda}] \;=\; \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_N \end{bmatrix} \tag{8.21}$$

and

$$[\mathbf{Z}] \;=\; \begin{bmatrix} [\mathbf{z}_1] & [\mathbf{z}_2] & \ldots & [\mathbf{z}_N] \end{bmatrix} \;, \tag{8.22}$$

so that the eigenvalue problem (8.20) may be conveniently rewritten in matrix form for all eigenpairs as

$$[\mathbf{M}][\mathbf{Z}][\mathbf{\Lambda}] \;=\; [\mathbf{K}][\mathbf{Z}] \;. \tag{8.23}$$

Given that $[\mathbf{M}]$ and $[\mathbf{K}]$ are symmetric, standard orthogonality properties of the eigenpairs $(\lambda_I, \mathbf{z}_I)$, $I = 1, 2, \ldots, N$, where $\mathbf{z}_I$ are linearly independent, yield the diagonalizations

$$[\mathbf{Z}]^T[\mathbf{M}][\mathbf{Z}] \;=\; \begin{bmatrix} m_1 & & & \\ & m_2 & & \\ & & \ddots & \\ & & & m_N \end{bmatrix} \tag{8.24}$$

and

$$[\mathbf{Z}]^T[\mathbf{K}][\mathbf{Z}] \;=\; \begin{bmatrix} k_1 & & & \\ & k_2 & & \\ & & \ddots & \\ & & & k_N \end{bmatrix} , \tag{8.25}$$

where $\lambda_I = \dfrac{k_I}{m_I}$, $m_I > 0$, and $k_I \geq 0$, thus also $\lambda_I \geq 0$. Indeed, starting from (8.20) with any eigenpair $(\lambda_I, \mathbf{z}_I)$ and premultiplying both sides by the eigenvector $[\mathbf{z}_J]^T$ yields

$$\lambda_I[\mathbf{z}_J]^T[\mathbf{M}][\mathbf{z}_I] \;=\; [\mathbf{z}_J]^T[\mathbf{K}][\mathbf{z}_I] \; . \tag{8.26}$$

Conversely, starting from (8.20) for another eigenpair $(\lambda_J, \mathbf{z}_J)$ and premultiplying both sides by the eigenvector $[\mathbf{z}_I]^T$ leads to

$$\lambda_J[\mathbf{z}_I]^T[\mathbf{M}][\mathbf{z}_J] \;=\; [\mathbf{z}_I]^T[\mathbf{K}][\mathbf{z}_J] \; . \tag{8.27}$$

Taking into account the symmetry of $[\mathbf{M}]$ and $[\mathbf{K}]$, equations (8.26) and (8.27) imply that

$$(\lambda_J - \lambda_I)[\mathbf{z}_J]^T[\mathbf{M}][\mathbf{z}_I] \;=\; 0 \; . \tag{8.28}$$

Equation (8.28) proves the diagonalization of $[\mathbf{M}]$, as in (8.24), provided that $\lambda_I \neq \lambda_J$. Either (8.26) or (8.27) may be subsequently invoked to deduce the simultaneous diagonalization of $[\mathbf{K}]$, as in (8.25).

It can be shown that the preceding diagonalization is possible even in the case of repeated eigenvalues. Indeed, since $\mathbf{M}$ is positive-definite, one may let $\mathbf{u} = \mathbf{M}^{-1/2}\mathbf{v}$ and substitute this in the original homogeneous system, which would take the form $\dot{\mathbf{v}} + \tilde{\mathbf{M}}\mathbf{v} = \mathbf{0}$, where $\tilde{\mathbf{M}} = \mathbf{M}^{-1/2}\mathbf{K}\tilde{\mathbf{M}}^{-1/2}$ is symmetric. It is subsequently easy to show that the eigenvectors $[\tilde{\mathbf{z}}_I]$ of this system can be always made orthogonal to each other with respect to $\tilde{\mathbf{M}}$ and are related to the eigenvalues of the original system problem as $[\mathbf{z}_I] = [\mathbf{M}^{-1/2}][\tilde{\mathbf{z}}_I]$. Once the eigenpairs are known, the constants c_I in (8.18) are determined from the initial conditions of the problem.

The solution of the non-homogeneous equations (8.17) is attained by employing the classical technique of *variation of parameters*, according to which it is assumed that

$$[\mathbf{u}(t)] \;=\; [\mathbf{Z}][\mathbf{v}(t)] \; , \tag{8.29}$$

where $[\mathbf{Z}]$ is defined in (8.22) for the homogeneous problem and $[\mathbf{v}(t)]$ is a vector function of time to be determined. Substituting (8.29) into (8.17) gives rise to

$$[\mathbf{M}][\mathbf{Z}][\dot{\mathbf{v}}] + [\mathbf{K}][\mathbf{Z}][\mathbf{v}] \;=\; [\mathbf{F}] \; , \tag{8.30}$$

subject to the initial condition $[\mathbf{Z}][\mathbf{v}(0)] = [\mathbf{u}(0)]$. Premultiplying this equation by $[\mathbf{Z}]^T$ leads to

$$[\mathbf{Z}]^T[\mathbf{M}][\mathbf{Z}][\dot{\mathbf{v}}] + [\mathbf{Z}]^T[\mathbf{K}][\mathbf{Z}][\mathbf{v}] \;=\; [\mathbf{Z}]^T[\mathbf{F}] \;. \tag{8.31}$$

Taking now into account the earlier orthogonality conditions (8.24) and (8.25) gives rise to

$$\begin{bmatrix} m_1 & & & \\ & m_2 & & \\ & & \ddots & \\ & & & m_N \end{bmatrix} \begin{bmatrix} \dot{v}_1 \\ \dot{v}_2 \\ \vdots \\ \dot{v}_N \end{bmatrix} + \begin{bmatrix} k_1 & & & \\ & k_2 & & \\ & & \ddots & \\ & & & k_N \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_N \end{bmatrix} \;, \tag{8.32}$$

where $g_I = [\mathbf{z}_I]^T[\mathbf{F}]$, $I = 1, 2, \ldots, N$. It is readily concluded from (8.32) that the original system of coupled linear ordinary differential equations (8.17) has been reduced to a set of N uncoupled scalar ordinary differential equations of the form

$$m_I\dot{v}_I + k_I v_I = g_I \;, \tag{8.33}$$

for $I = 1, 2, \ldots, N$. Therefore, in order to understand the behavior of the original N-dimensional system of ordinary differential equations in (8.17), it is sufficient to study the solution of a single scalar ordinary differential equation of the form

$$m\dot{v} + kv \;=\; g \;, \tag{8.34}$$

with initial condition $v(0) = v_0$.

The general solution of equation (8.34) can be obtained using the method of variation of parameters, and is given by

$$v(t) \;=\; e^{-\lambda t}y(t) \;, \tag{8.35}$$

where $\lambda = \dfrac{k}{m}$ and $y = y(t)$ is a function to be determined. Upon substituting the general solution (8.35) into (8.34), it follows that

$$\dot{y}(t) \;=\; \frac{1}{m}e^{\lambda t}g \;. \tag{8.36}$$

This equation may be integrated in the time interval $(t_n, t]$, which results in

$$y(t) \;=\; y_n + \int_{t_n}^{t} \frac{1}{m}e^{\lambda\tau}g(\tau)\,d\tau \;, \tag{8.37}$$

where $y_n = y(t_n)$, see Figure 8.2. Hence, one obtains from (8.35) the solution for $v(t)$ in

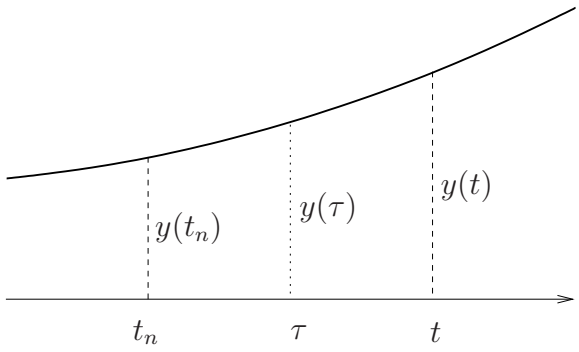


Figure 8.2. *Integration of* (8.36) *in the domain* $(t_n, t]$

convolution form as

$$v(t) \;=\; e^{-\lambda t} y_n + \int_{t_n}^{t} \frac{1}{m} e^{\lambda(\tau - t)} g(\tau)\, d\tau \;. \tag{8.38}$$

Noting from (8.35) that $v(t_n) = v_n = e^{-\lambda t_n} y_n$, it follows that the preceding solution can be also expressed as

$$v(t) \;=\; e^{-\lambda(t - t_n)} v_n + \int_{t_n}^{t} \frac{1}{m} e^{-\lambda(t - \tau)} g(\tau)\, d\tau \;. \tag{8.39}$$

Setting $t = t_{n+1}$, it is readily seen from (8.39) that

$$v_{n+1} \;=\; e^{-\lambda \Delta t_n} v_n + \int_{t_n}^{t_n + 1} \frac{1}{m} e^{-\lambda(t_{n+1} - \tau)} g(\tau)\, d\tau \;, \tag{8.40}$$

where $\Delta t_n = t_{n+1} - t_n$.

The ratio $r = \dfrac{v_{n+1}}{v_n}$ is termed *the amplification factor*. In the homogeneous case ($g = 0$), equation (8.40) immediately implies that $r = e^{-\lambda \Delta t_n}$, that is, the exact solution experiences exponential decay. This, in turn, implies that $r \to 1$ when $\lambda \Delta t_n \to 0$ and $r \to 0$ when $\lambda \Delta t_n \to \infty$.

8.2 Stability of classical time integrators

In this section, attention is focused on the application of certain discrete time integrators to the scalar first-order differential equation (8.34), which, as argued in the preceding section, fully represents the general system (8.17) obtained through the semi-discretization of the weak form (8.2).

The first discrete time integrator is the *forward Euler method*, according to which the time derivative $\dot{v}$ can be approximated at time t_n by using a Taylor series expansion of $v(t)$

around $t = t_n$ as

$$v_{n+1} \;=\; v_n + \Delta t_n \dot{v}_n + o(\Delta t_n^2) \;, \tag{8.41}$$

where, adopting a simplified notation, $v_n = v(t_n)$ and $\dot{v}_n = \dot{v}(t_n)$. Upon ignoring the second-order term in $\Delta t_n = t_{n+1} - t_n$, the preceding equation leads to

$$\dot{v}_n \;\doteq\; \frac{v_{n+1} - v_n}{\Delta t_n} \;. \tag{8.42}$$

Substituting $\dot{v}_n$ from (8.42) to the scalar equation (8.34) at t_n, it is concluded that

$$m\,\frac{v_{n+1} - v_n}{\Delta t_n} + k v_n \;=\; g_n \;. \tag{8.43}$$

This equation may be trivially rewritten as

$$v_{n+1} \;=\; (1 - \lambda \Delta t_n) v_n + \frac{\Delta t_n}{m}\, g_n \;, \tag{8.44}$$

where, again $\lambda = \dfrac{k}{m}$. In the homogeneous case ($g = 0$), it is seen from (8.44) that the discrete amplification ratio r_f of the forward Euler method is given by

$$r_f \;=\; 1 - \lambda \Delta t_n \;. \tag{8.45}$$

Equation (8.45) implies that for finite values of λ, the limiting case $\Delta t_n \to 0$ leads to $r_f \to 1$, which is consistent with the exact solution, as argued earlier in this section. However, the limiting case $\Delta t_n \to \infty$ leads to $r_f \to -\infty$, which reveals that the discrete solution does not predict exponential decay in the limit of an infinitely large time step Δt_n. Ignoring the inhomogeneous term in (8.44), it is clear that for $\lambda \Delta t_n > 1$, the discrete solution exhibits oscillations with respect to $v = 0$ (which are, of course, absent in the exact exponentially decaying solution). For $1 < \lambda \Delta t_n < 2$, these oscillations are decaying, hence the discrete solution is *stable*. However, for $\lambda \Delta t_n > 2$, the oscillations grow in magnitude with each time step and the solution becomes *unstable*, that is, instead of decaying, it artificially grows toward infinity. Therefore, the forward Euler method is referred to as a *conditionally stable* method, which means that its time step Δt_n needs to be controlled in order to satisfy the condition $\Delta t_n < \dfrac{2}{\lambda} = \Delta t_{cr}$, where Δt_{cr} is the *critical step-size*. In systems with many degrees of freedom, such as (8.17), the critical step-size may be defined as

$$\Delta t_{cr} \;=\; \frac{2}{\lambda_{max}} \;, \tag{8.46}$$

where λ_{max} is the maximum eigenvalue of problem (8.20). This implies that in order to guarantee stability for the forward Euler method, one needs to know (or estimate) the

maximum eigenvalue of (8.20). Fortunately, there exist inexpensive methods of estimating λ_{max} in finite element approximations, a fact that significantly enhances the usefulness of the forward Euler method.

A simple scaling argument can be made for the dependence of λ_{max} on the element size h. To this end, note, with the aid of equations (8.13) and (8.12), that, since the interpolation functions φ_I are dimensionless, the components $[K_{IJ}]$ of the stiffness matrix for the two-dimensional transient heat conduction problem are of order $o(1)$, while the components $[M_{IJ}]$ of the mass matrix for the same problem are of order $o(h^2)$. This implies that, by virtue of its definition, λ is of order $o(h^{-2})$, hence Δt_{cr} is of order $o(h^2)$. This means that, when using forward Euler integration in the solution of the two-dimensional transient heat conduction equation, the critical step-size must be reduced quadratically under mesh refinement, that is, halving the mesh-size necessitates reduction of the step-size by a factor of four. Similar scaling arguments can be made for one- or three-dimensional versions of the transient heat conduction equation.

An alternative discrete time integrator is the *backward Euler method*, which may be deduced by writing v_n using a Taylor series expansion around $t = t_{n+1}$ as

$$v_n \;=\; v_{n+1} - \Delta t_n \dot{v}_{n+1} + o(\Delta t_n^2) \;, \tag{8.47}$$

which, upon ignoring the second-order terms in Δt_n leads to

$$\dot{v}_{n+1} \;\doteq\; \frac{v_{n+1} - v_n}{\Delta t_n} \;. \tag{8.48}$$

Writing now (8.34) at t_{n+1}, with $\dot{v}_{n+1}$ estimated from (8.48), results in

$$m\,\frac{v_{n+1} - v_n}{\Delta t_n} + k v_{n+1} = g_{n+1} \tag{8.49}$$

or, upon solving for v_{n+1},

$$v_{n+1} \;=\; \frac{1}{1+\lambda\Delta t_n}\,v_n + \frac{\Delta t_n}{1+\lambda\Delta t_n}\,\frac{1}{m}\,g_{n+1} \;. \tag{8.50}$$

For the homogeneous problem, equation (8.50) implies that in the limiting cases $\Delta t_n \to 0$ and $\Delta t_n \to \infty$, the discrete amplification ratio r_b, defined as

$$r_b \;=\; \frac{1}{1+\lambda\Delta t_n} \;, \tag{8.51}$$

satisfies $r_b \to 1$ and $r_b \to 0$, respectively. This means that the backward Euler method is consistent with the exact solution in both extreme cases. In addition, as seen from (8.51),

this method is *unconditionally stable*, in the sense that it yields numerical approximations to $v(t)$ that are decaying in time (without any oscillations) regardless of the step-size Δt_n. Figure 8.3 shows the amplification factor for the two methods, as well as for the exact solution.

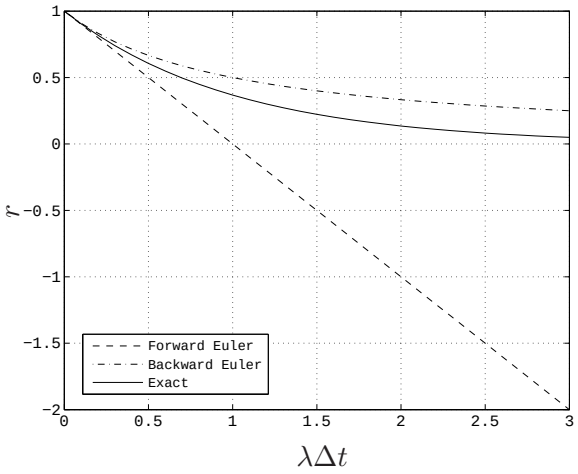


Figure 8.3. *Amplification factor r as a function of $\lambda\Delta t$ for forward Euler, backward Euler and the exact solution of the homogeneous counterpart of* (8.34)

Returning to the system of ordinary differential equations in (8.17), one may use forward Euler integration at time t_n, which leads to

$$[\mathbf{M}]\,\frac{[\mathbf{u}_{n+1}]-[\mathbf{u}_n]}{\Delta t_n}+[\mathbf{K}][\mathbf{u}_n] \;=\; [\mathbf{F}_n]\;, \tag{8.52}$$

hence

$$[\mathbf{M}][\mathbf{u}_{n+1}] \;=\; [\mathbf{M}][\mathbf{u}_n]-\Delta t_n[\mathbf{K}][\mathbf{u}_n]+\Delta t_n[\mathbf{F}_n]\;. \tag{8.53}$$

It is clear that computing $[\mathbf{u}_{n+1}]$ requires the factorization of $[\mathbf{M}]$, which may be performed once and be used repeatedly for $n = 1, 2, \ldots$. In fact, the factorization itself may become unnecessary if $[\mathbf{M}]$ is diagonal, in which case $[\mathbf{M}]^{-1}$ can be obtained from $[\mathbf{M}]$ by merely inverting its diagonal components. In this case, it is clear that the advancement of the solution from $[\mathbf{u}_n]$ to $[\mathbf{u}_{n+1}]$ does not require the solution of an algebraic system. For this reason, the resulting semi-discrete method is termed *explicit*. A diagonal approximation of the mass matrix $[\mathbf{M}]$ can be easily computed using nodal quadrature, that is, by evaluating the integral expression that defines its components using an integration rule that takes the element nodes as its sampling points. This observation can be readily justified by recalling

the definition of the components M_{IJ} of the mass matrix in (8.12) and property (5.27) of the element interpolation functions.

The backward Euler method can also be applied to (8.17) at time t_{n+1}, resulting in

$$[\mathbf{M}]\,\frac{[\mathbf{u}_{n+1}]-[\mathbf{u}_n]}{\Delta t_n}+[\mathbf{K}][\mathbf{u}_{n+1}] \;=\; [\mathbf{F}_{n+1}]\;, \tag{8.54}$$

which implies that

$$\big([\mathbf{M}]+\Delta t_n[\mathbf{K}]\big)[\mathbf{u}_{n+1}] \;=\; [\mathbf{M}][\mathbf{u}_n]+\Delta t_n[\mathbf{F}_{n+1}]\;. \tag{8.55}$$

The above system requires factorization of $[\mathbf{M}]+\Delta t_n[\mathbf{K}]$, which cannot be circumvented by diagonalization, as in the forward Euler case. Hence, the resulting semi-discrete method is termed *implicit*, in the sense that advancement of the solution from $[\mathbf{u}_n]$ to $[\mathbf{u}_{n+1}]$ cannot be achieved without the solution of algebraic equations.

Explicit and implicit semi-discrete methods give rise to vastly different computer code architectures. In the former case, emphasis is placed on the control of step-size Δt_n, so that is always remain below the critical value Δt_{cr}. In the latter, emphasis is placed on the efficient solution of the resulting algebraic equations.

8.3 Weighted-residual interpretation of classical time integrators

It is instructive to re-derive the discrete time integrators of the previous section using a weighted-residual formalization. To this end, start from equation (8.15) and consider the time interval $I=(t_n,t_{n+1}]$, where

$$\int_{t_n}^{t_{n+1}}[\mathbf{w}]^T\big([\mathbf{M}][\dot{\mathbf{u}}]+[\mathbf{K}][\mathbf{u}]-[\mathbf{F}]\big)\,dt \;=\; 0\;. \tag{8.56}$$

Now, choose a linear polynomial interpolation of $[\mathbf{u}]$ in time, namely

$$[\mathbf{u}] \;\doteq\; \left(1-\frac{t-t_n}{\Delta t_n}\right)[\mathbf{u}_n]+\frac{t-t_n}{\Delta t_n}\,[\mathbf{u}_{n+1}]\;, \tag{8.57}$$

where $[\mathbf{u}_n]$ is known from the integration in the previous time interval $(t_{n-1},t_n]$.

Different discrete time integrators can be deduced by appropriate choices of the weighting function $[\mathbf{w}]$. Specifically, let

$$[\mathbf{w}] \;\doteq\; \delta(t_n^+)[\mathbf{c}] \tag{8.58}$$

in $(t_n, t_{n+1}]$, where $\mathbf{c}$ is an arbitrary constant vector. Substituting (8.57) and (8.58) into (8.56), one obtains (8.52), thus recovering the semi-discrete equations of the forward Euler rule. Alternatively, setting

$$[\mathbf{w}] \doteq \delta(t_{n+1})[\mathbf{c}] \,, \tag{8.59}$$

one readily obtains (8.54), namely the semi-discrete equations of the backward Euler rule.

More generally, let

$$[\mathbf{w}] \doteq \delta(t_{n+\alpha})[\mathbf{c}] \,, \tag{8.60}$$

where $0 < \alpha \leq 1$ and $t_{n+\alpha} = (1-\alpha)t_n + \alpha t_{n+1}$. Substituting (8.57) and (8.60) into (8.56) leads to

$$[\mathbf{M}]\,\frac{[\mathbf{u}_{n+1}] - [\mathbf{u}_n]}{\Delta t_n} + [\mathbf{K}]\Big((1-\alpha)[\mathbf{u}_n] + \alpha[\mathbf{u}_{n+1}]\Big) = [\mathbf{F}_{n+\alpha}] \tag{8.61}$$

or

$$\big([\mathbf{M}] + \alpha\Delta t_n[\mathbf{K}]\big)[\mathbf{u}_{n+1}] = \big([\mathbf{M}] - (1-\alpha)\Delta t_n[\mathbf{K}]\big)[\mathbf{u}_n] + \Delta t_n[\mathbf{F}_{n+\alpha}] \,, \tag{8.62}$$

which corresponds to the *generalized trapezoidal rule*. For the special case $\alpha = 1/2$, one obtains the *Crank-Nicolson rule*.

Finally, one may choose to use a smooth interpolation for the weighting function $[\mathbf{w}]$ in $(t_n, t_{n+1}]$. Indeed, let

$$[\mathbf{w}] \doteq \frac{t - t_n}{\Delta t_n}[\mathbf{w}_{n+1}] \,, \tag{8.63}$$

where $\mathbf{w}_{n+1}$ is an arbitrary constant vector. In this case, one recovers the Bubnov-Galerkin method in time. In particular, substituting (8.57) and (8.63) into (8.56) leads to

$$\int_{t_n}^{t_{n+1}} \frac{t-t_n}{\Delta t_n}[\mathbf{w}_{n+1}]^T \left[[\mathbf{M}]\,\frac{[\mathbf{u}_{n+1}] - [\mathbf{u}_n]}{\Delta t_n} + [\mathbf{K}]\left\{\left(1 - \frac{t-t_n}{\Delta t_n}\right)[\mathbf{u}_n] + \frac{t-t_n}{\Delta t_n}[\mathbf{u}_{n+1}]\right\} - [\mathbf{F}]\right] dt = 0 \,. \tag{8.64}$$

Upon integrating (8.64) in time and recalling that $[\mathbf{w}_{n+1}]$ is arbitrary, one finds that

$$\frac{1}{2}[\mathbf{M}]\big([\mathbf{u}_{n+1}] - [\mathbf{u}_n]\big) + [\mathbf{K}]\left(\frac{1}{6}[\mathbf{u}_n] + \frac{1}{3}[\mathbf{u}_{n+1}]\right)\Delta t_n - \int_{t_n}^{t_{n+1}} \frac{t-t_n}{\Delta t_n}[\mathbf{F}]\,dt = \mathbf{0} \tag{8.65}$$

or

$$\left([\mathbf{M}] + \frac{2}{3}\Delta t_n[\mathbf{K}]\right)[\mathbf{u}_{n+1}] = \left([\mathbf{M}] - \frac{1}{3}\Delta t_n[\mathbf{K}]\right)[\mathbf{u}_n] + [\bar{\mathbf{F}}] \,, \tag{8.66}$$

where $[\bar{\mathbf{F}}] = \displaystyle\int_{t_n}^{t_{n+1}} 2\,\frac{t-t_n}{\Delta t_n}[\mathbf{F}]\,dt$. When $[\mathbf{F}]$ is independent of time, the Bubnov-Galerkin method coincides with the generalized trapezoidal rule with $\alpha = 2/3$, as is easily inferred from (8.62).

8.4 Exercises

Problem 1

Show that the mass matrix emanating from the finite element approximation of the two-dimensional transient heat conduction problem is positive-definite under the usual assumption that the interpolation functions are linearly independent.

Chapter 9

Hyperbolic Differential Equations

The classical Bubnov-Galerkin finite element method is optimal in the sense of the best approximation property for elliptic partial differential equations. In many problems of mechanics and convective heat transfer where convection dominates diffusion, this method ceases to be optimal. Rather, its solutions exhibit spurious oscillations in the dependent variable which tend to increase depending on the relative strength of the convective component. It is clear that another method has to be used in order to circumvent this problem. A concise discussion of this issue is the subject of the present chapter.

9.1 The one-dimensional convection-diffusion equation

The limitations of the classical Bubnov-Galerkin method and an alternative approach designed to address these limitations are discussed here in the context of the one-dimensional convection-diffusion equation

$$u_{,t} + \alpha u_{,x} = \epsilon u_{,xx} \qquad ; \qquad \alpha \geq 0 \quad , \quad \epsilon \geq 0 \; , \tag{9.1}$$

which was already encountered in Chapter 1. The steady solution of this equation in the domain $(0, L)$ with boundary conditions $u(0) = 0$ and $u(L) = \bar{u} > 0$ is

$$u(x) = \frac{1 - e^{\frac{\alpha}{\epsilon}x}}{1 - e^{\frac{\alpha}{\epsilon}L}} \, \bar{u} \; . \tag{9.2}$$

The non-dimensional number $Pe = \dfrac{\alpha}{\epsilon} L$, is known as the *Péclet number* and provides a measure of relative significance of convection and diffusion, such that convection dominates

if $Pe \gg 1$ and diffusion dominates if $Pe \ll 1$. Recalling the definition of the Péclet number, one may rewrite (9.2) as

$$u(x) \;=\; \frac{1 - e^{Pe\frac{x}{L}}}{1 - e^{Pe}}\,\bar{u}\;. \tag{9.3}$$

It is clear from (9.3) that when diffusion dominates, then $u(x) \doteq \frac{x}{L}\bar{u}$. This is because the two exponential terms in (9.3) have exponents that are much smaller than one, thus may be accurately approximated by a first-order Taylor series expansion. In contrast, when convection dominates, then the solution is nearly zero throughout the domain except for a small region near the boundary $x = L$, where it increases sharply to $\bar{u}$. The latter is due to the fact that in this case the exponential terms in (9.3) are much larger than one, so that $u(x) \doteq e^{Pe(\frac{x}{L}-1)}\bar{u}$. It is precisely this steep boundary layer that the classical Bubnov-Galerkin method fails to accurately resolve, as will be argued shortly. Figure 9.1 demonstrates the two distinct characters of $u(x)$ in (9.3) depending on the value of the Péclet number.

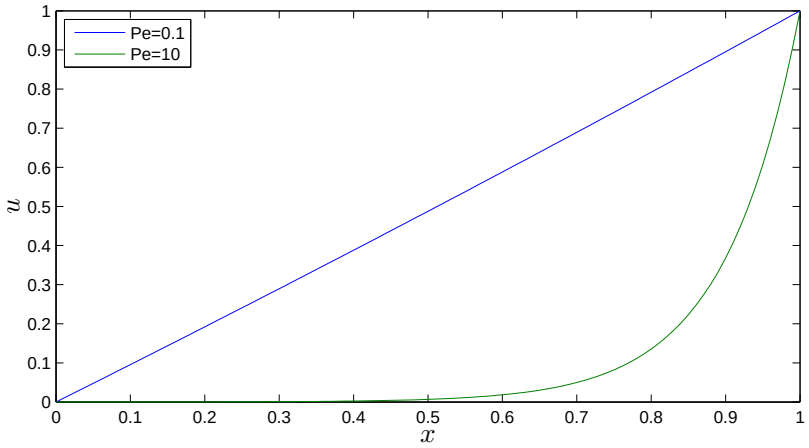


Figure 9.1. *Plots of the solution* (9.3) *of the steady-state convection-diffusion equation for* $L = 1$, $\bar{u} = 1$ *and Péclet numbers* $Pe = 0.1$ *and* $Pe = 10$.

Before proceeding further, it is important to note that the one-dimensional convection-diffusion equation is not substantially different in nature from the three-dimensional Navier-Stokes equations which govern the motion of a compressible Newtonian fluid, and which can be expressed as

$$-\nabla p + (\lambda + \mu)\nabla(\nabla \cdot \mathbf{v}) + \mu\nabla \cdot (\nabla \mathbf{v}) + \mathbf{f} \;=\; \rho(\frac{\partial \mathbf{v}}{\partial t} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}}\mathbf{v})\;. \tag{9.4}$$

Here, $p = p(\rho)$ is the pressure, $\mathbf{v}$ is the fluid velocity, ρ is the mass density, $\mathbf{f}$ is the body force per unit volume, and λ, μ are material constants. The second and third terms of the

left-hand side are diffusive and the last term of the right-hand side is convective. A similar conclusion can be reached for the incompressible case, that is when $\nabla \cdot \mathbf{v} = 0$. In the Navier-Stokes equations, the non-dimensional parameter that quantifies the relative significance of convection and diffusion is the *Reynolds number* Re, defined as $Re = \frac{\rho|\mathbf{v}|}{\mu}L$. Again, the classical Bubnov-Galerkin method performs poorly for $Re \gg 1$, while it yields good results for $Re \ll 1$.

Returning to the one-dimensional steady convection-diffusion equation, one may start by applying the Bubnov-Galerkin approximation method and subsequently discretize the resulting equations by $N+1$ equally-sized finite elements with linear interpolation functions for the dependent variable u, as in Figure 9.2. It is straightforward to show that the resulting

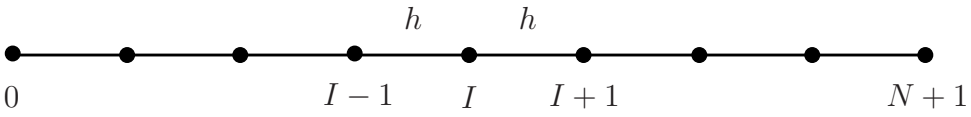


Figure 9.2. *Finite element discretization for the one-dimensional convection-diffusion equation*

system of linear algebraic equations is

$$\alpha\frac{1}{2h}(u_{I+1} - u_{I-1}) \;=\; \epsilon\frac{1}{h^2}(u_{I+1} - 2u_I + u_{I-1}) \quad , \quad I = 1, 2, \ldots, N \; , \tag{9.5}$$

where $h = \frac{L}{N+1}$. In fact, these equations coincide with those obtained by applying directly the finite difference method on the differential equation, as in Section 1.2.2. One may rewrite the above equations in the form

$$au_{I-1} + bu_I + cu_{I+1} \;=\; 0 \quad , \quad I = 1, 2, \ldots, N \; , \tag{9.6}$$

where

$$a \;=\; -(\frac{\alpha}{2} + \frac{\epsilon}{h}) \quad , \quad b \;=\; \frac{2\epsilon}{h} \quad , \quad c \;=\; \frac{\alpha}{2} - \frac{\epsilon}{h} \; . \tag{9.7}$$

The system of linear algebraic equations in (9.6) may be conveniently expressed in matrix form as

$$\begin{bmatrix} b & c & & & & \\ a & b & c & & & \\ & a & b & c & & \\ & & & \ddots & & \\ & & & a & b & c \\ & & & & a & b \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{N-1} \\ u_N \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ -c\bar{u} \end{bmatrix} \; . \tag{9.8}$$

Clearly, the $N \times N$ matrix in (9.8) is unsymmetric as long as $\alpha \neq 0$. If $c = 0$ (that is, if $\dfrac{\alpha h}{2\epsilon} = 1$), then one gets a "sharp" solution of the form $u_I = 0$ for $I = 0, 1, \ldots, N$ and $u_{N+1} = \bar{u}$, which is an accurate approximation of the exact solution, see Figure 9.3. It also

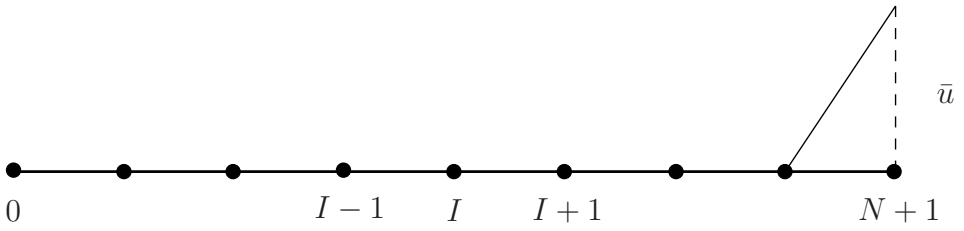


Figure 9.3. *Finite element solution for the one-dimensional convection-diffusion equation for* $c = 0$

follows from (9.8) that the numerical solution exhibits no oscillations when $c < 0$. Indeed, since $b > 0$ the first of the equations in (9.8) leads to $u_2 = -\frac{b}{c}u_1$, hence u_2 has the same sign as u_1. Next, it can be shown that the second of the equations in (9.8) implies that u_3 retain the sign of u_1 and u_2 as long as $b^2 > ac$, which can be confirmed from (9.7).

Likewise, the Bubnov-Galerkin solution of the convection-diffusion equation exhibits oscillations around zero when $c > 0$, see Figure 9.4. This is easy to argue by noting that u_2

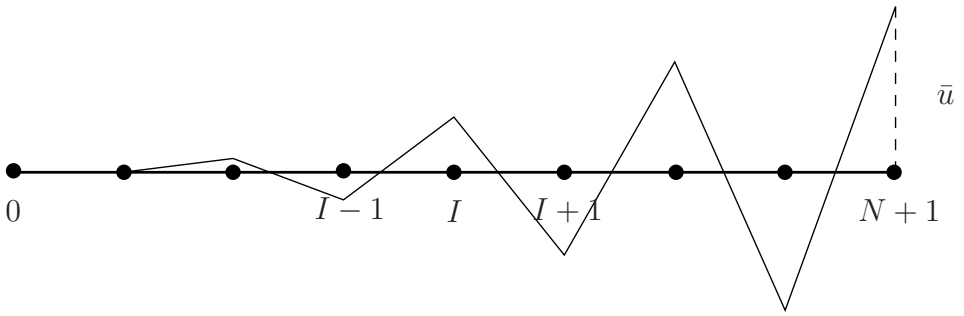


Figure 9.4. *Finite element solution for the one-dimensional convection-diffusion equation for* $c > 0$

changes sign relative to u_1, given that, as before, $u_2 = -\frac{b}{c}u_1$. Also, since $a < 0$ must have the sign of u_1 to satisfy the second of the equations in (9.8), and so on.

In conclusion, one may be able to accurately resolve the analytical solution so long as $c \leq 0$, or if, equivalently, the so-called *grid Péclet number* $Pe_h = \dfrac{\alpha}{\epsilon}\dfrac{h}{2} = Pe\dfrac{h/2}{L}$ is less or equal to one. If $Pe \gg 1$, then satisfying the condition $Pe_h \leq 1$ may require a prohibitively

small h. This is precisely why the Bubnov-Galerkin method is not a practical formulation for convection-dominated problems.

To remedy the oscillatory behavior of the Bubnov-Galerkin method, one may choose instead to employ an *upwinding method*. Since this problem is obviously caused by the convective (as opposed to the diffusive) part of the equation, one idea is to modify the spatial interpolation of the convective term by forcing it to use information which is taken to be preferentially upstream (that is, skew the interpolation toward the part of the domain where the solution is relatively constant). To this end, equation (9.5) may be replaced by

$$\alpha \frac{1}{h}(u_I - u_{I-1}) \;=\; \epsilon \frac{1}{h^2}(u_{I+1} - 2u_I + u_{I-1}) \quad , \quad I = 1, 2, \ldots, N \; , \tag{9.9}$$

which it tantamount to using an *upwind difference* approximation $\left.\frac{du}{dx}\right|_I \doteq \frac{1}{h}(u_I - u_{I-1})$ as opposed to a centered difference $\left.\frac{du}{dx}\right|_I \doteq \frac{1}{2h}(u_{I+1} - u_{I-1})$ for the convective term. One may interpret this upwind difference as the algebraic difference of the corresponding centered difference from a discrete artificial viscous term. Indeed, note that

$$\frac{\alpha}{h}(u_I - u_{I-1}) \;=\; \frac{\alpha}{2h}(u_{I+1} - u_{I-1}) - \frac{\alpha}{2h}(u_{I-1} - 2u_I + u_{I+1}) \; . \tag{9.10}$$

Clearly, the second term on the right-hand side of (9.10) would contribute additional diffusion, as it corresponds to the centered difference Laplace operator with a *grid diffusion* constant $k_h = \frac{\alpha h}{2}$. This is not necessarily an undesirable feature, as it is well-known that the centered-difference method under-diffuses (that is, its convergence is from below in the appropriate energy norm, see Chapter 7).

It may be shown in the context of the one-dimensional convection-diffusion equation that there exists an optimal amount $k_{h,opt}$ of diffusion that can be added to the problem by way of upwinding to render the numerical solution exact at the nodes of a uniformly discretized domain. This is, in fact, given by $k_{h,opt} = \frac{\alpha h}{2}\left[\coth Pe_h - \frac{1}{Pe_h}\right]$.

The upwinding method can be also interpreted as a Petrov-Galerkin method in which the weighting functions for a given node are preferentially weighing its upwind domain, see Figure 9.5 for a schematic depiction. To further appreciate this point, write the weak form of the steady convection-diffusion equation as

$$\int_0^L w(\alpha u_{,x} - \epsilon u_{,xx})\, dx \;=\; 0 \; . \tag{9.11}$$

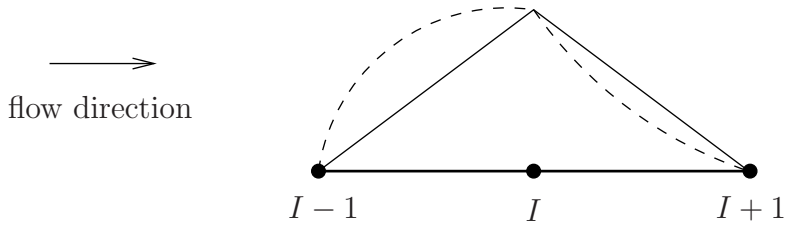


Figure 9.5. *A schematic depiction of the upwind Petrov-Galerkin method for the convection-diffusion equation (continuous line: Bubnov-Galerkin, broken line: Petrov-Galerkin)*

where the weighting function w satisfies $w_1(0) = w_1(L) = 0$. Using integration by parts, the weak form (9.11) can be rewritten as

$$\int_0^L w\alpha u_{,x}\, dx \; + \int_0^L w_{,x}\, \epsilon u_{,x}\, dx \;=\; 0\;. \tag{9.12}$$

Recalling now that upwinding can be interpreted as introducing artificial diffusion with constant k_h, one may modify the discrete counterpart of (9.12) so that it takes the form

$$\int_0^L w_h \alpha u_{h,x}\, dx \; + \int_0^L w_{h,x}\, (\epsilon + k_h) u_{h,x}\, dx \;=\; 0\;. \tag{9.13}$$

The sum of the convective and upwinding contributions in (9.13) may be expressed as

$$\int_0^L (w_h \alpha u_{h,x} + w_{h,x}\, k_h u_{h,x}\,)\, dx \;=\; \int_0^L \tilde{w}_h \alpha u_{h,x}\, dx\;, \tag{9.14}$$

where $\tilde{w}_h$ is the new weighting function for the convective part of the convection-diffusion equation in the spirit of the Petrov-Galerkin method. In this case, $\tilde{w}_h = w_h + \frac{k_h}{\alpha} w_{h,x}$.

Multi-dimensional generalizations of upwind finite element methods need special attention. This is because upwinding should only be effected in the direction of the flow (*streamline upwinding*). This is because the introduction of diffusion in directions other than the flow direction (*crosswind diffusion*) generates excessive errors.

9.2 Linear elastodynamics

The problem of linear elastostatics described in detail in Section 7.2 is extended here to include the effects of inertia. The resulting equations of motion take the form

$$\begin{aligned} \nabla\cdot\boldsymbol{\sigma} + \mathbf{f} &= \rho\mathbf{a} && \text{in } \Omega\times I\ , \\ \boldsymbol{\sigma}\mathbf{n} &= \bar{\mathbf{t}} && \text{on } \Gamma_q\times I\ , \\ \mathbf{u} &= \bar{\mathbf{u}} && \text{on } \Gamma_u\times I\ , \\ \mathbf{u}(x_1,x_2,x_3,0) &= \mathbf{u}_0(x_1,x_2,x_3) && \text{in } \Omega\ , \\ \mathbf{v}(x_1,x_2,x_3,0) &= \mathbf{v}_0(x_1,x_2,x_3) && \text{in } \Omega\ , \end{aligned} \tag{9.15}$$

where $\mathbf{u}=\mathbf{u}(x_1,x_2,x_3,t)$ is the unknown time-dependent displacement field, $\rho(>0)$ is the mass density, $\mathbf{a}(=\ddot{\mathbf{u}})$ is the acceleration, and $I=(0,T]$, with $T>0$ being a given end-time. Also, $\mathbf{u}_0$ and $\mathbf{v}_0$ are the prescribed initial displacement and velocity fields. Clearly, equations $(9.15)_{2,3}$ are two sets of time-dependent boundary conditions on Γ_q and Γ_u, respectively, which are assumed to hold throughout the time interval I. Likewise, two sets of initial conditions are set in $(9.15)_{4,5}$ for the whole domain Ω at time $t=0$. The strong form of the resulting initial/boundary-value problem is stated as follows: given functions $\mathbf{f}$, ρ, $\bar{\mathbf{t}}$, $\bar{\mathbf{u}}$, $\mathbf{u}_0$ and $\mathbf{v}_0$, as well as a constitutive equation (7.5) for $\boldsymbol{\sigma}$, find $\mathbf{u}$ in $\Omega\times I$, such that the equations (9.15) are satisfied.

A Galerkin-based weak form of the linear elastostatics problem has been derived in Section 7.2. In the elastodynamics case, the only substantial difference involves the inclusion of the term $\int_\Omega \mathbf{w}\cdot\rho\ddot{\mathbf{u}}\,d\Omega$, as long as one adopts the semi-discrete approach. As a result, the weak form at a given time can be expressed as

$$\int_\Omega \mathbf{w}\cdot\rho\mathbf{a}\,d\Omega + \int_\Omega \nabla_s\mathbf{w}:\boldsymbol{\sigma}\,d\Omega = \int_\Omega \mathbf{w}\cdot\mathbf{f}\,d\Omega + \int_{\Gamma_q}\mathbf{w}\cdot\bar{\mathbf{t}}\,d\Gamma\ . \tag{9.16}$$

The admissible displacement and test function fields are defined respectively as

$$\begin{aligned} \mathcal{U} = \big\{\mathbf{u}\in\mathbf{H}^1(\Omega\times I) \quad | \quad \mathbf{u}=\bar{\mathbf{u}} \text{ on } \Gamma_u\times I \quad , \quad \mathbf{u}(x_1,x_2,x_3,0)=\mathbf{u}_0\ , \\ \dot{\mathbf{u}}(x_1,x_2,x_3,0)=\mathbf{v}_0\big\}\ , \end{aligned} \tag{9.17}$$

and

$$\begin{aligned} \mathcal{W} = \big\{\mathbf{w}\in\mathbf{H}^1(\Omega\times I) \quad | \quad \mathbf{w}=\mathbf{0} \text{ on } \Gamma_u\times I \quad , \quad \mathbf{w}(x_1,x_2,x_3,0)=\mathbf{0}\ , \\ \dot{\mathbf{w}}(x_1,x_2,x_3,0)=\mathbf{0}\big\}\ , \end{aligned} \tag{9.18}$$

Following the development in Section 7.2, the discrete counterpart of (9.16) can be written in matrix form as

$$\int_\Omega [\mathbf{w}_h]^T \rho [\mathbf{a}_h]\, d\Omega + \int_\Omega \langle \boldsymbol{\epsilon}(\mathbf{w}_h)\rangle^T [\mathbf{D}] \langle \boldsymbol{\epsilon}(\mathbf{u}_h)\rangle \, d\Omega \;=\; \int_\Omega [\mathbf{w}_h]^T [\mathbf{f}]\, d\Omega + \int_{\Gamma_q} [\mathbf{w}_h]^T [\bar{\mathbf{t}}]\, d\Gamma \;, \quad (9.19)$$

where, in addition to (7.31),

$$[\mathbf{a}_h] \;=\; [\mathbf{N}^e][\mathbf{a}^e] \;. \quad (9.20)$$

A Galerkin approximation of (9.19) at the element level using the nomenclature of (7.30-7.43) leads to a system of ordinary differential equations of the form

$$[\mathbf{M}^e][\mathbf{a}^e] + [\mathbf{K}^e][\mathbf{u}^e] \;=\; [\mathbf{F}^e] + [\mathbf{F}^{\text{int},e}] \;, \quad (9.21)$$

where all quantities have already been defined in Section 7.2 except for the element mass matrix $[\mathbf{M}^e]$ which is given by

$$[\mathbf{M}^e] \;=\; \int_{\Omega^e} [\mathbf{N}^e]^T \rho [\mathbf{N}^e]\, d\Omega \;. \quad (9.22)$$

Clearly, the mass matrix is symmetric and also positive-definite. Following a standard procedure elaborated upon in Section 6.2, the contribution of the forcing vector $[\mathbf{F}^{\text{int},e}]$ due to interelement tractions is neglected upon assembly of the global equations. As a result, the equations (9.21) give rise to their assembled counterparts in the form

$$[\mathbf{M}][\hat{\mathbf{a}}] + [\mathbf{K}][\hat{\mathbf{u}}] \;=\; [\mathbf{F}] \;, \quad (9.23)$$

where $\hat{\mathbf{u}}$ and $\hat{\mathbf{a}}$ are the global unknown displacement and acceleration vectors, respectively.[1] The preceding differential equations are, of course, subject to initial conditions that can be written in vectorial form as $[\hat{\mathbf{u}}(0)] = [\hat{\mathbf{u}}_0]$ and $[\hat{\mathbf{v}}(0)] = [\hat{\mathbf{v}}_0]$.

The most commonly employed method for the numerical solution of the system of coupled linear second-order ordinary differential equations (9.23) is the *Newmark*[2] *method*. This is based on a time series expansion of $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}} = \dot{\hat{\mathbf{u}}}$. With reference to the time interval $(t_n, t_{n+1}]$, the Newmark method is defined by the equations

$$\begin{aligned} [\hat{\mathbf{u}}_{n+1}] \;&=\; [\hat{\mathbf{u}}_n] + [\hat{\mathbf{v}}_n]\Delta t_n + \frac{1}{2}\Big\{(1-2\beta)[\hat{\mathbf{a}}_n] + 2\beta[\hat{\mathbf{a}}_{n+1}]\Big\}\Delta t_n^2 \;, \\ [\hat{\mathbf{v}}_{n+1}] \;&=\; [\hat{\mathbf{v}}_n] + \Big\{(1-\gamma)[\hat{\mathbf{a}}_n] + \gamma[\hat{\mathbf{a}}_{n+1}]\Big\}\Delta t_n \;, \end{aligned} \quad (9.24)$$

[1] Note that the overhead "hat" symbol is used to distinguish between the vector field $\mathbf{u}$ and the solution vector $\hat{\mathbf{u}}$ emanating from the finite element approximation of the vector field $\mathbf{u}$.

[2] Nathan M. Newmark (1910-1981) was an American structural engineer.

where $\Delta t_n = t_{n+1} - t_n$, $[\hat{\mathbf{a}}] = [\hat{\ddot{\mathbf{u}}}]$, and β, γ are parameters chosen such that

$$0 \le \beta \le 0.5 \qquad , \qquad 0 < \gamma \le 1 \ . \tag{9.25}$$

The special case $\beta = 0.25$, $\gamma = 0.5$ corresponds to the trapezoidal rule. Indeed, in this case equations $(9.24)_{2,1}$ reduce to

$$\begin{aligned}
[\hat{\mathbf{v}}_{n+1}] &= [\hat{\mathbf{v}}_n] + \frac{1}{2}\Big\{[\hat{\mathbf{a}}_n] + [\hat{\mathbf{a}}_{n+1}]\Big\}\Delta t_n \ , \\
[\hat{\mathbf{u}}_{n+1}] &= [\hat{\mathbf{u}}_n] + [\hat{\mathbf{v}}_n]\Delta t_n + \frac{1}{4}\Big\{[\hat{\mathbf{a}}_n] + [\hat{\mathbf{a}}_{n+1}]\Big\}\Delta t_n^2 \\
&= [\hat{\mathbf{u}}_n] + [\hat{\mathbf{v}}_n]\Delta t_n + \frac{1}{2}\Big\{[\hat{\mathbf{v}}_{n+1}] - [\hat{\mathbf{v}}_n]\Big\}\Delta t_n \\
&= [\hat{\mathbf{u}}_n] + \frac{1}{2}\Big\{[\hat{\mathbf{v}}_n] + [\hat{\mathbf{v}}_{n+1}]\Big\}\Delta t_n \ ,
\end{aligned} \tag{9.26}$$

where $(9.26)_1$ is used in deriving $(9.26)_3$.

Likewise, the special case $\beta = 0$, $\gamma = 0.5$ corresponds to the centered-difference rule, provided the time step remains constant. To appreciate this point, first write (9.24) for this case as

$$\begin{aligned}
[\hat{\mathbf{u}}_{n+1}] &= [\hat{\mathbf{u}}_n] + [\hat{\mathbf{v}}_n]\Delta t_n + \frac{1}{2}[\hat{\mathbf{a}}_n]\Delta t_n^2 \ , \\
[\hat{\mathbf{v}}_{n+1}] &= [\hat{\mathbf{v}}_n] + \frac{1}{2}\big([\hat{\mathbf{a}}_n] + [\hat{\mathbf{a}}_{n+1}]\big)\Delta t_n \ .
\end{aligned} \tag{9.27}$$

Next, solve $(9.27)_2$ for $[\hat{\mathbf{a}}_{n+1}]$ and substitute $[\hat{\mathbf{a}}_n]$ with its equal from $(9.27)_1$,

$$\begin{aligned}
[\hat{\mathbf{a}}_{n+1}] &= \frac{2}{\Delta t_n}\big([\hat{\mathbf{v}}_{n+1}] - [\hat{\mathbf{v}}_n]\big) - [\hat{\mathbf{a}}_n] \\
&= \frac{2}{\Delta t_n}\big([\hat{\mathbf{v}}_{n+1}] - [\hat{\mathbf{v}}_n]\big) - \frac{2}{\Delta t_n^2}\big([\hat{\mathbf{u}}_{n+1}] - [\hat{\mathbf{u}}_n] - [\hat{\mathbf{v}}_n]\Delta t_n\big) \\
&= -\frac{2}{\Delta t_n^2}\big([\hat{\mathbf{u}}_{n+1}] - [\hat{\mathbf{u}}_n]\big) + \frac{2}{\Delta t_n}[\hat{\mathbf{v}}_{n+1}] \ .
\end{aligned} \tag{9.28}$$

The corresponding relation for the time domain $(t_{n-1}, t_n]$ takes the form

$$[\hat{\mathbf{a}}_n] = -\frac{2}{\Delta t_{n-1}^2}\big([\hat{\mathbf{u}}_n] - [\hat{\mathbf{u}}_{n-1}]\big) + \frac{2}{\Delta t_{n-1}}[\hat{\mathbf{v}}_n] \ . \tag{9.29}$$

Upon substituting $[\hat{\mathbf{a}}_n]$ from (9.29) to $(9.27)_1$ and setting $\Delta t_{n-1} = \Delta t_n = \Delta t$, it follows that

$$[\hat{\mathbf{v}}_n] = \frac{[\hat{\mathbf{u}}_{n+1}] - [\hat{\mathbf{u}}_{n-1}]}{2\Delta t} \ . \tag{9.30}$$

Finally, substituting $[\hat{\mathbf{v}}_n]$ from (9.30) to (9.29) leads to

$$[\hat{\mathbf{a}}_n] \;=\; \frac{[\hat{\mathbf{u}}_{n+1}] - 2[\hat{\mathbf{u}}_n] + [\hat{\mathbf{u}}_{n-1}]}{\Delta t^2} \;. \tag{9.31}$$

Equations (9.30) and (9.31) are indeed the standard first- and second-order centered-difference formulae.

It is clear that the Newmark equations (9.24) define a two-parameter family of time integrators. It is important to divide this family of integrators into two distinct categories, namely implicit and explicit integrators, corresponding to $\beta > 0$ and $\beta = 0$, respectively.

The general *implicit Newmark method* may be implemented as follows: First, solve $(9.24)_1$ for $[\hat{\mathbf{a}}_{n+1}]$, namely write

$$[\hat{\mathbf{a}}_{n+1}] \;=\; \frac{1}{\beta\Delta t_n^2}\big([\hat{\mathbf{u}}_{n+1}] - [\hat{\mathbf{u}}_n] - [\hat{\mathbf{v}}_n]\Delta t_n\big) - \frac{1-2\beta}{2\beta}[\hat{\mathbf{a}}_n] \;. \tag{9.32}$$

Then, substitute (9.32) into the semi-discrete form (9.23) evaluated at t_{n+1} to find that

$$\left\{\frac{1}{\beta\Delta t_n^2}[\mathbf{M}] + [\mathbf{K}]\right\}[\hat{\mathbf{u}}_{n+1}] \;=\; [\mathbf{F}_{n+1}] + [\mathbf{M}]\left\{\big([\hat{\mathbf{u}}_n] + [\hat{\mathbf{v}}_n]\Delta t_n\big)\frac{1}{\beta\Delta t_n^2} + \frac{1-2\beta}{2\beta}[\hat{\mathbf{a}}_n]\right\} \;. \tag{9.33}$$

After solving (9.33) for $[\hat{\mathbf{u}}_{n+1}]$, one may compute the acceleration $[\hat{\mathbf{a}}_{n+1}]$ from (9.32) and the velocity $[\hat{\mathbf{v}}_{n+1}]$ from $(9.24)_2$. It can be shown that the implicit Newmark method is unconditionally stable for this problem regardless of the values of β and γ within the range in (9.25).

The general *explicit Newmark method* may be implemented as follows: First, since $\beta = 0$, the displacements $[\hat{\mathbf{u}}_{n+1}]$ are immediately computed from $(9.24)_1$ independently of the accelerations $[\hat{\mathbf{a}}_{n+1}]$. Next, taking the semi-discrete equations (9.23) at t_{n+1}, one may substitute $[\hat{\mathbf{u}}_{n+1}]$ from $(9.24)_1$ to find that

$$[\mathbf{M}][\hat{\mathbf{a}}_{n+1}] \;=\; -[\mathbf{K}]\big([\hat{\mathbf{u}}_n] + [\hat{\mathbf{v}}_n]\Delta t_n + \frac{1}{2}[\hat{\mathbf{a}}_n]\Delta t_n^2\big) + [\mathbf{F}_{n+1}] \;. \tag{9.34}$$

If $[\mathbf{M}]$ is rendered diagonal (see discussion in Chapter 8), then $[\hat{\mathbf{a}}_{n+1}]$ can be determined without solving any coupled linear algebraic equations. Finally, the velocity $[\hat{\mathbf{v}}_{n+1}]$ is computed from $(9.24)_2$.

Before proceeding with the stability analysis of the Newmark method, recall, by way of background, that the general solution of the homogeneous counterpart of (9.23) is of the form

$$[\hat{\mathbf{u}}] \;=\; \sum_{j=1}^{N} c_j e^{i\omega_j t}[\mathbf{\Phi}_j] \;, \tag{9.35}$$

where N is the dimension of the vector $[\hat{\mathbf{u}}]$ and c_j are constants, while $\mathbf{\Phi}_j$ are N-dimensional vectors and ω_j are scalar parameters to be determined. Substituting a typical vector of (9.35) into the homogeneous counterpart of (9.23) leads to

$$\sum_{j=1}^{N} e^{i\omega_j t} \left\{ [\mathbf{K}] - \omega_j^2 [\mathbf{M}] \right\} [\mathbf{\Phi}_j] \;=\; [\mathbf{0}] \;. \tag{9.36}$$

It follows from (9.36) that the eigenpairs $(\omega_j^2, \mathbf{\Phi}_j)$, $j = 1, 2, \ldots, N$, are extracted from the eigenvalue problem

$$\left([\mathbf{K}] - \omega_j^2 [\mathbf{M}] \right) [\mathbf{\Phi}_j] \;=\; [\mathbf{0}] \;. \tag{9.37}$$

Setting

$$[\mathbf{\Phi}] \;=\; \left[[\mathbf{\Phi}_1] \; [\mathbf{\Phi}_2] \ldots [\mathbf{\Phi}_N] \right] \;, \tag{9.38}$$

the preceding eigenvalue problem can be expressed as

$$[\mathbf{M}][\mathbf{\Phi}][\mathbf{\Omega}] \;=\; [\mathbf{K}][\mathbf{\Phi}] \;, \tag{9.39}$$

where $\mathbf{\Omega}$ is a diagonal $N \times N$ matrix that contains all eigenvalues ω_j^2, $j = 1, 2, \ldots, N$.

The solution of the non-homogeneous problem (9.23) may be obtained by variation of parameters, according to which the solution vector is written as

$$[\hat{\mathbf{u}}] \;=\; [\mathbf{\Phi}][\hat{\mathbf{y}}] \;, \tag{9.40}$$

where $[\hat{\mathbf{y}}] = [\hat{\mathbf{y}}(t)]$ is an N-dimensional vector to be determined. Substituting (9.40) into (9.23) results in

$$[\mathbf{M}][\mathbf{\Phi}][\hat{\ddot{\mathbf{y}}}] + [\mathbf{K}][\mathbf{\Phi}][\hat{\mathbf{y}}] \;=\; [\mathbf{F}] \;. \tag{9.41}$$

Upon premultiplying the preceding equation by $[\mathbf{\Phi}]^T$, one finds that

$$[\mathbf{\Phi}]^T [\mathbf{M}][\mathbf{\Phi}][\hat{\ddot{\mathbf{y}}}] + [\mathbf{\Phi}]^T [\mathbf{K}][\mathbf{\Phi}][\hat{\mathbf{y}}] \;=\; [\mathbf{\Phi}]^T [\mathbf{F}] \;. \tag{9.42}$$

Appealing to the standard diagonalization property already derived for the parabolic case in Section 8.1, the equations (9.42) are decoupled and written as

$$m_j \ddot{y}_j + k_j y_j \;=\; g_j \quad , \quad j = 1, 2, \ldots, N \;, \tag{9.43}$$

where $g_j = [\mathbf{\Phi}_j]^T [\mathbf{F}]$.

The stability of the explicit Newmark method ($\beta = 0$) can be investigated for the homogeneous scalar equation

$$m\ddot{u} + ku \;=\; 0 \;, \tag{9.44}$$

derived from (9.43) by a simple change in notation. In this case, the explicit Newmark equations (9.24) may be written as

$$\begin{aligned} u_{n+1} &= u_n + v_n \Delta t_n + \frac{1}{2} a_n \Delta t_n^2 \\ v_{n+1} &= v_n + [(1-\gamma) a_n + \gamma a_{n+1}] \Delta t_n \end{aligned} \tag{9.45}$$

or, taking into account (9.44) at t_n and t_{n+1},

$$\begin{aligned} u_{n+1} &= u_n + v_n \Delta t_n + \frac{1}{2} \{ -\frac{k}{m} u_n \} \Delta t_n^2 \\ v_{n+1} &= v_n - \frac{k}{m} \big[(1-\gamma) u_n + \gamma u_{n+1} \big] \Delta t_n \\ &= v_n - \frac{k}{m} \Big[(1-\gamma) u_n + \gamma \big\{ u_n + v_n \Delta t_n + \frac{1}{2} \{ -\frac{k}{m} u_n \} \Delta t_n^2 \big\} \Big] \Delta t_n \ . \end{aligned} \tag{9.46}$$

Equations (9.46) can be put in matrix form as

$$\left[\begin{array}{c} u_{n+1} \\ v_{n+1} \end{array} \right] = \left[\begin{array}{cc} 1 - \frac{1}{2}\alpha & \Delta t_n \\ -\frac{\alpha}{\Delta t_n} + \frac{1}{2}\gamma \frac{\alpha^2}{\Delta t_n} & 1 - \gamma\alpha \end{array} \right] \left[\begin{array}{c} u_n \\ v_n \end{array} \right] \ , \tag{9.47}$$

where $\alpha = \frac{k}{m} \Delta t_n^2$.

It is easy to show that the stability of the explicit Newmark method depends on the spectral properties of the *amplification matrix* $[\mathbf{r}]$, defined with reference to (9.47) as

$$[\mathbf{r}] = \left[\begin{array}{cc} 1 - \frac{1}{2}\alpha & \Delta t_n \\ -\frac{\alpha}{\Delta t_n} + \frac{1}{2}\gamma \frac{\alpha^2}{\Delta t_n} & 1 - \gamma\alpha \end{array} \right] \ . \tag{9.48}$$

Specifically, for the method to be stable both eigenvalues of $[\mathbf{r}]$ need to less than or equal to one in absolute value. Here, these eigenvalues are given by

$$\lambda_{1,2} = 1 - \frac{1}{2} \left[(\frac{1}{2} + \gamma)\alpha \pm \sqrt{ \Big\{ (\frac{1}{2} + \gamma)\alpha \Big\}^2 - 4\alpha } \right] \ . \tag{9.49}$$

For the most common case of $\gamma = 0.5$, the eigenvalues of the amplification matrix reduce to

$$\lambda_{1,2} = 1 - \frac{1}{2} \left[\alpha \pm \sqrt{\alpha^2 - 4\alpha} \right] \ , \tag{9.50}$$

which is easily shown to be less than or equal to one in absolute value if $\alpha \leq 4$. This, in turn, implies that the critical step-size Δt_{cr} for explicit Newmark with $\gamma = 0.5$ is

$$\Delta t_{cr} = \frac{2}{\sqrt{\frac{k}{m}}} \ . \tag{9.51}$$

In a multi-dimensional setting, the preceding condition becomes

$$\Delta t_{cr} \;=\; \frac{2}{\max_j \sqrt{\frac{k_j}{m_j}}} \;, \tag{9.52}$$

where k_j and m_j are the diagonalized stiffness and mass components, respectively. Clearly, condition (9.52) places restrictions on the step-size Δt_n and, therefore, dictates the cost of the explicit computations. As in the parabolic problem of Section 8.2, a simple scaling argument shows that the stiffness is of order $o(1)$ and the mass is of order $o(h^2)$, hence the critical step in (9.52) is of order $o(h)$, where h is a linear measure of mesh size.

Following an analogous procedure, it can be shown that the implicit Newmark method is unconditionally stable provided that

$$0.5 \;\leq\; \gamma \;\leq\; 2\beta \;. \tag{9.53}$$

Otherwise, stability is conditional and the critical time step equals

$$\Delta t_{cr} \;=\; \frac{1/\sqrt{\gamma/2-\beta}}{\sqrt{\frac{k}{m}}} \;. \tag{9.54}$$

An extension to the multi-dimensional case is obtained along the lines of (9.52).

9.3 Exercises

Problem 1

Write a program to solve the one-dimensional convection-diffusion equation

$$u_{,t} \;+\; 10u_{,x} \;=\; 0.1u_{,xx}$$

in the domain $(0,1)$, with periodic boundary conditions $u(0,t) = u(1,t)$ and $u'(1,t) = 0$, and initial condition

$$u(x,0) \;=\; \begin{cases} 0 & \text{for } x \in (0,0.45) \\ 20(x-0.45) & \text{for } x \in [0.45,0.5] \\ 20(0.55-x) & \text{for } x \in (0.5,0.55] \\ 0 & \text{for } x \in (0.55,1) \end{cases} \;.$$

For the spatial discretization, use linear finite elements with 2-point integration and upwinding with optimal grid diffusion. For the temporal discretization, use the forward Euler method with constant time step-size $\Delta t = h^2/6$, where h is the length of a typical element (the choice of the step-size is critical for stability). Run the program for uniform meshes of 20 and 200 elements and plot the distribution of the dependent variable u at times $t = 0.01$, 0.05, and 0.1.

Hint: You may enforce the periodic boundary condition on the left end either by time-lagging, namely setting $u(0,t_{n+1}) = u(1,t_n)$, or by direct coupling between the unknowns $u(0,t)$ and $u(1,t)$.

Index