

ME280A

Introduction to the Finite Element Method

Panayiotis Papadopoulos

Department of Mechanical Engineering
University of California, Berkeley

Copyright ©2005 by Panayiotis Papadopoulos

Contents

1	INTRODUCTION TO THE FINITE ELEMENT METHOD	1
1.1	Historical perspective: the origins of the finite element method	1
1.2	Introductory remarks on the concept of discretization	2
1.2.1	Structural analogue substitution method	3
1.2.2	Finite difference method	4
1.2.3	Finite element method	5
1.3	Classifications of partial differential equations	6
1.4	Suggestions for further reading	9
2	MATHEMATICAL PRELIMINARIES	11
2.1	Linear function spaces, operators and functionals	11
2.2	Continuity and differentiability	15
2.3	Inner products, norms and completeness	16
2.3.1	Inner products	16
2.3.2	Norms	16
2.3.3	Banach spaces	18
2.3.4	Linear operators and bilinear forms in Hilbert spaces	21
2.4	Background on variational calculus	23
2.5	Suggestions for further reading	28
3	METHODS OF WEIGHTED RESIDUALS	29
3.1	Introduction	29
3.2	Galerkin methods	32
3.3	Collocation methods	39
3.3.1	Point-collocation method	40
3.3.2	Subdomain-collocation method	44

3.4	Least-squares methods	46
3.5	Composite methods	48
3.6	An interpretation of finite-difference methods	48
3.7	Suggestions for further reading	52
4	VARIATIONAL METHODS	53
4.1	Introduction to variational principles	53
4.2	Weak (variational) forms and variational principles	57
4.3	Rayleigh-Ritz method	61
4.4	Suggestions for further reading	66
5	CONSTRUCTION OF FINITE ELEMENT SUBSPACES	67
5.1	Introduction	67
5.2	Finite element spaces	74
5.3	Completeness property	78
5.4	Basic finite element shapes in one, two and three dimensions	81
5.4.1	One dimension	81
5.4.2	Two dimensions	82
5.4.3	Three dimensions	82
5.4.4	Higher dimensions	82
5.5	Polynomial shape functions	83
5.5.1	Interpolations in one dimension	83
5.5.2	Interpolations in two dimensions	89
5.5.3	Interpolations in three dimensions	100
5.6	The concept of isoparametric mapping	104
6	COMPUTER IMPLEMENTATION OF FINITE ELEMENT METHODS	113
6.1	Numerical integration of element matrices	113
6.2	Assembly of global element arrays	118
6.3	Algebraic equation solving by Gaussian elimination and its variants	121
6.4	Finite element modeling: mesh design and generation	123
6.4.1	Symmetry	124
6.4.2	Optimal node numbering	124
6.5	Computer program organization	125

7	ELLIPTIC DIFFERENTIAL EQUATIONS	127
7.1	The Laplace equation in two dimensions	127
7.2	Linear elastostatics	127
7.2.1	A Galerkin approximation to the weak form	132
7.2.2	On the order of numerical integration	136
7.2.3	The patch test	140
7.3	Best approximation property of the finite element method	142
7.4	Error sources and estimates	145
7.5	Application to incompressible elastostatics and Stokes' flow	148
8	PARABOLIC DIFFERENTIAL EQUATIONS	155
8.1	Standard semi-discretization methods	156
8.2	Stability of classical time integrators	162
8.3	Weighted-residual interpretation of classical time integrators	165
9	HYPERBOLIC DIFFERENTIAL EQUATIONS	167
9.1	The one-dimensional convection-diffusion equation	167
9.2	Linear elastodynamics	172

DRAFT

List of Figures

1.1	<i>An infinite degree-of-freedom system</i>	3
1.2	<i>A simple example of the structural analogue method</i>	3
1.3	<i>The finite difference method in one dimension</i>	4
1.4	<i>A two-dimensional domain and its finite element subdomains</i>	5
1.5	<i>A two-dimensional domain and its boundary element subdomains</i>	6
2.1	<i>Schematic depiction of a set</i>	11
2.2	<i>Example of a set that does not form a linear space</i>	12
2.3	<i>Mapping between two sets</i>	13
2.4	<i>A function of class $C^0(0, 2)$</i>	15
2.5	<i>Distance between two points in the classical Euclidean sense</i>	17
2.6	<i>A linear operator mapping \mathcal{U} to \mathcal{V}</i>	22
2.7	<i>A bilinear form on $\mathcal{U} \times \mathcal{V}$</i>	23
2.8	<i>A functional exhibiting a minimum, maximum or saddle point at $u = u^*$</i>	24
3.1	<i>An open and connected domain Ω with smooth boundary written as the union of boundary regions $\partial\Omega_i$</i>	29
3.2	<i>The domain Ω of the Laplace-Poisson equation with Dirichlet boundary Γ_u and Neumann boundary Γ_q</i>	32
3.3	<i>The point-collocation method</i>	40
3.4	<i>The point collocation method in a square domain</i>	42
3.5	<i>The subdomain collocation method</i>	44
3.6	<i>Lagrangian interpolation functions used in the weighted-residual interpretation of the finite difference method</i>	49
4.1	<i>Piecewise linear interpolations functions in one dimension</i>	64
4.2	<i>Comparison of exact and approximate solutions</i>	65

5.1	<i>A finite element mesh</i>	75
5.2	<i>A finite element-based interpolation function</i>	76
5.3	<i>Finite element vs. exact domain</i>	76
5.4	<i>Error in the enforcement of Dirichlet boundary conditions due to the difference between the exact and the finite element domain</i>	77
5.5	<i>A potential violation of the integrability (compatibility) requirement</i>	78
5.6	<i>Pascal triangle</i>	81
5.7	<i>Finite element domains in one dimension</i>	82
5.8	<i>Finite element domains in two dimensions</i>	82
5.9	<i>Finite element domains in three dimensions</i>	82
5.10	<i>Linear element interpolations in one dimension</i>	84
5.11	<i>Standard quadratic element interpolations in one dimension</i>	84
5.12	<i>Hierarchical quadratic element interpolations in one dimension</i>	86
5.13	<i>Hermitian interpolation functions in one dimension</i>	87
5.14	<i>A three-noded triangular element</i>	89
5.15	<i>Higher-order triangular elements</i>	91
5.16	<i>A transitional triangular element</i>	91
5.17	<i>Area coordinates in a triangular domain</i>	92
5.18	<i>Four-noded rectangular element</i>	93
5.19	<i>Three members of the serendipity family of rectangular elements</i>	94
5.20	<i>Pascal triangle for serendipity elements</i>	95
5.21	<i>Three members of the Lagrangian family of rectangular elements</i>	95
5.22	<i>Pascal triangle for Lagrangian elements</i>	96
5.23	<i>A general rectangular finite element domain</i>	96
5.24	<i>Rectangular finite elements made of two or four joined triangular elements</i>	97
5.25	<i>A simple potential 3- or 4-noded triangular element for the case $p = 2$</i>	98
5.26	<i>Illustration of violation of the integrability requirement for the 9- or 10-dof triangle for the case $p = 2$</i>	98
5.27	<i>12-dof triangular element for the case $p = 2$</i>	99
5.28	<i>Clough-Tocher triangular element for the case $p = 2$</i>	99
5.29	<i>The 4-noded tetrahedral element</i>	100
5.30	<i>The 10-noded tetrahedral element</i>	101
5.31	<i>The 6- and 15-noded pentahedral elements</i>	102
5.32	<i>The 8-noded hexahedral element</i>	103

5.33	<i>The 20- and 27-noded hexahedral elements</i>	103
5.34	<i>Schematic of a parametric mapping from Ω_{\square}^e to Ω^e</i>	105
5.35	<i>The 4-noded isoparametric quadrilateral</i>	105
5.36	<i>Geometric interpretation of one-to-one isoparametric mapping in the 4-noded quadrilateral</i>	108
5.37	<i>Convex and non-convex 4-noded quadrilateral element domains</i>	109
5.38	<i>Relation between area elements in the natural and physical domain</i>	109
5.39	<i>Isoparametric 6-noded triangle and 8-noded quadrilateral</i>	110
5.40	<i>Isoparametric 8-noded hexahedral element</i>	110
6.1	<i>Two-dimensional Gauss quadrature rules for $q_1, q_2 \leq 1$ (left), $q_1, q_2 \leq 3$ (center), and $q_1, q_2 \leq 5$ (right)</i>	116
6.2	<i>Integration rules in triangular domains for $q \leq 1$ (left) and $q \leq 2$ (right)</i>	117
6.3	<i>Finite element mesh depicting global node and element numbering, as well as global degree of freedom assignments</i>	120
6.4	<i>Profile of a typical finite element stiffness matrix (x denotes a non-zero entry)</i>	122
6.5	<i>Representative examples of symmetries in the domains of differential equations (corresponding symmetries in the boundary conditions, loading, and equations themselves are assumed)</i>	124
6.6	<i>Two possible ways of node numbering in a finite element mesh</i>	125
7.1	<i>The domain Ω of the linear elastostatics problem</i>	128
7.2	<i>Zero-energy modes for the 4-noded quadrilateral with 1×1 Gaussian quadrature</i>	138
7.3	<i>Zero-energy modes for the 8-noded quadrilateral with 2×2 Gaussian quadrature</i>	139
7.4	<i>Schematic of the patch test (form A)</i>	141
7.5	<i>Schematic of the patch test (form B)</i>	141
7.6	<i>Schematic of the patch test (form C)</i>	142
7.7	<i>Geometric interpretation of the best approximation property as a closest-point projection from u to \mathcal{U} in the sense of the energy norm</i>	145
7.8	<i>Illustration of volumetric locking in plane strain when using 3-noded triangular elements</i>	152
8.1	<i>Amplification factor r as a function of $\lambda\Delta t$ for forward Euler, backward Euler and the exact solution of the homogeneous counterpart of (8.18)</i>	164

9.1	<i>Finite element discretization for the one-dimensional convection-diffusion equation</i>	169
9.2	<i>Finite element solution for the one-dimensional convection-diffusion equation for $c = 0$</i>	169
9.3	<i>Finite element solution for the one-dimensional convection-diffusion equation for $c > 0$</i>	170
9.4	<i>A schematic depiction of the upwind Petrov-Galerkin method for the convection-diffusion equation (continuous line: Bubnov-Galerkin, broken line: Petrov-Galerkin)</i>	171

Introduction

This is a set of notes that has been 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 2005

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 (1870) and W. Ritz (1909) on variational methods and the weighted-residual approach taken by B. G. Galerkin (1915) and others form the theoretical framework to the finite element method. With a bit of a stretch, one may even claim that Schellbach's approximate solution to Plateau's problem (find a surface of minimum area enclosed by a given closed curve in three dimensions), which dates back to 1851 is a rudimentary application of the finite element method.

Most researchers agree that the era of the finite element method begins with a lecture presented in 1941 by R. Courant 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 Ray Clough, Jr. of the University of California at Berkeley, and

his colleagues essentially reinvented the finite element method as a natural extension of matrix structural analysis and published their first work in 1956. Clough himself attributes the introduction of the term “finite element” to M.J. Turner, one of his associates at that time. An apparently simultaneous effort by John Argyris at the University of London independently led to another successful introduction of the method. It should come as no surprise that, to a large extent, the finite element method appears to owe its reinvention to structural engineers. 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.

In the few years following its introduction to the engineering community, the finite element method has attracted the attention of applied mathematicians, particularly those interested in numerical solution of partial differential equations. In 1973, G. Strang 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 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 1.1. This is an infinite dimensional system with respect to the position of all of its particle points. If the same body 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)

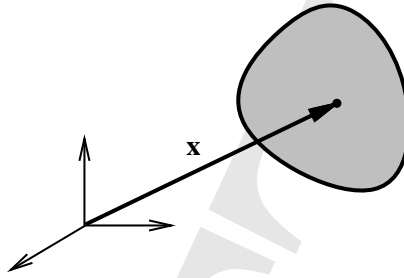


Figure 1.1: *An infinite degree-of-freedom system*

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.

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 1.2. 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).

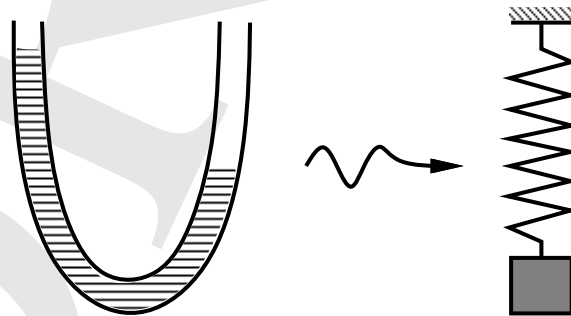


Figure 1.2: *A simple example of the structural analogue method*

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 Kron in the 1930s and 1940s is generally viewed as a typical example of the structural analogue approach.

1.2.2 Finite difference method

Consider the ordinary differential equation

$$k \frac{d^2 u}{dx^2} = f \text{ in } (0, L), \quad (1.1)$$

$$u(0) = u_0, \quad (1.2)$$

$$u(L) = u_L, \quad (1.3)$$

where k is a constant and $f = f(x)$ is a smooth function. Assume that N points are 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

$$\left. \frac{d^2 u}{dx^2} \right|_l \approx \frac{u_{l+1} - 2u_l + u_{l-1}}{\Delta x^2}, \quad (1.4)$$

with error $o(\Delta x^2)$. Indeed, employing twice a Taylor series expansion with remainder around

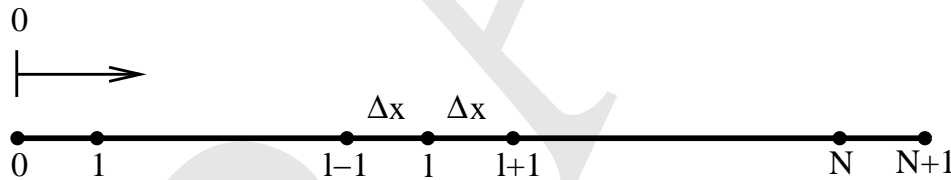


Figure 1.3: *The finite difference method in one dimension*

a typical point l in Figure 1.3, write

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

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

Adding the above equations results in

$$\left. \frac{d^2 u}{dx^2} \right|_l = \frac{u_{l+1} - 2u_l + u_{l-1}}{\Delta x^2} - \frac{\Delta x^2}{24} \left(\left. \frac{d^4 u}{dx^4} \right|_{l+\theta_1} + \left. \frac{d^4 u}{dx^4} \right|_{l-\theta_2} \right),$$

so that ignoring the second term of the right-hand side, the proposed approximation to the second derivative of u is recovered. Applying difference equation (1.4) to nodal points $1, 2, \dots, N$ and accounting for boundary conditions (1.2-3) gives rise to a system of N linear algebraic equations

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

with unknowns u_l , $l = 1, 2, \dots, N$. 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.

Remarks:

- The state equations are (approximately) satisfied only at discrete points $1, 2, \dots, N$.
- Serious difficulty exists in handling complex boundaries because of spatial regularity of the required grid of points.

1.2.3 Finite element method

Here, the domain Ω is divided (discretized) into connected elementary sub-domains (finite element domains) Ω_h^e , as shown in Figure 1.4. Thus, $\Omega \approx \Omega_h = \sum_e \Omega_h^e$.

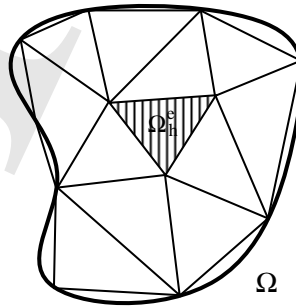


Figure 1.4: A two-dimensional domain and its finite element subdomains

A domain-based integral form of the state equations applies to each element and all elements are appropriately connected (assembled) with their neighboring elements to provide a global approximation of the original problem.

Remarks:

- State equations are satisfied (in an integral sense) over the whole domain with respect to a set of (simple) admissible functions.
- Boundary conditions are handled trivially.

The so-called boundary element method is a special finite element method, in which it is possible to write the state equations in boundary integral form. Consequently, the individual elements are situated at the boundary of the domain of analysis, as in Figure 1.5. The boundary element method is especially attractive for certain problems that involve infinite domains or local singularities. However, it does not have the broad applicability of the domain-based finite element method.

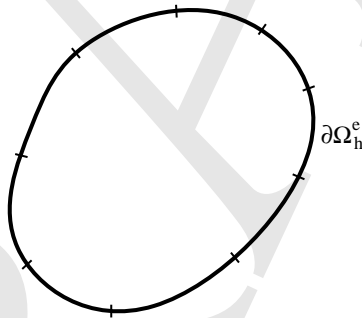


Figure 1.5: A two-dimensional domain and its boundary element subdomains

1.3 Classifications of partial differential equations

Consider a partial differential equation of the general form

$$F(x, y, \dots, u, u_x, u_y, \dots, u_{xx}, u_{xy}, u_{yy}, \dots) = 0, \quad (1.5)$$

where x, y, \dots are the independent variables, and $u = u(x, y, \dots)$ is the dependent variable.

Also,

$$u_{,x} = \frac{\partial u}{\partial x} \quad , \quad u_{,xx} = \frac{\partial^2 u}{\partial x^2} \quad , \quad \text{etc .}$$

Some useful definitions follow:

Order of a PDE: the order of the highest derivative of u in (1.5).

Linear vs. non-linear PDE: PDE is linear if F is linear in u and all of its derivatives, with coefficients depending on the independent variables x, y, \dots

Examples:

$$\begin{aligned} 3u_{,x} + u_{,y} - u &= 0 && \text{(linear - first order) ,} \\ xu_{,xx} + \frac{1}{y}u_{,yy} - 3u &= 0 && \text{(linear - second order) ,} \\ u_{,xx}^2 + u_{,yy} &= 0 && \text{(non-linear - second order) ,} \\ u_{,x} u_{,xxx}^2 + u_{,yy} &= 0 && \text{(non-linear - third order) .} \end{aligned}$$

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

$$au_{,xx} + bu_{,xy} + cu_{,yy} = d \quad , \quad (1.6)$$

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 (1.6) 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 phenomena (e.g., heat conduction, electrostatics), namely

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

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,

$$u_{,t} - ku_{,xx} = 0 \quad ; \quad k = k(x) \quad ,$$

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,

$$\alpha^2 u_{,xx} - u_{,tt} = 0 ,$$

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 (1.6) is not always an easy task. The elliptic, hyperbolic or parabolic nature of a partial differential equation is associated with the particular form of the characteristic curves. These are curves along which certain derivatives of a solution to the differential equation exhibit discontinuities.

The type of a partial differential equation determines the overall character of the expected solution and, to a great extent, dictates the choice of methodology used in its numerical approximation by the finite element or other methods.

Remarks:

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

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

The above equation is of hyperbolic type if $\epsilon = 0$ and $\alpha > 0$ (i.e., 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 its solution satisfies 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:

$$u_{,xx} + xu_{,yy} = 0 ,$$

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].

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].

DRAFT

Chapter 2

MATHEMATICAL PRELIMINARIES

2.1 Linear function spaces, operators and functionals

Consider a set \mathcal{V} whose members (typically called “points”) can be scalars, vectors or functions, 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.

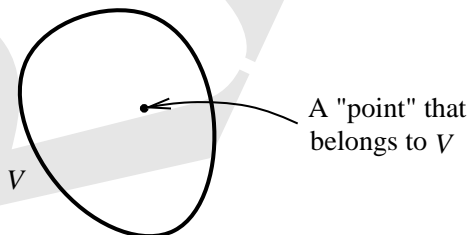


Figure 2.1: *Schematic depiction of a set*

A linear (or vector) space $\{\mathcal{V}, +; \mathbb{R}, \cdot\}$ is defined by the following properties for any $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathcal{V}$ 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) $\exists 0 \in \mathcal{V} \mid u + 0 = u$ (existence of null element),
- (iv) $\exists -u \in \mathcal{V} \mid u + (-u) = 0$ (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).

Examples:

- (a) $\mathcal{V} = P_2 := \{\text{all second degree polynomials } ax^2 + bx + c\}$ with the standard polynomial addition and scalar multiplication.

It can be trivially verified that $\{P_2, +; \mathbb{R}, \cdot\}$ is a linear function space. P_2 is also “equivalent” to an ordered triad $(a, b, c) \in \mathbb{R}^3$.

- (b) Define $\mathcal{V} = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$ with the standard addition and scalar multiplication for vectors. Notice that given $u = (x_1, y_1)$ and $v = (x_2, y_2)$ as in Figure 2.2,

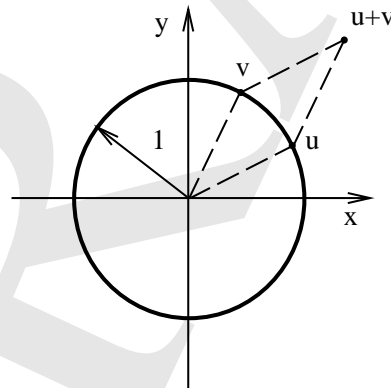


Figure 2.2: *Example of a set that does not form a linear space*

property (i) is violated, i.e., 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 sub-space of \mathcal{V} with respect to the same operations $(+)$ and (\cdot) , if, for any $u, v \in \mathcal{U}$ and $\alpha, \beta \in \mathbb{R}$

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

i.e., closure is maintained within \mathcal{U} .

Example:

- (a) Define the set P_n of all algebraic polynomials of degree smaller or equal to $n > 2$ and consider the linear space $\{P_n, +; \mathbb{R}, \cdot\}$ with the usual polynomial addition and scalar multiplication. Then, P_2 is a linear subspace of $\{P_n, +; \mathbb{R}, \cdot\}$.

Let \mathcal{U}, \mathcal{V} be two sets and define a *mapping* 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 .

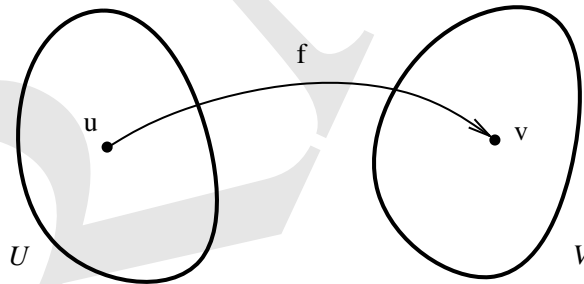


Figure 2.3: *Mapping between two sets*

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/vector points to a real number, i.e.

$$f : x \in \mathcal{U} \rightarrow f(x) \in \mathbb{R} \quad ; \quad \mathcal{U} = \mathbb{R}^n \quad , \quad n = 1, 2, \dots ,$$

- (2) *functional*: a mapping from a set with function points (namely points that correspond to functions) to the real numbers, i.e.

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

- (3) *operator*: a mapping from a set of functions to another set of functions, i.e.

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

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} as a simple function space. However, the distinction will be observed for didactic purposes.

Examples:

(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) $I[u] := \int_0^1 u(x)dx$ is a functional where u belongs to a function space, say $u(x) \in P_n$.

(c) $A[u] := \frac{d}{dx}u(x)$ 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 $A : \mathcal{U} \rightarrow \mathcal{V}$ is called linear, provided that, for all $u_1, u_2 \in \mathcal{U}$ and $\alpha, \beta \in \mathbb{R}$,

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

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 written as

$$A[u] = c ,$$

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 $f : \mathcal{U} \rightarrow \mathbb{R}$ 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

$$|f(x) - f(x_0)| < \epsilon,$$

provided that

$$|x - x_0| < \delta.$$

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.

Examples:

(a) The function $f : (0, 2) \rightarrow \mathbb{R}$ 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.

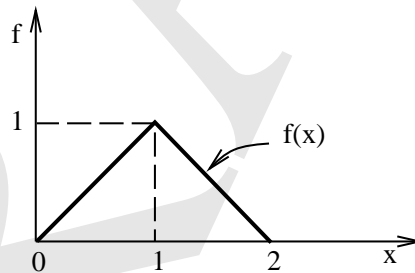


Figure 2.4: A function of class $C^0(0, 2)$

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

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

The “smoothness” of functions f plays a significant role in the proper construction of finite elements approximations.

2.3 Inner products, norms and completeness

2.3.1 Inner products

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

- (i) $\langle u + v, w \rangle = \langle u, w \rangle + \langle v, w \rangle$,
- (ii) $\langle u, v \rangle = \langle v, u \rangle$,
- (iii) $\langle \alpha \cdot u, v \rangle = \alpha \langle u, v \rangle$,
- (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} endowed with an inner product is called an *inner product space*. Further, two elements u, v of \mathcal{V} are *orthogonal* relative to the inner product $\langle \cdot, \cdot \rangle$ if $\langle u, v \rangle = 0$.

Example:

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

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

It is easy to show that the above mapping is an inner product on $\mathcal{V} \times \mathcal{V}$. This inner product-space is called the n -dimensional Euclidean vector space.

2.3.2 Norms

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 $d : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}_0^+$ as

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

We wish to establish a similar notion of proximity (“closeness”) between functions rather than merely between points of a Euclidean space. Moreover, we need to quantify the “largeness” of a function. The appropriate context for the above is provided by norms.

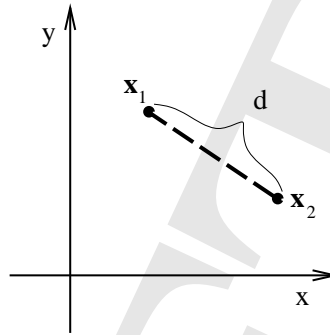


Figure 2.5: *Distance between two points in the classical Euclidean sense*

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

- (i) $\|u + v\| \leq \|u\| + \|v\|$ (triangular inequality),
- (ii) $\|\alpha \cdot u\| = |\alpha| \|u\|$,
- (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} endowed with a norm is called a *normed linear space* (NLS).

Examples:

- (a) Consider the n -dimensional Euclidean space \mathbb{R}^n and let point $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_n]^T \in \mathbb{R}^n$. Some standard norms in \mathbb{R}^n can be defined – for example,
 - the 1-norm: $\|\mathbf{x}\|_1 := \sum_{i=1}^n |x_i|$,
 - the 2-norm: $\|\mathbf{x}\|_2 := (\sum_{i=1}^n x_i^2)^{1/2}$,
 - 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 norms 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

$$d(u_n, u) := \|u_n - u\|.$$

We say that $u_n \rightarrow u \in \mathcal{U}$ if $\forall \epsilon > 0 \exists N(\epsilon)$, so that

$$d(u_n, u) < \epsilon \quad \forall n > N .$$

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 $\epsilon > 0 \exists N(\epsilon)$ such that

$$d(u_m, u_n) = \|u_m - u_n\| < \epsilon \quad \forall m, n > N .$$

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 necessarily true.

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

$$d(u, v) < r .$$

A subset \mathcal{V} of \mathcal{U} is termed *open* if, for each point $u \in \mathcal{V}$, there exists a neighborhood $\mathcal{N}_r(u)$ which is fully contained in \mathcal{V} .

Example:

- (a) Consider the set of real numbers \mathbb{R} equipped with the usual norm (i.e., the absolute value). The set \mathcal{V} defined as $\mathcal{V} := \{x \in \mathbb{R} \mid 0 < x < 1\} := (0, 1)$ is open.

2.3.3 Banach spaces

A linear space \mathcal{U} for which every Cauchy sequence converges to “point” $u \in \mathcal{U}$ is called a *complete* space. Complete normed linear spaces are also referred to as *Banach spaces*. Complete inner product spaces are called *Hilbert spaces*. Hilbert spaces form the proper functional context for the mathematical analysis of finite element methods. The basic goal of the 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.

Hilbert spaces are also Banach spaces, while the opposite is generally not true. Indeed, the inner product of a Hilbert space induces an associated norm (called the “natural norm”) given by

$$\|u\| := \langle u, u \rangle^{1/2} .$$

To prove that $\langle u, u \rangle^{1/2}$ is actually a norm, it is sufficient to show that the three defining properties of a norm hold. Properties (ii) and (iii) are easily verified using the fact that $\langle \cdot, \cdot \rangle$ is an inner product, i.e. for (ii)

$$\|\alpha \cdot u\| = \langle \alpha \cdot u, \alpha \cdot u \rangle^{1/2} = (\alpha^2 \langle u, u \rangle)^{1/2} = |\alpha| \|u\| ,$$

and for (iii)

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

Property (i) (i.e., the triangular inequality) merits more attention: in order to prove that it holds, we make use of the Cauchy-Schwartz inequality, which states that for any $u, v \in \mathcal{V}$

$$|\langle u, v \rangle| \leq \|u\| \|v\| . \quad (2.1)$$

To prove (2.1), first notice that it trivially holds as equality if $u = 0$ or $v = 0$. Then, define a function $F : \mathbb{R} \rightarrow \mathbb{R}_0^+$ as

$$F(\lambda) := \|u + \lambda \cdot v\|^2 ; \quad \lambda \in \mathbb{R} ,$$

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

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

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

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

inequality

$$\langle u, v \rangle^2 - \|u\|^2 \|v\|^2 \leq 0$$

must hold, thus yielding (2.1).

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

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

which implies that the triangular inequality holds.

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

$$L_2(\Omega) := \left\{ u : \Omega \rightarrow \mathbb{R} \mid \int_{\Omega} u^2 d\Omega < \infty \right\} .$$

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

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

$$H^m(\Omega) := \left\{ u : \Omega \rightarrow \mathbb{R} \mid D^\alpha u \in L_2(\Omega) \quad \forall \alpha \leq m \right\} ,$$

where

$$D^\alpha u := \frac{\partial^\alpha u}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \dots \partial x_n^{\alpha_n}} , \quad \alpha = \alpha_1 + \alpha_2 + \dots + \alpha_n$$

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

$$\langle u, v \rangle_{H^m(\Omega)} := \int_{\Omega} \left\{ \sum_{\alpha=0}^m \sum_{\beta=\alpha} D^\beta u D^\beta v \right\} d\Omega ,$$

and the corresponding natural norm as

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

Example:

(a) 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 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, defined on \mathbb{R}^n by the property

$$\int_{\Omega} \delta(x_1, x_2, \dots, x_n) f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n := f(0, 0, \dots, 0)$$

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

Given that a function belongs to $H^m(\Omega)$, it is important to obtain an estimate of its smoothness on the boundary $\partial\Omega$ of the domain of analysis. Denoting the outward normal to the boundary by n , define (to within some technicalities) the fractional space $H^{m-j-1/2}(\partial\Omega)$ as

$$H^{m-j-1/2}(\partial\Omega) := \{ \phi \in L_2(\partial\Omega) \mid \exists u \in H^m(\Omega) \mid \gamma_j u = \phi \text{ on } \partial\Omega \} ,$$

where the trace operator $\gamma_j : H^m(\Omega) \mapsto L_2(\partial\Omega)$ is given by

$$\gamma_j := \frac{\partial^j u}{\partial n^j} , \quad 0 \leq j \leq m - 1 .$$

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 these theorem states that given an open set $\Omega \subset \mathbb{R}^n$ with sufficiently smooth boundary, and $C_b^k(\Omega)$ is the space of bounded functions of class $C^k(\Omega)$, then

$$H^m(\Omega) \subset C_b^k(\Omega) ,$$

if, and only if, $m > k + n/2$. Setting $m = 2$, $k = 1$ and $n = 1$, it follows from the above 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

Consider a linear operator $A : \mathcal{U} \mapsto \mathcal{V}$, $v = A[u]$, where \mathcal{U} , \mathcal{V} are Hilbert spaces, as in Figure 2.6. Some important definitions follow:

A is *bounded* if there exists an $M > 0$ 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.

A is (uniformly) *continuous* if for any $\epsilon > 0$ and any $u_1, u_2 \in \mathcal{U}$, there is a $\delta = \delta(\epsilon)$ such that $\|A[u_1] - A[u_2]\|_{\mathcal{V}} < \epsilon$ for $\|u_1 - u_2\|_{\mathcal{U}} < \delta$.

It is easy to show that, in the context of linear operators, boundedness implies (uniform) continuity and vice-versa (i.e., the two properties are equivalent).

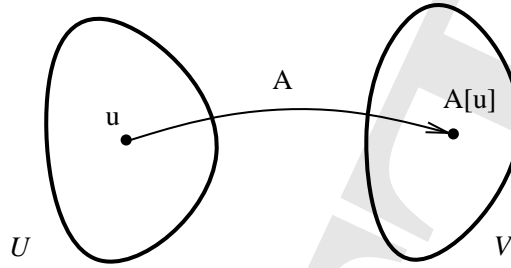


Figure 2.6: A linear operator mapping \mathcal{U} to \mathcal{V}

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

$$\langle u_1, A[u_2] \rangle = \langle A[u_1], u_2 \rangle ,$$

for all $u_1, u_2 \in \mathcal{U}$.

Example:

- (a) Let $U = \mathbb{R}^n$ 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

$$\langle \mathbf{x}, A[\mathbf{y}] \rangle := \mathbf{x}^T \mathbf{A} \mathbf{y} .$$

Then

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

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 *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

$$\langle A[u], v \rangle = \langle u, A^*[v] \rangle ,$$

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 inverse is not true.

Define an operator $B : \mathcal{U} \times \mathcal{V} \mapsto \mathbb{R}$ as in Figure 2.7, 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 *continuous* if there is an $M > 0$ such that for all $u \in \mathcal{U}$ and $v \in \mathcal{V}$

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

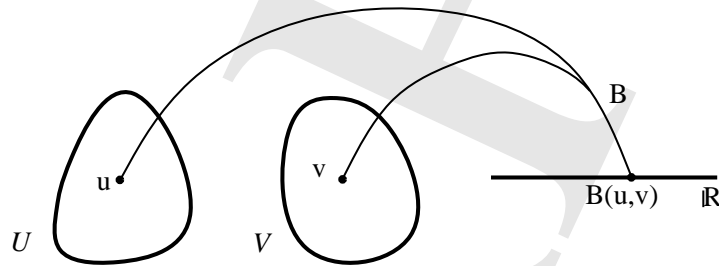


Figure 2.7: A bilinear form on $\mathcal{U} \times \mathcal{V}$

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

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

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

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

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 Chapter 4 of the notes. 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 $u = u(x, y, \dots)$ that can play the role of the dependent variable in a partial differential equation. The variation δu of u is a function defined on the same domain as u and represents admissible changes to the function u . Thus, if $\Omega \subset \mathbb{R}^n$ is the domain of $u \in \mathcal{U}$ with boundary $\partial\Omega$, where

$$\mathcal{U} := \left\{ u \in H^1(\Omega) \mid u = \bar{u} \text{ on } \partial\Omega \right\},$$

then δu is an *arbitrary* function that belongs to \mathcal{U}_0 , where

$$\mathcal{U}_0 := \left\{ u \in H^1(\Omega) \mid u = 0 \text{ on } \partial\Omega \right\}.$$

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 (i.e., minimum, maximum or a saddle point), as schematically indicated in Figure 2.8.

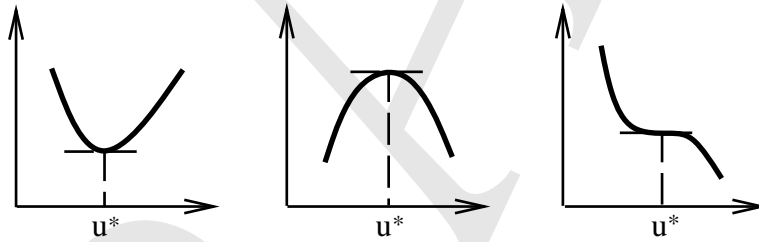


Figure 2.8: A functional exhibiting a minimum, maximum or saddle point at $u = u^*$

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

$$\delta I[u] := \lim_{w \rightarrow 0} \frac{I[u + w\delta u] - I[u]}{w}, \quad (2.2)$$

and, by induction, the k -th variation as

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

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

$$I[u + \delta u] = I[u] + \delta I[u] + \frac{1}{2!} \delta^2 I[u] + \frac{1}{3!} \delta^3 I[u] + \dots$$

Examples:

(a) Let I be defined on \mathbb{R}^n as

$$I[\mathbf{u}] := \frac{1}{2} \mathbf{u} \cdot \mathbf{A} \mathbf{u} - \mathbf{u} \cdot \mathbf{b} , \quad (2.3)$$

where \mathbf{A} is an $n \times n$ symmetric positive-definite matrix and \mathbf{b} belongs to \mathbb{R}^n . Using (2.2), it follows that

$$\delta I[\mathbf{u}] = \delta \mathbf{u} \cdot \mathbf{A} \mathbf{u} - \delta \mathbf{u} \cdot \mathbf{b}$$

and

$$\delta^2 I[\mathbf{u}] = \delta \mathbf{u} \cdot \mathbf{A} \delta \mathbf{u} .$$

Therefore, it is seen that 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

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

which coincides with the minimum of $I[\mathbf{u}]$. 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 (2.2). Thus,

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

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

and

$$\delta^k I[u] = 0 , \quad k = 3, 4, \dots .$$

Using the alternative definition 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.

Remarks:

- In the variation of $I[u]$, the independent variables x, y, \dots 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.
- 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, i.e. for $u = u(x)$, then

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

assuming continuity of $\frac{du}{dx}$, 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, \dots , then the variation of I obviously depends on the variations of all u, v, \dots , i.e.

$$\delta I[u, v, \dots] := \lim_{\omega \rightarrow 0} \frac{I[u + \omega \delta u, v + \omega \delta v, \dots] - I[u, v, \dots]}{\omega} ,$$

and

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

or, alternatively,

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

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

$$\delta I[u, u', u'', \dots] := \lim_{\omega \rightarrow 0} \frac{I[u + \omega \delta u, u' + \omega \delta u', u'' + \omega \delta u'', \dots] - I[u, u', u'', \dots]}{\omega}.$$

Note that if u^* is the function that extremizes $I[u]$, then for any variation δu around u^*

$$I[u^* + \delta u] = I[u^*] + \delta I[u^*] + \frac{1}{2!} \delta^2 I[u^*] + \dots \quad (2.4)$$

Equation (2.4) implies that necessary and sufficient condition for extremization of I at $u = u^*$ is that

$$\delta I[u^*] = 0.$$

A weaker 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 $DI[u] \cdot v$). This is defined as

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

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:

- (a) 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

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

From a previous exercise it has been concluded that

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

2.5 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 and connected set $\Omega \subset \mathbb{R}^n$ with boundary $\partial\Omega$, 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 , $i = 1, \dots, k$, involving traces γ_j with $j < p$ are defined on appropriate boundary function spaces. Further, the boundary $\partial\Omega$ is assumed to possess a unique outer unit normal vector \mathbf{n} at every point, and is decomposed (arbitrarily at present) into k parts $\partial\Omega_i$, $i = 1, \dots, k$, such that

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

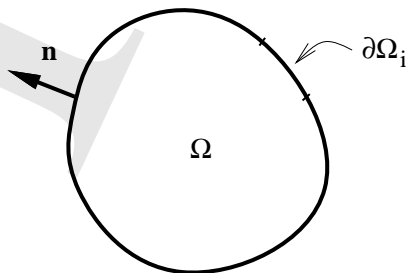


Figure 3.1: An open and connected domain Ω with smooth boundary written as the union of boundary regions $\partial\Omega_i$

Given functions f and g_i , $i = 1, \dots, k$, on Ω and $\partial\Omega_i$, respectively, a mathematical problem associated with a partial differential equation is described by the system

$$A[u] = f \quad \text{in } \Omega, \quad (3.1)$$

$$B_i[u] = g_i \quad \text{on } \partial\Omega_i, \quad i = 1, \dots, k. \quad (3.2)$$

With reference to equations (3.1-3.2), define functions w_Ω and w_i , $i = 1, \dots, k$, on Ω and $\partial\Omega_i$, respectively, such that the scalar quantity R , given by

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

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

Equations (3.1-3.2) constitute the *strong form* of the differential equation. The scalar equation

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

where functions w_Ω and w_i , $i = 1, \dots, k$, 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 (3.1-3.2) implies the general weighted-residual form. The converse is also true, conditional upon sufficient smoothness of the fields involved. 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 $\Omega \subset \mathbb{R}^n$. Show that

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

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

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

and without loss of generality, let $f_0 > 0$. It follows that, since f is continuous, there is an open “sphere” \mathcal{N} of radius $\delta > 0$ centered at \mathbf{x}_0 and defined by

$$\|\mathbf{x} - \mathbf{x}_0\| < \delta ,$$

such that

$$|f(\mathbf{x}) - f(\mathbf{x}_0)| < \epsilon = \frac{f_0}{2} \quad (3.4)$$

when \mathbf{x} is located anywhere inside \mathcal{N} . Thus, it is seen from (3.4) that

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

everywhere in \mathcal{N} , hence

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

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 (3.1-2) and (3.3), note that since the latter holds for arbitrary choices of w_Ω and w_i , let

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

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

$$w_i = 0 \quad , \quad i = 1, \dots, k .$$

Invoking the localization theorem, it is readily concluded that (3.1) 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 (3.2) 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 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 $0, \dots, 2q - 1$.

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:

$$\frac{\partial}{\partial x_1} \left(k \frac{\partial u}{\partial x_1} \right) + \frac{\partial}{\partial x_2} \left(k \frac{\partial u}{\partial x_2} \right) = f \quad \text{in } \Omega, \quad (3.5)$$

$$-k \frac{\partial u}{\partial n} = \bar{q} \quad \text{on } \Gamma_q, \quad (3.6)$$

$$u = \bar{u} \quad \text{on } \Gamma_u, \quad (3.7)$$

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

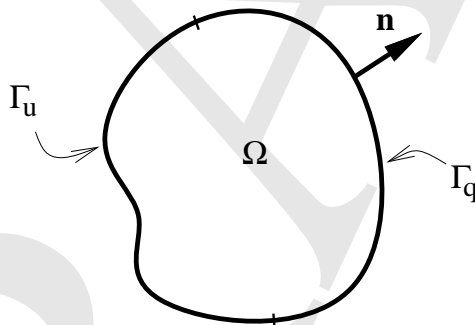


Figure 3.2: The domain Ω of the Laplace-Poisson equation with Dirichlet boundary Γ_u and Neumann boundary Γ_q

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 (i.e., they are known in advance). The boundary conditions (3.6) and (3.7) 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 *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

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

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) reads

$$\int_{\Omega} w_\Omega R_\Omega d\Omega + \int_{\Gamma_q} w_q R_q d\Gamma + \int_{\Gamma_u} w_u R_u d\Gamma = 0 , \quad (3.8)$$

and, 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 (3.8).

A series of assumptions are introduced in deriving the Galerkin method. First, assume that boundary condition (3.7) is satisfied at the outset, namely that the solution u is sought over a set of candidate functions that already satisfy (3.7). Hence, the third integral of the left-hand side of (3.8) vanishes and the choice of function w_u becomes irrelevant.

Observing that the two remaining integral terms in (3.8) 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

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

Substitution of the above expressions for the weighting functions into the reduced form of (3.8) yields

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

which, after integration by parts and use of the divergence theorem¹, is rewritten as

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

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

$$\frac{\partial u}{\partial n} := \frac{du}{d\mathbf{x}} \cdot \mathbf{n} = \frac{\partial u}{\partial x_1} n_1 + \frac{\partial u}{\partial x_2} n_2 ,$$

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

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

Here, an additional assumption is introduced, namely

$$w = 0 \quad \text{on } \Gamma_u . \quad (3.9)$$

This last assumption leads to the weighted residual equation

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

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

Alternatively, it is possible to assume that both (3.6) and (3.7) are satisfied at the outset and write the weighted residual statement for $w_{\Omega} = w$ as

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

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

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

which, in turn, becomes identical to (3.10) by imposing restriction (3.9) and making explicit use of the assumed condition (3.6).

¹This theorem states that given a closed smooth surface $\partial\Omega$ with interior Ω and a C^1 function $f : \Omega \rightarrow \mathbb{R}$, 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 $\partial\Omega$.

The weighted residual problem associated with equation (3.10) can be expressed operationally as follows: find $u \in \mathcal{U}$, such that for all $w \in \mathcal{W}$

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

where

$$\begin{aligned} \mathcal{U} &:= \left\{ u \in H^1(\Omega) \mid u = \bar{u} \text{ on } \Gamma_u \right\} , \\ \mathcal{W} &:= \left\{ w \in H^1(\Omega) \mid w = 0 \text{ on } \Gamma_u \right\} . \end{aligned}$$

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

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

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

$$(w, f) := \int_{\Omega} w f d\Omega$$

and

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

The identification of admissible solution fields \mathcal{U} and weighting function fields \mathcal{W} is dictated by restrictions placed during the derivation of (3.10) and by the requirement that the bi-linear form $B(w, u)$ be computable (i.e., that the integral be well-defined). Clearly, alternative definitions of \mathcal{U} and \mathcal{W} (with regards to smoothness) can also be acceptable.

A *Galerkin approximation* of (3.10) is obtained by restating the weighted-residual problem as follows: find $u_h \in \mathcal{U}_h$, such that for all $w_h \in \mathcal{W}_h$

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

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

$$u \approx u_h = \sum_{I=1}^N \alpha_I \varphi_I(x_1, x_2) + \varphi_0(x_1, x_2) , \quad (3.11)$$

$$w \approx w_h = \sum_{I=1}^N \beta_I \psi_I(x_1, x_2) . \quad (3.12)$$

In the above, $\varphi_I(x_1, x_2)$ and $\psi_I(x_1, x_2)$, $I = 1, 2, \dots, N$, are given functions (called *interpolation* or *basis functions*), which vanish on Γ_u , and $\varphi_0(x_1, x_2)$ is chosen so that u_h satisfy

boundary condition (3.7). Parameters $\alpha_I \in \mathbb{R}$, $I = 1, 2, \dots, N$, are to be determined by invoking (3.10), while parameters $\beta_I \in \mathbb{R}$, $I = 1, 2, \dots, N$, are arbitrary.

A *Bubnov-Galerkin* approximation is obtained from (3.11-3.12) by setting $\psi_I = \varphi_I$ for all $I = 1, 2, \dots, N$. This is the most popular version of the Galerkin method. Use of functions $\psi_I \neq \varphi_I$ yields a so-called *Petrov-Galerkin* approximation.

Substitution of u_h and w_h as defined in (3.11-3.12) into the weak form (3.10) results in

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

or, alternatively,

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

where

$$K_{IJ} := \int_{\Omega} \{ \psi_{I,1} \ \psi_{I,2} \} k \begin{Bmatrix} \varphi_{J,1} \\ \varphi_{J,2} \end{Bmatrix} d\Omega , \quad (3.13)$$

and

$$F_I := - \int_{\Omega} \psi_I f d\Omega - \int_{\Omega} \{ \psi_{I,1} \ \psi_{I,2} \} k \begin{Bmatrix} \varphi_{0,1} \\ \varphi_{0,2} \end{Bmatrix} d\Omega - \int_{\Gamma_q} \psi_I \bar{q} d\Gamma . \quad (3.14)$$

Since parameters β_I are arbitrary, it follows that

$$\sum_{J=1}^N K_{IJ} \alpha_J - F_I = 0 \quad , \quad I = 1, 2, \dots, N ,$$

or, in matrix form,

$$\mathbf{K}\boldsymbol{\alpha} = \mathbf{F} , \quad (3.15)$$

where \mathbf{K} is the $N \times N$ *stiffness matrix* with components given by (3.13), \mathbf{F} is the $N \times 1$ *forcing vector* with components as in (3.14), and $\boldsymbol{\alpha}$ is the $N \times 1$ vector of parameters α_I introduced in (3.11).

It is important to note that the Galerkin approximation (3.11-3.12) transforms the integro-differential equation (3.10) 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 (3.5-7) 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. 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

$$A[v] = f - A[w]$$

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

- It can be easily seen from (3.13) 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 (3.15) 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:

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

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

Hence, equation (3.10) takes the form

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

A one-parameter Bubnov-Galerkin approximation can be obtained by setting $N = 1$ in equations (3.11-3.12) and choosing

$$\varphi_0(x) = 0$$

and

$$\varphi_1(x) = x .$$

Substituting u_h and w_h into (\dagger) 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

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

Again, (\dagger) implies that

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

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 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.

In the remainder of this section, the Galerkin method is summarized in the context of the model problem

$$A[u] = f \quad \text{in } \Omega, \quad (3.16)$$

$$B[u] = g \quad \text{on } \Gamma_q, \quad (3.17)$$

$$u = \bar{u} \quad \text{on } \Gamma_u, \quad (3.18)$$

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

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

where the space of admissible solutions u satisfies (3.18) 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 (3.17), 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 of analysis. Starting from the general weighted-residual form given in (3.3), assume, as in the Galerkin method, that boundary condition (3.18) will be explicitly satisfied by the admissible functions u_h , and obtain the reduced form

$$\int_{\Omega} w_{\Omega}(A[u] - f) d\Omega + \int_{\Gamma_q} w_q(B[u] - g) d\Gamma = 0, \quad (3.19)$$

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

$$u(\mathbf{x}) \approx u_h(\mathbf{x}) = \sum_{I=1}^N \alpha_I \varphi_I(\mathbf{x}) + \varphi_o(\mathbf{x}), \quad (3.20)$$

with conditions on Γ_u set to $\varphi_I(\mathbf{x}) = 0$, $I = 1, \dots, N$, and $\varphi_o(\mathbf{x}) = \bar{u}$.

3.3.1 Point-collocation method

First, identify n interior points in Ω with coordinates \mathbf{x}_i , $i = 1, \dots, n$, and $N - n$ boundary points on Γ_q with coordinates \mathbf{x}_i , $i = n + 1, \dots, N$. These are referred to as domain and boundary collocation points, respectively, and are shown schematically in Figure 3.3.

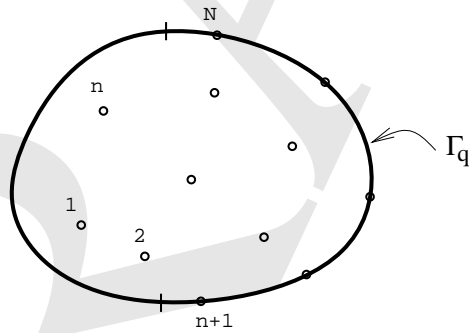


Figure 3.3: *The point-collocation method*

The interior and boundary weighting functions are respectively defined according to

$$w_{\Omega h}(\mathbf{x}) = \sum_{i=1}^n \beta_i \delta(\mathbf{x} - \mathbf{x}_i) \quad (3.21)$$

and

$$w_{qh}(\mathbf{x}) = \rho^2 \sum_{i=n+1}^N \beta_i \delta(\mathbf{x} - \mathbf{x}_i), \quad (3.22)$$

where the scalar parameter ρ is introduced in w_{qh} for unit consistency. Substitution of (3.20-22) into the weak form (3.19) yields

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

Recalling that A and B are linear in u , it follows that

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

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

$$\sum_{I=1}^N K_{iI} \alpha_I - F_i = 0 \quad , \quad i = 1, \dots, N \quad ,$$

where

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

and

$$F_i := \begin{cases} -A[\varphi_o(\mathbf{x}_i)] + f(\mathbf{x}_i) & , \quad 1 \leq i \leq n \quad , \\ -\rho^2 (B[\varphi_o(\mathbf{x}_i)] - g(\mathbf{x}_i)) & , \quad n+1 \leq i \leq N \end{cases} .$$

These equations are solved for parameters α_I , so that the approximate solution u_h is obtained from (3.20).

The particular choice of admissible fields renders the integrals in (3.19) well-defined, since products of Dirac-delta functions (from w_h) and smooth functions (from u_h) are always properly integrable.

Example:

Consider the partial differential equation

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

The domain of the problem is sketched in Figure 3.4. It is immediately concluded that the boundary $\partial\Omega$ 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.

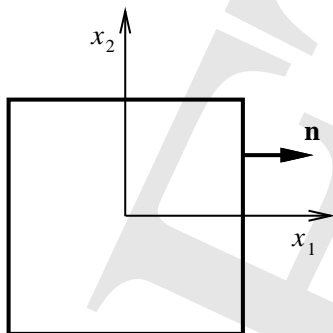


Figure 3.4: *The point collocation method in a square domain*

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, i.e., 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

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

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

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

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

$$\alpha_1 = 1/4 \quad , \quad \alpha_2 = 1/2,$$

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 $\Omega_i, i = 1, \dots, n$, and $\Gamma_{qi}, i = n + 1, \dots, N$, be mutually disjoint connected subsets of the domain Ω and the boundary Γ_q , respectively, as in Figure 3.5. It follows that

$$\bigcup_{i=1}^n \Omega_i \subset \Omega$$

and

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

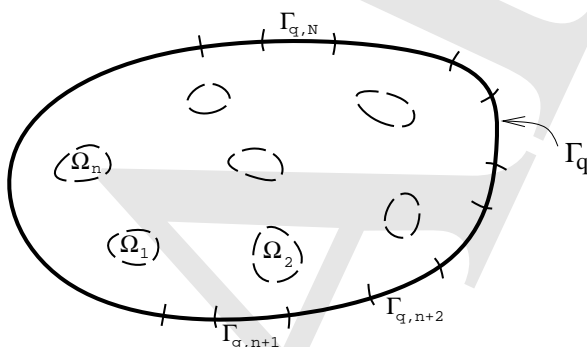


Figure 3.5: *The subdomain collocation method*

Recall the weighted residual form (3.19) and define the weighting function on Ω as

$$w_{\Omega h}(\mathbf{x}) = \sum_{i=1}^n \beta_i w_{\Omega_i}(\mathbf{x}) ,$$

with

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

Similarly, write on Γ_q

$$w_{qh}(\mathbf{x}) = \sum_{i=n+1}^N \beta_i w_{q_i}(\mathbf{x}) ,$$

with

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

Given the above weighting functions, the weighted-residual form (3.19) is rewritten as

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

Substitution of u_h from (3.20) into the above weak form yields

$$\begin{aligned} \sum_{i=1}^n \beta_i \int_{\Omega_i} \left(A \left[\sum_{I=1}^N \alpha_I \varphi_I(\mathbf{x}) + \varphi_o(\mathbf{x}) \right] - f \right) d\Omega \\ + \rho^2 \sum_{i=n+1}^N \beta_i \int_{\Gamma_{qi}} \left(B \left[\sum_{I=1}^N \alpha_I \varphi_I(\mathbf{x}) + \varphi_o(\mathbf{x}) \right] - g \right) d\Gamma = 0 . \end{aligned}$$

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

$$\begin{aligned} \sum_{i=1}^n \beta_i \int_{\Omega_i} \left(\sum_{I=1}^N \alpha_I A[\varphi_I(\mathbf{x})] + A[\varphi_o(\mathbf{x})] - f \right) d\Omega \\ + \rho^2 \sum_{i=n+1}^N \beta_i \int_{\Gamma_{qi}} \left(\sum_{I=1}^N \alpha_I B[\varphi_I(\mathbf{x})] + B[\varphi_o(\mathbf{x})] - g \right) d\Gamma = 0 , \end{aligned}$$

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

$$\sum_{I=1}^N K_{iI} \alpha_I - F_i = 0 , \quad i = 1, \dots, N ,$$

where

$$K_{iI} := \begin{cases} \int_{\Omega_i} A[\varphi_I(\mathbf{x}_i)] d\Omega , & 1 \leq i \leq n , \\ \rho^2 \int_{\Gamma_{qi}} B[\varphi_I(\mathbf{x}_i)] d\Gamma , & n+1 \leq i \leq N \end{cases} , \quad I = 1, \dots, N ,$$

and

$$F_i := \begin{cases} \int_{\Omega_i} \left(-A[\varphi_o(\mathbf{x}_i)] + f(\mathbf{x}_i) \right) d\Omega , & 1 \leq i \leq n , \\ \rho^2 \int_{\Gamma_{qi}} \left(-B[\varphi_o(\mathbf{x}_i)] + g(\mathbf{x}_i) \right) d\Gamma , & n+1 \leq i \leq N \end{cases} .$$

Remarks:

- The point-collocation method requires very small computational effort to form the stiffness matrix and the forcing vector.
- Collocation methods generally lead to unsymmetric stiffness matrices.
- Choice of collocation points is not arbitrary – for certain classes of differential equations, one may identify collocation points that yield optimal accuracy of the approximate solution.

3.4 Least-squares methods

Consider model problem (3.16-3.18) of the previous section, and assuming that (3.18) is satisfied by the space of admissible solutions u , form the “least-squares” functional $I[u]$, defined as

$$I[u] = \int_{\Omega} (A[u] - f)^2 d\Omega + \rho^2 \int_{\Gamma_q} (B[u] - g)^2 d\Gamma, \quad (3.23)$$

where ρ is a consistency parameter. Clearly, the functional attains an absolute minimum at the solution of (3.16-3.18). In order to obtain the extrema of the functional defined in (3.23), determine its first variation as

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

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

$$\begin{aligned} \delta A[u] &= \lim_{\omega \rightarrow 0} \frac{A[u + \omega \delta u] - A[u]}{\omega} \\ &= \lim_{\omega \rightarrow 0} \frac{A[u] + \omega A[\delta u] - A[u]}{\omega} = A[\delta u]. \end{aligned}$$

Consequently, extremization of $I[u]$ requires that

$$\int_{\Omega} A[\delta u](A[u] - f) d\Omega + \rho^2 \int_{\Gamma_q} B[\delta u](B[u] - g) d\Gamma = 0. \quad (3.24)$$

At this stage, introduce the finite-dimensional approximation for u_h as in (3.20), and, in addition, write

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

with $\delta\alpha_I$, $I = 1, \dots, N$, being arbitrary scalar parameters. Use of u_h and δu_h in (3.23) results in

$$\int_{\Omega} A\left[\sum_{I=1}^N \delta\alpha_I \varphi_I(\mathbf{x})\right] \left(A\left[\sum_{J=1}^N \alpha_J \varphi_J(\mathbf{x}) + \varphi_o(\mathbf{x})\right] - f\right) d\Omega \\ + \rho^2 \int_{\Gamma_q} B\left[\sum_{I=1}^N \delta\alpha_I \varphi_I(\mathbf{x})\right] \left(B\left[\sum_{J=1}^N \alpha_J \varphi_J(\mathbf{x}) + \varphi_o(\mathbf{x})\right] - g\right) d\Gamma = 0 .$$

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

$$\int_{\Omega} \sum_{I=1}^N \delta\alpha_I A[\varphi_I(\mathbf{x})] \left(\sum_{J=1}^N \alpha_J A[\varphi_J(\mathbf{x})] + A[\varphi_o(\mathbf{x})] - f\right) d\Omega \\ + \rho^2 \int_{\Gamma_q} \sum_{I=1}^N \delta\alpha_I B[\varphi_I(\mathbf{x})] \left(\sum_{J=1}^N \alpha_J B[\varphi_J(\mathbf{x})] + B[\varphi_o(\mathbf{x})] - g\right) d\Gamma = 0 ,$$

which, in turn, using the arbitrariness of $\delta\alpha_I$, gives rise to a system of linear algebraic equations of the form

$$\sum_{J=1}^N K_{IJ} \alpha_J - F_I = 0 \quad , \quad I = 1, \dots, N ,$$

where

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

and

$$F_I = \int_{\Omega} \left(A[\varphi_I] f - A[\varphi_I] A[\varphi_o]\right) d\Omega + \rho^2 \int_{\Gamma_q} \left(B[\varphi_I] g - B[\varphi_I] B[\varphi_o]\right) d\Gamma \quad , \quad I = 1, \dots, N .$$

It is important to note that the smoothness requirements for admissible fields u are governed by the integrals that appear in (3.23). It can be easily deduced that if A is a differential operator of second order (i.e., maps functions u to partial derivatives of second order), then for (3.23) to be well-defined, it is necessary that $u \in H^2(\Omega)$. This requirement can be contrasted with the one obtained in the Galerkin approximation of (3.5-3.7), 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 (3.19) is recovered as follows: choose $w_\Omega = A[w]$, $w_q = B[w]$ and set $w = 0$ on Γ_u . Then, the weak form in (3.24) is reproduced, where w appears in place of δu .

3.5 Composite methods

The Galerkin, collocation and least-squares methods can be appropriately combined to yield composite methods of approximation. 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 simple Galerkin/collocation method can be obtained for the model problem (3.16-18), with associated weighted-residual form (3.19), by defining the admissible solutions as in (3.20) and setting

$$w_\Omega(\mathbf{x}) \approx w_{\Omega h}(\mathbf{x}) = \sum_{I=1}^{m_1} \beta_I \psi_I(\mathbf{x}) + \sum_{I=m_1+1}^{n_1} \beta_I \rho^2 \delta(\mathbf{x} - \mathbf{x}_I),$$

In addition, on Γ_q ,

$$w_q(\mathbf{x}) \approx w_{qh}(\mathbf{x}) = \sum_{I=1}^{m_2} \beta_{I+n_1} \psi_I(\mathbf{x}) + \sum_{I=m_2+1}^{n_2} \beta_{I+n_1} \rho^2 \delta(\mathbf{x} - \mathbf{x}_I),$$

where ψ_I , $I = 1, \dots, N$ vanish on Γ_u and $n_1 + n_2 = N$.

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

$$w_\Omega(\mathbf{x}) \approx w_{\Omega h}(\mathbf{x}) = \sum_{I=1}^{m_1} \beta_I A[\varphi_I(\mathbf{x})] + \sum_{I=m_1+1}^{n_1} \beta_I \rho^2 \delta(\mathbf{x} - \mathbf{x}_I)$$

and

$$w_q(\mathbf{x}) \approx w_{qh}(\mathbf{x}) = \sum_{I=1}^{m_2} \beta_{I+n_1} B[\psi_I(\mathbf{x})] + \sum_{I=m_2+1}^{n_2} \beta_{I+n_1} \rho^2 \delta(\mathbf{x} - \mathbf{x}_I),$$

respectively, where, again, ψ_I , $I = 1, \dots, N$, vanish on Γ_u and $n_1 + n_2 = N$.

3.6 An interpretation of finite-difference methods

Finite-difference methods can be interpreted as 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 (1.1-1.3), and let grid points x_i , $i = 1, \dots, N$, be chosen in the interior of the domain $(0, L)$, as in Section 1.2.2. The system of equations

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

is obtained by applying the centered-difference operator

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

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

$$\int_{\Omega} w \left(k \frac{d^2 u}{dx^2} - f \right) d\Omega = 0, \quad (3.25)$$

where boundary conditions (1.2) and (1.3) are assumed to hold at the outset. Subsequently, define the approximate solution u_h within each sub-domain $(x_l - \frac{\Delta x}{2}, x_l + \frac{\Delta x}{2}]$, $l = 2, \dots, N-1$, as

$$u_h(x) = \sum_{i=l-1}^{l+1} N_i(x) \alpha_i, \quad (3.26)$$

where N_i are the standard Lagrangian polynomials of degree 2, namely,

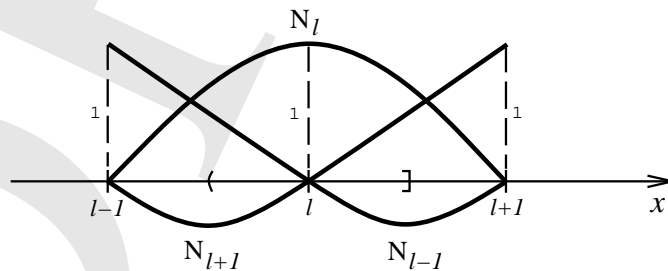


Figure 3.6: *Lagrangian interpolation functions used in the weighted-residual interpretation of the finite difference method*

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

Figure 3.6 illustrates the three interpolation functions in the representative sub-domain. It can be readily seen that $u_l = \alpha_l$, which implies that the parameters α_l can be interpreted as the values of the dependent variable at $x = x_l$.

Recalling the above remark, 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

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

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

The approximation for w_h over the domain is taken in the form

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

so that equation (3.25) is written as

$$\sum_{l=1}^N \beta_l \left(k \frac{d^2 u_h}{dx^2}(x_l) - f(x_l) \right) = 0. \quad (3.27)$$

It follows from (3.26) that at a representative sub-domain $(x_l - \frac{\Delta x}{2}, x_l + \frac{\Delta x}{2})$,

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

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

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

owing to the arbitrariness of parameters β_l , $l = 1, \dots, N$, in (3.27). Similarly, in sub-domain $[0, x_1 + \frac{\Delta x}{2}]$,

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

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

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

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

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.

Remark:

- The choice of admissible fields \mathcal{U}_h and \mathcal{W}_h is legitimate, since the integral on the left-hand side of (3.25) is always well-defined.

3.7 Suggestions for further reading

Sections 3.1-3.4

- [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 3.5

- [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].

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

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

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

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

where u is assumed continuously differentiable. Following the developments of Section 3.2, integration by parts and application of the divergence theorem on (4.2) yields

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

Recalling that

$$\frac{\partial u}{\partial x_1} n_1 + \frac{\partial u}{\partial x_2} n_2 = \frac{\partial u}{\partial n},$$

write

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

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

$$\frac{\partial}{\partial x_1} \left(k \frac{\partial u}{\partial x_1} \right) + \frac{\partial}{\partial x_2} \left(k \frac{\partial u}{\partial x_2} \right) = f \quad \text{in } \Omega$$

and

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

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

(i) Set

$$\delta u = 0 \quad \text{on } \partial\Omega.$$

This condition implies that the dependent variable u is prescribed throughout the boundary $\partial\Omega$. The space of admissible fields u is defined as

$$\mathcal{U} := \left\{ u \in H^1(\Omega) \mid u = \bar{u} \text{ on } \partial\Omega \right\},$$

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*).

(ii) Set

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

In this case, 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.

(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

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

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

$$\mathcal{U} := \left\{ u \in H^1(\Omega) \mid u = \bar{u} \text{ on } \Gamma_u \right\} .$$

It can be concluded from the above, with reference to (4.3) 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 (4.3) reveals that extremization of the functional in (4.1) yields a function u which satisfies the differential equation (3.5) and boundary conditions selected in conjunction to the space of admissible fields \mathcal{U} .

Following case (iii), note that the space of admissible variations \mathcal{U}_0 is defined as

$$\mathcal{U}_0 := \left\{ u \in H^1(\Omega) \mid u = 0 \text{ on } \Gamma_u \right\} = H_0^1(\Omega) .$$

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

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

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

$$\bar{I}[u] = \int_{\Omega} \left[\frac{k}{2} \left(\frac{\partial u}{\partial x_1} \right)^2 + \frac{k}{2} \left(\frac{\partial u}{\partial x_2} \right)^2 + fu \right] d\Omega + \int_{\Gamma_q} \bar{q}u \, d\Gamma , \quad (4.4)$$

where $\bar{q} = \bar{q}(x_1, x_2)$ is a continuous function on Γ_q .

Vanishing of the first variation of $\bar{I}[u]$ implies that

$$\int_{\Omega} \left[\frac{\partial u}{\partial x_1} k \frac{\partial \delta u}{\partial x_1} + \frac{\partial u}{\partial x_2} k \frac{\partial \delta u}{\partial x_2} + f \delta u \right] d\Omega + \int_{\Gamma_q} \bar{q} \delta u \, d\Gamma = 0 . \quad (4.5)$$

Equation (4.5) is termed the weak (variational) form of boundary-value problem (3.5-3.7). Comparing the above equation to (3.10), it is obvious that they are identical provided that the space of admissible field \mathcal{W} for w in (3.10) is identical to that of δu in (4.5).

The nature of the extremum point u (i.e., whether it renders $I[u]$ minimum, maximum or merely stationary) can be determined by means of the second variation of $I[u]$. Specifically, write

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

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

$$\delta^2 \bar{I}[u] \geq c \|\delta u\|^2,$$

where c is a positive constant.

The weak (variational) form of problem (3.5-3.7) can be stated operationally as follows: find $u \in \mathcal{U}$, such that for all $\delta u \in \mathcal{U}_0$

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

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

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

and the linear forms $(\delta u, f)$ and $(\delta u, \bar{q})_{\Gamma_q}$ are defined, respectively, as

$$(\delta u, f) := \int_{\Omega} \delta u f d\Omega$$

and

$$(\delta u, \bar{q})_{\Gamma_q} := \int_{\Gamma_q} \delta u \bar{q} d\Gamma.$$

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 (3.5-3.7).

In addition to the above weak (variational) form, there exists a variational *principle* associated with the solution u of problem (3.5-3.7). This can be stated as follows: find $u \in \mathcal{U}$, such that

$$\bar{I}[u] \leq \bar{I}[v],$$

for all $v \in \mathcal{U}$, where $\bar{I}[u]$ is defined in (4.4).

Remark:

- Directional derivatives can be used in deriving the weak (variational) equation (4.5) from functional $\bar{I}[u]$. Indeed, write

$$D_v \bar{I}[u] = 0 \Rightarrow \left[\frac{d}{d\omega} \bar{I}[u + \omega v] \right]_{\omega=0} = 0 \Rightarrow \int_{\Omega} \left[\frac{\partial u}{\partial x_1} k \frac{\partial v}{\partial x_1} + \frac{\partial u}{\partial x_2} k \frac{\partial v}{\partial x_2} + f v \right] d\Omega + \int_{\Gamma_q} \bar{q} v d\Gamma = 0,$$

where $v \in \mathcal{U}_0$.

4.2 Weak (variational) forms and variational principles

The analysis in Section 4.1, as applied to the model problem (3.5-3.7), provides an attractive perspective to the solution of certain partial differential equations: the solution is identified with a “point”, which minimizes an appropriately constructed functional over an admissible function space. Weak (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 weak (variational) forms and variational principles is not guaranteed: indeed, there exists no general method of constructing functionals $I[u]$, whose extremization recovers a desired weak (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 weak (variational) form to a functional extremization problem. If such an equivalence holds, the functional is referred to as a *potential*.

Theorem (Vainberg)

Consider a weak (variational) form

$$G(u, \delta u) := B(u, \delta u) + (f, \delta u) + (\bar{q}, \delta u)_{\Gamma_q} = 0 ,$$

where $u \in \mathcal{U}$, $\delta u \in \mathcal{U}_0$, and f and \bar{q} are independent of u . Assume that G possesses a Gâteaux derivative in a neighborhood \mathcal{N} of u , and the Gâteaux differential $D_{\delta u_1} B(u, \delta u_2)$ 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

$$D_{\delta u_1} G(u, \delta u_2) = D_{\delta u_2} G(u, \delta u_1) , \quad (4.6)$$

namely that $D_{\delta u_1} G(u, \delta u_2)$ be symmetric for all $\delta u_1, \delta u_2 \in \mathcal{U}_0$ and all $u \in \mathcal{N}$.

Preliminary to proving the above theorem, introduce the following two lemmas:

Lemma 1

Show that

$$D_v I[u] = \lim_{\Delta\omega \rightarrow 0} \frac{I[u + \Delta\omega v] - I[u]}{\Delta\omega} .$$

To prove that Lemma 1 holds, use the definition of the directional derivative of I in the direction v , so that

$$\begin{aligned} D_v I[u] &= \left\{ \frac{d}{d\omega} I[u + \omega v] \right\}_{\omega=0} \\ &= \left\{ \lim_{\Delta\omega \rightarrow 0} \frac{I[u + \omega v + \Delta\omega v] - I[u + \omega v]}{\Delta\omega} \right\}_{\omega=0} \\ &= \lim_{\Delta\omega \rightarrow 0} \left\{ \frac{I[u + \omega v + \Delta\omega v] - I[u + \omega v]}{\Delta\omega} \right\}_{\omega=0} \\ &= \lim_{\Delta\omega \rightarrow 0} \frac{I[u + \Delta\omega v] - I[u]}{\Delta\omega}. \end{aligned}$$

In the above derivation, note that operations $\frac{d}{d\omega}$ and $|_{\omega=0}$ are not interchangeable (as they both refer to the same variable ω), while $\lim_{\Delta\omega \rightarrow 0}$ and $|_{\omega=0}$ are interchangeable, conditional upon sufficient smoothness of $I[u]$.

Lemma 2 (Lagrange's formula)

Let $I[u]$ be a functional with Gâteaux derivatives everywhere, and $u, u + \delta u$ be any points of \mathcal{U} . Then,

$$I[u + \delta u] - I[u] = D_{\delta u} I[u + \epsilon \delta u] \quad , \quad 0 < \epsilon < 1 .$$

To prove Lemma 2, fix u and $u + \delta u$ in \mathcal{U} , and define function f on \mathbb{R} as

$$f(\omega) := I[u + \omega \delta u] .$$

It follows that

$$\begin{aligned} f' &:= \frac{df}{d\omega} = \lim_{\Delta\omega \rightarrow 0} \frac{f(\omega + \Delta\omega) - f(\omega)}{\Delta\omega} \\ &= \lim_{\Delta\omega \rightarrow 0} \frac{I[u + \omega \delta u + \Delta\omega \delta u] - I[u + \omega \delta u]}{\Delta\omega} = D_{\delta u} I[u + \omega \delta u] , \end{aligned}$$

where Lemma 1 was invoked. Then, using the standard mean-value theorem of calculus, write

$$\begin{aligned} I[u + \delta u] - I[u] &= f(1) - f(0) \\ &= \frac{f(1) - f(0)}{1 - 0} = f'(\epsilon) = D_{\delta u} I[u + \epsilon \delta u] , \end{aligned}$$

where $0 < \epsilon < 1$.

Given Lemma 2, one may proceed directly to the proof of Vainberg's theorem. First, prove the necessity of (4.6), namely that if an appropriate $I[u]$ exists, then (4.6) holds. To this end, fix $u \in \mathcal{N}$ and define the scalar quantity Δ as

$$\Delta := I[u + a\delta u_1 + b\delta u_2] - I[u + a\delta u_1] - I[u + b\delta u_2] + I[u],$$

where a, b are non-zero scalars, such that $u + \eta\delta u_1 + \theta\delta u_2 \in \mathcal{N}$, for all $0 \leq \eta \leq a$ and $0 \leq \theta \leq b$. Also, define functional $J[u]$ as

$$J[u] := I[u + a\delta u_1] - I[u],$$

so that

$$\Delta = J[u + b\delta u_2] - J[u].$$

Using Lemma 2 and the above definitions of $J[u]$ and Δ , write

$$\begin{aligned} \Delta = J[u + b\delta u_2] - J[u] &= D_{b\delta u_2}J[u + \epsilon_1 b\delta u_2] \\ &= D_{b\delta u_2}I[u + \epsilon_1 b\delta u_2 + a\delta u_1] - D_{b\delta u_2}I[u + \epsilon_1 b\delta u_2] \\ &= B(u + \epsilon_1 b\delta u_2 + a\delta u_1, b\delta u_2) + (f, b\delta u_2) + (\bar{q}, b\delta u_2)_{\Gamma_q} \\ &\quad - B(u + \epsilon_1 b\delta u_2, b\delta u_2) - (f, b\delta u_2) - (\bar{q}, b\delta u_2)_{\Gamma_q} \\ &= bB(a\delta u_1, \delta u_2) = abB(\delta u_1, \delta u_2). \end{aligned} \quad (4.7)$$

Alternatively, let functional $K[u]$ be defined as

$$K[u] := I[u + b\delta u_2] - I[u],$$

so that Δ is also written as

$$\Delta = K[u + a\delta u_1] - K[u].$$

Using the steps followed in the derivation of (4.7), it can be readily concluded that

$$\Delta = abB(\delta u_2, \delta u_1), \quad (4.8)$$

so that (4.7) and (4.8) lead to

$$B(\delta u_1, \delta u_2) = B(\delta u_2, \delta u_1).$$

Noting that, due to the linearity of B ,

$$\begin{aligned} D_{\delta u_1}B(u, \delta u_2) &= \left\{ \frac{d}{d\omega} B(u + \omega\delta u_1, \delta u_2) \right\}_{\omega=0} \\ &= \left\{ \frac{d}{d\omega} [B(u, \delta u_2) + \omega B(\delta u_1, \delta u_2)] \right\}_{\omega=0} = B(\delta u_1, \delta u_2), \end{aligned}$$

and, similarly,

$$D_{\delta u_2} B(u, \delta u_1) = B(\delta u_2, \delta u_1) ,$$

it follows that condition (4.6) holds.

In order to show the sufficiency of (4.6), namely prove that (4.6) implies the existence of an appropriate functional $I[u]$, define

$$I[u] := \int_0^1 B(tu, u) dt + (f, u) + (\bar{q}, u)_{\Gamma_q} . \quad (4.9)$$

Since

$$\frac{d}{d\omega} B(tu, u + \omega \delta u) = B(tu, \delta u) , \quad (4.10)$$

and

$$\frac{d}{d\omega} B(\omega \delta u, u + \omega \delta u) = B(\delta u, u) + 2\omega B(\delta u, \delta u) , \quad (4.11)$$

the directional derivative of $I[u]$ in the direction δu is written as

$$D_{\delta u} I[u] = D_{\delta u} \left(\int_0^1 B(tu, u) dt \right) + (f, \delta u) + (\bar{q}, \delta u)_{\Gamma_q} .$$

With the aid of (4.10) and (4.11), the first term on the right-hand side of the above equation may be rewritten as

$$\begin{aligned} D_{\delta u} \left(\int_0^1 B(tu, u) dt \right) &= \int_0^1 D_{\delta u} B(tu, u) dt \\ &= \int_0^1 \left\{ \frac{d}{d\omega} B(tu + t\omega \delta u, u + \omega \delta u) \right\}_{\omega=0} dt \\ &= \int_0^1 \left\{ \frac{d}{d\omega} B(tu, u + \omega \delta u) + \frac{d}{d\omega} B(t\omega \delta u, u + \omega \delta u) \right\}_{\omega=0} dt \\ &= \int_0^1 [B(tu, \delta u) + B(t\delta u, u)] dt . \end{aligned}$$

Exploiting the assumed symmetry of B , it follows from the above that

$$\begin{aligned} D_{\delta u} \int_0^1 B(tu, u) dt &= 2 \int_0^1 t B(u, \delta u) dt \\ &= 2B(u, \delta u) \left[\frac{t^2}{2} \right]_0^1 = B(u, \delta u) , \end{aligned}$$

which proves that $I[u]$, as defined in (4.9), is indeed an appropriate functional.

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 condition (4.6) is typically an easy task.
- Vainberg's theorem not only establishes a condition for potentiality of a weak form, but also provides a direct definition of the potential in the form of (4.9).

Example:

Recall the weak (variational) form of Section 3.2, which is associated with boundary-value problem (3.5-3.7). In this case,

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

$$(f, \delta u) := \int_{\Omega} f \delta u d\Omega ,$$

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 (4.9), is given by

$$I[u] = \frac{1}{2} B(u, u) + (f, u) + (\bar{q}, u)_{\Gamma_q}$$

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

whose extremization yield the above weak (variational) form.

4.3 Rayleigh-Ritz method

The Rayleigh-Ritz method provides approximate solutions to partial differential equations, whose weak (variational) form is derivable from a functional $I[u]$. 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

$$u \approx u_h = \sum_{I=1}^N \alpha_I \varphi_I + \varphi_o , \quad (4.12)$$

where φ_I , $I = 1, \dots, N$, is a specified family of interpolation functions that vanish where essential boundary conditions are enforced. In addition, function φ_o is defined so that u_h

satisfy identically the essential boundary conditions. Consequently, a proper N -dimensional subspace \mathcal{U}_h is completely defined by (4.12). Extremization of $I[u]$ over \mathcal{U}_h yields

$$\delta I[u_h] = \delta I\left[\sum_{I=1}^N \alpha_I \varphi_I + \varphi_o\right] = 0.$$

Instead of directly obtaining the weak (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 as a function of parameters α_I , $I = 1, \dots, N$, namely

$$\delta I[\alpha_1, \dots, \alpha_N] = 0. \quad (4.13)$$

It follows from (4.13) that $I[\alpha_1, \dots, \alpha_N]$ is an integral scalar function which attains an extremum over \mathcal{U}_h if, and only if,

$$\frac{\partial I}{\partial \alpha_1} \delta \alpha_1 + \frac{\partial I}{\partial \alpha_2} \delta \alpha_2 + \dots + \frac{\partial I}{\partial \alpha_N} \delta \alpha_N = 0.$$

Since the variations $\delta \alpha_I$, $I = 1, \dots, N$, are arbitrary, it may be immediately concluded that

$$\frac{\partial I}{\partial \alpha_I} = 0, \quad I = 1, \dots, N. \quad (4.14)$$

Equations (4.14) may be solved for parameters α_I , so that an approximate solution to the variational problem is expressed by means of (4.12).

Example:

Consider the functional $I[u]$ defined in the domain $(0, 1)$ as

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

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

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

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

$$\begin{aligned} I[u_2] &= \int_0^1 \left[\frac{1}{2} (\alpha_1 + 2\alpha_2 x)^2 + (\alpha_1 x + \alpha_2 x^2) \right] dx + 2(\alpha_1 + \alpha_2) \\ &= \frac{1}{2} \alpha_1^2 + \alpha_1 \alpha_2 + \frac{2}{3} \alpha_2^2 + \frac{5}{2} \alpha_1 + \frac{7}{3} \alpha_2. \end{aligned}$$

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

$$\varphi_1(x) = \begin{cases} 2x & \text{if } 0 \leq x \leq 0.5 \\ 2(1 - x) & \text{if } 0.5 < x \leq 1 \end{cases}$$

and

$$\varphi_2(x) = \begin{cases} 0 & \text{if } 0 \leq x \leq 0.5 \\ 2(x - \frac{1}{2}) & \text{if } 0.5 < x \leq 1 \end{cases},$$

where functions φ_1 and φ_2 are depicted in Figure 4.1. Then, $I[\tilde{u}_2]$ is written as

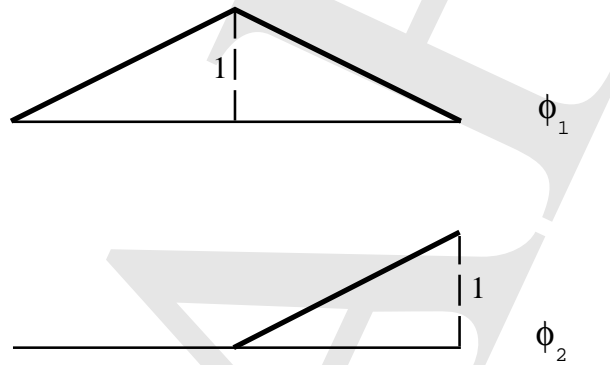


Figure 4.1: *Piecewise linear interpolations functions in one dimension*

$$\begin{aligned} I[\tilde{u}_2] &= \int_0^{0.5} \left[\frac{1}{2} (2\alpha_1)^2 + 2\alpha_1 x \right] dx \\ &+ \int_{0.5}^1 \left[\frac{1}{2} (-2\alpha_1 + 2\alpha_2)^2 + 2\alpha_1(1 - x) + 2\alpha_2(x - \frac{1}{2}) \right] dx + 2\alpha_2 \\ &= 2\alpha_1^2 - 2\alpha_1\alpha_2 + \alpha_2^2 + \frac{1}{2}\alpha_1 + \frac{9}{4}\alpha_2. \end{aligned}$$

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

$$\tilde{u}_2(x) = \begin{cases} -\frac{11}{4}x & \text{if } 0 \leq x \leq 0.5 \\ -\frac{1}{4}(1 + 9x) & \text{if } 0.5 < x \leq 1 \end{cases}.$$

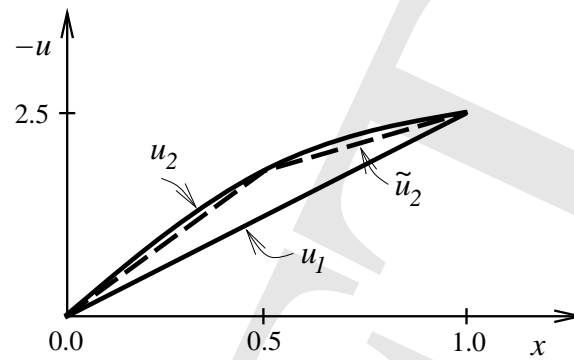


Figure 4.2: *Comparison of exact and approximate solutions*

Solutions u_1 , u_2 and \tilde{u}_2 are plotted in Figure 4.2.

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 Suggestions for further reading

Section 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

$$\text{supp } f := \overline{\{\mathbf{x} \in \Omega \mid f(\mathbf{x}) \neq 0\}} .$$

With reference to the general form of the approximation functions u_h and w_h given by equations (3.10) and (3.11), respectively, one may establish a distinction between *global* and *local* approximation methods. Local approximation methods are those for which $\text{supp } \varphi_I$ 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 (3.14), 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 (3.10) and (3.11) 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 $\{\varphi_1, \dots, \varphi_N\}$ 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 cannot be approximated.
- (b) All 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 spanned by functions $\{\varphi_1, \dots, \varphi_N\}$, namely

$$\mathcal{U}_h = \left\{ u_h \mid u_h = \sum_{I=1}^N \alpha_I \varphi_I, \quad \alpha_I \in \mathbb{R}, \quad I = 1, \dots, N \right\}.$$

Linear independence of the interpolation functions is equivalent to stating that given any $u_h \in \mathcal{U}_h$, there exists a unique set of parameters $\{\alpha_1, \dots, \alpha_N\}$, such that

$$u_h = \sum_{I=1}^N \alpha_I \varphi_I.$$

An alternative statement of linear independence of functions $\{\varphi_1, \dots, \varphi_N\}$ is that

$$\sum_{I=1}^N \alpha_I \varphi_I = 0 \quad \Leftrightarrow \quad \alpha_I = 0 \quad , \quad I = 1, \dots, N .$$

If property (c) holds, then functions $\{\varphi_1, \dots, \varphi_N\}$ are said to form a *basis* of \mathcal{U}_h . Linear independence of the interpolation functions is essential for the derivation of approximate solutions. Indeed, if parameters $\{\alpha_1, \dots, \alpha_N\}$ are not uniquely defined for any given $u_h \in \mathcal{U}_h$, then the linear algebraic system (3.14) does not possess a unique solution and, consequently, the discrete problem is ill-posed.

- (d) Interpolation functions must satisfy the integrability requirements emanating from the associated weak forms, as discussed in Chapters 3 and 4.
- (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

$$u_h = \sum_{I=1}^N \alpha_I \varphi_I$$

approximates a function $u \in \mathcal{U}$ as N increases. In order to address the above point, consider a set of functions $\{\varphi_1, \varphi_2, \dots, \varphi_N, \dots\}$, which are linearly independent in \mathcal{U} and, thus, form a countably infinite basis.¹ These functions are termed *orthonormal* in \mathcal{U} if

$$\langle \varphi_I, \varphi_J \rangle = \begin{cases} 1 & \text{if } I = J \\ 0 & \text{if } I \neq J \end{cases} .$$

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

$$\psi_1 := \frac{\varphi_1}{\|\varphi_1\|} ,$$

so that, clearly,

$$\langle \psi_1, \psi_1 \rangle = 1 .$$

Then, let

$$\psi_2 = a_2[\varphi_2 - \langle \varphi_2, \psi_1 \rangle \psi_1] , \tag{5.1}$$

¹Hilbert spaces can be shown to always possess such a basis.

where a_2 is a scalar parameter to be determined. It is immediately seen from (5.1) that

$$\begin{aligned} \langle \psi_1, \psi_2 \rangle &= \langle \psi_1, a_2 \varphi_2 - a_2 \langle \varphi_2, \psi_1 \rangle \psi_1 \rangle \\ &= a_2 \langle \psi_1, \varphi_2 \rangle - a_2 \langle \psi_1, \psi_1 \rangle \langle \psi_1, \varphi_2 \rangle = 0. \end{aligned}$$

The scalar parameter a_2 is determined so that $\|\psi_2\| = 1$, namely

$$a_2 = \frac{1}{\|\varphi_2 - \langle \varphi_2, \psi_1 \rangle \psi_1\|}.$$

In general, the function φ_{K+1} , $K = 1, 2, \dots$, gives rise to ψ_{K+1} defined as

$$\psi_{K+1} = a_{K+1} \left[\varphi_{K+1} - \sum_{I=1}^K \langle \varphi_{K+1}, \psi_I \rangle \psi_I \right], \quad (5.2)$$

where

$$a_{K+1} = \frac{1}{\|\varphi_{K+1} - \sum_{I=1}^K \langle \varphi_{K+1}, \psi_I \rangle \psi_I\|}.$$

To establish that $\{\psi_1, \psi_2, \dots, \psi_N, \dots\}$ are orthonormal, it suffices to show by induction that if $\{\psi_1, \psi_2, \dots, \psi_K\}$ are orthonormal, then ψ_{K+1} is orthonormal with respect to each of the first K members of the sequence. Indeed, using (5.2) it is seen that

$$\begin{aligned} \langle \psi_{K+1}, \psi_K \rangle &= \langle a_{K+1} \varphi_{K+1} - a_{K+1} \sum_{I=1}^K \langle \varphi_{K+1}, \psi_I \rangle \psi_I, \psi_K \rangle \\ &= \langle a_{K+1} \varphi_{K+1}, \psi_K \rangle - \sum_{I=1}^{K-1} a_{K+1} \langle \varphi_{K+1}, \psi_I \rangle \langle \psi_I, \psi_K \rangle \\ &\quad - a_{K+1} \langle \varphi_{K+1}, \psi_K \rangle \langle \psi_K, \psi_K \rangle = 0 \end{aligned}$$

and, for $N < K$,

$$\begin{aligned} \langle \psi_{K+1}, \psi_N \rangle &= \langle a_{K+1} \varphi_{K+1} - a_{K+1} \sum_{I=1}^K \langle \varphi_{K+1}, \psi_I \rangle \psi_I, \psi_N \rangle \\ &= \langle a_{K+1} \varphi_{K+1}, \psi_N \rangle - \sum_{I=1}^{N-1} a_{K+1} \langle \varphi_{K+1}, \psi_I \rangle \langle \psi_I, \psi_N \rangle \\ &\quad - \sum_{I=N+1}^K a_{K+1} \langle \varphi_{K+1}, \psi_I \rangle \langle \psi_I, \psi_N \rangle \\ &\quad - a_{K+1} \langle \varphi_{K+1}, \psi_N \rangle \langle \psi_N, \psi_N \rangle = 0, \end{aligned}$$

which establishes the desired result.

Since $\{\psi_1, \psi_2, \dots, \psi_N, \dots\}$ is an orthonormal basis in \mathcal{U} , one may uniquely write any $u \in \mathcal{U}$ in the form

$$u = \sum_{I=1}^{\infty} \alpha_I \psi_I, \quad (5.3)$$

which may be interpreted as meaning that given any $\epsilon > 0$, there exists a positive integer N and scalars α_I , such that

$$\|u - \sum_{I=1}^n \alpha_I \psi_I\| < \epsilon,$$

for all $n > N$. The coefficients α_I in (5.3) are known as the *Fourier coefficients* of u with respect to the given basis and can be easily determined by exploiting that orthogonality of ψ_I and noting that

$$\begin{aligned} \langle u, \psi_J \rangle &= \left\langle \sum_{I=1}^{\infty} \alpha_I \psi_I, \psi_J \right\rangle \\ &= \sum_{I=1}^{\infty} \alpha_I \langle \psi_I, \psi_J \rangle = \alpha_J. \end{aligned}$$

Therefore, one obtains the *Fourier representation* of u with respect to the given orthonormal basis as

$$u = \sum_{I=1}^{\infty} \langle u, \psi_I \rangle \psi_I. \quad (5.4)$$

It is noted that the natural norm of u satisfies *Parseval's identity*, namely,

$$\|u\|^2 = \sum_{I=1}^{\infty} |\alpha_I|^2.$$

Indeed,

$$\begin{aligned} \|u\|^2 = \langle u, u \rangle &= \left\langle \sum_{I=1}^{\infty} \langle u, \psi_I \rangle \psi_I, \sum_{J=1}^{\infty} \langle u, \psi_J \rangle \psi_J \right\rangle \\ &= \sum_{I=1}^{\infty} \langle u, \psi_I \rangle \sum_{J=1}^{\infty} \langle u, \psi_J \rangle \langle \psi_I, \psi_J \rangle \\ &= \sum_{I=1}^{\infty} \langle u, \psi_I \rangle^2 = \sum_{I=1}^{\infty} |\alpha_I|^2. \end{aligned}$$

If $\{\phi_1, \phi_2, \dots, \phi_N, \dots\}$ are merely a countably infinite orthonormal set (i.e., not necessarily a basis), then

$$\begin{aligned}
0 &\leq \left\| u - \sum_{I=1}^N \langle \phi_I, u \rangle \phi_I \right\|^2 \Leftrightarrow \\
0 &\leq \left\langle u - \sum_{I=1}^N \langle \phi_I, u \rangle \phi_I, u - \sum_{J=1}^N \langle \phi_J, u \rangle \phi_J \right\rangle \Leftrightarrow \\
0 &\leq \langle u, u \rangle - \left\langle u, \sum_{J=1}^N \langle \phi_J, u \rangle \phi_J \right\rangle - \left\langle \sum_{I=1}^N \langle \phi_I, u \rangle \phi_I, u \right\rangle \\
&\quad + \sum_{I=1}^N \langle \phi_I, u \rangle \sum_{J=1}^N \langle \phi_J, u \rangle \langle \phi_I, \phi_J \rangle \Leftrightarrow \\
0 &\leq \|u\|^2 - 2 \sum_{I=1}^N \langle \phi_I, u \rangle \langle \phi_I, u \rangle + \sum_{I=1}^N \langle \phi_I, u \rangle \langle \phi_I, u \rangle \Leftrightarrow \\
0 &\leq \|u\|^2 - \sum_{I=1}^N |\alpha_I|^2,
\end{aligned}$$

and, since u does not depend on N ,

$$\sum_{I=1}^{\infty} |\alpha_I|^2 \leq \|u\|^2. \quad (5.5)$$

The above result is known as *Bessel's inequality*.

The following theorem provides a clear connection between convergence in a Hilbert space and convergence of standard algebraic series.

Theorem

Let $\{\phi_1, \phi_2, \dots, \phi_N, \dots\}$ be a countably infinite orthonormal set in a Hilbert space \mathcal{U} . Then $\sum_{I=1}^{\infty} \alpha_I \phi_I$ converges in \mathcal{U} if, and only if, $\sum_{I=1}^{\infty} |\alpha_I|^2$ converges.

Proof

To prove the preceding theorem, assume that the series $\sum_{I=1}^{\infty} \alpha_I \phi_I$ converges and write

$$u = \sum_{I=1}^{\infty} \alpha_I \phi_I.$$

It follows from (5.5)

$$\sum_{I=1}^{\infty} |\alpha_I|^2 \leq \|u\|^2,$$

which implies that $\sum_{I=1}^{\infty} |\alpha_I|^2$ is a bounded series of non-negative numbers, therefore converges.

Conversely, assume that $\sum_{I=1}^{\infty} |\alpha_I|^2$ converges and set

$$u_N = \sum_{I=1}^N \alpha_I \phi_I .$$

It follows that for any $N > M$

$$\|u_N - u_M\|^2 = \langle u_N - u_M, u_N - u_M \rangle = \sum_{I=M+1}^N |\alpha_I|^2 ,$$

therefore

$$\lim_{N, M \rightarrow \infty} \|u_N - u_M\| = 0 ,$$

which implies that u_N is a Cauchy sequence. Since \mathcal{U} is a Hilbert space (therefore is complete), it follows that u_N converges, namely

$$\sum_{I=1}^{\infty} \alpha_I \phi_I$$

is convergent.

The interpolation functions φ_I used in the finite element method should satisfy the *completeness* property. This means that they have to be appropriately chosen from a family of functions which have the property that if $u \in \mathcal{U}$ and $\langle u, \varphi_I \rangle = 0$ for all $I = 1, 2, \dots$, then $u = 0$. It can be shown that in the context of Hilbert spaces, the completeness property is equivalent to satisfaction of Parseval's identity for any $u \in \mathcal{U}$. Also, completeness is equivalent to the requirement that any $u \in \mathcal{U}$ be expressed in a Fourier representation, as in (5.4).

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

$$|f(x) - P_N(x)| < \epsilon ,$$

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 can be potentially approximated by a polynomial u_h of degree N , so that

$$\lim_{N \rightarrow \infty} \|u - u_h\| = 0 .$$

Polynomials in \mathbb{R} (namely the sequence of functions $1, x, x^2, \dots$) satisfy the completeness property as stated earlier. The above theorem can be extended to polynomials defined in closed and bounded subsets of \mathbb{R}^n , as well as to trigonometric functions as evidenced by the classical Fourier representation of a continuous real function u in the form

$$u(x) = \sum_{k=0}^{\infty} (\alpha_k \sin kx + \beta_k \cos kx) .$$

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.

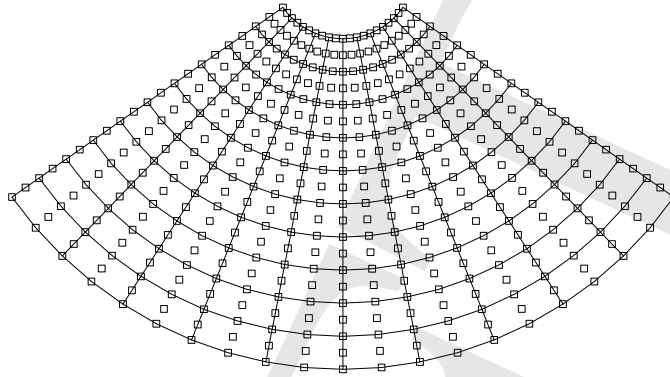
5.2 Finite element spaces

The finite element method is a rational procedure for constructing local piece-wise 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 domain Ω of analysis, admit the existence of finite element sub-domains Ω_e , such that

$$\Omega = \overline{\bigcup_e \Omega_e} , \quad (5.6)$$

as shown schematically in Figure 5.1. Similarly, the boundary $\partial\Omega$ is decomposed into sub-domains $\partial\Omega_e$ consistently with (5.6), so that

$$\partial\Omega = \overline{\bigcup_e \partial\Omega_e} ,$$

Figure 5.1: A *finite element mesh*

Also, admit the existence of points $I \in \Omega$, associated with sub-domains Ω_e . Points I have coordinates \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 Ω (i.e., accounting for the specific geometry of the sub-domains and nodes) 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 piece-wise polynomial interpolation functions φ_I are defined for each interior finite element node, so that, by convention,

$$\varphi_I(\mathbf{x}_J) = \begin{cases} 1 & \text{if } I = J \\ 0 & \text{otherwise} \end{cases} . \quad (5.7)$$

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 5.2.

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 .

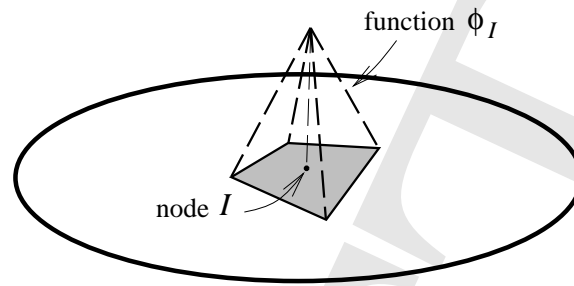


Figure 5.2: A finite element-based interpolation function

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 (5.6) holds only approximately, i.e. subdomains Ω_e only partially cover the domain Ω , as seen in Figure 5.3. In this case, property (a) may be violated in certain small regions on the domain, thus inducing an error in the approximation.

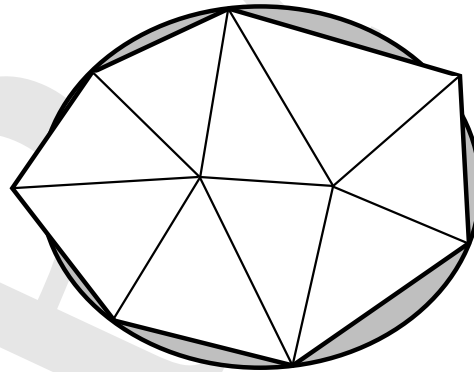


Figure 5.3: Finite element vs. exact domain

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 5.4.

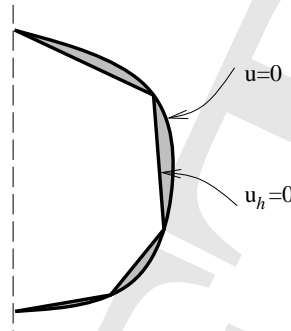


Figure 5.4: *Error in the enforcement of Dirichlet boundary conditions due to the difference between the exact and the finite element domain*

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 (5.7), one may immediately conclude that at any node I

$$u_h = \alpha_I \varphi_I(\mathbf{x}_I) = \alpha_I ,$$

hence $\alpha_I = 0$. Since nodal point I is chosen arbitrary, it follows that all α_I 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 5.5.

Property (e) and its implications within the context of the finite element method deserve special attention, and are discussed separately in the following section.

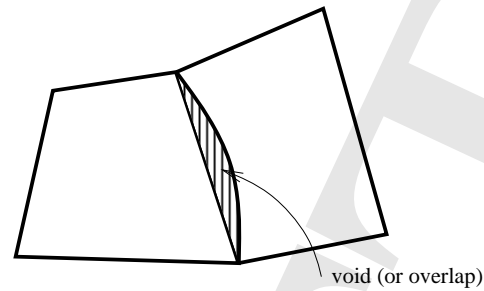


Figure 5.5: A potential violation of the integrability (compatibility) requirement

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 field, 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 “finesness” of the approximation. To see this point, consider a smooth real function u , and fix a point x in its domain. With reference to Taylor’s theorem, write

$$u(x+h) = u(x) + hu'(x) + \frac{1}{2!}h^2u''(x) + \dots + \frac{1}{q!}h^qu^{(q)}(x) + o(h^{q+1}),$$

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 $x + h$

$$u = u_h + o(h^{q+1}) . \quad (5.8)$$

Letting p be the order of the highest derivative of u in any weak form, it follows from (5.8) that

$$\frac{d^p u}{dx^p} = \frac{d^p u_h}{dx^p} + o(h^{q-p+1}) .$$

Thus, for \mathcal{U}_h to be a (polynomially) complete field, it suffices to establish that

$$q - p + 1 \geq 1 ,$$

or, equivalently,

$$q \geq p . \quad (5.9)$$

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 $h \rightarrow 0$ (h-refinement) or $q \rightarrow \infty$ (p-refinement). Indeed, consider a weak form

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

associated with a linear partial differential equation and let both u_h and w_h be refined in the same fashion (i.e. using h- or p-refinement). It is easily seen that

$$B(w, u) = B(w_h, u_h) + B(w - w_h, u - u_h) + B(w - w_h, u_h) + B(w_h, u - u_h) .$$

Owing to (5.9), 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

$$B(w, u) = \lim_{h \rightarrow 0} B(w_h, u_h) .$$

under h-refinement, and

$$B(w, u) = \lim_{q \rightarrow \infty} B(w_h, u_h) .$$

under p-refinement.

Similar conclusions can be reached for the linear forms (w, f) and $(w, \bar{q})_{\Gamma_q}$.

Examples:

(a) Consider the differential equation

$$k \frac{d^2 u}{dx^2} = f \quad \text{in } (0, 1),$$

where $p = 1$ when using the Galerkin method (see Chapter 3). Then, (5.9) 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 has been shown that a weak (variational) form is derivable such that $p = 2$. Then, the monomial terms that should be independently present in any complete approximation are $\{1, x_1, x_2, x_1^2, x_1 x_2, x_2^2\}$.

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 (5.9) yields more accurate approximations of the exact solution to a given problem.

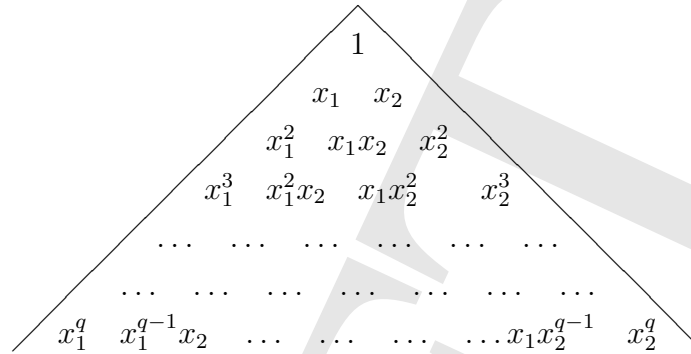
A polynomial approximation in \mathbb{R}^n is said to be *complete up to order q* , if it contains independently all monomials $x_1^{q_1} x_2^{q_2} \dots x_n^{q_n}$, where $q_1 + q_2 + \dots + q_n \leq q$. In \mathbb{R} , the above implies that terms $\{1, x, \dots, x^q\}$ should be independently represented. In \mathbb{R}^2 , completeness up to order q can be conveniently visualized by means of a *Pascal triangle*, as shown in Figure 5.6. In this case, the number of independent monomials is $\frac{(q+1)(q+2)}{2}$.

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

$$A[u] = f \tag{5.10}$$

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 (3.15). Therefore, for fixed h , one may define a *discrete* operator A_h associated with the operator A , so that

$$A_h[u] := A[u_h],$$

Figure 5.6: *Pascal triangle*

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

$$A_h[u] = f + o(h^\alpha) \quad ; \quad \alpha > 0 . \quad (5.11)$$

Assuming sufficient smoothness of u , equation (5.11) 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 can be fully determined by two sets of data:

- (i) The position of 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 given domain interpolation functions.

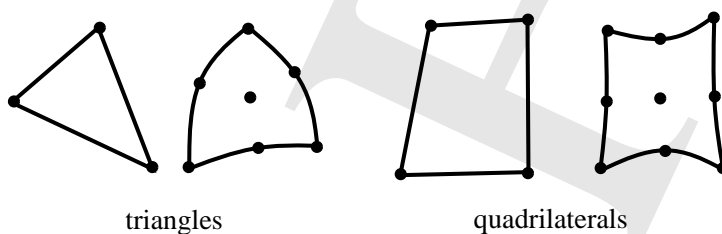
5.4.1 One dimension

One-dimensional finite element domains are line segments, straight or curved, as in Figure 5.7.

Figure 5.7: *Finite element domains in one dimension*

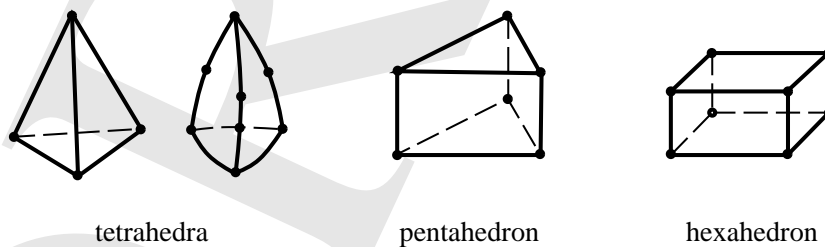
5.4.2 Two dimensions

Two-dimensional finite element domains are typically triangular or quadrilateral, with straight or curved edges, as in Figure 5.8. Elements with more complicated geometric shapes are rarely used in practice.

Figure 5.8: *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 5.9. Again, elements with more complicated geometric shapes are generally avoided.

Figure 5.9: *Finite element domains in three dimensions*

5.4.4 Higher dimensions

Elements of four dimensions or higher will not be discussed here.

5.5 Polynomial shape 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 $q \geq 1$.

The simplest finite element which satisfies the above integrability and completeness requirements is the 2-noded element of length Δx , as in Figure 5.10. Associated with every such element, 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

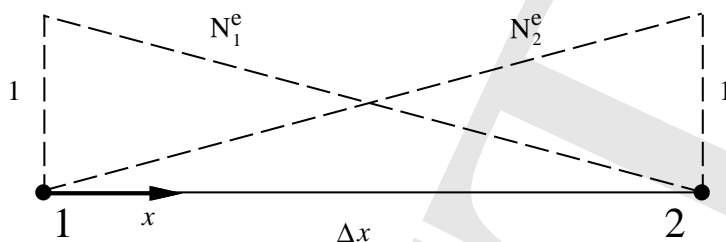
$$u_h(x) = N_1^e(x) u_1^e + N_2^e(x) u_2^e, \quad (5.12)$$

where the element interpolation functions N_1^e and N_2^e are defined as

$$N_1^e(x) = 1 - \frac{x}{\Delta x}, \quad N_2^e(x) = \frac{x}{\Delta x}.$$

It is immediately noted 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 (5.12) 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 5.10), respectively.

Clearly, the above finite element approximation is complete in 1 and x (i.e., $q = 1$). 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.

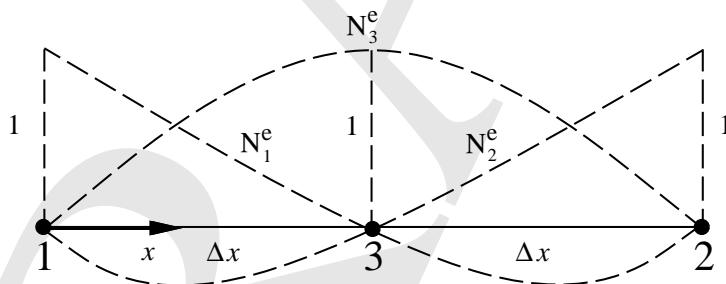
Figure 5.10: *Linear element interpolations in one dimension*

A complete quadratic interpolation can be obtained by constructing 3-noded elements as in Figure 5.11. Here, the dependent variable is given by

$$u_h(x) = N_1^e(x) u_1^e + N_2^e(x) u_2^e + N_3^e(x) u_3^e, \quad (5.13)$$

where

$$N_1^e(x) = \frac{(x - \Delta x)(x - 2\Delta x)}{2\Delta x^2}, \quad N_2^e(x) = \frac{x(x - \Delta x)}{2\Delta x^2}, \quad N_3^e(x) = -\frac{x(x - 2\Delta x)}{\Delta x^2}.$$

Figure 5.11: *Standard quadratic element interpolations in one dimension*

Again, compatibility and completeness (to degree $q = 2$) are satisfied by the interpolation in (5.13).

Generally, for an element with $q+1$ nodes having coordinates x_i , $i = 1, \dots, q+1$, one may obtain a *Lagrangian interpolation* of the form

$$u_h(x) = \sum_{i=1}^{q+1} N_i^e(x) u_i^e.$$

The generic element interpolation function N_i^e is a polynomial of degree q written as

$$N_i^e(x) = c_0 + c_1x + \dots + c_qx^q,$$

where

$$N_i^e(x_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}. \quad (5.14)$$

Conditions (5.14) 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

$$N_i^e(x) := l_i(x) = \frac{(x - x_1) \dots (x - x_{i-1})(x - x_{i+1}) \dots (x - x_{q+1})}{(x_i - x_1) \dots (x_i - x_{i-1})(x_i - x_{i+1}) \dots (x_i - x_{q+1})}. \quad (5.15)$$

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 *xhierarchical* interpolation. To illustrate an application of hierarchical interpolation, consider the 2-noded element discussed earlier in this section, and modify (5.12) so that

$$u_h(x) = N_1^e(x) u_1^e + N_2^e(x) u_2^e + N_3^e(x) \alpha^e,$$

where both the function N_3^e and the degree of freedom α^e are to be determined. Clearly, 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 (5.12). Therefore,

$$N_3^e(x) = c_0 + c_1x + c_2x^2,$$

where c_0 , c_1 and c_2 are parameters to be determined. In order to satisfy compatibility (i.e., continuity of u_h at interelement boundaries), it is sufficient to assume that $N_3^e(0) = 0$ and $N_3^e(\Delta x) = 0$. These conditions imply that

$$N_3^e(x) = \frac{cx}{\Delta x} \left(1 - \frac{x}{\Delta x}\right),$$

where c can be any non-zero constant. The three interpolation functions obtained by the above hierarchical procedure are depicted in Figure 5.12. In contrast with the standard interpolation, here the degree of freedom α^e is not associated with a finite element node. A

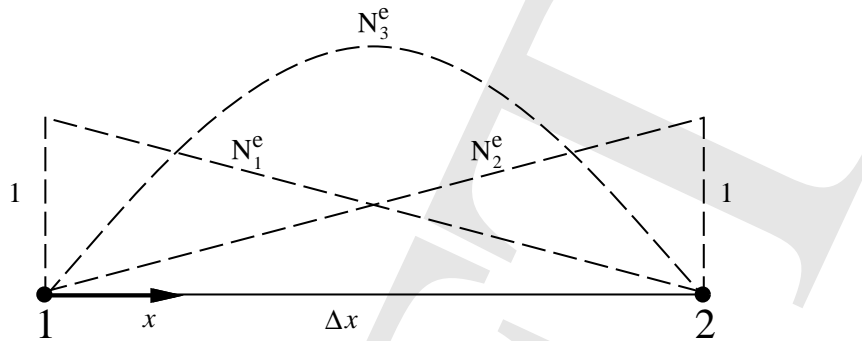


Figure 5.12: Hierarchical quadratic element interpolations in one dimension

simple algebraic interpretation of α^e can be obtained as follows: let the element interpolation function $N_e^e(x)$ take the specific form

$$N_3^e(x) = \frac{4x}{\Delta x} \left(1 - \frac{x}{\Delta x}\right).$$

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 often possible to determine its value locally, 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

$$\frac{d^4 u}{dx^4} = f,$$

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 5.13, in which each node is associated with two degrees of freedom, identified as the ordinates of the dependent variable u and its first derivative $\theta := \frac{du}{dx}$, respectively.

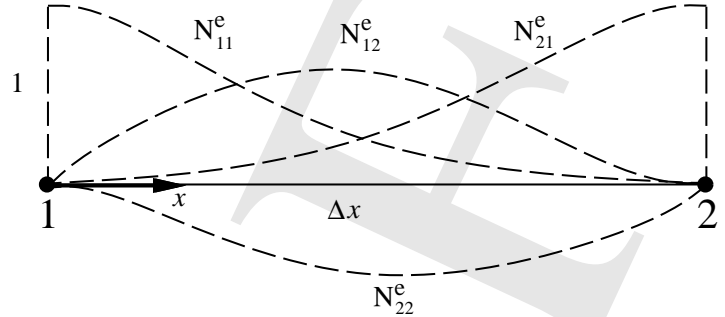


Figure 5.13: *Hermitian interpolation functions in one dimension*

Given that there are four degrees of freedom in each element, a cubic polynomial interpolation of the form

$$u_h(x) = c_0 + c_1x + c_2x^2 + c_3x^3$$

can be determined uniquely under the conditions

$$u_h(0) = u_1^e, \quad \frac{du}{dx}(0) = \theta_1^e, \quad u_h(\Delta x) = u_2^e, \quad \frac{du}{dx}(\Delta x) = \theta_2^e.$$

Solving the above equations for the four parameters c_0 to c_3 yields a standard Hermitian interpolation, in which

$$u_h(x) = \sum_{i=1}^2 N_{i1}^e(x) u_i^e + \sum_{i=1}^2 N_{i2}^e(x) \theta_i^e,$$

where

$$N_{11}^e = 1 - 3\left(\frac{x}{\Delta x}\right)^2 + 2\left(\frac{x}{\Delta x}\right)^3, \quad N_{21}^e = 3\left(\frac{x}{\Delta x}\right)^2 - 2\left(\frac{x}{\Delta x}\right)^3$$

and

$$N_{12}^e = \Delta x \left[\frac{x}{\Delta x} - 2\left(\frac{x}{\Delta x}\right)^2 + \left(\frac{x}{\Delta x}\right)^3 \right], \quad N_{22}^e = \Delta x \left[-\left(\frac{x}{\Delta x}\right)^2 + \left(\frac{x}{\Delta x}\right)^3 \right].$$

Generally, a Hermitian interpolation can be introduced for a $q+1$ -noded element, where each node i is associated with coordinate x_i and with degrees of freedom u_i^e and θ_i^e . It follows that u_h is a polynomial of degree $2q+1$ in the form

$$u_h(x) = \sum_{i=1}^{q+1} N_{i1}^e(x) u_i^e + \sum_{i=1}^{q+1} N_{i2}^e(x) \theta_i^e .$$

The element interpolation functions N_{i1}^e in the above equation satisfy

$$N_{i1}^e(x_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} , \quad \frac{dN_{i1}^e}{dx}(x_j) = 0 .$$

Similarly, the functions N_{i2}^e satisfy the conditions

$$\frac{dN_{i2}^e}{dx}(x_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} , \quad N_{i2}^e(x_j) = 0 .$$

It can be easily verified that the above Hermitian polynomials are defined as

$$N_{i1}^e(x) := [1 - 2l_i'(x_i)(x - x_i)] l_i^2(x) , \quad N_{i2}^e(x) := (x - x_i) l_i^2(x) ,$$

where $l_i(x)$ denotes the Lagrangian polynomial of degree q defined in (5.15). 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-noded element and adding hierarchical degrees of freedom. For example, one may assume a quartic interpolation of the form

$$u_h(x) = \sum_{i=1}^2 N_{i1}^e(x) u_i^e + \sum_{i=1}^2 N_{i2}^e(x) \theta_i^e + N_3^e(x) \alpha^e ,$$

where the interpolation function N_3^e is written as

$$N_3^e(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + c_4 x^4 .$$

Given the conditions $N_3^e(0) = N_3^e(\Delta x) = 0$ and $\frac{dN_3^e}{dx}(0) = \frac{dN_3^e}{dx}(\Delta x) = 0$, it follows that

$$N_3^e(x) = c \left[\left(\frac{x}{\Delta x} \right)^2 - 2 \left(\frac{x}{\Delta x} \right)^3 + \left(\frac{x}{\Delta x} \right)^4 \right] .$$

Finite element approximations which enforce continuity of higher-order derivatives are generally easy. 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 in 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 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-noded straight-edge triangle Ω^e with one degree-of-freedom per node, as seen in Figure 5.14. For this element, assume a linear poly-

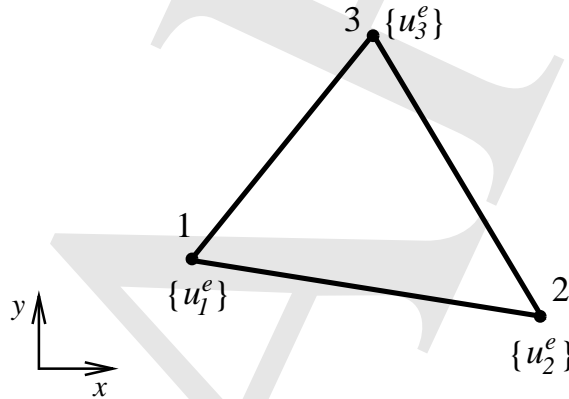


Figure 5.14: A three-noded triangular element

nomial interpolation u_h of the dependent variable u in the form

$$u_h(x, y) = \sum_{i=1}^3 N_i^e(x, y) u_i^e = c_0 + c_1 x + c_2 y, \quad (5.16)$$

with reference to a fixed Cartesian coordinate system (x, y) . Having identified the degree of freedom at node $i = 1, 2, 3$ with 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

$$\begin{aligned} u_1^e &= c_0 + c_1 x_1^e + c_2 y_1^e, \\ u_2^e &= c_0 + c_1 x_2^e + c_2 y_2^e, \\ u_3^e &= c_0 + c_1 x_3^e + c_2 y_3^e. \end{aligned} \quad (5.17)$$

Assuming that the solution of the above system is unique, one may write

$$\begin{aligned} c_0 &= \frac{1}{2A} \left[u_1^e(x_2y_3 - x_3y_2) + u_2^e(x_3y_1 - x_1y_3) + u_3^e(x_1y_2 - x_2y_1) \right], \\ c_1 &= \frac{1}{2A} \left[u_1^e(y_2 - y_3) + u_2^e(y_3 - y_1) + u_3^e(y_1 - y_2) \right], \\ c_2 &= \frac{1}{2A} \left[u_1^e(x_3 - x_2) + u_2^e(x_1 - x_3) + u_3^e(x_2 - x_1) \right], \end{aligned} \quad (5.18)$$

where

$$A := \frac{1}{2} \det \begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{pmatrix}.$$

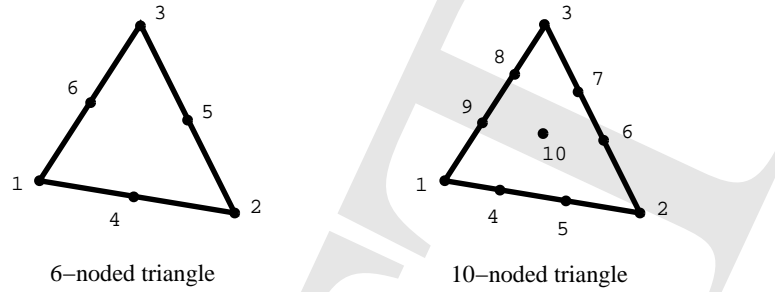
It is interesting to note that A represents the (signed) area of the triangle Ω^e . Therefore, the system (5.17) is solvable if, and only if, the nodes 1,2,3 do not lie on the same line. In addition, it can be easily concluded that the area A of a non-degenerate triangle is positive if, and only if, the nodes are numbered in a counter-clockwise manner, as in Figure 5.14.

Explicit polynomial expressions for the element interpolation functions are obtained from (5.16a) and (5.18) in the form

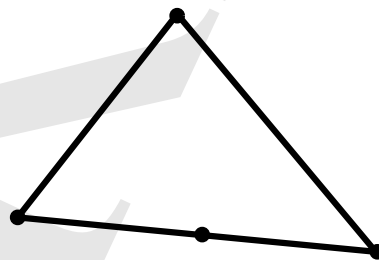
$$\begin{aligned} N_1^e &= \frac{1}{2A} \left[(x_2y_3 - x_3y_2) + (y_2 - y_3)x + (x_3 - x_2)y \right] \\ N_2^e &= \frac{1}{2A} \left[(x_3y_1 - x_1y_3) + (y_3 - y_1)x + (x_1 - x_3)y \right]. \\ N_3^e &= \frac{1}{2A} \left[(x_1y_2 - x_2y_1) + (y_1 - y_2)x + (x_2 - x_1)y \right] \end{aligned} \quad (5.19)$$

It can be noted from (5.19a) 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 (5.19).

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

Figure 5.15: *Higher-order triangular elements*

6- and 10-noded triangular elements which are polynomially complete to degree $q = 2$ and 3 are illustrated in Figure 5.15. It should be noted that the nodes are generally positioned with geometric regularity. Thus, for the 6-noded 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-noded triangle. Similarly, continuity of the dependent variable in these elements can be proved by arguments identical to those used for the 3-noded triangle. Elements featuring irregular positioning of the nodes, such as the 4-noded element in Figure 5.16 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-noded triangles with another consisting of 6-noded triangles).

Figure 5.16: *A transitional 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. 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 , A_2 and A_3 , as

shown in Figure 5-17. Noting that

$$A_1 = \frac{1}{2} \det \begin{pmatrix} 1 & x & y \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{pmatrix},$$

with similar expressions for A_2 and A_3 , define the so-called area coordinates of the point (x, y) as

$$L_i := \frac{A_i}{A}, \quad i = 1, 2, 3. \quad (5.20)$$

Clearly, only two of the three *area coordinates* are independent since from (5.20) it is seen

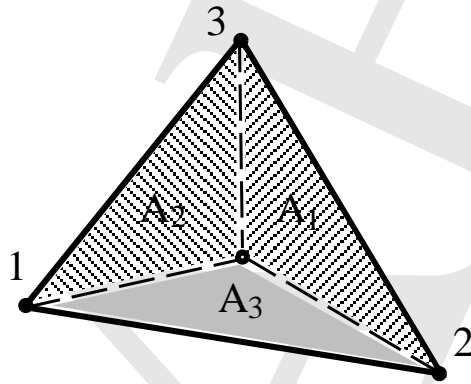


Figure 5.17: *Area coordinates in a triangular domain*

that $L_1 + L_2 + L_3 = 1$. Interestingly, comparing (5.19) to (5.20) 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 reference to the 6-noded element depicted in Figure 5.15, 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

$$\begin{aligned} N_1^e &= 2L_1(L_1 - \frac{1}{2}) & , & & N_2^e &= 2L_2(L_2 - \frac{1}{2}) & , & & N_3^e &= 2L_3(L_3 - \frac{1}{2}) \\ N_4^e &= 4L_1L_2 & , & & N_5^e &= 4L_2L_3 & , & & N_6^e &= 4L_3L_1 . \end{aligned}$$

An important formula for the integration of polynomial functions of the area coordinates over the region of a straight-edge triangle can be established in the form

$$\int_{\Omega^e} L_1^\alpha L_2^\beta L_3^\gamma dA = \frac{\alpha! \beta! \gamma!}{(\alpha + \beta + \gamma + 2)!} 2A ,$$

where α , β and γ are integers.

Quadrilateral elements are also used very 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-noded rectangle of Figure 5.18. Here, it is assumed that the dependent

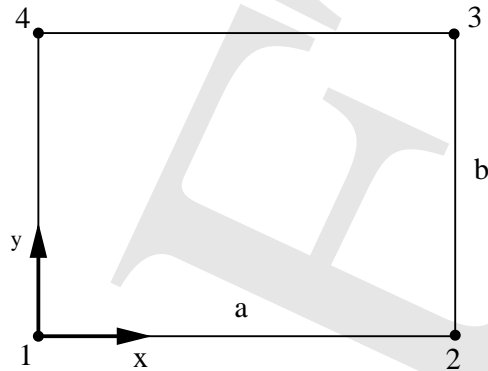


Figure 5.18: *Four-noded rectangular element*

variable is interpolated as

$$u_h = \sum_{i=1}^4 N_i^e u_i^e = c_0 + c_1 x + c_2 y + c_3 xy, \quad (5.21)$$

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 stipulating that

$$\begin{aligned} u_1^e &= c_0 + c_1 x_1^e + c_2 y_1^e + c_3 x_1^e y_1^e, \\ u_2^e &= c_0 + c_1 x_2^e + c_2 y_2^e + c_3 x_2^e y_2^e, \\ u_3^e &= c_0 + c_1 x_3^e + c_2 y_3^e + c_3 x_3^e y_3^e, \\ u_4^e &= c_0 + c_1 x_4^e + c_2 y_4^e + c_3 x_4^e y_4^e. \end{aligned}$$

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 (5.21) 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-noded rectangle of Figure 5.18, these

functions are given by

$$\begin{aligned}
 N_1^e &= \frac{1}{ab}(x - a)(y - b) , \\
 N_2^e &= -\frac{1}{ab}x(y - b) , \\
 N_3^e &= \frac{1}{ab}xy , \\
 N_4^e &= -\frac{1}{ab}(x - a)y .
 \end{aligned}$$

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-noded rectangle is common to both families. The next three elements of the serendipity family are the 8-, 12- and 17-noded elements, see Figure 5.19. These elements are polynomially complete to degree $q = 2, 3$ and 4 , respectively. The 8-noded rectangle may represent

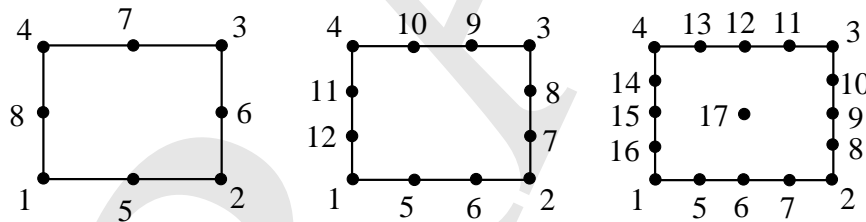


Figure 5.19: *Three members of the serendipity family of rectangular elements*

any 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-noded 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 shape functions using their fundamental property. Regrettably, this guessing becomes more difficult for the 12- and the 17-noded 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 5.20.

²Note that the degree of completeness is still only $q = 1$ despite the existence of the bilinear term xy .

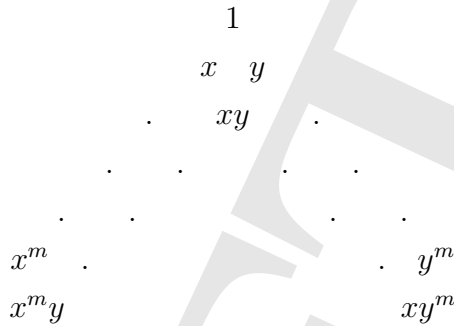


Figure 5.20: *Pascal triangle for serendipity elements*

The Lagrangian family of rectangular elements is comprised of the 4-noded element discussed earlier, followed by the 9-, 16- and 25-noded element, see Figure 5.21. The 9-noded

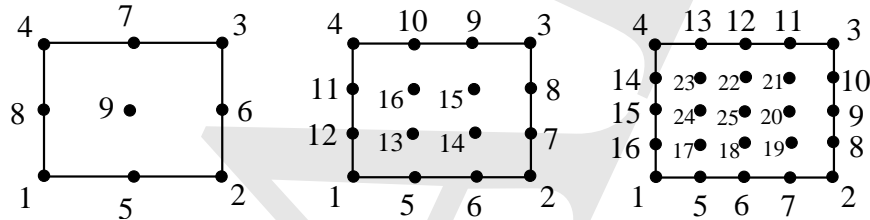


Figure 5.21: *Three members of the Lagrangian family of rectangular elements*

rectangle 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 shape functions of the Lagrangian elements can be determined trivially as products of one-dimensional Lagrange interpolation functions. As an example, consider the shape function N_{18}^e associated with node 18 of the 25-noded element of Figure 5.21. This can be written as

$$N_{18}^e = l_3(x)l_2(y) ,$$

where

$$l_3(x) = \frac{(x - x_{16})(x - x_{17})(x - x_{19})(x - x_8)}{(x_{18} - x_{16})(x_{18} - x_{17})(x_{18} - x_{19})(x_{18} - x_8)} ,$$

$$l_2(y) = \frac{(y - y_6)(y - y_{25})(y - y_{22})(y - y_{12})}{(y_{18} - y_6)(y_{18} - y_{25})(y_{18} - y_{22})(y_{18} - y_{12})} .$$

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 5.22.

connected triangles or a composite 5-noded triangle consisting of four connected triangles, as in Figure 5.24. In both cases, the interpolation in each triangle is linearly complete and continuity of the dependent variable is guaranteed at all interelement boundaries. In a subsequent section, the question of general quadrilateral elements will be revisited within the context of the so-called isoparametric mapping.

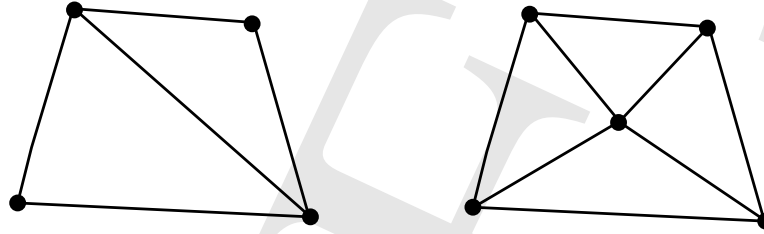


Figure 5.24: *Rectangular finite elements made of two or four joined triangular elements*

The construction of two-dimensional finite elements with $p = 2$ is substantially more complicated than the respective one-dimensional case. To illustrate this, consider a simple cubically complete interpolation of the dependent variable u_h as

$$u_h = c_0 + c_1x + c_2y + c_3x^2 + c_4xy + c_5y^2 + c_6x^3 + c_7x^2y + c_8xy^2 + c_9y^3.$$

One may choose to associate this interpolation with a 3-noded triangular element in Figure 5.25. Here, there are three degrees of freedom per node, namely the dependent variable u_h and its two partial derivatives $\frac{\partial u_h}{\partial x}$ and $\frac{\partial u_h}{\partial y}$. 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 assign 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, i.e., set $c_7 = c_8$.

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 as shown in Figure 5.26. It is clear from the original interpolation assumption that u_h varies cubically in edge 1-2. Hence, given that both u_h and $\frac{\partial u_h}{\partial s}$ are specified on this edge, it follows that u_h , as well as $\frac{\partial u_h}{\partial s}$ are continuous across 1-2. However, this is not the case for the normal derivative $\frac{\partial u_h}{\partial n}$, 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 $\frac{\partial u_h}{\partial n}$ is discontinuous across 1-2, therefore this simple element violates the integrability requirement for the case $p = 2$.

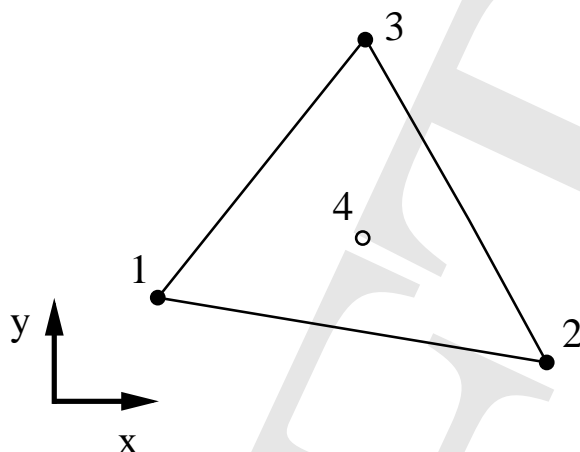


Figure 5.25: A simple potential 3- or 4-noded triangular element for the case $p = 2$

This implies that a simple extension of the one-dimensional Hermitian interpolation-based elements to the two-dimensional case is not permissible. To remedy this problem, one may

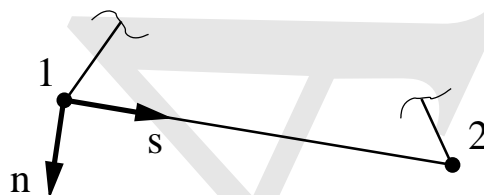
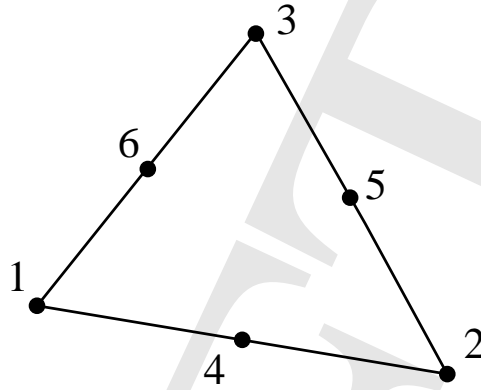


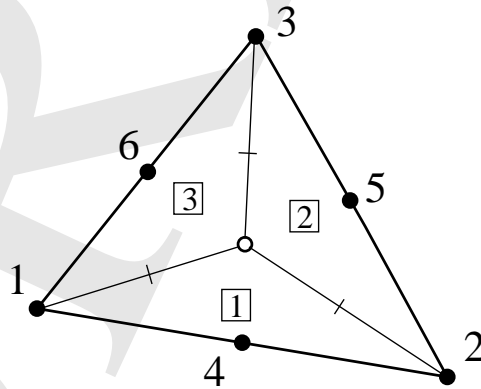
Figure 5.26: Illustration of violation of the integrability requirement for the 9- or 10-dof triangle for the case $p = 2$

resort to elements that have mid-edge degrees of freedom, such as the 6-noded triangle in Figure 5.27. 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.

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 that needs to be taken 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). Furthermore, it turns out that this element needs to employ algebraically complex rational polynomial interpolations for the mid-edge degrees of freedom. Composite triangles, such as the celebrated Clough-Tocher element, were developed to avoid these undesirable features. This element is comprised of three joined triangles, in each of which one employs a complete cubic interpolation of u_h ,

Figure 5.27: 12-dof triangular element for the case $p = 2$

see Figure 5.28. 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 are shared at each of the four vertices (three exterior and one interior), the total number of degrees of freedom is immediately reduced to 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 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. This element possesses piecewise cubic polynomial interpolation of the dependent variable in each triangular subdomain and satisfies both the integrability and the completeness requirement.

Figure 5.28: Clough-Tocher triangular element for the case $p = 2$

There are numerous triangular and quadrilateral elements for the case $p = 2$. However, their use has been gradually diminished in finite element practice. For this reason, they will

not be discussed in any further detail.

5.5.3 Interpolations in three dimensions

In this section, three dimensional polynomial interpolations are considered in connection with tetrahedral, pentahedral and hexahedral elements.

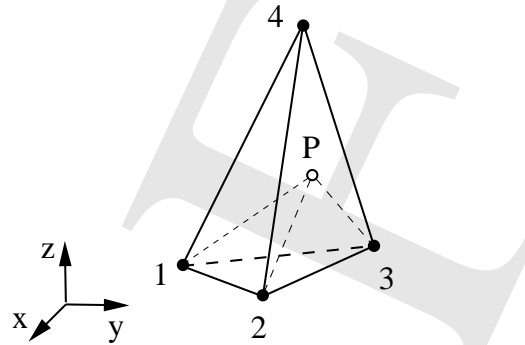


Figure 5.29: *The 4-noded tetrahedral element*

The simplest three-dimensional element is the four-noded tetrahedron with one node at each vertex, see Figure 5.29. This element has one degree-of-freedom at each node and the dependent variable u is interpolated as

$$u_h = \sum_{i=1}^4 N_i^e u_i^e = c_0 + c_1 x + c_2 y + c_3 z, \quad (5.22)$$

where u_i^e , $i = 1 - 4$, are the nodal degrees of freedom and $c_0 - c_3$ are constants. Recalling again that $u_i = u_h(x_i^e, y_i^e, z_i^e)$, i.e., that the degrees of freedom are identical in value to 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

$$\begin{aligned} u_1^e &= c_0 + c_1 x_1^e + c_2 y_1^e + c_3 z_1^e, \\ u_2^e &= c_0 + c_1 x_2^e + c_2 y_2^e + c_3 z_2^e, \\ u_3^e &= c_0 + c_1 x_3^e + c_2 y_3^e + c_3 z_3^e, \\ u_4^e &= c_0 + c_1 x_4^e + c_2 y_4^e + c_3 z_4^e. \end{aligned}$$

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 $p = 1$, i.e., 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-noded 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 $\{1, x, y, z, x^2, xy, y^2, xy, yz, zx\}$, see Figure 5.30.

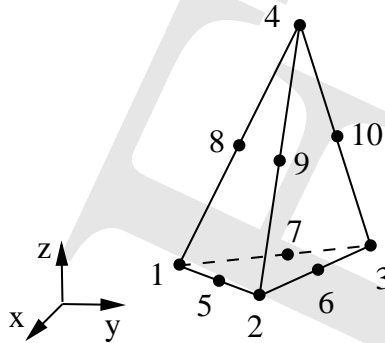


Figure 5.30: *The 10-noded tetrahedral element*

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. With reference to the 4-noded tetrahedral element of Figure 5.29, one may define the volume coordinate L_i of a typical point P in the interior of the tetrahedron as

$$L_i := \frac{V_i}{V} \quad , \quad i = 1 - 4 \quad ,$$

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-noded 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 tetrahedral regions, one may employ a convenient formula, according to which

$$\int_{\Omega^e} L_1^\alpha L_2^\beta L_3^\gamma L_4^\delta dV = \frac{\alpha! \beta! \gamma! \delta!}{(\alpha + \beta + \gamma + \delta + 3)!} 6V \quad ,$$

where α , β , γ and δ are integers.

The first two pentahedral elements of interest are the 6-noded and the 15-noded pentahedron, shown in Figure 5.31. The former is complete up to polynomial degree $q = 1$ and

its interpolation functions are capable of representing the monomial terms $\{1x, y, z, xz, yz\}$. 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.

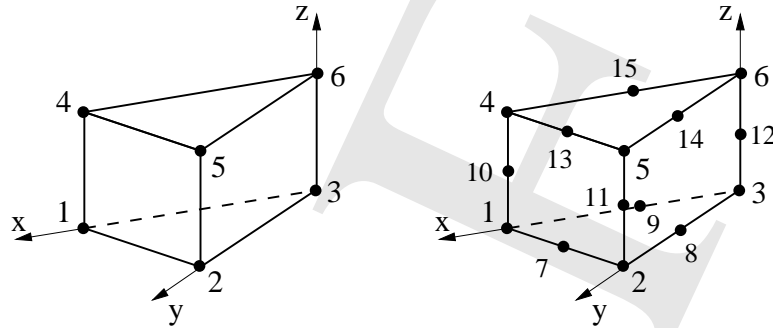


Figure 5.31: The 6- and 15-noded pentahedral elements

Hexahedral elements are very widely used in three-dimensional finite element analyses. The simplest such element is the 8-noded hexahedron with nodes at each of its vertices, see Figure 5.32. This element is polynomially complete up to degree $q = 1$ and its interpolation 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-noded hexahedron of Figure 5.32, can be written by inspection as

$$\begin{aligned}
 N_1^e &= -\frac{1}{abc}(x-a)(y-b)(z-c), \\
 N_2^e &= \frac{1}{abc}(x-a)y(z-c), \\
 N_3^e &= -\frac{1}{abc}(x-a)yz, \\
 N_4^e &= \frac{1}{abc}(x-a)(y-b)z, \\
 N_5^e &= \frac{1}{abc}x(y-b)(z-c), \\
 N_6^e &= -\frac{1}{abc}xy(z-c), \\
 N_7^e &= \frac{1}{abc}xyz, \\
 N_8^e &= -\frac{1}{abc}x(y-b)z.
 \end{aligned}$$

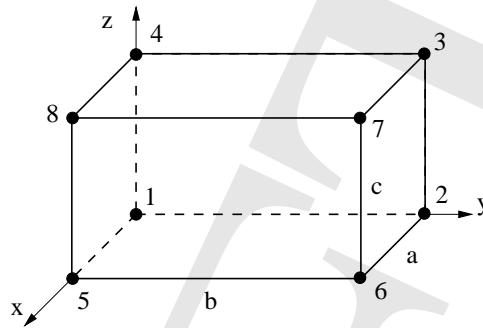


Figure 5.32: *The 8-noded hexahedral element*

The next two useful hexahedral elements are the 20- and the 27-noded element, see Figure 5.33. 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-noded 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-noded hexahedron can additionally represent the monomials

$$\{x^2y^2, y^2z^2, z^2x^2, x^2y^2z, y^2z^2x, z^2x^2y, x^2y^2z^2\}.$$

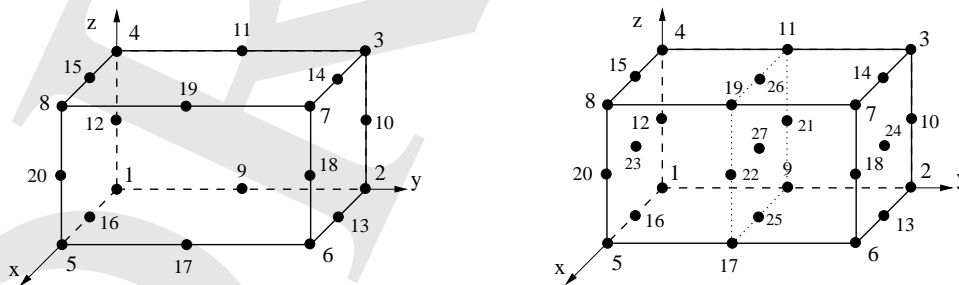


Figure 5.33: *The 20- and 27-noded hexahedral elements*

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 main idea of the isoparametric mapping is to construct the irregularly-shaped element in the *physical* domain (i.e., the domain of interest) as a mapping from a *parent* domain in which this same element has a regular shape. This mapping can be expressed in three-dimensions as

$$x = \hat{x}(\xi, \eta, \zeta) \quad , \quad y = \hat{y}(\xi, \eta, \zeta) \quad , \quad z = \hat{z}(\xi, \eta, \zeta) \quad , \quad (5.23)$$

where (ξ, η, ζ) and (x, y, z) are coordinates in the natural and physical domain, respectively. The mapping of equation (5.23) can be equivalently (and more succinctly) represented in vector form as

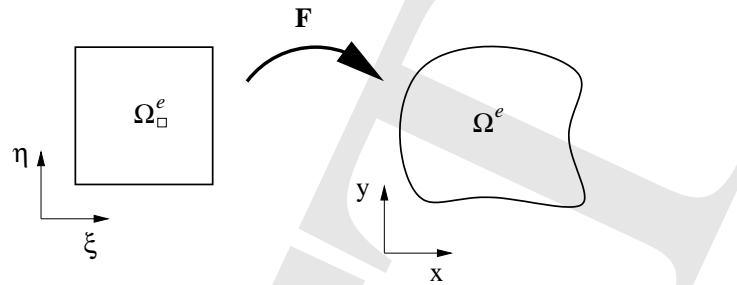
$$\mathbf{x} = \mathbf{F}(\boldsymbol{\xi}) \quad ,$$

where $\mathbf{x} = [x \ y \ z]^T$ and $\boldsymbol{\xi} = [\xi \ \eta \ \zeta]^T$. Here, \mathbf{F} maps the regular-shaped domain Ω_{\square}^e to the irregular-shaped domain Ω^e , see Figure 5.34. By way of background, the mapping \mathbf{F} is termed *one-to-one* (or *injective*) if for any two distinct points $\boldsymbol{\xi}_1 \neq \boldsymbol{\xi}_2$ in Ω_{\square}^e , their images \mathbf{x}_1 and \mathbf{x}_2 under \mathbf{F} satisfy $\mathbf{x}_1 \neq \mathbf{x}_2$. Further, the mapping \mathbf{F} is termed *onto* (or *surjective*) if $\mathbf{F}(\Omega_{\square}^e) = \Omega^e$, or, said equivalently, any point $\mathbf{x} \in \Omega^e$ is the image of some point $\boldsymbol{\xi} \in \Omega_{\square}^e$.

In order to define what constitutes an isoparametric mapping, let the dependent variable u be approximated inside the element Ω^e of interest as

$$u_h^e = \sum_{i=1}^n N_i^e u_i^e \quad , \quad (5.24)$$

where u_i^e , $i = 1 - n$, are the element degrees of freedom. Likewise, suppose that the geometry

Figure 5.34: Schematic of a parametric mapping from Ω_{\square}^e to Ω^e

of the element Ω^e is defined by the equations

$$\mathbf{x} = \sum_{j=1}^m N_j^e \mathbf{x}_j^e, \quad (5.25)$$

where \mathbf{x}_j^e , $j = 1 - m$, 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 Ω_{\square}^e , namely they are functions of the natural coordinates (ξ, η, ζ) .

With reference to equations (5.24) and (5.25), 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 same functions are employed to to define the element geometry and the interpolation of the dependent variable. The implications of this assumption will become apparent in the ensuing developments.

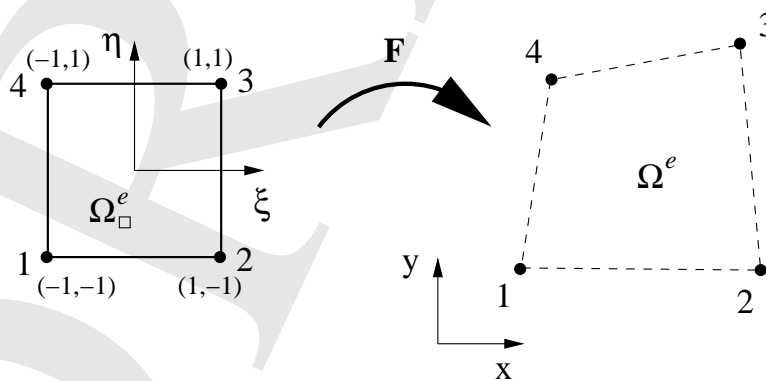


Figure 5.35: The 4-noded isoparametric quadrilateral

By way of a concrete example, consider in detail the isoparametric 4-noded quadrilateral element of Figure 5.35. The element interpolation functions in the parent domain are given

by

$$\begin{aligned}
 N_1^e(\xi, \eta) &= \frac{1}{4}(1 - \xi)(1 - \eta) , \\
 N_2^e(\xi, \eta) &= \frac{1}{4}(1 + \xi)(1 - \eta) , \\
 N_3^e(\xi, \eta) &= \frac{1}{4}(1 + \xi)(1 + \eta) , \\
 N_4^e(\xi, \eta) &= \frac{1}{4}(1 - \xi)(1 + \eta) .
 \end{aligned} \tag{5.26}$$

Given that the element is isoparametric, it follows that

$$u_h^e = \sum_{i=1}^4 N_i^e u_i^e \quad , \quad \mathbf{x} = \sum_{i=1}^4 N_i^e \mathbf{x}_i^e , \tag{5.27}$$

where \mathbf{x}_i^e are the vectors with coordinates (x_i^e, y_i^e) pointing to the positions of the four nodes 1 – 4 in the physical domain.

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 to $\xi \in (-1, 1)$ and $\eta = -1$. Given (5.27)₂, this means that the equations describing the edge 1-2 are:

$$\begin{aligned}
 x &= \frac{1}{2}(1 - \xi)x_1^e + \frac{1}{2}(1 + \xi)x_2^e = \frac{1}{2}(x_1^e + x_2^e) + \frac{1}{2}\xi(x_2^e - x_1^e) , \\
 y &= \frac{1}{2}(1 - \xi)y_1^e + \frac{1}{2}(1 + \xi)y_2^e = \frac{1}{2}(y_1^e + y_2^e) + \frac{1}{2}\xi(y_2^e - y_1^e) .
 \end{aligned}$$

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 (5.27)₁ needs to be able to exactly represent any polynomial of the form

$$u_h = c_0 + c_1x + c_2y . \tag{5.28}$$

However, equation (5.27)₁ implies that, if the four degrees of freedom u_i^e coincide with the nodal values of u_h , then

$$\begin{aligned}
 u_h &= \sum_{i=1}^4 N_i^e u_i^e = \sum_{i=1}^4 N_i^e u_h(x_i^e, y_i^e) = \sum_{i=1}^4 N_i^e (c_0 + c_1x_i^e + c_2y_i^e) \\
 &= \left(\sum_{i=1}^4 N_i^e\right)c_0 + \left(\sum_{i=1}^4 N_i^e x_i^e\right)c_1 + \left(\sum_{i=1}^4 N_i^e y_i^e\right)c_2 = \left(\sum_{i=1}^4 N_i^e\right)c_0 + c_1x + c_2y ,
 \end{aligned}$$

where equation (5.27)₂ is used. In view of equation (5.28), completeness of the 4-noded isoparametric quadrilateral is guaranteed as long as $\sum_{i=1}^4 N_i^e = 1$, which can be easily verified from equations (5.26).

Integrability for the case $p = 1$ can be established as follows: consider a typical element edge, say 1-2, along which

$$u_h(\xi, -1) = \frac{1}{2}(1 - \xi)u_1^e + \frac{1}{2}(1 + \xi)u_2^e,$$

as readily seen from equations (5.26) and (5.27)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. This, in turn, implies continuity of u_h across the edge 1-2, which is a sufficient condition for integrability.

One of the key questions associated with isoparametric finite elements is whether the isoparametric mapping \mathbf{F} , expressed here through equations (5.27)₂, is invertible. Said differently, the relevant question is whether one may uniquely associate points $(\xi, \eta) \in \Omega_{\square}^e$ with points $(x, y) \in \Omega^e$ 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 $\mathbf{F} : \Omega_{\square}^e \mapsto \Omega^e$ of class C^r , such that $\boldsymbol{\xi} \in \Omega_{\square}^e$ is mapped to $\mathbf{x} = \mathbf{F}(\boldsymbol{\xi}) \in \Omega^e$, where Ω_{\square}^e and Ω^e are open sets. If $J = \det \frac{\partial \mathbf{x}}{\partial \boldsymbol{\xi}} \neq 0$ at a point $\bar{\boldsymbol{\xi}} \in \Omega_{\square}^e$, then there is an open neighborhood around $\bar{\boldsymbol{\xi}}$, such that \mathbf{F} is one-to-one and onto an open subset of Ω^e containing the point $\bar{\mathbf{x}} = \mathbf{F}(\bar{\boldsymbol{\xi}})$ and the inverse function \mathbf{F}^{-1} exists and is of class C^r .

The derivative $\frac{\partial \mathbf{x}}{\partial \boldsymbol{\xi}}$ can be written in matrix form as

$$[J] = \begin{bmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial y}{\partial \xi} \\ \frac{\partial x}{\partial \eta} & \frac{\partial y}{\partial \eta} \end{bmatrix}, \quad (5.29)$$

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 $(\xi, \eta) \in \Omega_{\square}^e$ provided that the determinant J is non-zero everywhere in Ω_{\square}^e .³

Given equation (5.29), the Jacobian determinant J is given by

$$J = \frac{\partial x}{\partial \xi} \frac{\partial y}{\partial \eta} - \frac{\partial y}{\partial \xi} \frac{\partial x}{\partial \eta},$$

³Note that here \mathbf{F} is of class C^∞ .

which, taking into account equation (5.27)₂, leads, after some algebra, to

$$\begin{aligned}
 J = \frac{1}{8} & \left[(x_1^e y_2^e - x_2^e y_1^e + x_2^e y_3^e - x_3^e y_2^e + x_3^e y_4^e - x_4^e y_3^e + x_4^e y_1^e - x_1^e y_4^e) \right. \\
 & + \xi (x_1^e y_4^e - x_4^e y_1^e + x_2^e y_3^e - x_3^e y_2^e + x_3^e y_1^e - x_1^e y_3^e + x_4^e y_2^e - x_2^e y_4^e) \\
 & \left. + \eta (x_1^e y_3^e - x_3^e y_1^e + x_2^e y_1^e - x_1^e y_2^e + x_3^e y_4^e - x_4^e y_3^e + x_4^e y_2^e - x_2^e y_4^e) \right]. \quad (5.30)
 \end{aligned}$$

It is instructive to observe here that since J is linear in ξ and η , then if $J > 0$ at all four

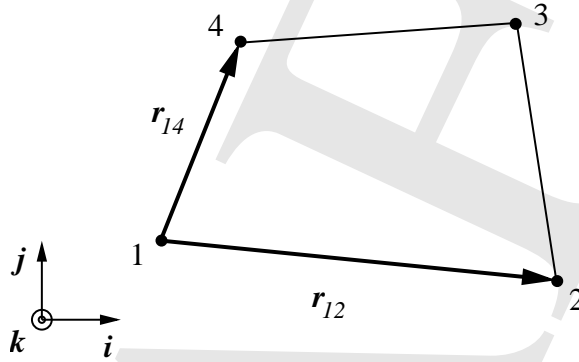


Figure 5.36: *Geometric interpretation of one-to-one isoparametric mapping in the 4-noded quadrilateral*

nodal points, then $J > 0$ everywhere in the interior of the domain Ω_{\square}^e . Now, consider node 1, with natural coordinates $(-1, -1)$ and conclude from equation (5.30) that at this node

$$J = \frac{1}{4} \left[(x_2^e - x_1^e)(y_4^e - y_1^e) - (x_4^e - x_1^e)(y_2^e - y_1^e) \right].$$

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 5.36, one may interpret the Jacobian determinant at node 1 according to

$$4J\mathbf{k} = \mathbf{r}_{12} \times \mathbf{r}_{14},$$

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 mapping for the 4-noded quadrilateral is guaranteed as long as the element domain Ω^e is convex, see Figure 5.37.

It is easy to see that the isoparametric mapping in Figure 5.35 is orientation-preserving, in the sense that the nodal sequencing (say 1-2-3-4, if following a counter-clockwise convention)

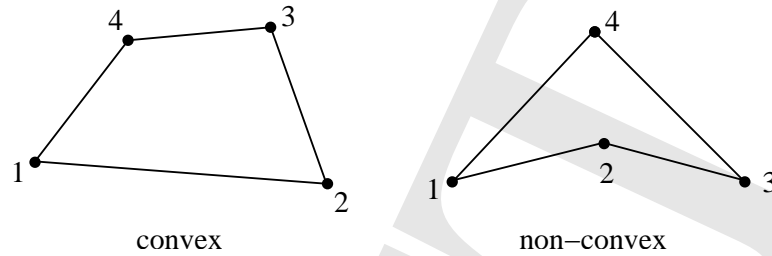


Figure 5.37: *Convex and non-convex 4-noded quadrilateral element domains*

is preserved under the mapping \mathbf{F} . 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 associated with weak forms is deduced for the 4-noded quadrilateral. To this end, note that

$$dx = \frac{\partial x}{\partial \xi} d\xi + \frac{\partial x}{\partial \eta} d\eta, \quad dy = \frac{\partial y}{\partial \xi} d\xi + \frac{\partial y}{\partial \eta} d\eta.$$

Hence, with reference to Figure 5.38, the infinitesimal vector area $d\mathbf{A}$ is written as

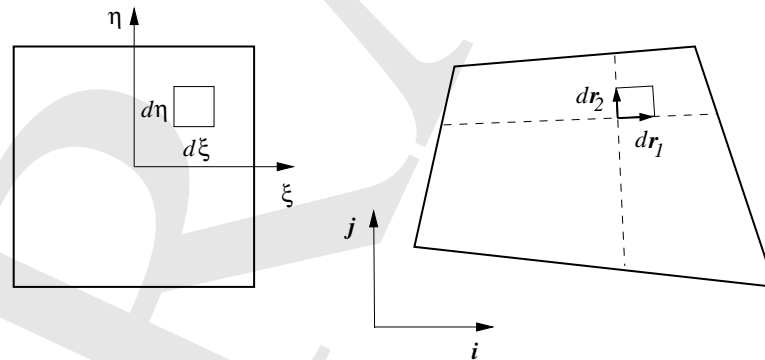


Figure 5.38: *Relation between area elements in the natural and physical domain*

$$d\mathbf{A} = d\mathbf{r}_1 \times d\mathbf{r}_2$$

where $d\mathbf{r}_1$ and $d\mathbf{r}_2$ are the infinitesimal vectors along lines of constant η and ξ , respectively. This implies that

$$d\mathbf{A} = \left(\frac{\partial x}{\partial \xi} d\xi \mathbf{i} + \frac{\partial y}{\partial \xi} d\xi \mathbf{j} \right) \times \left(\frac{\partial x}{\partial \eta} d\eta \mathbf{i} + \frac{\partial y}{\partial \eta} d\eta \mathbf{j} \right) = J d\xi d\eta \mathbf{k},$$

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

$$dA = J d\xi d\eta.$$

The above argument has not made specific use of the isoparametric mapping of the 4-noded quadrilateral element, hence it applies to all planar isoparametric elements. This argument can be easily extended to three-dimensional elements or restricted to one-dimensional elements of the isoparametric type.

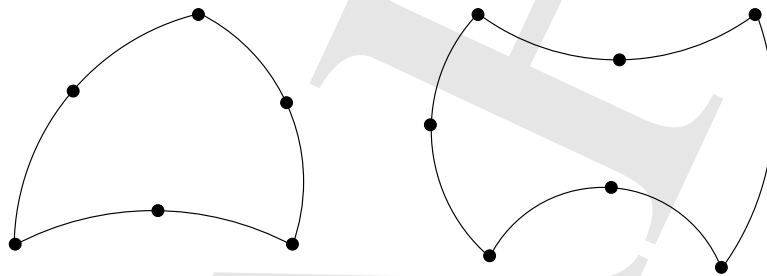


Figure 5.39: *Isoparametric 6-noded triangle and 8-noded quadrilateral*

The isoparametric approach can be applied to triangles and quadrilaterals without appreciable complication over what has been described for the 4-noded quadrilateral. For example, higher-order planar isoparametric elements can be constructed based on the 6-noded triangle and the 8-noded serendipity rectangle, see Figure 5.39. Both elements may have curved boundaries, which is a desirable feature when modeling arbitrary domains.

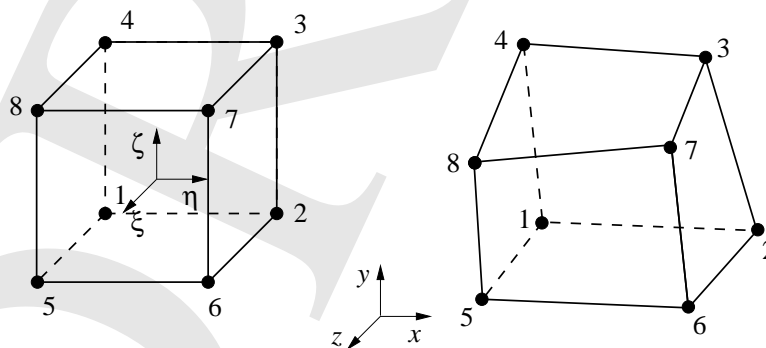


Figure 5.40: *Isoparametric 8-noded hexahedral element*

Three-dimensional isoparametric elements are also possible and, in fact, quite popular. The simplest such element is the 8-noded isoparametric brick of Figure 5.40. The geometry

of this element is defined in terms of the position vectors \mathbf{x}_i of its eight vertex nodes and the corresponding interpolation functions N_i^e in the natural domain. The latter can be written relative to the coordinate system shown in Figure 5.40 as

$$\begin{aligned} N_1^e &= \frac{1}{4}(1 - \xi)(1 - \eta)(1 - \zeta) & , & & N_2^e &= \frac{1}{4}(1 - \xi)(1 + \eta)(1 - \zeta) & , \\ N_3^e &= \frac{1}{4}(1 - \xi)(1 + \eta)(1 + \zeta) & , & & N_4^e &= \frac{1}{4}(1 - \xi)(1 - \eta)(1 + \zeta) & , \\ N_5^e &= \frac{1}{4}(1 + \xi)(1 - \eta)(1 - \zeta) & , & & N_6^e &= \frac{1}{4}(1 + \xi)(1 + \eta)(1 - \zeta) & , \\ N_7^e &= \frac{1}{4}(1 + \xi)(1 + \eta)(1 + \zeta) & , & & N_8^e &= \frac{1}{4}(1 + \xi)(1 - \eta)(1 + \zeta) & . \end{aligned} \quad (5.31)$$

All element edges in the 8-noded isoparametric brick are straight. Indeed, a typical edge, say 8-7 with coordinates $(1, \eta, 1)$, is described by the equations

$$\begin{aligned} x &= \frac{1}{2}(x_7^e + x_8^e) + \frac{1}{2}(x_7^e - x_8^e)\eta & , \\ y &= \frac{1}{2}(y_7^e + y_8^e) + \frac{1}{2}(y_7^e - y_8^e)\eta & , \\ z &= \frac{1}{2}(z_7^e + z_8^e) + \frac{1}{2}(z_7^e - z_8^e)\eta & , \end{aligned}$$

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

$$\begin{aligned} x &= \frac{1}{4}(1 - \xi)(1 + \eta)x_3^e + \frac{1}{4}(1 - \xi)(1 - \eta)x_4^e + \frac{1}{4}(1 + \xi)(1 + \eta)x_7^e + \frac{1}{4}(1 + \xi)(1 - \eta)x_8^e & , \\ y &= \frac{1}{4}(1 - \xi)(1 + \eta)y_3^e + \frac{1}{4}(1 - \xi)(1 - \eta)y_4^e + \frac{1}{4}(1 + \xi)(1 + \eta)y_7^e + \frac{1}{4}(1 + \xi)(1 - \eta)y_8^e & , \\ z &= \frac{1}{4}(1 - \xi)(1 + \eta)z_3^e + \frac{1}{4}(1 - \xi)(1 - \eta)z_4^e + \frac{1}{4}(1 + \xi)(1 + \eta)z_7^e + \frac{1}{4}(1 + \xi)(1 - \eta)z_8^e & , \end{aligned}$$

which contain a bilinear term $\xi\eta$ responsible for the non-flatness of the resulting surface.⁴ As in the case of planar elements, it is straightforward to formulate higher-order three-dimensional isoparametric elements based, e.g., on the 10-noded tetrahedron or the 20-noded brick. In general, these higher-order elements have both curved edges and non-flat faces.

⁴Another 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.

DRAFT

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 point collocation-based ones, 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 (3.10) and (3.24). These integrals are ultimately evaluated as sums over integrals at the single element level, i.e., 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

$$I := \int_a^b f(x) dx ,$$

where f is a given real-valued functions and a, b are constant integration limits. The domain (a, b) of integration can be readily transformed into a mapped domain $(-1, 1)$ relative to a

new coordinate ξ , by merely setting

$$x = \frac{1}{2}(a+b) + \frac{1}{2}(b-a)\xi .$$

This transformation is one-to-one and onto as long as $a \neq b$ and it establishes symmetry of the domain relative to the origin. Taking into account the preceding transformation, one may write the original integral as

$$I = \int_a^b f(x) dx = \int_{-1}^1 f\left(\frac{1}{2}(a+b) + \frac{1}{2}(b-a)\xi\right) \frac{1}{2}(b-a) d\xi := \int_{-1}^1 g(\xi) d\xi .$$

Now, the integral is evaluated numerically as

$$I = \int_{-1}^1 g(\xi) d\xi \approx \sum_{k=1}^L w_k g(\xi_k) . \quad (6.1)$$

Here, ξ_k are the *sampling points* and w_k the corresponding *weights*. Equation (6.1)₂ encompasses virtually all the numerical integration methods used in one-dimensional finite elements.

Examples:

- (a) The classical *trapezoidal rule* can be expressed, in view of equation (6.1), as

$$I = \int_{-1}^1 g(\xi) d\xi \approx \sum_{k=1}^2 w_k g(\xi_k) ,$$

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

$$I = \int_{-1}^1 g(\xi) d\xi \approx \sum_{k=1}^3 w_k g(\xi_k) ,$$

where $w_1 = w_3 = \frac{1}{3}$, $w_2 = \frac{4}{3}$, and also $\xi_1 = -1$, $\xi_2 = 0$, 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, i.e., whether they furnish the maximum possible accuracy for the given cost. It turns out that the answer to this question is, in fact, negative. This can be argued as follows: recalling equation (6.1)₂, it is clear that it 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 integrate with such a formula a polynomial of degree q of the form

$$P(\xi) = a_0 + a_1\xi + \dots + a_q\xi^q ,$$

over the canonical domain $(-1, 1)$. This implies that

$$\int_{-1}^1 P(\xi) d\xi = \sum_{k=1}^L w_k g(\xi_k) ,$$

namely

$$\int_{-1}^1 (a_0 + a_1\xi + \dots + a_q\xi^q) d\xi = \sum_{k=1}^L w_k (a_0 + a_1\xi_k + \dots + a_q\xi_k^q) .$$

Since the constant coefficients a_k , $k = 0, 1, \dots, q$ are independent of each other and arbitrary, it follows from the above equation that

$$\left[\frac{\xi^{i+1}}{i+1} \right]_{-1}^1 = \sum_{k=1}^L w_k \xi_k^i = \begin{cases} \frac{2}{i+1} & i \text{ even} \\ 0 & i \text{ odd} \end{cases} , \quad i = 0, 1, \dots, q . \quad (6.2)$$

The $q + 1$ equations in (6.2) 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 (6.2) possesses such a unique solution, which yields the *Gaussian quadrature* rules, which are optimal in terms of accuracy and are used extensively in finite element implementations.

Consider now the first three examples of Gaussian quadrature corresponding to the cases $L = 1, 2, 3$.

- (a) When $L = 1$, it is clear that, owing to symmetry, $\xi_1 = 0$ and $w_1 = 2$. 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 equation (6.2) 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 equation (6.2) leads to $-\xi_1 = \xi_2 = \frac{1}{\sqrt{3}}$.

- (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 (6.2), one may find that $w_1 = w_3 = \frac{5}{9}$, $w_2 = \frac{8}{9}$, and $-\xi_1 = \xi_3 = \sqrt{\frac{3}{5}}$.

Similar results may be obtained for higher-order accurate Gaussian quadrature 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 integral as

$$I = \int_{\Omega^e} f(x, y) dx dy = \int_{-1}^1 \int_{-1}^1 f(x(\xi, \eta), y(\xi, \eta)) J d\xi d\eta := \int_{-1}^1 \int_{-1}^1 g(\xi, \eta) d\xi d\eta.$$

Since the coordinates ξ and η are independent, one may write

$$\int_{-1}^1 \int_{-1}^1 g(\xi, \eta) d\xi d\eta \approx \int_{-1}^1 \left\{ \sum_{k=1}^L w_k g(\xi_k, \eta) \right\} d\eta \approx \sum_{l=1}^L w_l \sum_{k=1}^L w_k g(\xi_k, \eta_l) = \sum_{k=1}^L \sum_{l=1}^L w_k w_l g(\xi_k, \eta_l).$$

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.

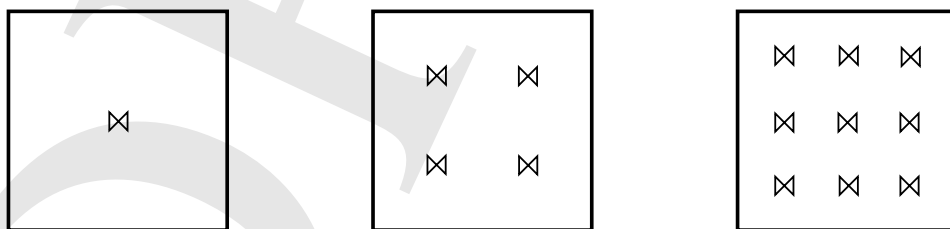


Figure 6.1: Two-dimensional Gauss quadrature rules for $q_1, q_2 \leq 1$ (left), $q_1, q_2 \leq 3$ (center), and $q_1, q_2 \leq 5$ (right)

Remark:

- It can be shown that the locations of the Gauss point in the domain $(-1, 1)$ are solutions of the *Legendre polynomials* P_k . These are defined by the recurrence formula

$$P_{k+1}(\xi) = \frac{(2k+1)\xi P_k(\xi) - kP_{k-1}(\xi)}{k+1},$$

with $P_0(\xi) = 1$ and $P_1(\xi) = \xi$. The Legendre polynomials satisfy the property

$$\int_{-1}^1 P_i(\xi)P_j(\xi) d\xi = \begin{cases} \frac{2}{i+1} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

This orthogonality 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

$$I = \int_{\Omega^e} g(L_1, L_2, L_3) dA \approx A \sum_{k=1}^L w_k g(L_{1k}, L_{2k}, L_{3k})$$

for straight-edge triangles of area A , or

$$I = \int_{\Omega^e} g(L_1, L_2, L_3, L_4) dA \approx V \sum_{k=1}^L w_k g(L_{1k}, L_{2k}, L_{3k}, L_{4k})$$

for straight-edge, flat face tetrahedra of volume V . Figure 6.2 depicts two simple integration rules for triangles.

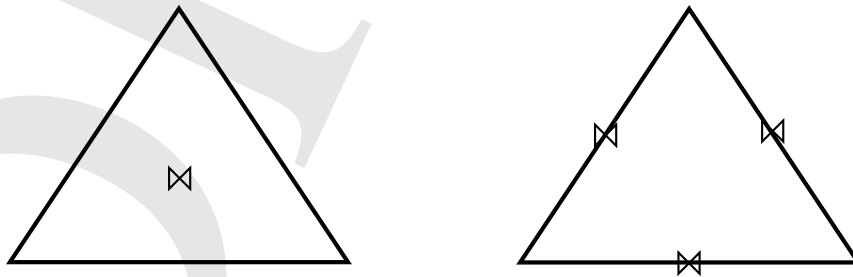


Figure 6.2: *Integration rules in triangular domains for $q \leq 1$ (left) and $q \leq 2$ (right)*

6.2 Assembly of global element arrays

The purpose of this section is to establish the manner in 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 applies over Ω , then it also applied 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 (3.10) can be written over the domain Ω^e of a given element, i.e.,

$$\int_{\Omega^e} \left[\frac{\partial w_h}{\partial x_1} k \frac{\partial u_h}{\partial x_1} + \frac{\partial w_h}{\partial x_2} k \frac{\partial u_h}{\partial x_2} + w_h f \right] d\Omega + \int_{\partial\Omega^e \cap \Gamma_q} w_h \bar{q} d\Gamma + \int_{\partial\Omega^e \setminus \partial\Omega} w_h q d\Gamma = 0. \quad (6.3)$$

The two boundary integral terms in (6.3) differ from those in (3.10). 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 of the domain Ω . The second boundary term refers to the interior part of the element boundary (i.e., the portion of the element boundary that is shared with other elements), which is subject to (yet unknown) Neumann boundary condition specifying the flux q between two neighboring elements.

Starting from equation (6.3), one may write corresponding element-wide weak forms for all finite elements and add them together. This leads to

$$\begin{aligned} \sum_e \int_{\Omega^e} \left[\frac{\partial w_h}{\partial x_1} k \frac{\partial u_h}{\partial x_1} + \frac{\partial w_h}{\partial x_2} k \frac{\partial u_h}{\partial x_2} + w_h f \right] d\Omega + \sum_e \int_{\partial\Omega^e \cap \Gamma_q} w_h \bar{q} d\Gamma \\ + \sum_{e,e' \text{ neighbors}} \int_{C_{e,e'}} w_h \llbracket q \rrbracket d\Gamma = 0, \quad (6.4) \end{aligned}$$

where $C_{e,e'}$ denotes an edge shared between two contiguous elements e and e' and $\llbracket q \rrbracket$ denotes the jump of the normal flux q from element e to element e' across $C_{e,e'}$. Equation (6.4) may be readily rewritten as

$$\begin{aligned} \int_{\Omega} \left[\frac{\partial w_h}{\partial x_1} k \frac{\partial u_h}{\partial x_1} + \frac{\partial w_h}{\partial x_2} k \frac{\partial u_h}{\partial x_2} + w_h f \right] d\Omega + \int_{\Gamma_q} w_h \bar{q} d\Gamma \\ + \sum_{e,e' \text{ neighbors}} \int_{C_{e,e'}} w_h \llbracket q \rrbracket d\Gamma = 0, \quad (6.5) \end{aligned}$$

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 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 (6.5) differs from the original form in (3.10) because of the introduction of the finite element fields (w_h, u_h) and the jump conditions in the last term of the left-hand side.

The finite element assembly operation entails the summation of all of the 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 (3.11). Hence, for each element, one may derive an equation of the form

$$\sum_{i=1}^n \beta_i \left(\sum_{j=1}^n K_{ij}^e u_j^e - F_i^e - F_i^{\text{int},e} \right) = 0, \quad (6.6)$$

where K_{ij}^e are the components of the element stiffness matrix, F_i^e are the components of the forcing vector contributed by the domain Ω^e and 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

$$\sum_{j=1}^n K_{ij}^e u_j^e - F_i^e - F_i^{\text{int},e} = 0, \quad i = 1, 2, \dots, n. \quad (6.7)$$

The assembly operation now amounts to combining all element-wide state equations of the form (6.7) into a global system that applies to the whole domain Ω . Symbolically, this can be represented by way of an assembly operator \mathbb{A}_e , such that

$$\mathbb{A}_e \left[\sum_{j=1}^n K_{ij}^e u_j^e - F_i^e - F_i^{\text{int},e} \right] = 0, \quad (6.8)$$

or

$$\sum_{J=1}^N K_{IJ} u_J - F_I = 0, \quad I = 1, 2, \dots, N. \quad (6.9)$$

Here,

$$[K_{IJ}] = \mathbb{A}_e[K_{ij}^e], \quad [F_I] = \mathbb{A}_e[F_i^e].$$

¹This is precisely why one cannot, in general, solve the original boundary-value problem on an direct element-by-element basis.

Note that the assembled contributions of the interior boundary fluxes are neglected, namely

$$[F_I^{\text{int}}] = \mathbb{A}_e[F_i^{\text{int},e}] = 0.$$

This assumption is made in finite element methods by necessity, although it is clear that it 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 (i.e., as the approximation becomes more accurate), the force contributions of these jumps tend to vanish.

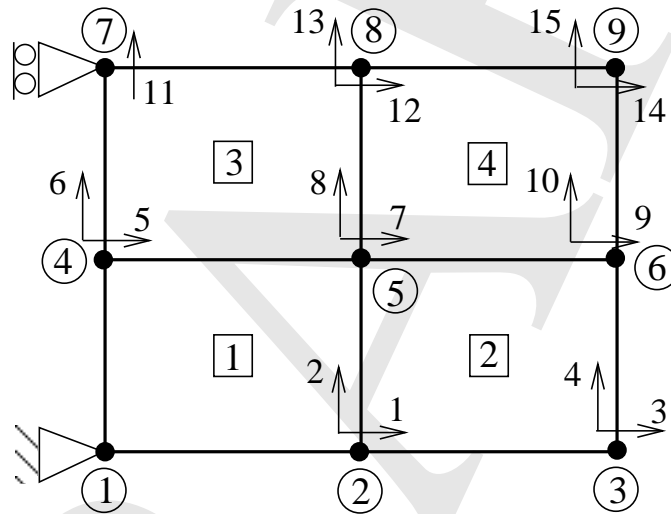


Figure 6.3: *Finite element mesh depicting global node and element numbering, as well as global degree of freedom assignments*

It is instructive here to illustrate the action of the assembly operator \mathbb{A}_e by means of an example. To this end, consider a simple finite element mesh of 4-noded rectangular elements with two-degrees of freedom per node, see Figure 6.3. All active degrees of freedom (i.e., those that are not fixed) are numbered in increasing order of the global node numbers. In this manner, one forms the ID array

$$[\text{ID}] = \begin{bmatrix} 0 & 1 & 3 & 5 & 7 & 9 & 0 & 12 & 14 \\ 0 & 2 & 4 & 6 & 8 & 10 & 11 & 13 & 15 \end{bmatrix} \quad (6.10)$$

by looping over all nodes. The dimension of this array is $\text{ndf} \times \text{numnp}$, 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. Taking into account the ID array and the local nodal numbering convention for 4-noded elements (see Figure 5.36), one may now generate the LM array as

$$[\text{LM}] = \begin{bmatrix} 0 & 1 & 5 & 7 \\ 0 & 2 & 6 & 8 \\ 1 & 3 & 7 & 9 \\ 2 & 4 & 8 & 10 \\ 7 & 9 & 12 & 14 \\ 8 & 10 & 13 & 15 \\ 5 & 7 & 0 & 12 \\ 6 & 8 & 11 & 13 \end{bmatrix} \quad (6.11)$$

by looping over all elements. The dimension of this array is $\mathbf{ndf} \times \mathbf{nen} \times \mathbf{nel}$, where \mathbf{nen} is the total number of nodes per element and \mathbf{nel} the total number of elements in the mesh. 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 $[K_{ij}^4]$ into the global stiffness array is reduced to identifying the correspondence between local and global degrees of freedom for each component of $[K_{ij}^4]$ by direct reference to the LM array. For instance, the entry K_{12}^4 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 (6.11). Likewise, the entry F_5^3 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 data structure shows that the task of assembling global arrays amounts to a simple reindexing of the local arrays with the aid of the LM array.

6.3 Algebraic equation solving by Gaussian elimination and its variants

The system of linear algebraic equations (6.9) 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.

approximately

$$\frac{\frac{2}{3}(10^5)^3}{100 \times 10^6} = \frac{2}{3}10^7 \text{ sec} \approx 77 \text{ days} .$$

On the other hand, using a banded Gauss elimination solver requires

$$\frac{2(10^3)^2(10^5)}{100 \times 10^6} = 2 \times 10^3 \text{ sec} \approx 1/2 \text{ hour} .$$

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.

6.4 Finite element modeling: mesh design and generation

Finite element modeling is a relatively complex undertaking. It requires:

- 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?).
- 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, 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 symmetry(ies) 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.

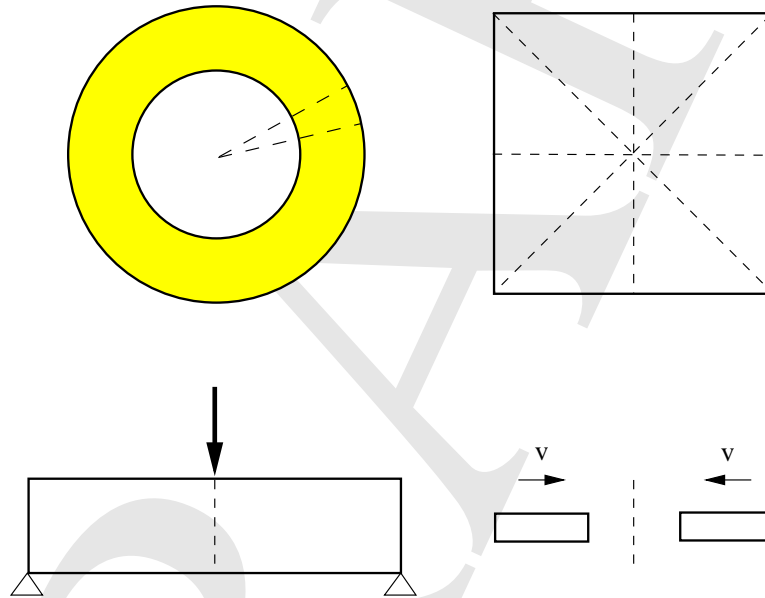


Figure 6.5: *Representative examples of symmetries in the domains of differential equations (corresponding symmetries in the boundary conditions, loading, and equations themselves are assumed)*

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, as in Figure 6.6, where the nodes of the same mesh are numbered in two distinct and regular ways, i.e., in 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 $b_B = 7$.

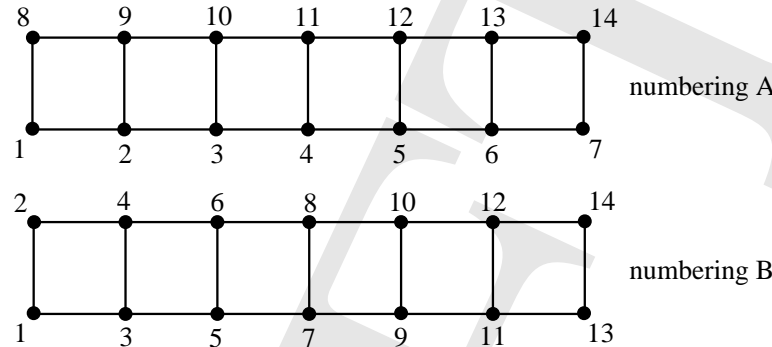


Figure 6.6: *Two possible ways of node numbering in a finite element mesh*

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-alone 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. At the same time, there is a reverse trend toward diversification, where analysts use different software products for different tasks (e.g., separate software for pre-processing, solution and post-processing).

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 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 $\partial\Omega$ be smooth with outward unit normal \mathbf{n} , and be decomposed into two regions Γ_u and Γ_q , such that $\partial\Omega = \overline{\Gamma_u \cup \Gamma_q}$. Further, assume that there is a vector function $\mathbf{u} : \bar{\Omega} \times \mathbb{R}^+ \mapsto \mathbb{R}^3$, such that the position vector \mathbf{x} 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

$$\mathbf{x}(\mathbf{X}, t) = \mathbf{X} + \mathbf{u}(\mathbf{X}, t) .$$

The vector function \mathbf{u} is referred to as the *displacement* field. The body is assumed to be

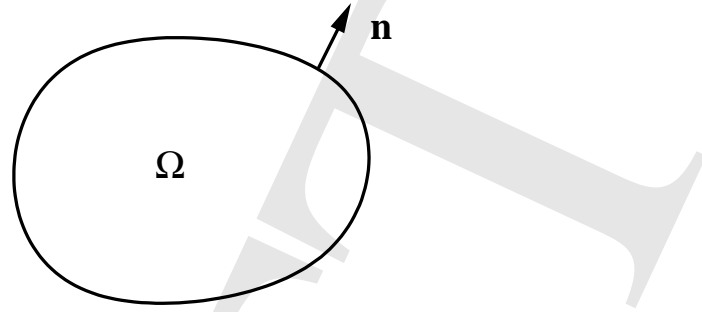


Figure 7.1: *The domain Ω of the linear elastostatics problem*

made of an elastic material and is subject to body force \mathbf{f} per unit 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 $\bar{\mathbf{u}}$ are assumed continuous.

The strong form of the equations of linear elastostatics are written as:

$$\begin{aligned} \nabla \cdot \boldsymbol{\sigma} + \mathbf{f} &= \mathbf{0} && \text{in } \Omega, \\ \boldsymbol{\sigma} \mathbf{n} &= \bar{\mathbf{t}} && \text{on } \Gamma_q, \\ \mathbf{u} &= \bar{\mathbf{u}} && \text{on } \Gamma_u, \end{aligned} \quad (6.5)$$

where $\boldsymbol{\sigma}$ is the stress tensor and $\nabla \cdot \boldsymbol{\sigma}$ denotes the divergence of $\boldsymbol{\sigma}$. Here, (7.1)₁ are the equations of equilibrium for the body, while (7.1)_{2,3} are the Neumann and Dirichlet boundary conditions, respectively.

Assuming that the material is isotropic and homogeneous, one may express the stress tensor as

$$\boldsymbol{\sigma} = \lambda \operatorname{tr} \boldsymbol{\epsilon} \mathbf{I} + 2\mu \boldsymbol{\epsilon}, \quad (7.2)$$

in terms of the Lamé constants λ and μ ($\lambda + \frac{2}{3}\mu > 0$, $\mu > 0$), the identity tensor \mathbf{I} , and the infinitesimal strain tensor $\boldsymbol{\epsilon}$. The latter is defined as

$$\boldsymbol{\epsilon} := \frac{1}{2}[\nabla \mathbf{u} + (\nabla \mathbf{u})^T] := \nabla_s \mathbf{u}, \quad (7.3)$$

where $\nabla \mathbf{u}$ is the gradient of \mathbf{u} expressed in Cartesian component form as

$$[\nabla \mathbf{u}] = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \begin{bmatrix} \frac{\partial}{\partial X_1} & \frac{\partial}{\partial X_2} & \frac{\partial}{\partial X_3} \end{bmatrix} = \begin{bmatrix} u_{1,1} & u_{1,2} & u_{1,3} \\ u_{2,1} & u_{2,2} & u_{2,3} \\ u_{3,1} & u_{3,2} & u_{3,3} \end{bmatrix},$$

where $u_{i,j} := \frac{\partial u_i}{\partial X_j}$. Taking into account equation (7.3), it follows that the components of the strain tensor are

$$[\boldsymbol{\epsilon}] = \begin{bmatrix} u_{1,1} & \frac{1}{2}(u_{1,2} + u_{2,1}) & \frac{1}{2}(u_{1,3} + u_{3,1}) \\ \frac{1}{2}(u_{1,2} + u_{2,1}) & u_{2,2} & \frac{1}{2}(u_{2,3} + u_{3,2}) \\ \frac{1}{2}(u_{1,3} + u_{3,1}) & \frac{1}{2}(u_{2,3} + u_{3,2}) & u_{3,3} \end{bmatrix}.$$

Clearly, the strain tensor is symmetric and, in view of equation (7.2), so is the stress tensor.

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 (7.1) are satisfied. The precise sense in which equations (7.1) 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}_Ω and \mathbf{w}_q are chosen to be identical, i.e., that $\mathbf{w}_\Omega = \mathbf{w}_q = \mathbf{w}$, and (c) that $\mathbf{w} = \mathbf{0}$ on Γ_u . Taking into account the preceding conditions, the weak form may be written as

$$\int_{\Omega} \mathbf{w} \cdot (-\nabla \cdot \boldsymbol{\sigma} - \mathbf{f}) \, d\Omega + \int_{\Gamma_q} \mathbf{w} \cdot (\boldsymbol{\sigma} \mathbf{n} - \bar{\mathbf{t}}) \, d\Gamma = 0. \quad (7.4)$$

Concentrating on the first term of the left-hand side of equation (7.4), use the Einsteinian summation convention¹ to write

$$\begin{aligned} \int_{\Omega} \mathbf{w} \cdot (\nabla \cdot \boldsymbol{\sigma}) \, d\Omega &= \int_{\Omega} w_i \sigma_{ij,j} \, d\Omega \\ &= \int_{\Omega} (w_i \sigma_{ij})_{,j} \, d\Omega - \int_{\Omega} w_{i,j} \sigma_{ij} \, d\Omega \\ &= \int_{\partial\Omega} w_i \sigma_{ij} n_j \, d\Gamma - \int_{\Omega} w_{i,j} \sigma_{ij} \, d\Omega \\ &= \int_{\Gamma_q} w_i \sigma_{ij} n_j \, d\Gamma - \int_{\Omega} w_{i,j} \sigma_{ij} \, d\Omega, \end{aligned} \quad (7.5)$$

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

$$w_{i,j} \sigma_{ij} = \left[\frac{1}{2}(w_{i,j} + w_{j,i}) + \frac{1}{2}(w_{i,j} - w_{j,i}) \right] \sigma_{ij} = \frac{1}{2}(w_{i,j} + w_{j,i}) \sigma_{ij}. \quad (7.6)$$

¹According 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.

Taking into account equations (7.5) and (7.6), it follows that

$$\int_{\Omega} \mathbf{w} \cdot (\nabla \cdot \boldsymbol{\sigma}) d\Omega = \int_{\Gamma_q} \mathbf{w} \cdot (\boldsymbol{\sigma} \mathbf{n}) d\Gamma - \int_{\Omega} \nabla_s \mathbf{w} : \boldsymbol{\sigma} d\Omega, \quad (7.7)$$

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 (7.7), the weak form of equation (7.4) can be rewritten as

$$\int_{\Omega} \nabla_s \mathbf{w} : \boldsymbol{\sigma} d\Omega = \int_{\Omega} \mathbf{w} \cdot \mathbf{f} d\Omega + \int_{\Gamma_q} \mathbf{w} \cdot \bar{\mathbf{t}} d\Gamma. \quad (7.8)$$

Given \mathbf{f} , $\bar{\mathbf{t}}$, $\bar{\mathbf{u}}$, and the stress-strain law (7.2), the weak form of the problem of linear elastostatics amounts to finding $\mathbf{u} \in \mathcal{U}$ such that equation (7.8) holds for all admissible $\mathbf{w} \in \mathcal{W}$. Here, the admissible spaces \mathcal{U} and \mathcal{W} are defined as

$$\begin{aligned} \mathcal{U} &= \left\{ \mathbf{u} \in H^1(\Omega) \mid \mathbf{u} = \bar{\mathbf{u}} \text{ on } \Gamma_u \right\}, \\ \mathcal{W} &= \left\{ \mathbf{w} \in H^1(\Omega) \mid \mathbf{w} = \mathbf{0} \text{ on } \Gamma_u \right\}. \end{aligned}$$

Adopting the terminology of “virtual” displacements, equation (7.8) can be viewed as a statement of the theorem of virtual work, according to which the work done by the actual internal forces (i.e., the stress $\boldsymbol{\sigma}$) over the “virtual” strains $\nabla_s \mathbf{w}$ is equal to the work done by the actual external forces (i.e., the body force \mathbf{f} and surface traction $\bar{\mathbf{t}}$) over the “virtual” displacement \mathbf{w} .

The weak form (7.8) can be written operationally as

$$B(\mathbf{w}, \mathbf{u}) = (\mathbf{w}, \mathbf{f}) + (\mathbf{w}, \bar{\mathbf{t}})_{\Gamma_q},$$

where

$$\begin{aligned} B(\mathbf{w}, \mathbf{u}) &:= \int_{\Omega} \nabla_s \mathbf{w} : \boldsymbol{\sigma} d\Omega, \\ (\mathbf{w}, \mathbf{f}) &:= \int_{\Omega} \mathbf{w} \cdot \mathbf{f} d\Omega, \\ (\mathbf{w}, \bar{\mathbf{t}})_{\Gamma_q} &:= \int_{\Gamma_q} \mathbf{w} \cdot \bar{\mathbf{t}} d\Gamma. \end{aligned} \quad (7.9)$$

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

$$[\boldsymbol{\epsilon}] = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{bmatrix}$$

and rewrite them with a slight abuse of notation as

$$[\boldsymbol{\epsilon}] = \begin{bmatrix} \epsilon_{11} & \epsilon_{22} & \epsilon_{33} & 2\epsilon_{12} & 2\epsilon_{23} & 2\epsilon_{31} \end{bmatrix}^T .$$

Likewise, the 3×3 symmetric matrix of components of the stress tensor

$$[\boldsymbol{\sigma}] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}$$

are written in vector form as

$$[\boldsymbol{\sigma}] = \begin{bmatrix} \sigma_{11} & \sigma_{22} & \sigma_{33} & \sigma_{12} & \sigma_{23} & \sigma_{31} \end{bmatrix}^T .$$

Note that the factor “2” in the last three rows of the strain vector is included in order to ensure that the contraction $\boldsymbol{\sigma} : \boldsymbol{\epsilon} = \sigma_{ij}\epsilon_{ij}$ is defined consistently when employing the vector convention. The preceding vector notation is employed in the remainder of this section.

The stress-strain law (7.2) can be written using the vector convention as

$$[\boldsymbol{\sigma}] = [\mathbf{D}][\boldsymbol{\epsilon}] , \quad (7.10)$$

where $[\mathbf{D}]$ is a (6×6) elasticity matrix such that

$$\begin{bmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{33} \\ \sigma_{12} \\ \sigma_{23} \\ \sigma_{31} \end{bmatrix} = \begin{bmatrix} \lambda + 2\mu & \lambda & \lambda & 0 & 0 & 0 \\ \lambda & \lambda + 2\mu & \lambda & 0 & 0 & 0 \\ \lambda & \lambda & \lambda + 2\mu & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu \end{bmatrix} \begin{bmatrix} \epsilon_{11} \\ \epsilon_{22} \\ \epsilon_{33} \\ 2\epsilon_{12} \\ 2\epsilon_{23} \\ 2\epsilon_{31} \end{bmatrix} .$$

In the special case of plane strain on the $(1, 2)$ plane, the preceding system reduces to

$$\begin{bmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{12} \end{bmatrix} = \begin{bmatrix} \lambda + 2\mu & \lambda & 0 \\ \lambda & \lambda + 2\mu & 0 \\ 0 & 0 & \mu \end{bmatrix} \begin{bmatrix} \epsilon_{11} \\ \epsilon_{22} \\ 2\epsilon_{12} \end{bmatrix} .$$

Lastly, in the special case of plane stress on the $1 - 2$ plane, one may write the stress-strain relations in reduced matrix form as

$$\begin{bmatrix} \sigma_{11} \\ \sigma_{22} \\ \sigma_{12} \end{bmatrix} = \begin{bmatrix} \frac{4\mu(\lambda + \mu)}{\lambda + 2\mu} & \frac{2\lambda\mu}{\lambda + 2\mu} & 0 \\ \frac{2\lambda\mu}{\lambda + 2\mu} & \frac{4\mu(\lambda + \mu)}{\lambda + 2\mu} & 0 \\ 0 & 0 & \mu \end{bmatrix} \begin{bmatrix} \epsilon_{11} \\ \epsilon_{22} \\ 2\epsilon_{12} \end{bmatrix} ,$$

and $\sigma_{33} = \lambda(\epsilon_{11} + \epsilon_{22})$.

Since the matrix $[\mathbf{D}]$ is always symmetric, it follows that the integrand of the bilinear form in (7.9) can be written with the aid of (7.10) as

$$\nabla_s \mathbf{w} : \boldsymbol{\sigma} = [\boldsymbol{\epsilon}(\mathbf{w})][\mathbf{D}][\boldsymbol{\epsilon}(\mathbf{u})] := \boldsymbol{\epsilon}(\mathbf{w}) \cdot \mathbf{D}\boldsymbol{\epsilon}(\mathbf{u}) ,$$

which shows that the bilinear form in (7.9) is indeed symmetric. This, in turn, implies that Vainberg's theorem is applicable and that there exists a functional $I[\mathbf{u}]$, given by

$$\begin{aligned} I[\mathbf{u}] &= \frac{1}{2}B(\mathbf{u}, \mathbf{u}) - (\mathbf{u}, \mathbf{f}) - (\mathbf{u}, \bar{\mathbf{t}})_{\Gamma_q} \\ &= \frac{1}{2} \int_{\Omega} \boldsymbol{\epsilon}(\mathbf{u}) \cdot \mathbf{D}\boldsymbol{\epsilon}(\mathbf{u}) d\Omega - \int_{\Omega} \mathbf{u} \cdot \mathbf{f} d\Omega - \int_{\Gamma_q} \mathbf{u} \cdot \bar{\mathbf{t}} d\Gamma . \end{aligned} \quad (7.11)$$

The first term on the right-hand side of (7.11) 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 (7.11) is referred to as the *total potential energy* of the 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

$$I[\mathbf{u}] = \int_{\Omega} \boldsymbol{\epsilon}(\delta\mathbf{u}) \cdot \mathbf{D}\boldsymbol{\epsilon}(\mathbf{u}) d\Omega - \int_{\Omega} \delta\mathbf{u} \cdot \mathbf{f} d\Omega - \int_{\Gamma_q} \delta\mathbf{u} \cdot \bar{\mathbf{t}} d\Gamma = 0 , \quad (7.12)$$

which coincides with the weak form (7.8). Furthermore, given any $\delta\mathbf{u} \in \mathcal{W}$, it is easy to conclude with the aid of (7.12) that

$$I[\mathbf{u} + \delta\mathbf{u}] - I[\mathbf{u}] = \int_{\Omega} \boldsymbol{\epsilon}(\delta\mathbf{u}) \cdot \mathbf{D}\boldsymbol{\epsilon}(\delta\mathbf{u}) d\Omega . \quad (7.13)$$

The right-hand side of (7.13) is necessarily positive for $\delta\mathbf{u} \neq \mathbf{0}$, given that \mathbf{D} is a positive-definite matrix, as its eigenvalues are $\lambda_1 = 3\lambda + 2\mu (> 0)$, are $\lambda_{2,3} = 2\mu (> 0)$, and $\lambda_{4,5,6} = \mu (> 0)$.

7.2.1 A Galerkin approximation to the weak form

The discrete counterpart of (7.8) can be written as

$$\int_{\Omega} \boldsymbol{\epsilon}(\mathbf{w}_h) \cdot \mathbf{D}\boldsymbol{\epsilon}(\mathbf{u}_h) d\Omega = \int_{\Omega} \mathbf{w}_h \cdot \mathbf{f} d\Omega + \int_{\Gamma_q} \mathbf{w}_h \cdot \bar{\mathbf{t}} d\Gamma , \quad (7.14)$$

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

$$\mathbf{u}_h = \sum_{i=1}^{\text{nen}} N_i^e \mathbf{u}_i^e, \quad \mathbf{w}_h = \sum_{i=1}^{\text{nen}} N_i^e \mathbf{w}_i^e, \quad (7.15)$$

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 . Equations (7.15) can be written compactly as

$$\mathbf{u}_h = \mathbf{N}^e \mathbf{u}^e, \quad \mathbf{w}_h = \mathbf{N}^e \mathbf{w}^e, \quad (7.16)$$

in terms of the $\text{ndf} \times \text{nen}$ vectors

$$\mathbf{u}^e := \begin{bmatrix} \mathbf{u}_1^e \\ \mathbf{u}_2^e \\ \cdot \\ \cdot \\ \mathbf{u}_{\text{nen}}^e \end{bmatrix}, \quad \mathbf{w}^e := \begin{bmatrix} \mathbf{w}_1^e \\ \mathbf{w}_2^e \\ \cdot \\ \cdot \\ \mathbf{w}_{\text{nen}}^e \end{bmatrix},$$

and the $\text{ndf} \times \text{ndf} \times \text{nen}$ matrix

$$\mathbf{N}^e := \begin{bmatrix} N_1^e \mathbf{I}_{\text{ndf}} & N_2^e \mathbf{I}_{\text{ndf}} & \cdot & \cdot & N_{\text{nen}}^e \mathbf{I}_{\text{ndf}} \end{bmatrix},$$

and \mathbf{I}_{ndf} is the $\text{ndf} \times \text{ndf}$ identity matrix.

The strain tensor, expressed in vector form, can be also written as

$$\begin{bmatrix} \epsilon_{11} \\ \epsilon_{22} \\ \epsilon_{33} \\ 2\epsilon_{12} \\ 2\epsilon_{23} \\ 2\epsilon_{31} \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x_1} & 0 & 0 \\ 0 & \frac{\partial}{\partial x_2} & 0 \\ 0 & 0 & \frac{\partial}{\partial x_3} \\ \frac{\partial}{\partial x_2} & \frac{\partial}{\partial x_1} & 0 \\ \frac{\partial}{\partial x_2} & \frac{\partial}{\partial x_1} & 0 \\ 0 & \frac{\partial}{\partial x_3} & \frac{\partial}{\partial x_2} \\ \frac{\partial}{\partial x_3} & 0 & \frac{\partial}{\partial x_1} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}.$$

This implies that the strains in Ω^e are given by

$$\boldsymbol{\epsilon}(\mathbf{u}_h) = \sum_{i=1}^{\text{nen}} \mathbf{B}_i^e \mathbf{u}_i^e, \quad \boldsymbol{\epsilon}(\mathbf{w}_h) = \sum_{i=1}^{\text{nen}} \mathbf{B}_i^e \mathbf{w}_i^e, \quad (7.17)$$

in terms of the $6 \times \text{ndf}$ strain-displacement matrix

$$[\mathbf{B}_i^e] := \begin{bmatrix} \frac{\partial N_i^e}{\partial x_1} & 0 & 0 \\ 0 & \frac{\partial N_i^e}{\partial x_2} & 0 \\ 0 & 0 & \frac{\partial N_i^e}{\partial x_3} \\ \frac{\partial N_i^e}{\partial x_2} & \frac{\partial N_i^e}{\partial x_1} & 0 \\ 0 & \frac{\partial N_i^e}{\partial x_3} & \frac{\partial N_i^e}{\partial x_2} \\ \frac{\partial N_i^e}{\partial x_3} & 0 & \frac{\partial N_i^e}{\partial x_1} \end{bmatrix} .$$

Again, resorting to compact notation, equation (7.17) can be recast in the form

$$\boldsymbol{\epsilon}(\mathbf{u}_h) = \mathbf{B}^e \mathbf{u}^e \quad , \quad \boldsymbol{\epsilon}(\mathbf{w}_h) = \mathbf{B}^e \mathbf{w}^e \quad , \quad (7.18)$$

where \mathbf{B}^e is a $6 \times \text{ndf} \times \text{nen}$ matrix defined as

$$\mathbf{B}^e := \left[\mathbf{B}_1^e \quad \mathbf{B}_2^e \quad \cdot \quad \cdot \quad \mathbf{B}_{\text{nen}}^e \right] .$$

The weak form (7.14) can be applied to element e , such that

$$\int_{\Omega^e} \boldsymbol{\epsilon}(\mathbf{w}_h) \cdot \mathbf{D} \boldsymbol{\epsilon}(\mathbf{u}_h) d\Omega = \int_{\Omega^e} \mathbf{w}_h \cdot \mathbf{f} d\Omega + \int_{\partial\Omega^e \cap \Gamma_q} \mathbf{w}_h \cdot \bar{\mathbf{t}} d\Gamma + \int_{\partial\Omega^e \setminus \partial\Omega} \mathbf{w}_h \cdot \mathbf{t} d\Gamma . \quad (7.19)$$

Appealing to equations (7.16) and (7.18), the preceding weak form is written as

$$\begin{aligned} & \int_{\Omega^e} (\mathbf{B}^e \mathbf{w}^e)^T \mathbf{D} (\mathbf{B}^e \mathbf{u}^e) d\Omega \\ & = \int_{\Omega^e} (\mathbf{N}^e \mathbf{w}^e)^T \mathbf{f} d\Omega + \int_{\partial\Omega^e \cap \Gamma_q} (\mathbf{N}^e \mathbf{w}^e)^T \bar{\mathbf{t}} d\Gamma + \int_{\partial\Omega^e \setminus \partial\Omega} (\mathbf{N}^e \mathbf{w}^e)^T \mathbf{t} d\Gamma \end{aligned} \quad (7.20)$$

or

$$\begin{aligned} & \mathbf{w}^{e,T} \left[\left\{ \int_{\Omega^e} \mathbf{B}^{e,T} \mathbf{D} \mathbf{B}^e d\Omega \right\} \mathbf{u}^e \right. \\ & \quad \left. - \int_{\Omega^e} \mathbf{N}^{e,T} \mathbf{f} d\Omega - \int_{\partial\Omega^e \cap \Gamma_q} \mathbf{N}^{e,T} \bar{\mathbf{t}} d\Gamma - \int_{\partial\Omega^e \setminus \partial\Omega} \mathbf{N}^{e,T} \mathbf{t} d\Gamma \right] = 0 . \end{aligned} \quad (7.21)$$

Given the arbitrariness of \mathbf{w}^e , equation (7.21) leads to the linear system

$$\mathbf{K}^e \mathbf{u}^e = \mathbf{F}^e + \mathbf{F}^{\text{int},e} ,$$

where

$$\begin{aligned}
 \mathbf{K}^e &= \int_{\Omega^e} \mathbf{B}^{e,T} \mathbf{D} \mathbf{B}^e d\Omega, \\
 \mathbf{F}^e &= \int_{\Omega^e} \mathbf{N}^{e,T} \mathbf{f} d\Omega + \int_{\partial\Omega^e \cap \Gamma_q} \mathbf{N}^{e,T} \bar{\mathbf{t}} d\Gamma, \\
 \mathbf{F}^{\text{int},e} &= \int_{\partial\Omega^e \setminus \partial\Omega} \mathbf{N}^{e,T} \mathbf{t} d\Gamma.
 \end{aligned} \tag{7.22}$$

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: (4-noded isoparametric quadrilateral in plane strain)

The element interpolation functions N_i^e , $i = 1 - 4$, for this element are given in equation (5.26) relative to the natural coordinates (ξ, η) . The element interpolation array \mathbf{N}^e is now given by

$$\mathbf{N}^e := \left[N_1^e \mathbf{I}_2 \quad N_2^e \mathbf{I}_2 \quad N_3^e \mathbf{I}_2 \quad N_4^e \mathbf{I}_2 \right],$$

and is of dimension 2×8 (note that here $\text{nen} = 4$ and $\text{ndf} = 2$). Moreover, the strain-displacement array B_i^e is given by

$$[\mathbf{B}_i^e] = \begin{bmatrix} \frac{\partial N_i^e}{\partial x_1} & 0 \\ 0 & \frac{\partial N_i^e}{\partial x_2} \\ \frac{\partial N_i^e}{\partial x_2} & \frac{\partial N_i^e}{\partial x_1} \end{bmatrix},$$

hence

$$\mathbf{B}^e = \begin{bmatrix} \frac{\partial N_1^e}{\partial x_1} & 0 & \frac{\partial N_2^e}{\partial x_1} & 0 & \frac{\partial N_3^e}{\partial x_1} & 0 & \frac{\partial N_4^e}{\partial x_1} & 0 \\ 0 & \frac{\partial N_1^e}{\partial x_2} & 0 & \frac{\partial N_2^e}{\partial x_2} & 0 & \frac{\partial N_3^e}{\partial x_2} & 0 & \frac{\partial N_4^e}{\partial x_2} \\ \frac{\partial N_1^e}{\partial x_2} & \frac{\partial N_1^e}{\partial x_1} & \frac{\partial N_2^e}{\partial x_2} & \frac{\partial N_2^e}{\partial x_1} & \frac{\partial N_3^e}{\partial x_2} & \frac{\partial N_3^e}{\partial x_1} & \frac{\partial N_4^e}{\partial x_2} & \frac{\partial N_4^e}{\partial x_1} \end{bmatrix}. \tag{7.23}$$

Given that the elasticity matrix \mathbf{D} is of dimension 3×3 (see earlier discussion of the plane strain case), it follows from (7.23) that the element stiffness matrix \mathbf{K}^e in (7.22)₁ is of dimension 8×8 .

7.2.2 On the order of numerical integration

The stiffness matrix and forcing vector in equations (7.22)_{1,2} require the evaluation of domain and boundary integrals. In the case of isoparametric elements, the stiffness matrix 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 with respect to the physical coordinates (x_1, x_2, x_3) . Appealing to the chain rule, a typical such derivative $\frac{\partial N_i^e}{\partial x_j}$ can be written as

$$\frac{\partial N_i^e}{\partial x_j} = \frac{\partial N_i^e}{\partial \xi} \frac{\partial \xi}{\partial x_j} + \frac{\partial N_i^e}{\partial \eta} \frac{\partial \eta}{\partial x_j} + \frac{\partial N_i^e}{\partial \zeta} \frac{\partial \zeta}{\partial x_j}. \quad (7.24)$$

While terms of the type $\frac{\partial N_i^e}{\partial \xi}$ in equation (7.24) are clearly polynomial in (ξ, η, ζ) , this is not the case with terms of the type $\frac{\partial \xi}{\partial x_j}$, which are, in fact, inverse polynomial in (ξ, η, ζ) .

To find an analytical expression for derivatives of the type $\frac{\partial N_i^e}{\partial x_j}$, write

$$\begin{aligned} \frac{\partial N_i^e}{\partial \xi} &= \frac{\partial N_i^e}{\partial x_1} \frac{\partial x_1}{\partial \xi} + \frac{\partial N_i^e}{\partial x_2} \frac{\partial x_2}{\partial \xi} + \frac{\partial N_i^e}{\partial x_3} \frac{\partial x_3}{\partial \xi}, \\ \frac{\partial N_i^e}{\partial \eta} &= \frac{\partial N_i^e}{\partial x_1} \frac{\partial x_1}{\partial \eta} + \frac{\partial N_i^e}{\partial x_2} \frac{\partial x_2}{\partial \eta} + \frac{\partial N_i^e}{\partial x_3} \frac{\partial x_3}{\partial \eta}, \\ \frac{\partial N_i^e}{\partial \zeta} &= \frac{\partial N_i^e}{\partial x_1} \frac{\partial x_1}{\partial \zeta} + \frac{\partial N_i^e}{\partial x_2} \frac{\partial x_2}{\partial \zeta} + \frac{\partial N_i^e}{\partial x_3} \frac{\partial x_3}{\partial \zeta}, \end{aligned}$$

or, in matrix form

$$\begin{bmatrix} \frac{\partial N_i^e}{\partial \xi} \\ \frac{\partial N_i^e}{\partial \eta} \\ \frac{\partial N_i^e}{\partial \zeta} \end{bmatrix} = \underbrace{\begin{bmatrix} \frac{\partial x_1}{\partial \xi} & \frac{\partial x_2}{\partial \xi} & \frac{\partial x_3}{\partial \xi} \\ \frac{\partial x_1}{\partial \eta} & \frac{\partial x_2}{\partial \eta} & \frac{\partial x_3}{\partial \eta} \\ \frac{\partial x_1}{\partial \zeta} & \frac{\partial x_2}{\partial \zeta} & \frac{\partial x_3}{\partial \zeta} \end{bmatrix}}_{\mathbf{J}^e} \begin{bmatrix} \frac{\partial N_i^e}{\partial x_1} \\ \frac{\partial N_i^e}{\partial x_2} \\ \frac{\partial N_i^e}{\partial x_3} \end{bmatrix}. \quad (7.25)$$

Equation (7.25) demonstrates that the computation of partial derivatives of the type $\frac{\partial N_i^e}{\partial x_j}$ requires inversion of the 3×3 Jacobian matrix $[\mathbf{J}^e]$, which is equal to $\frac{1}{J^e} \text{adj}(\mathbf{J}^e)$, where $\text{adj}(\mathbf{J}^e)$ is the adjugate of \mathbf{J}^e . Given that J^e is the product of polynomials in (ξ, η, ζ) , the presence of the determinant J^e in the denominator of $[\mathbf{J}^e]^{-1}$ establishes the rational polynomial form of $\frac{\partial N_i^e}{\partial x_j}$ (hence, also of \mathbf{B}^e and \mathbf{K}^e).

Exact integration of \mathbf{K}^e is possible, yet 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.

(a) *Minimum order of integration for completeness*

Recalling the definition of the bilinear form $B(\cdot, \cdot)$ in equation (7.9)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 $q \geq 1$. 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 is constant within the element, the bilinear form becomes

$$B(\mathbf{w}_h, \mathbf{u}_h) = \int_{\Omega^e} \bar{\boldsymbol{\epsilon}}^T(\mathbf{w}) \mathbf{D} \bar{\boldsymbol{\epsilon}}(\mathbf{u}) d\Omega = \bar{\boldsymbol{\epsilon}}^T(\mathbf{w}) \mathbf{D} \bar{\boldsymbol{\epsilon}}(\mathbf{u}) \int_{\Omega^e} d\Omega ,$$

where $\bar{\boldsymbol{\epsilon}}(\mathbf{w})$ and $\bar{\boldsymbol{\epsilon}}(\mathbf{u})$ are constant. 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

$$\int_{\Omega^e} d\Omega = \int_{\Omega_0^e} J^e d\xi d\eta d\zeta ,$$

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-noded 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 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.

To illustrate the above point, consider again the 4-noded 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

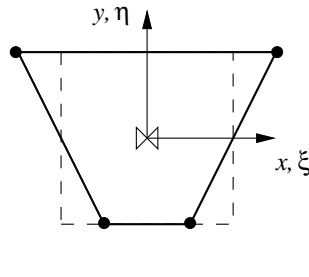


Figure 7.2: Zero-energy modes for the 4-noded quadrilateral with 1×1 Gaussian quadrature

energy. Indeed, letting for simplicity the natural and physical domains and coordinates coincide, the displacement and strain vector associated with one of these modes is

$$[\mathbf{u}_h] = \begin{bmatrix} \alpha \xi \eta \\ 0 \end{bmatrix}, \quad [\boldsymbol{\epsilon}] = \begin{bmatrix} \alpha \eta \\ 0 \\ \alpha \xi \end{bmatrix}, \quad (7.26)$$

where $\alpha (> 0)$ is a constant. The strain energy of this deformation mode is

$$\begin{aligned} W &= \frac{1}{2} \int_{\Omega^e} \boldsymbol{\epsilon}^T(\mathbf{u}_h) \mathbf{D} \boldsymbol{\epsilon}(\mathbf{u}_h) d\Omega \\ &= \frac{\alpha^2}{2} \int_{\Omega^e} \begin{bmatrix} \eta & 0 & \xi \end{bmatrix} \begin{bmatrix} \lambda + 2\mu & \lambda & 0 \\ \lambda & \lambda + 2\mu & 0 \\ 0 & 0 & \mu \end{bmatrix} \begin{bmatrix} \eta \\ 0 \\ \xi \end{bmatrix} d\Omega > 0. \end{aligned}$$

However, when using 1×1 Gauss quadrature, it is readily seen that the strain energy of this mode is zero (recall that the 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*. This zero energy mode of equation (7.26) disappears upon using 2×2 Gaussian quadrature, since, in this case its strain energy is approximated by

$$W \approx \frac{\alpha^2}{2} \sum_{l=1}^4 \begin{bmatrix} \eta_l & 0 & \xi_l \end{bmatrix} \begin{bmatrix} \lambda + 2\mu & \lambda & 0 \\ \lambda & \lambda + 2\mu & 0 \\ 0 & 0 & \mu \end{bmatrix} \begin{bmatrix} \eta_l \\ 0 \\ \xi_l \end{bmatrix} > 0,$$

where $(\xi_l, \eta_l) = (\pm \frac{1}{\sqrt{3}}, \pm \frac{1}{\sqrt{3}})$. 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-noded isoparametric quadrilateral elements with 2×2 Gaussian quadrature. Here, the deformation mode

$$[\mathbf{u}_h] = \begin{bmatrix} \alpha\xi(\eta^2 - \frac{1}{3}) \\ -\alpha\eta(\xi^2 - \frac{1}{3}) \end{bmatrix}, \quad [\boldsymbol{\epsilon}] = \begin{bmatrix} \alpha(\eta^2 - \frac{1}{3}) \\ -\alpha(\xi^2 - \frac{1}{3}) \\ 0 \end{bmatrix}, \quad (7.27)$$

with $\alpha > 0$, is obviously associated with positive strain energy, see Figure 7.3. However, using 2×2 Gaussian quadrature

$$W \approx \frac{\alpha^2}{2} \sum_{l=1}^4 \begin{bmatrix} \eta_l^2 - \frac{1}{3} & -(\xi_l^2 - \frac{1}{3}) & 0 \end{bmatrix} \begin{bmatrix} \lambda + 2\mu & \lambda & 0 \\ \lambda & \lambda + 2\mu & 0 \\ 0 & 0 & \mu \end{bmatrix} \begin{bmatrix} \eta_l^2 - \frac{1}{3} \\ -(\xi_l^2 - \frac{1}{3}) \\ 0 \end{bmatrix} = 0,$$

which means that the mode of equation (7.27) 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 .

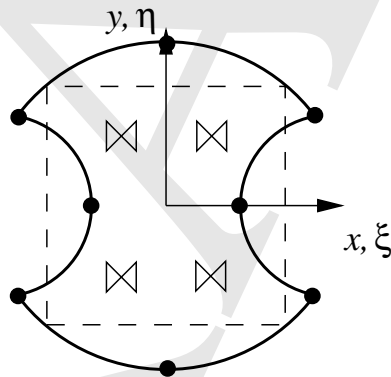


Figure 7.3: Zero-energy modes for the 8-noded quadrilateral with 2×2 Gaussian quadrature

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, considering again the 4-noded isoparametric quadrilateral

with 1×1 Gaussian quadrature, write its integrated stiffness directly as

$$\mathbf{K}^e \approx 4J(0,0)\mathbf{B}^{e,T}(0,0)\mathbf{D}\mathbf{B}^e(0,0) .$$

Since the dimension of this matrix is 8×8 , its maximum rank is 3 (why?), and two-dimensional motions include 3 rigid-body modes, it follows that this matrix has at least 2 zero-energy modes. Likewise, for the 8-noded isoparametric quadrilateral with 2×2 Gaussian quadrature, the stiffness matrix is given by

$$\mathbf{K}^e \approx \sum_{l=1}^2 \sum_{m=1}^2 w(\xi_l)w(\eta_m)J(\xi_l, \eta_m)\mathbf{B}^{e,T}(\xi_l, \eta_m)\mathbf{D}\mathbf{B}^e(\xi_l, \eta_m) .$$

Since the maximum rank of this 16×16 matrix is $4 \times 3 = 12$ and there exist exactly 3 rigid-body modes, it follows that there is at least one zero-energy mode.

Zero-energy modes are often suppressed by the actual deformation of the body, i.e., 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 at the outset.

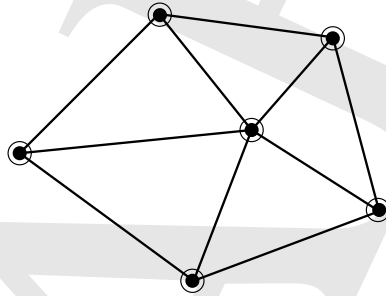
7.2.3 The patch test

A simple test for completeness of finite element approximations was proposed in 1965 by B. Irons. The idea is that any patch of elements that is used to generate piecewise polynomial approximations of the dependent variable should be unconditionally capable of reproducing a constant field of the p -th derivative of this variable when subject to appropriate boundary conditions. Irons argued somewhat heuristically that the above requirement is necessary to guarantee that the error in approximating the p -th derivative (which is the highest derivative that appears in the given weak form, with $p = 1$ for linear elastostatics) is at most of order $o(h)$. Under mesh refinement (i.e., as h approaches zero), this produces a sequence of solution that converges 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.

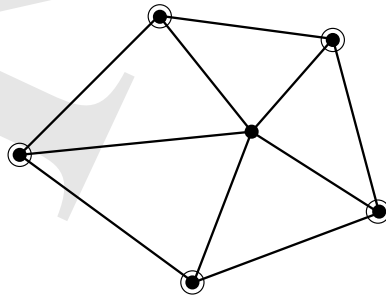
By way of background, consider the differential equation described operationally in (5.10). Three separate forms of the patch test that feature an increasing degree of severity are identified as follows:

Form A (full nodal restraint)

The values of all dependent variables in the finite element approximation are prescribed at every element according to a specified global polynomial field u_h of degree less than or equal to p which satisfies (5.10), see Figure 7.4. Since all degrees of freedom are prescribed, the solution to this problem merely consists of evaluating the “forces” corresponding to u_h . The test is designed to provide a comparison of $A[u_h]$ (which is directly available, since u_h is given) with $A_h[u_h]$, which is the finite element counterpart of A . Operators A and A_h should be identical, when applied to the given polynomial u_h of degree less or equal to p , as seen from (5.8) and (5.9).

Figure 7.4: *Schematic of the patch test (form A)***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 (5.10), 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, and is subjected to natural 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.

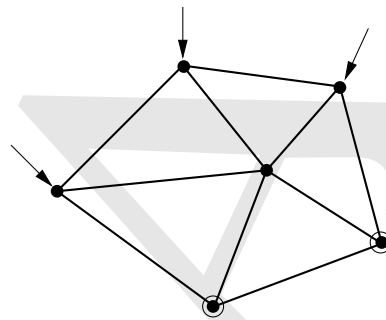


Figure 7.6: *Schematic of the patch test (form C)*

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. The degree of severity of the patch test increases from Form A to 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

$$B(w, u) = (w, f), \quad (7.29)$$

for all $w \in \mathcal{W}$, where $B(\cdot, \cdot)$ is a symmetric bilinear form and (w, 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

$$B(u, u) \geq \alpha \|u\|^2, \quad (7.30)$$

where $\|\cdot\|$ is the norm associated with the inner product of \mathcal{U} . In a discrete setting, the above condition translates to

$$\mathbf{u}^T \mathbf{K} \mathbf{u} \geq \alpha \mathbf{u}^T \mathbf{u},$$

where \mathbf{u} is a 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} .

Examples:

- (a) Consider 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 $\Gamma_u \neq \emptyset$. Here, *V-ellipticity* of B translates to the existence of a positive α , such that

$$\int_{\Omega} \left(\frac{\partial u}{\partial x_1} k \frac{\partial u}{\partial x_1} + \frac{\partial u}{\partial x_2} k \frac{\partial u}{\partial x_2} \right) d\Omega \geq \alpha \int_{\Omega} \left[u^2 + \left(\frac{\partial u}{\partial x_1} \right)^2 + \left(\frac{\partial u}{\partial x_2} \right)^2 \right] d\Omega. \quad (7.31)$$

The preceding result can be proved by appealing to the celebrated *Poincaré inequality*, according to which there exists a constant $c > 0$, such that

$$\int_{\Omega} u^2 d\Omega \leq c \int_{\Omega} \left[\left(\frac{\partial u}{\partial x_1} \right)^2 + \left(\frac{\partial u}{\partial x_2} \right)^2 \right] d\Omega, \quad (7.32)$$

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.

Taking into account (7.32), and assuming without loss of generality that $k > 0$ is constant, it follows that

$$\frac{c+1}{k} \int_{\Omega} \left[k \left(\frac{\partial u}{\partial x} \right)^2 + k \left(\frac{\partial u}{\partial y} \right)^2 \right] d\Omega \geq \int_{\Omega} \left[u^2 + \left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 \right] d\Omega,$$

$$\text{hence } \alpha = \frac{k}{c+1}.$$

- (b) Consider the problem of linear elastostatics in a domain Ω , where now the space of admissible displacements is defined as $\mathcal{U} = \{\mathbf{u} \in H^1(\Omega) \mid \mathbf{u} = \mathbf{0} \text{ on } \Gamma_u\}$ and, again,

$\Gamma_u \neq \emptyset$. Here, V -ellipticity can be proved by means of *Korn's inequality*, which states that there exists a constant $c > 0$, such that

$$\int_{\Omega} \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} \, d\Omega \geq c \|\mathbf{u}\|_{H^1(\Omega)}^2 ,$$

assuming that $\lambda > 0$ and $\mu > 0$.

When $B(\cdot, \cdot)$ is V -elliptic, it is easy to see that it induces an inner product. Indeed, $B(\cdot, \cdot)$ is bilinear, symmetric, and

$$B(u, u) \geq \alpha \|u\|_{H^1(\Omega)}^2 \geq 0$$

and $B(u, u) = 0 \Leftrightarrow \|u\|_{H^1(\Omega)} = 0 \Leftrightarrow u = 0$. The natural norm associated with this inner product is defined by

$$\|u\|_E := [B(u, u)]^{1/2} ,$$

for any $u \in \mathcal{U} = \{u \in H^1(\Omega) \mid u = 0 \text{ on } \Gamma_u\}$. This is referred to as the *energy norm*, due to its physical interpretation of being equal to twice the strain energy in the case of linear elastostatics, where

$$B(\mathbf{u}, \mathbf{u}) := \|\mathbf{u}\|_E^2 = \int_{\Omega} \boldsymbol{\epsilon}^T(\mathbf{u}) \mathbf{D} \boldsymbol{\epsilon}(\mathbf{u}) \, d\Omega .$$

Returning to (7.29), write its discrete counterpart as: find $u_h \in \mathcal{U}_h \subset \mathcal{U}$, such that

$$B(w_h, u_h) = (w_h, f) , \tag{7.33}$$

for all $w \in \mathcal{W}_h \subset \mathcal{W}$. Since $\mathcal{W}_h \subset \mathcal{W}$, one may also write

$$B(w_h, u) = (w_h, f) \tag{7.34}$$

for all $w_h \in \mathcal{W}_h$. Subtracting (7.33) from (7.34), it follows that

$$B(w_h, u - u_h) = 0 , \tag{7.35}$$

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 $w_h \in \mathcal{W}_h$.

Now consider any function $\tilde{u} \in \mathcal{U}_h$, and minimize the energy norm of the error $u - \tilde{u}$ over all $\tilde{u} \in \mathcal{U}_h$. This implies that

$$\left[\frac{d}{d\omega} B(u - \tilde{u} + \omega w_h, u - \tilde{u} + \omega w_h) \right]_{\omega=0} = 0$$

or

$$B(w_h, u - \tilde{u}) = 0 . \tag{7.36}$$

Comparing (7.36) to (7.35), it is seen that $\tilde{u} = u_h$. In conclusion the finite element solution u_h minimizes the error in the energy norm and, in this sense, it constitutes the *best approximation* to the exact solution u , see Figure 7.7 for a geometric interpretation. Note that the existence and uniqueness of the solution to either (7.29) or (7.33) is guaranteed by the *Lax-Milgram theorem*, assuming a continuous and V -elliptic bilinear form $B(\cdot, \cdot)$ and a continuous linear form (\cdot, f) on Hilbert spaces \mathcal{U} (or \mathcal{U}_h).

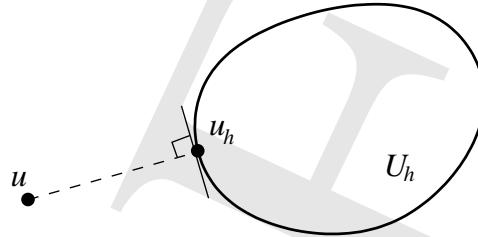


Figure 7.7: *Geometric interpretation of the best approximation property as a closest-point projection from u to \mathcal{U} in the sense of the energy norm*

An important corollary of the orthogonality condition (7.35) is noted here: let u be the solution to an elliptic problem of the type (7.29) and u_h be the finite element approximation to this solution. It follows that

$$B(u, u) = B(u_h + e, u_h + e) = B(u_h, u_h) + B(e, e) + 2B(u_h, e), \quad (7.37)$$

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 (7.35) implies that $B(u_h, e) = 0$, hence

$$B(u, u) = B(u_h, u_h) + B(e, e) \geq B(u_h, u_h) \quad (7.38)$$

since $B(e, e) \geq 0$. The inequality (7.38) shows that the energy is underestimated by the finite element method. 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

$$\Omega_h \approx \Omega \quad , \quad \partial\Omega_h \approx \partial\Omega .$$

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.,

$$\int_{\square} f(\xi, \eta, \zeta) d\xi d\eta d\zeta \approx \sum_{k=1}^L \sum_{l=1}^L \sum_{m=1}^L w_k w_l w_m f(\xi_k, \eta_l, \zeta_m),$$

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, see 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

$$\kappa = \|\mathbf{K}\| \|\mathbf{K}^{-1}\|, \quad (7.39)$$

where $\|\cdot\|$ is any matrix norm. If the matrix norm is taken to be the spectral norm, defined as

$$\|\mathbf{K}\| := \max\{\sqrt{\rho} \mid \rho : \text{eigenvalue of } \mathbf{K}^T \mathbf{K}\},$$

then the condition number of equation (7.39) takes the particular form

$$\kappa = \left| \frac{\rho_{max}}{\rho_{min}} \right|,$$

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 can be obtained by starting with the orthogonality condition (7.35) and writing

$$\begin{aligned} B(u - u_h, u - u_h) &= B(u - u_h, u - u_h) + B(u - u_h, w_h) \\ &= B(u - u_h, u - u_h + w_h) \\ &= B(u - u_h, u - v) , \end{aligned} \tag{7.40}$$

where v is an arbitrary element of \mathcal{U}_h written as $v = u_h - w_h$. Recalling that B is continuous in both of its arguments, it follows that there is a constant $M > 0$, such that

$$B(u - u_h, u - v) \leq M \|u - u_h\| \|u - v\| , \tag{7.41}$$

for all $u \in \mathcal{U}$, and $u_h, v \in \mathcal{U}_h$. Furthermore, taking into account (7.40) and (7.41), the V -ellipticity condition (7.30) leads to

$$M \|u - u_h\| \|u - v\| \geq \alpha \|u - u_h\|^2 ,$$

which, in turn, implies that

$$\|u - u_h\| \leq \frac{M}{\alpha} \|u - v\| , \tag{7.42}$$

for all $v \in \mathcal{U}_h$.

The error estimate (7.42) bounds the error from above by the difference between the exact solution u and any other element v of \mathcal{U}_h . Although interesting in its own right, this result is of limited practical significance, as it involves the (unknown) exact solution on both sides of the inequality.

Another type of error estimate applicable to the case of h -adaptivity can be deduced for elliptic problems in the form

$$\|u - u_h\|_E \leq C_1 h^{q-p+1} |u|_{q+1} , \tag{7.43}$$

where

$$|u|_{q+1}^2 := \int_{\Omega} \sum_{\alpha_1 + \alpha_2 + \dots + \alpha_n = q+1} \left| \frac{d^{q+1}u}{dx_1^{\alpha_1} dx_2^{\alpha_2} \dots dx_n^{\alpha_n}} \right|^2 d\Omega .$$

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 a positive constant that is independent of h . Also, the term $|u|_{q+1}$ is a measure of smoothness of the exact solution.

In the case of p -adaptivity, a typical error estimate is of the form

$$\|u - u_h\|_E \leq C_2 q^{-(p-1)} |u|_p , \quad (7.44)$$

where $p > 1$ and C_2 is a constant independent of q .

The error estimates (7.43) and (7.44) are of practical use because they establish the rate of convergence of the finite element approximation under mesh refinement (i.e., when $h \mapsto 0$), or under increase of the degree of polynomial completeness (i.e., when $q \mapsto \infty$), respectively. Although, again, it contains on the right-hand side a term that depends on the exact solution, this does not limit its 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

$$\boldsymbol{\epsilon} = \mathbf{e} + \frac{1}{3}(\text{tr } \boldsymbol{\epsilon})\mathbf{I} \quad , \quad \boldsymbol{\sigma} = \mathbf{s} + \frac{1}{3}(\text{tr } \boldsymbol{\sigma})\mathbf{I} , \quad (7.44)$$

where \mathbf{e} and \mathbf{s} are the *deviatoric* strain and stress, respectively. It follows from (7.44) that the deviatoric strain and stress are traceless, namely that $\text{tr } \mathbf{e} = 0$ and $\text{tr } \mathbf{s} = 0$. Also, the *volumetric* strain θ and the pressure p are defined as

$$\theta = \text{tr } \boldsymbol{\epsilon} = \nabla \cdot \mathbf{u} \quad , \quad p = \frac{1}{3} \text{tr } \boldsymbol{\sigma} . \quad (7.45)$$

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 (7.3) that

$$dv = (1 + \epsilon_{11})(1 + \epsilon_{22})(1 + \epsilon_{33})dV$$

or, upon ignoring higher-order terms,

$$\frac{dv}{dV} \approx 1 + \epsilon_{11} + \epsilon_{22} + \epsilon_{33} = 1 + \theta .$$

Taking into account (7.44) and (7.45), the isotropic stress-strain relation (7.2) can be rewritten as

$$\mathbf{s} = 2\mu\boldsymbol{\epsilon} \quad , \quad p = \left(\lambda + \frac{2}{3}\mu\right)\theta := K\theta , \quad (7.46)$$

where K is the *bulk* modulus. As seen from (7.46), the original stress-strain relation (7.2) 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 $K = \frac{E}{3(1-2\nu)}$, 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 (7.46)₂ applies and the pressure is computed from it. In the latter case, (7.46)₂ ceases to apply and the pressure becomes indeterminate from it as $K \rightarrow \infty$ and $\theta \rightarrow 0$. In this case, 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

$$\begin{aligned} \nabla \cdot \boldsymbol{\sigma} + \mathbf{f} &= \mathbf{0} && \text{in } \Omega , \\ \boldsymbol{\sigma} \mathbf{n} &= \bar{\mathbf{t}} && \text{on } \Gamma_q , \\ \mathbf{u} &= \bar{\mathbf{u}} && \text{on } \Gamma_u , \\ \nabla \cdot \mathbf{u} &= 0 && \text{in } \Omega , \end{aligned} \quad (7.47)$$

where now the stress tensor $\boldsymbol{\sigma}$ is given by

$$\boldsymbol{\sigma} = p\mathbf{I} + 2\mu\boldsymbol{\epsilon} = p\mathbf{I} + 2\mu\boldsymbol{\epsilon} . \quad (7.48)$$

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 (7.1), where the only unknown is the displacement \mathbf{u} , while the pressure p is determined by the constitutive equation (7.46)₂.

The strong form (7.47) 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

$$\int_{\Omega} \mathbf{w} \cdot (-\nabla \cdot \boldsymbol{\sigma} - \mathbf{f}) d\Omega + \int_{\Gamma_q} \mathbf{w} \cdot (\boldsymbol{\sigma} \mathbf{n} - \bar{\mathbf{t}}) d\Gamma + \int_{\Omega} q \nabla \cdot \mathbf{u} d\Omega = 0, \quad (7.49)$$

where the weighting functions (\mathbf{w}, q) belong to $\mathcal{W} \times \mathcal{Q}$. Upon following the standard process of employing integration by parts and the divergence theorem as in Section 7.1, equation (7.49) transforms into

$$\int_{\Omega} \nabla_s \mathbf{w} : \boldsymbol{\sigma} d\Omega + \int_{\Omega} q \nabla \cdot \mathbf{u} d\Omega = \int_{\Omega} \mathbf{w} \cdot \mathbf{f} d\Omega + \int_{\Gamma_q} \mathbf{w} \cdot \bar{\mathbf{t}} d\Gamma. \quad (7.50)$$

Recalling (7.44), (7.45), (7.46) and (7.47)₄, one may write

$$\nabla_s \mathbf{w} : \boldsymbol{\sigma} = \boldsymbol{\epsilon}(\mathbf{w}) : (\mathbf{s} + p\mathbf{I}) = \boldsymbol{\epsilon}(\mathbf{w}) : 2\mu\mathbf{e} + p\nabla \cdot \mathbf{w} = \boldsymbol{\epsilon}(\mathbf{w}) : 2\mu\boldsymbol{\epsilon}(\mathbf{u}) + p\nabla \cdot \mathbf{w}. \quad (7.51)$$

Hence, the weak form (7.50) can be expressed equivalently as

$$\int_{\Omega} \nabla_s \mathbf{w} : 2\mu\nabla_s \mathbf{u} d\Omega + \int_{\Omega} p\nabla \cdot \mathbf{w} d\Omega + \int_{\Omega} q \nabla \cdot \mathbf{u} d\Omega = \int_{\Omega} \mathbf{w} \cdot \mathbf{f} d\Omega + \int_{\Gamma_q} \mathbf{w} \cdot \bar{\mathbf{t}} d\Gamma, \quad (7.52)$$

or, resorting to the vector representation,

$$\int_{\Omega} \boldsymbol{\epsilon}^T(\mathbf{w}) 2\mu\boldsymbol{\epsilon}(\mathbf{u}) d\Omega + \int_{\Omega} p\nabla \cdot \mathbf{w} d\Omega + \int_{\Omega} q \nabla \cdot \mathbf{u} d\Omega = \int_{\Omega} \mathbf{w}^T \mathbf{f} d\Omega + \int_{\Gamma_q} \mathbf{w}^T \bar{\mathbf{t}} d\Gamma. \quad (7.53)$$

The weighted-residual problem amounts to finding $(\mathbf{u}, p) \in \mathcal{U} \times \mathcal{P}$, such that (7.53) hold for all $(\mathbf{w}, q) \in \mathcal{W} \times \mathcal{Q}$. The spaces of admissible displacements and associated weighting functions are

$$\begin{aligned} \mathcal{U} &= \left\{ \mathbf{u} \in H^1(\Omega) \mid \mathbf{u} = \bar{\mathbf{u}} \text{ on } \Gamma_u \right\}, \\ \mathcal{W} &= \left\{ \mathbf{w} \in H^1(\Omega) \mid \mathbf{w} = \mathbf{0} \text{ on } \Gamma_u \right\}, \end{aligned}$$

whereas the spaces of admissible pressures and associated weighting functions are

$$\mathcal{P} = \mathcal{Q} = \left\{ p \in H^0(\Omega) \right\}.$$

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 \bar{p} is a constant over the domain Ω , then

$$\int_{\Omega} (p+\bar{p}) \nabla \cdot \mathbf{w} \, d\Omega = \int_{\partial\Omega} (p+\bar{p}) \mathbf{w} \cdot \mathbf{n} \, d\Gamma - \int_{\Omega} \nabla(p+\bar{p}) \cdot \mathbf{w} \, d\Omega = - \int_{\Omega} \nabla p \cdot \mathbf{w} \, d\Omega = \int_{\Omega} p \nabla \cdot \mathbf{w} \, d\Omega .$$

To remove this indeterminacy, one may define \mathcal{P} in such cases as

$$\mathcal{P} = \mathcal{Q} = \left\{ p \in H^0(\Omega) \mid \int_{\Omega} p \, d\Omega = 0 \right\} .$$

The preceding problem can be written in operational form as

$$\begin{aligned} B(\mathbf{w}, \mathbf{u}) + C(\mathbf{w}, p) &= (\mathbf{w}, \mathbf{f}) \\ C(\mathbf{u}, q) &= 0 , \end{aligned} \tag{7.54}$$

where

$$B(\mathbf{w}, \mathbf{u}) := 2\mu \int_{\Omega} \epsilon^T(\mathbf{w}) \epsilon(\mathbf{u}) \, d\Omega$$

and

$$C(\mathbf{w}, p) := \int_{\Omega} p \nabla \cdot \mathbf{w} \, d\Omega .$$

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 directly attempt to incorporate the constraint (7.47)₄ directly into the space of admissible displacements \mathcal{U}_c , namely define

$$\mathcal{U}_c = \left\{ \mathbf{u} \in H^1(\Omega) \mid \mathbf{u} = \bar{\mathbf{u}} \text{ on } \Gamma_u \text{ , } \nabla \cdot \mathbf{u} = \mathbf{0} \text{ in } \Omega \right\} .$$

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 (7.53) leads to *dual* methods, which are generally simpler to implement.

The weak form (7.54) constitutes the basis for a finite element approximation of the constrained problem. Indeed, such an approximation amounts to defining discrete admissible fields $\mathcal{U}_h \in \mathcal{U}$, $\mathcal{W}_h \in \mathcal{W}$ and $\mathcal{P}_h \in \mathcal{P}$ and seeking a solution to (7.54) within these fields. The discrete problem leads to a global system of algebraic equations of the form

$$\begin{bmatrix} \mathbf{K}_{uu} & \mathbf{K}_{up} \\ \mathbf{K}_{pu}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}} \\ \hat{\mathbf{p}} \end{bmatrix} = \begin{bmatrix} \mathbf{F} \\ \mathbf{0} \end{bmatrix} ,$$

where $\hat{\mathbf{u}}$ and $\hat{\mathbf{p}}$ are the displacement and pressure degrees of freedom. Recalling the structure of (7.54), 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-noded triangular elements with linear displacement \mathbf{u}_h and constraint pressure p_h in each element, see Figure 7.8.

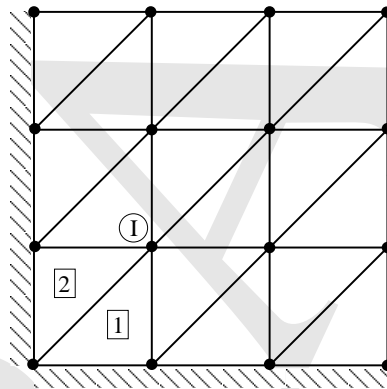


Figure 7.8: *Illustration of volumetric locking in plane strain when using 3-noded triangular elements*

The constraint of incompressibility can be expressed at the element level as

$$\int_{\Omega^e} q_h \nabla \cdot \mathbf{u}_h \, d\Omega = 0 ,$$

or, since the pressure (hence also the weighting function q_h) is assumed piecewise constant,

$$\int_{\Omega^e} \nabla \cdot \mathbf{u}_h \, d\Omega = 0 .$$

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 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 to infinity. To understand the nearly incompressible case, recall that the total potential energy $I[\mathbf{u}]$ of equation (7.11) attains an absolute minimum at the equilibrium point, and write the strain energy as

$$W[\mathbf{u}] = \frac{1}{2} \int_{\Omega} \boldsymbol{\epsilon} : \boldsymbol{\sigma} \, d\Omega = \frac{1}{2} \int_{\Omega} [2\mu \mathbf{e} : \mathbf{e} + K\theta^2] \, d\Omega .$$

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 .

DRAFT

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, i.e., spatial variables (e.g., x_i , $i = 1, 2, 3$) and temporal variables (t). In the context of the finite element method, there are two general approaches in dealing with these 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 the spatial or 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.

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

$$\frac{\partial}{\partial x_1} \left(k \frac{\partial u}{\partial x_1} \right) + \frac{\partial}{\partial x_2} \left(k \frac{\partial u}{\partial x_2} \right) - f = \rho c \frac{\partial u}{\partial t} \quad \text{in } \Omega \times I, \quad (7.54)$$

$$-k \frac{\partial u}{\partial n} = \bar{q} \quad \text{on } \Gamma_q \times I, \quad (7.54)$$

$$u = \bar{u} \quad \text{on } \Gamma_u \times I, \quad (7.54)$$

$$u(x_1, x_2, 0) = u_0(x_1, x_2) \quad \text{in } \Omega, \quad (7.54)$$

where $u = u(x_1, x_2, t)$ is the (yet unknown) solution and $I = (0, T]$, with T being a given 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 on Ω and a continuous function $f = f(x_1, x_2, t)$ is defined in $\Omega \times I$. Further, continuous 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 (8.2) and (8.3) are the *time-dependent Neumann* and *time dependent Dirichlet* conditions, respectively. Finally, equation (8.4) is the *initial condition*. 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)-(8.4).

A Galerkin-based weighted-residual form of the above problem can be deduced 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, i.e., $w_u = 0$ 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* on 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

$$\int_{\Omega \times I} w \left[-\rho c \frac{\partial u}{\partial t} + \frac{\partial}{\partial x_1} \left(k \frac{\partial u}{\partial x_1} \right) + \frac{\partial}{\partial x_2} \left(k \frac{\partial u}{\partial x_2} \right) - f \right] d(\Omega \times I) - \int_{\Gamma_q \times I} w \left[k \frac{\partial u}{\partial n} + \bar{q} \right] d(\Gamma \times I) = 0. \quad (8.5)$$

Clearly, equation (8.5) involves a space-time integral. Since the spatial and temporal dimensions are independent of each other, they may be readily decoupled, so that (8.5) can be

rewritten as

$$\int_I \int_{\Omega} w \left[-\rho c \frac{\partial u}{\partial t} + \frac{\partial}{\partial x_1} \left(k \frac{\partial u}{\partial x_1} \right) + \frac{\partial}{\partial x_2} \left(k \frac{\partial u}{\partial x_2} \right) - f \right] d\Omega dt - \int_I \int_{\Gamma_q} w \left[k \frac{\partial u}{\partial n} + \bar{q} \right] d\Gamma dI = 0. \quad (8.6)$$

One may taking advantage of this decoupling and “freeze” time in order to first operate on the space integrals, i.e., on the integro-differential equation

$$\int_{\Omega} w \left[-\rho c \frac{\partial u}{\partial t} + \frac{\partial}{\partial x_1} \left(k \frac{\partial u}{\partial x_1} \right) + \frac{\partial}{\partial x_2} \left(k \frac{\partial u}{\partial x_2} \right) - f \right] d\Omega - \int_{\Gamma_q} w \left[k \frac{\partial u}{\partial n} + \bar{q} \right] d\Gamma = 0, \quad (8.7)$$

which, upon using integration by parts, the divergence theorem, and assumption (iii) takes the form

$$\int_{\Omega} w \rho c \frac{\partial u}{\partial t} d\Omega + \int_{\Omega} \left[\frac{\partial w}{\partial x_1} k \frac{\partial u}{\partial x_1} + \frac{\partial w}{\partial x_2} k \frac{\partial u}{\partial x_2} + w f \right] d\Omega + \int_{\Gamma_q} w \bar{q} d\Gamma = 0. \quad (8.8)$$

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

$$\int_I \left[\int_{\Omega} w \rho c \frac{\partial u}{\partial t} d\Omega + \int_{\Omega} \left[\frac{\partial w}{\partial x_1} k \frac{\partial u}{\partial x_1} + \frac{\partial w}{\partial x_2} k \frac{\partial u}{\partial x_2} + w f \right] d\Omega + \int_{\Gamma_q} w \bar{q} d\Gamma \right] dt = 0, \quad (8.9)$$

for all $w \in \mathcal{W}$. Here, the spaces \mathcal{U} and \mathcal{W} of admissible solution and weighting functions are defined as

$$\mathcal{U} := \left\{ u \in H^1(\Omega \times I) \mid u = \bar{u} \text{ on } \Gamma_u \times I, \quad u(x_1, x_2, 0) = u_0 \right\},$$

and

$$\mathcal{W} := \left\{ w \in H^1(\Omega \times I) \mid w = 0 \text{ on } \Gamma_u \times I, \quad w(x_1, x_2, 0) = 0 \right\}.$$

A Bubnov-Galerkin approximation of the weak form (8.9) can be effected by writing

$$u \approx u_h = \sum_{I=1}^N \varphi_I(x_1, x_2) u_I(t) + u_b(x_1, x_2, t), \quad (7.54)$$

$$w \approx w_h = \sum_{I=1}^N \varphi_I(x_1, x_2) w_I(t), \quad (7.54)$$

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 Neumann condition (8.3) and the initial condition (8.4). It is

clear from (8.10) and (8.11) that the approximation induces a separation of spatial and temporal variables, which plays an essential role in the ensuing developments.

Substituting of u_h and w_h into the weak form (8.9) leads to

$$\begin{aligned} \int_I \left[\sum_{I=1}^N w_I \int_{\Omega} \varphi_I \rho c (\sum_{J=1}^N \varphi_J \dot{u}_J + \dot{u}_b) d\Omega \right. \\ \left. + \sum_{I=1}^N w_I \int_{\Omega} \{\varphi_{I,1} \varphi_{I,2}\} k \left(\sum_{J=1}^N \begin{Bmatrix} \varphi_{J,1} \\ \varphi_{J,2} \end{Bmatrix} u_J + \begin{Bmatrix} u_{b,1} \\ u_{b,2} \end{Bmatrix} \right) d\Omega \right. \\ \left. + \sum_{I=1}^N w_I \int_{\Omega} \varphi_I f d\Omega + \sum_{I=1}^N w_I \int_{\Gamma_q} \varphi_I \bar{q} d\Gamma \right] dt = 0. \quad (7.54) \end{aligned}$$

This equation may be rewritten as

$$\int_I \left[\sum_{I=1}^N w_I \left\{ \sum_{J=1}^N (M_{IJ} \dot{u}_J + K_{IJ} u_J) - F_I \right\} \right] = 0, \quad (8.13)$$

where

$$\begin{aligned} M_{IJ} &:= \int_{\Omega} \varphi_I \rho c \varphi_J d\Omega, \\ K_{IJ} &:= \int_{\Omega} \{\varphi_{I,1} \varphi_{I,2}\} k \begin{Bmatrix} \varphi_{J,1} \\ \varphi_{J,2} \end{Bmatrix} d\Omega, \end{aligned}$$

and

$$F_I := - \int_{\Omega} \varphi_I \rho c \dot{u}_b d\Omega - \int_{\Omega} \varphi_I f d\Omega - \int_{\Omega} \{\varphi_{I,1} \varphi_{I,2}\} k \begin{Bmatrix} u_{b,1} \\ u_{b,2} \end{Bmatrix} d\Omega - \int_{\Gamma_q} \varphi_I \bar{q} d\Gamma.$$

The arrays $[M_{IJ}]$, $[K_{IJ}]$ and $[F_I]$ are termed the *mass* (or *capacitance*) matrix, the *stiffness* matrix and the *forcing* vector, respectively. Clearly, both $[M_{IJ}]$ and $[K_{IJ}]$ are symmetric, while it is easy to establish that $[M_{IJ}]$ is also positive-definite and $[K_{IJ}]$ is positive-semidefinite (as in the steady problem).

In conclusion, one has arrived at the semi-discrete form (8.13), which may be also written as

$$\int_I \mathbf{w}^T (\mathbf{M} \dot{\mathbf{u}} + \mathbf{K} \mathbf{u} - \mathbf{F}) dt = 0, \quad (8.14)$$

where $\mathbf{u} := [u_1(t) \ u_2(t) \ \dots \ u_N(t)]^T$ and $\mathbf{w} := [w_1(t) \ w_2(t) \ \dots \ w_N(t)]^T$. Equation (8.14) is now an integro-differential equation in time only, as all the spatial derivatives and integrals have been evaluated and “stored” in the arrays $[M_{IJ}]$, $[K_{IJ}]$ and $[F_I]$.

Once the spatial problem has been discretized, one may proceed to the temporal problem. Here, there are two distinct options:

- (a) Discretize \mathbf{u} and \mathbf{w} in time according to some polynomial series, i.e.,

$$\mathbf{u} \approx \hat{\mathbf{u}} = \sum_{n=1}^M \boldsymbol{\alpha}_n t^n, \quad \mathbf{w} \approx \hat{\mathbf{w}} = \sum_{n=1}^M \boldsymbol{\beta}_n t^n,$$

where $\boldsymbol{\alpha}_n$ is a vector to be determined and $\boldsymbol{\beta}_n$ is an arbitrary vector. These approximate functions are then substituted into the semi-discrete form (8.14) and the resulting system is solved for the values of α_n . This is essentially a Bubnov-Galerkin approximation in time.

- (b) Apply a standard discrete time integrator directly on the semi-discrete form (8.14). This amounts to choosing \mathbf{w} to consist of Dirac-delta functions at discrete times $t_1, t_2, \dots, t_n, t_{n+1}, \dots$, which would imply that the system of ordinary differential equations

$$\mathbf{M}\dot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{F} \quad (8.15)$$

is to be exactly satisfied at these times.

In what follows, the second option is pursued. To this end, recall that the general solution of the homogeneous counterpart of (8.15), i.e., when $\mathbf{F} = \mathbf{0}$, is of the form

$$\mathbf{u}(t) = \sum_{I=1}^N c_I e^{-\lambda_I t} \mathbf{z}_I,$$

where the pairs $(\lambda_I, \mathbf{z}_I)$, $I = 1, 2, \dots, N$ are to be determined. Upon substituting a typical such pair (λ, \mathbf{z}) into the homogeneous equation, one gets

$$e^{-\lambda t} (-\lambda \mathbf{M} + \mathbf{K}) \mathbf{z} = \mathbf{0},$$

hence,

$$\lambda \mathbf{M} \mathbf{z} = \mathbf{K} \mathbf{z}. \quad (8.16)$$

Equation (8.16) corresponds to the general symmetric linear eigenvalue problem, which can be solved for the eigenpairs $(\lambda_I, \mathbf{z}_I)$, $I = 1, 2, \dots, N$. For notational simplicity, define the $(N \times N)$ arrays

$$\boldsymbol{\Lambda} := \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_N \end{bmatrix}$$

and

$$\mathbf{Z} := [\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_N],$$

so that the eigenvalue problem (8.16) may be conveniently rewritten as

$$\mathbf{M}\mathbf{Z}\mathbf{\Lambda} = \mathbf{K}\mathbf{Z}.$$

Given that \mathbf{M} and \mathbf{K} are symmetric, standard orthogonality properties of the eigenpairs $(\lambda_I, \mathbf{z}_I)$, $I = 1, 2, \dots, N$, require that

$$\mathbf{Z}^T \mathbf{M} \mathbf{Z} = \begin{bmatrix} m_1 & & & \\ & m_2 & & \\ & & \ddots & \\ & & & m_N \end{bmatrix}$$

and

$$\mathbf{Z}^T \mathbf{K} \mathbf{Z} = \begin{bmatrix} k_1 & & & \\ & k_2 & & \\ & & \ddots & \\ & & & k_N \end{bmatrix},$$

where $\lambda_I = \frac{k_I}{m_I}$, and $m_I > 0$, $k_I \geq 0$, thus $\lambda_I \geq 0$.

The solution of the non-homogeneous equations (8.15) is attained by employing the classical technique of variation of parameters, according to which it is assumed that

$$\mathbf{u}(t) = \mathbf{Z}\mathbf{v}(t), \quad (8.17)$$

where \mathbf{Z} is obtained from the homogeneous problem. Substituting (8.17) into (8.15) gives rise to

$$\mathbf{M}\mathbf{Z}\dot{\mathbf{v}} + \mathbf{K}\mathbf{Z}\mathbf{v} = \mathbf{F}.$$

Premultiplying this equation by \mathbf{Z}^T leads to

$$\mathbf{Z}^T \mathbf{M} \mathbf{Z} \dot{\mathbf{v}} + \mathbf{Z}^T \mathbf{K} \mathbf{Z} \mathbf{v} = \mathbf{Z}^T \mathbf{F},$$

or, taking into account the earlier orthogonality conditions,

$$\begin{bmatrix} m_1 & & & \\ & m_2 & & \\ & & \ddots & \\ & & & m_N \end{bmatrix} \begin{bmatrix} \dot{v}_1 \\ \dot{v}_2 \\ \vdots \\ \dot{v}_N \end{bmatrix} + \begin{bmatrix} k_1 & & & \\ & k_2 & & \\ & & \ddots & \\ & & & k_N \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_N \end{bmatrix},$$

where $g_i := \mathbf{z}_i^T \mathbf{F}$. It is now concluded that the original system of coupled linear ordinary differential equations (8.15) has been reduced to a set of N uncoupled scalar ordinary differential equations of the form

$$m_I \dot{v}_I + k_I v_I = g_I ,$$

where $I = 1, 2, \dots, N$. Therefore, in order to understand the behavior of the original system, one only needs to study the solution of a single scalar ordinary differential equation of the form

$$m\dot{v} + kv = g , \quad (8.18)$$

with initial condition $v(0) = v_0$.

The general solution of equation (8.18) can be obtained using the method of variation of parameters, and is given by

$$v(t) = e^{-\lambda t} y(t) , \quad (8.19)$$

where $\lambda = \frac{k}{m}$ and $y = y(t)$ is a function to be determined. Upon substituting the general solution into (8.18) and simplifying, the resulting expression leads to

$$\dot{y}(t) = \frac{1}{m} e^{\lambda t} g .$$

This equation can be integrated in the time interval $I_{n+1} := (t_n, t_{n+1}]$ and results in

$$y(t) = y_n + \int_{t_n}^t \frac{1}{m} e^{\lambda \tau} g(\tau) d\tau ,$$

where $y_n := y(t_n)$. Hence, one obtains from (8.19) the solution for $v(t)$ in convolution form as

$$v(t) = e^{-\lambda t} y_n + \int_{t_n}^t \frac{1}{m} e^{\lambda(\tau-t)} g(\tau) d\tau .$$

Noting, further, that $v(t_n) := v_n = e^{-\lambda t_n} y_n$, it follows that the preceding solution can be also expressed as

$$v(t) = e^{-\lambda(t-t_n)} v_n + \int_{t_n}^t \frac{1}{m} e^{-\lambda(t-\tau)} g(\tau) d\tau . \quad (8.20)$$

Setting $t = t_{n+1}$, it is readily seen from (8.20) that

$$v_{n+1} = e^{-\lambda \Delta t_n} v_n + \int_{t_n}^{t_{n+1}} \frac{1}{m} e^{-\lambda(t_{n+1}-\tau)} g(\tau) d\tau ,$$

where $\Delta t_n := t_{n+1} - t_n$. The ratio $\frac{v_{n+1}}{v_n} = r$ is termed *the amplification factor*. In the homogeneous case ($g = 0$), equation (8.20) immediately implies that $r = e^{-\lambda \Delta t_n}$, i.e., the exact solution experiences exponential decay. This, in turn, implies that $r \rightarrow 1$ when $\lambda \Delta t_n \rightarrow 0$ and $r \rightarrow 0$ when $\lambda \Delta t_n \rightarrow \infty$.

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 (8.18), which, as argued in the preceding section, fully represents the general system (8.15) obtained through the semi-discretization of the weak form (8.5).

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+1} by using a Taylor series expansion of $v(t)$ at t_n as

$$v(t_{n+1}) = v(t_n) + \Delta t_n \dot{v}(t_n) + o(\Delta t_n^2),$$

which, upon ignoring the second-order terms in $\Delta t_n := t_{n+1} - t_n$, leads to

$$\dot{v}(t_n) \approx \frac{v_{n+1} - v_n}{\Delta t_n}. \quad (8.21)$$

Upon writing (8.15) at t_n with $\dot{v}(t_n)$ computed from (8.21), it is concluded that

$$m \frac{v_{n+1} - v_n}{\Delta t_n} + kv_n = g_n,$$

where $v_k := v(t_k)$. This equation may be trivially rewritten as

$$v_{n+1} = (1 - \lambda \Delta t_n) v_n + \frac{\Delta t_n}{m} g_n, \quad (8.22)$$

where, again $\lambda := \frac{k}{m}$. In the homogeneous case ($g = 0$), it is seen from (8.22) that the discrete amplification ratio r_f of the forward Euler method is given by

$$r_f := 1 - \lambda \Delta t_n. \quad (8.23)$$

Equation (8.23) implies that for finite values of λ , the limiting case $\Delta t_n \rightarrow 0$ leads to $r_f \rightarrow 1$, which is consistent with the exact solution, as argued earlier in this section. However, the limiting case $\Delta t_n \rightarrow \infty$ leads to $r_f \rightarrow -\infty$, which reveals that the discrete solution does not predict exponential decay in the limit of an infinitely large time step Δt_n . As can be easily inferred from (8.22), the forward Euler method is only conditionally stable. Indeed, ignoring the inhomogeneous term, 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*, i.e., 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 $\Delta t_n < \frac{2}{\lambda} =: \Delta t_{cr}$. In systems with many degrees of freedom, such as (8.15), the *critical step-size* Δt_{cr} is defined as

$$\Delta t_{cr} := \frac{2}{\lambda_{max}},$$

where λ_{max} is the maximum eigenvalue of problem (8.16). This implies that in order to guarantee stability for the forward Euler method, one needs to know (or estimate) the maximum eigenvalue of (8.16). 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, recall that, since the interpolation functions φ_I are dimensionless, the components of the stiffness matrix in the two-dimensional transient heat conduction problem are of order $o(1)$, while the components $[M_{IJ}]$ of the mass matrix 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, the critical step-size must be reduced quadratically under mesh refinement, i.e., halving the mesh-size necessitates reduction of the step-size by a factor of four.

An alternative discrete time integrator is the *backward Euler method*, which can be deduced by writing $v(t_n)$ using a Taylor series expansion at t_{n+1} as

$$v(t_n) = v(t_{n+1}) - \Delta t_n \dot{v}(t_{n+1}) + o(\Delta t_n^2),$$

which, upon ignoring the second-order terms in Δt_n leads to

$$\dot{v}(t_{n+1}) \approx \frac{v_{n+1} - v_n}{\Delta t_n}. \quad (8.24)$$

Writing now (8.15) at t_{n+1} , with $\dot{v}(t_{n+1})$ estimated from (8.24), results in

$$m \frac{v_{n+1} - v_n}{\Delta t_n} + kv_{n+1} = g_{n+1}$$

or, upon solving for v_{n+1} ,

$$v_{n+1} = \frac{1}{1 + \lambda \Delta t_n} v_n + \frac{\Delta t_n}{1 + \lambda \Delta t_n} \frac{1}{m} g_{n+1}. \quad (8.25)$$

For the homogeneous problem, equation (8.25) implies that in the limiting cases $\Delta t_n \rightarrow 0$ and $\Delta t_n \rightarrow \infty$, the discrete amplification ratio r_b , defined as

$$r_b := \frac{1}{1 + \lambda \Delta t_n}, \quad (8.26)$$

satisfies $A \rightarrow 1$ and $A \rightarrow 0$, respectively. This means that the backward Euler method is consistent with the exact solution in both extreme cases. In addition, as seen from (8.26), 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 8.1 shows the amplification factor for the two methods, as well as for the exact solution.

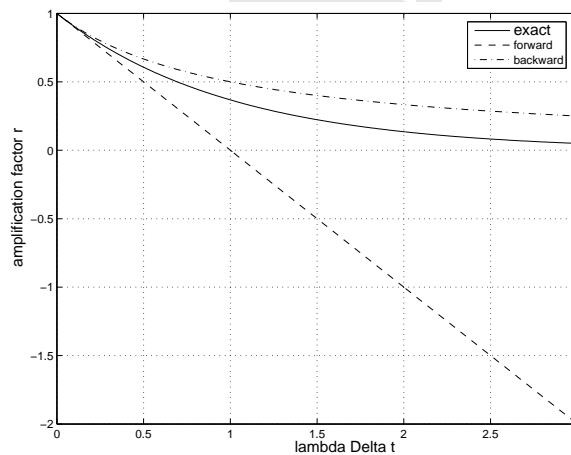


Figure 8.1: Amplification factor r as a function of $\lambda \Delta t$ for forward Euler, backward Euler and the exact solution of the homogeneous counterpart of (8.18)

Returning to the system of ordinary differential equations in (8.15), one may use forward Euler integration, which leads to

$$\mathbf{M} \frac{\mathbf{u}_{n+1} - \mathbf{u}_n}{\Delta t_n} + \mathbf{K} \mathbf{u}_n = \mathbf{F}_n, \quad (8.27)$$

hence

$$\mathbf{M} \mathbf{u}_{n+1} = \mathbf{M} \mathbf{u}_n - \Delta t_n \mathbf{K} \mathbf{u}_n + \Delta t_n \mathbf{F}_n. \quad (8.28)$$

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 $n = 1, 2, \dots$. 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 estimate of the mass matrix \mathbf{M} can be easily computed using nodal quadrature, i.e., 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 and the properties of the element interpolation functions.

The backward Euler method can also be applied to (8.15), resulting in

$$\mathbf{M} \frac{\mathbf{u}_{n+1} - \mathbf{u}_n}{\Delta t_n} + \mathbf{K} \mathbf{u}_{n+1} = \mathbf{F}_{n+1}, \quad (8.29)$$

which implies that

$$(\mathbf{M} + \Delta t_n \mathbf{K}) \mathbf{u}_{n+1} = \mathbf{M} \mathbf{u}_n + \Delta t_n \mathbf{F}_n. \quad (8.30)$$

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 it 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 interesting to rederive the discrete time integrators of the previous section using a weighted-residual formalization. To this end, start from equation (8.14) and consider the time interval $I = (t_n, t_{n+1}]$, where

$$\int_{t_n}^{t_{n+1}} \mathbf{w}^T (\mathbf{M} \dot{\mathbf{u}} + \mathbf{K} \mathbf{u} - \mathbf{F}) dt = 0. \quad (8.31)$$

Now, choose a linear polynomial interpolation of \mathbf{u} with time, i.e.,

$$\mathbf{u} \approx \hat{\mathbf{u}} = \left(1 - \frac{t - t_n}{\Delta t_n}\right) \mathbf{u}_n + \frac{t - t_n}{\Delta t_n} \mathbf{u}_{n+1}, \quad (8.32)$$

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

$$\mathbf{w} \approx \hat{\mathbf{w}} = \delta(t_n^+) \mathbf{c}, \quad (8.33)$$

in (t_n, t_{n+1}) , where \mathbf{c} is an arbitrary constant vector. Substituting (8.32) and (8.33) into (8.31), one obtains (8.27), thus recovering the semi-discrete equations of the forward Euler rule. Alternatively, setting

$$\mathbf{w} \approx \hat{\mathbf{w}} = \delta(t_{n+1}) \mathbf{c}, \quad (8.34)$$

one readily obtains (8.29), namely the semi-discrete equations of the backward Euler rule.

More generally, let

$$\mathbf{w} \approx \hat{\mathbf{w}} = \delta(t_{n+\alpha}) \mathbf{c}, \quad (8.35)$$

where $0 < \alpha \leq 1$. Substituting (8.32) and (8.35) into (8.31) leads to

$$\mathbf{M} \frac{\mathbf{u}_{n+1} - \mathbf{u}_n}{\Delta t_n} + \mathbf{K} [(1 - \alpha) \mathbf{u}_n + \alpha \mathbf{u}_{n+1}] = \mathbf{F}_{n+\alpha},$$

which corresponds to the *generalized trapezoidal rule*. For the special case $\alpha = 1/2$, one recovers 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

$$\mathbf{w} \approx \hat{\mathbf{w}} = \frac{t - t_n}{\Delta t_n} \mathbf{w}_{n+1}, \quad (8.36)$$

where \mathbf{w}_{n+1} is an arbitrary constant vector. In this case, one recovers the Bubnov-Galerkin method in time. Substituting (8.32) and (8.36) into (8.31) leads to

$$\int_{t_n}^{t_{n+1}} \frac{t - t_n}{\Delta t_n} \mathbf{w}_{n+1}^T \left[\mathbf{M} \frac{\mathbf{u}_{n+1} - \mathbf{u}_n}{\Delta t_n} + \mathbf{K} \left\{ \left(1 - \frac{t - t_n}{\Delta t_n} \right) \mathbf{u}_n + \frac{t - t_n}{\Delta t_n} \mathbf{u}_{n+1} \right\} - \mathbf{F} \right] dt = 0. \quad (8.37)$$

Upon integrating (8.37) in time and recalling that \mathbf{w}_{n+1} is arbitrary, one finds that

$$\frac{1}{2} \mathbf{M} (\mathbf{u}_{n+1} - \mathbf{u}_n) + \mathbf{K} \left(\frac{1}{6} \mathbf{u}_n + \frac{1}{3} \mathbf{u}_{n+1} \right) \Delta t_n - \int_{t_n}^{t_{n+1}} \frac{t - t_n}{\Delta t_n} \mathbf{F} dt = 0$$

or

$$\left(\mathbf{M} + \frac{2}{3} \Delta t_n \mathbf{K} \right) \mathbf{u}_{n+1} = \left(\mathbf{M} - \frac{1}{3} \Delta t_n \mathbf{K} \right) \mathbf{u}_n + \bar{\mathbf{F}}, \quad (8.38)$$

where $\bar{\mathbf{F}} := \int_{t_n}^{t_{n+1}} 2 \frac{t - t_n}{\Delta t_n} \mathbf{F} dt$. When \mathbf{F} is a constant, the Bubnov-Galerkin method coincides with the generalized trapezoidal rule with $\alpha = 2/3$.

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 is discussed here in the context of the one-dimensional convection-diffusion equation

$$u_{,t} + \alpha u_{,x} = \epsilon u_{,xx} \quad ; \quad \alpha \geq 0 \quad , \quad \epsilon \geq 0 \quad , \quad (9.1)$$

which is 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

$$u(x) = \frac{1 - e^{\frac{\alpha}{\epsilon}x}}{1 - e^{\frac{\alpha}{\epsilon}L}} \bar{u} \quad . \quad (9.2)$$

The non-dimensional number $Pe := \frac{\alpha}{\epsilon} L$, is known as the *Péclet number* and provides a measure of relative significance of convection and diffusion, such that convection dominates

if $Pe \ll 1$ and diffusion dominates if $Pe \gg 1$. Recalling the definition of the Péclet number, one may rewrite (9.2) as

$$u(x) = \frac{1 - e^{Pe \frac{x}{L}}}{1 - e^{Pe}} \bar{u}. \quad (9.3)$$

It is clear from (9.3) that when diffusion dominates, then $u(x) \simeq \frac{x}{L} \bar{u}$. This is because, in this case the exponentials in (9.3) have exponents that are much less than one, thus can be accurately approximated by their first order Taylor expansion. In contrast, when convection dominates, then the solution is nearly zero throughout the domain except for a thin boundary layer close to $x = L$ 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 $u(x) \simeq e^{\frac{x}{L}-1} \bar{u}$. It is precisely this steep boundary layer that the classical Bubnov-Galerkin method fails to accurately resolve, as will be argued shortly.

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 three-dimensional compressible Newtonian fluid, and which can be expressed as

$$-\nabla p + (\lambda + \mu)\nabla(\nabla \cdot \mathbf{v}) + \mu\nabla \cdot (\nabla \mathbf{v}) + \rho \mathbf{b} = \rho \left(\frac{\partial \mathbf{v}}{\partial t} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}} \right).$$

Here, $p = p(\rho)$ is the pressure, \mathbf{v} is the fluid velocity, ρ is the mass density, \mathbf{b} is the applied body force, 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. 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 $Re := \frac{\rho \|\mathbf{v}\| L}{\mu}$. 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 method and subsequently discretize the resulting equations by $N + 1$ equally-sized finite elements with linear interpolation functions for the dependent variable u , see Figure 9.1. This readily leads to the system of linear algebraic equations

$$\alpha \frac{1}{2h} (u_{I+1} - u_{I-1}) = \epsilon \frac{1}{h^2} (u_{I+1} - 2u_I + u_{I-1}) \quad , \quad I = 1, 2, \dots, N, \quad (9.4)$$

where $h = \frac{1}{N+1}$. In fact, these equations coincide with those obtained by applying the centered-difference method directly on the differential equation. One may rewrite the above equations in the form

$$au_{I-1} + bu_I + cu_{I+1} = 0 \quad , \quad I = 1, 2, \dots, N, \quad (9.5)$$

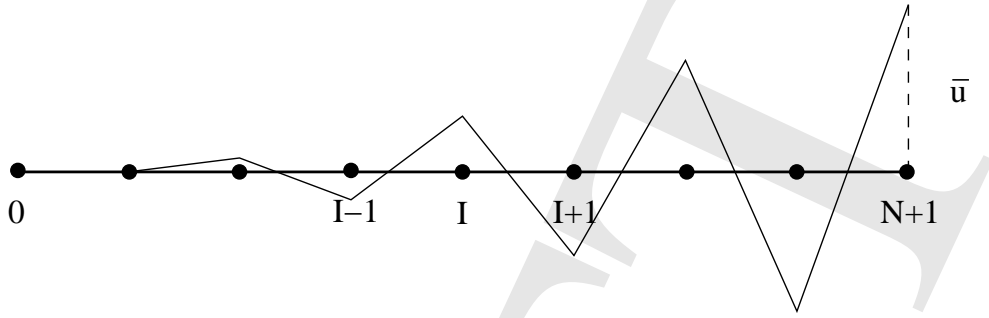


Figure 9.3: *Finite element solution for the one-dimensional convection-diffusion equation for $c > 0$*

To remedy the oscillatory behavior of the Bubnov-Galerkin method, one may choose instead to employ an *upwinding method*. This can be understood by considering again equation (9.5) and writing it for $I = N$ in the form

$$au_{N-1} + bu_N = -c\bar{u} < 0, \quad (9.7)$$

where it is assumed that $c > 0$ (which implies oscillations of the Bubnov-Galerkin solution). With reference to (9.7), one may argue that the term au_{N-1} is small compared to the others, and b is positive, hence $u_N < 0$. Similarly, one may write (9.5) for $I = N - 1$ and use the same argument to conclude that $u_{N-1} > 0$, etc., which essentially explains the presence of oscillations when $c > 0$. 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 (i.e., skew the interpolation toward the part of the domain where the solution is relatively constant). To this end, equation (9.4) may be replaced by

$$\alpha \frac{1}{2h}(u_I - u_{I-1}) = \epsilon \frac{1}{h^2}(u_{I+1} - 2u_I + u_{I-1}) \quad , \quad I = 1, 2, \dots, N, \quad (9.8)$$

which is tantamount to using an *upwind difference* approximation $\frac{du}{dx}\Big|_I \simeq \frac{1}{h}(u_I - u_{I-1})$ as opposed to a centered difference $\frac{du}{dx}\Big|_I \simeq \frac{1}{2h}(u_{I+1} - u_{I-1})$ for the convective term. One may interpret this upwind difference as the sum of the corresponding centered difference and an artificial viscosity term. Indeed, note that

$$\frac{1}{h}(u_I - u_{I-1}) - \frac{1}{2h}(u_{I+1} - u_{I-1}) = -\frac{h}{2}(u_{I-1} - 2u_I + u_{I+1}). \quad (9.9)$$

Clearly, the right-hand side of (9.9) is a term that would contribute additional diffusion, as it corresponds to the discrete Laplacian operator with a *grid diffusion* constant $k_h := \frac{h}{2}$. This is not necessarily an undesirable feature as it is well-known that the centered difference method underdiffuses (i.e., its convergence is from below in the appropriate energy norm, see Chapter 7).

It has been shown in the context of the one-dimensional convection-diffusion equation that there exists an optimal amount 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 corresponds to grid diffusion $k_{h,opt} := \frac{h}{2} \left[\coth Pe_h - \frac{1}{Pe_h} \right]$.

The upwinding method can be interpreted as a Petrov-Galerkin method in which the weighting functions for a given node are preferentially weighing its upwind domain, see Figure 9.4 for a schematic depiction. To further appreciate this point, write the weak form

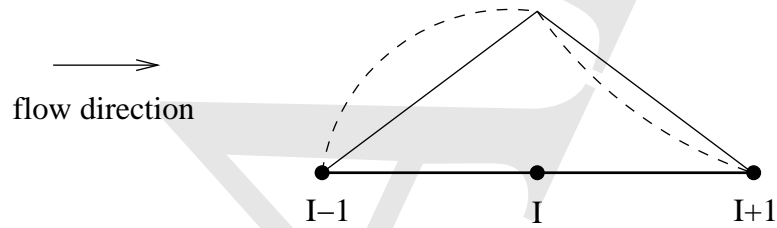


Figure 9.4: A schematic depiction of the upwind Petrov-Galerkin method for the convection-diffusion equation (continuous line: Bubnov-Galerkin, broken line: Petrov-Galerkin)

of the convection-diffusion equation as

$$\int_0^L w(\alpha u_{,x} - \epsilon u_{,xx}) dx = 0 . \tag{9.10}$$

where the weighting function w satisfies $w_1(0) = w_1(L) = 0$. Using integration by parts, the weak form (9.10) can be rewritten as

$$\int_0^L w \alpha u_{,x} dx + \int_0^L w_{,x} \epsilon u_{,x} dx . \tag{9.11}$$

Recalling now that upwinding can be interpreted as artificial diffusion with constant $k_{h,opt}$, one may modify (9.11) so that it takes the form

$$\int_0^L w \alpha u_{,x} dx + \int_0^L w_{,x} (\epsilon + k_{h,opt}) u_{,x} dx . \tag{9.12}$$

If the total diffusive contribution in (9.12) is expressed as

$$w_{,x} (\epsilon + k_{h,opt}) u_{,x} = w'_{,x} \epsilon u_{,x} ,$$

then w' is the new weighting function for the diffusive part of the convection-diffusion equation in the spirit of the Petrov-Galerkin method.

Multi-dimensional generalizations of upwind finite element methods need special case. 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 7.3 can be extended to include the effects of inertia. The resulting equations of motion take the form

$$\begin{aligned}
 \nabla \cdot \boldsymbol{\sigma} + \mathbf{f} &= \rho \ddot{\mathbf{u}} && \text{in } \Omega \times I, \\
 \boldsymbol{\sigma} \mathbf{n} &= \bar{\mathbf{t}} && \text{on } \Gamma_q \times I, \\
 \mathbf{u} &= \bar{\mathbf{u}} && \text{on } \Gamma_u \times I, \\
 \mathbf{u}(x_1, x_2, x_3, 0) &= \mathbf{u}_0(x_1, x_2, x_3) && \text{in } \Omega, \\
 \mathbf{v}(x_1, x_2, x_3, 0) &= \mathbf{v}_0(x_1, x_2, x_3) && \text{in } \Omega,
 \end{aligned} \tag{7.54}$$

where $\mathbf{u} = \mathbf{u}(x_1, x_2, x_3, t)$ is the unknown displacement field, ρ is the mass density, and $I = (0, T)$ with T being a given time. Also, \mathbf{u}_0 and \mathbf{v}_0 are the prescribed initial displacement and velocity fields. Clearly, two sets of boundary conditions are set on Γ_u and Γ_q , respectively, and are assumed to hold throughout the time interval I . Likewise, two sets of initial conditions are set 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} , $\bar{\mathbf{t}}$, $\bar{\mathbf{u}}$, \mathbf{u}_0 and \mathbf{v}_0 , as well as a constitutive equation (7.2) for $\boldsymbol{\sigma}$, find \mathbf{u} in $\Omega \times I$, such that the equations (9.13) are satisfied.

A Galerkin-based weak form of the linear elastostatics problem has been derived in Section 7.3. 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 fixed time can be expressed as

$$\int_{\Omega} \mathbf{w} \cdot \rho \ddot{\mathbf{u}} d\Omega + \int_{\Omega} \nabla_s \mathbf{w} : \boldsymbol{\sigma} d\Omega = \int_{\Omega} \mathbf{w} \cdot \mathbf{f} d\Omega + \int_{\Gamma_q} \mathbf{w} \cdot \bar{\mathbf{t}} d\Gamma. \tag{9.14}$$

Following the development of Section 7.3, the discrete counterpart of (9.14) can be written as

$$\int_{\Omega} \mathbf{w}_h \cdot \rho \ddot{\mathbf{u}}_h d\Omega + \int_{\Omega} \boldsymbol{\epsilon}(\mathbf{w}_h) \cdot \mathbf{D} \boldsymbol{\epsilon}(\mathbf{u}_h) d\Omega = \int_{\Omega} \mathbf{w}_h \cdot \mathbf{f} d\Omega + \int_{\Gamma_q} \mathbf{w}_h \cdot \bar{\mathbf{t}} d\Gamma. \tag{9.15}$$

A Galerkin approximation of (9.15) using the nomenclature of (7.15)-(7.16) leads to a system of the form

$$\mathbf{M}^e \ddot{\mathbf{u}}^e + \mathbf{K}^e \mathbf{u}^e = \mathbf{F}^e + \mathbf{F}^{\text{int},e}, \quad (9.16)$$

where all quantities have already been defined in Section 7.3 except for the element mass matrix \mathbf{M}^e which is given by

$$\mathbf{M}^e := \int_{\Omega^e} \mathbf{N}^{e,T} \rho \mathbf{N}^e d\Omega .$$

Following a standard procedure, 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 (9.16) give rise to their assembled counterparts in the form

$$\mathbf{M}\hat{\mathbf{u}} + \mathbf{K}\hat{\mathbf{u}} = \mathbf{F}, \quad (9.17)$$

where $\hat{\mathbf{u}}$ is the global unknown displacement vector¹. The preceding 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 (9.17) is the *Newmark method*. This method is based on a time series expansion of \hat{u} and $\hat{\mathbf{u}} := \hat{\mathbf{v}}$. Concentrating on the time interval $(t_n, t_{n+1}]$, the Newmark method is defined by the equations

$$\begin{aligned} \hat{\mathbf{u}}_{n+1} &= \hat{\mathbf{u}}_n + \hat{\mathbf{v}}_n \Delta t_n + \frac{1}{2} [(1 - 2\beta)\hat{\mathbf{a}}_n + 2\beta\hat{\mathbf{a}}_{n+1}] \Delta t_n^2, \\ \hat{\mathbf{v}}_{n+1} &= \hat{\mathbf{v}}_n + [(1 - \gamma)\hat{\mathbf{a}}_n + \gamma\hat{\mathbf{a}}_{n+1}] \Delta t_n, \end{aligned} \quad (7.54)$$

where $\Delta t_n := t_{n+1} - t_n$, $\hat{\mathbf{a}} := \hat{\ddot{\mathbf{u}}}$, and β, γ are parameters chosen such that

$$0 \leq \beta \leq \frac{1}{2}, \quad 0 < \gamma \leq 1.$$

The special choice $\beta = \frac{1}{4}$ and $\gamma = \frac{1}{2}$ corresponds to the trapezoidal rule. Likewise the special choice $\beta = 0$ and $\gamma = \frac{1}{2}$ corresponds to the centered-difference rule.

It is clear that the Newmark equations (9.18) define a whole family of time integrators. It is important to distinguish this family into two categories, namely implicit and explicit integrators, corresponding to $\beta > 0$ and $\beta = 0$, respectively.

¹The 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} .

The general implicit Newmark integration method may be implemented as follows: first, solve (9.18)₁ for $\hat{\mathbf{a}}_{n+1}$, namely write

$$\hat{\mathbf{a}}_{n+1} = \frac{1}{\beta\Delta t_n^2}(\hat{\mathbf{u}}_{n+1} - \hat{\mathbf{u}}_n - \hat{\mathbf{v}}_n\Delta t_n) - \frac{1-2\beta}{2\beta}\hat{\mathbf{a}}_n. \quad (9.19)$$

Then, substitute (9.19) into the semi-discrete form (9.17) evaluated at t_{n+1} to find that

$$\left[\frac{1}{\beta\Delta t_n^2}\mathbf{M} + \mathbf{K} \right] \hat{\mathbf{u}}_{n+1} = \mathbf{F}_{n+1} + \mathbf{M} \left[(\hat{\mathbf{u}}_n + \hat{\mathbf{v}}_n\Delta t_n) \frac{1}{\beta\Delta t_n^2} + \frac{1-2\beta}{2\beta}\hat{\mathbf{a}}_n \right]. \quad (9.20)$$

After solving (9.20) for $\hat{\mathbf{u}}_{n+1}$, one may compute the acceleration $\hat{\mathbf{a}}_{n+1}$ from (9.19) and the velocity $\hat{\mathbf{v}}_{n+1}$ from (9.18)₂.

Finally, the general explicit Newmark integration method may be implemented as follows: starting from the semi-discrete equations (9.17) evaluated at t_{n+1} , one may substitute $\hat{\mathbf{u}}_{n+1}$ from (9.18)₁ to find that

$$\mathbf{M}\hat{\mathbf{a}}_{n+1} = -\mathbf{K}(\hat{\mathbf{u}}_n + \hat{\mathbf{v}}_n\Delta t_n + \frac{1}{2}\hat{\mathbf{a}}_n\Delta t_n^2) + \mathbf{F}_{n+1}. \quad (9.21)$$

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. Then, the velocities $\hat{\mathbf{v}}_{n+1}$ are immediately computed from (9.18)₂. Also, the displacements $\hat{\mathbf{u}}_{n+1}$ are computed from (9.18)₁ independently of the accelerations $\hat{\mathbf{a}}_{n+1}$.

Index

- amplification factor, 161
- approximation
 - complete, 80
 - dual, 151
 - global, 67
 - global-local, 68
 - local, 67
 - primal, 151
- backward Euler method, 163
- Banach space, 18
- basis functions, 35, 69
- best approximation property, 145
- bilinear form, 23
 - V -elliptic, 143
 - continuous, 23
- boundary conditions
 - Dirichlet, 32, 156
 - essential, 54
 - natural, 54
 - Neumann, 32, 156
- Bubnov-Galerkin approximation, 36
- bulk modulus, 149
- compatibility condition, 77
- completeness, 73
- condition number, 146
- conditionally stable, 163
- coordinates
 - area coordinates, 92
 - volume coordinates, 101
- Crank-Nicolson rule, 166
- critical step-size, 163
- directional differential, 27
- displacement, 127
- domain
 - parent, 104
 - physical, 104
- energy norm, 144
- explicit, 165
- finite element
 - definition, 75
- finite elements
 - isoparametric, 105
 - Lagrangian, 94
 - serendipity, 94
 - space-time, 155
 - subparametric, 105
 - superparametric, 105
- first fundamental error, 145
- flop, 122
- forcing vector, 36, 158
- formal adjoint, 23
- formal operator, 23
- forward Euler method, 162

- Fourier coefficients, 71
- Fourier representation, 71
- function, 13
 - continuous, 15
 - continuous t point, 15
 - support, 67
- functional, 14
- functions
 - linearly independent, 68
 - orthogonal, 69
- Galerkin approximation, 35
- Galerkin formulation, 33
- Gaussian quadrature, 115
- grid diffusion, 171
- Hilbert space, 18
- implicit, 165
- incompressibility
 - exact, 149
 - near, 149
- index
 - dummy, 129
 - free, 129
- initial condition, 156
- inner product, 16
 - orthogonality, 16
- inner product space, 16
- interpolation, 35
 - hierarchical, 85
 - Lagrangian, 84
 - standard, 85
- inverse function theorem, 107
- Korn's inequality, 144
- Laplace-Poisson equation, 32
- Lax-Milgram theorem, 145
- Legendre polynomials, 117
- linear operator
 - adjoint, 22
 - bounded, 21
 - continuous, 21
 - positive, 22
 - self-adjoint, 22
 - symmetric, 22
- linear space
 - complete, 18
- mapping, 13
 - one-to-one, 104
 - onto, 104
- mass matrix, 158
- mesh, 75
 - structured, 104
 - unstructured, 104
- mid-point rule, 115
- Minimum Total Potential Energy theorem, 132
- Neumann problem, 42
- Newmark method, 173
- Newton-Cotes closed integration, 115
- nodal points, 75
- norm, 17
- normed linear space, 17
- operator, 14
- Péclet number, 167
 - grid, 169
- Pascal triangle, 80

-
- patch test, 140
 - PDE
 - linear vx. non-linear, 7
 - order, 7
 - penalty parameter, 153
 - penalty regularization, 153
 - Petrov-Galerkin approximation, 36
 - Poincaré inequality, 143
 - refinement
 - h-refinement, 78
 - hp-refinement, 78
 - p-refinement, 78
 - r-refinement, 78
 - Reynolds number, 168
 - sampling points, 114
 - semi-discretization, 155
 - sequence
 - Cauchy onvergent, 18
 - set
 - open, 18
 - shape functions, 83
 - Simpson's rule, 114
 - Sobolev space, 20
 - stable, 162
 - static condensation, 86
 - stiffness matrix, 36, 158
 - Stokes' flow, 150
 - strain
 - deviatoric, 148
 - volumetric, 148
 - test functions, 30
 - total potential energy, 132
 - trapezoidal rule, 114
 - generalized, 166
 - unconditionally stable, 164
 - unstable, 163
 - upwind difference, 170
 - upwinding, 170
 - variation, 24
 - volumetric locking, 153
 - weighting functions, 30
 - weights, 114
 - zero-energy modes, 138
-