# FINITE ELEMENT METHOD IN MECHANICS

**Spring term 2002**

**Eero-Matti Salonen and Jouni Freund**

## CONTENTS (23.4.2002)

# 1 INTRODUCTION

## 1.1 BASICS OF THE FINITE ELEMENT METHOD

The *finite element method* (elementtimenetelmä) is a general and powerful numerical method to solve field problems (mainly ordinary and partial differential equations). It has become possible in practice with the advent of the digital computer. An essential feature of the method is its systematic way to approximate functions by a *discrete model* (diskreetti malli). The model is generated by dividing the domain of the function under consideration in sub-domains or *finite elements* (elementti) (total number $n_e$). On the boundaries of the elements and often also inside them certain points, *nodal points* or shortly *nodes* (solmupiste, solmu) (total number $n_n$), are further selected. The resulting configuration is called the *element mesh* (elementtiverkko). The function is approximated in each element with simple functions — usually polynomials — by which it is interpolated inside the element employing its values at the nodes, the *nodal values* (solmuarvo).

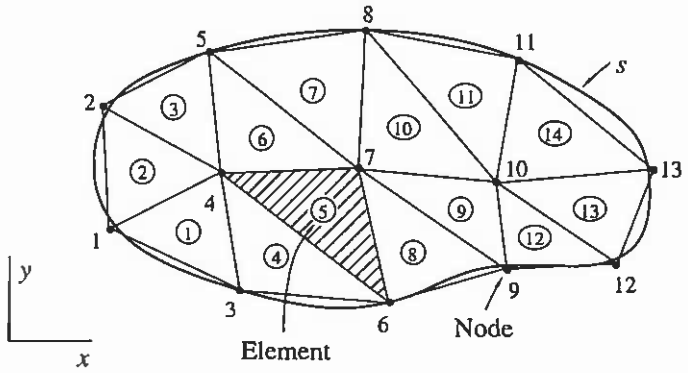



**Figure 1.1** Division of a two-dimensional domain into triangular elements ($n_e = 14$, $n_n = 13$).

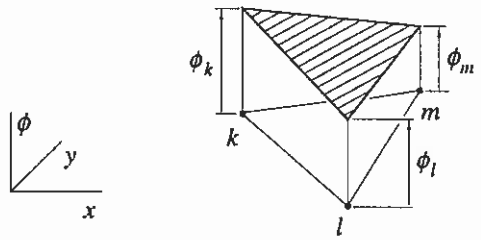



**Figure 1.2** Linear approximation of $\phi(x, y)$ in an element having nodes $k$, $l$, $m$.





This procedure is illuminated in Figures 1.1, 1.2 and 1.3 for a function $\phi(x, y)$ of two independent variables $x$ and $y$. The elements, shown here, are called three-noded triangular elements or linear triangular elements.

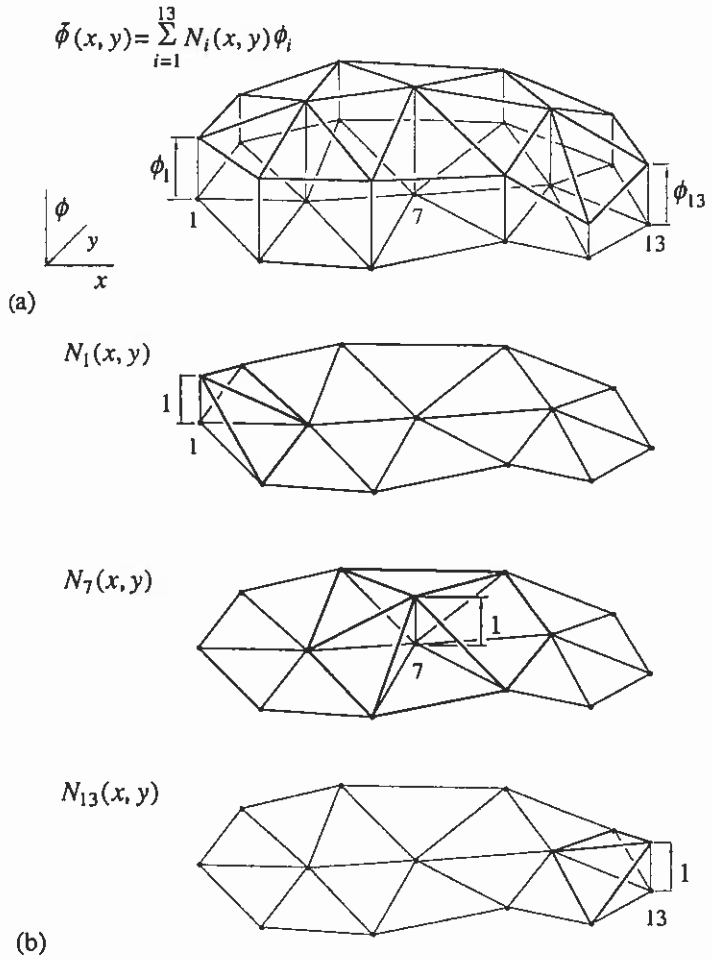



**Figure 1.3 (a)** Approximation $\tilde{\phi}$ of $\phi$. **(b)** Three global shape functions.

We realize with the help of Figure 1.3 that it is possible to define interpolation functions or in the finite element terminology *shape functions* (muotofunktio) $N_i(x, y)$, $i = 1, 2, \cdots, n_n$ so that the approximation in the whole domain can be expressed in the linear form (with respect to the nodal values)

$$\boxed{\bar{\phi}(x,y)=\sum_{i=1}^{n_n} N_i(x,y)\phi_i} = N_1(x,y)\phi_1 + N_2(x,y)\phi_2 + \cdots \qquad (1)$$

A shape function *obtains the value one at the node corresponding to its index and the value zero at all other nodes* and differs from zero at most in the elements connected to the node in question.

According to (1), after a certain element mesh with its corresponding shape functions has been selected, *the approximation is wholly determined by fixing the discrete nodal values* $\phi_i$.

The way of presentation (1) is suitable for theoretical considerations but in practical calculations these so-called *global shape functions* (gobaali muotofunktio) $N_i$ are not used. Namely, in the domain of a certain element $e$, approximation (1) can be clearly given simply as

$$\boxed{\phi^e(x,y)=\sum_{i=1}^{n_n^e} N_i^e(x,y)\phi_i^e} = N_1^e(x,y)\phi_1^e + N_2^e(x,y)\phi_2^e + \cdots \qquad (2)$$

where the quantities $N_i^e$ are so-called *local* or *element shape functions* (lokaalinen muotofunktio, paikallinen muotofunktio, elementtimuotofunktio) which have been defined only in the domain of element $e$. (They coincide with the global shape functions in the element domain because the global shape functions are obtained in a piecewise manner from the local ones.) $n_n^e$ is the total number of nodes of element $e$. The values $1,2,\cdots,n_n^e$ of index $i$ refer to the *local node numbers* or local indices or node identifiers (sisäinen, paikallinen, lokaali solmunumero). At a local node $r$

$$\phi_r^e = \phi_i \qquad (3)$$

where $i$ is the *global node number* (ulkoinen, globaali solmunumero) corresponding to the local node number $r$. In (3), we can similarly also speak about *local and global nodal values*. The global and local numbering of the nodes and also the numbering of the elements is performed normally starting from number 1 without "gaps". Figure 1.4 describes the local shape functions in our example case in a generic element $e$.

**Remark 1.1.** The presentation above has been for continuous functions. Naturally, discontinuous functions can also be described by finite elements. The simplest case would consist of constant function values in the elements. □



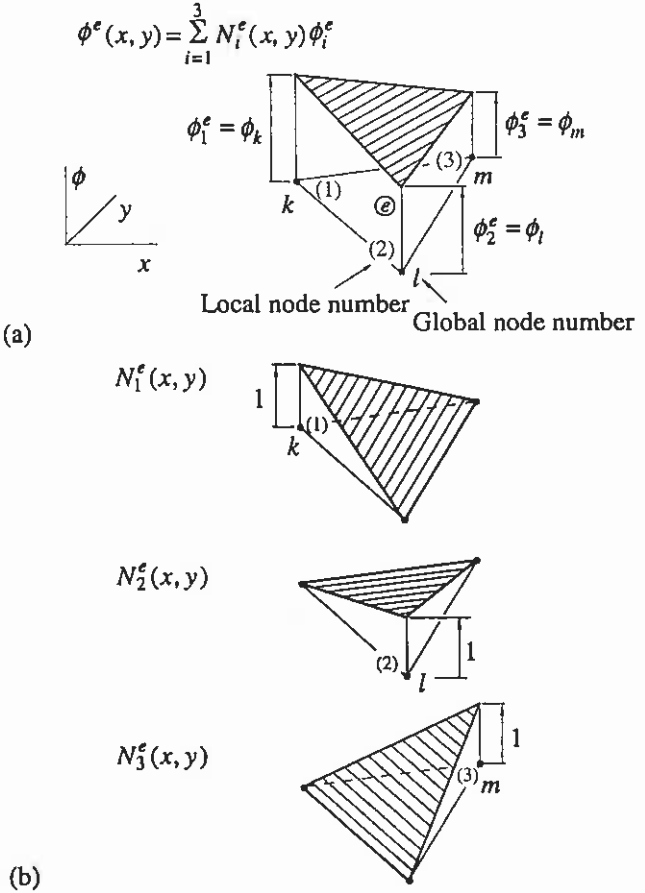



**Figure 1.4 (a)** Approximation of $\phi$ in element $e$. **(b)** Element shape functions.

Let us consider as a specific example the element 5 of Figure 1.1. Let the local node numbering for it be that shown in Figure 1.5. Thus according to (2),

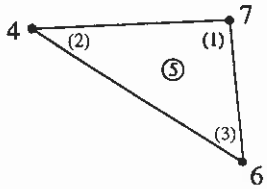



**Figure 1.5** Element 5.

$$\phi^5(x,y) = N_1^5(x,y)\phi_1^5 + N_2^5(x,y)\phi_2^5 + N_3^5(x,y)\phi_3^5 \tag{4}$$

and as additional information there exists the correspondence given in the following table:

| Local number $r$ | | Global number $i$ |
|---|---|---|
| 1 | $\triangleq$ | 7 |
| 2 | $\triangleq$ | 4 |
| 3 | $\triangleq$ | 6 |

(5)

This type of data for each element is enough to describe the connection between the global and local shape functions and between the global and local nodal values. For instance, we now know based on the table and equation (3) that

$$\phi_1^5 = \phi_7, \quad \phi_2^5 = \phi_4, \quad \phi_3^5 = \phi_6 \tag{6}$$

**Remark 1.2.** The shape functions of the finite element method can be considered as special cases of the so-called *trial functions* (yritefunktio) or co-ordinate functions used classically for example in the Ritz (1909) method. Classical trial functions are usually (smooth and) non-zero nearly everywhere in the whole domain of the problem. The global shape functions, on the other hand, are usually (non-smooth and) non-zero in a rather small part of the whole domain; cf. Figure 1.3 changed to have a more realistic dense mesh. □

In the finite element method definite integrals over the domain under consideration and over its boundary are constantly needed. For instance in the case of Figure 1.1 the area integral of a function $\phi(x,y)$

$$\int_A \phi \, dA \approx \sum_{e=1}^{n_e} \int_{A^e} \phi \, dA \approx \sum_{e=1}^{n_e} \int_{A^e} \phi^e \, dA \tag{7}$$

According to (2) and (7) the integration over the domain of an element *can be performed totally independent of the other elements* and the *final integral is obtained as a sum from the individual element contributions*. These two items may be considered as basic properties of the finite element method whose importance becomes apparent later.

The first approximation in (7) is geometrical in nature and due to the fact that the curved boundary $s$ of the domain in Figure 1.1 is replaced by a polygon. (To be more precise, function $\phi(x,y)$ is not defined outside the domain $A$. Thus if some of the elements are partly outside the domain as in the figure, equation (7) is not meaningful unless $\phi(x,y)$ is thought to be somehow extrapolated outside.) It is intuitively obvious that the geometric error decreases when the number of elements is increased and also if more refined elements with curved sides are used. The second approximation comes from using the representation



(2) instead of the exact one. Again it seems obvious that the errors must decrease when the number of elements is increased.

This far the finite element method has been described mainly from the point of view of an interpolation method for a function thought to be given. The finite element approximation can be used and is used in this sense in many applications, say to evaluate complicated integrals, to represent complicated surfaces and volumes, to smoothen results from experimental data, etc. The finite element method proper is arrived at, however, when it is used to determine the unknown function in a problem where the information on the function is based only on the governing differential equation and boundary conditions. The task is to select the approximation so that it is in some sense near the exact solution. This can happen in principle by two different ways. The first is called *residual formulation* (jäännösformulaatio) and the second *variational formulation* (variaatioformulaatio). They are described later. The finite element method transforms a differential equation problem to a system of $n_n$ algebraic equations from which the unknown nodal values are determined. If the differential equation is linear, so is the algebraic system. This operation producing the transformation is called *discretization* (diskretointi). The discretization transforms the study of a continuous function (which is unknown at an infinite number of points) to the study of $n_n$ values at separate or discrete points.

Some features of the finite element method have been described above in a very elementary form. Generalizations are given in the following chapters. Let us however shortly mention that in this text the two basic forms emerging through the discretization are the *linear system of algebraic equations*

$$[K]\{a\} = \{b\} \tag{8}$$

and to lesser extent the *linear system of first order differential equations*

$$[M]\dot{\{a\}} + [K]\{a\} = \{b\} \tag{9}$$

The meaning of the notations can be found from the NOMENCLATURE section. Quantity $\{a\}$ consists of the unknown nodal values. In practical problems the number of unknowns can easily be of the order of thousands, even of millions. Efficient algorithms taking into account the specific features of the finite element method have been developed to solve systems like (8) and (9).

To assimilate the finite element method may be said to consist of roughly *three levels*. The *first level* consists of those general mathematical principles on which the discretization is based. The *second level* consists of that detailed bookkeeping by which the information is transferred to the computer. The *third*

*level* consists of the algorithms to solve discrete equation systems like (8) and (9). In this text the emphasis is on the first and second level.

**Remark 1.3.** Let us comment the following with respect to notation. The finite element approximation of a quantity is denoted in this text usually by a tilde over the symbol of the quantity; say $\phi \approx \tilde{\phi}$. To simplify the formulas this rule is, however, not followed consistently if the meaning is quite obvious from the context. Thus the finite element approximation in an element $e$ is denoted as in (2) without the tilde and similarly a nodal value with $\phi_i$ even if they are in general approximate. Often also the shape functions of a separate element are written without the element number superscript. Rather common alternative notations for the approximation in the literature are $\hat{\phi}$ and $\phi^h$ or $\phi^N$. In the latter two, the superscripts are to remind that the approximation depends on the density of the element mesh as $h$ is the conventional symbol for a typical element size and $N$ on the other hand tells that the approximation depends on the number of nodal values used. □

## 1.2 HISTORICAL BACKGROUND

The finite element method originated in the area of structural mechanics in the aerospace industry. This is explainable perhaps by: great demand for accurate stress analysis of complicated structures, early availability of digital computers.

The article by Turner, Clough, Martin and Topp (1956) is generally considered as the birth paper of the finite element method in engineering. The three-noded triangular element was presented and applied to plane stress analysis. The mathematician Courant (1943), however, had already given a formulation in connection of warping function determination in torsion problems containing all the main ingredients of the finite element method.

Hrennikoff (1941) introduced his framework method in which continuum structures are discretized by replacing them with bar frameworks where the elastic properties of the bars are selected so that the resulting structure in some sense simulates the behavior of the original one. Elements generated from bars are still in use in some practical structural applications. One phase in the development of the finite element method may be considered as an attempt to mathematize the physical approach of Hrennikoff.

The term "finite element" was coined by Clough (1960).

Gradually it was realized that the finite element method for structures was just a special case of the much older Ritz method, Ritz (1909) — also called the Rayleigh-Ritz method, Strutt (1870) — applied to the principle of minimum potential energy. This discovery gave mathematical credibility to the method and generated the idea to try to apply it to other problems of physics where a variational principle was known to exist.

After that the progress of the method has been fast and the application areas have become vast in various physical problems. The first basic version, residual



formulation (historically following the variational formulation in the field of finite elements), mentioned in Section 1.1, has further essentially increased the applicability of the method. Only this version, in fact, has made possible successful solution of general fluid mechanics problems by the finite element method. The finite element method as such is, however, not in any way tied just to physics. A more general point of view is to consider it as a very general computational method of applied mathematics.

Let us add the following quotation from Cook (1981): "As late as 1967, engineers and mathematicians worked with finite elements in apparent ignorance of each other. (Today the two camps are aware of one another, but mathematicians are rarely interested in engineering problems, and engineers are rarely able to understand mathematicians.)" Also, Cook, Malkus and Plesha (1989): "Ten papers about finite elements were published in 1961, 134 in 1968, and 844 in 1971. By 1976, two decades after engineering applications began, the cumulative total of publications about finite elements exceeded 7000. By 1986, the total was about 20,000." Finally, Liu, Belytschko, Oden (Comput. Methods Appl. Mech. Engrg., 139, 1996, 1-2): "There is no dispute that the single most important advance in numerical methods in the twentieth century is the invention of the finite element method."

## REFERENCES

At present, there exist tens of textbooks on the finite element method. A sample of references is given in the following.

**Early references**

Clough, R. W. (1960). The Finite Element Method in Plane Stress Analysis, *J. Struct. Div. ASCE, Proc. 2 nd Conf. Electronic Computation*, 345 - 378.

Courant, R. (1943). Variational Methods for the Solution of Problems of Equilibrium and Vibrations, *Bull. Amer. Math. Soc.*, Vol. 49, 1 - 23.

Hrennikoff, A. (1941). Solution of Problems in Elasticity by the Framework Method, *J. Appl. Mech.*, Vol. 8, A169 - A175.

Ritz, W. (1909).Über eine neue Methode zur Lösung gewisser Variations-Probleme der mathematischen Physik, *J. Reine Angew. Math.*, Vol. 135, 1 - 61.

Strutt, J. W. (Lord Rayleigh) (1870). On the Theory of Resonance, *Trans. Roy. Soc.* (London), Vol. A161, 77 – 118.

Turner, M. J., Clough, R. W., Martin, H. C. and Topp, L. (1956). Stiffness and Deflection Analysis of Complex Structures, *J. Aero. Sci.* Vol. 23, No. 9, 805 - 823.

Zienkiewicz, O. C. and Cheung, Y. K. (1967). *The Finite Element Method in Structural and Continuum Mechanics*, McGraw-Hill, London. The first textbook on finite elements.

**Books with mathematical emphasis**

Ainsworth, M. and Oden, J. T. (2000). A Posteriori Error Estimation in Finite Element Analysis, Wiley, New York, ISBN 0-471-29411-X.

Eriksson, K., Estep, K. D., Hansbo, P. and Johnson, C. (1996). *Computational Differential Equations*, Studentlitteratur, Lund, ISBN 91-44-49311-8.

Reddy, B. D. (1986). *Functional Analysis and Boundary Value Problems: an Introductory Teatment*, Longman, New York, ISBN 0-470-20384-6. A very pleasant and clear presentation of the basic tools needed to understand the deeper mathematics of finite elements.

Strang, G. and Fix, G. J. (1973). *An Analysis of the Finite Element Method*, Prentice-Hall, Englewoods Cliffs, New Jersey, ISBN 0-13-032946-0. One of the first books dealing with the mathematics of finite elements.

Zienkiewicz, O. C. and Morgan, K. (1983). *Finite Elements and Approximation*, Wiley, Chichester, ISBN 0-471-89089-8.

**Books with wide application area**

Huebner, K. H. and Thornton, E. A. (1982). *The Finite Element Method for Engineers*, 2nd ed., Wiley, Chichester.

Zienkiewicz, O. C. and Taylor, R. L. (2000). *The Finite Element Method*, 5th ed., Butterworth-Heineman, Oxford. Vol. 1: *The Basis*, ISBN 0 7506 5049 4. Vol 2: *Solid Mechanics*, ISBN 0 7506 5055 9. Vol 3: *Fluid Dynamics*, ISBN 0 7506 5050 8. A monumental work by two distinguished experts.

**Books on solid mechanics**

Cook, R. D. (1981). *Concepts and Applications of Finite Element Analysis*, 2nd ed., Wiley, New York.

Cook, R. D., Malkus, D. S. and Plesha, M. E. (1989). *Concepts and Applications of Finite Element Analysis*, 3nd ed., Wiley, New York. An excellent "starting" text for applications in structural mechanics.

Bathe, K. J. (1996). *Finite Element Procedures*, Prentice-Hall, Englewood Cliffs, New Jersey, ISBN 0-13-301458-4.

Belytschko, T., Liu, W. K. and Moran, B. (2000). *Nonlinear Finite Elements for Continua and Structures*, Wiley, Chichester, ISBN 471-988774-3.

Hughes, T. J. R. (1987). *The Finite Element Method — Linear Static and Dynamic Finite Element Analysis*, Prentice-Hall, Englewood Cliffs, New Jersey, ISBN 0-13-317017-9.

Irons, B. and Ahmad, S. (1980). *Techniques of Finite Elements*, Wiley, Ellis Horwood, Chichester, ISBN 0-85312-130-3. A personal representation with strong physical insight. Irons' influence on the development on some of the basic concepts of the finite element method has been very important.

**Books on fluid mechanics and heat transfer**

Baker, A. J. (1983). *Finite Element Computational Fluid Mechanics*, Hemisphere. Washington, D. C., ISBN 0-07-Y66187-1.

Baker, A. J. and Pepper, D. W. (1991): *Finite Elements 1.2.3.*, McGraw-Hill, New York, ISBN 0-07-909 975-0.

Comini, G., Del Giudice, S. and Nonino, C. (1994). *Finite Element Analysis in Heat Transfer, Basic Formulation and Linear Problems*, Taylor & Francis, London, ISBN 1-56032-354-X.

Connor, J. J. and Brebbia, C. A. (1976). *Finite Element Techniques for Fluid*, Newnes-Butterworths, London, ISBN 0-408-00176-3. Probably the first textbook on applications in fluid mechanics.

Hämäläinen, J. and Järvinen, J. (1994): *Elementtimenetelmä virtauslaskennassa*, CSC - Tieteellinen laskenta Oy, ISBN 952-9821-07-7. First textbook in Finnish on finite elements in fluid mechanics.



Lewis, R. W., Morgan, K., Thomas, H. R. and Seetharamu, K. N. (1996). *The Finite Element Method in Heat Transfer Analysis*, Wiley, Chichester, ISBN 0-471-94362-2.

Pironneau, O. (1989). *Finite Element Method for Fluids*, Wiley, Chichester, ISBN 0-471-92255-2.

Reddy, J. N. and Gartling, D. K. (2001). *The Finite Element Method in Heat Transfer and Fluid Dynamics*,2nd ed., CRC Press, Boca Raton, ISBN 0-8493-2355-X.

**Books with emphasis on programming**

Akin, J. E. (1994). *Finite Elements for Analysis and Design*, Academic Press, London, ISBN 0-12-047654-1.

Kwon, Y. W. and Bang, H. (2000). *The Finite Element Method Using* MATLAB, 2nd ed., CRC Press, Boca Raton, ISBN 0-8493-0096-7.

Smith, I. M. (1982). *Programming the Finite Element Method with Application to Geomechanics*, Wiley, Chichester.

# 2 DIFFUSION

## 2.1 DISCRETIZATION PRINCIPLES

The basic ideas for achieving the discretization are explained first in a simple setting using classical trial functions (see Remark 1.2). The introduction of the finite element approximation thereafter only means an additional bookkeeping effort but no new principles.

### 2.1.1 Heat conduction model example

Let us consider the simple *heat conduction problem* (lämmönjohtumisprobleema) described by the differential equation

$$R(T) \equiv \boxed{\frac{\mathrm{d}}{\mathrm{d}x}\left(-k\frac{\mathrm{d}T}{\mathrm{d}x}\right) - s = 0} \qquad \text{in } \Omega = ]a,b[ \tag{1}$$

and the boundary conditions

$$R_{\mathrm{D}}(T) \equiv \boxed{T - \overline{T} = 0} \qquad \text{on } \Gamma_{\mathrm{D}} = \{a\} \tag{2}$$

$$R_{\mathrm{N}}(T) \equiv \boxed{-k\frac{\mathrm{d}T}{\mathrm{d}x} - \overline{q} = 0} \qquad \text{on } \Gamma_{\mathrm{N}} = \{b\} \tag{3}$$

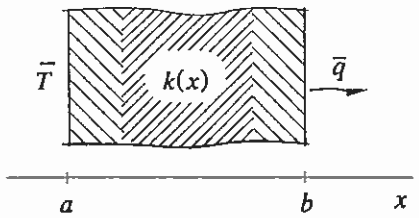



**Figure 2.1** One-dimensional heat conduction through a wall.

$T(x)$ is the unknown *temperature* (lämpötila) to be determined as a function of the independent variable $x$ in the closed interval $\overline{\Omega} = [a,b]$ of the $x$-axis (Figure 2.1). The *thermal conductivity* (lämmönjohtavuus) $k(x)$, and the *heat source rate per volume* (lämpölähteen antoisuus) $s(x)$ are given quantities. The boundary conditions mean that on the left-hand boundary $x = a$ the value of the temperature is given $= \overline{T}$ and that on the right-hand boundary $x = b$ the value of the *heat flow rate density* (lämpövirran tiheys) — considered positive when the flow is out of the body — is given $= \overline{q}$. These equations are a simple





special case — steady pure diffusion problem — of the general diffusion-convection-reaction problem described in detail in Appendix A ($\phi \triangleq T$, $D \triangleq k$, $s \triangleq f$, $\overline{q} \triangleq \overline{j}^{\mathrm{d}}$).

**Remark 2.1.** Equations (1), (2), (3) are written here and similarly often in the following in a standard form — looking perhaps somewhat awkward — but proving from the point of view of describing the ideas behind the residual formulation convenient:

$$\boxed{\text{Something on the left - hand side} = 0} \tag{4}$$

This form makes the initial derivations to proceed systematically. However, later we usually give up this practice with respect to the boundary conditions. □

**Remark 2.2.** Most of the notation is explained in the NOMENCLATURE section. The general symbol for an open domain will be $\Omega$, for its boundary $\Gamma$ and the closure of $\Omega$ is denoted $\overline{\Omega}$. These notations have been employed already here to get the reader accustomed to them. The symbol $\{\cdot\}$ is used in addition for sets also for column matrices. Here we can for instance write

$$\begin{aligned} &\Omega = ]a,b[ = \{x : x \in R,\ a < x < b\} \\ &\Gamma = \Gamma_{\mathrm{D}} \cup \Gamma_{\mathrm{N}} = \{a,b\}, \qquad \Gamma_{\mathrm{D}} \cap \Gamma_{\mathrm{N}} = \varnothing \\ &\overline{\Omega} = \Omega \cup \Gamma = [a,b] = \{x : x \in R,\ a \le x \le b\} \end{aligned} \tag{5}$$

The standard three boundary conditions, frequently occurring, concern the value of a function, roughly the normal derivative of a function and the linear combination of them. They are called the *Dirichlet condition*, the *Neumann condition* and the *Robin condition*, respectively. They are referred to here with the subscripts D, N, R. Similarly we will speak about the Dirichlet, Neumann and Robin boundary. □

### 2.1.2 Residual formulation

**Starting point.** We now proceed to derive a residual formulation corresponding to our model problem described by equations (1), (2), (3). We select an *arbitrary* (smooth enough) function $w(x)$ and two *arbitrary* constants $w_{\mathrm{D}}$ and $w_{\mathrm{N}}$. We multiply (1), (2), (3) by them, respectively, further integrate the first equation generated over the domain and finally add the resulting equations together *to still obtain an equation*

$$F \equiv \int_a^b w\left[\frac{\mathrm{d}}{\mathrm{d}x}\left(-k\frac{\mathrm{d}T}{\mathrm{d}x}\right) - s\right]\mathrm{d}x + w_{\mathrm{D}}\left(T - \overline{T}\right)\Big|_{x=a} + w_{\mathrm{N}}\left(-k\frac{\mathrm{d}T}{\mathrm{d}x} - \overline{q}\right)\Big|_{x=b} = 0 \tag{6}$$

Employing the $R$-shorthand notations in (1), (2), (3) and some of the general domain notations we get more concisely

$$F \equiv \boxed{\int_\Omega wR\,\mathrm{d}\Omega + w_{\mathrm{D}}R_{\mathrm{D}} + w_{\mathrm{N}}R_{\mathrm{N}} = 0} \tag{7}$$

This equation followed by manipulating the governing equations (1), (2), (3). Thus,

$$\boxed{\begin{array}{l}\text{If equations (1), (2) and (3) are satisfied, equation (7)}\\ \text{is also satisfied for arbitrary } w, w_{\mathrm{D}} \text{ and } w_{\mathrm{N}}.\end{array}} \tag{8}$$

More importantly, we can state conversely:

$$\boxed{\begin{array}{l}\text{If equation (7) is satisfied for arbitrary } w, w_{\mathrm{D}} \text{ and } w_{\mathrm{N}},\\ \text{equations (1), (2) and (3) are also satisfied.}\end{array}} \tag{9}$$

The truth of (9) is based crucially on the arbitrariness of the quantities $w$, $w_{\mathrm{D}}$ and $w_{\mathrm{N}}$. For instance, one can first, say, select $w(x) \equiv 0$ and $w_{\mathrm{N}} = 0$. Equation (7) gives $w_{\mathrm{D}} R_{\mathrm{D}} = 0$ and as $w_{\mathrm{D}}$ is arbitrary there follows

$$R_{\mathrm{D}} = 0 \tag{10}$$

Thereafter, say, the selection $w(x) \equiv 0$ and $w_{\mathrm{D}} = 0$ gives $w_{\mathrm{N}} R_{\mathrm{N}} = 0$ and there follows

$$R_{\mathrm{N}} = 0 \tag{11}$$

Finally taking $w_{\mathrm{D}} = 0$, $w_{\mathrm{N}} = 0$ gives

$$\int_{\Omega} w R \, \mathrm{d}\Omega = 0 \tag{12}$$

for an arbitrary $w(x)$. Using the way of thinking explained in Appendix D (the fundamental lemma of variational calculus with the role of $\delta\phi$ replaced here by $w$) there follows also the field equation

$$R = 0 \quad \text{in } \Omega \tag{13}$$

An equation like (7) is called in the literature the *weak form* or *weak statement* (heikko muoto) corresponding to a differential equation formulation such as (1), (2), (3) which is, as an opposite, called in this connection as the *strong form* (vahva muoto).

**Remark 2.3.** In the mathematical literature instead of the term weak form quite often the name *variational form* or *variational statement* (variaatiomuoto) is used. In this text, however, we reserve the word "variational" in the normal custom of engineering only to those cases where a variational principle is used, i.e., stationarity of a functional is involved (see Appendix D and Chapter 5). □



The quantities $w$ used to multiply field equations and boundary conditions to obtain the weak form are called *weighting functions* or *test functions* (painofunktio; oikeastaan paremmin painotusfunktio, testifunktio). The latter name is especially appropriate; we test the satisfaction of field equations and boundary conditions with respect to the test functions: is the weak form satisfied? The word "weak" (does not mean here the opinion of your colleague about the standard of your work) refers to the fact that the solution in the formulation, cast into an integral form, needs often less (weaker) smoothness than the solution of the differential equation formulation. (Mathematicians represent these facts by carefully defining those function spaces where the solution and the test functions must be.) This is especially the case when the weak form has been further transformed by integrations by parts.

We now present the idea behind the approximate solution of the model problem of Section 2.1.1 based on the use of the weak form. The starting point is to assume an approximation of the type

$$\boxed{\tilde{T}(x) = \sum_{i=1}^{N} a_i \varphi_i(x)} = a_1\varphi_1(x) + a_2\varphi_2(x) + \cdots + a_N\varphi_N(x) \tag{14}$$

This in principle of the same type as (1.1.1). The trial functions $\varphi_i(x)$, equivalents of the global shape functions $N_i(x, y)$, are given and the multipliers $a_i$, often called *undetermined parameters* (määräämätön parametri), equivalents to the nodal values $\phi_i$, are the unknowns. (For reasons of convention, we have reversed the order of the terms in the representation.) Approximation (14) is mathematically a *linear combination* (lineaarikombinaatio) of the trial functions. To use more precise terminology, we could call $\tilde{T}(x)$ the *trial solution* (yriteratkaisu tai lyhyesti yrite) which is *spanned* (virittää) by the *trial basis functions* (yritekantafunktiot) $\varphi_i(x)$. This type of representation abounds in mathematics. For instance, Fourier series is one important example, the basis functions consisting of sines and cosines (cf. also end of Remark 2.15). After agreeing on a suitable form of (14), the only task left is to select "good" values for the unknown parameters $a_i$.

Substituting (14) in (1), (2), (3) gives as the left-hand sides

$$\begin{aligned}\tilde{R}(\{a\}; x) &\equiv R(\tilde{T})\\ \tilde{R}_{\mathrm{D}}(\{a\}) &\equiv R_{\mathrm{D}}(\tilde{T}(a))\\ \tilde{R}_{\mathrm{N}}(\{a\}) &\equiv R_{\mathrm{N}}(\tilde{T}(b))\end{aligned} \tag{15}$$

(The notation $\{a\}$ refers here to a $N\times 1$ column matrix consisting of the matrix elements $a_1, a_2, \cdots, a_N$ and just means that the quantities (15) depend on them.) The quantities (15) do not in general vanish with any selection of $\{a\}$ — as we would like to — but obtain non-zero values, errors or so called *residuals* (jäännös). In connection of (15), we may speak quite obviously about the *field equation residual* and about the *boundary equation residuals*.

Similarly, after substituting the approximation into the weak form, its left-hand side *cannot vanish* in general any more for any, however good, choice of the unknown parameters *with respect to arbitrary weighting functions*. We can, however, write the expression

$$\boxed{F_i \equiv \int_\Omega w_i \tilde{R}\,\mathrm{d}\Omega + w_{\mathrm{D}i}\bar{R}_{\mathrm{D}} + w_{\mathrm{N}i}\bar{R}_{\mathrm{N}}} \tag{16}$$

and demand

$$\boxed{F_i = 0} \quad i = 1,2,\cdots,N \tag{17}$$

In each case $i$ we select a "suitable" combination of $w_i$, $w_{\mathrm{D}i}$, $w_{\mathrm{N}i}$. We realize that after the integration has been performed with respect to $x$,

$$F_i = F_i(\{a\}) \tag{18}$$

i.e., the dependence on $x$ disappears, and equations (17) form an algebraic set from which the undetermined parameters can be hopefully determined. The idea behind the *residual method* or the *weighed residual method* or the *residual formulation* (jäännösmenetelmä, painotettujen jäännösten menetelmä, jäännösformulaatio) is contained in equations (16) and (17). The discrete equations (17) are called here the *system equations* (systeemiyhtälöt).

Recapitulating and generalizing:

> After substituting the approximation, the field equations and boundary conditions of a problem cannot in general be satisfied exactly for any selection of the undetermined parameters. The parameters can be selected, however, so that the equations are satisfied in some average, integral sense through satisfying the weak form with respect to some suitable weighting functions. (19)



The idea behind the residual formulation is thus seen to be very simple and very general. It is also quite apparent that the approximations and in the weightings must be based on some reasonable logic to obtain accurate and converging results.

Expression (16) could be called, say, the *weighted total residual expression* (painotettu kokonaisjäännös). (This kind of term is not in general use.) Different versions of the residual formulation are obtained according to type of weighting functions used. The most common versions are the *Galerkin method* (Galerkinin keino), the *subdomain method* or subdomain collocation (osa-aluekeino), *collocation* or point collocation (kollokaatio) and the *least squares method* (pienimmän neliön keino). These are explained in an admirably way in Crandall (1956).

In the Galerkin method the weighting functions are taken from the set of trial basis functions. What is meant by this in a general case of several unknown functions with different type of approximations is not necessarily quite obvious. The Galerkin method has been the most useful version in connection with finite elements. We describe in this section only the Galerkin method and the least squares method.

**Example 2.1.** We consider the model problem of Section 2.1.1 with the data $a = 0$, $b = L$, $k$ and $s$ constants. The problem is

$$R \equiv -k\frac{\mathrm{d}^2T}{\mathrm{d}x^2} - s = 0 \qquad 0 < x < L \tag{a}$$

$$R_{\mathrm{D}} \equiv T - \bar{T} = 0 \qquad x = 0 \tag{b}$$

$$R_{\mathrm{N}} \equiv -k\frac{\mathrm{d}T}{\mathrm{d}x} - \bar{q} = 0 \qquad x = L \tag{c}$$

The exact temperature distribution is found to be

$$T(x) = \bar{T} + \left(\frac{sL^2}{k} - \frac{\bar{q}\,L}{k}\right)\frac{x}{L} - \frac{sL^2}{2k}\left(\frac{x}{L}\right)^2 \tag{d}$$

and the corresponding *heat flux* (lämpövuo)

$$q_x(x) \equiv -k\frac{\mathrm{d}T}{\mathrm{d}x} = \bar{q} - \left(1 - \frac{x}{L}\right)sL \tag{e}$$

Guided by the exact solution, we take in this demonstration example the simple polynomial approximation of the type (14):

$$\tilde{T}(x) = \sum_{i=1}^{3} a_i\varphi_i(x) = a_1\varphi_1(x) + a_2\varphi_2(x) + a_3\varphi_3(x) = a_1\cdot 1 + a_2\cdot x + a_3\cdot x^2 \tag{f}$$

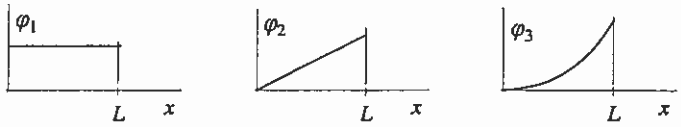



**Figure (a)**

The trial basis functions are shown in Figure (a).

From (f),

$$\frac{d\tilde{T}}{dx} = a_2 \cdot 1 + a_3 \cdot 2x = a_2 + 2a_3 x \tag{g}$$

and

$$\frac{d^2\tilde{T}}{dx^2} = a_3 \cdot 2 = 2a_3 \tag{h}$$

The residuals (15) are thus

$$\begin{aligned} \tilde{R} &= -k\frac{d^2\tilde{T}}{dx^2} - s = -k(2a_3) - s = -2ka_3 - s \\ \tilde{R}_D &= \tilde{T}\big|_{x=0} - \bar{T} = a_1 - \bar{T} \\ \tilde{R}_N &= -k\frac{d\tilde{T}}{dx}\bigg|_{x=L} - \bar{q} = -k(a_2 + 2a_3 L) - \bar{q} = -ka_2 - 2kLa_3 - \bar{q} \end{aligned} \tag{i}$$

In this simple case the field equation residual happens not to depend on $x$.

The system equations from (16) and (17) are

$$F_i \equiv \int_0^L w_i(-2ka_3 - s)dx + w_{Di}(a_1 - \bar{T}) + w_{Ni}(-ka_2 - 2kLa_3 - \bar{q}) = 0 \tag{j}$$

with $i = 1,2,3$. We take the following selections:

1. equation: $w_1 = 0, \quad w_{D1} = \varphi_1\big|_{x=0} = 1, \quad w_{N1} = 0, \Rightarrow$

$$F_1 \equiv 1\cdot(a_1 - \bar{T}) = a_1 - \bar{T} = 0 \tag{k}$$

2. equation: $w_2 = \varphi_2 = x, \quad w_{D2} = 0, \quad w_{N2} = 0, \Rightarrow$

$$F_2 \equiv \int_0^L x(-2ka_3 - s)dx = (-2ka_3 - s)\Big|_0^L \frac{x^2}{2} = -kL^2 a_3 - \frac{sL^2}{2} = 0 \tag{l}$$

3. equation: $w_3 = 0, \quad w_{D2} = 0, \quad w_{N3} = \varphi_3\big|_{x=L} = L^2, \Rightarrow$

$$F_3 \equiv L^2\cdot(-ka_2 - 2kLa_3 - \bar{q}) = -kL^2 a_2 - 2kL^3 a_3 - \bar{q}L^2 = 0 \tag{m}$$



As is seen, we have been using some kind of Galerkin method since the weighting is taken from the trial basis.

We have obtained thus the algebraic system equations

$$\begin{aligned} &a_1 - \bar{T} = 0 \\ &-kL^2 a_3 - \frac{sL^2}{2} = 0 \\ &-kL^2 a_2 - 2kL^3 a_3 - \bar{q}L^2 = 0 \end{aligned} \tag{n}$$

which are moreover linear and can be represented in matrix notation as

$$\begin{Bmatrix} F_1 \\ F_2 \\ F_3 \end{Bmatrix} \equiv \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -kL^2 \\ 0 & -kL^2 & -2kL^3 \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \\ a_3 \end{Bmatrix} + \begin{Bmatrix} -\bar{T} \\ -sL^2/2 \\ -\bar{q}L^2 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ 0 \end{Bmatrix} \tag{o}$$

This is in the spirit "something on the left hand side equals zero". A more natural form at the end is of course

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -kL^2 \\ 0 & -kL^2 & -2kL^3 \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \\ a_3 \end{Bmatrix} = \begin{Bmatrix} \bar{T} \\ sL^2/2 \\ \bar{q}L^2 \end{Bmatrix} \tag{p}$$

This is seen to be of the type (1.1.8). The solution is

$$a_1 = \bar{T}, \qquad a_2 = \frac{sL}{k} - \frac{\bar{q}}{k}, \qquad a_3 = -\frac{s}{2k} \tag{q}$$

and substitution into representation (f) gives here the exact result (d) as was to be expected.

**Integration by parts.** The starting point weak form described above for demonstration purposes is hardly ever used in practice. For problems including second order derivatives in the field equations, normally manipulaton with integration by parts is first performed.

Let us recall the weak form (6):

$$\int_a^b w\left[\frac{d}{dx}\left(-k\frac{dT}{dx}\right) - s\right]dx + w_D(T - \bar{T})\big|_{x=a} + w_N\left(-k\frac{dT}{dx} - \bar{q}\right)\Big|_{x=b} = 0 \tag{20}$$

Employing formula (B.1.1) with

$$g \triangleq w, \qquad h \triangleq -k\frac{dT}{dx} \tag{21}$$

gives

$$\int_a^b w\frac{\mathrm{d}}{\mathrm{d}x}\left(-k\frac{\mathrm{d}T}{\mathrm{d}x}\right)\mathrm{d}x = \int_a^b \frac{\mathrm{d}w}{\mathrm{d}x}k\frac{\mathrm{d}T}{\mathrm{d}x}\mathrm{d}x - \Big|_a^b wk\frac{\mathrm{d}T}{\mathrm{d}x}$$
$$= \int_a^b \frac{\mathrm{d}w}{\mathrm{d}x}k\frac{\mathrm{d}T}{\mathrm{d}x}\mathrm{d}x - \left(wk\frac{\mathrm{d}T}{\mathrm{d}x}\right)\Big|_{x=b} + \left(wk\frac{\mathrm{d}T}{\mathrm{d}x}\right)\Big|_{x=a} \qquad (22)$$

When this is substituted in (20) we obtain

$$\int_a^b \frac{\mathrm{d}w}{\mathrm{d}x}k\frac{\mathrm{d}T}{\mathrm{d}x}\mathrm{d}x - \int_a^b ws\,\mathrm{d}x$$
$$+\left[w_{\mathrm{D}}\left(T-\bar{T}\right)+wk\frac{\mathrm{d}T}{\mathrm{d}x}\right]\Big|_{x=a} + \left[w_{\mathrm{N}}\left(-k\frac{\mathrm{d}T}{\mathrm{d}x}-\bar{q}\right)-wk\frac{\mathrm{d}T}{\mathrm{d}x}\right]\Big|_{x=b} = 0 \qquad (23)$$

One may well ask: why to perform the integration by parts? First, for smooth enough trial and weighting functions the same system equations are finally arrived at from either of the weak forms. Second, however, it turns out that for the conventional $C^0$ continuous trial functions (meaning here that the function is continuous but the first derivative is not; cf. Remark B.1) used in the finite element method, weak form (23) works well but form (6) is useless (cf. Remark 2.17). This is based on the fact that form (6) contains the second derivative of $T$ but form (23) only the first. Third, employing the boundary terms produced by integration by parts, important simplifications are achieved with respect to the Neumann boundary condition as is soon seen. Thus in practice, *integration by parts is an essential mathematical tool in the finite element method.*

We now continue as follows. The weak form (23) is written first as

$$\int_\Omega \frac{\mathrm{d}w}{\mathrm{d}x}k\frac{\mathrm{d}T}{\mathrm{d}x}\mathrm{d}\Omega - \int_\Omega ws\,\mathrm{d}\Omega$$
$$+\left[w_{\mathrm{D}}\left(T-\bar{T}\right)+wk\frac{\mathrm{d}T}{\mathrm{d}x}\right]\Big|_{\Gamma_{\mathrm{D}}} + \left[w_{\mathrm{N}}\left(-k\frac{\mathrm{d}T}{\mathrm{d}x}-\bar{q}\right)-wk\frac{\mathrm{d}T}{\mathrm{d}x}\right]\Big|_{\Gamma_{\mathrm{N}}} = 0 \qquad (24)$$

so that looks a little bit more general. Especially in the finite element method it is usually very easy to satisfy the Dirichlet boundary conditions in advance by the trial solution; at least in a pointwise (nodalwise) way. If we thus restrict function $T$ in (24) to satisfy in advance the Dirichlet condition

$$T|_{\Gamma_{\mathrm{D}}} = \bar{T} \qquad (25)$$



the corresponding term in (24) disappears and there is no more need to worry about suitable values for the weighting constant $w_{\mathrm{D}}$. We further make the clever (conventional) selections

$$w|_{\Gamma_{\mathrm{D}}} = 0 \qquad (26)$$

and

$$w_{\mathrm{N}} = -w|_{\Gamma_{\mathrm{N}}} \qquad (27)$$

Selection (26) can be justified roughly as follows. The outward heat flow rate density $q = k\,\mathrm{d}T/\mathrm{d}x$ at the left-hand boundary is unknown and to approximate it we would need the value of the derivative $\mathrm{d}\tilde{T}/\mathrm{d}x$. This is somewhat awkward at least in a $C^0$ continuous finite element representation as nodal values from nodes other than those on the boundary would be needed (cf. Figure 2.10). (In fact, trying to make use of the derivative $\mathrm{d}\tilde{T}/\mathrm{d}x$ in this way is found to produce an inconsistent discrete equation.) This problem can be avoided altogether by the selection (26). Now the boundary term at $\Gamma_{\mathrm{D}}$ has disappeared from (24). Finally, selection (27) is seen to cancel the term $k\,\mathrm{d}T/\mathrm{d}x$ at $\Gamma_{\mathrm{N}}$ and we have arrived at a very simple form

$$F \equiv \boxed{\int_\Omega \frac{\mathrm{d}w}{\mathrm{d}x}k\frac{\mathrm{d}T}{\mathrm{d}x}\mathrm{d}\Omega - \int_\Omega ws\,\mathrm{d}\Omega + w\bar{q}\,|_{\Gamma_{\mathrm{N}}} = 0} \qquad (28)$$

with

$$T = \bar{T}, \quad w = 0 \quad \text{on} \quad \Gamma_{\mathrm{D}} \qquad (29)$$

We call this the *standard weak form* for heat conduction problems or for one-dimensional steady pure diffusion problems in general with new appropriate interpretations for the quantities appearing. Its generalization to two or three dimensions is going to look essentially the same. To be exact, we should remember the constraints (29) when applying the weak form although they will not always be explicitly stated.

The standard weak form has only one weighting function and no more any weighting constants. This makes life easier for the applier as fewer decisions have to be made. The crucial practical advantage of a weak form like (28) is, however, that the satisfaction of the Neumann condition is so to say implicitly buried in it. The case is opposite say when applying the standard finite difference method where the Neumann condition must be simulated directly which is often very inconvenient.

**Remark 2.4.** We have arrived at the standard weak form starting from the equation set (1), (2), (3). The signs of these equations can be of course changed in any combination. That would not change the content of the resulting weak form (20) as the weighting quantities can be taken with arbitrary signs. This also means that starting with different signs and in striving to the standard form, the right hand side of (27) might need the plus sign. □

**Remark 2.5.** The standard weak form (28) can be derived alternatively — and this is the way it is usually presented in textbooks and how we are going to proceed later — than we have done above by starting from the sole field equation weak form

$$\int_\Omega w\left[\frac{\mathrm{d}}{\mathrm{d}x}\left(-k\frac{\mathrm{d}T}{\mathrm{d}x}\right)-s\right]\mathrm{d}\Omega=0 \tag{30}$$

and integrating by parts (see (22)) to obtain first

$$\int_\Omega \frac{\mathrm{d}w}{\mathrm{d}x}k\frac{\mathrm{d}T}{\mathrm{d}x}\mathrm{d}\Omega-\int_\Omega ws\,\mathrm{d}\Omega-wk\frac{\mathrm{d}T}{\mathrm{d}x}\bigg|_{\Gamma_N}+wk\frac{\mathrm{d}T}{\mathrm{d}x}\bigg|_{\Gamma_D}=0 \tag{31}$$

Information about the boundary conditions

$$T=\bar{T} \qquad \text{on } \Gamma_D \tag{32}$$

and

$$-k\frac{\mathrm{d}T}{\mathrm{d}x}=\bar{q} \quad \text{on } \Gamma_N \tag{33}$$

must be included in the formulation. This is achieved by first demanding $T$ in (31) to satisfy (32) in advance. (It is often said that this boundary condition is now satisfied in a *strong form* (vahva muoto).) Information about the Neumann boundary condition comes by substitution of expression (33) into (31). Taking finally the selection (26) produces the standard weak form. (28). This manipulation is clearly more straightforward and easier to follow than the one, which was used earlier to arrive at (28). □

The system equations are obtained by assuming approximation (14), by defining

$$\boxed{F_i\equiv\int_\Omega \frac{\mathrm{d}w_i}{\mathrm{d}x}k\frac{\mathrm{d}\tilde{T}}{\mathrm{d}x}\mathrm{d}\Omega-\int_\Omega w_i s\,\mathrm{d}\Omega+w_i\bar{q}\,\big|_{\Gamma_N}} \tag{34}$$

and by demanding

$$\boxed{F_i=0} \quad i=1,2,\cdots,N \tag{35}$$

for suitable $w_i$ similarly as with the original weak form (see (16) and (17)).

In the *Galerkin method* we take here simply



$$\boxed{w_i=\varphi_i} \tag{36}$$

so the *system equations* are

$$\boxed{F_i\equiv\int_\Omega \frac{\mathrm{d}\varphi_i}{\mathrm{d}x}k\frac{\mathrm{d}\tilde{T}}{\mathrm{d}x}\mathrm{d}\Omega-\int_\Omega \varphi_i s\,\mathrm{d}\Omega+\varphi_i\bar{q}\,\big|_{\Gamma_N}=0} \qquad i=1,2,\cdots,N \tag{37}$$

They can be developed further into a more detailed form but this is postponed to be done later in connection with the finite element approximation.

**Example 2.2.** We repeat Example 2.1 using now the standard weak form. The problem was

$$-k\frac{\mathrm{d}^2T}{\mathrm{d}x^2}-s=0 \qquad 0<x<L \tag{a}$$

$$T=\bar{T} \qquad x=0 \tag{b}$$

$$-k\frac{\mathrm{d}T}{\mathrm{d}x}=\bar{q} \qquad x=L \tag{c}$$

The approximation used in Example 2.1 was

$$\tilde{T}(x)=a_1\varphi_1(x)+a_2\varphi_2(x)+a_3\varphi_3(x)=a_1\cdot 1+a_2\cdot x+a_3\cdot x^2 \tag{d}$$

The Dirichlet condition is clearly satisfied in advance by taking $a_1=\bar{T}$ as $\varphi_2$ and $\varphi_3$ vanish at $x=0$. There are now only two undetermined parameters: $a_2$ and $a_3$.

The system equations (37) are

$$F_i\equiv k\int_0^L\frac{\mathrm{d}\varphi_i}{\mathrm{d}x}\frac{\mathrm{d}\tilde{T}}{\mathrm{d}x}\mathrm{d}x-s\int_0^L\varphi_i\,\mathrm{d}x+\varphi_i\bar{q}\big|_{x=L}=0, \qquad i=2,3 \tag{e}$$

We have from (d)

$$\frac{\mathrm{d}\tilde{T}}{\mathrm{d}x}=a_2\cdot 1+a_3\cdot 2x=a_2+2a_3x \tag{f}$$

and

$$\frac{\mathrm{d}\varphi_2}{\mathrm{d}x}=1, \qquad \frac{\mathrm{d}\varphi_3}{\mathrm{d}x}=2x \tag{g}$$

Thus, the first system equation is

$$\begin{aligned}F_2&\equiv k\int_0^L 1\cdot(a_2+2a_3x)\mathrm{d}x-s\int_0^L x\,\mathrm{d}x+L\bar{q}\\&=k\Big|_0^L\left(a_2x+a_3x^2\right)-s\Big|_0^L\tfrac{1}{2}x^2+\bar{q}\,L\end{aligned}$$

$$= kL\cdot a_2 + kL^2\cdot a_3 - \frac{sL^2}{2} + \bar{q}\,L = 0 \tag{h}$$

and the second

$$\begin{aligned} F_3 &\equiv k\int_0^L 2x\cdot(a_2 + 2a_3 x)\,\mathrm{d}x - s\int_0^L x^2\,\mathrm{d}x + L^2\bar{q} \\ &= k\Big|_0^L\left(a_2x^2 + \tfrac{4}{3}a_3x^3\right) - s\Big|_0^L \tfrac{1}{3}x^3 + \bar{q}\,L^2 \\ &= kL^2\cdot a_2 + \frac{4kL^3}{3}\cdot a_3 - \frac{sL^3}{3} + \bar{q}L^2 = 0 \end{aligned} \tag{i}$$

Together, using matrix notation:

$$\begin{Bmatrix} F_2 \\ F_3 \end{Bmatrix} \equiv \begin{bmatrix} kL & kL^2 \\ kL^2 & 4kL^3/3 \end{bmatrix}\begin{Bmatrix} a_2 \\ a_3 \end{Bmatrix} + \begin{Bmatrix} -sL^2/2 + \bar{q}\,L \\ -sL^3/3 + \bar{q}\,L^2 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix} \tag{j}$$

The solution is

$$a_2 = \frac{sL}{k} - \frac{\bar{q}}{k}, \qquad a_3 = -\frac{s}{2k} \tag{k}$$

and the exact result is again achieved.

**Least squares method.** We consider still the model problem (1), (2), (3) and write down the expression

$$\begin{aligned} \Pi(T) = \frac{1}{2}\int_a^b \alpha\left[\frac{\mathrm{d}}{\mathrm{d}x}\left(-k\frac{\mathrm{d}T}{\mathrm{d}x}\right) - s\right]^2 \mathrm{d}x \\ + \frac{1}{2}\alpha_{\mathrm{D}}\left(T - \bar{T}\right)^2\Big|_{x=a} + \frac{1}{2}\alpha_{\mathrm{N}}\left(-k\frac{\mathrm{d}T}{\mathrm{d}x} - \bar{q}\right)^2\Big|_{x=b} \end{aligned} \tag{38}$$

or shortly

$$\boxed{\Pi(T) = \frac{1}{2}\int_\Omega \alpha R^2\,\mathrm{d}\Omega + \frac{1}{2}\alpha_{\mathrm{D}}R_{\mathrm{D}}^2 + \frac{1}{2}\alpha_{\mathrm{N}}R_{\mathrm{N}}^2} \tag{39}$$

This is called the *least squares expression* (pienimmän neliön lauseke) corresponding to the differential equation formulation. The factors $\alpha(x)$, $\alpha_{\mathrm{D}}$, $\alpha_{\mathrm{N}}$ are *given positive quantities* which can be called, say, *weight factors* (painotekijä) to distinguish them from the weighting functions or constants considered earlier.



The exact solution $T(x)$ with $R=0$, $R_{\mathrm{D}}=0$, $R_{\mathrm{N}}=0$ gives clearly the minimum value zero to the least squares expression. An approximate $\tilde{T} = \tilde{T}(\{a\};x)$ with errors $\tilde{R}$, $\tilde{R}_{\mathrm{D}}$, $\tilde{R}_{\mathrm{N}}$ gives a positive value

$$\tilde{\Pi}\left(\{a\}\right) = \frac{1}{2}\int_\Omega \alpha\tilde{R}^2\,\mathrm{d}x + \frac{1}{2}\alpha_{\mathrm{D}}\tilde{R}_{\mathrm{D}}^2 + \frac{1}{2}\alpha_{\mathrm{N}}\tilde{R}_{\mathrm{N}}^2 \tag{40}$$

(It is realized again that after the integration with respect to $x$ has been thought to be performed, $x$ disappears from the expression.) To determine the parameters $a_i$, it is natural to demand that the weighted "error expression" (40) obtains a minimum value. The necessary conditions (stationarity conditions) for this to happen are

$$\boxed{F_i \equiv \frac{\partial\tilde{\Pi}}{\partial a_i} = 0} \qquad i = 1,2,\cdots,N \tag{41}$$

These are the system equations produced by the least squares method.

As the parameters $a_i$ do not depend on $x$, we can bring the differentiations inside the integral sign to obtain in more detail (employing chain differentiation)

$$\boxed{F_i \equiv \frac{\partial\tilde{\Pi}}{\partial a_i} = \int_\Omega \alpha\tilde{R}\frac{\partial\tilde{R}}{\partial a_i}\,\mathrm{d}\Omega + \alpha_{\mathrm{D}}\tilde{R}_{\mathrm{D}}\frac{\partial\tilde{R}_{\mathrm{D}}}{\partial a_i} + \alpha_{\mathrm{N}}\tilde{R}_{\mathrm{N}}\frac{\partial\tilde{R}_{\mathrm{N}}}{\partial a_i} = 0} \tag{42}$$

for $i = 1,2,\cdots,N$. Comparison with (16) shows now that the least squares method can indeed be considered as a weighted residual formulation where the weighting functions are "selected" as

$$w_i = \alpha\frac{\partial\tilde{R}}{\partial a_i}, \qquad w_{\mathrm{D}i} = \alpha_{\mathrm{D}}\frac{\partial\tilde{R}_{\mathrm{D}}}{\partial a_i}, \qquad w_{\mathrm{N}i} = \alpha_{\mathrm{N}}\frac{\partial\tilde{R}_{\mathrm{N}}}{\partial a_i} \tag{43}$$

The least squares method can be applied in principle for any differential equation system. At first look, it seems very promising as the weighting functions are provided for the applier so to say automatically. There are, however, two rather serious drawbacks. First, to obtain good results the weight factors must have some correct proportions. So the problem of selecting good weighting functions is transferred in fact to the problem of selecting good weight factors. This is normally based on numerical experiments and can be awkward when there are many unknown functions (and corresponding differential equations). Second, if second order derivatives appear in the

differential equations, the conventional and convenient $C^0$ continuous finite element approximation does not work and complicated $C^1$ continuous finite elements have to be used. Alternatively, this difficulty can be evaded by introducing new unknown functions. In heat conduction we can for instance replace equation $\mathrm{d}(-k\,\mathrm{d}T/\mathrm{d}x)/\mathrm{d}x - s = 0$ with $\mathrm{d}q_x/\mathrm{d}x - s = 0$ and $q_x = -k\mathrm{d}T/\mathrm{d}x$, but the new variables and equations bring now forth the first drawback mentioned and in addition increase the number of discrete variables.

**Remark 2.6.** The multipliers $1/2$ in the least squares expressions above have no final effect and have been included just for aesthetic reasons so that the system equations do not contain the multipliers 2. In fact, it is the ratios between the weight factors which only matters and which puts emphasis on the different terms in the expression. It should be noted that each of the weight factors must have such a dimension that all the terms in the least squares expression have the same dimension, in other words, the least squares expression must be *dimensionally homogeneous* (dimensiohomogeeninen) to make sense. If the governing equations are first made dimensionless, which is as such a good practice, the weight factors can all be pure numbers. Then often the weight factors are missing (they are taken to have the value one) and consequently some writers then claim that the least squares method does not need any tuning parameters to work. This is however not true because the weight factors are then buried implicitly in the formulation with the selection of the characteristic measures used in making the equations dimensionless. Finally, the weight factors can of course all be negative. Then we are just trying to maximize the corresponding expression. □

**Remark 2.7.** The least squares expression is sometimes called the *least squares functional* (pienimmän neliön funktionaali) as by putting any (smooth enough) $T(x)$ in it, it produces a real number $\Pi$ (equipped maybe with some dimension) as the output (see Appendix D). However, the Euler equations corresponding to this functional are unfortunately not directly the original differential equations of the problem at hand. □

**Example 2.3.** Again the problem treated in Examples 2.1 and 2.2 is considered now using the least squares method. The approximation is also the same as before:

$$\tilde{T}(x) = a_1\varphi_1(x) + a_2\varphi_2(x) + a_3\varphi_3(x) = a_1\cdot 1 + a_2\cdot x + a_3\cdot x^2 \tag{a}$$

We take the system equations directly from (42):

$$F_i \equiv \alpha\int_0^L \tilde{R}\frac{\partial\tilde{R}}{\partial a_i}\mathrm{d}x + \alpha_{\mathrm{D}}\tilde{R}_{\mathrm{D}}\frac{\partial\tilde{R}_{\mathrm{D}}}{\partial a_i} + \alpha_{\mathrm{N}}\tilde{R}_{\mathrm{N}}\frac{\partial\tilde{R}_{\mathrm{N}}}{\partial a_i} = 0, \quad i = 1,2,3 \tag{b}$$

where we have put $\alpha$ as a constant. The residuals have been evaluated already in Example 2.1 from where we get (Formulas (i))

$$\begin{aligned}\tilde{R} &= -2ka_3 - s\\ \tilde{R}_{\mathrm{D}} &= a_1 - \bar{T}\\ \tilde{R}_{\mathrm{N}} &= -ka_2 - 2kLa_3 - \bar{q}\end{aligned} \tag{c}$$

The partial derivatives needed are thus



$$\begin{aligned}&\frac{\partial\tilde{R}}{\partial a_1} = 0, && \frac{\partial\tilde{R}_{\mathrm{D}}}{\partial a_1} = 1, && \frac{\partial\tilde{R}_{\mathrm{N}}}{\partial a_1} = 0\\ &\frac{\partial\tilde{R}}{\partial a_2} = 0, && \frac{\partial\tilde{R}_{\mathrm{D}}}{\partial a_2} = 0, && \frac{\partial\tilde{R}_{\mathrm{N}}}{\partial a_2} = -k\\ &\frac{\partial\tilde{R}}{\partial a_3} = -2k, && \frac{\partial\tilde{R}_{\mathrm{D}}}{\partial a_3} = 0, && \frac{\partial\tilde{R}_{\mathrm{N}}}{\partial a_3} = -2kL\end{aligned} \tag{d}$$

The system equations (b) are

$$\begin{aligned}F_1 &\equiv \alpha_{\mathrm{D}}(a_1 - \bar{T})\cdot 1 = \alpha_{\mathrm{D}}(a_1 - \bar{T}) = 0\\ F_2 &\equiv \alpha_{\mathrm{N}}(-ka_2 - 2kLa_3 - \bar{q})(-k) = -\alpha_{\mathrm{N}}k(-ka_2 - 2kLa_3 - \bar{q}) = 0\\ F_3 &\equiv \alpha\int_0^L(-2ka_3 - s)(-2k)\mathrm{d}x + \alpha_{\mathrm{N}}(-ka_2 - 2kLa_3 - \bar{q})(-2kL)\\ &= -2\alpha kL(-2ka_3 - s) - 2\alpha_{\mathrm{N}}kL(-ka_2 - 2kLa_3 - \bar{q}) = 0\end{aligned} \tag{e}$$

Without developing them further, it is seen quite readily that the system equations reduce to an equivalent form

$$\begin{aligned}&a_1 - \bar{T} = 0\\ &-ka_2 - 2kLa_3 - \bar{q} = 0\\ &-2ka_3 - s = 0\end{aligned} \tag{f}$$

This actually means that we just put here the constant residuals (c) equal to zero to obtain again

$$a_1 = \bar{T}, \qquad a_2 = \frac{sL}{k} - \frac{\bar{q}}{k}, \qquad a_3 = -\frac{s}{2k} \tag{g}$$

So in this extremely simple case the weight factor selection problem did not emerge.

## 2.2 ONE-DIMENSIONAL ELEMENTS

We describe here only the two most usual finite elements in one dimension: the linear and the quadratic element. The presentation will be very short. More details on the theory are given in Chapter 3.

### 2.2.1 Two-noded element

Figure 2.2 shows a *two-noded* or *linear* (one-dimensional) *element* (kaksisolmuinen tai lineaarinen elementti; tarkemmin janaelementti) in the so-called reference space. Section 2.2.3 explains how the corresponding element in the actual space (in the so-called global space) where the problem at hand is originally described can be generated. The element considered here is thus called the reference element.

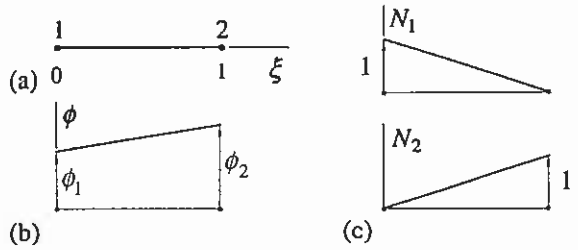



**Figure 2.2 (a)** Linear reference element. **(b)** Approximation. **(c)** Shape functions.

The shape function expressions are

$$\boxed{\begin{aligned} N_1 &= 1-\xi \\ N_2 &= \xi \end{aligned}} \tag{1}$$

and the approximation or interpolation in the element for a function $\phi$ is (see (1.1.2) with $n_n^e = 2$)

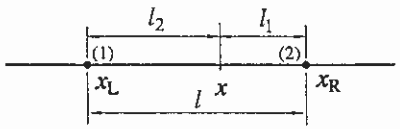
$$\phi = \sum_{i=1}^{2} N_i \phi_i = N_1\phi_1 + N_2\phi_2 = (1-\xi)\phi_1 + \xi\phi_2 \tag{2}$$

where $\phi_1$ and $\phi_2$ are the nodal values of $\phi$.

It is customary and useful to employ so-called *natural* or *intrinsic* or *non-dimensional coordinates* (luonnollinen koordinaatti) within an element. They are selected according to the nature of the basic shape of the element and their range is taken usually to be $[-1,1]$ or $[0,1]$. Here the natural coordinate is denoted $\xi$ and its range has been selected to be $[0,1]$. Nodes 1 and 2 have the coordinate values $\xi = 0$ and $\xi = 1$, respectively.

An alternative natural coordinate system representation is based on the concept of *length coordinates* $L_1$ and $L_2$ (pituuskoordinaatti) defined as (Figure 2.3)

$$L_1 = \frac{l_1}{l}, \quad L_2 = \frac{l_2}{l} \tag{3}$$




$$L_1 = \frac{l_1}{l} = \frac{x_R - x}{l} = 1-\xi$$

$$L_2 = \frac{l_2}{l} = \frac{x - x_L}{l} = \xi$$



**Figure 2.3** Some notations for an element in the global space.

It is seen that $L_1 = 1-\xi$ and $L_2 = \xi$. This formulation has some aesthetic appeal as the formulas become symmetric; for instance the shape function expressions are

$$N_1 = L_1, \qquad N_2 = L_2 \tag{4}$$

A slight drawback is that the coordinates are not independent as they must clearly satisfy the constraint equation

$$L_1 + L_2 = 1 \tag{5}$$

For triangles and tetrahedrons the equivalents of length coordinates are area coordinates $L_1$, $L_2$, $L_3$ and volume coordinates $L_1$, $L_2$, $L_3$, $L_4$.

### 2.2.2 Three-noded element

Figure 2.4 shows a *three-noded* or *quadratic* (one-dimensional) *element* (kolmisolmuinen tai kvadraattinen elementti; tarkemmin janaelementti) in the reference space.

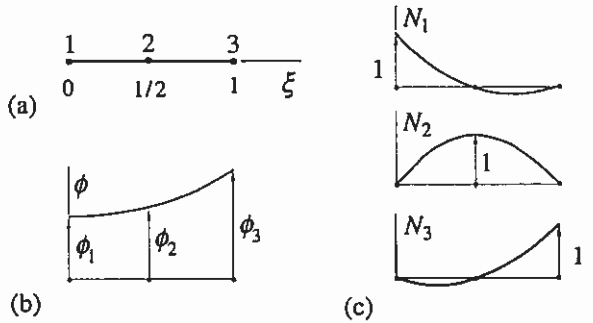



**Figure 2.4 (a)** Quadratic reference element. **(b)** Approximation. **(c)** Shape functions.

The shape function expressions are

$$\boxed{\begin{aligned} N_1 &= (1-\xi)(1-2\xi) = 1-3\xi+2\xi^2 \\ N_2 &= 4\xi(1-\xi) = 4\xi-4\xi^2 \\ N_3 &= \xi(2\xi-1) = -\xi+2\xi^2 \end{aligned}} \tag{6}$$

and the approximation is

$$\phi = \sum_{i=1}^{3} N_i \phi_i = N_1\phi_1 + N_2\phi_2 + N_3\phi_3$$
$$= \left(1 - 3\xi + 2\xi^2\right)\phi_1 + \left(4\xi - 4\xi^2\right)\phi_2 + \left(-\xi + 2\xi^2\right)\phi_3 \tag{7}$$

where $\phi_1$, $\phi_2$ and $\phi_3$ are the nodal values of $\phi$. The natural coordinate $\xi$ has as for the linear element the range [0,1]. The coordinate values for nodes 1, 2, and 3 are $\xi = 0$, $\xi = 1/2$ and $\xi = 1$, respectively.

As mentioned in Section 1.1, the shape functions have the value one at the node corresponding to its index and the value zero at the other nodes. This is seen to be true from Figures 2.2 and 2.4 and also by direct calculation from expressions (1) and (6). The shape functions described above are in fact special cases of the classical Lagrange interpolation polynomials familiar from textbooks on numerical analysis and closed form expressions for any degree polynomials are available; see for instance Conte and de Boor (1972).

### 2.2.3 Mapping

Let us consider as a demonstration case Figure 2.5 (a) showing schematically a finite element mesh of two-noded elements for the $x$-axis interval $[a,b]$.

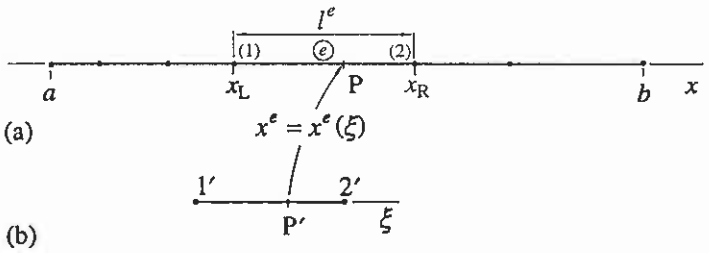



**Figure 2.5** (a) Element mesh. (b) Reference element.

It proves very useful to consider each element of a certain element type as generated from one and only typical element of that type "living" in its natural coordinate space as in Figure 2.5 (b) by a suitable mapping. We might call the element in its natural $\xi$-coordinate system as the *reference* or *natural* or *parent element* and we may speak correspondingly about the *reference or natural space*. This kind of concepts have already been touched upon shortly in the previous sections. We equip the quantities associated with the reference element temporarily with dashes. A mapping from the reference space to the *global* or *physical space*, which maps a generic point P′ to a generic point P (and the reference element to a global or physical element) would be in principle a relationship



$$x = x(\xi) \tag{8}$$

In the finite element method a very beautiful and important innovation has been to use for this purpose the mapping (we now include the element number $e$ superscript temporarily for definiteness)

$$\boxed{x^e = \sum_{i=1}^{n_n^e} N_i^e x_i^e} \tag{9}$$

where $x_i^e$ are the coordinates of the nodes of the mapped element in the global space. In other words, the same type of expression is used for the geometry description as for the functions

$$\phi^e = \sum_{i=1}^{n_n^e} N_i^e \phi_i^e \tag{10}$$

to be approximated. Elements generated this way are called for obvious reason *isoparametric elements* (isoparametrinen elementti). In the one-dimensional case considered this far the advantages of the isoparametric formulation are not evident but they will become clear later.

In connection with Figure 2.5 we have for the generic element $e$ with its left node and right node global coordinate values $x_L$ and $x_R$, the simple isoparametric mapping

$$x = N_1^e x_1^e + N_2^e x_2^e = N_1^e x_L + N_2^e x_R = (1-\xi)x_L + \xi x_R$$
$$= x_L + \xi(x_R - x_L) = x_L + \xi l^e \tag{11}$$

where $l^e$ is the length of the element:

$$l^e = x_R - x_L \tag{12}$$

We obtain from (11) we easily the inverse mapping $\xi = \xi(x)$:

$$\xi = \frac{x - x_L}{l^e} \tag{13}$$

We have presented the shape functions in Section 1.1 as functions of the global coordinates. However, in practice *the shape functions are nearly always expressed as functions of the natural coordinates* as we have done above in connection with the linear and quadratic element. But the existence of the inverse mappings such as (13) means that the shape functions can be finally

thought to be functions of the global coordinates. We will later see that explicit inverse mappings are fortunately not in fact needed in practical calculations.

Figure 2.6 repeats the presentation of Figure 2.5 now for three-noded elements. The isoparametric mapping for a generic element $e$ with the notations of the figure is

$$\begin{aligned} x &= N_1^e x_1^e + N_2^e x_2^e + N_3^e x_3^e = N_1^e x_L + N_2^e x_M + N_3^e x_R \\ &= \left(1-3\xi+2\xi^2\right)x_L + \left(4\xi-4\xi^2\right)x_M + \left(-\xi+2\xi^2\right)x_R \end{aligned} \tag{14}$$

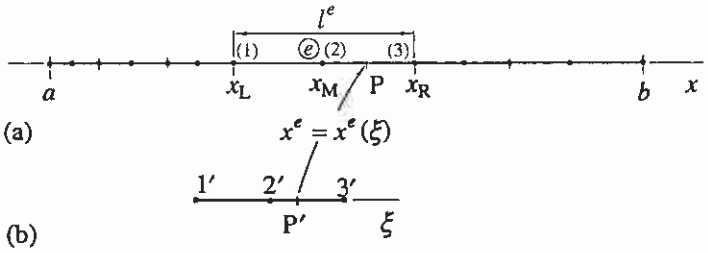



**Figure 2.6 (a)** Element mesh. **(b)** Reference element.

The inverse mapping $\xi = \xi(x)$ becomes involved in the general case as a second degree equation should be solved. If the middle node is put at the exact midpoint of the element, that is,

$$x_M = \frac{x_L + x_R}{2} \tag{15}$$

as is usual, there is obtained simply

$$x = x_L + (x_R - x_L)\xi = x_L + \xi l^e \tag{16}$$

The inverse mapping is then again (13).

One-dimensional elements can exist also in two- and three-dimensional spaces. For instance, the isoparametric mapping

$$\begin{aligned} x &= \left(1-3\xi+2\xi^2\right)x_L + \left(4\xi-4\xi^2\right)x_M + \left(-\xi+2\xi^2\right)x_R \\ y &= \left(1-3\xi+2\xi^2\right)y_L + \left(4\xi-4\xi^2\right)y_M + \left(-\xi+2\xi^2\right)y_R \end{aligned} \tag{17}$$

produces a curved three-noded element in the $x,y$-plane (Figure 2.7). In addition to being useful in the finite element method, this element is also very popular in the applications of the *boundary element method* (reunaelementtimenetelmä); see for instance Becker (1992), Beer and Watson (1992), where the accurate representation of the domain boundary is of importance. The boundary element method is not considered in this text. Approximation (7) is still valid and can be used for representing say the temperature and heat flow rate density at the boundary.



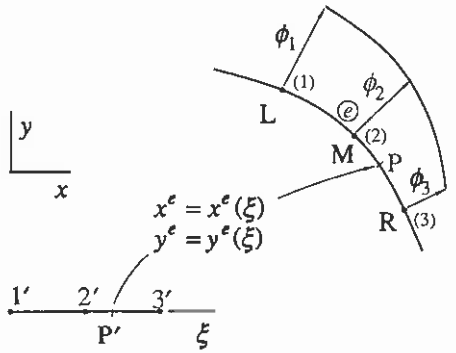



**Figure 2.7** Curved three-noded isoparametric one-dimensional element.

**Remark 2.8.** For boundary condition implementation, in addition to the definite order numbering of the nodes of a reference element, the separate boundary parts or "element sides" of the element must be numbered in an agreed way. For the one-dimensional element the sides are just the two end points and they are numbered here 1 and 2 with $\xi = 0$ and $\xi = 1$, respectively. □

## 2.3 FINITE ELEMENT SOLUTION

### 2.3.1 Discretization

We consider the model problem given by equations (2.1.1), (2.1.2) and (2.1.3) with the simplifications stated in Example 2.1. Thus repeating, the problem is

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}x}\left(-k\frac{\mathrm{d}T}{\mathrm{d}x}\right)-s=0 \quad && \text{in } \Omega=]0,L[ \\ &T=\bar{T} && \text{on } \Gamma_{\mathrm{D}}=\{0\} \\ &-k\frac{\mathrm{d}T}{\mathrm{d}x}=\bar{q} && \text{on } \Gamma_{\mathrm{N}}=\{L\} \end{aligned} \tag{1}$$

with $k$ and $s$ constants.

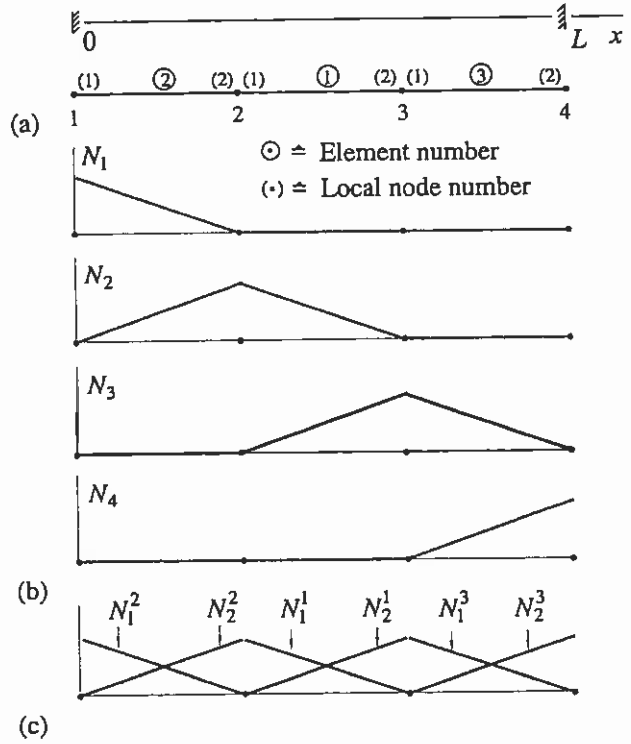



**Figure 2.8 (a)** Domain and the element mesh; $n_e=3$, $n_n=4$. **(b)** Global shape functions. **(c)** Local shape functions.



We employ a uniform example mesh of three two-noded elements and four global nodes shown in Figure 2.8 (a). The numbering of the elements has been done on purpose in a slightly odd manner for demonstration purposes for the assembly process. The figure also shows the symbols to be used in this text for element and local node numbering.

We employ the Galerkin method. The general system equations based on the standard weak form (2.1.28) have been derived in Section 2.1.2. They are (see (2.1.37))

$$F_i \equiv \int_\Omega \frac{\mathrm{d}\varphi_i}{\mathrm{d}x} k \frac{\mathrm{d}\tilde{T}}{\mathrm{d}x}\mathrm{d}\Omega - \int_\Omega \varphi_i s\,\mathrm{d}\Omega + \varphi_i \bar{q}\big|_{\Gamma_{\mathrm{N}}} = 0 \tag{2}$$

We remember that the quantities $\varphi$ are the trial basis functions, that is, the approximation is of the form

$$\tilde{T}(x)=\sum_{j=1}^{N}\varphi_j(x)a_j \tag{3}$$

(For later purposes, we have changed the summation index from $i$ to $j$.) In our present case with finite elements, the equivalent of (3) is

$$\boxed{\tilde{T}(x)=\sum_{j=1}^{n_n}N_j(x)T_j} \tag{4}$$

The trial basis functions $N_j$ are the global shape functions. They are shown in our example case in Figure 2.8 (b). The undetermined parameters are the nodal values, nodal temperatures $T_j$, to be determined. Taking these notations into account in (2) gives the *finite element system equations*

$$\boxed{F_i \equiv \int_\Omega \frac{\mathrm{d}N_i}{\mathrm{d}x} k \frac{\mathrm{d}\tilde{T}}{\mathrm{d}x}\mathrm{d}\Omega - \int_\Omega N_i s\,\mathrm{d}\Omega + N_i \bar{q}\big|_{\Gamma_{\mathrm{N}}} = 0} \quad i=1,2,\cdots,n_n \tag{5}$$

We now develop these equations further. From (4),

$$\frac{\mathrm{d}\tilde{T}}{\mathrm{d}x}=\sum_{j=1}^{n_n}\frac{\mathrm{d}N_j}{\mathrm{d}x}T_j \tag{6}$$

and

$$F_i = \int_\Omega \frac{dN_i}{dx} k \left( \sum_{j=1}^{n_n} \frac{dN_j}{dx} T_j \right) d\Omega - \int_\Omega N_i s \, d\Omega + N_i \bar{q} \big|_{\Gamma_N} \tag{7}$$

Still further development is possible. From the properties of the definite integral we have, say, for two functions $f_1(x)$ and $f_2(x)$ and for two constants $c_1$ and $c_2$,

$$\int_\Omega \left[ f_1(x) c_1 + f_2(x) c_2 \right] d\Omega = \left[ \int_\Omega f_1(x) d\Omega \right] c_1 + \left[ \int_\Omega f_2(x) d\Omega \right] c_2 \tag{8}$$

or more generally

$$\int_\Omega \left[ \sum_j f_j(x) c_j \right] d\Omega = \sum_j \left[ \int_\Omega f_j(x) d\Omega \right] c_j \tag{9}$$

This means just that summation inside an integral can be taken outside the integral and that constants can also be taken outside the integral. Taking this into account in the first integral of (7) gives

$$\begin{aligned} \int_\Omega \frac{dN_i}{dx} k \left( \sum_{j=1}^{n_n} \frac{dN_j}{dx} T_j \right) d\Omega &= \int_\Omega \left( \sum_{j=1}^{n_n} \frac{dN_i}{dx} k \frac{dN_j}{dx} T_j \right) d\Omega \\ &= \sum_{j=1}^{n_n} \left( \int_\Omega \frac{dN_i}{dx} k \frac{dN_j}{dx} d\Omega \right) T_j \end{aligned} \tag{10}$$

The final form of the system equations is thus

$$\boxed{F_i \equiv \sum_{j=1}^{n_n} \left( \int_\Omega \frac{dN_i}{dx} k \frac{dN_j}{dx} d\Omega \right) T_j - \int_\Omega N_i s \, d\Omega + N_i \bar{q} \big|_{\Gamma_N} = 0} \tag{11}$$

for $i = 1, 2, \cdots, n_n$. This is a linear system of equations with respect to the unknowns $T_j$, which can be expressed also as

$$\sum_{j=1}^{n_n} K_{ij} T_j - b_i = 0, \quad i = 1, 2, \cdots, n_n \tag{12}$$

or by employing matrix notation as



$$\underset{n_n \times n_n}{[K]} \underset{n_n \times 1}{\{a\}} - \underset{n_n \times 1}{\{b\}} = \underset{n_n \times 1}{\{0\}} \tag{13}$$

or in a cleaner way

$$\boxed{[K]\{a\} = \{b\}} \tag{14}$$

We call $[K]$ and $\{b\}$ *system matrices* (systeemimatriisi). The matrix elements are

$$\begin{aligned} K_{ij} &= \int_\Omega \frac{dN_i}{dx} k \frac{dN_j}{dx} d\Omega, \qquad i = 1, 2, \cdots, n_n, \quad j = 1, 2, \cdots, n_n \\ b_i &= \int_\Omega N_i s \, d\Omega - N_i \bar{q} \big|_{\Gamma_N}, \qquad i = 1, 2, \cdots, n_n \end{aligned} \tag{15}$$

As seen, we have denoted the column matrix consisting of the nodal temperatures with the general symbol $\{a\}$. Changing the indices $i$ and $j$ in (15) gives the result $K_{ji} = K_{ij}$, that is, the system *coefficient matrix is symmetric* which is advantageous from the computational point of view.

With our example mesh we obtain in principle the system equations

$$\begin{bmatrix} K_{11} & K_{12} & K_{13} & K_{14} \\ K_{21} & K_{22} & K_{23} & K_{24} \\ K_{31} & K_{32} & K_{33} & K_{34} \\ K_{41} & K_{42} & K_{43} & K_{44} \end{bmatrix} \begin{Bmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \end{Bmatrix} = \begin{Bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{Bmatrix} \tag{16}$$

and we have detailed recipes (15) for evaluating the matrix elements. After this has been done, we can solve the set for the nodal temperatures.

**Remark 2.9.** We have cheated a little bit in the derivations above. In fact when employing the basic formula (11) we should have the approximation for the temperature to satisfy in advance the Dirichlet boundary condition and the weighting functions should be zero at the Dirichlet boundary (see formulas (2.1.29)). But we can correct for this at this phase just by discarding the first equation, which is the only one (see Figure 2.8 (b)) produced by a weighting (shape function $N_1$) differing from zero at the Dirichlet boundary $x = 0$. Further, putting $T_1 = \bar{T}$ in (4) makes the approximation satisfy the boundary condition. Thus, taking the above into account in (16) gives a new system

$$\begin{bmatrix} K_{21} & K_{22} & K_{23} & K_{24} \\ K_{31} & K_{32} & K_{33} & K_{34} \\ K_{41} & K_{42} & K_{43} & K_{44} \end{bmatrix} \begin{Bmatrix} \bar{T} \\ T_2 \\ T_3 \\ T_4 \end{Bmatrix} = \begin{Bmatrix} b_2 \\ b_3 \\ b_4 \end{Bmatrix} \tag{17}$$

or with rearrangement

$$\begin{bmatrix} K_{22} & K_{23} & K_{24} \\ K_{32} & K_{33} & K_{34} \\ K_{42} & K_{43} & K_{44} \end{bmatrix} \begin{Bmatrix} T_2 \\ T_3 \\ T_4 \end{Bmatrix} = \begin{Bmatrix} b_2 - K_{21}\bar{T} \\ b_3 - K_{31}\bar{T} \\ b_4 - K_{41}\bar{T} \end{Bmatrix} \tag{18}$$

which can be solved for the remaining nodal temperatures. Actually, the first equation discarded here contains important information, which can made use of in post-processing (see Section 2.4.2). □

### 2.3.2 Assembly process

The previous section gave the formulas to build the system equations from the global shape functions. (In fact, the formulas are still in a form which could be used in connection with classical trial basis functions with approriate interpretations for the variables.) As was mentioned in Section 1.1, the global shape functions are not employed in actual calculations and we now make use of the local shape functions. The starting point is simple. First,

$$\boxed{\text{When } x \in \Omega^e,\ \tilde{T}(x) = T^e(x)} \tag{19}$$

that is, when we consider a certain element $e$, the global approximation $\tilde{T}$ is obtained from the local one $T^e$.

Second, weak forms contain definite integrals over $\Omega$. Due to the properties of the definite integral,

$$\boxed{\int_\Omega (\cdot)\,\mathrm{d}\Omega = \sum_{e=1}^{n_e} \int_{\Omega^e} (\cdot)\,\mathrm{d}\Omega} \tag{20}$$

This simple fact is fundamental in the finite element method: *the integral over $\Omega$ can be evaluated by evaluating separate integrals over the subdomains $\Omega^e$ of $\Omega$ and by summing the contributions.* (A similar statement can be given with respect to $\Gamma$.)

Of course the subdomains must cover the whole domain without gaps and overlaps. Mathematically this is indicated with

$$\bar{\Omega} = \cup_e \bar{\Omega}^e, \qquad \cap_e \Omega^e = \varnothing \tag{21}$$

where the meaning of the set notations are explained in the NOMENCLATURE section. In the finite element method the subdomains are called finite elements. To be more exact, when we speak about an element, we sometimes consider it



depending on the context as a general concept having for instance an element number and sometimes as a domain (set of points) as in formula (20). It should be mentioned that with complicated geometries the boundary of the domain cannot usually be quite exactly followed by finite elements and the first of (21) is then not necessarily strictly valid.

**Remark 2.10.** Considering notations (21) in detail, we realize that we have defined the element domains in this connection to consist of *open* domains; the element boundaries (here the endpoints of the elements) are not included. This means in fact, that in evaluating the right-hand side of (20), the integrals over the element boundaries (or so-called element interfaces inside the mesh) are neglected. However, as the "measures" of the element boundaries are zero compared to the domains, the left- and right-hand sides in (20) are still equal. Similarly, we do not need the value of quantity $(\cdot)$ on the element boundaries in (20). This fact is reflected also in (19), where the element boundaries are missing. Further, with possible jumps in the value of $(\cdot)$ over interfaces, we have no need to worry about the value of $(\cdot)$ on element interfaces; it is enough to deal with the limiting values from both sides. □

The best starting point is equation (5), which is repeated here:

$$F_i \equiv \int_\Omega \frac{\mathrm{d}N_i}{\mathrm{d}x} k \frac{\mathrm{d}\tilde{T}}{\mathrm{d}x}\,\mathrm{d}\Omega - \int_\Omega N_i s\,\mathrm{d}\Omega + N_i \bar{q}\big|_{\Gamma_N} = 0 \tag{22}$$

According to (19) and (20) we can now write

$$\boxed{F_i = \sum_{e=1}^{n_e} F_i^e} \tag{23}$$

where

$$F_i^e = \int_{\Omega^e} \frac{\mathrm{d}N_i}{\mathrm{d}x} k \frac{\mathrm{d}T^e}{\mathrm{d}x}\,\mathrm{d}\Omega - \int_{\Omega^e} N_i s\,\mathrm{d}\Omega + N_i \bar{q}\big|_{\Gamma_N^e} \tag{24}$$

This means that the *left-hand side $F_i$ of a system equation is obtained by summation from the element contributions (elementtiosuus) $F_i^e$ over the number of elements.*

The notation $\Gamma_N^e$ refers to the Neumann boundary part of element $e$. In our example case this exists only for element 3 and on its right-hand end.

In a generic element $e$ the equivalents of (4) and (6) are

$$\boxed{T^e(x)=\sum_{j=1}^{n_n^e} N_j^e(x)T_j^e} \tag{25}$$

and

$$\frac{dT^e}{dx}=\sum_{j=1}^{n_n^e}\frac{dN_j^e}{dx}T_j^e \tag{26}$$

Substitution of the last expression in the integral in (24) gives

$$\int_{\Omega^e}\frac{dN_i}{dx}k\frac{dT^e}{dx}d\Omega=\int_{\Omega^e}\frac{dN_i}{dx}k\left(\sum_{j=1}^{n_n^e}\frac{dN_j^e}{dx}T_j^e\right)d\Omega=$$

$$\int_{\Omega^e}\sum_{j=1}^{n_n^e}\left(\frac{dN_i}{dx}k\frac{dN_j^e}{dx}T_j^e\right)d\Omega=\sum_{j=1}^{n_n^e}\left(\int_{\Omega^e}\frac{dN_i}{dx}k\frac{dN_j^e}{dx}d\Omega\right)T_j^e \tag{27}$$

Similar manipulative steps as in obtaining result (10) have been used. The element contribution is thus

$$F_i^e=\sum_{j=1}^{n_n^e}\left(\int_{\Omega^e}\frac{dN_i}{dx}k\frac{dN_j^e}{dx}d\Omega\right)T_j^e-\int_{\Omega^e}N_i s\,d\Omega+N_i\bar{q}\big|_{\Gamma_N^e} \tag{28}$$

This formula could be used for evaluating the contributions. However, a look at the global and local shape functions in Figure 2.8 shows that this is not practical. As an example, let us consider the generation of the third system equation $F_3=0$. It is obtained by taking $N_3$ as the weighting function. This is non-zero only in elements 1 and 3 so non-zero contributions $F_3^1$ and $F_3^3$ come only from elements 1 and 3 and they are in fact obtained by the local weightings $N_2^1$ and $N_1^3$, respectively. It is realized immediately that however dense the mesh, only few of the elements (for two-noded elements only two at the inner nodes and one at the boundary nodes in one dimension) in the sum over the elements can give non-zero contributions.

Repeating the considerations discussed above with a slightly different wording we can say the following. Each $N_i$ used as the weighting function to generate the $i$:th *system equation* is in a generic element $e$ either zero (if node $i$ is "far from element $e$") or consists of one of the element shape functions — here $N_1^e$



or $N_2^e$ — (if element $e$ is connected to node $i$). *Thus all possible non-zero contributions from a generic element e are*

$$F_i^e=\sum_{j=1}^{2}\left(\int_{\Omega^e}\frac{dN_i^e}{dx}k\frac{dN_j^e}{dx}d\Omega\right)T_j^e-\int_{\Omega^e}N_i^e s\,d\Omega+N_i^e\bar{q}\big|_{\Gamma_N^e},\quad i=1,2 \tag{29}$$

(We have written this for the example case; in the general case put $2\rightarrow n_n^e$.) When $i=1$, $F_1^e$ gives contributions to a certain system equation and similarly when $i=2$, $F_2^e$ gives contributions to another system equation; to which equations is to be described in detail shortly. It is important to realize, that the $F_i^e$-quantity defined above differs from the one defined by formula (24), as here the element contributions appear in a *condensed form* and they cannot be directly summed into the system equations but must be placed in right places in them. We do not want, however, to introduce new notation as the difference can be understood from the context.

The *system equation assembly* (systeemiyhtälöiden kokoaminen) can now be described as follows. If the system equations are written as

$$\boxed{F_i=0}\quad i=1,2,\cdots,n_n \tag{30}$$

the left-hand sides are obtained by the summation formula

$$\boxed{F_i=\sum_{e=1}^{n_e}F_r^e} \tag{31}$$

*where element e gives a contribution (or can give at most a non-zero contribution) to the quantity $F_i$ if the element has the global node i. Then the local node number r must be given the value corresponding to i. The element contribution becomes $F_r^e$.*

Let us consider the example case shown in Figure 2.8 to illustrate the formula. The correspondence between the global and local node numbers as found from the figure is shown in Table 2.1.

**Table 2.1** Element nodal data

| Element number $e$ | Local node number $r$ | Global node number $i$ |
|---|---|---|
| ① | (1) (2) | 2 3 |
| ② | (1) (2) | 1 2 |
| ③ | (1) (2) | 3 4 |

The same data can be given in a more concise form using Table 2.2. The meaning of the contents should be self-explanatory. This latter type of presentation will be employed later in this text.

**Table 2.2** Element nodal data

| | (1) | (2) |
|---|---|---|
| ① | 2 | 3 |
| ② | 1 | 2 |
| ③ | 3 | 4 |

By giving the consecutive values $i=1,2,3,4$ and by going for each $i$ through the elements in Table 2.1, we obtain the system equations

$$\begin{aligned} F_1 &= F_1^2 = 0 \\ F_2 &= F_1^1 + F_2^2 = 0 \\ F_3 &= F_2^1 + F_1^3 = 0 \\ F_4 &= F_2^3 = 0 \end{aligned} \tag{32}$$

The assembly rule (31) and its application (32) should now be quite obvious on the basis of how a typical system equation is obtained: through weighting by a specific global shape function which is composed in each element of a specific local shape function (or more often in general vanishes). Finally also the local nodal values must be expressed in the global nodal values. This knowledge is also contained in Table 2.1 or 2.2 and we immediately obtain

$$\begin{aligned} T_1^1 &= T_2, \quad & T_2^1 &= T_3 \\ T_1^2 &= T_1, \quad & T_2^2 &= T_2 \\ T_1^3 &= T_3, \quad & T_2^3 &= T_4 \end{aligned} \tag{33}$$

This completes the idea of the assembly.



As the original problem is here linear (and in nonlinear cases the solution proceeds any way in practice iteratively with linear subproblems) in the unknown function $T$, the system equations and the element contributions are similarly also linear in the nodal values. This means that we arrive at still more detailed prescriptions for the assembly. The system equations (30) are here (see (12))

$$F_i \equiv \sum_{j=1}^{n_n} K_{ij} T_j - b_i = 0, \quad i = 1,2,\cdots,n_n \tag{34}$$

or if matrix notation is used

$$\underset{n_n\times 1}{\{F\}} \equiv \underset{n_n\times n_n}{[K]}\ \underset{n_n\times 1}{\{a\}} - \underset{n_n\times 1}{\{b\}} = \underset{n_n\times 1}{\{0\}}. \tag{35}$$

The element contributions (29) are similarly of the form

$$F_i^e \equiv \sum_{j=1}^{n_n^e} K_{ij}^e T_j^e - b_i^e, \quad i = 1,2,\cdots,n_n^e \tag{36}$$

or using matrix notation

$$\underset{n_n^e\times 1}{\{F\}^e} = \underset{n_n^e\times n_n^e}{[K]^e}\ \underset{n_n^e\times 1}{\{a\}^e} - \underset{n_n^e\times 1}{\{b\}^e} \tag{37}$$

We call $[K]^e$ and $\{b\}^e$ as the *element matrices* (elementtimatriisi). The matrix elements are

$$\begin{aligned} K_{ij}^e &= \int_{\Omega^e} \frac{\mathrm{d}N_i^e}{\mathrm{d}x} k \frac{\mathrm{d}N_j^e}{\mathrm{d}x} \mathrm{d}\Omega, \qquad i = 1,2,\cdots,n_n^e \quad j = 1,2,\cdots,n_n^e \\ b_i^e &= \int_{\Omega^e} N_i^e s\, \mathrm{d}\Omega - N_i^e \overline{q}\Big|_{\Gamma_N^e}, \quad i = 1,2,\cdots,n_n^e \end{aligned} \tag{38}$$

**Remark 2.11.** It is seen that the expressions for the element matrices (38) can in fact be written down directly from the corresponding expressions (15) for the system matrices by just adding the superscript $e$ for the shape functions and by changing the ranges of $i$ and $j$. This is a general result which is made use of later to shorten the derivations. One way to justify and remember this is just to consider for a while the whole domain to be one element and to apply expressions (15). □

Some further consideration on the formulas given above shows that the final detailed assembly process can be given now as

$$K_{ij} = \sum_{e=1}^{n_e} K_{rs}^e, \quad b_i = \sum_{e=1}^{n_e} b_r^e \tag{39}$$

The formulas describe how the matrix elements of the system matrices are obtained by summation from the matrix elements of the element matrices. Element $e$ gives a contribution to term $K_{ij}$ at most, if the element has global nodes $i$ and $j$. The local node numbers $r$ and $s$ must then be given the values corresponding to $i$ and $j$. Term $b_i$ obtains a contribution at most, if the element has global node $i$. The local node number $r$ must then be given the value corresponding to $i$.

The finite element programs operate usually so that the program proceeds in the assembly in an element by element fashion. When a certain element $e$ is reached, the program first adds the term $K_{11}^e$ on the right place, then term $K_{12}^e$ and so on until the last term $b_2^e$ has been added and then proceeds to the next element. This procedure differs in fact in spirit from formulas (39), because in them $i$ and $j$ are considered first as fixed, as $e$ goes through the values $1,2,\cdots,n_e$. The final outcome does of course not depend on the order of performing the assembly as the operation is simply that of addition. Formulas (39) can be applied directly, for instance, when we want produce one typical system equation for closer examination. In the assembly in the element by element fashion the formulas are applied so that we always have a certain $K_{rs}^e$ or $b_r^e$ and we search for global node numbers $i$ and $j$ corresponding to the local numbers $r$ and $s$, which then determine the positions of the terms in the system matrices.

Figure 2.9 indicates visually as an example the allocation of the terms $K_{21}^1$ and $b_1^1$ of element 1 in the system matrices. The element mesh is the same as in Figure 2.8. The crosses indicate that the corresponding terms are (usually) non-zero.



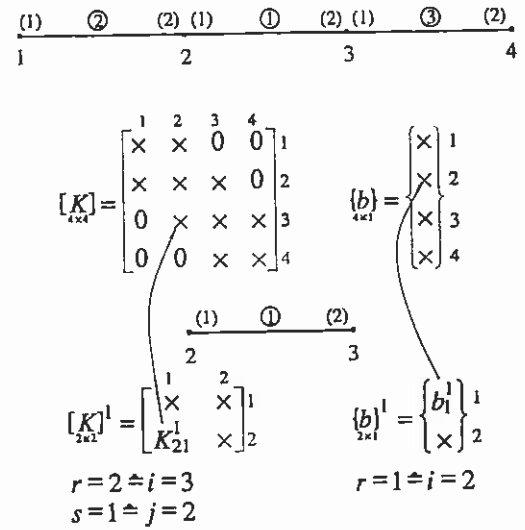



**Figure 2.9** Some example details of the assembly process.

We continue by first deriving the detailed element matrices for our example problem. The expressions for two-noded elements are (see (38))

$$K_{ij}^e = \int_{\Omega^e} \frac{\mathrm{d}N_i^e}{\mathrm{d}x} k \frac{\mathrm{d}N_j^e}{\mathrm{d}x} \mathrm{d}\Omega, \qquad i=1,2 \quad j=1,2$$
$$b_i^e = \int_{\Omega^e} N_i^e s \,\mathrm{d}x - N_i^e \bar{q}\Big|_{\Gamma_N^e}, \qquad i=1,2 \tag{40}$$

The element shape functions are given in the natural coordinate $\xi$:

$$N_1 = 1-\xi, \qquad N_2 = \xi \tag{41}$$

Even the given functions such as $k(x)$ or $s(x)$ are usually expressed for convenience by a finite element approximation so that finally we have the representations $k(\xi)$ and $s(\xi)$. It is thus quite natural to try to evaluate the element contributions in the reference element space. The transformation formulas needed in this connection are explained in detail in Appendix E. We have for a generic function $f$ here

$$\int_{\Omega^e} f \,\mathrm{d}\Omega = \int_{\Omega^{\prime e}} f \frac{\mathrm{d}x}{\mathrm{d}\xi} \mathrm{d}\xi = \int_0^1 f \frac{\mathrm{d}x}{\mathrm{d}\xi} \mathrm{d}\xi \tag{42}$$

and

$$\frac{dN_i}{dx}=\frac{dN_i}{d\xi}\frac{d\xi}{dx} \tag{43}$$

For the two-noded element (see (2.2.11) and (2.2.13))

$$x=x_L+\xi l^e, \qquad \xi=\frac{x-x_L}{l^e} \tag{44}$$

so

$$\frac{dx}{d\xi}=l^e, \qquad \frac{d\xi}{dx}=\frac{1}{l^e} \tag{45}$$

Thus

$$\int_{\Omega^e} f\,d\Omega=l^e\int_0^1 f\,d\xi \tag{46}$$

and

$$\frac{dN_i^e}{dx}=\frac{1}{l^e}\frac{dN_i^e}{d\xi} \tag{47}$$

where

$$\frac{dN_1^e}{d\xi}=-1, \qquad \frac{dN_2^e}{d\xi}=1 \tag{48}$$

Thus first ($k$ is assumed constant)

$$\begin{aligned}K_{ij}^e&=k\int_{\Omega^e}\frac{dN_i^e}{dx}\frac{dN_j^e}{dx}d\Omega=kl^e\int_0^1\frac{1}{l^e}\frac{dN_i^e}{d\xi}\frac{1}{l^e}\frac{dN_j^e}{d\xi}d\xi\\&=\frac{k}{l^e}\int_0^1\frac{dN_i^e}{d\xi}\frac{dN_j^e}{d\xi}d\xi\end{aligned} \tag{49}$$

and further



$$\begin{aligned}K_{11}^e&=\frac{k}{l^e}\int_0^1(-1)(-1)\,d\xi=\frac{k}{l^e}\\K_{12}^e&=\frac{k}{l^e}\int_0^1(-1)\cdot 1\,d\xi=-\frac{k}{l^e}\\K_{21}^e&=\frac{k}{l^e}\int_0^1 1\cdot(-1)\,d\xi=-\frac{k}{l^e}\\K_{22}^e&=\frac{k}{l^e}\int_0^1 1\cdot 1\,d\xi=\frac{k}{l^e}\end{aligned} \tag{50}$$

Similarly ($s$ is assumed constant)

$$b_i^e=s\int_{\Omega^e}N_i^e\,d\Omega-N_i^e\bar{q}\Big|_{\Gamma_N}=sl^e\int_0^1 N_i^e\,d\xi-\bar{q}\delta_{e3}\delta_{i2} \tag{51}$$

The term from the Neumann boundary appears only in the third element and due to the second shape function $N_2^3$ as the weighting ($N_2^3(1)=1$). This is taken here into account in ad hoc manner using the Kronecker deltas. Thus the contributions from the heat source without the boundary term are

$$\begin{aligned}\left(b_1^e\right)_s&=sl^e\int_0^1 N_1^e\,d\xi=sl^e\int_0^1(1-\xi)\,d\xi=sl^e\Big|_0^1\left(\xi-\tfrac{1}{2}\xi^2\right)=\frac{sl^e}{2}\\\left(b_2^e\right)_s&=sl^e\int_0^1 N_2^e\,d\xi=sl^e\int_0^1\xi\,d\xi=sl^e\Big|_0^1\tfrac{1}{2}\xi^2=\frac{sl^e}{2}\end{aligned} \tag{52}$$

These expressions could have been deduced directly without actual calculations by looking at the shape function graphs. The element matrices are thus

$$\begin{aligned}\underset{2\times 2}{[K]^e}&=\frac{k}{l^e}\begin{bmatrix}1&-1\\-1&1\end{bmatrix}\\\underset{2\times 1}{\{b\}^e}&=\frac{sl^e}{2}\begin{Bmatrix}1\\1\end{Bmatrix}-\begin{Bmatrix}0\\\bar{q}\delta_{e3}\delta_{i2}\end{Bmatrix}\end{aligned} \tag{53}$$

In the example case $l^1=l^2=l^3=L/3$ and

$$\begin{aligned}[K]^1=[K]^2=[K]^3&=\frac{3k}{L}\begin{bmatrix}1&-1\\-1&1\end{bmatrix}\\\{b\}^1=\{b\}^2=\frac{sL}{6}\begin{Bmatrix}1\\1\end{Bmatrix},\quad\{b\}^3&=\frac{sL}{6}\begin{Bmatrix}1\\1\end{Bmatrix}-\begin{Bmatrix}0\\\bar{q}\end{Bmatrix}\end{aligned} \tag{54}$$

Using the assembly rules (39) gives the system matrices

$$[K]_{4\times4}=\begin{matrix}1&2&3&4\end{matrix}\begin{bmatrix} K_{11}^2 & K_{12}^2 & 0 & 0 \\ K_{21}^2 & K_{11}^1+K_{22}^2 & K_{12}^1 & 0 \\ 0 & K_{21}^1 & K_{22}^1+K_{11}^3 & K_{12}^3 \\ 0 & 0 & K_{21}^3 & K_{22}^3 \end{bmatrix}\begin{matrix}1\\2\\3\\4\end{matrix}$$

$$\{b\}_{4\times1}=\begin{Bmatrix} b_1^2 \\ b_1^1+b_2^2 \\ b_2^1+b_1^3 \\ b_2^3 \end{Bmatrix}\begin{matrix}1\\2\\3\\4\end{matrix} \tag{55}$$

and collecting the terms from (54) produces the preliminary system equations

$$\frac{3k}{L}\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}\begin{Bmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \end{Bmatrix}=\frac{sL}{6}\begin{Bmatrix} 1 \\ 2 \\ 2 \\ 1 \end{Bmatrix}-\begin{Bmatrix} 0 \\ 0 \\ 0 \\ \bar{q} \end{Bmatrix} \tag{56}$$

At this phase we continue according to Remark 2.9 and put $T_1=\bar{T}$ to obtain the final system equations (after division by $3k/L$)

$$\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}\begin{Bmatrix} T_2 \\ T_3 \\ T_4 \end{Bmatrix}=\begin{Bmatrix} \bar{T} \\ 0 \\ 0 \end{Bmatrix}+\frac{sL^2}{18k}\begin{Bmatrix} 2 \\ 2 \\ 1 \end{Bmatrix}-\frac{\bar{q}L}{3k}\begin{Bmatrix} 0 \\ 0 \\ 1 \end{Bmatrix} \tag{57}$$

The right-hand side consists of three separate type forcing terms generating the temperature field.

The solution is

$$\begin{Bmatrix} T_2 \\ T_3 \\ T_4 \end{Bmatrix}=\bar{T}\begin{Bmatrix} 1 \\ 1 \\ 1 \end{Bmatrix}+\frac{sL^2}{18k}\begin{Bmatrix} 5 \\ 8 \\ 9 \end{Bmatrix}-\frac{\bar{q}L}{3k}\begin{Bmatrix} 1 \\ 2 \\ 3 \end{Bmatrix} \tag{58}$$

The nodal temperatures obtained happen to be here exact. This is shown in Figure 2.10 for the case $\bar{T}=0$, $\bar{q}=0$.



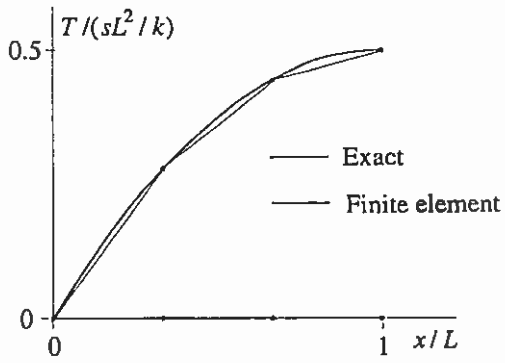



**Figure 2.10** Temperature distribution in the wall ($s>0$).

**Remark 2.12.** It is customary and natural to generate the system equations in such an order that the first equation corresponds to the first nodal variable (meaning that the weighting can be associated to that variable as has been the case here) and so on. Therefore the way the global nodes are numbered in a mesh affects the structure of the system coefficient matrix. For instance, if the nodes are numbered similarly as in the example case of Figure 2.8 continuously from left to right (or from right to left) the coefficient matrix keeps the so-called tri-diagonal form, an example of which can be seen in (55). A more irregular ordering of the nodes distributes the non-zero terms all around the matrix although the number of non-zero terms remains the same. This obviously is apt to make the solution more expensive as less regularity is available to be made use of. There exist algorithms to optimize the node numbering order for finite element meshes. □

**Remark 2.13.** An alternative way from the one described in Remark 2.9 to treat given nodal values due to the Dirichlet boundary which keeps the preliminary system size unaltered is quite often used. We explain it in connection with the set (16). Instead of the form (17) we can write

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & K_{22} & K_{23} & K_{24} \\ 0 & K_{32} & K_{33} & K_{34} \\ 0 & K_{42} & K_{43} & K_{44} \end{bmatrix}\begin{Bmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \end{Bmatrix}=\begin{Bmatrix} \bar{T} \\ b_2-K_{21}\bar{T} \\ b_3-K_{31}\bar{T} \\ b_4-K_{41}\bar{T} \end{Bmatrix} \tag{59}$$

That is, the first equation is replaced by the identity $T_1=\bar{T}$ and the terms $K_{21}\bar{T}$, ... in the rest of the equations are transferred on the right-hand side. This kind of manipulation keeps the matrix again symmetric if it is originally symmetric. □

**Remark 2.14.** If the boundary conditions consist of Neumann conditions only, or in other words, there is no Dirichet boundary in the problem, the analytical temperature distribution solution is not unique. If one solution is $T(x)$, the function $T(x)+c$ where $c$ is a constant, is also a solution. This can be seen by substitution in the governing equations. Physically the reason is that in this case only the gradient of the temperature, not the temperature itself controls the temperature distribution. This fact is reflected in the discrete model. The

corresponding system matrix $[K]$ proves to be singular. One — and only one — nodal value can be arbitrarily given to change the matrix non-singular and to fix the solution. □

**Remark 2.15.** The Dirichlet boundary condition can be taken care of in an alternative fashion by representing the solution in the form

$$T(x) = \bar{T}(x) + \Delta T(x) \tag{60}$$

where $\bar{T}(x)$ is a *given* function in $\bar{\Omega}$, an "extension" of the boundary data into the domain, satisfying $\bar{T}\big|_{\Gamma_D} = \bar{T}$ (the same symbol is used here for the function in $\bar{T}(x)$ in $\bar{\Omega}$ and for $\bar{T}$ on $\Gamma_D$ but this should not cause much confusion as the meaning can be inferred from the context) and $\Delta T(x)$ a *new unknown* function to be determined. Now function $\Delta T(x)$ has to satisfy only (see (60))

$$\Delta T = 0 \quad \text{on} \quad \Gamma_D \tag{61}$$

Introducing (60) in the standard weak form (2.1.28) gives an alternative weak form

$$\int_\Omega \frac{dw}{dx} k \frac{d\Delta T}{dx} d\Omega + \int_\Omega \frac{dw}{dx} k \frac{d\bar{T}}{dx} d\Omega - \int_\Omega ws\, d\Omega + w\bar{q}\big|_{\Gamma_N} = 0 \tag{62}$$

with

$$\Delta T = 0, \qquad w = 0 \quad \text{on} \quad \Gamma_D \tag{63}$$

This formulation has firstly the theoretical advantage that the function set for the trial functions $\Delta T$ and the test functions $w$ is the same and is in fact a linear space (see Appendix C). Secondly, in the finite element approximation

$$\Delta\tilde{T}(x) = \sum_j N_j(x)\, \Delta T_j \tag{64}$$

the step described in Remark 2.9 or in Remark 2.13 to modify the preliminary equations is not needed at all as the given nodal values $\Delta T_j$ of $\Delta T(x)$ at the Dirichlet boundary are zero and thus no contributions are generated from them. Thirdly, in non-linear problems (to be considered in Chapter 11) $\bar{T}(x)$ may represent conveniently the initial solution guess or the current updated solution in an iterative procedure, which is continued until the norm of $\Delta T(x)$ falls below a given limit.

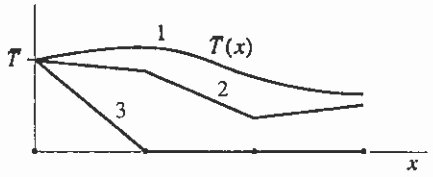



**Figure 2.11** Possible extensions of the Dirichlet data corresponding to problem (1).



The given $\bar{T}(x)$ can be here in principle any $C^0$ continuous function satisfying the Dirichlet boundary condition (see Figure 2.11). When using the finite element method it is natural to represent also $\bar{T}(x)$ by finite elements (function 2 in Figure 2.11):

$$\bar{T}(x) = \sum_j N_j(x)\, \bar{T}_j \tag{65}$$

with $\bar{T}_I = \bar{T}$ or in practice here in the simplest form (function 3 in Figure 2.11)

$$\bar{T}(x) = N_I(x)\bar{T} \tag{66}$$

It should be noted that $\bar{T}(x)$ in formulas (65) and (66) needs no approximation symbol as we can just take the given $\bar{T}(x)$ to have the forms indicated. The final nodal values $T_i = \bar{T}_i + \Delta T_i$ are not effected by the selection of the specific finite element form for $\bar{T}(x)$.

The system equations based on the weak form (62) are (the column matrix consisting of the nodal values of $\Delta T(x)$ is denoted $\{\Delta a\}$)

$$[K]\{\Delta a\} = \{b\} \tag{67}$$

with

$$\begin{aligned} K_{ij} &= \int_\Omega \frac{dN_i}{dx} k \frac{dN_j}{dx} d\Omega \\ b_i &= -\int_\Omega \frac{dN_i}{dx} k \frac{d\bar{T}}{dx} d\Omega + \int_\Omega N_i s\, d\Omega - N_i \bar{q}\big|_{\Gamma_N} \end{aligned} \tag{68}$$

Comparison with (15) shows that only the latter expression is affected.

We will call in what follows formulations like (60) and (62) as the *deltaform* (deltamuoto) to express the difference from the — what we call — standard form taking the non-zero Dirichlet boundary data into account the way described in Remarks 2.9 and 2.13. We are going to favour mostly the standard form in theoretical presentations but the reader should always be prepared to make the minor alterations needed to use the advantageous deltaform in actual practical calculations.

We finally note that often also in classical type of approximations such as (2.1.14), an additional given term, say $\bar{\varphi}(x)$, is present having a similar role as $\bar{T}(x)$ has here. □

**Remark 2.16.** We introduce still some more terminology. The given nodal values, usually due to the Dirichlet boundary, are often called *fixed nodal values* (kiinnitetty solmuarvo) and the rest *free* or *active nodal values* (vapaa tai aktiivinen solmuarvo). Similarly, the system equations corresponding to free or active nodal values are called *free* or *active equations* (vapaa tai aktiivinen yhtälö). The superfluous system equations corresponding to fixed nodal values — which are in fact not strictly correct (see Remark 2.9) formed or not depending on the program formulation — are called here as the *non-active equations* (epäaktiivinen yhtälö). The concept of nodal value is extended in Remark 3.9 to the more general concept of nodal parameter. □

**Example 2.4.** The demonstration problem (1) with the mesh of Figure 2.8 leading to active system equations (57) and to solution (58) is treated again here now with the deltaform.

Instead of the element contributions (38) of the standard form:

$$K_{ij}^e = \int_{\Omega^e} \frac{dN_i^e}{dx} k \frac{dN_j^e}{dx} d\Omega$$
$$b_i^e = \int_{\Omega^e} N_i^e s \, d\Omega - N_i^e \bar{q}\Big|_{\Gamma_N^e} \tag{a}$$

we have now from (68) and making use of Remark 2.11 the terms

$$K_{ij}^e = \int_{\Omega^e} \frac{dN_i^e}{dx} k \frac{dN_j^e}{dx} d\Omega$$
$$b_i^e = -\int_{\Omega^e} \frac{dN_i^e}{dx} k \frac{d\bar{T}^e}{dx} d\Omega + \int_{\Omega^e} N_i^e s \, d\Omega - N_i^e \bar{q}\Big|_{\Gamma_N^e} \tag{b}$$

Function $\bar{T}(x)$ is taken here according to the choice 3 in Figure 2.11. Thus it is non-zero only in element 2 where it has the form

$$\bar{T}^2(x) = N_1^2(x)\bar{T} \tag{c}$$

and its derivative

$$\frac{d\bar{T}^2}{dx} = \frac{dN_1^2}{dx}\bar{T} \tag{d}$$

Thus the only change for the element contributions comes from element 2 for which we obtain the additional terms

$$(b_1^2)_{\bar{T}} \equiv -\int_{\Omega^e} \frac{dN_1^2}{dx} k \frac{d\bar{T}^2}{dx} d\Omega = -k\bar{T}\int_{\Omega^e} \frac{dN_1^2}{dx}\frac{dN_1^2}{dx} d\Omega = -\frac{k\bar{T}}{l^{(2)}} = -\frac{3k\bar{T}}{L}$$
$$(b_2^2)_{\bar{T}} \equiv -\int_{\Omega^e} \frac{dN_2^2}{dx} k \frac{d\bar{T}^2}{dx} d\Omega = -k\bar{T}\int_{\Omega^e} \frac{dN_2^2}{dx}\frac{dN_1^2}{dx} d\Omega = \frac{k\bar{T}}{l^{(2)}} = \frac{3k\bar{T}}{L} \tag{e}$$

Formulas (F.1.1) have been made use of. The element column matrices are thus (see formulas (54))

$$\{b\}^1 = \frac{sL}{6}\begin{Bmatrix}1\\1\end{Bmatrix}, \quad \{b\}^2 = \frac{sL}{6}\begin{Bmatrix}1\\1\end{Bmatrix} + \frac{3k\bar{T}}{L}\begin{Bmatrix}-1\\1\end{Bmatrix}, \quad \{b\}^3 = \frac{sL}{6}\begin{Bmatrix}1\\1\end{Bmatrix} - \begin{Bmatrix}0\\\bar{q}\end{Bmatrix} \tag{f}$$

Assembly of the system column matrix gives (see formula (55))



$$\{b\} = \begin{Bmatrix} b_1^2 \\ b_1^1 + b_2^2 \\ b_2^1 + b_1^3 \\ b_2^3 \end{Bmatrix} = \frac{sL}{6}\begin{Bmatrix}1\\2\\2\\1\end{Bmatrix} + \frac{3k\bar{T}}{L}\begin{bmatrix}-1\\1\\0\\0\end{bmatrix} - \begin{Bmatrix}0\\0\\0\\\bar{q}\end{Bmatrix} \tag{g}$$

The system coefficient matrix does not change. The preliminary system equations (67) are thus

$$\frac{3k}{L}\begin{bmatrix}1&-1&0&0\\-1&2&-1&0\\0&-1&2&-1\\0&0&-1&1\end{bmatrix}\begin{Bmatrix}\Delta T_1\\\Delta T_2\\\Delta T_3\\\Delta T_4\end{Bmatrix} = \frac{3k\bar{T}}{L}\begin{bmatrix}-1\\1\\0\\0\end{bmatrix} + \frac{sL}{6}\begin{Bmatrix}1\\2\\2\\1\end{Bmatrix} - \begin{Bmatrix}0\\0\\0\\\bar{q}\end{Bmatrix} \tag{h}$$

The first equation is non-active with the fixed nodal value $\Delta T_1 = 0$. The remaining active equations are (after division by $3k/L$)

$$\begin{bmatrix}2&-1&0\\-1&2&-1\\0&-1&1\end{bmatrix}\begin{Bmatrix}\Delta T_2\\\Delta T_3\\\Delta T_4\end{Bmatrix} = \begin{Bmatrix}\bar{T}\\0\\0\end{Bmatrix} + \frac{sL^2}{18k}\begin{Bmatrix}2\\2\\1\end{Bmatrix} - \frac{\bar{q}L}{3k}\begin{Bmatrix}0\\0\\1\end{Bmatrix} \tag{i}$$

The set is identical with (57) except for the unknowns. However, as here

$$\bar{T}(x_2) = \bar{T}(x_3) = \bar{T}(x_4) = 0 \tag{j}$$

there follows

$$T_2 = \Delta T_2, \quad T_3 = \Delta T_3, \quad T_4 = \Delta T_4 \tag{k}$$

and we have obtained in fact again the solution (58).

**Remark 2.17.** This remark concerns the important point of smoothness of the functions when performing integrations by parts. In the manipulations to obtain the standard weak form (2.1.28), integration by parts gave

$$\int_\Omega w\frac{d}{dx}\left(-k\frac{dT}{dx}\right)d\Omega = \int_\Omega \frac{dw}{dx}k\frac{dT}{dx}d\Omega + \left(wk\frac{dT}{dx}\right)\Big|_{x=a} - \left(wk\frac{dT}{dx}\right)\Big|_{x=b} \tag{69}$$

According to Section B.1 the functions involved ($w$ and $-k\,dT/dx$ here) must be smooth enough — at least $C^0$ functions — for the above to be valid. This condition is fulfilled for the exact solution as the flux $-k\,dT/dx$ is indeed continuous for energy balance reasons even for a discontinuous $k$. But finally we are applying the weak form for the finite element approximation $\bar{T}$ whose derivative $d\bar{T}/dx$ has jumps also at those points where $k$ is continuous. The *approximate flux is thus no more continuous* at the element interfaces. To be honest, we should still be able to perform the integration by parts also in this case and to be on the safe side we may start by performing the manipulations in an element by element fashion and obtain first

$$\int_\Omega w\frac{\mathrm{d}}{\mathrm{d}x}\left(-k\frac{\mathrm{d}T}{\mathrm{d}x}\right)\mathrm{d}\Omega = \sum_e \int_{\Omega^e} w\frac{\mathrm{d}}{\mathrm{d}x}\left(-k\frac{\mathrm{d}T}{\mathrm{d}x}\right)\mathrm{d}\Omega$$

$$= \sum_e \int_{\Omega^e} \frac{\mathrm{d}w}{\mathrm{d}x}k\frac{\mathrm{d}T}{\mathrm{d}x}\mathrm{d}\Omega + \left(wk\frac{\mathrm{d}T}{\mathrm{d}x}\right)\Big|_{x=a} - \sum_I w_I\left[\!\left[-k\frac{\mathrm{d}T}{\mathrm{d}x}\right]\!\right]_I - \left(wk\frac{\mathrm{d}T}{\mathrm{d}x}\right)\Big|_{x=b} \tag{70}$$

The summation over $I$ means sum over the element interfaces (here the nodes inside the domain, for instance, in the mesh of Figure 2.8 nodes 2 and 3). The meaning of the jump bracket notation is the following:

$$[\![f]\!] = f^+ - f^- \tag{71}$$

where the minus and plus values refer to the left- and right-handed limit values of function $f$ at an element interface point. For the exact solution with continuous flux the jump terms in (70) are seen to vanish. Result (70) is correct for a $C^0$ continuous approximation for $T$, for which $k\mathrm{d}\bar{T}/\mathrm{d}x$ is a $C^{-1}$ function, but if applied with finite elements it is found to produce completely useless results. We can find a remedy, however. Instead of the field equation representation (2.1.1), we start with, see e.g. Salonen (1991),

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(-k\frac{\mathrm{d}T}{\mathrm{d}x}\right) - s = 0 \qquad \text{in } \Omega^e,\ \ e = 1,2,\cdots \tag{72}$$

and

$$\left[\!\left[-k\frac{\mathrm{d}T}{\mathrm{d}x}\right]\!\right] = 0 \qquad \text{for } I = 1,2,\cdots \tag{73}$$

This means that we have anticipated that the solution may not be smooth particularly on the element interfaces and we thus replace the field equation at these finite points with the continuity condition for the heat flux. The corresponding starting formulation for the weak form is thus

$$\sum_e \int_{\Omega^e} w\left[\frac{\mathrm{d}}{\mathrm{d}x}\left(-k\frac{\mathrm{d}T}{\mathrm{d}x}\right) - s\right]\mathrm{d}\Omega + \sum_I v_I\left[\!\left[-k\frac{\mathrm{d}T}{\mathrm{d}x}\right]\!\right]_I = 0 \tag{74}$$

Comparison with expression (70) shows that if we select we weighting constants $v_I$ by

$$v_I = w(x_I) = w_I \tag{75}$$

the jump terms cancel after integration by parts has been applied on (74) and *we end up with the standard weak form* (2.1.28) written here just as a summation over the elements:

$$\sum_e \int_{\Omega^e} \frac{\mathrm{d}w}{\mathrm{d}x}k\frac{\mathrm{d}T}{\mathrm{d}x}\mathrm{d}\Omega - \sum_e \int_{\Omega^e} ws\,\mathrm{d}\Omega + \sum_e w\bar{q}\big|_{\Gamma_N^e} = 0 \tag{76}$$

The above lines of thought can be applied similarly also in two and three dimensions. Reference Belytschko et.al. (2000) contains detailed considerations on this theme in connection with the principle of virtual work (see Remark 3.1). In what follows we do not care to go through these extra details, in fact needed to proceed quite correctly, but perform the integrations by parts manipulations directly over the whole domain assuming the functions



to be smooth enough and "forget" on purpose the violations brought in via the approximations. □

## 2.4 PRE- AND POST-PROCESSING

The solution of a problem by the finite element method can be divided roughly into three separate phases: (1) pre-prosessing, (2) generation and solution of the discrete equations, (3) post-processing.

### 2.4.1 Pre-processing

*Pre-processing* (esikäsittely) means the preparation of the data needed by the program to produce the discrete equations. This includes the generation of the finite element mesh, information about the terms in the weak form and about the Dirichlet boundary conditions. It is clear that in practice with thousands of elements this phase must be effected mostly automatically. Thus, a more or less automatic mesh generation algorithm must be available in any useful commercial finite element package. Here, in the Mathematica based, Wolfram (1999), demonstration program MATHFEM used in this text (Appendix G) the degree of automatization in pre-processing is rather mild and will be explained later.

*Adaptive procedures* (adaptiivinen menettely) in connection with the finite element method are becoming more and more important. It must be remembered that the finite element method is an approximate procedure to solve a problem. The early applications of the finite element method were based mainly — and still often are — on the practical intuitive knowledge of the applier on the accuracy achievable with a given type of mesh in a given type of problem. However, to proceed logically, some systematic criteria to study the quality of a finite element solution must preferably be available. Different *error estimators* (virhe-estimaattori) have been developed in the literature for this purpose. A recent reference is Ainsworth and Oden (2000). If the errors associated with a tentative finite element solution are found to be too large, the problem must be solved again with a new mesh. How the initial mesh should be locally *refined* (tihentää) or perhaps *de-refined* or *coarsened* (harventaa) for best economy is based on *error indicators* (virheindikaattori). Several consecutive meshes may have to be used in a problem to achieve the accuracy needed. Further, it is obvious that if an iterative method is employed to solve the discrete equations, the previous solutions with earlier meshes can be made use of to produce an initial guess for the new solution. This means that the pre-processing phases and the equation generation and solution phases get mixed.

### 2.4.2 Post-processing

*Post-processing* (jälkikäsittely) in general means that supplementary information needed by the applier are extracted from the "raw material" of the finite element solution consisting of the nodal values of the approximation. This information should be presented preferably mostly in graphical form. Similarly as with pre-processing, more or less automatic post-processing algorithms are in practice necessary to cope with the huge amount of data produced by the solution. The options available in MATHFEM will be described later.

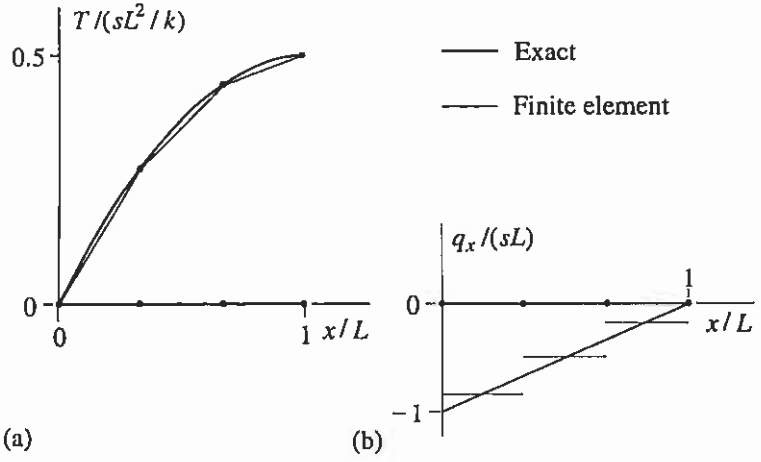



**Figure 2.12 (a)** Temperature. **(b)** Heat flux ($s>0$).

Figure 2.12 (a) shows again the exact and the finite element temperature distribution obtained in the example problem with three linear elements in the wall in the case $\bar{T}=0$, $\bar{q}=0$ considered in the previous section. Figure (b) shows the corresponding heat flux

$$q_x=-k\frac{\mathrm{d}T}{\mathrm{d}x} \tag{1}$$

The heat flux concept is discussed in more detail in Section 3.1.1. In practical problems the heat flux and especially the heat flow rate (3.1.4):

$$\dot{Q}=\int_\Gamma q\,\mathrm{d}\Gamma \tag{2}$$

through certain surfaces are of interest in addition to the temperature. The situation is similar, say, in applications of structural mechanics. The basic unknown is usually the displacement field and the corresponding nodal variables are thus the displacement components. Often, however, also the stresses induced are of basic importance. Stresses correspond to the heat flux as



they are obtained through differentiation from the displacement field (multiplication by some material property data similarly as in (1) is also needed).

In the example problem the finite element solution for the heat flux is unrealistically discontinuous. The jumps in the values are clearly some measures of the errors in the solution. A simple and common practice to refine the heat flux solution is to take as the nodal value the average of the values at both sides of a node and then use again the finite element approximation now for the flux. This would give here the exact flux in the middle element but it cannot improve the result at the domain boundary where only "one-sided" information is available. It is just there where accurate results would be valuable. Here the heat flow rate density out of the body has the expressions (see Section 3.1.1)

$$\begin{aligned} q=-q_x&=k\frac{\mathrm{d}T}{\mathrm{d}x} \quad \text{at } x=0\\ q=q_x&=-k\frac{\mathrm{d}T}{\mathrm{d}x} \quad \text{at } x=L \end{aligned} \tag{3}$$

At the right-hand Neumann boundary $x=L$ the heat flow rate density is given, $\bar{q}=0$, and we in fact need no approximation. The value obtained directly from the finite element solution (see (2.3.58))

$$\begin{aligned} \tilde{q}\big|_{x=L}&=\tilde{q}_x(L)=-k\frac{\mathrm{d}\tilde{T}}{\mathrm{d}x}(L)=-k\frac{\mathrm{d}T^{(3)}}{\mathrm{d}x}(L)=-k\frac{\mathrm{d}\xi}{\mathrm{d}x}\frac{\mathrm{d}T^{(3)}}{\mathrm{d}\xi}\bigg|_{\xi=1}\\ &=-k\frac{1}{L/3}(-T_3+T_4)=-k\frac{3}{L}\frac{sL^2}{18k}(-8+9)=-\frac{1}{6}sL\approx-0.167sL \end{aligned} \tag{4}$$

can be considered as a measure of the error in the solution. At the left-hand Dirichlet boundary $x=0$ the heat flow rate density evaluated directly from the finite element solution is similarly not very accurate:

$$\begin{aligned} \tilde{q}\big|_{x=0}&=-\tilde{q}_x(0)=k\frac{\mathrm{d}\tilde{T}}{\mathrm{d}x}(0)=k\frac{\mathrm{d}T^{(2)}}{\mathrm{d}x}(0)=-k\frac{\mathrm{d}\xi}{\mathrm{d}x}\frac{\mathrm{d}T^{(3)}}{\mathrm{d}\xi}\bigg|_{\xi=0}\\ &=k\frac{1}{L/3}(-T_1+T_2)=k\frac{3}{L}\frac{sL^2}{18k}(-0+5)=\frac{5}{6}sL\approx0.833sL \end{aligned} \tag{5}$$

as the exact value is $1\cdot sL$. This result can be made here rather easily more accurate by post-processing as follows.

The basic energy equation weak form (3.1.19) presented in Section 3.1.1 is here in one dimension

$$-\int_\Omega \frac{\mathrm{d}w}{\mathrm{d}x} q_x \mathrm{d}\Omega - \int_\Omega ws\,\mathrm{d}\Omega + \int_\Gamma wq\,\mathrm{d}\Gamma = 0 \tag{6}$$

The standard weak form (2.1.28) is arrived at by introducing the constitutive relation (1) and by restricting the weighting function $w$ to disappear on the Dirichlet boundary. The boundary consist here of the points $x=0$ and $x=L$ and (6) is in more detail

$$-\int_\Omega \frac{\mathrm{d}w}{\mathrm{d}x} q_x \mathrm{d}\Omega - \int_\Omega ws\,\mathrm{d}\Omega + wq\big|_{x=0} + w\bar{q}\big|_{x=L} = 0 \tag{7}$$

when the Neumann boundary condition (2.3.1) expressed in the form $q(L)=\bar{q}$ is made use of. It is of interest to note that the selection $w\equiv 1$ in (7) gives the exact heat flow rate density at the left-hand boundary:

$$q\big|_{x=0} = \int_\Omega s\,\mathrm{d}\Omega - \bar{q}\big|_{x=L} \tag{8}$$

This result, extracted from the weak form, has an obvious physical content: the heat flow rate out of the body through the left hand boundary equals the heat rate generated inside the body minus the heat flow rate out of the body through the right-hand boundary. Taking the example case, $s$ is constant and $\bar{q}=0$, which gives the exact value $q\big|_{x=0} = sL$.

For instance in the case, where the boundary conditions are of the Dirichlet type at the both ends, the selection $w\equiv 1$ gives correctly the total heat outflow rate but we cannot any more find the separate portions at the ends and some alternative procedure is to be used. The obvious choice is to select the weighting to disappear on those boundary parts over which the heat flow rate is not to be evaluated and to have a constant value (say 1) on those parts over which the heat flow rate is wanted to be known. This selection transforms (7) into the form

$$F \equiv -\int_\Omega \frac{\mathrm{d}w}{\mathrm{d}x} q_x \mathrm{d}\Omega - \int_\Omega ws\,\mathrm{d}\Omega + wq\big|_{x=0} = 0 \tag{9}$$

where now $w(0)=1$. This could be used again to evaluate $q\big|_{x=0}$ if the exact $q_x$ would be known which is however not the case in general. We can try to simulate (9) in the finite dimensional case by replacing the exact $q_x$ with the finite element approximation $\tilde{q}_x = -k\mathrm{d}\tilde{T}/\mathrm{d}x$ and by selecting the weighting to



be the global weighting function $N_1$ which clearly satisfies the conditions $w(0)=1$ and $w(L)=0$. The analogue of (9) is thus

$$\hat{F}_1 \equiv \int_\Omega \frac{\mathrm{d}N_1}{\mathrm{d}x} k \frac{\mathrm{d}\tilde{T}}{\mathrm{d}x} \mathrm{d}\Omega - \int_\Omega N_1 s\,\mathrm{d}\Omega + N_1 \hat{q}\big|_{x=0} = 0 \tag{10}$$

We have denoted the heat flow rate density in (10) at $x=0$ as $\hat{q}$ to emphasize the possible difference with the exact quantity $q$ and with the direct approximation $\tilde{q} = -\tilde{q}_x = k\mathrm{d}\tilde{T}/\mathrm{d}x$. Comparison, say, with equation (2.3.22) with $i=1$ shows that we have produced the result ($N_1(0)=1$)

$$\hat{F}_1 \equiv F_1 + \hat{q}\big|_{x=0} = 0 \tag{11}$$

which is almost the first system equation or in more detail

$$K_{11}T_1 + K_{12}T_2 + K_{13}T_3 + K_{14}T_4 - b_1 + \hat{q}\big|_{x=0} = 0 \tag{12}$$

If the terms $K_{1j}$ and $b_1$ for the first preliminary discrete equation have been assembled and stored we can now evaluate $\hat{q}$ from (12):

$$\begin{aligned}\hat{q}\big|_{x=0} &= -\sum K_{1j}T_j + b_1 = -\frac{3k}{L}\frac{sL^2}{18k}(1\cdot 0 - 1\cdot 5 + 0\cdot 8 + 0\cdot 9) + \frac{sL}{6} \\ &= \frac{5}{6}sL + \frac{1}{6}sL = sL\end{aligned} \tag{13}$$

Formulas (2.3.56) and (2.3.58) have been applied. The exact result happened to be recovered by this procedure and thus the generation of the first equation which was later discarded was not in vain as the terms generated could be made use of in accurate evaluation of the heat flow rate. This kind of procedure is to be preferred over the direct calculation from the approximation if accurate results are needed. An intuitive explanation for the increased accuracy of (13) over (5) is in the fact that expression (13) contains information in addition to the nodal values also about the source term. In two and three dimensional cases analogical procedures to calculate heat flow rates through Dirichlet boundaries can be devised, e.g. Gresho et al. (1987). Reference Zienkiewicz and Taylor (2000) explains also useful practical recovery processes.

**Remark 2.18.** On Neumann boundaries the heat flux rate density is given and the temperature is unknown. On Dirichlet boundaries the temperature is given and the heat flux rate density is unknown. Starting from the basic form (6) we can write instead of (2.3.5) the typical discrete equation always first as

$$F_i \equiv \int_\Omega \frac{\mathrm{d}N_i}{\mathrm{d}x} k \frac{\mathrm{d}\tilde{T}}{\mathrm{d}x} \mathrm{d}\Omega - \int_\Omega N_i s\,\mathrm{d}\Omega + N_i q\big|_\Gamma = 0 \tag{14}$$

If node $i$ is on the Neumann boundary, we can replace $q$ with the given value and are back at formula (2.3.5). If node $i$ is on the Dirichlet boundary, the correct discrete equation should include the term

$$N_i q\big|_{\Gamma_D} \tag{15}$$

arising from the unknown heat flow rate density. This means that the discrete equations produced by the standard weak form where the contributions from the Dirichlet boundary are missing are in fact incorrect as mentioned earlier. This is quite obvious because the condition of the weighting function to disappear on the Dirichlet boundary used to obtain the weak form is later violated in generating the discrete equation. As these wrong equations — if generated — are later discarded no harm is done. Term (15) is sometimes called the *thermal reaction* (terminen reaktio), Akin (1994). Equation (10) is an example of a formulation where the thermal reaction is included for post-processing purposes.

The name "reaction" comes from the terminology of structural mechanics. There roughly, on Neumann boundaries the traction (stress vector) is given and the displacements are unknown. On Dirichlet boundaries the displacements are given and the traction (reaction forces per unit area from the surroundings) is unknown. In the so-called displacement formulation where the basic unknowns are the displacements, correct discrete equations corresponding to nodes on the Dirichlet boundary contain reaction force terms, which can be evaluated by post-processing.

We can proceed in theory alternatively by producing all the system equations right from the beginning using instead of the standard the basic energy equation weak form, that is, employing here formulas (14) written in more detail as (cf. (7))

$$F_i \equiv \int_\Omega \frac{\mathrm{d}N_i}{\mathrm{d}x} k \frac{\mathrm{d}\bar{T}}{\mathrm{d}x}\,\mathrm{d}\Omega - \int_\Omega N_i s\,\mathrm{d}\Omega + N_i q\big|_{\Gamma_D} + N_i \bar{q}\big|_{\Gamma_N} = 0 \tag{16}$$

We show this in connection with the demonstration problem described in Figure 2.8. Looking at equations (2.3.56), we see that only the first equation changes and obtains the form

$$\frac{3k}{L}(T_1 - T_2) + N_1(0)\,\hat{q}\big|_{x=0} = \frac{sL}{6} \tag{17}$$

The unknown approximative heat flow rate density at $x = 0$ has been denoted similarly as in (10). Taking further into account that $T_1 = \bar{T}$ we obtain instead of (2.3.56) the system equations

$$\frac{3k}{L}\begin{bmatrix} L/(3k) & -1 & 0 & 0 \\ 0 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}\begin{Bmatrix} \hat{q}\big|_{x=0} \\ T_2 \\ T_3 \\ T_4 \end{Bmatrix} = \frac{3k}{L}\begin{Bmatrix} -\bar{T} \\ \bar{T} \\ 0 \\ 0 \end{Bmatrix} + \frac{sL}{6}\begin{Bmatrix} 1 \\ 2 \\ 2 \\ 1 \end{Bmatrix} - \begin{Bmatrix} 0 \\ 0 \\ 0 \\ \bar{q} \end{Bmatrix} \tag{18}$$

Now the unknowns consist of three nodal temperatures and of one nodal flux. This is a kind of "mixed formulation". The solution for the temperatures is the same as before and the nodal flux is also given by (13). The system matrix in (18) is no more symmetric. This is certainly one of the reasons why this kind of formulation is not used in practice. □

## REFERENCES




Ainsworth, M. and Oden, J. T. (2000). A Posteriori Error Estimation in Finite Element Analysis, Wiley, New York, ISBN 0-471-29411-X.

Akin, J. E. (1994). *Finite Elements for Analysis and Design*, Academic Press, London, ISBN 0-12-047654-1.

Becker, A. A. (1992). *The Boundary Element Method in Engineering: a complete course*, Mc-Graw-Hill, London, ISBN 0-07-707415-7.

Beer, G. and Watson J. O. (1992). *Introduction to Finite and Boundary Element Methods for Engineers*, Wiley, Chichester, ISBN 0-471-92813-5.

Belytschko, T., Liu, W. K. and Moran, B. (2000). *Nonlinear Finite Elements for Continua and Structures*, Wiley, Chichester, ISBN 471-988774-3.

Conte, S. D. and de Boor, C. (1972). *Elementary Numerical Analysis, An algorithmic Approach*, 2nd ed., Mc-Graw-Hill, Tokyo.

Crandall, S. H. (1956). *Engineering Analysis, A Survey of Numerical Procedures*, Mc-Graw-Hill, New York.

Gresho, P. M., Lee, R. L., Sani, R. L., Maslanik, M. K. and Eaton, B. E. (1987). The Consistent Galerkin FEM for Computing Derived Boundary Quantities in Thermal and/or Fluids Problems, *Int. j. numer. methods fluids*, Vol. 7, 371 - 394.

Salonen, E.-M. (1991). A note on the derivation of weak forms, (Onate, E., Periaux, J. and Samuelsson, A. ed.) *Finite Elements in the 90's*, Springer/CIMNE, Barcelona, 549 - 553, ISBN 84-87867-04-9.

Wolfram, S. (1999). *The Mathematica Book*, 4th ed., Wolfram Media/Cambridge University Press . ISBN 0-521-64314-7.

Zienkiewicz, O. C. and Taylor, R. L. (2000). *The Finite Element Method*, 5th ed., Butterworth-Heineman, Oxford. Vol. 1: *The Basis*, ISBN 0 7506 5049 4. Vol 2: *Solid Mechanics*, ISBN 0 7506 5055 9. Vol 3: *Fluid Dynamics*, ISBN 0 7506 5050 8.


## PROBLEMS

# 3 MORE DIFFUSION

## 3.1 HEAT CONDUCTION

In this chapter we expand on the presentation of Chapter 2 still without dealing with convection and only slightly with reaction terms. The weak form introduced in Section 2.1.2 in one dimension is treated here from a more general and physical point of view.

### 3.1.1 Energy equation weak form

By applying the principle of balance of energy (= the first law of thermodynamics) (energian taseen periaate, termodynamiikan ensimmäinen pääsääntö) to a differential continuum volume element, it is found that a quantity, called the *heat flux vector* (lämpövuovektori)

$$\mathbf{q} = q_x\mathbf{i} + q_y\mathbf{j} + q_z\mathbf{k} \tag{1}$$

($[\mathbf{q}] = \mathrm{W/m^2}$) can be defined, which has the following property. Consider a differential *material* surface element $\mathrm{d}S$ in the continuum or on its surface (Figure 3.1 (a)) with the unit normal vector

$$\mathbf{n} = n_x\mathbf{i} + n_y\mathbf{j} + n_z\mathbf{k} \tag{2}$$

The differential heat which flows through the surface element (positive towards the side given by $\mathbf{n}$) during a time differential $\mathrm{d}t$ is $q\mathrm{d}S\mathrm{d}t$ ($[q\mathrm{d}S\mathrm{d}t] = \mathrm{J}$) where $q$ is called the *heat flow rate density* (lämpövirran tiheys) ($[q] = \mathrm{W/m^2}$). This and the heat flux vector are connected by

$$\boxed{q = \mathbf{n}\cdot\mathbf{q}} = n_x q_x + n_y q_y + n_z q_z \tag{3}$$

This relationship is similar to the connection between the stress vector and stress tensor at a point but simpler as here $q$ is a scalar and $\mathbf{q}$ is a vector. It should be noted that the terminology used in connection with $q$ seems to vary much in the literature.

The *heat flow rate* (lämpövirta, lämpöteho) $\dot{Q}$ ($[\dot{Q}] = \mathrm{W}$) through a finite surface (Figure 3.1 (b)) is thus given by

$$\dot{Q} = \int_S q\mathrm{d}S = \int_S \mathbf{n}\cdot\mathbf{q}\mathrm{d}S \tag{4}$$

where the meaning of the notations is rather obvious. (As heat is not a thermodynamic state function, the differential heat $\bar{\mathrm{d}}Q$ crossing the surface





during the time increment $\mathrm{d}t$ is not a total differential and thus we use for the ratio $\bar{\mathrm{d}}Q/\mathrm{d}t$ instead of the more familiar symbol $\dot{Q}$ the symbol $\overset{*}{Q}$ .)

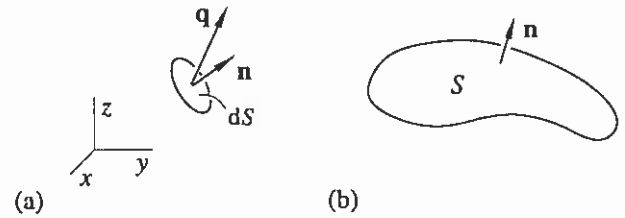



**Figure 3.1 (a)** Heat flow through a differential surface element. **(b)** Heat flow through a finite surface.

Figure 3.2 shows a three-dimensional body and some notations. The boundary $\Gamma = S$ of the domain $\Omega = V$ is assumed to consist here in general of three different types depending on the boundary conditions. These are the Dirichlet, the Neumann and the Robin boundaries. Applications of the two first have appeared already in Chapter 2.

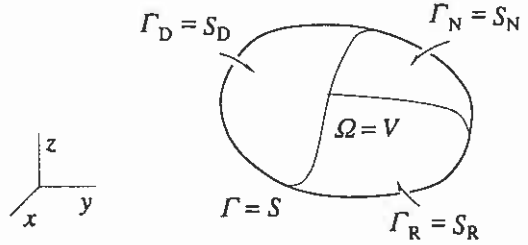



**Figure 3.2** Three-dimensional domain $\Omega = V$ and its boundary $\Gamma = S$.

On the *Dirichlet boundary* $\Gamma_D = S_D$ the temperature is given:

$$T = \overline{T} \tag{5}$$

This is called the *Dirichlet boundary condition.*

On the *Neumann boundary* $\Gamma_N = S_N$ the heat flow rate density is given:

$$q \equiv \mathbf{n}\cdot\mathbf{q} = \overline{q}, \quad q \equiv \mathbf{n}\cdot\left(-k\nabla T\right) = \overline{q} \tag{6}$$

This is called the *Neumann boundary condition.*

On the *Robin boundary* $\Gamma_R = S_R$ convective heat transfer is taking place:

$$q \equiv \mathbf{n}\cdot\mathbf{q} = h(T - T_\infty), \quad q \equiv \mathbf{n}\cdot(-k\nabla T) = h(T - T_\infty) \tag{7}$$

This is called the *Robin boundary condition.*

In the above, $\overline{T}$, $\mathbf{n}$, $\overline{q}$, $k$, $h$, $T_\infty$ are given quantities, the four first having appeared earlier. The quantity $h$ is called convection *heat transfer coefficient* (konvektion lämmönsiirtymiskerroin) ($[h] = \mathrm{W/(m^2K)}$). $T_\infty$ is a reference temperature, say, the temperature of a fluid moving past the body at some distance from the body surface, the so-called free stream temperature. The relationship (7) is often also called *Newton's law of cooling* (Newtonin jäähtymislaki). The value of $h$ depends strongly on the specific application, especially on the flow speed of the surrounding fluid. It should be noted that even when the term "convective heat transfer" is used here, it does not mean that convection, which is due to bulk movement of the body itself and which will be considered later.

For some purposes it is convenient to use the more specific alternatives $V$ and $S$ instead of the general symbols $\Omega$ and $\Gamma$ meant to be valid in any dimension. The forms (6b) and (7b) containing the temperature are obtained assuming as the constitutive relation the *Fourier law* (here for isotropic material) (Fourierin lämmönjohtumislaki)

$$\boxed{\mathbf{q} = -k\nabla T} = -k\left(\frac{\partial T}{\partial x}\mathbf{i} + \frac{\partial T}{\partial y}\mathbf{j} + \frac{\partial T}{\partial z}\mathbf{k}\right) \tag{8}$$

or equivalently

$$q_x = -k\frac{\partial T}{\partial x}, \quad q_y = -k\frac{\partial T}{\partial y}, \quad q_z = -k\frac{\partial T}{\partial z} \tag{9}$$

to be valid.

An alternative expression for the term $\mathbf{n}\cdot(-k\nabla T)$ is

$$\mathbf{n}\cdot(-k\nabla T) = -k\mathbf{n}\cdot\nabla T = -k\left(n_x\frac{\partial T}{\partial x} + n_y\frac{\partial T}{\partial y} + n_z\frac{\partial T}{\partial z}\right) = -k\frac{\partial T}{\partial n} \tag{10}$$

where $\partial T/\partial n$ is the *normal derivative* (normaaliderivaatta) of $T$. On the boundary of a body, $\mathbf{n}$ is always taken to be the outward directed unit normal vector so that in what follows, $q = \mathbf{n}\cdot\mathbf{q}$ is the heat flow rate density positive out of the body under consideration.

The boundary domains cover the whole boundary without gaps and overlaps:



$$\Gamma = \overline{\Gamma}_D \cup \overline{\Gamma}_N \cup \overline{\Gamma}_R, \qquad \Gamma_D \cap \Gamma_N \cap \Gamma_R = \varnothing \tag{11}$$

and similarly

$$\int_\Gamma (\cdot)\,\mathrm{d}\Gamma = \int_{\Gamma_D} (\cdot)\,\mathrm{d}\Gamma + \int_{\Gamma_N} (\cdot)\,\mathrm{d}\Gamma + \int_{\Gamma_R} (\cdot)\,\mathrm{d}\Gamma \tag{12}$$

The governing field equation — obtainable again from the first law of thermodynamics — is in the steady case in a continuum at rest

$$\boxed{\nabla\cdot\mathbf{q} - s = 0} \tag{13}$$

or

$$\frac{\partial q_x}{\partial x} + \frac{\partial q_y}{\partial y} + \frac{\partial q_z}{\partial z} - s = 0 \tag{14}$$

Substitution of the Fourier law gives the conventional *heat conduction equation* (lämmönjohtumisyhtälö)

$$\boxed{\nabla\cdot(-k\nabla T) - s = 0} \tag{15}$$

or

$$\frac{\partial}{\partial x}\left(-k\frac{\partial T}{\partial x}\right) + \frac{\partial}{\partial y}\left(-k\frac{\partial T}{\partial y}\right) + \frac{\partial}{\partial z}\left(-k\frac{\partial T}{\partial z}\right) - s = 0 \tag{16}$$

We derived in Section 2.1 the weak form starting from the one-dimensional counterpart (2.1.1) of field equation (16). However, to obtain the weak form in its "purest shape" we can start directly from (13) or (14) and take the Fourier law into account later. Starting from (14), the obvious steps are (cf. Remark 2.5)

$$\int_\Omega w(x,y,z)\left(\frac{\partial q_x}{\partial x} + \frac{\partial q_y}{\partial y} + \frac{\partial q_z}{\partial z} - s\right)\mathrm{d}\Omega = 0 \tag{17}$$

$$-\int_\Omega \left(\frac{\partial w}{\partial x}q_x + \frac{\partial w}{\partial y}q_y + \frac{\partial w}{\partial z}q_z\right)\mathrm{d}\Omega - \int_\Omega ws\,\mathrm{d}\Omega + \int_\Gamma w(q_x n_x + q_y n_y + q_z n_z)\,\mathrm{d}\Gamma = 0 \tag{18}$$

$$\boxed{-\int_\Omega \nabla w\cdot\mathbf{q}\,\mathrm{d}\Omega - \int_\Omega ws\,\mathrm{d}\Omega + \int_\Gamma wq\,\mathrm{d}\Gamma = 0} \tag{19}$$

The three-dimensional integration by parts manipulations employed in the step between (17) and (18) are based on equation (B.3.1). The step between (18) and (19) is based simply on the definition of the dot product and on formula (3). Equation (19) is called here the *basic energy equation weak form* (energiayhtälön heikko perusmuoto).

**Remark 3.1.** The weak form (19) does not contain any information on the material properties of the body (continuum) under consideration. Thus *it is valid irrespective of the material type* of the body. For readers familiar with strength of materials or structural mechanics the situation is analogous to the use of a weak form called the *principle of virtual work* (virtuaalisen työn periaate) called also the *principle of virtual displacements* (virtuaalisten siirtymien periaate). The ancient principle of virtual work — having history long before finite elements were conceived — is extremely useful in applications; for instance, it is an immediate starting point for discretization with finite elements. *Weak form* (19) *can be considered to have a similar status in the field of heat conduction.* Unfortunately there seems to exist no settled terminology for this weak form or for the terms in it. In lack of anything better we shall call equation (19) as *the energy equation weak form* (energiayhtälön heikko muoto) or in more detail we include the attribute "basic" as above. At the final stage, the material properties must be introduced — here in the form of the Fourier law $\mathbf{q} = -k\nabla T$ — to have a solvable problem. The situation is again similar to that existing in the principle of virtual work. Information about the material under consideration must naturally be finally given. If elastic material, for instance, is assumed, the material properties are given in the form of Hooke's law. Another possible name for the weak form (19) could be the *principle of virtual temperatures* (virtuaalisten lämpötilojen periaate). This name might be justified as follows. In the principle of virtual work or virtual displacements the weighting functions used are classically called *virtual displacements*, that is, they are interpreted physically to be arbitarary infinitesimal displacements or mathematically variations $\delta\mathbf{u}$ of the displacements $\mathbf{u}$ in a body under loading. This physical interpretation for the weighting has made the principle of virtual work more concrete for appliers. The considerations in Section D.3 and especially equation (D.3.30) indicate that we could interprete here without any harm the weighting function $w$ in (19) to be a variation $\delta T$ or "virtual temperature" of the temperatue field $T$ in the body. □

When the identity (12), the Neumann condition (6), the Robin condition (7) and equation (3) are used, the basic weak form (19) can be written in a somewhat more detailed style:

$$-\int_\Omega \nabla w \cdot \mathbf{q}\,\mathrm{d}\Omega - \int_\Omega ws\,\mathrm{d}\Omega + \int_{\Gamma_D} wq\,\mathrm{d}\Gamma + \int_{\Gamma_N} w\bar{q}\,\mathrm{d}\Gamma + \int_{\Gamma_R} wh(T-T_\infty)\,\mathrm{d}\Gamma = 0 \tag{20}$$

**Remark 3.2.** It is not necessary to restrict the weighting function to be zero on the Dirichlet boundary — as was done for instance in connection with the derivation of the standard weak form (2.1.28) in one dimension — it depends on the application purposes if this restriction is useful or not. This theme has been discussed already in Section 2.4.2 and especially in Remark 2.18. There is an analogous situation in the principle of virtual work; the virtual displacements can be of two types depending on the application: virtual displacements satisfying the constraints (kinemaattisesti luvallinen virtuaalinen siirtymä) and virtual displacements violating the constraints (kinemaattisesti luvaton virtuaalinen siirtymä). The



standard practice with the energy equation weak form when generating the system equations is, however, to demand the weighting function to disappear on the Dirichlet boundary (and at the same time demand the temperature to satisfy the Dirichlet condition). This gives the *standard energy equation weak form* (energiayhtälon heikko standardimuoto)

$$-\int_\Omega \nabla w \cdot \mathbf{q}\,\mathrm{d}\Omega - \int_\Omega ws\,\mathrm{d}\Omega + \int_{\Gamma_N} w\bar{q}\,\mathrm{d}\Gamma + \int_{\Gamma_R} wh(T-T_\infty)\,\mathrm{d}\Gamma = 0 \tag{21}$$

For post-processing purposes, for example to evaluate the heat flow through the Dirichlet boundary in an accurate way it is essential not to restrict the weighting function to vanish there. Then the basic form (20) is the one to be used. Often, when we speak about the standard form, we further assume the constitutive relation such as (8) already substituted into it. □

### 3.1.2 Specific cases of the weak form

**Plane case.** For specialization to two- and one-dimensional cases we write the three-dimensional basic energy equation weak form (19) with the alternative domain and boundary symbols $V$ and $S$ to retain $\Omega$ and $\Gamma$ for use in the specialized meanings later:

$$-\int_V \nabla w \cdot \mathbf{q}\,\mathrm{d}V - \int_V ws\,\mathrm{d}V + \int_S wq\,\mathrm{d}S = 0 \tag{22}$$

**Remark 3.3.** It should be emphasized that the integrands in (22) can be shown to be so-called *invariant scalars* (skalaari-invariantti), that is, their values do not depend on the coordinate system used. This is another extremely beautiful and useful property of the weak form, as we can apply (22) directly by just substituting well-known expressions for the vectors $\mathbf{q}$ and $\mathbf{n}$ and for the gradient vector $\nabla w$ in the coordinate system at hand as given by mathematics to obtain the specific weak form. Thus we do not even need to see the specific forms of the governing differential equation as the numerical solution is generated directly from the weak form. (If the differential equation is desired, it can, however, be derived analytically from the weak form; cf. Example 3.1.) We shall present the weak form versions here free of material properties as long as appropriate to have short and clean looking formulas. □

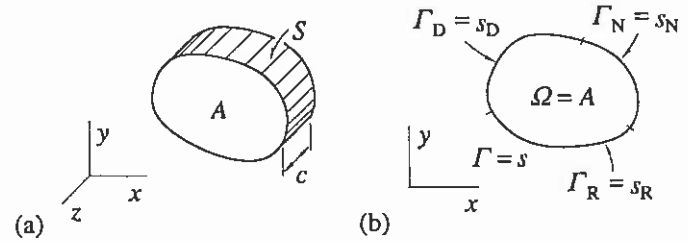



**Figure 3.3** Two-dimensional plane case. **(a)** Three-dimensional body. **(b)** Section $z$ =constant.

Figure 3.3 describes the steps leading to the two-dimensional plane case. The three-dimensional cylindrical body of length $c$ in the $z$-direction depicted in

Figure (a) is employed as the starting point. No dependence on any variable on coordinate $z$ is assumed. The volume and surface element expressions are $dV = dx\,dy\,dz = dz\,dA$ and $dS = ds\,dz = dz\,ds$ and thus

$$\int_V (\cdot)\,dV = \int_z dz \int_A (\cdot)\,dA = c\int_A (\cdot)\,dA$$
$$\int_S (\cdot)\,dS = \int_z dz \int_s (\cdot)\,ds = c\int_s (\cdot)\,ds \tag{23}$$

The meaning of the notations should be clear from the figure. The surface integrals on the two surfaces $z$ = constant of the three-dimensional body do not give any contributions. This follows from $q$ being zero there because of the assumption of no dependence on $z$. Introduction of the results (23) into (22) and division by the common multiplier $c$ gives the weak form

$$-\int_A \nabla w \cdot \mathbf{q}\,dA - \int_A ws\,dA + \int_s wq\,ds = 0 \tag{24}$$

or

$$-\int_\Omega \nabla w \cdot \mathbf{q}\,d\Omega - \int_\Omega ws\,d\Omega + \int_\Gamma wq\,d\Gamma = 0 \tag{25}$$

*valid in the two-dimensional plane case.* When heat flow rate through a certain surface is needed, multiplication of the corresponding boundary line integral by $c$ must of course be performed.

If rectangular Cartesian coordinates are employed,

$$\mathbf{q} = q_x\mathbf{i} + q_y\mathbf{j}, \quad \mathbf{n} = n_x\mathbf{i} + n_y\mathbf{j}, \quad \nabla w = \frac{\partial w}{\partial x}\mathbf{i} + \frac{\partial w}{\partial y}\mathbf{j} \tag{26}$$

and the weak form looks in detail as

$$-\int_A \left(\frac{\partial w}{\partial x}q_x + \frac{\partial w}{\partial y}q_y\right)dA - \int_A ws\,dA + \int_s wq\,ds = 0 \tag{27}$$

or

$$-\int_A \left(\frac{\partial w}{\partial x}q_x + \frac{\partial w}{\partial y}q_y\right)dA - \int_A ws\,dA$$
$$+\int_{s_D} wq\,ds + \int_{s_N} w\bar{q}\,ds + \int_{s_R} wh(T - T_\infty)\,ds = 0 \tag{28}$$

Finally the relationships



$$q_x = -k\frac{\partial T}{\partial x}, \quad q_y = -k\frac{\partial T}{\partial y} \tag{29}$$

can be substituted.

These versions are of course easily seen directly from the three-dimensional forms but we wanted to describe the steps in a similar way as in the axisymmetric case to follow, where the details are somewhat more involved.

**Axisymmetric case.** Figure 3.4 describes the steps leading to the axisymmetric case.

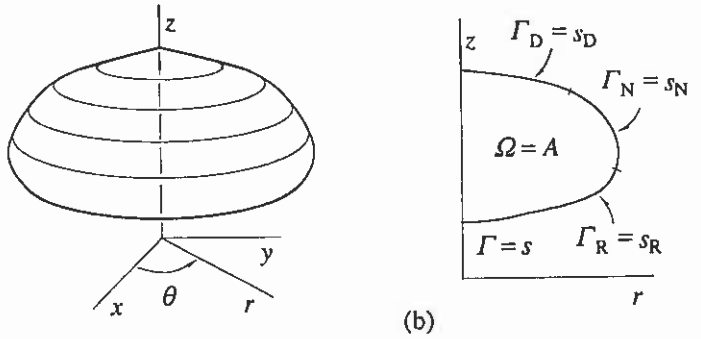



**Figure 3.4** Axisymmetric case. **(a)** Three-dimensional body. **(b)** Section $\theta$ = constant.

The three-dimensional axisymmetric body depicted in Figure (a) is employed as the starting point. No dependence of any variable on the polar angle coordinate $\theta$ is assumed. The volume and surface element expressions are $dV = r\,d\theta\,dr\,dz = d\theta\,r\,dA$ and $dS = r\,d\theta\,ds = d\theta\,r\,ds$ and thus

$$\int_V (\cdot)\,dV = \int_\theta d\theta \int_A (\cdot)\,r\,dA = 2\pi\int_A (\cdot)\,r\,dA$$
$$\int_S (\cdot)\,dS = \int_\theta d\theta \int_s (\cdot)\,r\,ds = 2\pi\int_s (\cdot)\,r\,ds \tag{30}$$

The meaning of the notations should be obvious from the figure. Introduction of the results (30) into (22) and division by the common multiplier $2\pi$ gives the weak form

$$-\int_A \nabla w \cdot \mathbf{q}\,r\,dA - \int_A ws\,r\,dA + \int_s wq\,r\,ds = 0 \tag{31}$$

or

$$-\int_\Omega \nabla w \cdot \mathbf{q}\, r\,\mathrm{d}\Omega - \int_\Omega ws r\,\mathrm{d}\Omega + \int_\Gamma wq r\,\mathrm{d}\Gamma = 0 \tag{32}$$

*valid in the axisymmetric case.* The only change compared to the plane case is seen to be the appearance of the factor $r$ in the integrals. When heat flow rate through a certain surface is needed, multiplication of the corresponding boundary line integral by $2\pi$ must be performed.

In cylindrical coordinates

$$\mathbf{q} = q_r \mathbf{e}_r + q_z \mathbf{e}_z, \quad \mathbf{n} = n_r \mathbf{e}_r + n_z \mathbf{e}_z, \quad \nabla w = \frac{\partial w}{\partial r}\mathbf{e}_r + \frac{\partial w}{\partial z}\mathbf{e}_z \tag{33}$$

and the weak form looks in detail as

$$-\int_A \left( \frac{\partial w}{\partial r} q_r + \frac{\partial w}{\partial z} q_z \right) r\,\mathrm{d}A - \int_A ws\, r\,\mathrm{d}A + \int_s wq\, r\,\mathrm{d}s = 0 \tag{34}$$

or

$$\begin{aligned} &-\int_A \left( \frac{\partial w}{\partial r} q_r + \frac{\partial w}{\partial z} q_z \right) r\,\mathrm{d}A - \int_A ws\, r\,\mathrm{d}A \\ &+\int_{s_\mathrm{D}} wq\, r\,\mathrm{d}s + \int_{s_\mathrm{N}} w\bar{q}\, r\,\mathrm{d}s + \int_{s_\mathrm{R}} wh\left(T - T_\infty\right) r\,\mathrm{d}s = 0 \end{aligned} \tag{35}$$

Finally the relationships

$$q_r = -k\frac{\partial T}{\partial r}, \quad q_z = -k\frac{\partial T}{\partial z} \tag{36}$$

can be substituted.

**Fin.** Next a plate- or finlike geometry is considered (Figure 3.5).

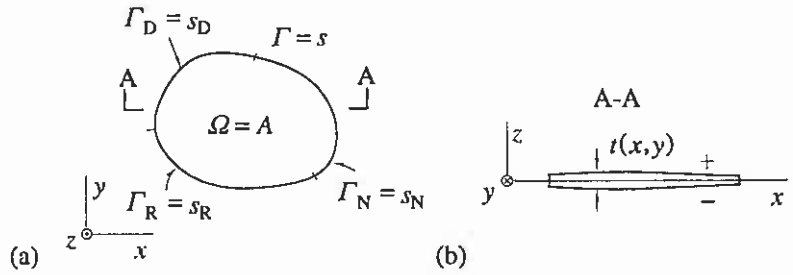



**Figure 3.5** **(a)** Plate geometry. **(b)** Section A-A.



It is first assumed that the middle surface of the plate coincides with the $xy$-plane. The thickness $t$ of the plate can depend on $x$ and $y$:

$$t = t(x, y) \tag{37}$$

We shall use the shorthand notation

$$\int_z (\cdot)\mathrm{d}z \equiv \int_{-t/2}^{t/2} (\cdot)\mathrm{d}z \tag{38}$$

Thus

$$\begin{aligned} &\int_V (\cdot)\mathrm{d}V = \int_A \int_z (\cdot)\mathrm{d}z\mathrm{d}A \\ &\int_S (\cdot)\mathrm{d}S \approx \int_A (\cdot)^+ \mathrm{d}A + \int_A (\cdot)^- \mathrm{d}A + \int_s \int_z (\cdot)\mathrm{d}z\,\mathrm{d}s \end{aligned} \tag{39}$$

The plus- and minus- superscripts refer to the values of the integrand in question at the arbitrary selected plus- and minus-sides of the plate (Figure (b)). As the integration is over the middle plane and not over the curved surfaces, an approximation, which is however acceptable for mild thickness variations and considering the other assumptions to be made later, is included in the latter formula.

It is assumed here that the temperature distribution does not depend on $z$:

$$T(x, y, z) \approx T(x, y) \tag{40}$$

This common assumption cannot of course be in general strictly true, because some gradient must be available in the $z$-direction for heat transfer to be possible through the plate plus- and minus-surfaces. The quantity on the right-hand side in (40) is to be considered to represent some average temperature value through the thickness. The next sharpened assumption would consist of having the second term taken into account in the Taylor expansion of $T$ with respect to $z$:

$$T(x, y, z) \approx T(x, y, 0) + \frac{\partial T}{\partial z}(x, y, 0) z \equiv T(x, y) + U(x, y) z \tag{41}$$

Then we would have to determine two unknown functions: $T$ and $U$. Assumptions (40) and (41) lead to formulations, whose analogs in structural mechanics would be plate stretching and plate bending, respectively. However, here we shall use the simpler expression (40). Corresponding to the aim to obtain a two-dimensional theory, the *weighting function $w$ is similarly chosen to depend only on $x$ and $y$.* The energy equation weak form (22) is now first

$$-\int_A\int_z\left(\frac{\partial w}{\partial x}q_x+\frac{\partial w}{\partial y}q_y\right)\mathrm{d}z\,\mathrm{d}A-\int_A\int_z ws\,\mathrm{d}z\,\mathrm{d}A$$
$$+\int_A wq^+\mathrm{d}A+\int_A wq^-\mathrm{d}A+\int_s\int_z wq\,\mathrm{d}z\,\mathrm{d}s=0 \tag{42}$$

$$-\int_A\left(\frac{\partial w}{\partial x}\int_z q_x\mathrm{d}z+\frac{\partial w}{\partial y}\int_z q_y\mathrm{d}z\right)\mathrm{d}A-\int_A w\int_z s\,\mathrm{d}z\,\mathrm{d}A$$
$$+\int_A w\left(q^++q^-\right)\mathrm{d}A+\int_s w\int_z q\,\mathrm{d}z\,\mathrm{d}s=0 \tag{43}$$

The second form is arrived at by taking into account that the weighting function and its derivatives do not depend on z. We next define the quantities

$$Q_x=\int_z q_x\mathrm{d}z,\quad Q_y=\int_z q_y\mathrm{d}z,\quad Q=\int_z q\,\mathrm{d}z,\quad S=\int_z s\,\mathrm{d}z \tag{44}$$

$Q_x$, $Q_y$ and $Q$ are the heat flow rates per unit lengths along the $y$-, $x$- or $s$-directions ($[Q]=\mathrm{W/m}$). These quantities are the analogs of, say, the stress resultants per unit length much used in structural mechanics. $S$ is the heat source per unit area along the plate middle surface ($[S]=\mathrm{W/m^2}$). Using these notations the weak form gets the outlook

$$-\int_A\left(\frac{\partial w}{\partial x}Q_x+\frac{\partial w}{\partial y}Q_y\right)\mathrm{d}A-\int_A wS\,\mathrm{d}A+\int_A w\left(q^+-q^-\right)\mathrm{d}A+\int_s wQ\,\mathrm{d}s=0 \tag{45}$$

which does not differ too much in structure from the two-dimensional form (27). The third integral is a new type of term. If convective heat transfer, for instance, is assumed to take place on the plus- and minus-sides, we have

$$q^+=h^+\left(T-T_\infty^+\right)$$
$$q^-=h^-\left(T-T_\infty^-\right) \tag{46}$$

where the meaning of the notations should be obvious. Thus the integral

$$\int_A w\left(q^+-q^-\right)\mathrm{d}A=\int_A\left[w\left(h^++h^-\right)T-w\left(h^+T_\infty^++h^-T_\infty^-\right)\right]\mathrm{d}A \tag{47}$$

By deriving similarly as in Example 3.1 the field equation corresponding to weak form (45) with (47) it is easily seen that now for the first time a reaction type term $(h^++h^-)T$ (cf. Section A.1) emerges. How to treat reaction terms accurately in the finite element method is dealt with in Chapter 7.



Often

$$h^+=h^-=h$$
$$T_\infty^+=T_\infty^-=T_\infty \tag{48}$$

and the expressions simplify. The last integral in (45) can be put similarly as before into the form

$$\int_s wQ\,\mathrm{d}s=\int_{s_\mathrm{D}} wQ\,\mathrm{d}s+\int_{s_\mathrm{N}} w\bar{Q}\,\mathrm{d}s+\int_{s_\mathrm{R}} wH\left(T-T_\infty\right)\mathrm{d}s \tag{49}$$

Here the notation is understandable from the definition of $Q$ in (44) and from the Neumann and Robin boundary conditions (6) and (7). $H$ is the integral of $h$ over the thickness or if $h$ is constant we have simply $H=ht$.

**Anisotropy.** Let us consider in this connection the possibility of an *anisotropic* (anisotrooppinen) heat conducting material. The so-called *generalized Fourier law* is then instead of (5)

$$\boxed{\mathbf{q}=-\mathbf{k}\cdot\nabla T} \tag{50}$$

where $\mathbf{k}$ is the heat conduction tensor which is symmetric. Employing index notation and summation convention, we have the Cartesian form

$$q_\alpha=-k_{\alpha\beta}\frac{\partial T}{\partial x_\beta} \tag{51}$$

or using matrices:

$$\begin{Bmatrix}q_x\\q_y\\q_z\end{Bmatrix}=-\begin{bmatrix}k_{xx}&k_{xy}&k_{xz}\\k_{yx}&k_{yy}&k_{yz}\\k_{zx}&k_{zy}&k_{zz}\end{bmatrix}\begin{Bmatrix}\partial T/\partial x\\\partial T/\partial y\\\partial T/\partial z\end{Bmatrix} \tag{52}$$

In the isotropic case

$$\begin{Bmatrix}q_x\\q_y\\q_z\end{Bmatrix}=-\begin{bmatrix}k&0&0\\0&k&0\\0&0&k\end{bmatrix}\begin{Bmatrix}\partial T/\partial x\\\partial T/\partial y\\\partial T/\partial z\end{Bmatrix} \tag{53}$$

we are back at the conventional Fourier law.

Returning to the present problem and substituting the expressions

$$q_x = -k_{xx}\frac{\partial T}{\partial x} - k_{xy}\frac{\partial T}{\partial y} - k_{xz}\frac{\partial T}{\partial z}$$
$$q_y = -k_{yx}\frac{\partial T}{\partial x} - k_{yy}\frac{\partial T}{\partial y} - k_{yz}\frac{\partial T}{\partial z} \qquad (54)$$

into the first two equations (44) gives (here $T$, $\partial T/\partial x$ and $\partial T/\partial y$ do not depend on $z$)

$$Q_x = -K_{xx}\frac{\partial T}{\partial x} - K_{xy}\frac{\partial T}{\partial y}$$
$$Q_y = -K_{yx}\frac{\partial T}{\partial x} - K_{yy}\frac{\partial T}{\partial y} \qquad (55)$$

where

$$K_{xx} \equiv \int_z k_{xx}\,\mathrm{d}z, \quad K_{yy} \equiv \int_z k_{yy}\,\mathrm{d}z, \quad K_{xy} = K_{yx} \equiv \int_z k_{xy}\,\mathrm{d}z \qquad (56)$$

Relationships (55) are the constitutive equations to be employed finally in (45). For composite plates the conductivities depend strongly on $z$. In the isotropic and homogeneous case

$$K_{xx} = kt \equiv K, \quad K_{yy} = kt \equiv K, \quad K_{xy} = K_{yx} = 0 \qquad (57)$$

The formulation which has been presented above in connection with the geometry of Figure 3.5 can obviously be employed with reasonable accuracy also in such cases where the fin middle surface is not strictly a plane, that is, for so-called shallow geometries. The domain $A$ of the problem is then the projection of the fin middle surface on a suitable selected plane.

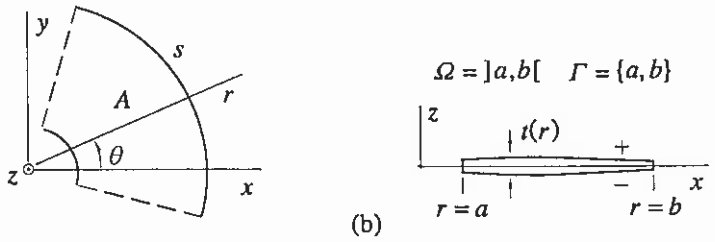



**Figure 3.6** **(a)** Axisymmetric plate geometry. **(b)** Section $\theta$ = constant.

Figure 3.6 describes the axisymmetric special case obtained from the platelike geometry of Figure 3.5. Thus again, no dependence on any variable on coordinate $\theta$ is assumed. The problem becomes mathematically one-



dimensional, the coordinate $r$ being the independent variable. The analogs of equations (37) and (40) are simply

$$t = t(r) \qquad (58)$$

and

$$T(x,y,z) \approx T(r) \qquad (59)$$

and the weighting function $w$ is similarly chosen to depend only on $r$. It is easiest to make the changes needed in the already processed weak form (48) rather than to start from the two-dimensional axisymmetric form (37) or (38). The middle surface area element and boundary curve element (Figure (a)) expressions are $\mathrm{d}A = r\mathrm{d}\theta\mathrm{d}r$ and $\mathrm{d}s = a\mathrm{d}\theta$ or $\mathrm{d}s = b\mathrm{d}\theta$ and thus

$$\int_A (\cdot)\mathrm{d}A = \int_\theta \mathrm{d}\theta \int_r (\cdot) r\,\mathrm{d}r = 2\pi \int_r (\cdot) r\,\mathrm{d}r$$
$$\int_s (\cdot)\mathrm{d}s = \int_\theta \mathrm{d}\theta\, r(\cdot)\big|_{r=a} + \int_\theta \mathrm{d}\theta\, r(\cdot)\big|_{r=b} = 2\pi\left[r(\cdot)\big|_{r=a} + r(\cdot)\big|_{r=b}\right] \qquad (60)$$

Introduction of these results into (45) and division by the common multiplier $2\pi$ gives the weak form

$$-\int_r \frac{\mathrm{d}w}{\mathrm{d}r} Q_r r\,\mathrm{d}r - \int_r wSr\,\mathrm{d}r + \int_r w\left(q^+ - q^-\right) r\,\mathrm{d}r + wrQ\big|_{r=a} + wrQ\big|_{r=b} = 0 \qquad (61)$$

*valid in the axisymmetric fin case.* The first integral may need an explanation. The invariant integrand in the first integral of (45) can be evaluated here by letting, say, the $x$-direction coincide locally with the $r$-direction. Then $Q_y = Q_\theta = 0$ due to the axisymmetry. Further,

$$Q_r = \int_z q_r\,\mathrm{d}z \qquad (62)$$

and the Fourier law obtains the form

$$Q_r = -K_{rr}\frac{\mathrm{d}T}{\mathrm{d}r} \qquad (63)$$

where

$$K_{rr} \equiv \int_z k_{rr}\,\mathrm{d}z \qquad (64)$$

Finally, the comments above concerning the formulation when the fin middle surface is not strictly a plane apply also here. Further, the true heat flow rate evaluations again need multiplication by the factor $2\pi$.

**Example 3.1.** We derive the governing field equation and boundary conditions in the axisymmetric case from the corresponding weak form (35):

$$-\int_A\left(\frac{\partial w}{\partial r}q_r+\frac{\partial w}{\partial z}q_z\right)r\,dA-\int_A ws\,r\,dA$$
$$+\int_{s_D} wq\,r\,ds+\int_{s_N} w\bar{q}\,r\,ds+\int_{s_R} wh(T-T_\infty)r\,ds=0 \qquad \text{(a)}$$

Again, all is based on the fact that the weighting function is arbitrary. However, as both $w$ and its derivatives appear simultaneously in (a) we cannot draw yet directly any conclusions. The strategy is to go in the opposite direction than in the generation of the weak form and perform first suitable integration by parts to get rid of the derivatives on $w$ in the area integral. Thus, employing formula (B.2.1a), we get

$$-\int_A\left(\frac{\partial w}{\partial r}q_r+\frac{\partial w}{\partial z}q_z\right)r\,dA=-\int_A\left(\frac{\partial w}{\partial r}rq_r+\frac{\partial w}{\partial z}rq_z\right)dA$$
$$=\int_A\left(w\frac{\partial}{\partial r}(rq_r)+w\frac{\partial}{\partial z}(rq_z)\right)dA-\int_s(wrq_rn_r+wrq_zn_z)ds$$
$$=\int_A w\left(\frac{\partial}{\partial r}(rq_r)+\frac{\partial}{\partial z}(rq_z)-sr\right)dA-\int_s wqr\,ds \qquad \text{(b)}$$

In this application of formula (B.2.1a), $x$ and $y$ have here the roles $r$ and $z$, respectively, and the situation is still described in rectangular coordinates as seen from Figure 3.4 (b). Further, we have used according to (33) the relation

$$q\equiv\mathbf{n}\cdot\mathbf{q}=n_rq_r+n_zq_z \qquad \text{(c)}$$

Equation (a) becomes thus (cf. formula (11a))

$$\int_A w\left(\frac{\partial}{\partial r}(rq_r)+\frac{\partial}{\partial z}(rq_z)-sr\right)dA$$
$$+\int_{s_N} w(\bar{q}-q)r\,ds+\int_{s_R} w\left[h(T-T_\infty)-q\right]r\,ds=0 \qquad \text{(d)}$$

Using now a similar logic as in connection with the derivation of statement (2.1.9), the following results are obtained:

$$\frac{\partial}{\partial r}(rq_r)+\frac{\partial}{\partial z}(rq_z)-sr=0 \quad \text{in } A$$
$$q=\bar{q} \quad \text{on } s_N \qquad \text{(e)}$$
$$q=h(T-T_\infty) \quad \text{on } s_R$$

As $r$ does not depend on z, the term $\partial(rq_z)/\partial z$ can also put into the form $r\partial q_z/\partial z$. It is to be noted — as discussed earlier — that the possible Dirichlet boundary condition does not follow from the weak form; in fact, the temperature does not appear at all in the domain $A$ in equation (a). Finally, constitutive relations (36) can be introduced.



### 3.1.3 Generalization

We have considered above the steady state heat conduction problem. It is a special case of the steady pure diffusion problem discussed in Appendix A and given there with the field equation

$$\frac{\partial j_\alpha^d}{\partial x_\alpha}-f=0 \quad \text{in } \Omega \qquad (65)$$

and with the boundary conditions

$$\phi=\bar{\phi} \quad \text{on } \Gamma_D$$
$$j^d=\bar{j}^d \quad \text{on } \Gamma_N \qquad (66)$$
$$j^d=a\phi+b \quad \text{on } \Gamma_R$$

The diffusion flux density

$$j^d=n_\alpha j_\alpha^d, \qquad j^d=-n_\alpha D_{\alpha\beta}\frac{\partial\phi}{\partial x_\beta} \qquad (67)$$

A common constitutive law for the diffusion flux vector $j_\alpha^d$ is

$$j_\alpha^d=-D_{\alpha\beta}\frac{\partial\phi}{\partial x_\beta} \qquad (68)$$

which has been employed already in (67b).

The basic weak form corresponding to equations (65) and (67) is

$$-\int_\Omega\frac{\partial w}{\partial x_\alpha}j_\alpha^d\,d\Omega-\int_\Omega wf\,d\Omega+\int_\Gamma wj^d\,d\Gamma=0 \qquad (69)$$

It is the analog of (19). The standard weak form corresponding to (21) with the constitutive relation included is

$$\int_\Omega\frac{\partial w}{\partial x_\alpha}D_{\alpha\beta}\frac{\partial\phi}{\partial x_\beta}d\Omega-\int_\Omega wf\,d\Omega+\int_{\Gamma_N} w\bar{j}^d\,d\Gamma+\int_{\Gamma_R} w(a\phi+b)\,d\Gamma=0 \qquad (70)$$

with

$$\phi = \bar{\phi}, \quad w = 0 \ \text{ on } \Gamma_D \tag{71}$$

The derivation of these forms can be performed similarly as with (21). A large number of physical phenomena are covered by this formulation with different interpretations for $\phi$, some of which having been explained in Appendix A.



## 3.2 TWO-DIMENSIONAL ELEMENTS

The most conventional two-dimensional elements are described. In all cases isoparametric mapping is employed to map the reference element to the global space.

### 3.2.1 Triangular elements

**Three-noded element.** Figure 3.7 shows a *three-noded* or *linear triangular element* (kolmisolmuinen tai lineaarinen kolmioelementti).

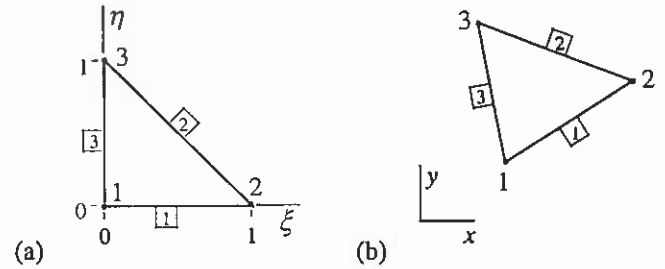



**Figure 3.7 (a)** Linear reference element. **(b)** Linear element in global space.

Figure 3.7 (a) fixes the node and side numbering order (symbol ⊡ ) selected. The numbering order is taken here to grow in the counterclockwise direction. The independent natural coordinates are $\xi \in [0,1]$, $\eta \in [0,1]$. The shape function expressions are

$$\boxed{\begin{aligned} N_1 &= L_1 = 1 - \xi - \eta \\ N_2 &= L_2 = \xi \\ N_3 &= L_3 = \eta \end{aligned}} \tag{1}$$

The formulas include the alternative forms in *area coordinates* (pinta-alakoordinaatti)

$$L_1 = \frac{A_1}{A}, \quad L_2 = \frac{A_2}{A}, \quad L_2 = \frac{A_3}{A} \tag{2}$$

which were referred to in Section 2.2.1. The meaning of the notations should be clear from Figure 3.8. The area coordinates $L_i \in [0,1]$ must satisfy the constraint equation

$$L_1 + L_2 + L_3 = 1 \tag{3}$$

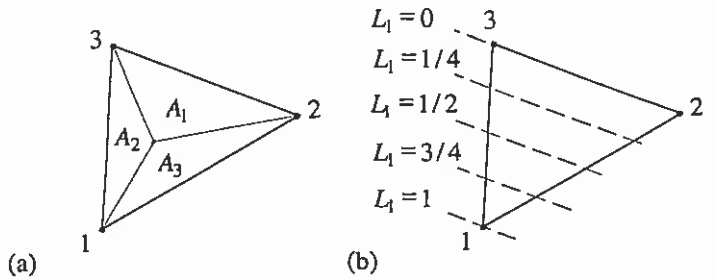



**Figure 3.8 (a)** Notations for area coordinates. **(b)** Some coordinate lines $L_1$ = constant.

Typical shape functions are sketched in Figure 1.4 (b).

The approximation is

$$\phi = \sum_{i=1}^{3} N_i \phi_i = (1-\xi-\eta)\phi_1 + \xi\phi_2 + \eta\phi_3 \tag{4}$$

The isoparametric mapping

$$\begin{aligned} x &= \sum_{i=1}^{3} N_i x_i = (1-\xi-\eta)x_1 + \xi x_2 + \eta x_3 \\ y &= \sum_{i=1}^{3} N_i y_i = (1-\xi-\eta)y_1 + \xi y_2 + \eta y_3 \end{aligned} \tag{5}$$

gives arbitrarily shaped straight sided elements in the $xy$-plane (Figure 3.7 (b)). The nodes 1, 2, 3 must follow in the counterclockwise order determined in Figure 3.7 (a) but the nodal numbering may start from any vertex. It is not considered necessary here to equip the reference element quantities with dashes and similarly the element superscript $e$ can be dropped without confusion.

In what follows we do not any more give the element approximation expressions and the isoparametric mappings as these features should be now quite obvious.

**Six-noded element.** Figure 3.9 shows a *six-noded* or *quadratic triangular element* (kuusisolmuinen tai kvadraattinen kolmioelementti). The midside nodes in the reference element are at the midpoints of the sides. The independent natural coordinates $\xi$ and $\eta$ have the same ranges and the side numbering is also the same as for the linear element.



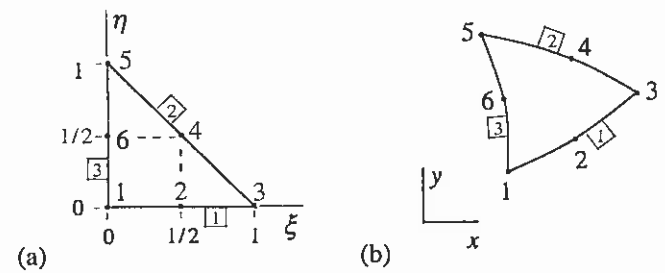



**Figure 3.9 (a)** Quadratic reference element. **(b)** Quadratic element in global space.

The shape function expressions are

$$\begin{aligned} N_1 &= (2L_1-1)L_1 = 1-3\xi-3\eta+2\xi^2+4\xi\eta+2\eta^2 \\ N_2 &= 4L_1L_2 = 4\xi-4\xi^2-4\xi\eta \\ N_3 &= (2L_2-1)L_2 = -\xi+2\xi^2 \\ N_4 &= 4L_2L_3 = 4\xi\eta \\ N_5 &= (2L_3-1)L_3 = -\eta+2\eta^2 \\ N_6 &= (2L_1-1)L_1 = 4\eta-4\eta^2-4\xi\eta \end{aligned} \tag{6}$$

The forms employing area coordinates associate the variables $L_1$, $L_2$, $L_3$ with the vertex numbers 1, 3, 5, respectively. These formulas are rather easy to write down after some practice directly by inspection without any serious calculations. The main idea is to make use of the following obvious fact: *an expression containing two or more product factors is zero at all those points where the factors are separately zero.* In addition, the level lines of constant area coordinates values such as shown in Figure 3.8 (b) are of help. For instance, the factor $L_1$ is zero at the nodes 3, 4, 5 and the factor $L_1 - 1/2$ is zero at the nodes 2 and 6. Thus the product $L_1(L_1-1/2)$ is zero at all nodes except at node 1. By now multiplying this product by a suitable scalar factor so that it gets the value 1 at node 1 we have obtained the first shape function (6).

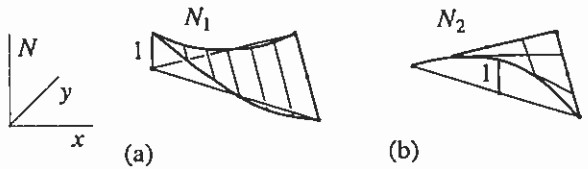


**Figure 3.10 (a)** Shape function for a corner node. **(b)** Shape function for a midside node.

The isoparametric mapping can produce elements in the global space having curved sides (Figure 3.9 (b)). Two typical shape functions are sketched in Figure 3.10.

**Remark 3.4.** Triangular elements and corresponding closed form shape functions can be generated in principle for any complete polynomial degree in $\xi$ and $\eta$. The next third degree or cubic element following the quadratic one has ten nodes one of them situating already inside the element at the point $L_1 = 1/3$, $L_2 = 1/3$, $L_3 = 1/3$. There is, however, usually in practice no need to go for these high degree elements. This is the case especially if the solution has boundary or internal layer type behavior. At these non-smooth solution areas it is roughly saying better to have many crude small elements than few large refined elements. □

### 3.2.2 Quadrilateral elements

**Four-noded element.** Figure 3.11 shows a *four-noded* or *bilinear quadrilateral element* (nelisolmuinen tai bilineaarinen nelikulmioelementti).

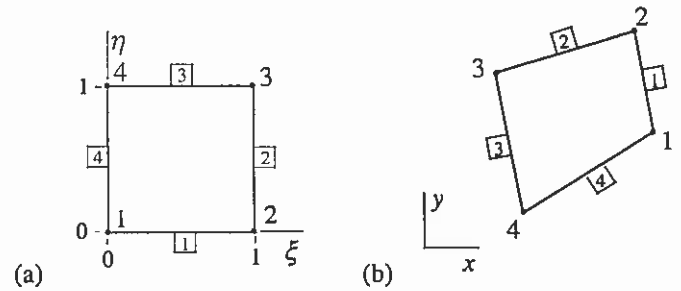



**Figure 3.11 (a)** Bilinear reference element. **(b)** Bilinear element in global space.

Figure 3.11 (a) fixes the node and side numbering order selected. The independent natural coordinates are again $\xi \in [0,1]$, $\eta \in [0,1]$. The shape function expressions are

$$\begin{aligned} N_1 &= (1-\xi)(1-\eta) \\ N_2 &= \xi(1-\eta) \\ N_3 &= \xi\eta \\ N_4 &= (1-\xi)\eta \end{aligned} \tag{7}$$



They are obtained by multiplying together one-dimensional linear shape functions treated in Section 2.2.1. These product shape functions are sometimes called tensor product shape functions. The name bilinear comes from the property that the functions are linear in $\xi$ or in $\eta$ at the lines $\eta$ = constant or $\xi$ = constant, respectively. This terminology has nothing to do with the concept bilinear form introduced in Appendix C.

The isoparametric mapping gives arbitrary shaped straight-sided quadrilaterals in the global space (Figure 3.11 (b)). One typical shape function is sketched in Figure 3.12.

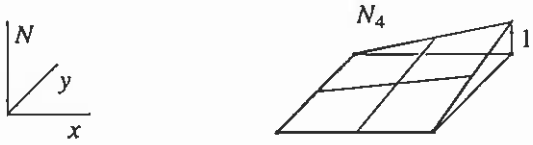



**Figure 3.12** Bilinear shape function.

**Nine-noded element.** Figure 3.13 shows a *nine-noded* or *biquadratic quadrilateral element* (yhdeksänsolmuinen tai bikvadraattinen nelikulmioelementti).

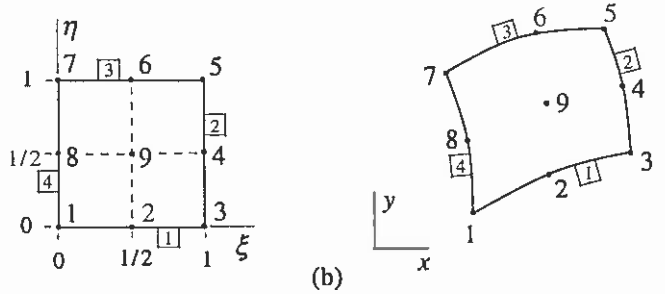



**Figure 3.13 (a)** Biquadratic reference element. **(b)** Biquadratic element in global space.

Figure 3.13 (a) fixes the node and side numbering order selected. The midside nodes and the middle node in the reference element are at the midpoints of the sides and at the midpoint of the element, respectively. The independent natural coordinates $\xi$ and $\eta$ have the same ranges and the side numbering is also the same as for the bilinear element. The shape function expressions are

$$
\begin{aligned}
N_1 &= \left(1-3\xi+2\xi^2\right)\left(1-3\eta+2\eta^2\right)\\
N_2 &= \left(4\xi-4\xi^2\right)\left(1-3\eta+2\eta^2\right)\\
N_3 &= \left(-\xi+2\xi^2\right)\left(1-3\eta+2\eta^2\right)\\
N_4 &= \left(-\xi+2\xi^2\right)\left(4\eta-4\eta^2\right)\\
N_5 &= \left(-\xi+2\xi^2\right)\left(-\eta+2\eta^2\right)\\
N_6 &= \left(4\xi-4\xi^2\right)\left(-\eta+2\eta^2\right)\\
N_7 &= \left(1-3\xi+2\xi^2\right)\left(-\eta+2\eta^2\right)\\
N_8 &= \left(1-3\xi+2\xi^2\right)\left(4\eta-4\eta^2\right)\\
N_9 &= (4\xi-4\xi^2)(4\eta-4\eta^2)
\end{aligned}
\tag{8}
$$

They are obtained by multiplying together one-dimensional quadratic shape functions given in Section 2.2.2. The terminology is explained by analogy to the bilinear case.

The isoparametric mapping can produce curved-sided quadrilaterals in the global space (Figure 3.13 (b)). Three typical shape functions are sketched in Figure 3.14.

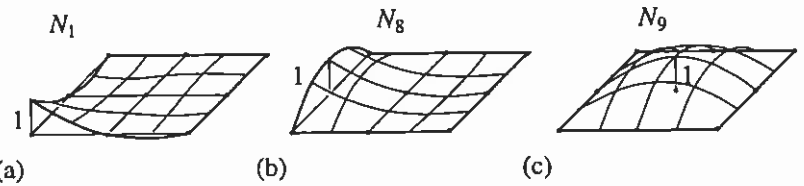



**Figure 3.14** **(a)** Shape function for a corner node. **(b)** Shape function for a midside node. **(c)** Shape function for the middle node.

**Remark 3.5.** The ninth node (Figure 3.13) is missing in the rather popular *eight-noded Serendipity quadrilateral element*; see e.g. Zienkiewicz and Taylor (2000). We do not consider it here. (Serendipity means roughly the faculty of making important discoveries by chance.) The bilinear and biquadratic quadrilateral elements are called sometimes as Lagrangian elements to discern them from the Serendipity elements. □

### 3.2.3 Element properties



All the elements described above are of the simple conventional $C^0$ continuous type (shortly, $C^0$ elements), that is, the finite element approximation in a mesh of these elements is continuous but the derivatives with respect to the global coordinates are no more continuous at the element boundaries. Refined $C^1$ (meaning that also the first derivatives are continuous) elements in two dimensions have been developed with considerable pain. They have been used mainly in structural mechanics in the analysis of bending of plates and will not be considered in this text. In Section 5.2.3 we will deal shortly with a $C^1$ element in one dimension.

The $C^0$ continuity of the elements described here is based simply on the fact that the approximation in a natural coordinate is a first or second degree polynomial along an element side in a linear or quadratic element, respectively. Two or three nodes, respectively, on a side with identical nodal values from the neighboring elements fix the neighboring polynomials to be identical so there are no jumps in the approximation along the element interfaces.

Linear triangular elements and bilinear quadrilateral elements can be mixed in a mesh and similarly with the quadratic triangles and biquadratic quadrilaterals. So-called transition elements, missing some of the midside nodes, can be rather easily devised. With them linear and quadratic elements can be mixed in a mesh.

The element shape functions have some interesting properties in addition to their 1- or 0- value behavior at the nodes discussed earlier. Some convergence criteria demand — and common sense would consider it favorable — that an element approximation should be able to represent a constant function and also a linear function in the global coordinates exactly (at least in the limit when the element size gets smaller and smaller).

Let us consider the consequences of the above criteria. We assume a given linear polynomial

$$\bar{\phi}(x,y) = \alpha + \beta x + \gamma y \tag{9}$$

where the coefficients $\alpha$, $\beta$, $\gamma$ can be arbitrary. This function gives the nodal values ($x_i$, $y_i$) are the global coordinates of element node $i$)

$$\phi_i = \bar{\phi}\left(x_i, y_i\right) = \alpha + \beta x_i + \gamma y_i \tag{10}$$

From these nodal values follows the element approximation

$$\phi(x,y) = \sum_i N_i \phi_i = \sum_i N_i \left(\alpha + \beta x_i + \gamma y_i\right)$$

$$
\begin{aligned}
&= \sum_i N_i \alpha + \sum_i N_i \beta x_i + \sum_i N_i \gamma y_i \\
&= \alpha \sum_i N_i + \beta \sum_i N_i x_i + \gamma \sum_i N_i y_i
\end{aligned} \tag{11}
$$

For this to represent expression (9) exactly, the conditions

$$
\boxed{\begin{aligned}
&\sum_i N_i = 1 \\
&\sum_i N_i x_i = x \\
&\sum_i N_i y_i = y
\end{aligned}} \tag{12}
$$

must be satisfied. The first condition (12) can be seen to be valid for the elements described in this text by direct summation from the given expressions. But the two latter conditions are just the formulas used in the isoparametric mapping and they are thus also satisfied. This is one further indication of the advantageous properties of isoparametric elements.

### 3.2.4 Global derivatives

Derivatives of shape functions with respect to the global coordinates $x$ and $y$ appear in the element contribution expressions. The shape functions are, however, represented in the natural coordinates $\xi$ and $\eta$: $N_i = N_i(\xi,\eta)$. The isoparametric mapping is of the form $x = x(\xi,\eta)$, $y = y(\xi,\eta)$. The inverse mapping $\xi = \xi(x,y)$, $\eta = \eta(x,y)$ is needed in principle but to find that in closed form with general geometries would lead to extremely complicated expressions. The way to proceed in practice is the following. Chain differentiation gives

$$
\begin{aligned}
\frac{\partial N_i}{\partial \xi} &= \frac{\partial N_i}{\partial x}\frac{\partial x}{\partial \xi} + \frac{\partial N_i}{\partial y}\frac{\partial y}{\partial \xi} \\
\frac{\partial N_i}{\partial \eta} &= \frac{\partial N_i}{\partial x}\frac{\partial x}{\partial \eta} + \frac{\partial N_i}{\partial y}\frac{\partial y}{\partial \eta}
\end{aligned} \tag{13}
$$

or in matrix notation

$$
\begin{Bmatrix} \dfrac{\partial N_i}{\partial \xi} \\ \dfrac{\partial N_i}{\partial \eta} \end{Bmatrix} = \begin{bmatrix} \dfrac{\partial x}{\partial \xi} & \dfrac{\partial y}{\partial \xi} \\ \dfrac{\partial x}{\partial \eta} & \dfrac{\partial y}{\partial \eta} \end{bmatrix} \begin{Bmatrix} \dfrac{\partial N_i}{\partial x} \\ \dfrac{\partial N_i}{\partial y} \end{Bmatrix} \equiv [J]^{\mathrm{T}} \begin{Bmatrix} \dfrac{\partial N_i}{\partial x} \\ \dfrac{\partial N_i}{\partial y} \end{Bmatrix} \tag{14}
$$

It is easy to differentiate the shape functions with respect to the natural coordinates and similarly the elements $\partial x/\partial\xi$ ... of the *Jacobian matrix* (Jacobin matriisi) $[J]$ can be easily evaluated from the isoparametric mappings. The global unknown derivatives can now be solved from the system (14):

$$\begin{Bmatrix} \dfrac{\partial N_i}{\partial x} \\ \dfrac{\partial N_i}{\partial y} \end{Bmatrix} = [J]^{-\mathrm{T}} \begin{Bmatrix} \dfrac{\partial N_i}{\partial \xi} \\ \dfrac{\partial N_i}{\partial \eta} \end{Bmatrix} \tag{15}$$

The complete form of the Jacobian matrix is

$$[J] \equiv \left[\frac{\partial(x,y)}{\partial(\xi,\eta)}\right] = \begin{bmatrix} \dfrac{\partial x}{\partial \xi} & \dfrac{\partial x}{\partial \eta} \\ \dfrac{\partial y}{\partial \xi} & \dfrac{\partial y}{\partial \eta} \end{bmatrix} = \begin{bmatrix} \sum \dfrac{\partial N_i}{\partial \xi} x_i & \sum \dfrac{\partial N_i}{\partial \eta} x_i \\ \sum \dfrac{\partial N_i}{\partial \xi} y_i & \sum \dfrac{\partial N_i}{\partial \eta} y_i \end{bmatrix} \tag{16}$$

Given a point $\mathrm{P}' : (\xi, \eta)$, the corresponding global derivatives can be evaluated from (15). This can be done for any point, which is enough when numerical integration is used.

**Remark 3.6.** The Jacobian matrix is defined often as the transpose of the one given here in which case the superscript T does not appear in (14) and (15). This does not affect the value of the determinant $\det[J]$ often called just the *Jacobian*. The definition in the beginning of (16) seems to be the normal one in mathematics texts. As $([J]^{\mathrm{T}})^{-1} = ([J]^{-1})^{\mathrm{T}}$, we have employed for this matrix for simplicity the notation appearing in (15). □

**Remark 3.7.** Making the notational changes $x \to x_1$, $y \to x_2$, $\xi \to x_{1'}$, $\eta \to x_{2'}$ and employing the summation convention gives a compact form of (14):

$$\frac{\partial N_i}{\partial x_{\beta'}} = J_{\alpha\beta'} \frac{\partial N_i}{\partial x_\alpha} \tag{17}$$

where the elements of the Jacobian matrix are

$$J_{\alpha\beta'} \equiv \frac{\partial x_\alpha}{\partial x_{\beta'}} \tag{18}$$

The inverse form of (17) corresponding to (15) is

$$\frac{\partial N_i}{\partial x_\alpha} = J_{\beta'\alpha} \frac{\partial N_i}{\partial x_{\beta'}} \tag{19}$$

where



$$J_{\beta'\alpha} \equiv \frac{\partial x_{\beta'}}{\partial x_\alpha} \tag{20}$$

As indicated above, the elements $J_{\beta'\alpha}$ can, however, be evaluated in general only in a pointwise manner.

These forms are valid also in three dimensions just by letting the range of $\alpha$ and $\beta$ extend from 1 and 2 to 3. □

**Example 3.2.** We consider the four-noded isoparametric quadrilateral element in Figure (a).

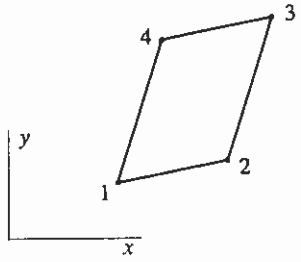



**Figure (a)**

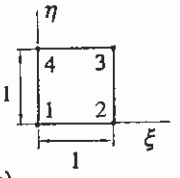



**Figure (b)**

The nodal coordinates are

$$\begin{aligned} x_1 &= 1\cdot a, & x_2 &= 2\cdot a, & x_3 &= 2.4\cdot a, & x_4 &= 1.4\cdot a \\ y_1 &= 0.5\cdot a, & y_2 &= 0.7\cdot a, & y_3 &= 2.1\cdot a, & y_4 &= 1.9\cdot a \end{aligned} \tag{a}$$

where $a$ is a measure having the dimension of length. We evaluate the global derivatives of the element shape functions.

The shape functions (7) are

$$\begin{aligned} N_1 &= (1-\xi)(1-\eta) = 1-\xi-\eta+\xi\eta \\ N_2 &= \xi(1-\eta) = \xi-\xi\eta \\ N_3 &= \xi\eta \\ N_4 &= (1-\xi)\eta = \eta-\xi\eta \end{aligned} \tag{b}$$

The isoparametric mapping from the reference element in Figure (b) gives the expressions

$$\begin{aligned} x &= (1-\xi-\eta+\xi\eta)x_1 + (\xi-\xi\eta)x_2 + \xi\eta\, x_3 + (\eta-\xi\eta)x_4 \\ &= (1-\xi-\eta+\xi\eta)a + (\xi-\xi\eta)2a + \xi\eta\, 2.4a + (\eta-\xi\eta)1.4a \\ &= a(1+\xi+0.4\eta) \end{aligned} \tag{c}$$

$$y=(1-\xi-\eta+\xi\eta)y_1+(\xi-\xi\eta)y_2+\xi\eta\, y_3+(\eta-\xi\eta)y_4$$
$$=(1-\xi-\eta+\xi\eta)0.5a+(\xi-\xi\eta)0.7a+\xi\eta\, 2.1a+(\eta-\xi\eta)1.9a$$
$$=a(0.5+0.2\xi+1.4\eta)$$

The nodal coordinates have been selected so that the element is a parallelogram. This simplifies the expressions so that the term $\xi\eta$ in (c) is missing and the elements of the Jacobian matrix become constants. We obtain

$$\frac{\partial x}{\partial \xi}=a, \quad \frac{\partial x}{\partial \eta}=0.4a$$
$$\frac{\partial y}{\partial \xi}=0.2a, \quad \frac{\partial y}{\partial \eta}=1.4a \tag{d}$$

The inverse of the Jacobian matrix

$$[J]=\begin{bmatrix} a & 0.4a \\ 0.2a & 1.4a \end{bmatrix}=\begin{bmatrix} 1 & 0.4 \\ 0.2 & 1.4 \end{bmatrix}a \tag{e}$$

is (found easily e.g. using Cramer's formula)

$$[J]^{-1}=\frac{1}{1.32a}\begin{bmatrix} 1.4 & -0.4 \\ -0.2 & 1 \end{bmatrix} \tag{f}$$

Formula (15) gives thus here

$$\begin{Bmatrix} \frac{\partial N_i}{\partial x} \\ \frac{\partial N_i}{\partial y} \end{Bmatrix}=\frac{1}{1.32a}\begin{bmatrix} 1.4 & -0.2 \\ -0.4 & 1 \end{bmatrix}\begin{Bmatrix} \frac{\partial N_i}{\partial \xi} \\ \frac{\partial N_i}{\partial \eta} \end{Bmatrix} \tag{g}$$

or in more detail

$$\frac{\partial N_i}{\partial x}=\frac{1}{1.32a}\left(1.4\frac{\partial N_i}{\partial \xi}-0.2\frac{\partial N_i}{\partial \eta}\right)$$
$$\frac{\partial N_i}{\partial y}=\frac{1}{1.32a}\left(-0.4\frac{\partial N_i}{\partial \xi}+1\frac{\partial N_i}{\partial \eta}\right) \tag{h}$$

For instance,

$$\frac{\partial N_1}{\partial x}=\frac{1}{1.32a}\left[1.4(-1+\eta)-0.2(-1+\xi)\right]$$
$$\frac{\partial N_1}{\partial y}=\frac{1}{1.32a}\left[-0.4(-1+\eta)+1(-1+\xi)\right] \tag{i}$$

and the rest of the derivatives are obtained similarly. For a more general geometry the step between (e) and (f) would become analytically cumbersome.



## 3.3 FINITE ELEMENT SOLUTION

### 3.3.1 Discretization

A two-dimensional basic energy equation weak form was given in Section 3.1 as (3.1.28):

$$-\int_A\left(\frac{\partial w}{\partial x}q_x+\frac{\partial w}{\partial y}q_y\right)\mathrm{d}A-\int_A ws\,\mathrm{d}A$$
$$+\int_{s_D} wq\,\mathrm{d}s+\int_{s_N} w\overline{q}\,\mathrm{d}s+\int_{s_R} wh(T-T_\infty)\mathrm{d}s=0 \tag{1}$$

The constitutive law for a thermally isotropic material is (3.1.29):

$$q_x=-k\frac{\partial T}{\partial x}, \quad q_y=-k\frac{\partial T}{\partial y} \tag{2}$$

Figure 3.15 shows some of the relevant notations.

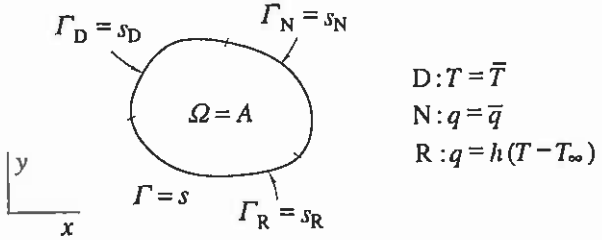



**Figure 3.15** Two-dimensional plane case.

Combining (1) and (2), taking the conventional restrictions on the Dirichlet boundary and using the general domain and boundary symbols gives the *standard weak form*

$$\boxed{\begin{aligned}&\int_\Omega\left(\frac{\partial w}{\partial x}k\frac{\partial T}{\partial x}+\frac{\partial w}{\partial y}k\frac{\partial T}{\partial y}\right)\mathrm{d}\Omega+\int_{\Gamma_R} whT\,\mathrm{d}\Gamma\\ &-\int_\Omega ws\,\mathrm{d}\Omega+\int_{\Gamma_N} w\overline{q}\,\mathrm{d}\Gamma-\int_{\Gamma_R} whT_\infty\mathrm{d}\Gamma=0\end{aligned}} \tag{3}$$

with

$$T=\overline{T}, \quad w=0 \ \text{ on } \Gamma_D \tag{4}$$

This may be compared with the corresponding one-dimensional form (2.1.28). Now partial derivatives appear, the boundary terms are integrals and the treatment has been expanded by the Robin boundary. The first line in (3) contains the unknown function $T(x,y)$ to be determined and the second line consists of given quantities.

Application of the finite element Galerkin method proceeds quite the same way as in one dimension. The finite element approximation is

$$\boxed{\bar{T}(x,y)=\sum_{j=1}^{n_n} N_j(x,y)T_j} \tag{5}$$

and the system equations are

$$\begin{aligned} F_i \equiv & \int_\Omega \left( \frac{\partial N_i}{\partial x} k \frac{\partial \bar{T}}{\partial x} + \frac{\partial N_i}{\partial y} k \frac{\partial \bar{T}}{\partial y} \right) \mathrm{d}\Omega + \int_{\Gamma_R} N_i h \bar{T} \, \mathrm{d}\Gamma \\ & - \int_\Omega N_i s \, \mathrm{d}\Omega + \int_{\Gamma_N} N_i \bar{q} \, \mathrm{d}\Gamma - \int_{\Gamma_R} N_i h T_\infty \, \mathrm{d}\Gamma = 0, \quad i = 1,2,\cdots,n_n \end{aligned} \tag{6}$$

Substitution of (5) and similar manipulations as in one dimension gives the final detailed system equations

$$\boxed{\begin{aligned} F_i \equiv & \sum_{j=1}^{n_n} \left[ \int_\Omega \left( \frac{\partial N_i}{\partial x} k \frac{\partial N_j}{\partial x} + \frac{\partial N_i}{\partial y} k \frac{\partial N_j}{\partial y} \right) \mathrm{d}\Omega + \int_{\Gamma_R} N_i h N_j \mathrm{d}\Gamma \right] T_j \\ & - \int_\Omega N_i s \, \mathrm{d}\Omega + \int_{\Gamma_N} N_i \bar{q} \, \mathrm{d}\Gamma - \int_{\Gamma_R} N_i h T_\infty \, \mathrm{d}\Gamma = 0 \end{aligned}} \tag{7}$$

This is of course still a linear system of equations:

$$\boxed{[K]\{a\}=\{b\}} \tag{8}$$

and the matrix elements are now

$$\begin{aligned} K_{ij} &= \int_\Omega \left( \frac{\partial N_i}{\partial x} k \frac{\partial N_j}{\partial x} + \frac{\partial N_i}{\partial y} k \frac{\partial N_j}{\partial y} \right) \mathrm{d}\Omega + \int_{\Gamma_R} N_i h N_j \mathrm{d}\Gamma \\ b_i &= \int_\Omega N_i s \, \mathrm{d}\Omega - \int_{\Gamma_N} N_i \bar{q} \, \mathrm{d}\Gamma + \int_{\Gamma_R} N_i h T_\infty \, \mathrm{d}\Gamma \end{aligned} \tag{9}$$

The coefficient matrix is again symmetric. The Robin condition is seen to give terms also to the coefficient matrix. The preliminary system equations must finally be processed with respect to the given nodal temperatures from the Dirichlet boundary; see Remarks 2.9 and 2.13.



**Remark 3.8.** If the Dirichlet boundary condition is taken into account according to the alternative procedure described in Remark 2.15, the deltaform, we write first

$$T(x,y)=\bar{T}(x,y)+\Delta T(x,y) \tag{10}$$

where $\bar{T}|_{\Gamma_D}=\bar{T}$ and substitute this into the weak form (3) to give the alternative form

$$\begin{aligned} & \int_\Omega \left( \frac{\partial w}{\partial x} k \frac{\partial \Delta T}{\partial x} + \frac{\partial w}{\partial y} k \frac{\partial \Delta T}{\partial y} \right) \mathrm{d}\Omega + \int_{\Gamma_R} wh\Delta T \, \mathrm{d}\Gamma \\ & + \int_\Omega \left( \frac{\partial w}{\partial x} k \frac{\partial \bar{T}}{\partial x} + \frac{\partial w}{\partial y} k \frac{\partial \bar{T}}{\partial y} \right) \mathrm{d}\Omega + \int_{\Gamma_R} wh\bar{T} \, \mathrm{d}\Gamma \\ & - \int_\Omega ws \, \mathrm{d}\Omega + \int_{\Gamma_N} w\bar{q} \, \mathrm{d}\Gamma - \int_{\Gamma_R} whT_\infty \mathrm{d}\Gamma = 0 \end{aligned} \tag{11}$$

with

$$\Delta T = 0, \quad w = 0 \quad \text{on} \quad \Gamma_D \tag{12}$$

The approximation for the new unknown function $\Delta T(x,y)$ is

$$\Delta\bar{T}(x,y)=\sum_{j=1}^{n_n} N_j(x,y)\Delta T_j \tag{13}$$

The matrix elements of the resulting system equations

$$[K]\{\Delta a\}=\{b\} \tag{14}$$

are easily found as a modification of the second formula (9). The first formula (9) remains unchanged. □

### 3.3.2 Assembly process

The element contributions can be written down immediately (see Remark 2.11) from (9):

$$\begin{aligned} K_{ij}^e &= \int_{\Omega^e} \left( \frac{\partial N_i^e}{\partial x} k \frac{\partial N_j^e}{\partial x} + \frac{\partial N_i^e}{\partial y} k \frac{\partial N_j^e}{\partial y} \right) \mathrm{d}\Omega + \int_{\Gamma_R^e} N_i^e h N_j^e \mathrm{d}\Gamma \\ b_i^e &= \int_{\Omega^e} N_i^e s \, \mathrm{d}\Omega - \int_{\Gamma_N^e} N_i^e \bar{q} \, \mathrm{d}\Gamma + \int_{\Gamma_R^e} N_i^e h T_\infty \, \mathrm{d}\Gamma \end{aligned} \tag{15}$$

The assembly process remains naturally the same as in one dimension (see (2.3.39)):

$$K_{ij} = \sum_{e=1}^{n_e} K_{rs}^e, \quad b_i = \sum_{e=1}^{n_e} b_r^e \tag{16}$$

**Remark 3.9.** This far we have called the discrete unknowns in the finite element method as nodal values (solmuarvo) and the indices in formulas such as (16) have referred to global and local node numbers (solmunumero). This has been appropriate, as we have had only one unknown quantity per node, the nodal value of the temperature. In more general situations we may have several unknowns per node, say two velocity components, the pressure etc. The discrete unknowns are numbered starting from number one (usually in some fashion following the nodal numbering both for the mesh and for an element) and we shall call them here in general as *nodal parameters* (solmuparametri). A much-used synonym in the literature is *degree of freedom* (vapausaste) but this is not very pertinent, as the proper term from classical mechanics would be generalized coordinate. The nodal parameters can be classified again in an obvious way as global and local ones. *The assembly process implied by formulas (16) and explained earlier stay valid if we just replace in the interpretations the words global and local node numbers with global and local nodal parameters, respectively.* It should be remarked that so-called nodeless discrete unknowns are sometimes also used in the finite element method, especially with *hierarchical elements* (hierarkkinen elementti). For bookkeeping purposes these variables can, however, always be associated with certain artificial nodes and thus the term nodal parameter can still be used. □

The assembly and some applications of the formulas developed above are now explained in connection with a simple demonstration case; Example 3.3. The finite element method is meaningful only when used with computers. However, some practice with hand calculations is necessary to become comfortable with the concepts introduced by the theory. Even with very simple example cases the demonstration hand calculations inevitably tend to become rather heavy.

**Example 3.3.** Two-dimensional plane heat conduction according to the theory presented in Section 3.3.1 is considered. The problem domain is shown in Figure (a) and the finite element mesh in Figure (b). The mesh consists of two ($n_e = 2$) identical square bilinear elements ($n_n^1 = n_n^2 = 4$). The total number of nodes is eight ($n_n = 8$). The task is to generate the element contributions and to assemble and solve the system equations.

The boundary conditions are

$$\begin{aligned} &T = \overline{T} = \frac{x}{L}\overline{T}_B && \text{on } \Gamma_D \ (y = 0) \\ &T = \overline{T} = \left(1 - \frac{x}{L}\right)\overline{T}_D && \text{on } \Gamma_D \ (y = 2L) \\ &q = \overline{q} && \text{on } \Gamma_N \\ &q = h(T - T_\infty) && \text{on } \Gamma_R \end{aligned} \tag{1}$$

The given quantities $k$, $s$, $\overline{q}$, $h$, $T_\infty$ are assumed to be constants for simplicity. $\overline{T}_B$ and $\overline{T}_D$ are given temperatures at points B and D.



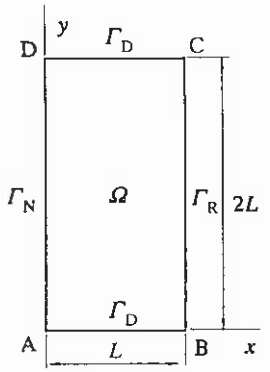



**Figure (a)**

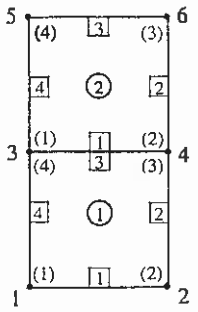



**Figure (b)**

The finite element contributions are (formulas (15))

$$\begin{aligned} K_{ij}^e &= \int_{\Omega^e} \left( \frac{\partial N_i^e}{\partial x} k \frac{\partial N_j^e}{\partial x} + \frac{\partial N_i^e}{\partial y} k \frac{\partial N_j^e}{\partial y} \right) d\Omega + \int_{\Gamma_R^e} N_i^e h N_j^e d\Gamma \\ b_i^e &= \int_{\Omega^e} N_i^e s \, d\Omega - \int_{\Gamma_N^e} N_i^e \overline{q} \, d\Gamma + \int_{\Gamma_R^e} N_i^e h T_\infty d\Gamma \end{aligned} \tag{2}$$

We first derive some results for the bilinear element.

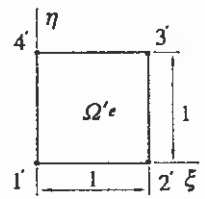



**Figure (c)**

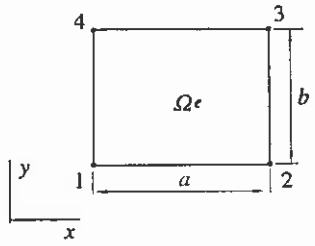



**Figure (d)**

Figure (c) shows the reference element (see Figure 3.11) and Figure (d) the element in the global space with a simple rectangular geometry the sides of the element aligned with the coordinate lines.

The element shape functions (3.2.7) are (for simplicity of presentation, the element number superscript $e$ can be dropped here mostly in what follows without danger of confusion)

$$
\begin{aligned}
N_1 &= (1-\xi)(1-\eta) = 1-\xi-\eta+\xi\eta \\
N_2 &= \xi(1-\eta) = \xi-\xi\eta \\
N_3 &= \xi\eta \\
N_4 &= (1-\xi)\eta = \eta-\xi\eta
\end{aligned}
\tag{3}
$$

The isoparametric mapping

$$
\begin{aligned}
x &= (1-\xi-\eta+\xi\eta)x_1 + (\xi-\xi\eta)x_2 + \xi\eta\, x_3 + (\eta-\xi\eta)x_4 \\
y &= (1-\xi-\eta+\xi\eta)y_1 + (\xi-\xi\eta)y_2 + \xi\eta\, y_3 + (\eta-\xi\eta)y_4
\end{aligned}
\tag{4}
$$

gives here using the data of Figure (d)

$$
\begin{aligned}
&x_2 = x_1 + a, \quad x_3 = x_1 + a, \quad x_4 = x_1, \\
&y_2 = y_1, \quad y_3 = y_1 + b, \quad y_4 = y_1 + b
\end{aligned}
\tag{5}
$$

simply

$$
x = x_1 + \xi a, \quad y = y_1 + \eta b \tag{6}
$$

In this case this is trivial to invert:

$$
\xi = \frac{x-x_1}{a}, \quad \eta = \frac{y-y_1}{b} \tag{7}
$$

The Jacobian matrix (3.2.16)

$$
[J] = \begin{bmatrix} \partial x/\partial\xi & \partial x/\partial\eta \\ \partial y/\partial\xi & \partial y/\partial\eta \end{bmatrix} = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \tag{8}
$$

and the magnification factor (see Section E.1.2)

$$
M = \det[J] = ab \tag{9}
$$

Thus

$$
\int_{\Omega^e} f\,\mathrm{d}\Omega = \int_{\Omega'^e} f \det[J]\,\mathrm{d}\Omega' = ab\int_{\Omega'^e} f\,\mathrm{d}\Omega' = ab\int_0^1\int_0^1 f(\xi,\eta)\mathrm{d}\xi\mathrm{d}\eta \tag{10}
$$

We continue to apply the general formulas of Section 3.2.4 for demonstration purposes. We have

$$
[J]^{-\mathrm{T}} = \left(\begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}^{-1}\right)^{\mathrm{T}} = \left(\begin{bmatrix} 1/a & 0 \\ 0 & 1/b \end{bmatrix}\right)^{\mathrm{T}} = \begin{bmatrix} 1/a & 0 \\ 0 & 1/b \end{bmatrix} \tag{11}
$$

and from (3.2.15)

$$
\begin{Bmatrix} \partial N_i/\partial x \\ \partial N_i/\partial y \end{Bmatrix} = \begin{bmatrix} 1/a & 0 \\ 0 & 1/b \end{bmatrix} \begin{Bmatrix} \partial N_i/\partial\xi \\ \partial N_i/\partial\eta \end{Bmatrix} \tag{12}
$$

or simply



$$
\frac{\partial N_i}{\partial x} = \frac{1}{a}\frac{\partial N_i}{\partial\xi}, \quad \frac{\partial N_i}{\partial y} = \frac{1}{b}\frac{\partial N_i}{\partial\eta} \tag{13}
$$

The element shape function derivatives with respect to the natural coordinates are

$$
\begin{aligned}
&\frac{\partial N_1}{\partial\xi} = -1+\eta, \quad &&\frac{\partial N_1}{\partial\eta} = -1+\xi, \\
&\frac{\partial N_2}{\partial\xi} = 1-\eta, \quad &&\frac{\partial N_2}{\partial\eta} = -\xi, \\
&\frac{\partial N_3}{\partial\xi} = \eta, \quad &&\frac{\partial N_3}{\partial\eta} = \xi, \\
&\frac{\partial N_4}{\partial\xi} = -\eta, \quad &&\frac{\partial N_4}{\partial\eta} = 1-\xi
\end{aligned}
\tag{14}
$$

We can now start to evaluate the terms in (2).

The term

$$
\begin{aligned}
\left(K_{ij}^e\right)_k &\equiv \int_{\Omega^e}\left(\frac{\partial N_i}{\partial x}k\frac{\partial N_j}{\partial x} + \frac{\partial N_i}{\partial y}k\frac{\partial N_j}{\partial y}\right)\mathrm{d}\Omega \\
&= kab\int_{\Omega'^e}\left(\frac{1}{a}\frac{\partial N_i}{\partial\xi}k\frac{1}{a}\frac{\partial N_j}{\partial\xi} + \frac{1}{b}\frac{\partial N_i}{\partial\eta}k\frac{1}{b}\frac{\partial N_j}{\partial\eta}\right)\mathrm{d}\Omega' \\
&= k\frac{b}{a}\int_0^1\int_0^1\frac{\partial N_i}{\partial\xi}\frac{\partial N_j}{\partial\xi}\mathrm{d}\xi\mathrm{d}\eta + k\frac{a}{b}\int_0^1\int_0^1\frac{\partial N_i}{\partial\eta}\frac{\partial N_j}{\partial\eta}\mathrm{d}\xi\mathrm{d}\eta
\end{aligned}
\tag{15}
$$

Let us evaluate as an example the term $(K_{12}^e)_k$:

$$
\begin{aligned}
\left(K_{12}^e\right)_k &= k\frac{b}{a}\int_0^1\int_0^1\left(-1+2\eta-\eta^2\right)\mathrm{d}\xi\mathrm{d}\eta + k\frac{a}{b}\int_0^1\int_0^1\left(\xi-\xi^2\right)\mathrm{d}\xi\mathrm{d}\eta \\
&= k\frac{b}{a}1\left|_0^1\left(-\eta+\eta^2-\frac{1}{3}\eta^3\right) + k\frac{a}{b}1\right|_0^1\left(\frac{1}{2}\xi^2-\frac{1}{3}\xi^3\right) = -\frac{1}{3}k\frac{b}{a} + \frac{1}{6}k\frac{a}{b}
\end{aligned}
\tag{16}
$$

These type of integrals have been collected in formulas (F.2.3) and we have with $a = b$ altogether (using directly the right-hand side on the first line in (15) and taking into account that $k$ is constant)

$$
\begin{aligned}
[K]_k^e &= \frac{k}{6}\begin{bmatrix} 2 & -2 & -1 & 1 \\ -2 & 2 & 1 & -1 \\ -1 & 1 & 2 & -2 \\ 1 & -1 & -2 & 2 \end{bmatrix} + \frac{k}{6}\begin{bmatrix} 2 & 1 & -1 & -2 \\ 1 & 2 & -2 & -1 \\ -1 & -2 & 2 & 1 \\ -2 & -1 & 1 & 2 \end{bmatrix} \\
&= \frac{k}{6}\begin{bmatrix} 4 & -1 & -2 & -1 \\ -1 & 4 & -1 & -2 \\ -2 & -1 & 4 & -1 \\ -1 & -2 & -1 & 4 \end{bmatrix}
\end{aligned}
\tag{17}
$$

In the next term

$$\left(K_{ij}^{e}\right)_{\mathrm{R}} \equiv \int_{\Gamma_{\mathrm{R}}^{e}} N_i h N_j \,\mathrm{d}\Gamma \tag{18}$$

the Robin boundary is for both elements on side 2 (Figure (b)). Only the element shape functions $N_2$ and $N_3$ are non-zero there and further as $\xi = 1$ they have the expressions

$$N_2 = 1-\eta, \qquad N_3 = \eta \tag{19}$$

On the boundary

$$\mathrm{d}\Gamma = \mathrm{d}s = \mathrm{d}y = b\,\mathrm{d}\eta \tag{20}$$

and thus

$$\left(K_{ij}^{e}\right)_{\mathrm{R}} = hb\int_0^1 N_i N_j \,\mathrm{d}\eta \tag{21}$$

The non-zero terms are in detail

$$\begin{aligned}
\left(K_{22}^{e}\right)_{\mathrm{R}} &= hb\int_0^1 \left(1-2\eta+\eta^2\right)\mathrm{d}\eta = \frac{1}{3}hb \\
\left(K_{23}^{e}\right)_{\mathrm{R}} &= \left(K_{32}^{e}\right)_{\mathrm{R}} = hb\int_0^1 \left(\eta-\eta^2\right)\mathrm{d}\eta = \frac{1}{6}hb \\
\left(K_{33}^{e}\right)_{\mathrm{R}} &= hb\int_0^1 \eta^2 \mathrm{d}\eta = \frac{1}{3}hb
\end{aligned} \tag{22}$$

These results could have been picked directly from formulas (F.1.1) as the shape functions on the boundary are those of the two-noded line element. Together we have ($b = L$)

$$\left[K\right]_{\mathrm{R}}^{e} = \frac{hL}{6}\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{23}$$

Column matrix $\{b\}^e$ consists of three separate terms. First

$$\left(b_i^e\right)_s \equiv \int_{\Omega^e} N_i s \,\mathrm{d}\Omega = s\int_{\Omega^e} N_i \,\mathrm{d}\Omega = sab\int_0^1\int_0^1 N_i \,\mathrm{d}\xi\,\mathrm{d}\eta \tag{24}$$

The integrals can be found from (F.2.3) and we obtain

$$\{b\}_s^e = \frac{ab}{4}\begin{Bmatrix}1\\1\\1\\1\end{Bmatrix} = \frac{L^2}{4}\begin{Bmatrix}1\\1\\1\\1\end{Bmatrix} \tag{25}$$

Second, in the terms



$$\left(b_i^e\right)_{\mathrm{N}} \equiv -\int_{\Gamma^e} N_i \overline{q}\,\mathrm{d}\Gamma \tag{26}$$

the Neumann boundary is for both elements on side 4 (Figure (b)). Only the element shape functions $N_1$ and $N_4$ are non-zero there and further as $\xi = 0$ they have the expressions

$$N_1 = 1-\eta, \qquad N_4 = \eta \tag{27}$$

The relations (20) are valid giving

$$\left(b_i^e\right)_{\mathrm{N}} = -\overline{q}b\int_0^1 N_i \mathrm{d}\eta \tag{28}$$

and the non-zero terms are

$$\begin{aligned}
\left(b_1^e\right)_{\mathrm{N}} &= -\overline{q}b\int_0^1 (1-\eta)\mathrm{d}\eta = -\frac{1}{2}\overline{q}b \\
\left(b_4^e\right)_{\mathrm{N}} &= -\overline{q}b\int_0^1 \eta\,\mathrm{d}\eta = -\frac{1}{2}\overline{q}b
\end{aligned} \tag{29}$$

Thus

$$\{b\}_{\mathrm{N}}^{e} = -\frac{\overline{q}L}{2}\begin{Bmatrix}1\\0\\0\\1\end{Bmatrix} \tag{30}$$

Formulas (F.1.1) could have again been made use of.

Third, the contribution from the Robin boundary

$$\left(b_i^e\right)_{\mathrm{R}} \equiv \int_{\Gamma^e} N_i h T_\infty \,\mathrm{d}\Gamma \tag{31}$$

can be treated similarly as (26) and we find

$$\{b\}_{\mathrm{R}}^{e} = \frac{hT_\infty L}{2}\begin{Bmatrix}0\\1\\1\\0\end{Bmatrix} \tag{32}$$

Collecting all the contributions together we obtain the element matrices

$$\left[K\right]^{\mathrm{I}} = \left[K\right]^{2} = \frac{k}{6}\begin{bmatrix} 4 & -1 & -2 & -1 \\ -1 & 4 & -1 & -2 \\ -2 & -1 & 4 & -1 \\ -1 & -2 & -1 & 4 \end{bmatrix} + \frac{hL}{6}\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{33}$$

$$\{b\}^1 = \{b\}^2 = \frac{sL^2}{2}\begin{Bmatrix}1\\1\\1\\1\end{Bmatrix} - \frac{\bar{q}L}{2}\begin{Bmatrix}1\\0\\0\\1\end{Bmatrix} + \frac{hT_\infty L}{2}\begin{Bmatrix}0\\1\\1\\0\end{Bmatrix}$$

We can now start the assembly. We get from from Figure (b) the table

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| ① | 1 | 2 | 4 | 3 |
| ② | 3 | 4 | 6 | 5 |

(34)

It contains similar data as Table 2.2. Using the assembly rules (16) gives the system matrices

$$[K]_{6\times6} = \begin{bmatrix} K_{11}^1 & K_{12}^1 & K_{14}^1 & K_{13}^1 & 0 & 0 \\ K_{21}^1 & K_{22}^1 & K_{24}^1 & K_{23}^1 & 0 & 0 \\ K_{41}^1 & K_{42}^1 & K_{44}^1+K_{11}^2 & K_{43}^1+K_{12}^2 & K_{14}^2 & K_{13}^2 \\ K_{31}^1 & K_{32}^1 & K_{34}^1+K_{21}^2 & K_{33}^1+K_{22}^2 & K_{24}^2 & K_{23}^2 \\ 0 & 0 & K_{41}^2 & K_{42}^2 & K_{44}^2 & K_{43}^2 \\ 0 & 0 & K_{31}^2 & K_{32}^2 & K_{34}^2 & K_{33}^2 \end{bmatrix}$$

(35)

$$[b]_{6\times1} = \begin{Bmatrix} b_1^1 \\ b_2^1 \\ b_4^1+b_1^2 \\ b_3^1+b_2^2 \\ b_4^2 \\ b_3^2 \end{Bmatrix}$$

We write down in detail only the two active equations corresponding to nodes 3 and 4:

$$\begin{aligned} K_{31}T_1 + K_{32}T_2 + K_{33}T_3 + K_{34}T_4 + K_{35}T_5 + K_{36}T_6 &= b_3 \\ K_{41}T_1 + K_{42}T_2 + K_{43}T_3 + K_{44}T_4 + K_{45}T_5 + K_{46}T_6 &= b_4 \end{aligned} \tag{36}$$

The Dirichlet data is taken into account according to Remark 2.9. We have

$$T_1 = 0, \quad T_2 = \bar{T}_B, \quad T_5 = \bar{T}_D, \quad T_6 = 0 \tag{37}$$

As is seen, we have given here preference at points A, D and B, C, where the Dirichlet and the Neumann or Robin boundary parts meet, to the Dirichlet data. The final active equations are thus

$$\begin{aligned} K_{33}T_3 + K_{34}T_4 &= b_3 - K_{32}\bar{T}_B - K_{35}\bar{T}_D \\ K_{43}T_3 + K_{44}T_4 &= b_4 - K_{42}\bar{T}_B - K_{45}\bar{T}_D \end{aligned} \tag{38}$$



where

$$\begin{aligned} K_{33} &= K_{44}^1 + K_{11}^2 = 8\frac{k}{6} \\ K_{34} &= K_{43}^1 + K_{12}^2 = K_{43} = K_{34}^1 + K_{21}^2 = -2\frac{k}{6} \\ K_{44} &= K_{33}^1 + K_{22}^2 = 8\frac{k}{6} + 4\frac{hL}{6} \\ K_{32} &= K_{42}^1 = -2\frac{k}{6}, \qquad K_{35} = K_{14}^2 = -1\frac{k}{6} \\ K_{42} &= K_{32}^1 = -1\frac{k}{6} + 1\frac{hL}{6}, \qquad K_{45} = K_{24}^2 = -2\frac{k}{6} \\ b_3 &= b_4^1 + b_1^2 = 2\frac{sL^2}{4} - 2\frac{\bar{q}L}{2} \\ b_4 &= b_3^1 + b_2^2 = 2\frac{sL^2}{4} + 2\frac{hT_\infty L}{2} \end{aligned} \tag{39}$$

Equations (38) are thus in detail

$$\frac{k}{6}\begin{bmatrix}8 & -2\\ -2 & 8+4\frac{hL}{k}\end{bmatrix}\begin{Bmatrix}T_3\\T_4\end{Bmatrix} = \frac{\bar{T}_B k}{6}\begin{Bmatrix}2\\1-\frac{hL}{k}\end{Bmatrix} + \frac{\bar{T}_D k}{6}\begin{Bmatrix}1\\2\end{Bmatrix} + \frac{sL^2}{2}\begin{Bmatrix}1\\1\end{Bmatrix} - \bar{q}L\begin{Bmatrix}1\\0\end{Bmatrix} + hT_\infty L\begin{Bmatrix}0\\1\end{Bmatrix} \tag{40}$$

The right-hand side consists of five forcing terms generating the temperature field. The quantity

$$\mathrm{Nu} = \frac{hL}{k} \tag{41}$$

is clearly dimensionless. It is called the *Nusselt number* (Nusseltin luku) in heat transfer literature. It is a dimensionless measure for the ratio of surface convection to heat conduction.

A slight further manipulation of (40) gives the set

$$\begin{bmatrix}4 & -1\\ -1 & 4+2\frac{hL}{k}\end{bmatrix}\begin{Bmatrix}T_3/T_\infty\\T_4/T_\infty\end{Bmatrix} = \frac{1}{2}\frac{\bar{T}_B}{T_\infty}\begin{Bmatrix}2\\1-\frac{hL}{k}\end{Bmatrix} + \frac{1}{2}\frac{\bar{T}_D}{T_\infty}\begin{Bmatrix}1\\2\end{Bmatrix} + \frac{3}{2}\frac{sL^2}{kT_\infty}\begin{Bmatrix}1\\1\end{Bmatrix} - 3\frac{\bar{q}L}{kT_\infty}\begin{Bmatrix}1\\0\end{Bmatrix} + 3\frac{hL}{k}\begin{Bmatrix}0\\1\end{Bmatrix} \tag{42}$$

Four additional dimensionless quantities appear on the right hand side. For example in the case

$$\frac{hL}{k}=1,\quad \frac{\bar{T}_B}{T_\infty}=1,\quad \frac{\bar{T}_D}{T_\infty}=1,\quad \frac{sL^2}{kT_\infty}=1,\quad \frac{\bar{q}L}{kT_\infty}=1 \tag{43}$$

the set is

$$\begin{bmatrix}4 & -1\\ -1 & 6\end{bmatrix}\begin{Bmatrix}T_3/T_\infty\\ T_4/T_\infty\end{Bmatrix}=\begin{Bmatrix}0\\ 11/2\end{Bmatrix} \tag{44}$$

and the solution is

$$\frac{T_3}{T_\infty}=\frac{11/2}{23}\approx 0.239,\qquad \frac{T_4}{T_\infty}=\frac{22}{23}\approx 0.957 \tag{45}$$

It is a good practice to operate with non-dimensional quantities. The step to obtain a non-dimensional presentation should actually be performed preferable already in the differential equation formulation stage. In this way the numerical calculations can made more use of as fewer variables appear in the problem. In fact, the computer works with pure numbers, not with kilograms or meters. However, we often put $L=1$ etc. for shortness of presentation. It must then be understood that we are using a consistent set of units and we work with the numerical measures only.

We now describe shortly how the data is presented in MATHFEM in this example case. MATHFEM is described in more detail in Section 3.4.

Roughly speaking, the discrete problem consists of two main parts: (1) the finite element approximation and (2) the weak form definition. The approximation is represented by the list of the global nodal numbers of each element, by the list of nodal coordinates, and by the list consisting of an initial guess of the unknown function nodal parameter values (MATHFEM employs the deltaform version). The weak form definition consists of the list of fixity codes for the nodal parameters, of the list of weak form integrands, and of a list telling which integrand to use on the element domains and their edges.

In Mathematica notation, the approximation definition is

**apr = {nod, crd, fun}** (46)

The global element nodal numbers are given in **nod**, the nodal coordinates in **crd** in the order of global nodal numbers, and the function nodal parameters in **fun** in the same order. Once this data is defined, one is able to plot the approximation, the finite element mesh and perform manipulations such as taking derivatives and so on.

In Mathematica notation, the weak form definition is

**prb = {fix, atr, exp}** (47)

The fixity codes are given in **fix**. This list consists of ones and zeros and is of the same size as **fun**. The code one means that the corresponding nodal parameter is allowed to change its value and the code zero that the parameter should keep its value. As the number of different types of integrand expressions is usually limited only to a few, the expressions are given in list **exp** in some convenient order. The list **atr** tells which integrand to use on the element domains and their edges. The sublist of **atr** in the position indicated by the element number consists of as many numbers as there are regions in the element. These are defined to be the element domain and its edges. The



number in the position corresponding to the local number of the region (Figure 3.18) is the location of the integrand expression in **exp**.

We consider next the data when the element mesh and the boundary conditions are as given in Figure (b) and by formulas (1) and the Dirichlet conditions are satisfied in the strong sense. The finite element approximation is given by

```
nod = {{1, 2, 4, 3}, {3, 4, 6, 5}};
crd = {{0, 0}, {1, 0}, {0, 1}, {1, 1}, {0, 2}, {1, 2}}*L;
fun = {{φa},{ φb}, {0}, {0}, {φd}, {φc}};
```
(48)

The list **nod** contains the global nodal numbers of the elements in the order determined by the local nodal numbering. The nodal coordinate list **crd** contains the nodal coordinates in the order determined by the global numbering The function nodal parameter list **fun** contains the given values (from the Dirichlet conditions). The unknown parameters can be given any values in the linear case, but a good guess may decrease the computational work in the non-linear case.

The function set $V$ and the weak form expression, i.e., the weak formulation is defined by

```
fix = {{0}, {0}, {1}, {1}, {0}, {0}};
atr = {{2, 1, 3, 1, 4}, {2, 1, 3, 1, 4}};
exp = {0, w[1]*d*φ[1]+w[2]*d*φ[2]-w[0]*f,
       w[0]*h*(φ[0]-φr), w[0]*q};
```
(49)

The value zero in the fixity code table **fix**, denoting a fixed nodal parameter, appears at locations corresponding to the nodes on $\Gamma_D$. Note that due to the restriction $\Delta\phi \in V \Rightarrow \Delta\phi|\Gamma_D = 0$, the Dirichlet condition "wins" at the nodes where the boundary condition type changes. In the MATHFEM code, the fixity code table is walked through replacing each occurrence of number 1 by an integer number starting from 1, using then 2 and so on to get an unique numbering for the free nodal parameters. The final value $n$ thus obtained is the total number of unknowns of the problem. In the example case the outcome of the modification step is $n = 2$ and

```
fix={{0}, {0}, {1}, {2}, {0}, {0}};
```
(50)

The members of the attribute list **atr** contain integer numbers referring to the integrand expressions to be used for the regions of the elements. For bookkeeping the different regions of the elements are ordered as follows. Domain proper $\Omega$ of an element comes first. The four edges (see Figure (b)) of the element $\Gamma_1$, $\Gamma_2$, $\Gamma_3$ and $\Gamma_4$ follow in that order. Similar ordering convention applies also in connection with line and triangular elements. The different integrand expressions appearing in the weak form are listed in **exp** in principle in an arbitrary order. Here the first member is the zero expression ($\hat{=}\,\Gamma_D$ or element sides inside the mesh), the second the domain expression ($\hat{=}\,\Omega$), the third the Robin boundary expression ($\hat{=}\,\Gamma_R$) and the fourth the Neumann boundary expression ($\hat{=}\,\Gamma_N$). The first number of a sublist of **atr**, corresponding to an element, gives the location of the expression of **exp** to be used in the domain, the next number gives the location of the expression to be used on the first edge and so on. To simplify the

numerical treatment, one-index notation is used for derivatives from zero to the first order (and also higher if necessary). The one-index notation for the derivatives has the meaning: $\phi[0]=\phi$, $\phi[1]=\partial\phi/\partial x_1$, and $\phi[2]=\partial\phi/\partial x_2$.

Figure (e) shows the solution for the temperature obtained by MATHFEM. The numerical values for the given data correspond to those of (43).

```
<<mathfem.m;
nod = {{1, 2, 4, 3}, {3, 4, 6, 5}};
crd = {{0, 0}, {1, 0}, {0, 1}, {1, 1}, {0, 2}, {1, 2}}*L;
fun = {{ϕa}, {ϕb}, {0}, {0}, {ϕd}, {ϕc}};
fix = {{0}, {0}, {1}, {1}, {0}, {0}};
atr = {{2, 1, 3, 1, 4}, {2, 1, 3, 1, 4}};
exp = {0, w[1]*d*ϕ[1]+w[2]*d*ϕ[2]-w[0]*f,
       w[0]*h*(ϕ[0]-ϕr), w[0]*q};
apr = {nod, crd, fun};
prb = {fix, atr, exp};

ϕa = ϕc = 0; ϕb = ϕd = 1; L = 1; d = f = 1; h = ϕr = 1; q = 1;
newapr = LINEAR[{apr, prb}];
SHOW3D[PLOT[newapr]];
Print["fun=", newapr[[3]]];
```

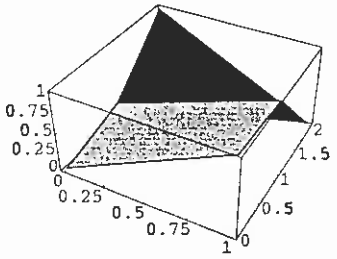



**Figure (e)**

```
fun = {{0.}, {1.}, {0.23913}, {0.956522}, {1.}, {0.}}
```

### 3.3.3 Numerical quadrature



The element contributions consist of definite integrals over the reference element domain and boundary. Some relevant formulas are explained in Appendix E. The integrands are usually so complicated that *numerical integration* (numeerinen integrointi) which is often also called *numerical quadrature* is necessary or convenient. This is a theme of classical numerical analysis and just a brief review is given here. We shall employ mostly the name "numerical quadrature" here as the term "numerical integration" is often used also in many other meanings, for instance for the numerical prediction of a solution in a time dependent phenomenon.

The integrals in one or two dimensions are replaced by sums:

$$\int f(r)\,\mathrm{d}r \approx \sum_i W_i f(r_i)$$
$$\iint f(r,s)\,\mathrm{d}r\,\mathrm{d}s \approx \sum_i W_i f(r_i,s_i) \tag{17}$$

Here the quantities $W_i$ are *weights* (paino) or to discern them from weighting functions in weak forms or from weight factors in the least squares method we can call them *weight coefficients* (painokerroin). The indices $i$ refer to the *integration points* or *sampling points* (integrointipiste, näytepiste) where the value of the integrand $f$ is to be evaluated for the sum. The forms (17) are quite transparent as a kind of Riemann sums where the weight coefficients can be interpreted as some measures of the sizes of the subdomains associated with the sampling points.

Let us consider first the one-dimensional case and the integral

$$\int_{-1}^{1} f(r)\,\mathrm{d}r \tag{18}$$

Most of the formulas in the literature are given for the standard non-dimensional interval $[-1,1]$. An arbitrary one-dimensional integral

$$\int_a^b f(x)\,\mathrm{d}x \tag{19}$$

can be transformed to form (18) by the mapping

$$x=\frac{a+b}{2}+\frac{b-a}{2}r \tag{20}$$

giving (cf. formula (E.1.3))

$$\int_a^b f(x)\,\mathrm{d}x=\frac{b-a}{2}\int_{-1}^{1} f\big(x(r)\big)\,\mathrm{d}r \tag{21}$$

We have employed above the symbols $r$ and $s$ for the independent dimensionless variables. These notations are used quite widely in the finite element literature for the local reference element coordinates. As our reference element coordinates $\xi$ and $\eta$ have the range $[0,1]$, we have used on purpose this double notation in this connection. For instance, with our notations we have the application

$$\int_0^1 f(\xi)\mathrm{d}\xi = \frac{1}{2}\int_{-1}^1 f(\xi(r))\mathrm{d}r \tag{22}$$

of formula (21) with the mapping (20) now

$$\xi = \frac{1}{2} + \frac{1}{2}r \tag{23}$$

Let us continue with the integral (18). The numerical integration formulas are derived in principle by approximating the integrand $f(r)$ by a polynomial passing through the function values $f(r_i)$ at the integration points and by evaluating the integral from the polynomial. There are basically two different ways to select the integration points leading to the Newton-Cotes formulas and to the Gauss formulas, respectively.

In the *Newton-Cotes formulas* the positions of the integration points are fixed in advance. Normally they are put uniformly in the domain. Figure 3.16 (a) shows the situation in the case the number of integration points $n = 4$.

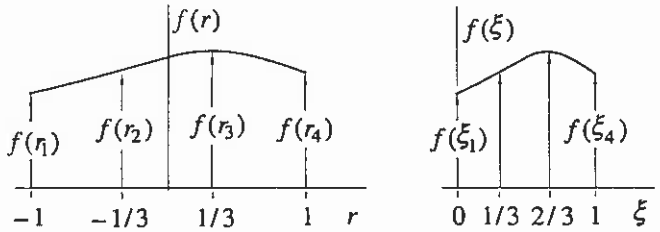



**Figure 3.16 (a)** Four integration points in the interval $r \in [-1,1]$. **(b)** The same in the interval $\xi \in [0,1]$.

The Lagrange approximating polynomial is

$$f(r) \approx \sum_i L_i^n(r) f(r_i) = L_1^4(r) f(r_1) + L_2^4(r) f(r_2) + L_3^4(r) f(r_3) + L_4^4(r) f(r_4) \tag{24}$$



or in more detail

$$\begin{aligned} f(r) \approx & \frac{(r+1/3)(r-1/3)(r-1)}{(-1+1/3)(-1-1/3)(-1-1)} f(r_1) \\ & + \frac{(r+1)(r-1/3)(r-1)}{(-1/3+1)(-1/3-1/3)(-1/3-1)} f(r_2) \\ & + \frac{(r+1)(r+1/3)(r-1)}{(1/3+1)(1/3+1/3)(1/3-1)} f(r_3) \\ & + \frac{(r+1)(r+1/3)(r-1/3)}{(1+1)(1+1/3)(1-1/3)} f(r_4) \end{aligned} \tag{25}$$

The Lagrange interpolation polynomials $L_i^n(r) = L_i^4(r)$ are here in the language of the finite element method the shape functions for a four-noded line element. They were referred to in Section 2.2.2. The general procedure for writing down the expressions for the interpolation functions can be detected from the example formula (25). Finally the integral

$$\int_{-1}^1 f(r)\mathrm{d}r \approx \int_{-1}^1 \left[\sum_i L_i^n(r) f(r_i)\right]\mathrm{d}r = \sum_i \left[\int_{-1}^1 L_i^n(r)\mathrm{d}r\right] f(r_i) \tag{26}$$

where summation and integration orders can and have been changed. Comparison with (17) shows that the weight coefficients can be evaluated from

$$W_i = \int_{-1}^1 L_i^n(r)\mathrm{d}r \tag{27}$$

The weight coefficients for the most conventional cases have been determined and documented in the literature.

Case $n = 2$ gives the *trapezoidal formula* (trapetsikaava)

$$\int_{-1}^1 f(r)\mathrm{d}r \approx 1 \cdot f(-1) + 1 \cdot f(1) \tag{28}$$

Case $n = 3$ gives the *Simpson's formula* (Simsonin kaava)

$$\int_{-1}^1 f(r)\mathrm{d}r \approx \frac{1}{3} \cdot f(-1) + \frac{4}{3} \cdot f(0) + \frac{1}{3} \cdot f(1) \tag{29}$$

and in the case $n = 4$, we have

$$\int_{-1}^{1} f(r)\mathrm{d}r \approx \frac{1}{4}\cdot f(-1)+\frac{3}{4}\cdot f\left(-\frac{1}{3}\right)+\frac{3}{4}\cdot f\left(\frac{1}{3}\right)+\frac{1}{4}\cdot f(1) \qquad (30)$$

A rough check on the weight coefficients is obtained by applying the formulas to the constant function 1 in which case we expect to obtain the exact integral — here 2 — so the sum of the weights should also equal 2. This result is indeed true with formulas (28) to (30).

Figure 3.16 (b) shows how the integration points from the standard interval $[-1,1]$ in Figure (a) are mapped according to (23) on the interval $[0,1]$. A look at formula (22) gives now a modified integration formula, say of (30):

$$\int_{0}^{1} f(\xi)\mathrm{d}\xi \approx \frac{1}{8}\cdot f(0)+\frac{3}{8}\cdot f\left(\frac{1}{3}\right)+\frac{3}{8}\cdot f\left(\frac{2}{3}\right)+\frac{1}{8}\cdot f(1) \qquad (31)$$

which could be used in connection with our notation.

In general, if the number of integration points is $n$, the corresponding Newton-Cotes formula integrates exactly a polynomial of degree $n-1$ (or lower). This is because $n$ function values determine uniquely a polynomial of degree $n-1$ and errors are generated if the integrand function is of higher degree. However, when the integration points are uniformly spaced and when $n$ is in addition odd, the formulas integrate "by chance" exactly still a polynomial of degree $n$. Simpson's formula is for this reason especially popular; it is simple enough and still integrates exactly a third degree polynomial.

In reality the functions to be integrated are of course seldom polynomials. But when we consider a function expanded by Taylor's formula, which consists of a polynomial and of the remainder, it is obvious that the formulas are more accurate in general, the higher degree polynomial they can integrate exactly.

If the number of integration points needed to achieve a given accuracy is rather high (say more than five) it is often not wise to apply the Lagrange interpolation polynomials for the whole interval. Instead the interval can be divided into subintervals and a lower degree formula can be applied separately for each part. The formulas generated in this way are called sometimes *composite formulas* (yhdistetty kaava). If for instance uniformly spaced subintervals and the trapezoidal formula are combined, we obtain

$$\int_{-1}^{1} f(r)\mathrm{d}r \approx \frac{2}{n-1}\left[\frac{1}{2}f(r_1)+f(r_2)+f(r_3)+\cdots+f(r_{n-1})+\frac{1}{2}f(r_n)\right] \qquad (32)$$



where the meaning of the notations is obvious. The use of composite formulas may be advantageous for instance in those cases where the function to be integrated is known to behave unsmoothly.

In the *Gauss formulas* the positions of the integration points are not fixed in advance. Instead, they are selected in an optimal way so that the resulting formula integrates exactly as high a degree polynomial as possible. When there are $n$ integration points, there are $2n$ unknown quantities: the weights $W_i$ and the coordinates $r_i$, which can be determined so that the formula integrates exactly a polynomial of degree $2n-1$. The unknowns are obtained from a system of equations, which is genereted by demanding that the formula should give exact integrals separately for the functions $r^0, r^1, r^2, \cdots, r^{2n-1}$. As the formula is linear in the function values, it then gives an exact integral for an arbitrary linear combination of these functions. The system of equations becomes non-linear with respect to the coordinate values $r_i$. In the solution the Legendre polynomials can be made use of. The Gauss formulas are therefore often also called Gauss-Legendre formulas. Table 3.1 gives the data for the cases $n = 1,2,3$. Reference Stroud and Secrest (1966) gives the data for $n$ up to 512 with coordinates and weight coefficients with the accuracy of 30 significant figures!

**Table 3.1** Integration point coordinates $r_i$ and weight coefficients $W_i$ in the Gauss quadrature formula

$$\int_{-1}^{1} f(r)\mathrm{d}r \approx \sum_{i=1}^{n} W_i f(r_i)$$

| $n$ | $i$ | $r_i$ | $W_i$ |
|---|---|---|---|
| 1 | 1 | 0 | 2 |
| 2 | 1 | $-1/\sqrt{3}$ | 1 |
| | 2 | $1/\sqrt{3}$ | 1 |
| 3 | 1 | $-\sqrt{3/5}$ | 5/9 |
| | 2 | 0 | 8/9 |
| | 3 | $\sqrt{3/5}$ | 5/9 |

It should be noted that when the values $r_i$ are known the weight coefficients can still be determined also from formula (27). Gauss type formulas have found

much use in the finite element method because they give good accuracy with a small number of sampling points and are thus cost effective.

Combination of the Newton-Cotes and Gauss type formulas can be devised — so-called *Lobatto formulas* (see e.g. Akin (1994)) — where some of the integration points are fixed in advance (the interval end points) and only the rest is optimized for position.

We have discussed this far mainly one-dimensional numerical integration. The Newton-Cotes and the Gauss formulas can be extended to two (and to three) dimensions. There are two possibilities: we can try to formulate the theory anew in two dimensions or we can just apply the one-dimensional formulas directly in a double integration. The latter alternative is normally used in connection with quadrilateral elements.

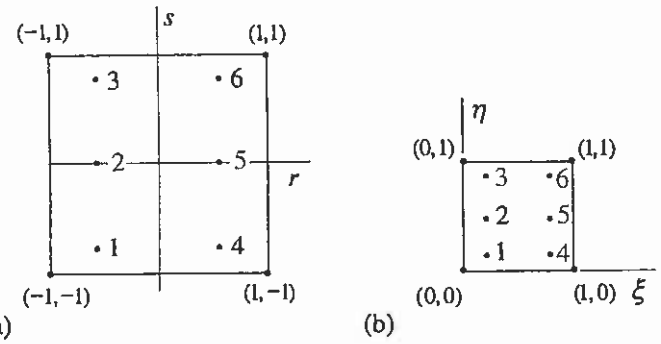



**Figure 3.17 (a)** Domain $r \in [-1,1]$, $s \in [-1,1]$. **(b)** Domain $\xi \in [0,1]$, $\eta \in [0,1]$. Integration points for $2 \times 3$ Gaussian integration.

Let us consider the integral

$$\int_{-1}^{1}\int_{-1}^{1} f(r,s)\,dr\,ds \tag{33}$$

over the standard domain shown in Figure 3.17 (a). The mapping

$$\xi = \frac{1}{2} + \frac{1}{2}r, \qquad \eta = \frac{1}{2} + \frac{1}{2}s \tag{34}$$

gives the corresponding reference element domain with our notation in Figure 3.17 (b) and we have the analog of (22):

$$\int_{0}^{1}\int_{0}^{1} f(\xi,\eta)\,d\xi\,d\eta = \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1} f\big(\xi(r),\eta(s)\big)\,dr\,ds \tag{35}$$



The plane integral (33) is evaluated analytically — if possible — as a double integral

$$\int_{-1}^{1}\left[\int_{-1}^{1} f(r,s)\,ds\right]dr \tag{36}$$

keeping $r$ constant in the inner integral so that the integrand is only function of $s$ and performing the integration first with respect to $s$. Let us do the same now numerically:

$$\int_{-1}^{1} f(r,s)\,ds \approx \sum_{l=1}^{n} W_l^{*} f(r,s_l) \tag{37}$$

Here the star refers to the weight coefficients in the corresponding one-dimensional formula with $n$ sampling points in the $s$-direction. After this we evaluate the integral

$$\int_{-1}^{1}\sum_{l=1}^{n} W_l^{*} f(r,s_l)\,dr \tag{38}$$

again numerically using one-dimensional formula now with $r$ as the integration variable:

$$\int_{-1}^{1}\sum_{l=1}^{n} W_l^{*} f(r,s_l)\,dr \approx \sum_{k=1}^{m} W_k^{*}\left(\sum_{l=1}^{n} W_l^{*} f(r_k,s_l)\right) = \sum_{k=1}^{m}\sum_{l=1}^{n} W_k^{*} W_l^{*} f(r_k,s_l) \tag{39}$$

We have arrived at the numerical integration formula

$$\int_{-1}^{1}\int_{-1}^{1} f(r,s)\,dr\,ds = \sum_{k=1}^{m}\sum_{l=1}^{n} W_k^{*} W_l^{*} f(r_k,s_l) \tag{40}$$

The weight coefficients are thus (see (17)) the products of the corresponding one-dimensional weights. This result has some similarity with the procedure of generating shape functions for quadrilateral elements by multiplication together corresponding one-dimensional shape functions (cf. Section 3.2.2).

Formula (40) is in much use especially with the Gauss formula and with the same number of integration points in both directions. Sometimes there may be physical reasons for using different number of integration points in the two directions. Figure 3.17 shows a case with $m = 2$ and $n = 3$. In this case formula (40) will integrate exactly a polynomial of third degree in $r$ and of fifth degree in $s$.

For triangular domains the application of the double integration procedure described above leads clearly to biased distribution of the integration points. This is aesthetically not quite satisfactory although this procedure works in practice. General Gauss type formulas starting with unknown coordinates and weights have been derived for triangles; for instance Hammer et al. (1956). Table 3.2 gives some data. The column headed "$p$" refers to the highest degree of complete polynomial the formula still integrates exactly. The data is given in area coordinates. When applied, say, for the reference domain shown in Figure 3.7 (a), the triangle area $A = 1/2$ and the values of $\xi$ and $\eta$ corresponding to the area coordinates can be found from (3.2.1).

**Table 3.2** Integration point coordinates $(L_1)_i$, $(L_2)_i$, $(L_3)_i$ and weight coefficients $W_i$ in the quadrature formula for a triangle:

$$\int_A f(L_1, L_2, L_3)\mathrm{d}A \approx 2A\left[\sum_{i=1}^{n} W_i f((L_1)_i, (L_2)_i, (L_3)_i)\right]$$

| $p$ | $n$ | $i$ | $(L_1)_i$ | $(L_2)_i$ | $(L_3)_i$ | $W_i$ |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1/3 | 1/3 | 1/3 | 1/2 |
| 2 | 3 | 1 | 1/2 | 1/2 | 0 | 1/6 |
| | | 2 | 0 | 1/2 | 1/2 | 1/6 |
| | | 3 | 1/2 | 0 | 1/2 | 1/6 |
| 3 | 4 | 1 | 1/3 | 1/3 | 1/3 | −9/32 |
| | | 2 | 3/5 | 1/5 | 1/5 | 25/96 |
| | | 3 | 1/5 | 3/5 | 1/5 | 25/96 |
| | | 4 | 1/5 | 1/5 | 3/5 | 25/96 |

The procedures explained above extend in an obvious way to three dimensions, say for hexahedra, tetraedra, triangular prisms, etc; Irons (1971), Hellen (1972). Reference Akin (1994) contains a large amount of data on numerical quadrature with finite elements.

Numerical quadrature of the element contributions leads for distorted elements with a non-constant Jacobian usually inevitably to some errors. However, since the finite element method is as such already an approximate procedure, this is not dangerous so long as the errors vanish fast enough when the element sizes tend to zero. In fact, numerical quadrature can be considered as an essential



ingredient in the finite element model of a problem. The model now "can feel what is happening only at the integration points". In certain cases so-called *reduced integration* (redusoitu integrointi) or *selectetive integration* (selektiivinen integrointi) can be employed to take advantage of this feature, e.g., Zienkiewicz and Taylor (2000) and Belytschko et al. (2000). We may call the integration rule for an element to be a *full integration* (täysi integrointi) rule if the finite element contributions are evaluated exactly by the rule when the element is not distorted. That is, the mapping from the reference element to the global element is at most linear in the natural coordinates and thus the Jacobian is a constant. (For instance for the three-noded triangle the isoparametric mapping is always of this type but not in general for example for the four-noded quadrilateral.) In this we further assume the data such as the thermal conductivity to be a constant in the element. For vanishing integration errors in the limit, full integration seems to be enough. In this text we will always apply full integration if nothing else is mentioned. Roughly, in reduced integration a less accurate rule than the full integration rule is used. In selective integration, reduced integration is applied only for certain terms in the element contribution expressions. Reduced and selective integration can in certain problems at the first look quite surprisingly increase the simulation capacity and this further with the benefit of lesser computational effort. One such problem is in elasticity with a nearly incompressible material.

**Example 3.4.** We consider the element of Example 3.2 and evaluate the term

$$K_{11} = \int_\Omega \left( \frac{\partial N_1}{\partial x} k \frac{\partial N_1}{\partial x} + \frac{\partial N_1}{\partial y} k \frac{\partial N_1}{\partial y} \right) \mathrm{d}\Omega \tag{a}$$

in the element coefficient matrix as given in (15). For simplicity, the element number superscript $e$ can be dropped here without danger of confusion. We perform the calculations (1) analytically, (2) with numerical quadrature.

(1) Taking into account formula (E.1.3) and with constant $k$, we have first

$$K_{11} = k \int_0^1 \int_0^1 \left( \frac{\partial N_1}{\partial x} \frac{\partial N_1}{\partial x} + \frac{\partial N_1}{\partial y} \frac{\partial N_1}{\partial y} \right) \det[J]\, \mathrm{d}\xi\, \mathrm{d}\eta \tag{b}$$

The global derivatives have been evaluated in Example 3.2:

$$\begin{aligned} \frac{\partial N_1}{\partial x} &= \frac{1}{1.32a}\left[1.4(-1+\eta) - 0.2(-1+\xi)\right] \\ \frac{\partial N_1}{\partial y} &= \frac{1}{1.32a}\left[-0.4(-1+\eta) + 1(-1+\xi)\right] \end{aligned} \tag{c}$$

and we have from there further

$$\det[J] = \det\left( \begin{bmatrix} 1 & 0.4 \\ 0.2 & 1.4 \end{bmatrix} a \right) = 1.32a^2 \tag{d}$$

Expression (b) is now in more detail

$$K_{11} = \frac{k}{1.32}\int_0^1\int_0^1 \Big[\big(1.4(-1+\eta)-0.2(-1+\xi)\big)^2 + \big(-0.4(-1+\eta)+1(-1+\xi)\big)^2\Big]\mathrm{d}\xi\mathrm{d}\eta$$

$$= \frac{k}{1.32}\int_0^1\int_0^1 \big(1.8-0.72\xi+2.88\eta+1.04\xi^2-1.36\xi\eta+2.12\eta^2\big)\mathrm{d}\xi\mathrm{d}\eta \qquad \text{(e)}$$

As the integrand happens to be here of a simple form and as the limits are simple, there is obtained with a relatively small effort the analytical result

$$K_{11} = \frac{k}{1.32}\left(1.8-0.72\frac{1}{2}+2.88\frac{1}{2}+1.04\frac{1}{3}-1.36\frac{1}{4}+2.12\frac{1}{3}\right) \approx 0.540\,404\,k \qquad \text{(f)}$$

(2) We start from formula (e) written as

$$K_{11} = \frac{k}{1.32}\int_0^1\int_0^1 f(\xi,\eta)\mathrm{d}\xi\mathrm{d}\eta \qquad \text{(g)}$$

with

$$f(\xi,\eta) = 0.8-0.72\xi+2.88\eta+1.04\xi^2-1.36\xi\eta+2.12\eta^2 \qquad \text{(h)}$$

(In an actual numerical quadrature in a finite element program, the treatment would naturally begin earlier without the tedious analytical manipulations performed above.) The integrand is of second degree $\xi$ and $\eta$. The one point $(1\times1)$ Gauss formula (40) will integrate exactly only up to the first degree terms in (h). The next four point $(2\times2)$ Gauss formula integrates exactly up to the third degree terms and is thus here enough. Thus $2\times2$ integration means here full integration for an arbitrary four-noded quadrilateral element in connection with a diffusion problem. Use of formulas (35), (40), (34), (17) and Table 3.1 gives

$$\int_0^1\int_0^1 f(\xi,\eta)\mathrm{d}\xi\mathrm{d}\eta = \frac{1}{4}\big[1\cdot1\cdot f(\xi_1,\eta_1)+1\cdot1\cdot f(\xi_2,\eta_2) + 1\cdot1\cdot f(\xi_3,\eta_3)+1\cdot1\cdot f(\xi_4,\eta_4)\big] \qquad \text{(i)}$$

with

$$\begin{aligned}
\xi_1 &= \frac{\sqrt{3}-1}{2\sqrt{3}} \approx 0.2113249, \quad \eta_1 = \frac{\sqrt{3}-1}{2\sqrt{3}} \approx 0.2113249\\
\xi_2 &= \frac{\sqrt{3}-1}{2\sqrt{3}} \approx 0.2113249, \quad \eta_2 = \frac{\sqrt{3}-1}{2\sqrt{3}} \approx 0.7886751\\
\xi_3 &= \frac{\sqrt{3}-1}{2\sqrt{3}} \approx 0.7886751, \quad \eta_3 = \frac{\sqrt{3}-1}{2\sqrt{3}} \approx 0.2113249\\
\xi_4 &= \frac{\sqrt{3}-1}{2\sqrt{3}} \approx 0.7886751, \quad \eta_4 = \frac{\sqrt{3}-1}{2\sqrt{3}} \approx 0.7886751
\end{aligned} \qquad \text{(j)}$$

We obtain

$$f(\xi_1,\eta_1)=1.1196152 \quad f(\xi_2,\eta_2)=0.5148975$$
$$f(\xi_3,\eta_3)=1.1384358 \quad f(\xi_4,\eta_4)=0.0803848 \tag{k}$$

and

$$K_{11}=\frac{k}{1.32}\int_0^1\int_0^1 f(\xi,\eta)\,d\xi\,d\eta$$
$$=\frac{k}{1.32}\frac{1}{4}\left[f(\xi_1,\eta_1)+f(\xi_2,\eta_2)+f(\xi_3,\eta_4)+f(\xi_4,\eta_4)\right]$$
$$=\frac{k}{1.32}\,0.71333333=0.540404\,k \tag{l}$$

Results (f) and (l) are seen to be equal as predicted by theory.



## 3.4 MATHFEM CODE

### 3.4.1 Introduction

In this section some properties of the Mathematica based program MATHFEM (Appendix G) used in this text are described. The rather simple code is meant for demonstration of the working principles of practical codes. As discussed in Section 2.4, it is convenient to think that a solution of a problem by the finite element method consists of three relatively independent parts — (1) pre-processing, (2) generation and solution of the discrete equations, and (3) post-processing — and the descriptions here follow that order.

Pre-processing means the phase consisting of the preparation of the input data. This phase, usually performed partly by some of the well-known mesh generation techniques, is not given much room here but it is assumed that the domain is a line or a rectangle or something that can be obtained by simple mapping from them.

The solver implements a rather general algorithm and it can be used for linear and non-linear problems, and for several unknown functions. Also non-stationary problems can be treated by using the time-discontinuous Galerkin method (Section 9.3). There are no "a priori" limitations on the problem size but the solution times tend to be acceptable only for problems with about 100 unknowns.

The post-processing phase, meaning transforming the outcome of the calculations into a figure of some kind, has been paid also some attention to. Although the post-processing functions are not particularly sophisticated in implementation, they should give some idea about the possibilities.

### 3.4.2 Data structure

Roughly speaking, the discrete formulation consists of the finite element approximation and of the weak form. The MATHFEM representation of the finite element approximation is given by (Some features of the program data structure have appeared already in Example 3.3.)

**apr = {nod, crd, fun}** (1)

where the element nodal numbers are given in **nod**, the nodal coordinates in **crd** and the function nodal parameters in **fun**. With this data one is able to plot the functions, the element mesh and perform manipulations such as taking derivatives of the functions and so on.

The representation of the weak form is

**prb = {fix, atr, exp}** (2)

where the fixity codes (see Example 3.1) and the expressions of the integrands are given in **fix** and **exp**, respectively. The second component **atr** contains the information: which integrand expression of **exp** is applied on which part of the element region. It should be noticed that the element region consists here, in addition of the element domain itself, also of its boundaries. The fixity code table **fix** is of the same size as **fun**. Code zero means that the corresponding nodal parameter is fixed i.e., the value is known from the Dirichlet boundary condition, and code one that the nodal parameter will be treated as unknown.

The numbers of **atr** indicating the selection of combination of the expressions of **exp** to be used are obtained in the following manner. First the expressions are put in list **exp** in some order. The first element of the list is zero as the zero expression is needed almost always. Thereafter the locations of the expressions to be used on the element and on its boundaries are represented in the form of a list in the order of the domain and boundary numbering order indicated in Figure 3.18. So the first element of the list gives the location of the expression to be used in the domain, the second digit the location of the expression to be used on the first edge and so on.

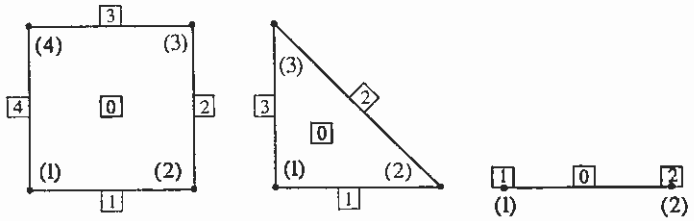



**Figure 3.18.** The local numbering of the nodes and the element domain and its boundaries.

**Example 3.5.** Definition of the **apr** and **prb** structure in the case of two quadrilateral four-noded elements (Figure (a)) and the appproximation problem (least squares fitting) for a function $\sin x \cdot \sin y$ in $(x, y) \in [0, L] \times [0, 2L]$.

```
nod = {{1, 2, 4, 3 }, {3, 4, 6, 5}};
crd = {{0, 0}, {1, 0}, {0, 1}, {1, 1},
       {0, 2}, {1, 2}}*L;
fun = {{1}, {1}, {1}, {1}, {1}, {1}}*f0;
apr = {nod, crd, fun};
fix = {{1}, {1}, {1}, {1}, {1}, {1}};
```



```
atr = {{2, 1, 1, 1, 1}, {2, 1, 1, 1, 1}};
exp = {0, w[0]*(ϕ[0]-
      Sin[x[1]] *Sin[x[2]])};
prb={fix, atr, exp};
```

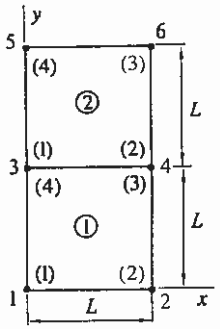
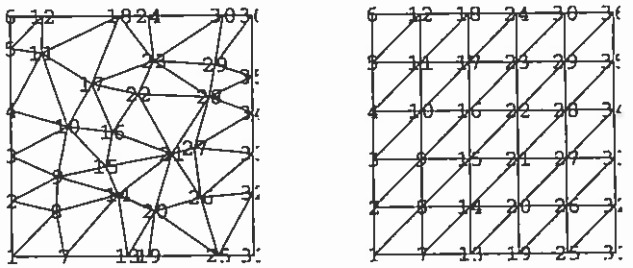



**Figure (a)**

**Remark 3.10.** Internally the basis functions are expressed in terms of polynomials of the global coordinates in contrast to the usual treatment based on the elementwise local system. This introduces, in certain cases, a minor modification over the usual continuous approximation (cf. Appendix G). □

### 3.4.3 Mesh generation

An automatic mesh generation scheme is an essential part of any code meant for practical applications, as the input data may consist of very large tables. Discussion of the techniques for unstructured and structured meshes in common use can be found for example in Akin (1994), Beer & Watson (1992) and references therein.



(a) (b)

**Figure 3.19 (a)** Unstructured and **(b)** structured mesh for a polygonal domain.

The generation schemes for *unstructured* (rakenteeton, epäsäännöllinen) meshes are best suited for the so-called simplex elements (triangle, tetrahedron). Typically the input data consist of specification of the boundary of the domain

in some convenient way and the number of elements wanted. Thereafter the generation, guided by some kind of mesh density information, proceeds more or less automatically. For example the Delaunay method can be used to generate good quality meshes with a minimum amount of input data. For a thorough discussion of the method and a Fortran library we refer to Joe (1986).

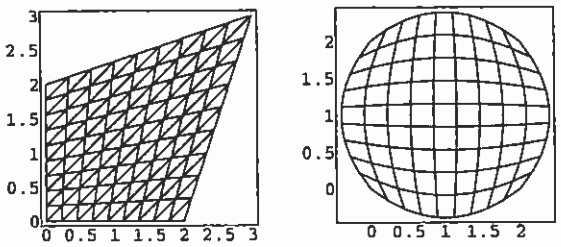
The generation methods for *structured* (rakenteinen, säännöllinen) meshes make use of mapping and interpolation techniques. As simple and straightforward, the method is widely used although the meshing of a complex region may involve a tedious manual stage. The simple version of the interpolation technique of MATHFEM is based on the use of an isoparametric mapping: the nodal coordinates of the so-called macro element are chosen to represent the true domain in the global coordinate system and the local coordinates are given values on a regular grid. The MATHFEM function

**msh = MSH[crd, n, nne]** (3)

generates the mesh for a single macro. The input parameters are the list of vertex coordinates in **crd** (in counterclockwise order), and **n** is the number of nodes wanted in the coordinate directions of the local system. The last argument **nne** is the number of nodes in an element in the mesh.

**Example 3.6.** Generation of simple meshes in two dimensions.

```
<<mathfem.m;
SHOW2D[MESH[MSH[{{0, 0}, {2, 0}, {3, 3}, {0, 2}},
{10, 10}, 3]]];
SHOW2D[MESH[MSH[{{0, 0}, {1,- 0.4}, {2, 0}, {2.4, 1},
{2, 2 }, {1, 2.4}, {0, 2}, {-.4, 1}}, {10, 10}, 4]]];
```




**Remark 3.11.** The mapping technique extends easily to several macros. Then the mesh is generated first for each macro and after that the approximation is forced to be continuous at the interfaces of the macros by using the fixity code table. Naturally the meshes of the macros must be compatible for overall integrity. □



**Remark 3.12.** The trend toward adaptive methods makes automatic mesh generation a part of the solution algorithm. Then the mesh is re-constructed during the solution several times to produce a numerical solution whose error (in some sense) is distributed uniformly in the domain. □

### 3.4.4 Data generation

Only the **nod** and **crd** lists of **apr** have been discussed this far. For a complete definition of a discrete problem one has to provide also the initial values of the unknown functions **fun**, the fixity codes of the nodal parameters **fix**, the integrand expression **exp** and the way these are applied **atr**.

The generation schemes of MATHFEM are applicable in one- and two-dimensional polygonal domains. The initial values of the unknown function can be generated by using the function

**apr = {nod, crd, fun} = APR[{nod, crd}, map]** (4)

taking as the input arguments the nodal numbers of the elements **nod**, nodal coordinates **crd** and a pure function **map**. The pure function specifies the initial values as functions of the global coordinates.

The structure representing the problem can be built using the function

**prb = PRB[apr, exp]** (5)

Initializing the members of list **fix** to have the value one meaning that all the nodal parameters are free to change their values. The **atr** list is initialized to use the second member of list **exp** for all the element domains and the first (zero) expression for all the element edges.

The way to modify the fixity codes is based on checking whether a node lies on a given polygonal curve. If a point is on the curve, the fixity code is zeroed and otherwise it is left untouched. The same scheme applies also for modifying the initial values. The function

**prb = FIX[prb, pol, val]** (6)

returns a **prb** representing a problem with fixed values for the nodal parameters on the nodes lying on the given polygonal curve **pol**. The values given in **val** are interpolated between the points of **pol**.

### 3.4.5 Finite element solver

**General.** The second phase, solution of the problem defined by **apr** and **prb**, involves many topics such as quadratures, numerical differentiation, linear equation system solving, non-linear equation system solving etc. To write an efficient solver means also considerations having to do with complexity of calculations (in storage and time) which are far beyond the scope of this text.

Although a solver intended for problems combining time-dependency, several unknowns and non-linearity can be rather complex on the algorithmic level, it is quite simple from the functional point of view: it is just a mapping transforming the finite element approximation **apr** into a new one with modified nodal parameter values. One may think that the problem represented by **prb** acts as a kind of parameter which tunes the mapping. This point of view omitting the internal structure will be adopted in the following sections.

**Linear case.** The generation and solution of the discrete equations is often the most time consuming phase. During this stage the element contributions are formed and added to the matrix and the vector of the system equations. After that some of the well-known techniques are applied to get the solution. The solver of the MATHFEM

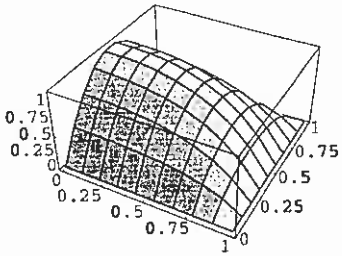
**apr = LINEAR[{apr, prb}]** (7)

takes as the input the problem **{apr, prb}** and returns a modified finite element approximation **apr** representing the numerical solution.

**Example 3.7.** Second order problem $-\partial^2\phi/\partial x^2 - \partial^2\phi/\partial y^2 - 10 = 0$ in $(x, y) \in [0,1]\times[0,1]$ with the zero Neumann condition on $x = 0$, $y \in [0,1]$ and the zero Dirichlet condition on the remaining part of the boundary.

```
<<mathfem.m;
dom = {{0, 0}, {1, 0}, {1, 1}, {0, 1}};
msh = MSH[ dom, {10, 10}, 4 ]];
apr = APR[msh, {0}&];
fep=PRB[apr,{0, w[1]*ϕ[1]+w[2]* ϕ[2]-w[0]*10}];
fep = FIX[fep, dom, {{0}, {0}, {0}, {0}}];
SHOW3D[PLOT[LINEAR[fep]]];
```





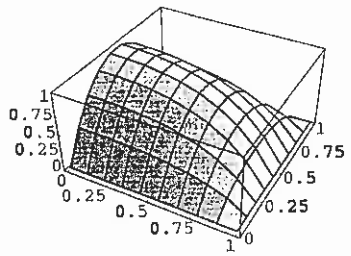

**Non-linear case.** It is clear from the way the weak form is derived that it is always linear in the weighting function (functions) but it may well be non-linear in the function (functions) to be determined. In practice one does not try to solve a non-linear problem directly, but the problem is rather replaced by a series of linear problems that can be solved by standard techniques. The function of MATHFEM implementing a combined Picard-quasi-Newton technique

**apr = NONLINEAR[{apr, prb}, err]** (8)

takes as the input the problem and it returns a modified finite element approximation **apr** representing the numerical solution. The second argument **err** is the relative error allowed.

**Example 3.8.** Second order problem $-\partial^2\phi/\partial x^2 - \partial^2\phi/\partial y^2 + \phi^2 - 10 = 0$ in $(x, y) \in [0,1]\times[0,1]$ with the zero Neumann condition on $x = 0$, $y \in [0,1]$ and the zero Dirichlet condition on the remaining part of the boundary.

```
<<mathfem.m;
dom = {{0, 0}, {1, 0}, {1,1}, {0,1}};
msh = MSH[dom, {10,10}, 4]];
apr = APR[msh, {0}&];
fep = PRB[apr, {0, w[1]*ϕ[1]+w[2]* ϕ[2]+w[0]* ϕ[0]^2-w[0] *10}];
fep = FIX[fep, dom, {{0}, {0}, {0}, {0}}];
SHOW3D[PLOT[NONLINEAR[fep, 0.01]]];
```




### 3.4.6 Function plot

The simplest way to illustrate the finite element approximation is a function plot, where the unknown function is plotted against the global coordinates. Besides that, one may also plot quantities derivable from the solution. An example is the unknown function derivative. It is noteworthy that the derivative mapping produces a function outside the finite-element space.

The natural way to proceed with the finite element interpolant is like assembling the global system. The basic scheme can be stated as: For each element (1) form the approximation as function of the global coordinates, (2) apply the given operator to get the wanted quantity, (3) calculate the nodal values of that quantity and plot as a polygon or a line. Finally, the set of polygons when rendered with hidden surface removal gives an illustration of the quantity wanted. Function

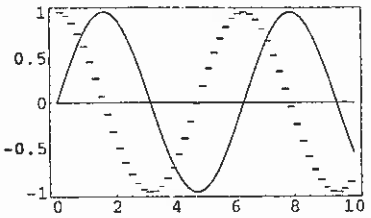
**graphics = PLOT[apr, map]** (9)

takes as input the finite element approximation **apr** and returns a Mathematica graphics statement. The second argument **map** is a pure function acting on the approximation in the way described above. The following examples show the basic features. It is noteworthy that in the one-dimensional case the first derivative is constant in each element and the second derivative is identically zero.

**Example 3.9.** A plot of function $\phi(x) = \sin x$, $x \in [0,10]$ with its first and second derivatives.

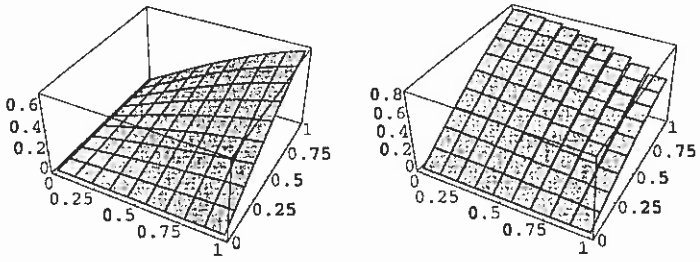
```
<<mathfem.m;
msh = MSH[{{0.}, {10.}}, {51}, 2];
apr = APR[msh, Sin[#1]&];
SHOW1D[{PLOT[apr], PLOT[apr, D[#, x[1]]&],
PLOT[apr, D[#, {x[1], 2}]&]}]
```






**Example 3.10.** Simple plot of function $\phi(x, y) = \sin x \cdot \sin y$ with its partial derivative with respect to $x$ in $(x, y) \in [0,1] \times [0,1]$.

```
<<mathfem.m;
msh = MSH[{{0, 0}, {1, 0}, {1, 1}, {0, 1}}, {10, 10}, 4];
apr = APR[msh, {Sin[#[[1]]Sin[#[[2]]}&];
SHOW3D[PLOT[MAP[apr, Identity]]];
SHOW3D[PLOT[apr, D[#, x [1]]&]];
```




### 3.4.7 Vector plot

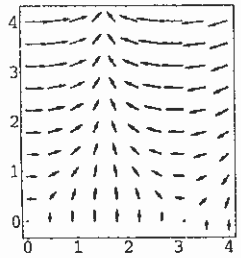
An arrow plot is often the most useful way for visualizing vector data such as a velocity field. The basic data consists then not of the approximation but only of the (nodal) coordinates and the corresponding function values. The function

**graphics = VECT[{crd, fun}]** (10)

plots the vectors at the nodal points. If a plot on some other set of points is wanted, one may use the **SCAN** function to get the values of the finite element approximation on that set of points.

**Example 3.11.** Plot of the gradient field of function $\phi(x, y) = y \cdot \sin x$ in $(x, y) \in [0,4] \times [0,4]$.

```
<<mathfem.m;
msh = MSH[{{0, 0}, {1, 0}, {1, 1}, {0, 1}}*4, {10, 10}, 4];
apr = APR[msh, {Cos[#[[1]]]*#[[2]],Sin[#[[1]]] }&];
SHOW2D[VECT[{apr[[2]], apr[[3]]}]];
```




### 3.4.8 Density plot

A density plot can be used for visualizing scalar data such as temperature. In the simplest form a density plot is nothing but the mesh plot where, instead of drawing the element boundaries, the elements are filled with a given color or pattern determined by the mean value of the function inside the element. For a smoother illustration one may divide each element until the jump in the mean value between the neighboring elements falls below a given limit. This option is, however, not included in

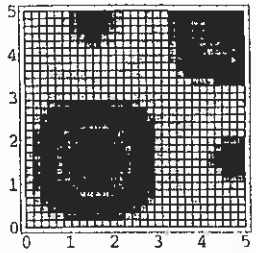
**graphics = POLY[apr, n]** (11)

of MATHFEM. The alternative approach based on the manipulating of the approximation before plotting will be discussed in the next section. The following example illustrates the drawback of the simple approach: the resolution is tied with the grid giving an impression of a more or less discontinuous solution.

**Example 3.12.** Density plot of $\phi(x, y) = \sin x \cdot \sin y$ in $(x, y) \in [0,5]\times[0,5]$ combined with the mesh plot.

```
<<mathfem.m;
msh = MSH[{{0, 0}, {1,0}, {1, 1}, {0, 1}}*5, {10, 10}, 4];
apr = APR[msh, {Sin[#[[1]]]Sin[#[[2]]]}&];
SHOW2D[{POLY[apr, 30],MESH[msh]}];
```






### 3.4.9 Manipulation on the approximation

Before proceeding to other ways of visualization, we discuss some elementary ways to manipulate the finite element approximation.

**Mapping back to $C^0(\bar{\Omega})$.** The figures in Section 3.4.6 illustrate the discontinuous behavior of the derivative. Although this is the true nature obtained from the $C^0$ finite element approximation, one would often like to illustrate the derivatives using a continuous approximation. For that task one has to map the discontinuous function somehow back on the finite element space. A simple — although not necessarily the best — method is the averaging of the nodal values from the function obtained through differentiation (see Section 2.4.2). The nodal values from each element are calculated separately, the values associated to a given node are added and the result is divided by the number of elements having the node in common. The function

**{crd, fun} = MAP[apr, crd]** (12)

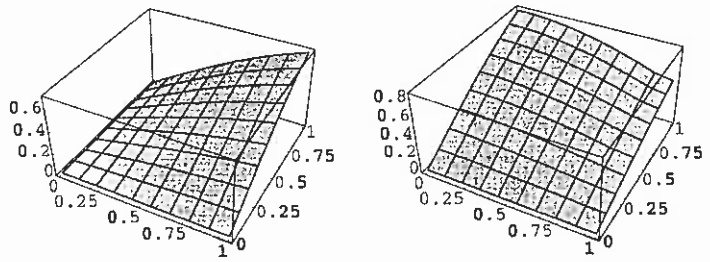
for simple averaging takes as input a finite element approximation **apr** and a list of coordinates **crd**. The list returned consists of the given coordinates and the approximation values on this list.

The figures below illustrate use of the function. When the approximation belongs to the finite element space the mapping performed is identity. The figure on the right is, however, recognizable more easily as the first partial derivative with respect to $x$.

**Example 3.13.** Plot of function $\phi(x, y) = \sin x \cdot \sin y$ and its partial derivative with respect to $x$ in $(x, y) \in [0,1]\times[0,1]$.

```
<<mathfem.m;
msh = MSH[{{0, 0}, {1, 0}, {1, 1}, {0, 1}}, {10, 10}, 4];
apr = APR[msh, {Sin[#[[1]]]Sin[#[[2]]]}&];
```

```
SHOW3D[PLOT[MAP[apr, Identity]]];
SHOW3D[PLOT[MAP[apr, D[#, x[1]]&]]];
```




**Splitting of the mesh.** The previous operation modified the finite element derivative approximation itself. A useful operation having no effect on the approximation but only on its elements **nod**, **crd** and **fun** is called here splitting. The operation divides the elements in such a way that a given *level curve* (tasa-arvokäyrä) of a function $\phi$

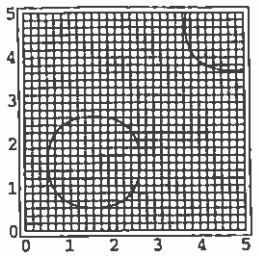
$$C = \{ (x,y) \in \Omega : \phi(x,y) = c = \text{a given constant} \} \tag{13}$$

does not intersect any of the elements. Depending on the original approximation and the element types, the polygons thus produced are not necessarily triangles or quadrilaterals but arbitrary polygons with known vertex coordinates and function nodal parameters (values in most cases). The splitting function finds use when one wants to plot only some subregion of the domain. The most important use is, however, in connection with contour plots to be discussed in the next section.

**Example 3.14.** Splitting of the mesh using the level curve $\phi(x,y) = 0.5$ of function $\phi(x,y) = \sin x \cdot \sin y$ in $(x,y) \in [0,5] \times [0,5]$.

```
<<mathfem.m;
msh = MSH[{{0,0}, {1,0}, {1,1}, {0,1}}*5, {30, 30}, 4];
apr = APR[msh, {Sin[#[[1]]]Sin[#[[2]]]}&];
apr2 = SPLIT[apr, 0.5];
msh2 = apr2[[{1, 2}]];
SHOW2D[MESH[msh2]];
```






### 3.4.10 Contour plot

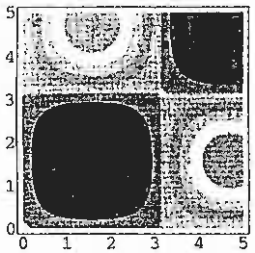
In several dimensions illustration of the level sets (13) for one or more selections of $c$ may give a satisfactory description about the solution. Often the level sets are curves but they can also be sub-regions. The so-called contour plot is useful, in particularly, if one wants to combine different types of plots e.g., to illustrate a scalar and a vector field in the same figure. The function

**graphics = CONT[apr, n]** (14)

produces a plot of **n** level curves of the approximation **apr**. The regions between the curves are also colored. In the approximate approach the function **SPLIT** is used together with the density plot function **POLY**. First the function **SPLIT** is applied iteratively for all the level set values wanted and after that the resulting modified approximation is plotted using **POLY** in such a way that the elements having the mean value of $\phi$ between two given subsequent level set values share the same color.

**Example 3.15.** Contour plot of $\phi(x,y) = \sin x \cdot \sin y$ in $(x,y) \in [0,5] \times [0,5]$.

```
<<mathfem.m;
msh = MSH[{{0, 0}, {1, 0}, {1, 1}, {0, 1}}*5, { 30, 30}, 4];
apr = APR[msh, {Sin[#[[1]]]Sin[#[[2]]]}&];
SHOW2D[CONT[apr, 10]];
```






## 3.5 APPLICATIONS

Some data and some parts of the example geometries in the following applications are adapted from Incropera and DeWitt (1996) and from Çengel (1998). The main practical interest in the examples is to determine the heat flow rates in the systems. The calculations have been performed using a Fortran program having a similar structure as MATHFEM.

### 3.5.1 Two fins

Two fin configurations are shown in Figure 3.20. Fin 1 (Figure (a)) is of uniform thickness and fin 2 (Figure (b)) is tapered. The temperature at the fin base is given and on the rest of the surface of the fins convective heat transfer is taking place.

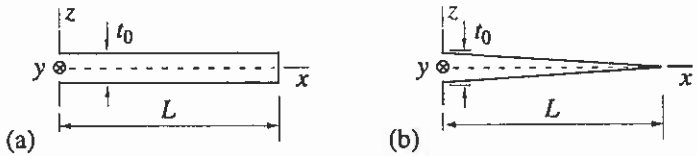



**Figure 3.20** **(a)** Cross section of fin 1. **(b)** Cross section of fin 2.

We apply here the (approximate) fin theory presented in Section 3.1.2. The properties and the conditions of the fins in the transverse ( $y$-axis) direction are assumed to be uniform so the problems become finally one-dimensional in $x$. We make the following assumptions on the data: constant (isotropic) heat conductivity $k$, no heat sources, constant heat transfer coefficient $h = h^+ = h^-$, constant reference temperature $T_\infty$. Combination of formulas (3.1.45), (3.1.47), (3.1.49) (3.1.55), (3.1.57) gives first a standard weak form

$$\int_A \left( \frac{\partial w}{\partial x} kt \frac{\partial T}{\partial x} + \frac{\partial w}{\partial y} kt \frac{\partial T}{\partial y} \right) \mathrm{d}A + 2\int_A wh\left(T - T_\infty\right) \mathrm{d}A + \int_{s_R} wht\left(T - T_\infty\right) \mathrm{d}s = 0 \tag{1}$$

As no dependence on coordinate $y$ is assumed, (1) becomes (We take a width $b$ in the $y$-axis direction. Then formally $\mathrm{d}A = b\mathrm{d}x$, $\int \mathrm{d}s = b$ and we divide by $b$.)

$$\int_0^L \frac{\mathrm{d}w}{\mathrm{d}x} kt \frac{\mathrm{d}T}{\mathrm{d}x} \mathrm{d}x + 2\int_0^L wh\left(T - T_\infty\right) \mathrm{d}x + wht\left(T - T_\infty\right)\Big|_{x=L} = 0 \tag{2}$$

For fin 1 the thickness is constant:

$$t = t_0 \tag{3}$$

and for fin 2 the fin is tapered according to

$$t = t_0\left(1-\frac{x}{L}\right) \tag{4}$$

The temperature at the base of the fin is given

$$T = T_b \quad \text{at } x = 0 \tag{5}$$

This is the Dirichlet boundary condition.

For fin 1 a closed form solution can be found for comparison purposes, e.g. Incropera and DeWitt (1996, p. 118)

$$T = T_\infty + \frac{\cosh m(L-x) + (h/mk)\sinh m(L-x)}{\cosh mL + (h/mk)\sinh mL}(T_b - T_\infty) \tag{6}$$

where

$$m = \sqrt{\frac{2h}{kt_0}} \tag{7}$$

The heat flow rate per unit length in the fin transverse direction is given by

$$\dot{Q}' = 2\int_0^L h(T-T_\infty)\,dx + ht_0(T-T_\infty)\big|_{x=L} = 0 \tag{8}$$

The analytical result is

$$\dot{Q}' = \sqrt{2hkt_0}\,\frac{\sinh mL + (h/mk)\cosh mL}{\cosh mL + (h/mk)\sinh mL}(T_b - T_\infty) \tag{9}$$

With finite elements we evaluate the heat flow rate by post-processing from the element contributions using (8) with $T$ replaced by $\tilde{T}$. For fin 2 the last terms in (2) and (8) are seen to vanish as $t = 0$ at $x = L$.

We perform the calculations using the data



$$L = 0.020\,\text{m}, \quad t_0 = 0.003\,\text{m},$$
$$k\,(\text{stainless steel}) = 14\frac{\text{W}}{\text{m}\cdot\text{K}}, \quad h = 30\frac{\text{W}}{\text{m}^2\cdot\text{K}} \tag{10}$$
$$T_b = 373\,\text{K}, \quad T_\infty = 293\,\text{K}$$

These give

$$mL = 0.7559, \quad \frac{h}{mk} = 0.05669, \quad \sqrt{2hkt_0} = 1.587\frac{\text{W}}{\text{m}\cdot\text{K}} \tag{11}$$

and the heat flow rate from (9) becomes

$$\dot{Q}' = 85.22\frac{\text{W}}{\text{m}} \tag{12}$$

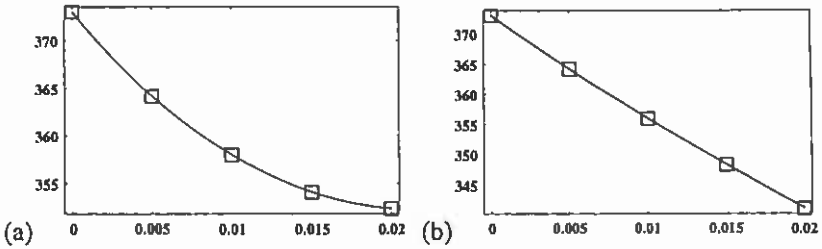



**Figure 3.21 (a)** Exact temperature distribution and finite element nodal temperature values for fin 1 (uniform fin). **(b)** "Exact" temperature distribution and finite element nodal temperature values for fin 2 (tapered fin).

We use two-noded line elements. Results for the temperature distribution with a uniform mesh of only four elements are shown in Figure 3.21. The in practice exact solution for the tapered fin is obtained by a mesh of 128 elements. The nodal values are seen to be very accurate already with this crude mesh as the exact temperature distribution is varying mildly.

The heat flow rates are given in Table 3.3. The results by a uniform mesh of eight elements are given for comparison purposes. The values obtained for the heat flow rates with the two meshes are seen to differ very little from each other. In engineering practice a quite usual — although mathematically of course not rigorous — way to operate is to solve a problem consecutively by a rather crude and by a more refined mesh. If the results by the two meshes do not differ "too much" from each other, the result by the denser mesh are considered adequate for the problem at hand.

**Table 3.3** Heat flow rates per unit transverse direction (W/m)

| | Fin 1 | Fin 2 |
|---|---|---|
| 4 elements | 85.42 | 76.082 |
| 8 elements | 85.27 | 76.078 |
| Exact | 85.22 | 76.077 |

### 3.5.2 Engine head

Figure 3.22(a) presents a part of a cross section of an engine head.

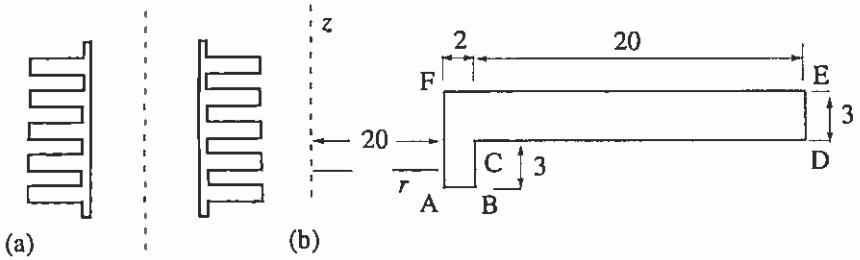



**Figure 3.22 (a)** Cross section of a cylindrical motor head. **(b)** Computational domain (measures in mm).

We consider as an initial study the performance of one typical repeating part in the middle area of the head assuming axial symmetry. We have the situation shown in Figure 3.22 (b) consisting of cross section of half of a fin and half of a base part. We assume in addition symmetry in solution with respect to planes perpendicular to the cylinder axis ($z$-axis) and cutting the fins and the bases at the middle (lines AB and FE). We do not try to use the approximate fin theory as in the previous example. Instead, we make direct use of the two-dimensional axisymmeric theory presented in Section 3.1.2. Thus the solution domain in the $r,z$-plane is defined by the boundary line ABCDEF in Figure 3.22 (b). We make the following assumptions on the data: the material (aluminum alloy) has $k = 190\,\mathrm{W/(m\cdot K)}$. The temperature at the inner wall $s_{\mathrm{FA}} = s_{\mathrm{D}}$ is $\bar{T} = 500\,\mathrm{K}$. On the surface part $s_{\mathrm{BCDE}} = s_{\mathrm{R}}$ convective heat transfer is taking place with $T_\infty = 300\,\mathrm{K}$ and $h = 40\,\mathrm{W/(m^2\cdot K)}$. Due to assumed symmetry in the cylinder axis direction, parts $s_{\mathrm{AB}}$ and $s_{\mathrm{EF}}$ are Neumann boundaries with the given $\bar{q} = 0$, so they give no terms into the weak form. The standard weak form using formulas (3.1.35) and (3.1.37) becomes



$$\int_A \left( \frac{\partial w}{\partial r} k \frac{\partial T}{\partial r} + \frac{\partial w}{\partial z} k \frac{\partial T}{\partial z} \right) r\,\mathrm{d}A + \int_{s_{\mathrm{R}}} wh(T - T_\infty) r\,\mathrm{d}s = 0 \tag{13}$$

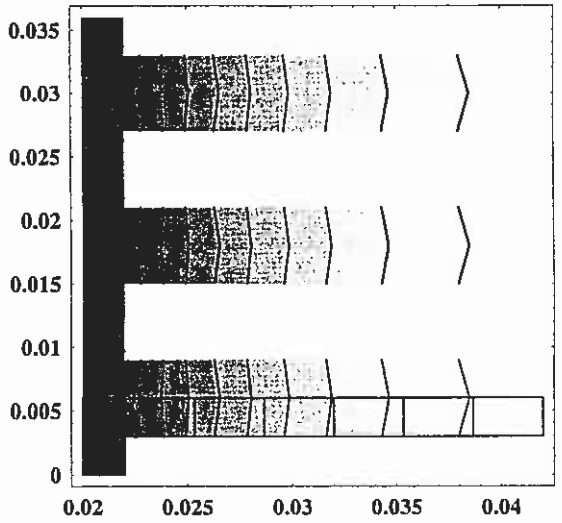
The heat flow rate is given by

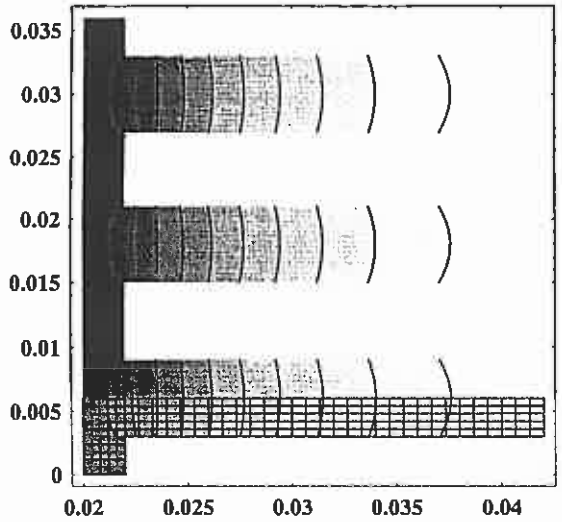
$$\mathring{Q} = 2\pi \int_{s_{\mathrm{R}}} h(T - T_\infty) r\,\mathrm{d}s \tag{14}$$

with finite elements $T$ above replaced by $\bar{T}$.

Some results obtained by two meshes are shown in Figure 3.23. Although the calculations are performed for the part shown in Figure 3.22 (b), the results are shown for clarity for a larger region by repeating the solution few times. In addition, the scale in the $z$-axis direction has been compressed somewhat to get the pictures to fit better on the pages.

The heat flow rates found are $\mathring{Q}$ =43.084W and $\mathring{Q}$ =43.064W for the coarse and refined mesh, respectively. Here it is clearly seen from the figures — as commented on in connection with formula (3.1.40) — that the assumption of constant temperature distribution in the thickness direction in a finlike configuration is not strictly true.






(b)

**Figure 3.23** **(a)** Mesh with 8 four-noded rectangular elements and some temperature contours. **(b)** Mesh with 200 four-noded rectangular elements and some temperature contours.

### 3.5.3 Wall

A cross section of a wall under study is shown in Figure 3.24 (a). The wall consists of long 20cm×20cm lightweight concrete blocks ($k=0.8\,\mathrm{W/(m\cdot K)}$) with centrally situated circular holes (diameter 15cm) filled with rigid foam ($k=0.03\,\mathrm{W/(m\cdot K)}$). The inner side of the wall consists of a 5cm thick rigid foam plate ($k=0.03\,\mathrm{W/(m\cdot K)}$). The inside room temperature $T_\infty^+=20\,^\circ\mathrm{C}$ and the outside temperature $T_\infty^-=-10\,^\circ\mathrm{C}$ and the corresponding heat transfer coefficients are $h^+=10\,\mathrm{W/(m^2\cdot K)}$ and $h^-=30\,\mathrm{W/(m^2\cdot K)}$.



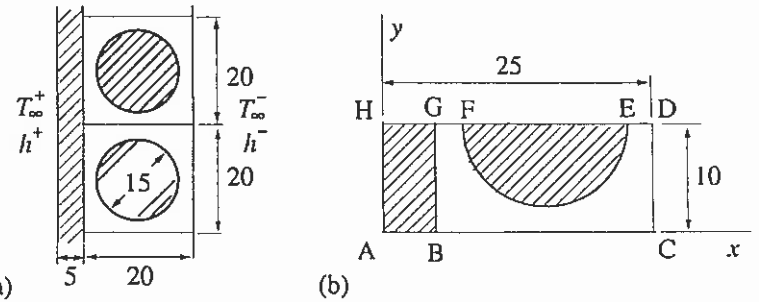



**Figure 3.24** **(a)** Cross section of a wall. **(b)** Computational domain. (measures in cm).

Due to the assumed obvious symmetries, we can restrict the calculations to a repeating wall part shown in Figure 3.24 (b). The boundary parts $s_{\mathrm{AC}}$ and $s_{\mathrm{DH}}$ are Neumann boundaries with the given heat flow rate density $\overline{q}=0$ due to symmetry so these parts again give no terms to the weak form. Parts $s_{\mathrm{AH}}$ and $s_{\mathrm{CD}}$ form together the Robin boundary. The Dirichlet boundary is missing. The weak form (3.3.3) becomes thus

$$\int_A\left(\frac{\partial w}{\partial x}k\frac{\partial T}{\partial x}+\frac{\partial w}{\partial y}k\frac{\partial T}{\partial y}\right)\mathrm{d}A$$
$$+\int_{s_{\mathrm{AH}}} wh^+\left(T-T_\infty^+\right)\mathrm{d}s+\int_{s_{\mathrm{CD}}} wh^-\left(T-T_\infty^-\right)\mathrm{d}s=0 \qquad (15)$$

With diffusivity having jumps — as here the heat conductivity between the two material interfaces — it is understandably wise to try construct the mesh so that the element boundaries follow the interfaces. The exact solution heat flow rate density is continuous over the interfaces but the derivatives of temperature have in general jumps there. It is realized that the conventional $C^0$ elements can then model this jump and have in fact in such situations advantage over possible $C^1$ elements which are too smooth. The mesh used and some temperature contour plots obtained are shown in Figure 3.25 (a). Again, although the calculations are performed for the part shown in Figure 3.24 (b), the results are shown for clarity for a larger region by repeating the solution a few times. It is seen that the temperature change happens practically already in the foam plate. To obtain a more interesting temperature distribution, the problem has been solved again with the heat conductivity of the foam material increased artificially to half of that of concrete. It is now clearly seen how the temperature contours get kinks on the material interfaces.

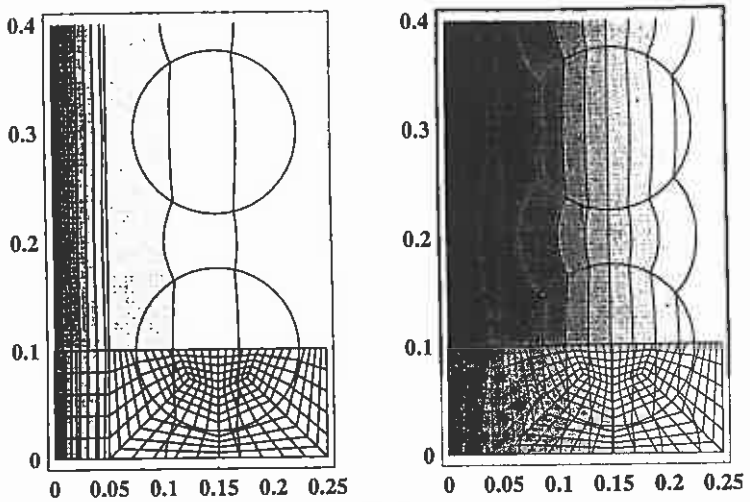



**Figure 3.25 (a)** Mesh with four-noded quadrilateral elements and some temperatrure contour plots **(b)** The same with the heat conductivity of the foam changed to $(k = 0.4\,\mathrm{W/(m \cdot K)})$.

The heat flow rate through the part per unit length in the wall transverse direction out through the wall is obtained here either from

$$\dot{Q}^* = -\int_{s_{AH}} wh^+ \left(T - T_\infty^+\right) \mathrm{d}s \tag{16}$$

or from

$$\dot{Q}^* = \int_{s_{CD}} wh^- \left(T - T_\infty^-\right) \mathrm{d}s \tag{17}$$

with $T$ replaced by $\bar{T}$. Use of formula (17) gives 1.24W/m and 3.46W/m as the heat flow rate per unit length in case (a) and (b), respectively.

## REFERENCES

Akin, J. E. (1994). *Finite Elements for Analysis and Design*, Academic Press, London, ISBN 0-12-047654-1.

Beer, G. and J. O. Watson. (1992). *Introduction to Finite and Boundary Element Methods for Engineers*, Wiley, Chichester, ISBN 0-471-92813-5.

Belytschko, T., Liu, W. K. and Moran, B. (2000). *Nonlinear Finite Elements for Continua and Structures*, Wiley, Chichester, ISBN 471-988774-3.



Çengel, Y. A. (1998). *Heat Transfer, A Practical Approach*, McGraw-Hill, Boston, ISBN 0-07-115223-7.

Hammer, P. C., Marlowe, O. P. and Stroud, A. H. (1956). Numerical integration over simplexes and cones, *Math. Tables Aids Comp.*, Vol. 10, 130-137.

Hellen, T. K. (1972). Effective quadrature rules for quadratic solid isoparametric finite elements, *Int. J. Num. Meth. Eng.*, Vol. 4, 597-600.

Incropera, F. P. and DeWitt, D. P. (1996). *Fundamentals of Heat Transfer*, 4th ed.,Wiley, New York, ISBN 0-471-30460-3.

Irons, B. M. (1971). Quadrature rules for brick based finite elements, *Int. J. Num. Meth. Eng.*, Vol. 3.

Joe, B. (1986): Delaunay triangular meshes in convex polygons, *SIAM J. Sci. Stat. Comput.*, Vol. 7, 514-539.

Stroud, A. H. and Secrest, D. (1966). *Gaussian Quadrature Formulas*, Prentice-Hall, Englewoods Cliffs

Zienkiewicz, O. C. and Taylor, R. L. (2000). *The Finite Element Method*, 5th ed., Butterworth-Heineman, Oxford. Vol. 1: *The Basis*, ISBN 0 7506 5049 4. Vol 2: *Solid Mechanics*, ISBN 0 7506 5055 9. Vol 3: *Fluid Dynamics*, ISBN 0 7506 5050 8.

## PROBLEMS

# 4 CONVERGENCE AND ERROR ANALYSIS

## 4.1 INTRODUCTION

The shape functions $N_j$ in the finite element approximation

$$\tilde{\phi} = \sum_j N_j \phi_j \tag{1}$$

for a quantity $\phi$ must presumably satisfy some general requirements for the finite element method to give converging results when the element mesh is made more and more dense. Some properties of the shape functions have been considered already in Section 3.2.3. Here the matter is studied from the point of view of evaluation of the integrals in the weak forms. The treatment is not rigorous and it is just meant to give a qualitative idea of the main factors behind convergence.

As a simple example we recall the one-dimensional heat conduction weak form (2.1.28):

$$F \equiv \int_\Omega \frac{\mathrm{d}w}{\mathrm{d}x} k \frac{\mathrm{d}T}{\mathrm{d}x} \mathrm{d}\Omega - \int_\Omega ws \,\mathrm{d}\Omega + w\bar{q}\big|_{\Gamma_N} = 0 \tag{2}$$

and the analogue for obtaining the discrete equations (2.3.5):

$$\tilde{F} \equiv \int_\Omega \frac{\mathrm{d}\tilde{w}}{\mathrm{d}x} k \frac{\mathrm{d}\tilde{T}}{\mathrm{d}x} \mathrm{d}\Omega - \int_\Omega \tilde{w}s \,\mathrm{d}\Omega + \tilde{w}\bar{q}\big|_{\Gamma_N} = 0 \tag{3}$$

We have introduced some new notation in the latter equation, which will be explained in more detail in the next section. The finite dimensional weighting function $\tilde{w}$ is taken consecutively as $N_i$ in generating the discrete equations.

To obtain a notationally more general presentation we, however, continue here by expressing (2) and (3) with the general D-C-R equation notation:

$$F \equiv \int_\Omega \frac{\mathrm{d}w}{\mathrm{d}x} D \frac{\mathrm{d}\phi}{\mathrm{d}x} \mathrm{d}\Omega - \int_\Omega wf \,\mathrm{d}\Omega + w\bar{j}^{\mathrm{d}}\big|_{\Gamma_N} = 0 \tag{4}$$

and

$$\tilde{F} \equiv \int_\Omega \frac{\mathrm{d}\tilde{w}}{\mathrm{d}x} D \frac{\mathrm{d}\tilde{\phi}}{\mathrm{d}x} \mathrm{d}\Omega - \int_\Omega \tilde{w}f \,\mathrm{d}\Omega + \tilde{w}\bar{j}^{\mathrm{d}}\big|_{\Gamma_N} = 0 \tag{5}$$





The left-hand side of a weak form is generally (in one dimension and for one unknown function) of the type

$$F = \int_\Omega f\left( x, w, \frac{\mathrm{d}w}{\mathrm{d}x}, \cdots, \phi, \frac{\mathrm{d}\phi}{\mathrm{d}x}, \frac{\mathrm{d}^2\phi}{\mathrm{d}x^2}, D, f, \cdots \right) \mathrm{d}\Omega \tag{6}$$

(The general integrand function notation $f$ and the source term notation $f$ should not produce confusion here.) It seems obvious that the terms in the integrand should tend to those of the exact solution when the mesh gets denser.

Engineering literature dealing with the finite element method gives roughly the following two main requirements for convergence, Zienkiewicz (1971), (1975):

> *Condition 1* (completeness condition): The element shape functions should be able to present with a suitable selection of the nodal parameter values in the limit as the element size tends to zero at least any given constant value in the element for any of the derivatives of $\phi$ appearing in the weak form. (7)

> *Condition 2* (continuity condition): The finite element approximation must be at least $C^{m-1}$ continuous where $m$ is the order of the highest order derivative of $\phi$ appearing in the weak form. (8)

Conditions 1 and 2 are considered in general to be sufficient to guarantee the convergence of the formulation. They are, however, not necessary, as there exist converging formulations violating the second condition.

The completeness condition means roughly the following. When the element mesh is made denser, the values of $\phi$ and its derivatives in the weak form should tend to the exact values. When the elements get smaller and smaller, the values corresponding to the exact solution change less and less in each element and in other words these values can be approximated more and more accurately by constants in each element. (This resembles the concept of the Riemann sum in the definition of the definite integral.) If the finite element is not even able to produce in the limit these constant values, it cannot give convergence.

**Remark 4.1.** Often some useful information can be obtained by assuming a tentative finite element solution, which is *nodally exact*. It can also be called the *interpolant to the exact solution* (tarkan ratkaisun interpolantti). This is in fact the goal we strive for constantly in the formulations in the following chapters. A nodally exact solution *cannot be unstable* and it is a very good starting point for *adaptive procedures* or for *post-processing*. At least in one dimension we can then consider each element as a new solution domain with *exact boundary conditions*. □

Let us consider as a further illustration of the completeness condition the situation in Figure 4.1. In Figure (a) the element is two-noded (linear) line element and we assume in the spirit of Remark 4.1 that the approximate solution is nodally exact. It is realized from the figure that function $\phi$ and its first derivative $\mathrm{d}\phi/\mathrm{d}x$ are approximated more and more accurately as the element size gets smaller but the approximation for the second derivative $\mathrm{d}^2\phi/\mathrm{d}x^2$ remains all the time at the unrealistic value zero. If the second derivative $\mathrm{d}^2\phi/\mathrm{d}x^2$ is present the weak form, this element would not satisfy the completeness condition. It should be noted from the figure that if $\phi(x)$ is smooth enough in the element (of class $C^1$) the finite element constant value $\mathrm{d}\bar{\phi}/\mathrm{d}x$ in the element coincides with the exact derivative value $\mathrm{d}\phi/\mathrm{d}x$ at least at one point (point P) in the element. Thus also that part of the weak form integrand containing the first derivative is evaluated correctly in the limit. For instance a three-noded (quadratic) line element could be able to approximate a non-zero second derivative and would now satisfy the completeness condition (but not in fact the continuity condition). In the diffusion problem weak form (4) only the first derivative $\mathrm{d}\phi/\mathrm{d}x$ appears and the linear line element clearly satisfies the completeness condition.

Figure (b) shows the case where the element is one-noded (constant) element. The element satisfies the completeness condition only if no derivatives of $\phi$ appear in the weak form.

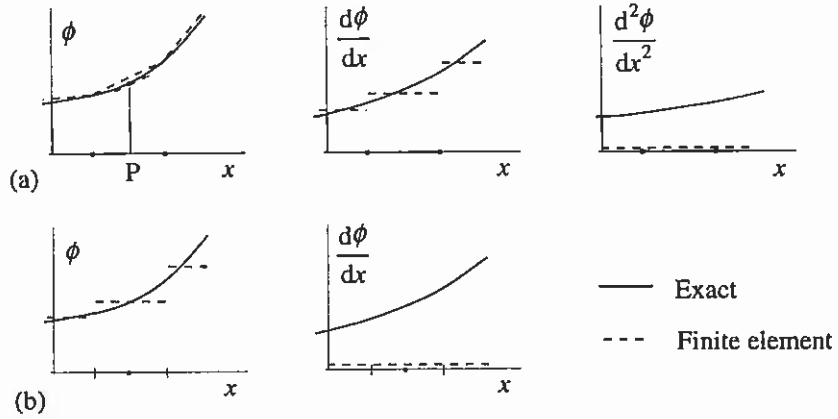



**Figure 4.1 (a)** Linear line elements. **(b)** Constant line elements.

The completeness condition is fulfilled in most elements already with a finite size. It is recalled from Section 3.2.3 that the isoparametric elements considered



there can represent any first degree function in the global coordinates and thus also any constant first derivative value.

The weak form integrands contain in addition to the unknown function and its derivatives given functions such as the diffusivity $D$, the source term $f$ etc. Approximation of given functions is of course in principle not necessary. In practice they are however normally approximated to ease the computations. It is obvious from the previous discussions that it is enough to take a suitable constant approximation (the value say at the midpoint could be used) — naturally a more accurate presentation is allowed — to achieve convergence in this respect.

The continuity condition is critically present at the element interfaces. Inside the elements this condition is normally automatically satisfied. For instance with the conventional polynomial approximation the function and all of its derivatives are continuous to any order; the representation is of class $C^\infty$ inside element. To achieve $C^1$ continuity at the element interfaces is already awkward in two and three dimensions. $C^0 \equiv C$ continuity is easy to achieve as we have seen with isoparametric elements. This is enough for problems where the highest derivatives in the weak form are of first order as will be the case in the problems considered in this text.

Elements satisfying the continuity condition are often called *conforming elements* (konforminen elementti) and those violating it are called *non-conforming* elements (epäkonforminen elementti). Sometimes the name *compatible element* (yhteensopiva elementti) is used for a conforming element. The term has its origin in solid mechanics where the displacement components are the basic unknowns. If the approximation is continuous enough, the elements fit to each other after deformation without gaps; we have compatible elements.

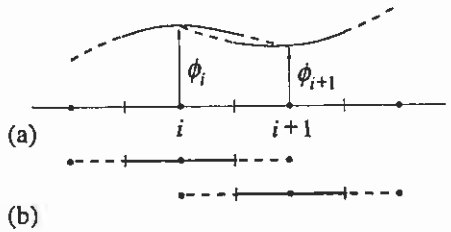



**Figure 4.2 (a)** Approximation. **(b)** Two elements.

Figure 4.2 shows an example of a rather exotic non-conforming element. The element is a three-noded line element but the two end nodes are outside the element domain. This is a one-dimensional version of a triangular element used

in plate bending, Ney and Utku (1972). The element does not even give a $C^0$ continuous approximation but seems still to work satisfactory in cases where the weak form contains second order derivatives.

The continuity condition can be illustrated in the following way, Zienkiewicz (1971). The question is still of the evaluation of the terms in the weak form but now we study the integrability of $\phi$ and its derivatives. Let us first think the element interfaces as thin but finite strips on which the element approximation is completed smoothly so that the integrals over the whole domain can be safely evaluated taking the contribution from the strips into account. The thicknesses of the strips are now let to approach zero. If the contributions from the strips tend to zero, the integrals can be correctly evaluated just from the elements as is assumed for instance in the basic formula (2.3.20). The presentation of Figure 4.3 shows that in the case of $C^0$ elements the first derivative stays finite in the strip but the second derivative grows without limit. If the integrand contains only the first derivative, the contribution from the strips goes to zero. The second derivative leads to an undefined form of the type $\infty \cdot 0$, and its contribution may be non-zero so the correctness on the basic formula is unsure in this case.

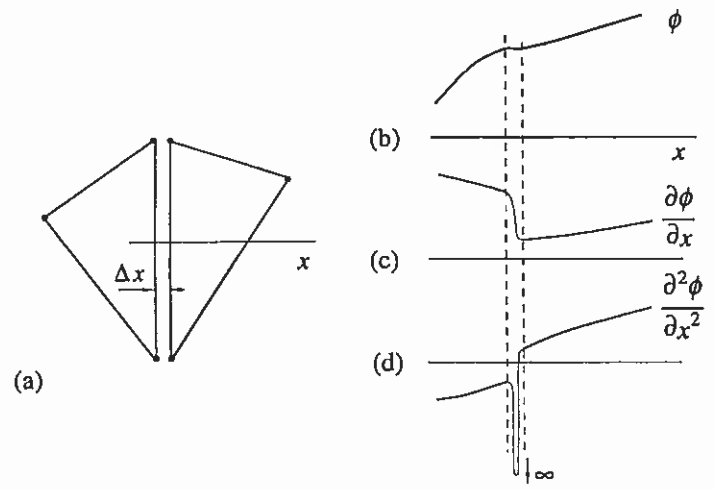



**Figure 4.3** **(a)** A strip of thickness $\Delta x$. **(b)** $C^0$ continuous function. **(c)** First derivative. **(d)** Second derivative.

It should be noted that the continuity condition must be applied also to the weighting function to assure the correct evaluation of the appropriate terms just from the elements.



As conforming elements are in some problems complicated to generate, the temptation is great to use non-conforming elements. To deal with them mathematically is not an easy theme. However, a useful engineering approach to clarify their acceptability has been developed. It is called the *patch test* (tilkkutesti).

The idea of the patch test is as follows. Let us consider as an example the part of a mesh of triangular elements shown in Figure 4.4 where all the elements connected to a node $i$ are present. The set of elements is called in this connection a *patch* (tilkku).

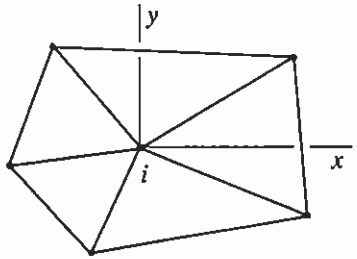



**Figure 4.4** Patch.

Let the governing linear (or linearized) field equation be

$$L(\phi) - f = 0 \tag{9}$$

with appropriate linear boundary conditions. $L$ is a linear differential operator, for instance in the two-dimensional pure diffusion case with isotropic diffusivity

$$L = \frac{\partial}{\partial x}\left(-D\frac{\partial}{\partial x}\right) + \frac{\partial}{\partial y}\left(-D\frac{\partial}{\partial y}\right) \tag{10}$$

We think the exact solution $\phi(x, y)$ expanded by Taylor's formula with node $i$ as the expansion center:

$$\phi(x,y) = \phi_0 + \left(\frac{\partial\phi}{\partial x}\right)_0 x + \left(\frac{\partial\phi}{\partial y}\right)_0 y + \frac{1}{2}\left(\frac{\partial^2\phi}{\partial x^2}\right)_0 x^2 + \cdots \tag{11}$$

(The subscript 0 indicates to a value evaluated at the local origin = node $i$ in Figure 4.4.) We have a polynomial representation. It seems obvious that the finite element solution is the better the higher order polynomial solution it can simulate by producing the exact nodal values. This is studied with the help of

the patch by first taking consecutive polynomials (we change the notation from (11) for convenience of presentation)

$$\phi = \alpha, \quad \phi = \alpha x + \beta y, \quad \phi = \alpha x^2 + \beta xy + \gamma y^2, \cdots \tag{12}$$

the coefficients of which are arbitrarily selected and by fixing the nodal values of all the nodes except of node $i$ with the values calculated directly from the polynomial. The nodal value $\phi_i$ of node $i$ is determined from the finite element system equation

$$\sum_j K_{ij}\phi_j = b_i \tag{13}$$

for the patch where thus the values $\phi_j$, $j \neq i$ are known. *If the nodal value $\phi_i$ obtained is equal to the nodal value calculated from the corresponding polynomial, the patch test is passed for the polynomial degree in question.*

> A necessary and sufficient condition for convergence is that the element must pass the patch test up to polynomial of degree $m$ where $m$ is the order of the highest order derivative of the unknown function appearing in the weak form. (14)

(This is the general opinion in the engineering literature although the exact mathematical proof seems to be missing.)

For the higher degree $\geq m$ polynomial the patch test is passed, the better rate of convergence is to be expected when the mesh gets denser.

**Remark 4.2.** It should be noted that the test is performed in principle for an infinitesimal patch. In practice the patch is naturally of finite size. Because of this, if the coefficients of the derivatives in the differential equation depend on position, they should be given some constant values to be used in the test. For instance, in axisymmetric formulations the radial coordinate $r$ often appears as a coefficient. It must thus be given a constant value when applying the test. Similarly the source term $f$ in the differential equation (9) (now with a constant $D$) must be taken according to the assumed polynomial solution; see Example 4.1. □

**Remark 4.3.** Let us consider for instance the quadratic expression in (12). In fact, it is enough to perform the patch test first only for $\phi = x^2$, then for $\phi = xy$ and finally for $\phi = y^2$. The multipliers $\alpha$, $\beta$, $\gamma$ are needed in general in theory to give the right physical dimension for $\phi$. The multipliers are, however, seen just to multiply both the nodal values and the possible source term equally and they cancel in the test. Further, as the finite element system equations are linear with respect to the nodal values and with respect to the source term, the patch test is then seen to be passed for the linear combination $\alpha x^2 + \beta xy + \gamma y^2$ if it is passed separately for $x^2$, $xy$ and $y^2$. Similar themes will be discussed in Remarks 5.9 and 5.11. □



For instance some quadrilateral elements have been found to fail in the test when the elements are of arbitrary shape but to pass if they are parallelograms. Thus when applying the test, one should select some arbitrary irregular geometry so that it is then highly unprobable that the test would be passed by chance.

The patch test was originally an ingenious insight by Irons (1972). It can be applied in addition to testing the convergence of non-conforming elements also say for checking the effect of a possibly too coarse numerical quadrature rule or just to check the working of a new program using standard elements. Reference Taylor et al. (1986) discusses many features of the patch test and also deals with boundary conditions.

**Example 4.1.** We consider the steady one-dimensional D-C-R equation

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(-D\frac{\mathrm{d}\phi}{\mathrm{d}x}\right) + \frac{\mathrm{d}}{\mathrm{d}x}(u\phi) + c\phi - f = 0 \tag{a}$$

and demonstrate the application of the patch test with two-noded line elements and the standard Galerkin formulation.

According to Remark 4.2 we can continue by taking constant values for $D$, $u$ and $c$ to obtain the field equation

$$-D\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2} + u\frac{\mathrm{d}\phi}{\mathrm{d}x} + c\phi - f = 0 \tag{b}$$

The weak form is (This is considered in more detail later. In any case, it is quite obvious how to proceed: field equation (b) is multiplied by a weighting function, integrated over the domain and integration by parts is effected with respect to the diffusion term. Finally information about boundary conditions is introduced which phase is not considered here.)

$$D\int_\Omega \frac{\mathrm{d}w}{\mathrm{d}x}\frac{\mathrm{d}\phi}{\mathrm{d}x}\,\mathrm{d}\Omega + u\int_\Omega w\frac{\mathrm{d}\phi}{\mathrm{d}x}\,\mathrm{d}\Omega + c\int_\Omega w\phi\,\mathrm{d}\Omega - \int_\Omega wf\,\mathrm{d}\Omega + \mathrm{bt} = 0 \tag{c}$$

As we do not deal with the boundary conditions, the exact content of the notation bt referring to some boundary terms is not needed here. A typical system equation is

$$\sum_j K_{ij}\phi_j = b_i \tag{d}$$

with (for a node inside the mesh)

$$\begin{aligned} K_{ij} &= D\int_\Omega \frac{\mathrm{d}N_i}{\mathrm{d}x}\frac{\mathrm{d}N_j}{\mathrm{d}x}\,\mathrm{d}\Omega + u\int_\Omega N_i\frac{\mathrm{d}N_j}{\mathrm{d}x}\,\mathrm{d}\Omega + c\int_\Omega N_i N_j\,\mathrm{d}\Omega \\ b_i &= \int_\Omega N_i f\,\mathrm{d}\Omega \end{aligned} \tag{e}$$

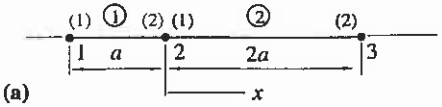



**Figure (a)**

We take the patch shown in Figure (a) consisting of two elements with the lengths

$$h^{(1)} = a, \qquad h^{(2)} = 2a \tag{f}$$

Of course any other combination for the lengths could have been taken but this should suffice as an example.

We first write down the element contributions. Remembering Remark 2.10 and making use of formulas (F.1.1) gives

$$\begin{aligned} [K]^e &= \frac{D}{h^e}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \frac{u}{2}\begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix} + \frac{c h^e}{6}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \\ \{b\}^e &= \int_{\Omega^e} \begin{Bmatrix} N_1^e f \\ N_2^e f \end{Bmatrix} \mathrm{d}\Omega \end{aligned} \tag{g}$$

The column matrix depends on the expression for $f$ and is dealt with later.

In the system equation

$$K_{21}\phi_1 + K_{22}\phi_2 + K_{23}\phi_3 = b_2 \tag{h}$$

for the internal patch node 2, the assembly process gives the expressions

$$\begin{aligned} &K_{21} = K_{21}^1, \quad K_{22} = K_{22}^1 + K_{11}^2, \quad K_{23} = K_{12}^2 \\ &b_2 = b_2^1 + b_1^2 \end{aligned} \tag{i}$$

Taking expressions (f) into account, there is obtained

$$\begin{aligned} K_{21} &= -\frac{D}{a} - \frac{u}{2} + \frac{ca}{6} \\ K_{22} &= \frac{D}{a} + \frac{u}{2} + \frac{ca}{3} + \frac{D}{2a} - \frac{u}{2} + \frac{2ca}{3} = \frac{3D}{2a} + ca \\ K_{23} &= -\frac{D}{2a} + \frac{u}{2} + \frac{ca}{3} \end{aligned} \tag{j}$$

The expression for the nodal value $\phi_2$, determined from the system equation (h) is thus

$$\phi_2 = \frac{1}{3D/2a + ca}\left[\left(\frac{D}{a} + \frac{u}{2} - \frac{ca}{6}\right)\phi_1 + \left(\frac{D}{2a} - \frac{u}{2} - \frac{ca}{3}\right)\phi_3 + b_2\right] \tag{k}$$

We can now start to apply the patch test.

(1) The first polynomial is the constant (see Remark 4.3)

$$\phi(x) = 1 \tag{l}$$



giving the nodal values

$$\phi_1 = 1, \quad \phi_3 = 1 \tag{m}$$

and to pass the patch test we should finally get from (k) the result $\phi_2 = 1$.

For the assumed exact solution (l) we determine from (b) according to Remark 4.2 the corresponding source term

$$f = c \tag{n}$$

It should be noted that it is a useful trick in general to introduce exact solution benchmark problems by taking any expression for the unknown and then to calculate from the field equation the source term needed to satisfy the equation.

We obtain making use of formulas (F.1.1)

$$\begin{aligned} \{b\}^1 &= c\int_{\Omega^1} \begin{Bmatrix} N_1^1 \\ N_2^1 \end{Bmatrix} \mathrm{d}\Omega = \frac{ca}{2}\begin{Bmatrix} 1 \\ 1 \end{Bmatrix} \\ \{b\}^2 &= c\int_{\Omega^2} \begin{Bmatrix} N_1^2 \\ N_2^2 \end{Bmatrix} \mathrm{d}\Omega = ca\begin{Bmatrix} 1 \\ 1 \end{Bmatrix} \end{aligned} \tag{o}$$

and

$$b_2 = \frac{ca}{2} + ca = \frac{3ca}{2} \tag{p}$$

Substitution of (m) and (p) into (k) gives

$$\begin{aligned} \phi_2 &= \frac{1}{3D/2a + ca}\left[\left(\frac{D}{a} + \frac{u}{2} - \frac{ca}{6}\right) + \left(\frac{D}{2a} - \frac{u}{2} - \frac{ca}{3}\right) + \frac{3ca}{2}\right] \\ &= \frac{1}{3D/2a + ca}\left(\frac{D}{a} + \frac{u}{2} - \frac{ca}{6} + \frac{D}{2a} - \frac{u}{2} - \frac{ca}{3} + \frac{3ca}{2}\right) \\ &= \frac{1}{3D/2a + ca}\left(\frac{3D}{2a} + ca\right) = 1 \end{aligned} \tag{q}$$

The patch test is thus passed.

(2) The next polynomial is

$$\phi(x) = x \tag{r}$$

giving

$$\phi_1 = -a, \quad \phi_3 = 2a, \quad (\phi_2 = 0) \tag{s}$$

and

$$f = u + cx \tag{t}$$

After some steps we get

(u)

and the patch test is again found to be passed. Thus according to the patch test criterion (14) we have a convergent formulation ($m = 1$).

(3) The patch test is no more passed for the quadratic polynomial $\phi(x) = x^2$. Instead of the correct value $\phi_2 = 0$, there is obtained

$$\phi_2 = -\frac{1}{3D/2a + ca}\left(\frac{ua^2}{2} + \frac{3ca^3}{4}\right) \tag{v}$$

Without convection and reaction terms the patch test clearly would have been passed (the numerator of (v) goes to zero) even for the quadratic polynomial; in fact, according to Section 4.2.5 it will be satisfied for any degree polynomial.

It should be realized that the patch test is normally performed numerically. In this very simple demonstration example we have proceeded with closed form formulas which reveal more.

In practice it is not enough that a formulation is convergent. The important thing is that accurate enough results are obtained with reasonable meshes. The formulation in Example 4.1 was found to be convergent but it will be seen in Chapter 6 that when convection is large, the formulation is totally useless.

The term *discretization error* (diskretointivirhe) refers to those errors which are due to the finite element model imagining that all calculations have been performed with infinite accuracy. Discretization error includes thus the errors due to approximation of the unknown and given functions, approximation of the boundary, use of numerical quadrature, etc. Additional errors due to the fact that the calculations are in reality performed with finite accuracy in a computer are called *round-off errors* (pyöristysvirhe). Their effect normally grows when the mesh is made denser so that a theoretically convergent formulation can in fact lead at some point to divergence. In practice only an additional calculation with double precision can give some information about the round-off.

## 4.2 THEORETICAL BASIS

The main idea here is to give the reader some familiarity with the basic concepts and shorthand notations used in formal mathematical treatments of the finite element method. This may help in following more advanced literature on the subject. The discussion will concern mainly problems in one dimension.

### 4.2.1 Convergence rate

In addition to the pointwise *error* (virhe) $e(x)$ appearing in



$$\phi(x) = \tilde{\phi}(x) + e(x) \tag{1}$$

where $\phi(x)$ is the exact solution and $\tilde{\phi}(x)$ the finite element solution, more complicated error measures called error norms (see Section C.4) are in use in the finite element mathematics. The standard notation for a norm is $\|\cdot\|$. A suitable error norm

$$\|e\| = \|\phi - \tilde{\phi}\| \tag{2}$$

describes in some abstract manner the distance of the two functions $\phi$ and $\tilde{\phi}$ in the function space (see Section C.1).

The so-called *a priori error estimates* (etukäteisvirhearvio) in the finite element method are typically of the form

$$\|e\| \le Ch^q \tag{3}$$

Here $C$ is a problem dependent positive coefficient depending for instance on the element type, the smoothness of the exact solution, etc. Quantity $h$ is an agreed linear measure, the so-called *mesh parameter* (verkkoparametri), describing the density of the finite element mesh. The denser the mesh, the smaller the $h$. In two dimensions $h$ could be for example the diameter of the smallest circle containing the largest element in the mesh. The exponent $q$, which must be positive for a convergent formulation, gives the so-called *asymptotic rate of convergence* (asymptoottinen suppenemisnopeus) of the method. The larger the $q$, the faster the solution converges with the refinement of the mesh (Figure 4.5).

The theory gives the value of the exponent $q$ but in general not a practically useful estimate for the coefficient $C$. In practice it would be important to know the smallest possible $C$ for which (3) is still valid, that is, when it becomes an equation

$$\|e\| = Ch^q \tag{4}$$

The content of formula (3) or (4) is often given also in the form

$$\|e\| = O\left(h^q\right) \tag{5}$$

The notation $O(h^q)$ is in words "order of magnitude $h^q$ term" (kertaluokkaa $h^q$ oleva termi), that is, it is a term which behaves essentially as a constant multiplied by $h^q$ when $h$ is small enough.

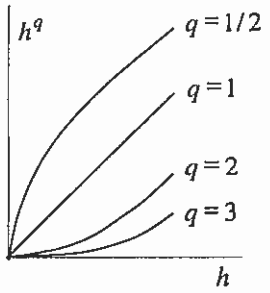



**Figure 4.5** $h^q$ versus $h$.

The rate of convergence can be studied also by numerical experiments in cases where the analytical solution is known by determining the finite element solution say with three similar meshes of different density. The error $\|e\|$ can then be calculated directly afterwards. The examination of the results gets easier, when the assumed relationship (4) is transformed by taking the logarithm:

$$\log\|e\| = \log C + q\log h \tag{6}$$

This is the equation for a line in the $\log h, \log\|e\|$-coordinate system and $q$ is clearly the slope of the line (Figure 4.6 (a)). This makes it possible to determine the rate of convergence conveniently experimentally. A prerequisite is, that the meshes are dense enough so that the asymptotic rate of convergence behavior has been reached and that round-off has not spoiled the results, so that the three points calculated stay roughly on a line. In physical problems $\|e\|$ and $h$ are in general dimensional quantities — if the problem has not been cast beforehand into a dimensionless form — so that it is necessary to use the type of representation given in Figure 4.6 (b), where $\|e\|_r$ and $h_r$ are some suitable reference values. (It is to be noticed that the logarithm cannot be taken of a dimensional quantity.)

Knowledge of the rate of convergence is important for instance in connection with adaptive procedures, where the mesh is refined for further calculations locally in different way in different parts of the domain based on previous results.



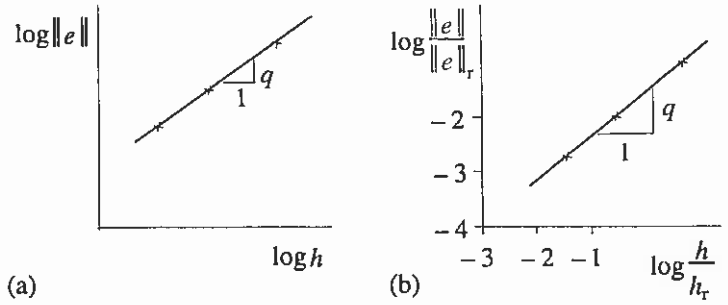



**Figure 4.6** Numerical determination of the rate of convergence.

The so-called *Richardson extrapolation*, e.g., Crandall (1956, p.171), can be made use of in extrapolating results from consecutive calculations towards the limit $h \to 0$. Let us consider say the determination of the value of a quantity $\phi$ at a certain point. Let the values obtained by using consecutive mesh parameters $h_1, h_2, \cdots$ be $\tilde{\phi}_1, \tilde{\phi}_2, \cdots$, respectively. We mark the points $(h_1, \tilde{\phi}_1)$, $(h_2, \tilde{\phi}_2)$, $\cdots$ on the $h,\phi$-plane and construct a polynomial going through them and finally evaluate it at $h=0$ to get an estimate on the exact $\phi$. This procedure in fact does not demand information about the exponent $q$. If it is known, however, we can write for two consecutive calculations for instance

$$\begin{aligned} e_1 &= \phi - \tilde{\phi}_1 = Ch_1^q \\ e_2 &= \phi - \tilde{\phi}_2 = Ch_2^q = \alpha^q Ch_1^q \end{aligned} \tag{7}$$

in which $h_2 = \alpha h_1$. $\alpha$ is a dimensionless number describing the refinement, often $\alpha = 1/2$. By eliminating the unknown $Ch_1^q$ from (7), we obtain the extrapolated estimate

$$\phi = \frac{\tilde{\phi}_2 - \alpha^q \tilde{\phi}_1}{1-\alpha^q} \tag{8}$$

(Formulas (7) are of the form (4). As we are not using absolute values on the left-hand sides, the multiplier $C$ can be here also negative.) We can also deduce from (7) the result

$$e_2 = \frac{\alpha^q}{1-\alpha^q}\left(\tilde{\phi}_2 - \tilde{\phi}_1\right) \tag{9}$$

These formulas can be employed to estimate exact values without knowing the exact solution.

### 4.2.2 Weak forms formalized

Let us first consider the one-dimensional pure diffusion problem with the boundary consisting of the Dirichlet type only; equation (4.1.4) without $\Gamma_N$. (Some of the notations to be introduced are explained in the NOMENCLATURE section.) The weak formulation is: Find $\phi \in S$ such that

$$\boxed{\int_\Omega \frac{dw}{dx} D \frac{d\phi}{dx} d\Omega - \int_\Omega w f \, d\Omega = 0} \quad \forall w \in V \tag{10}$$

Here $V$ is called the *weighting or test function space* (paino- eli testifunktioavaruus), that is, roughly the set

$$V = \left\{ u : u \in C(\bar{\Omega}),\ u = 0 \text{ on } \Gamma_D \right\} \tag{11}$$

The function $\phi$ to be determined is a member of the *trial function set* (yritefunktiojoukko) $S$ where

$$S = \left\{ u : u \in C(\bar{\Omega}),\ u = \bar{u} \text{ on } \Gamma_D \right\} \tag{12}$$

It is seen that the difference between $V$ and $S$ consists merely of the conditions on $\Gamma_D$; the members of $S$ must satisfy the Dirichlet boundary conditions and the members of $V$ the same but only in the homogeneous (= right-hand side is zero) form. This means that $V$ is a linear space (see Section C.2) as the sum $u + v$ of two members $u$ and $v$ is seen to be again a member of the space. The set $S$ is not, however, a linear space if the given Dirichlet boundary data $\bar{u} \neq 0$ as the sum of two members does not clearly satisfy the boundary condition.

**Remark 4.4.** A more precise characterization of the sets $V$ and $S$ than that given above consists of saying that $u$ belongs to the set $H^1$ of functions having the $L_2$-norms of the function itself and of its first (generalized) derivative bounded, Hughes (1987). Often the notation $H_0^1$ is used for the weighting function space to indicate that the satisfaction of the homogeneous Dirichlet boundary condition is included. In two- or three-dimensional cases all the first order (generalized) derivatives are included in the definition. □

**Remark 4.5.** Using the deltaform described in Remark 2.15 simplifies the presentation. Instead of (10) we can state: Find $\phi - \bar{\phi}\ (= \Delta\phi) \in V$ such that

$$\boxed{\int_\Omega \frac{dw}{dx} D \frac{d\phi}{dx} d\Omega - \int_\Omega w f \, d\Omega = 0} \quad \forall w \in V \tag{13}$$

where $\phi = \bar{\phi} + \Delta\phi$. Now we can operate with only one set $V$. □



The boundary conditions that the members of the trial function set $S$ must satisfy *a priori* (etukäteen) in a weak formulation are called *essential boundary conditions* (oleellinen reunehto) similarly as for functionals (see Appendix D). In the standard energy equation weak form the essential boundary conditions are the Dirchlet conditions. Those boundary condition that are consequences of the satisfaction of the weak form are called *natural boundary conditions* (luonnollinen reunaehto). In the standard energy equation weak form the natural boundary conditions are the Neumann and the Robin conditions. The fact that the trial functions have to satisfy beforehand only the essential boundary conditions is of utmost importance in the practical application of the finite element method.

In the finite element method, we define corresponding to (11) and (12) the sets

$$\bar{V} = \left\{ \tilde{u} : \tilde{u} \in C(\bar{\Omega}),\ u = 0 \text{ on } \Gamma_D,\ \tilde{u} \text{ is represented by the current finite element mesh} \right\} \tag{14}$$

$$\bar{S} = \left\{ \tilde{u} : \tilde{u} \in C(\bar{\Omega}),\ u = \bar{u} \text{ on } \Gamma_D,\ \tilde{u} \text{ is represented by the current finite element mesh} \right\} \tag{15}$$

Thus $\bar{V} \subset V$ and $\bar{S} \subset S$, that is, if $\tilde{u} \in \bar{V}$, then also $\tilde{u} \in V$ and similarly with $\bar{S}$ and $S$. $\bar{V}$ is called a *finite dimensional subspace* (äärellisdimensioinen aliavaruus) of $V$.

The finite element method analogue of (10) is: Find $\tilde{\phi} \in \bar{S}$ such that

$$\boxed{\int_\Omega \frac{d\tilde{w}}{dx} D \frac{d\tilde{\phi}}{dx} d\Omega - \int_\Omega \tilde{w} f \, d\Omega = 0} \quad \forall \tilde{w} \in \bar{V} \tag{16}$$

The system equations in the Galerkin method were obtained by selecting $\tilde{w}$ to be consecutively the global shape functions $N_i$. Since any $\tilde{w}$ is obtained as a linear combination of the shape functions ($\tilde{w} = \alpha N_1 + \beta N_2 + \cdots$), the weak form (16) is then clearly satisfied for any $\tilde{w}$ if the system equations are satisfied and the other way round.

The finite element method analogue of (13) is: Find $\tilde{\phi} - \bar{\phi}\ (= \Delta\tilde{\phi}) \in \bar{V}$ such that

$$\boxed{\int_\Omega \frac{d\tilde{w}}{dx} D \frac{d\tilde{\phi}}{dx} d\Omega - \int_\Omega \tilde{w} f \, d\Omega = 0} \quad \forall \tilde{w} \in \bar{V} \tag{17}$$

where $\tilde{\phi} = \bar{\phi} + \Delta\tilde{\phi}$.

**Remark 4.6.** In the literature, superscript $h$ is very often used in connection with the finite dimensional weighting function and the approximation. This convention reminds us of the fact that the sets defined depend on the mesh. To be precise, it should be noted that in our notation the tilde above the symbol of the function to be determined means approximation *but in connection with the weighting function it means the finite dimensional case*. That is, we do not approximate the weighting function, we are just forced to select it from a finite dimensional set. □

After these formal statements an important result is derived. As $\tilde{w}$ belongs to $\tilde{V}$, it also belongs to $V$ and thus (10) must hold also in the form

$$\int_\Omega \frac{\mathrm{d}\tilde{w}}{\mathrm{d}x} D \frac{\mathrm{d}\phi}{\mathrm{d}x} \mathrm{d}\Omega - \int_\Omega \tilde{w} f \,\mathrm{d}\Omega = 0 \quad \forall \tilde{w} \in \tilde{V} \tag{18}$$

Subtraction of (16) from (18) gives

$$\int_\Omega \frac{\mathrm{d}\tilde{w}}{\mathrm{d}x} D \left( \frac{\mathrm{d}\phi}{\mathrm{d}x} - \frac{\mathrm{d}\tilde{\phi}}{\mathrm{d}x} \right) \mathrm{d}\Omega = 0 \quad \forall \tilde{w} \in \tilde{V} \tag{19}$$

or ($e = \phi - \tilde{\phi}$)

$$\int_\Omega \frac{\mathrm{d}\tilde{w}}{\mathrm{d}x} D \frac{\mathrm{d}e}{\mathrm{d}x} \mathrm{d}\Omega = 0 \quad \forall \tilde{w} \in \tilde{V} \tag{20}$$

It is noticed that the source term has disappeared through this manipulation.

We now generalize and consider the full standard weak form (3.1.70): Find $\phi \in S$ such that

$$\boxed{\int_\Omega \frac{\partial w}{\partial x_\alpha} D_{\alpha\beta} \frac{\partial \phi}{\partial x_\beta} \mathrm{d}\Omega - \int_\Omega w f \,\mathrm{d}\Omega + \int_{\Gamma_\mathrm{N}} w \bar{j}^\mathrm{d} \mathrm{d}\Gamma + \int_{\Gamma_\mathrm{R}} w \left( a\phi + b \right) \mathrm{d}\Gamma = 0} \tag{21}$$

$\forall w \in V$. Here the sets $V$ and $S$ must be redefined in an obvious way from those of (11) and (12). The only change is that all partial derivatives of $u$ must be at least piecewise continuous. See, however, Remark 4.3. The weak form (21) consists of linear and constant terms with respect to function $\phi$ to be determined. Let us use, correspondingly, the shorthand notations

$$a(u,v) \equiv \int_\Omega \frac{\partial u}{\partial x_\alpha} D_{\alpha\beta} \frac{\partial v}{\partial x_\beta} \mathrm{d}\Omega + \int_{\Gamma_\mathrm{R}} u a v \mathrm{d}\Gamma \tag{22}$$

and



$$b(u) \equiv \int_\Omega u f d\Omega - \int_{\Gamma_\mathrm{N}} u \bar{j}^\mathrm{d} \mathrm{d}\Gamma - \int_{\Gamma_\mathrm{R}} u b \mathrm{d}\Gamma \tag{23}$$

These are a *bilinear form* and a *linear form*, respectively (see Section C.5). The weak form (21) is then simply

$$a(w,\phi) - b(w) = 0 \quad \forall w \in V \tag{24}$$

This type of concise representations abound in the mathematics literature on finite elements.

We may now repeat the steps to arrive at the equivalent of equation (20). The finite element system equations are obtained from

$$a\left(\tilde{w},\tilde{\phi}\right) - b\left(\tilde{w}\right) = 0 \quad \forall \tilde{w} \in \tilde{V} \tag{25}$$

Equation (24) can be written also for $\tilde{w}$:

$$a\left(\tilde{w},\phi\right) - b\left(\tilde{w}\right) = 0 \quad \forall \tilde{w} \in \tilde{V} \tag{26}$$

Subtraction of (25) from (26) gives

$$a\left(\tilde{w},\phi - \tilde{\phi}\right) = 0 \quad \forall \tilde{w} \in \tilde{V} \tag{27}$$

or

$$\boxed{a\left(\tilde{w},e\right) = 0} \quad \forall \tilde{w} \in \tilde{V} \tag{28}$$

or in detail taking the notation (22) into account

$$\int_\Omega \frac{\partial w}{\partial x_\alpha} D_{\alpha\beta} \frac{\partial e}{\partial x_\beta} \mathrm{d}\Omega + \int_{\Gamma_\mathrm{R}} \tilde{w} a e \mathrm{d}\Gamma = 0 \quad \forall \tilde{w} \in \tilde{V} \tag{29}$$

This is the analogue of (20) in the general case.

The bilinear form (22) is here also an inner product (see Section C.3). (We assume that the diffusivity tensor is symmetric and positive definite as is usually the case due to physics.) According to (28) the error $e$ of the finite element solution is orthogonal to each weighting function.

The energy norm

$$\| u \|_a \equiv a(u,u)^{1/2} = \left( \int_\Omega \frac{\partial u}{\partial x_\alpha} D_{\alpha\beta} \frac{\partial u}{\partial x_\beta} \mathrm{d}\Omega + \int_{\Gamma_\mathrm{R}} u a u \,\mathrm{d}\Gamma \right)^{1/2} \tag{30}$$

$$(u,v)\equiv\int_\Omega uv\,\mathrm{d}\Omega \tag{38}$$

and

$$\|u\|^2\equiv\int_\Omega u^2\mathrm{d}\Omega \tag{39}$$

without any subscripts. It is realized that, in one dimension, using these notations, $|u|_0=\|u\|$, $|u|_1=\|u'\|$, etc.

In the one-dimensional case the proof is rather straightforward. Let us divide $\Omega=]0,L[$ into elements $\Omega^e$ with $h=\max(h^e)$. The obvious relationship

$$f(x)=\int_{x_i}^{x}f'\mathrm{d}x\quad\left(=\Big|_{x_i}^{x}f=f(x)-f(x_i)=f(x)\right) \tag{40}$$

is valid in a typical $\Omega^e$ *for any* $C^0$ *function* $f$ *vanishing at* $x_i\in\Omega^e$. Taking absolute values on both sides of this and continuing gives first

$$\begin{aligned}|f|&=\left|\int_{x_i}^{x}f'\mathrm{d}x\right|=\left|\int_{x_i}^{x}1\cdot f'\mathrm{d}x\right|\overset{(1)}{\le}\left[\int_{x_i}^{x}1\cdot1\mathrm{d}x\right]^{1/2}\left[\int_{x_i}^{x}f'\cdot f'\mathrm{d}x\right]^{1/2}\\&\overset{(2)}{\le}(x-x_i)^{1/2}\left[\int_{\Omega^e}(f')^2\mathrm{d}x\right]^{1/2}\overset{(3)}{\le}\left(h^e\right)^{1/2}\|f'\|^e\overset{(4)}{\le}h^{1/2}\|f'\|^e\end{aligned} \tag{41}$$

The steps indicated above are in detail: (1): the Schwarz inequality for the $L_2$ inner product over $]x_i,x[$, (2): $\int_{x_i}^{x}\mathrm{pos.d}x\le\int_{\Omega^e}\mathrm{pos.d}x$, (3): $(x-x_i)^{1/2}\le(h^e)^{1/2}$, (4): $h^e\le h$. Raising both sides of (41) to the power 2 gives

$$f^2(x)\le h\left(\|f'\|^e\right)^2 \tag{42}$$

and integrating both sides of this over the element gives further (the right-hand side is a constant and $h^e\le h$)

$$\left(\|f\|^e\right)^2\le h^2\left(\|f'\|^e\right)^2 \tag{43}$$

For norms consisting of integrals the square of the total norm is the sum of the squares of the norms over the elements:



$$\|\cdot\|^2=\sum_{e=1}^{n_e}\left(\|\cdot\|^e\right)^2 \tag{44}$$

Thus summing both sides of (43) over the elements and taking the square root finally leads to

$$\|f\|\le h\|f'\| \tag{45}$$

It should be emphasized that the derivation demanded $f$ only to be continuous inside each element and to vanish at least once there.

We can now apply (45) to the function $\hat{e}=\phi-\hat{\phi}$. Assuming that *the interpolant is piecewise constant* in such a way that the value at some point in each element coincides with the exact value, we obtain

$$\|\hat{e}\|\le h\|\hat{e}'\|=h\left\|\phi'-\hat{\phi}'\right\|=h\|\phi'\| \tag{46}$$

which corresponds to (35) with $s=1$. (Note that $\hat{e}'=\phi'$ as the interpolant is piecewise constant.)

The finite element approximation consists more often of *piecewise linears* rather than of piecewise constants. More useful inequalities can be derived easily by noting that the essential point was that the function was known to vanish at one point in each element. If one considers linear interpolants, which coincide with the exact function at the nodal points, $\hat{e}'$ vanishes according to the so-called Rolle's theorem at least at one point inside the element. (This is obvious as it means that there is a tangent to $u$, which is parallel to the interpolating chord, see Figure 4.1 (b).) Thus (45) holds true here in addition to the form

$$\|\hat{e}\|\le h\|\tilde{e}'\| \tag{47}$$

also if $f$ has the role of the error derivative, i.e.,

$$\|\tilde{e}'\|\le\|\hat{e}''\|=h\left\|\phi''-\hat{\phi}''\right\|=h\|\phi''\| \tag{48}$$

(Note that $\hat{e}''=\phi''$ as the interpolant is linear.) Combining (47) and (48) gives

$$\|\hat{e}\|\le h^2\|\phi''\| \tag{49}$$

Formulas (49) and (48) correspond to (35) and (36) with $s=2$. Similar derivations can be employed for higher degree interpolants.

### 4.2.4 Error estimate

is generated by the inner product. (The energy norm is actually the square root of the quadratic part (multiplied by two) of the corresponding functional. See for instance expression (D.3.21) as a special case of (30).) Let us consider the square of the energy norm of the finite element error. We have

$$\begin{aligned}\|e\|_a^2 &= \|\phi-\bar{\phi}\|_a^2 = a(\phi-\bar{\phi},\phi-\bar{\phi}) = a(\phi-\tilde{\phi}+\tilde{\phi}-\bar{\phi},\phi-\bar{\phi}) \\ &= a(\phi-\tilde{\phi}+\tilde{\phi}-\bar{\phi},\phi-\bar{\phi}) = a(\phi-\tilde{\phi},\phi-\bar{\phi})+a(\tilde{\phi}-\bar{\phi},\phi-\bar{\phi}) \\ &= a(\phi-\tilde{\phi},\phi-\bar{\phi}) \le \left|a(\phi-\tilde{\phi},\phi-\bar{\phi})\right| \\ &\le a(\phi-\tilde{\phi},\phi-\tilde{\phi})^{1/2}\, a(\phi-\bar{\phi},\phi-\bar{\phi})^{1/2} \end{aligned} \tag{31}$$

On the first line an arbitrary function $\tilde{\phi} \in \bar{S}$ has been added and subtracted on the first argument of the bilinear form. The second line is achieved by taking into account the linearity of the bilinear form with respect of its arguments; here the first one. It is then realized that the difference function $\tilde{\phi}-\bar{\phi} \in \bar{V}$ and thus because of (28) or (29)

$$a(\tilde{\phi}-\bar{\phi},\phi-\tilde{\phi}) = a(\tilde{\phi}-\bar{\phi},e) = 0 \tag{32}$$

The fourth line is finally achieved by applying the Schwarz inequality (C.3.4). Dividing both sides of the inequality contained in (31) by the non-negative number $a(\phi-\tilde{\phi},\phi-\tilde{\phi})^{1/2}$, we obtain the important result

$$a(\phi-\tilde{\phi},\phi-\tilde{\phi})^{1/2} \le a(\phi-\bar{\phi},\phi-\bar{\phi})^{1/2} \tag{33}$$

or

$$\boxed{\|\phi-\tilde{\phi}\|_a \le \|\phi-\bar{\phi}\|_a} \tag{34}$$

It means that there is no member $\bar{\phi}$ of $\bar{S}$ that is a better approximation to $\phi$ in the sense of the energy norm measure than the Galerkin finite element solution $\tilde{\phi}$. This result is referred to as the *best approximation property* (paras-approksimaatio-ominaisuus). In pure diffusion problems the energy norm (30) is seen to consist of the first derivatives of $\phi$ (if the Robin boundary vanishes). Thus good values for the derivatives, or consequently say in heat conduction because of the Fourier law, to the components of the heat flux vector are to be expected. These quantities are in fact often more important in practice than the temperature itself.



### 4.2.3 Interpolation results

To proceed further in the estimations some results from classical interpolation theory are needed. Figure 4.7 shows examples of polynomial interpolation.

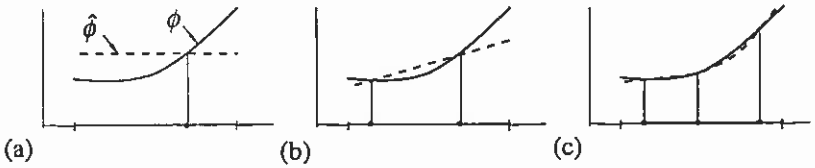



**Figure 4.7** **(a)** Constant interpolation. **(b)** Linear interpolation. **(c)** Quadratic interpolation.

The interpolating function or the so-called *interpolant* (interpolantti) to the given function $\phi$ is equipped here with the caret symbol: $\hat{\phi}$. By a constant, linear, quadratic, etc., polynomial interpolant we mean polynomials of the degree mentioned, which coincide in value with the given function $\phi$ (in one dimension) at least once, two times, three times, etc. in the interval under consideration. From Figure 4.7 it is intuitively obvious that if function $\phi$ behaves reasonably smoothly, its interpolant must be in some sense near the function itself and that the higher the degree of the interpolating polynomial the better the fit. Following basic estimates for the difference $\hat{e} = \phi - \hat{\phi}$ are valid, Johnson (1987, p. 84):

$$|\hat{e}|_0 \le Ch^s |\phi|_s \tag{35}$$

$$|\hat{e}|_1 \le Ch^{s-1} |\phi|_s \tag{36}$$

where the square of the seminorm (see Section C.4)

$$|u|_s^2 \equiv \sum_{i+j\cdots=s} \int_\Omega \left( \frac{\partial^s u}{\partial x^i \partial y^j \cdots} \right)^2 \mathrm{d}x\mathrm{d}y\cdots \tag{37}$$

and where $h$ is the mesh parameter and $C$ a constant independent of $h$. The highest complete degree of the interpolating polynomial in the above formulas is $s-1$.

To simplify the formulas to follow we shall henceforth denote the $L_2$ inner product and norm just by

We are now able to proceed from the best approximation inequality (34) or

$$\|e\|_a \le \|\hat{e}\|_a \tag{50}$$

In $\hat{e}=\phi-\hat{\phi}$, $\hat{\phi}\in\tilde{S}$ means now *the finite element interpolant* (elementti-interpolantti) to the exact solution, that is, it coincides with the exact solution at the nodal points. This concept has been discussed already in Remark 4.1. As the exact solution is unknown, so is the interpolant, but that does not prevent us from making use of the interpolation results derived above. The main thing is that also the interpolant $\hat{\phi}\in\tilde{S}$ as the theory demands. Let us consider (30) without the Robin boundary and for simplicity of presentation with an isotropic diffusivity and in the two-dimensional case:

$$\begin{aligned}\|\hat{e}\|_a^2 &= a(\hat{e},\hat{e}) = \int_\Omega \frac{\partial\hat{e}}{\partial x_\alpha} D\delta_{\alpha\beta} \frac{\partial\hat{e}}{\partial x_\beta}\mathrm{d}\Omega = \int_\Omega \frac{\partial\hat{e}}{\partial x_\alpha} D \frac{\partial\hat{e}}{\partial x_\alpha}\mathrm{d}\Omega \\ &= \int_\Omega D\left[\left(\frac{\partial\hat{e}}{\partial x}\right)^2+\left(\frac{\partial\hat{e}}{\partial y}\right)^2\right]\mathrm{d}\Omega \le D_{\max}\int_\Omega\left[\left(\frac{\partial\hat{e}}{\partial x}\right)^2+\left(\frac{\partial\hat{e}}{\partial y}\right)^2\right]\mathrm{d}\Omega \\ &\le D_{\max}|\hat{e}|_1^2 \le D_{\max}C^2h^{2(s-1)}|\phi|_s^2\end{aligned} \tag{51}$$

The steps used are rather obvious. Formula (36) has been finally made use of. Combining (51) with (50) and denoting $D_{\max}^{1/2}C$ again as $C$ gives the estimate

$$\|e\|_a \le Ch^{s-1}|\phi|_s \tag{52}$$

This estimate can be shown to remain valid even for the full expression (30) ($a$ positive) and even when the term $\int_\Omega \phi c\phi\mathrm{d}\Omega$ due to reaction part is included ($c$ positive) with the right-hand side in the form $C\sqrt{D_{\max}+ch^2+ah}\,h^{s-1}|\phi|_s$ where $D_{\max}$ is now the maximum eigenvalue of the diffusivity tensor.

For example, if the problem is solved with linear elements, estimate (52) gives

$$\|e\|_a \le Ch|\phi|_2 \tag{53}$$

Thus the exponent of the rate of the asymptotic convergence is 1. For quadratic and cubic elements the exponents are 2 and 3, respectively.

Let us consider still the energy inner product (22):



$$a(u,v) \equiv \int_\Omega \frac{\partial u}{\partial x_\alpha} D_{\alpha\beta} \frac{\partial v}{\partial x_\beta} d\Omega + \int_{\Gamma_R} uav\mathrm{d}\Gamma \tag{54}$$

We evaluate this for $u=v=\phi-\bar{\phi}$. Here function $\bar{\phi}$ is taken to be the finite element extension of the Dirichlet boundary data discussed in Remark 2.15. It may be taken to be non-zero only in the first element layer around the Dirichlet boundary. (It is assumed that the Dirichlet condition is then exactly satisfied everywhere on $\Gamma_D$ which may not be strictly true.) It is realized that the difference function $\phi-\bar{\phi}$ belongs to the linear space $V$. We obtain making use of the bilinearity and symmetry of the inner product

$$a(\phi-\bar{\phi},\phi-\bar{\phi}) = a(\phi,\phi)-2a(\phi,\bar{\phi})+a(\bar{\phi},\bar{\phi}) \tag{55}$$

The same quantity evaluated using the representation

$$\phi-\bar{\phi} = \tilde{\phi}+\phi-\tilde{\phi}-\bar{\phi} = \tilde{\phi}+e-\bar{\phi} = \tilde{\phi}-\bar{\phi}+e \tag{56}$$

gives similarly

$$\begin{aligned}a(\tilde{\phi}-\bar{\phi}+e,\tilde{\phi}-\bar{\phi}+e) &= a(\tilde{\phi}-\bar{\phi},\tilde{\phi}-\bar{\phi})-2a(\tilde{\phi}-\bar{\phi},e)+a(e,e) \\ &= a(\tilde{\phi}-\bar{\phi},\tilde{\phi}-\bar{\phi})+a(e,e) \\ &= a(\tilde{\phi},\tilde{\phi})-2a(\tilde{\phi},\bar{\phi})+a(\bar{\phi},\bar{\phi})+a(e,e)\end{aligned} \tag{57}$$

It is noticed that here $\tilde{\phi}-\bar{\phi}\in\tilde{V}$ and the middle term on the right-hand side on the first line disappears because of (28). Equating the right-hand sides of (55) and (57) and some manipulation gives the result

$$a(e,e) = a(\phi,\phi)-a(\tilde{\phi},\tilde{\phi})-2a(e,\bar{\phi}) \tag{58}$$

With homogeneous Dirichlet conditions, when the last term disappears (again due to (28) as then $\bar{\phi}\in V$), this is referred to in the literature as the *Pythagorean Theorem* and said in words: "*the energy of the error equals the error of the energy*". In the homogeneous case we have also (as $a(e,e)\ge 0$ because of the positive-definiteness of the inner product)

$$\boxed{a(\tilde{\phi},\tilde{\phi}) \le a(\phi,\phi)} \tag{59}$$

or in words: "*the approximate solution underestimates the energy*".

It can be shown that the left-hand side of (59) can be evaluated from

$$a(\tilde{\phi},\tilde{\phi})=\frac{1}{2}\{a\}^{\mathrm{T}}[K]\{a\}=\sum_{e}\{a\}^{e\mathrm{T}}[K]^{e}\{a\}^{e} \tag{60}$$

### 4.2.5 Pointwise error estimate

Here we need some results concerned with Green's functions. *Green's function* is the solution for a linear boundary value problem when the source term (forcing function, loading function) is the so-called *Dirac delta function* and the boundary conditions are homogeneous.

Let us consider as a simple example the one-dimensional pure diffusion problem with constant diffusivity, the Neumann condition $\phi'(0)=0$ and the Dirichlet condition $\phi(L)=0$. The source is first assumed to be concentrated around a given point $y$ on a length $\Delta x$ so that the intensity $f=F/\Delta x$ where $F$ is a constant. The solution is sketched in Figure 4.8 (a).

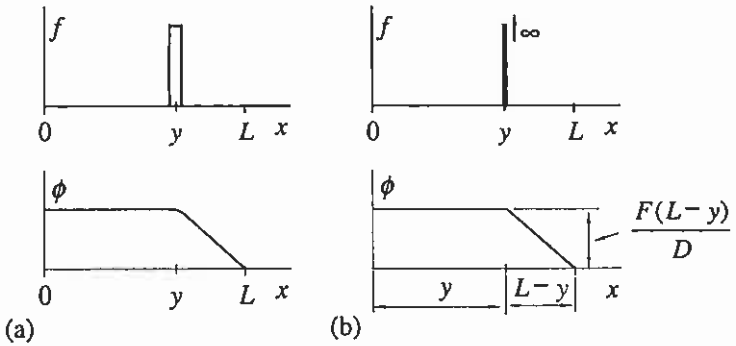



**Figure 4.8 (a)** Finite source intensity and the corresponding solution. **(b)** $F$ times the Dirac delta source and the corresponding solution.

When $F$ is kept fixed and $\Delta x\to 0$, $f\to\infty$ and the source distribution is denoted

$$f(x)=F\delta_y(x)\equiv F\delta(x-y) \tag{61}$$

where $\delta_y(x)\equiv\delta(x-y)$ is the Dirac delta function. Subscript $y$ refers to the point of action of the pointwise source. The Dirac delta is not a function in the classical sense but rather an operator defined by its action on continuous functions. Let $h(x)$ be continuous, then

$$\int_0^L h(x)\delta(x-y)\mathrm{d}x=h(y) \tag{62}$$



that is, the Dirac delta picks the value of $h$ at $y$. The total integrated source in the case of (61) is thus

$$\int_0^L f\,\mathrm{d}x=\int_0^L F\delta(x-y)\mathrm{d}x=F\int_0^L\delta(x-y)\mathrm{d}x=F\cdot 1=F \tag{63}$$

The solution of the problem with this source distribution is shown in Figure 4.8 (b). When $F=1$, it is the Green's function $G_y(x)$:

$$\begin{aligned} G_y(x)&=\frac{L-y}{k} && x\le y\\ G_y(x)&=\frac{L-y}{k}\frac{L-x}{L-y} && x\ge y \end{aligned} \tag{64}$$

The solution is obtained by solving the diffusion problem in a piecewise manner from the differential equation. But the solution must also be obtainable from the weak form or using the present notation: Find $G_y\in V$ (the boundary conditions are homogeneous so that $S=V$) such that

$$a(w,G_y)-(w,\delta_y)=0\quad\forall w\in V \tag{65}$$

But from the definition (62) $(w,\delta_y)=w(y)$ and

$$a(w,G_y)=w(y)\quad\forall w\in V \tag{66}$$

Thus *the inner product of w with the Green's function picks the value of w at y*. This relationship is seen to be similar to (62).

The finite element solution error $e$ belongs also to $V$ and (66) gives for it

$$a(G_y,e)=e(y) \tag{67}$$

From (28)

$$a(\tilde{w},e)=0\quad\forall\tilde{w}\in\tilde{V} \tag{68}$$

Subtraction of (68) from (67) gives finally

$$e(y)=a(G_y-\tilde{w},e)\quad\forall\tilde{w}\in\tilde{V} \tag{69}$$

This gives the possibility to estimate the pointwise value of the error appearing on the left-hand side. Taking absolute values of both sides, applying the

Schwarz inequality for the inner product and using the best approximation result (50) gives

$$|e(y)| = |a(G_y - \tilde{w}, e)| \le \|G_y - \tilde{w}\|_a \|e\|_a \le \|G_y - \tilde{w}\|_a \|\hat{e}\|_a \quad \forall \tilde{w} \in \tilde{V} \tag{70}$$

$\forall \tilde{w} \in \tilde{V}$. The last term can be bounded directly by using interpolation results. To achieve a good estimate we naturally try to select a $\tilde{w}$ which is as close as possible to $G_y$ to make the norm small. However, how well the Green's function can be followed by a $\tilde{w}$ belonging to $\tilde{V}$ depends strongly on the problem.

As an application let us consider the one-dimensional pure diffusion problem with constant diffusivity solved using linear elements. The interpolation estimate (51) is, when applied in one dimension and with slight changes in notation,

$$\|\hat{e}\|_a \le Ch\|\phi''\| \tag{71}$$

and (70)) reduces first into the form

$$|e(y)| \le \|G_y - \tilde{w}\|_a \, Ch\|\phi''\| \quad \forall \tilde{w} \in \tilde{V} \tag{72}$$

In order to proceed one needs at least some knowledge of the Green's function. Here it is seen from Figure 4.8 (b) — and this is true for any boundary conditions — that the function is piecewise linear and has a kink at $y$. There are two possibilities:

(a) The point $y$ is located at a node. Then the Green's function clearly belongs to the $\tilde{V}$ space and a $\tilde{w}$ can be selected to coincide with $G_y$ so that the corresponding norm disappears and the result is

$$e(y) = 0, \quad y \text{ coincides with a node} \tag{73}$$

Thus the exact nodal values obtained and shown in Figure 2.10 are not accidental but are in fact *valid for any source distribution.* (The same conclusion can be obtained for quadratic and cubic elements at the element endpoint nodes but not in general for the inner nodes.)

The exactness of the nodal values indicated by (73) is due to the fact that the Green's function was simple enough to belong to $\tilde{V}$. If the field equation contains a reaction term or if the diffusivity depends on position, the Green's function is more complicated than indicated in Figure 4.8 (b) and it does not belong any more in general to a polynomial weighting function space. The kink



still existing in the Green's function coincides with the node and the function is smooth everywhere else. Then one can show that the energy norm of the difference between the Green's function and its interpolant $\le Ch$ and the result is

$$|e(y)| \le Ch\|\phi''\|, \quad y \text{ coincides with a node} \tag{74}$$

(b) The point $y$ is not located at a node. The error estimate can be obtained for example by starting with the aid of the triangle inequality for absolute values. Namely

$$\begin{aligned} |e(y)| &= |\phi(y) - \tilde{\phi}(y)| = |\phi(y) - \hat{\phi}(y) + \hat{\phi}(y) - \tilde{\phi}(y)| \\ &\le |\phi(y) - \hat{\phi}(y)| + |\hat{\phi}(y) - \tilde{\phi}(y)| \end{aligned} \tag{75}$$

The first term on the right hand side is the absolute value of the interpolation error $\bar{e} = \phi - \bar{\phi}$. It can be estimated by using a finite Taylor expansion (See Remark 4.7 at the end of this derivation.):

$$|\phi(y) - \hat{\phi}(y)| \le h^2 \sup_{x \in \Omega} \|\phi''\| \tag{76}$$

The second term is the absolute value of the difference between the interpolant $\hat{\phi}$ and the finite element solution $\tilde{\phi}$. It is realized from Figure 4.9 that this term varies linearly (for linear elements) between the nodes and the value at $y$ is bounded from above by the value say at node $i$. Thus

$$\begin{aligned} |\hat{\phi}(y) - \tilde{\phi}(y)| &\le |\hat{\phi}(x_i) - \tilde{\phi}(x_i)| = |\phi(x_i) - \tilde{\phi}(x_i)| \le Ch^2\|\phi''\| \\ &\le Ch^2\sqrt{L} \sup_{x \in \Omega} |\phi''| \end{aligned} \tag{77}$$

Use have been made of the estimate (74) at a node and of the definition of the $L_2$-norm. Altogether (75), (76) and (77) give

$$|e(y)| \le \bar{C}h^2 \sup_{x \in \Omega} |\phi''| \tag{78}$$

which holds true for any fixed point in the domain. (Formula (74) can be developed further to have the sup-expression on the right-hand side if so wanted.) One should note, however, that the multiplier $\bar{C}$ may depend strongly on position and in order to verify (78) by numerical experiments one has to use similar meshes where the position of the point in the local elementwise coordinate system is fixed.

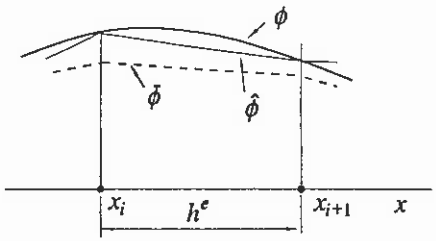



**Figure 4.9** The exact solution $\phi$, the finite element interpolant $\hat{\phi}$ and the finite element solution $\tilde{\phi}$.

**Remark 4.7.** The interpolation error $\hat{e}(y) = \phi(y) - \hat{\phi}(y)$ vanishes at points $x_i$ and $x_{i+1}$ (Figure 4.9). We assume $\phi$ to be at least a $C^2$ function in $\bar{\Omega}^e$. Then $\hat{e}$ is clearly also at least a $C^2$ function in $\bar{\Omega}^e$. Taylor's formula with remainder expanded at $x_i$ gives

$$\hat{e}(y) = \hat{e}'(x_i)(y - x_i) + \hat{e}''(\xi)(y - x_i)^2 / 2 \tag{79}$$

where $\xi \in [x_i, y]$. Similarly for $y = x_{i+1}$:

$$0 = \hat{e}'(x_i)(x_{i+1} - x_i) + \hat{e}''(\eta)(x_{i+1} - x_i)^2 / 2 = \hat{e}'(x_i)h^e + \hat{e}''(\eta)\left(h^e\right)^2 / 2 \tag{80}$$

where $\eta \in [x_i, x_{i+1}]$. Together (79) and (80) give

$$\hat{e}(y) = -\hat{e}''(\eta)h^e(y - x_i)/2 + \hat{e}''(\xi)(y - x_i)^2 / 2 \tag{81}$$

Taking absolute values on both sides of this, using the triangle inequality on the right hand side, making use of the facts $\hat{e}'' = \phi''$, $|y - x_i| \le h^e \le h$ etc., there is finally obtained

$$\left|\hat{e}(y)\right| \le h^2 \sup_{x \in \Omega^e} \left|\phi''\right| \le h^2 \sup_{x \in \Omega} \left|\phi''\right| \tag{82}$$

This is the result, which was used in the derivation above. □

Points, where the order of convergence is higher than in general, are referred to as *super-convergence points* (superkonvergenssipiste) in the literature. (In connection with the result (73), a node certainly deserves the name of a superconvergence point!) For say the derivatives of the unknown function, some other points (here the element centroids) may give superconvergence.

## REFERENCES

Crandall, S. (1956). *Engineering Analysis*, McGraw-Hill, New York.

Hughes, T. J. R. (1987). *The Finite Element Method — Linear Static and Dynamic Finite Element Analysis*, Prentice-Hall, Englewood Cliffs, New Jersey, ISBN 0-13-317017-9.



Irons, B. M. and Razzaque, A. (1972). Experience with Patch Test for Convergence of Finite Elements, (Aziz, A. K., ed.), *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, Academic Press.

Johnson, C. (1987). *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Studentlitteratur, Lund, ISBN 91-44-25241-1.

Ney, R. A. and Utku, S. (1972). An Alternative for the Finite Element Method, *Symp. Variational Methods*, Univ. of Southampton, Southampton University Press.

Taylor, R. L., Simo, J. C., Zienkiewicz, O. C. and Chan, C. H. (1986). The Patch Test — a Condition for Assessing FEM Convergence, *Int. J. num. Meth. Engng.* Vol. 22, pp. 39 - 62.

Zienkiewicz, O. C. (1971). *The Finite Element Method in Engineering Science*, McGraw-Hill, London.

Zienkiewicz, O. C. (1975). Why Finite Elements, Chapter 1 in ( Gallagher, R. H., Oden, J. T., Taylor, C. and Zienkiewicz, O. C., ed.), *Finite Elements in Fluids — Volume 1*, Wiley.

## PROBLEMS

# 5 SENSITIZED FORMULATION

## 5.1 INTRODUCTION

The sensitized formulation is an essential feature in the rest of this text and thus a general background chapter is considered here useful to introduce the basic ideas. These are approached through a structural mechanics problem employing a variational principle. Further, the assembly process with more than one nodal parameter per node is dealt with in this chapter.

### 5.1.1 Historical background

Historically, the finite element method was applied originally with very encouraging results to structural mechanics problems having available an associated variational principle. These problems are mathematically often kind of diffusion cases; roughly the second order derivatives in the differential equations are dominant. When more general types of problems with no associated variational principle available — especially fluid dynamics problems with dominant convection — were first attacked using the Galerkin finite element method, the results were not at all satisfactory: very dense meshes were needed to suppress the unphysical "wiggles" appearing in the discrete solutions. On the other hand, even in structural problems for flexible bodies using simple $C^0$ elements — thin beams, plates and shells — and still having an associated variational principle — impracticably dense meshes were again needed to achieve reasonable accuracy. Otherwise the displacements obtained were quite too small. This is called *locking* (lukkiutuminen). Many more or less useful tricks have been invented to try to circumvent these difficulties. It seems now that a rather simple to understand and theoretically sound procedure has finally emerged by which the standard Galerkin method can be modified to work well also in those cases where the standard version is performing poorly. This modification consists in effect of a combination of the Galerkin method and the least squares method. This methodology is called usually *stabilized formulation* (stabiloitu formulaatio) but we prefer to call it *sensitized formulation* (sensitoitu formulaatio). This latter terminology stems from the remarkable article by Courant (1943) which is in addition considered in the mathematics literature as the birth paper of the finite element method. It seems that researchers on stabilized formulations have not been aware the ideas suggested much earlier by Courant and we therefore want to start the description using these earlier presentations as the starting point. We quote from Courant (1943):

> "These facts which are intimately related to more profound questions in the general theory of the variational calculus have suggested the following method for obtaining better convergence in the Rayleigh-Ritz method. Instead of considering the simple variational problem for





> the corresponding boundary value problem, we modify the former problem without changing the solution of the latter. This is accomplished by adding to the original variational expression terms of higher order which vanish for the actual solution $u$. For example, we may formulate the equilibrium problem for a membrane under the external pressure $f$ as follows:
>
> $$I(v) = \iint_B (v_x^2 + v_y^2 + vf)dxdy + \iint_B k(\Delta v - f)^2 dxdy = \min.,$$
>
> where $k$ is an arbitrary positive constant or function. Such additional terms make $I(v)$ more *sensitive* to the variations of $v$ without changing the solution. In other words, minimizing sequences attached to such a "sensitized" functional will by force behave better as regards convergence [7].
>
> The practical value of the method of sensitizing the integral by the addition of terms of higher order has not yet been sufficiently explored. Certainly the sensitizing terms will lead to a more complicated system of equations for the $c_i$. This means that a compromise must be made for a suitable choice of the arbitrary positive function $k$ so that good convergence is assured while the necessary labor is kept within bounds."

**Remark 5.1.** The term *Rayleigh-Ritz method* is used in the literature usually simply in the same meaning as the term *Ritz-method* (cf. Remark 1.2). This terminology normally also implies that the discretization is performed via a variational formulation as is here clearly the case in the quotation above. □

**Remark 5.2.** In the formula of the quotation above, there is obviously a slight misprint and there should probably read

$$I(v) = \iint_B (v_x^2 + v_y^2 + 2vf)dxdy + \iint_B k(\Delta v - f)^2 dxdy = \min., \qquad (1)$$ □

Reference [7], in Courant's article of year (1943), is concerned with the possibility to append the variational integral in addition to the integral of the field equation residual squared considered above with an arbitrary number of similar terms consisting of integrals of derivatives of the field equation residual squared. We quote from Courant (1923):

> "Zur Erläuterung behandeln wir die Randwertaufgabe der Potentialtheorie für einen Bereich $G$ in der $xy$-Ebene. Die vorgegebenen Randwerte mögen identisch sein mit den Werten, die ein Polynom $p(x,y)$ auf dem Rand annimmt. Der Rand von $G$ möge abgesehen von endlich vielen Ecken eine sich stetig drehende Tangente besitzen. Wir betrachen das Integral

$$D[\varphi]=\int_G\int\{\varphi_x^2+\varphi_y^2+(\Delta\varphi)^2+(\Delta\varphi_x)^2+(\Delta\varphi_y)^2+(\Delta\varphi_{xx})^2+\cdots\}dxdy,$$

wobei rechts über alle in Frage kommenden Ableitungen zu summieren ist, und fordern, das Integral $D[\varphi]$ zum Minimum zu machen, wenn zum Vergleiche alle in $G$ mit ihren Ableitungen stetigen Funktionen $\varphi$ zugelassen werden, welche die vorgeschriebenen Randwerte besitzen."

**Remark 5.3.** From formula (1) of the latter quotation, it is apparent that the presentation there is assumed to be in a dimensionless form as otherwise the total expression would not be dimensionally homogeneous. □

The example cases in the quotations concern the solution of the Poisson and the Laplace differential equations $\Delta v-f=0$ and $\Delta\varphi=0$, respectively. The $\Delta$-notation refers to the Laplace differential operator $\Delta=\partial^2/\partial x^2+\partial^2/\partial y^2$. Further, $()_x=\partial()/\partial x$ and $()_y=\partial()/\partial y$. The conventional functional expressions in the first parts of the formulas are easily discerned from the least squares type appended terms.

**Remark 5.4.** It is realized as commented on in Courant (1923) that the possibility to append the sensitizing terms is not limited to linear problems. □

### 5.1.2 Timoshenko beam

Although we later concentrate on the use of weak forms and applications in heat transfer, we start here by presenting the important ideas of Courant employing a variational principle and a structural mechanics problem. This way to proceed hopefully proves to give finally an easy to understand explanation for the maybe rather mystical extra terms appearing in some presentations in the literature. As a side product we find how sensitizing can dramatically improve the discrete solution behavior also in structural mechanics problems. The membrane case considered by Courant is not very suitable as a demonstration example here as the standard finite element version is known to work well with it. The Timoshenko beam problem described in the following proves to be a more illustrative model for our purposes. For readers not familiar with variational calculus and functionals, certain relevant concepts are given in Appendix D. In what follows we borrow much from Freund and Salonen (2000).

Figure 5.1 describes some notations for a so-called Timoshenko beam. The beam axis is straight and the beam displacement is assumed to take place in the $xy$-plane. The maybe somewhat odd selection of the coordinate direction — $y$-axis downwards — is in rather general use for historical reasons in beam bending. Here we are using in principle contrary to the rest of the text the Lagrangian description for the kinematics common to solids (cf. Section 6.1.1).



However, as we restrict the study to the small displacement case, the difference in the descriptions does not come up (the domain after the deformation may be considered to coincide with the initial domain). The intensity of the given distributed transverse loading is $q(x)$ ($[q]=$ N/m).

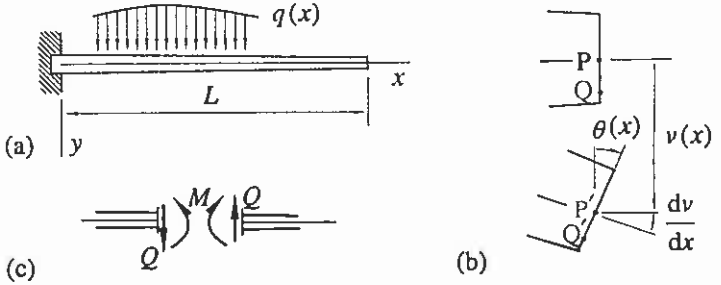



**Figure 5.1** **(a)** Beam under distributed loading. **(b)** Transverse displacement $v$ and cross section rotation $\theta$ greatly exaggerated. **(c)** Shearing force $Q$ and bending moment $M$ at a cross section.

In the Timoshenko beam theory the following kinematical assumption is made: beam cross-section material planes perpendicular to the beam axis before the deformation move as rigid planes during the deformation. This means that for small displacements the displacement components $u$ and $v$ in the $x$- and $y$-axis directions, respectively, for a generic material point Q are obtained by the formulas (Figure 5.1(b))

$$\begin{aligned}u(x,y)&=-y\theta(x)\\ v(x,y)&=v(x)\end{aligned}\qquad(2)$$

where $v(x)$ is the beam axis transverse displacement or *deflection* (taipuma) (displacement of point P) and $\theta(x)$ is the beam cross section rotation. The continuum problem is thus reduced to the determination of two functions. This is the first occurrence in this text where *two unknown functions* appear simultaneously in a problem.

From the free-body diagram for a differential beam element the following equilibrium equations

$$\begin{aligned}&\frac{dQ}{dx}+q=0\\ &Q-\frac{dM}{dx}=0\end{aligned}\qquad\text{in } \Omega=]0,L[\qquad(3)$$

are obtained where $Q$ is the *shearing force* (leikkausvoima) and $M$ the *bending moment* (taivutusmomentti) (Figure 5.1(c)). These are the important quantities in the design of the beam with respect to stresses. The beam can also be loaded by a distributed couple loading which would add a term in (3b). This refinement is usually needed only in dynamic problems, which are not considered here.

When the beam material is assumed to be elastic, the overall material properties are found to be given by the *shearing stiffness* (leikkausjäykkyys) $GA(x)$ ($[GA]=\mathrm{N}$) and the *bending stiffness* (taivutusjäykkyys) $EI(x)$ ($[EI]=\mathrm{Nm}^2$) so that

$$\begin{aligned} Q &= kGA\left(\frac{\mathrm{d}v}{\mathrm{d}x}-\theta\right) \\ M &= -EI\frac{\mathrm{d}\theta}{\mathrm{d}x} \end{aligned} \tag{4}$$

$GA$ and $EI$ are to understood to be in the general case just double letter symbols. For a homogeneous material in the $y$-direction with a constant *Young's modulus* $E$ (kimmokerroin) and a constant *shear modulus* $G$ (liukukerroin), $GA$ and $EI$ can be interpreted so that $A$ is the cross section area and $I$ is the cross-sectional *second moment* (pintaneliömomentti, pintahitausmomentti). For a rectangular cross section with breadth $b$ and height $t$ and isotropic homogeneous material

$$\begin{aligned} GA &= \frac{E}{2(1+\nu)}bt \\ EI &= \frac{Ebt^3}{12} \end{aligned} \tag{5}$$

where $\nu$ is the so-called Poisson's ratio; a material constant whose value for most structural materials is between 0 and $1/2$. The dimensionless multiplier $k$ in (4) is called the *shear correction factor* (liukumakorjauskerroin). The need for this multiplier is explained as follows. The kinematical assumptions (2) mean that the *shearing strain* (liukuma)

$$\gamma \equiv \frac{\partial v(x,y)}{\partial x}+\frac{\partial u(x,y)}{\partial y}=\frac{\mathrm{d}v(x)}{\mathrm{d}x}-\theta(x) \tag{6}$$

is constant on a cross section. In reality the cross section does not remain exactly plane but warps somewhat and $\theta$ must be considered to represent some kind of average rotation of the cross section. Similarly, the shearing strain is not constant on a cross section and this can be taken approximately into account by



the shear correction factor. Literature contains procedures to determine $k$. For instance the value $k=5/6$ is used for a rectangular cross section for a homogeneous material.

**Remark 5.5.** Comparison of (4a) and (6) shows that the shearing force is obtained as the product of the modified shearing stiffness and the shearing strain. When the beam gets thinner or specifically, when the ratio $t/L$ gets smaller and smaller, based on expressions (5), the relative value of the shearing stiffness compared to the bending stiffness gets larger and larger. From physical reasons the shearing force remains finite in a problem even when the beam gets thin. The shearing strain approaches then zero, as the shearing force must then be the product of a large and a small quantity. Thus we get from (6) in the limit

$$\frac{\mathrm{d}v}{\mathrm{d}x}-\theta=0 \tag{7}$$

or

$$\theta=\frac{\mathrm{d}v}{\mathrm{d}x} \tag{8}$$

In the so-called *Bernoulli beam theory*, expression (8) is used from the outset as one of the assumptions to form the beam model. Looking at Figure 5(b), this means that the cross section is assumed to remain perpendicular to the deformed beam axis after the deformation. Due to relation (8), in the Bernoulli beam model only one function, the beam axis deflection $v(x)$, remains to be determined. However, the corresponding differential equation is of the fourth order and the model is not realistic for thick beams. Thus the finite element applications have lately concentrated on the Timoshenko model (and to its analogues in plate and shell problems). This affords simple $C^0$ continuous approximations but the locking phenomenon mentioned in Section 5.1.1 for thin cases has been a problem. □

Substitution of the constitutive relations (4) into (3) gives the equilibrium equations expressed in displacement quantities:

$$\begin{aligned} R_v(v,\theta) &\equiv L_v(v,\theta)+q \equiv \frac{\mathrm{d}}{\mathrm{d}x}kGA\left(\frac{\mathrm{d}v}{\mathrm{d}x}-\theta\right)+q=0 \\ R_\theta(v,\theta) &\equiv L_\theta(v,\theta) \equiv kGA\left(\frac{\mathrm{d}v}{\mathrm{d}x}-\theta\right)-\frac{\mathrm{d}}{\mathrm{d}x}\left(-EI\frac{\mathrm{d}\theta}{\mathrm{d}x}\right)=0 \end{aligned} \quad \text{in } \Omega=]0,L[ \tag{9}$$

We have introduced some new notation for later use. The subscripts $v$ and $\theta$ are employed just to discern between the two field equation residuals. (Based on the equilibrium equations one may assume that the first equations is associated with the transverse displacement direction and the second with the cross section rotation direction which somewhat explains the notation.) The terms denoted by the symbol $L$ are defined implicitly by the middle forms. The unknown functions are included as argument symbols to make the expressions more transparent. We could define alternatively a linear matrix operator

$$[L]=\begin{bmatrix} \frac{\mathrm{d}}{\mathrm{d}x}kGA\frac{\mathrm{d}}{\mathrm{d}x} & -\frac{\mathrm{d}}{\mathrm{d}x}kGA \\ kGA\frac{\mathrm{d}}{\mathrm{d}x} & -kGA+\frac{\mathrm{d}}{\mathrm{d}x}EI\frac{\mathrm{d}}{\mathrm{d}x} \end{bmatrix} \tag{10}$$

and write equations (9) as

$$\begin{Bmatrix} R_v(v,\theta) \\ R_\theta(v,\theta) \end{Bmatrix} \equiv [L]\begin{Bmatrix} v \\ \theta \end{Bmatrix} + \begin{Bmatrix} q \\ 0 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix} \tag{11}$$

The boundary conditions for the example case shown in Figure 5.1(a) are

$$\begin{matrix} v=0 \\ \theta=0 \end{matrix} \quad \text{at } \Gamma_{\mathrm{D}}=\{0\} \tag{12}$$

and

$$\begin{aligned} &Q=0, \ \Rightarrow \ kGA\left(\frac{\mathrm{d}v}{\mathrm{d}x}-\theta\right)\Big|_{x=L}=0 \\ &M=0, \ \Rightarrow -EI\frac{\mathrm{d}\theta}{\mathrm{d}x}\Big|_{x=L}=0 \end{aligned} \qquad \text{at } \Gamma_{\mathrm{N}}=\{L\} \tag{13}$$

In the structural mechanics terminology the structure here is called a *cantilever beam* (ulokepalkki) with the left-hand end $x=0$ of the beam *clamped* (jäykästi kiinnitetty) and the right-hand end $x=L$ *free* (vapaa). Boundary conditions (12) are clearly of the Dirichlet type and latter conditions (13) can be in some sense be considered to be of the Neumann type.

Equations (9), (12), (13) describe the structural problem in a strong form. We, however, consider here the corresponding variational formulation. The most important variational principle in elasticity is the *principle of stationary potential energy* (potentiaalienergian stationaarisuuden periaate):

> When a conservative system is in equilibrium, the potential energy of the system has a stationary value. (14)

An elastic body forms a conservative system if the external forces acting on the body are conservative. We refer to Washizu (1982), for an especially valuable source on variational principles in solid mechanics.

The potential energy functional of the Timoshenko beam is, e.g., Dym and Shames (1973),



$$\Pi(v,\theta)=\int_\Omega\left[\frac{1}{2}kGA\left(\frac{\mathrm{d}v}{\mathrm{d}x}-\theta\right)^2+\frac{1}{2}EI\left(\frac{\mathrm{d}\theta}{\mathrm{d}x}\right)^2-qv\right]\mathrm{d}\Omega+\mathrm{bt} \tag{15}$$

The admissible argument functions $v(x)$ and $\theta(x)$ in (15) must satisfy the essential Dirichlet type boundary conditions; here conditions (12). The shorthand notation bt includes the possible additional terms arising from the natural Neumann type boundary conditions concerning given shearing force $Q$ or bending moment $M$. Because here the given $Q$ and $M$ are zero in (13), the term bt vanishes. Proceeding similarly as in Section D.3.1, we find that the differential equations (9) and the boundary conditions (13) indeed follow from the condition $\delta\Pi=0$.

### 5.1.3 Preliminary considerations

We shall first generate a conventional finite element solution for comparison with the later sensitized formulation. Before actual calculations we perform in this section some preliminary manipulations. We have had some practice how to discretize a functional in Section 2.1.1 where the least squares expression was considered. The approach was first to substitute the approximation. The system equations were then obtained by differentiations of the discretized functional with respect to the nodal parameters and putting the results equal to zero. Samples of this way to proceed are given in Examples 2.3 and D.1. For later purposes we proceed here by a slightly alternative route: we first perform analytically the variation to obtain the equation $\delta\Pi=0$ and only then substitute the approximation. The stationarity condition $\delta\Pi=0$ can then be interpreted as a weak form with the variations of the argument functions having the roles of the weighting functions (see Remark D.2). This route is convenient for example if we have at our use a computer program which is intended to be used only with weak forms as is the case with MATHFEM. The final resulting system equations become the same using either of the two routes.

Applying the calculation rules of Table D.1 on (15) gives the variation

$$\delta\Pi=\int_\Omega\left[kGA\left(\frac{\mathrm{d}v}{\mathrm{d}x}-\theta\right)\left(\frac{\mathrm{d}\delta v}{\mathrm{d}x}-\delta\theta\right)+EI\frac{\mathrm{d}\theta}{\mathrm{d}x}\frac{\mathrm{d}\delta\theta}{\mathrm{d}x}-q\delta v\right]\mathrm{d}\Omega \tag{16}$$

We use the notations

$$\delta v=w_v, \quad \delta\theta=w_\theta \tag{17}$$

where $w_v(x)$ and $w_\theta(x)$ are now called weighting functions and write the stationarity condition $\delta\Pi=0$ as a weak form

$$F \equiv \int_\Omega \left[ kGA\left(\frac{dv}{dx}-\theta\right)\left(\frac{dw_v}{dx}-w_\theta\right)+EI\frac{d\theta}{dx}\frac{dw_\theta}{dx}-qw_v \right] d\Omega = 0 \tag{18}$$

Because conditions (12) are essential, the variations $\delta v$ and $\delta\theta$ must vanish there (see Section D.3.1). This is also the case even when the right-hand sides of (12) would be non-zero. Thus in the weak form (18)

$$\begin{aligned} w_v &= 0 \\ w_\theta &= 0 \end{aligned} \quad \text{at } \Gamma_D = \{0\} \tag{19}$$

and the admissible $v$ and $\theta$ must satisfy (12).

As this is the first place where we deal with two unknown functions, before proceeding to approximations, we shall derive the weak form (18) alternatively directly from the governing differential equations (9). The procedure is a rather obvious generalization from what has been done earlier with one unknown function the aim being one scalar equation. The manipulations become simpler when we start from (3). The first equation is multiplied by an arbitrary weighting function $w_v(x)$, the second by $w_\theta(x)$ and the resulting equations are integrated over the domain $\Omega$ and added together to produce a preliminary weak form

$$\int_\Omega \left[ w_v\left(\frac{dQ}{dx}+q\right)+w_\theta\left(Q-\frac{dM}{dx}\right)\right] d\Omega = 0 \tag{20}$$

To lower the final order of derivatives appearing, we integrate the terms containing $dQ/dx$ and $dM/dx$:

$$\int_\Omega \left( -\frac{dw_v}{dx}Q+w_v q+w_\theta Q+\frac{dw_\theta}{dx}M \right) d\Omega + \Big|_0^L w_v\, Q - \Big|_0^L w_\theta\, M = 0 \tag{21}$$

Taking the restriction (19) into account and making use of the (here homogeneous) boundary conditions (13) shows that the boundary terms disappear and the weak form is now

$$\int_\Omega \left[ \left( w_\theta - \frac{dw_v}{dx}\right)Q+\frac{dw_\theta}{dx}M+w_v q \right] d\Omega = 0 \tag{22}$$

This is actually an application of the principle of virtual work discussed in Remark 3.1. It is seen that no material data is yet included and this form can thus be employed for example in connection with a plastic body where the principle of stationary potential energy would no more be valid. Substitution of the elastic constitutive relations (4) gives the final form here:



$$\int_\Omega \left[ \left( -\frac{dw_v}{dx}+w_\theta\right) kGA\left(\frac{dv}{dx}-\theta\right)-\frac{dw_\theta}{dx}EI\frac{d\theta}{dx}+w_v q \right] d\Omega = 0 \tag{23}$$

This is seen to be equivalent to (18). (If wanted, we can always change the signs in (23) to make it (18) by taking new arbitrary weighting functions of opposite signs to those used here.)

**Remark 5.6.** The above type of manipulations to produce a weak form generalize in an obvious way to any number of unknown functions and differential equations. The resulting weak form is always just *one scalar equation* and we have as many arbitrary weighting functions as there are independent scalar differential equations. □

### 5.1.4 Standard Galerkin finite element solution

We take as the basis of the discretization expression (18) written in a more conventional form so that the weighting functions are put first in the products in the integrand:

$$\boxed{F \equiv \int_\Omega \left[ \left( \frac{dw_v}{dx}-w_\theta\right) kGA\left(\frac{dv}{dx}-\theta\right)+\frac{dw_\theta}{dx}EI\frac{d\theta}{dx}-w_v q \right] d\Omega = 0} \tag{24}$$

We employ the simplest possible discretization:

$$\begin{aligned} \tilde{v}(x) &= \sum_j N_j(x) v_j \\ \tilde{\theta}(x) &= \sum_j N_j(x) \theta_j \end{aligned} \tag{25}$$

using two-noded linear elements. The nodal parameters are thus the deflection and the cross section rotation at a node. The system equations using the Galerkin method are obtained with a similar logic as earlier: the approximations (25) are substituted in (24) and the specific two discrete equations corresponding to node $i$ are found by taking first $w_v = N_i$, $w_\theta = 0$ and then $w_v = 0$, $w_\theta = N_i$ to obtain

$$\begin{aligned} &\int_\Omega \left[ \frac{dN_i}{dx} kGA\left(\frac{d\tilde{v}}{dx}-\tilde{\theta}\right)-N_i q \right] d\Omega = 0 \\ &\int_\Omega \left[ -N_i kGA\left(\frac{d\tilde{v}}{dx}-\tilde{\theta}\right)+\frac{dN_i}{dx}EI\frac{d\tilde{\theta}}{dx} \right] d\Omega = 0 \end{aligned} \tag{26}$$

We proceed to evaluate the element contributions and to perform the assembly using the following simple demonstration example case.

**Example 5.1.** A uniform property cantilever beam under constant distributed loading $q$ is analyzed (Figure (a)). The beam cross section is rectangular and the material is isotropic with Poisson's ratio $\nu = 1/3$ and we take $k = 5/6$.

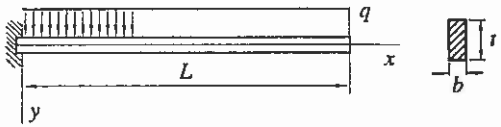



**Figure (a)**

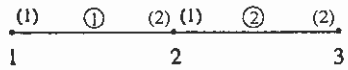



**Figure (b)**

A very crude uniform two element mesh is used (Figure (b)). A natural listing order of global nodal parameters is here

$$\underset{6\times1}{\{a\}} = [v_1\ \theta_1\ v_2\ \theta_2\ v_3\ \theta_3]^T \tag{a}$$

Due to the clamped end boundary conditions, we shall finally put

$$v_1 = 0, \quad \theta_1 = 0 \tag{b}$$

Notations and approximations for a generic element are shown in Figure (c). The listing order of the local nodal parameters is taken to be (we leave the element superscript $e$ for simplicity of notation mostly from the expressions to follow)

$$\underset{4\times1}{\{a\}} = [v_1\ \theta_1\ v_2\ \theta_2]^T \tag{c}$$

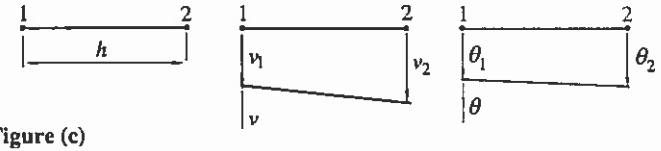



**Figure (c)**

The element approximations are

$$\begin{aligned} \bar{v} &= \sum_j N_j v_j = N_1 v_1 + N_2 v_2 = v_1 + \xi v_2 \\ \bar{\theta} &= \sum_j N_j \theta_j = N_1 \theta_1 + N_2 \theta_2 = (1-\xi)\theta_1 + \xi \theta_2 \end{aligned} \tag{d}$$

Making use of formulas (26) on the element level gives the element contributions

$$F_1 = \int_{\Omega^e} \left[ \frac{dN_1}{dx} kGA \left( \frac{d\bar{v}}{dx} - \bar{\theta} \right) - N_1 q \right] d\Omega$$



$$\begin{aligned} F_2 &= \int_{\Omega^e} \left[ -N_1 kGA \left( \frac{d\bar{v}}{dx} - \bar{\theta} \right) + \frac{dN_1}{dx} EI \frac{d\bar{\theta}}{dx} \right] d\Omega \\ F_3 &= \int_{\Omega^e} \left[ \frac{dN_2}{dx} kGA \left( \frac{d\bar{v}}{dx} - \bar{\theta} \right) - N_2 q \right] d\Omega \\ F_4 &= \int_{\Omega^e} \left[ -N_2 kGA \left( \frac{d\bar{v}}{dx} - \bar{\theta} \right) + \frac{dN_2}{dx} EI \frac{d\bar{\theta}}{dx} \right] d\Omega \end{aligned} \tag{e}$$

We evaluate in detail the first contribution (e):

$$\begin{aligned} F_1 &= \int_{\Omega^e} \left[ \frac{dN_1}{dx} kGA \left( \sum_j \frac{dN_j}{dx} v_j - \sum_j N_j \theta_j \right) - N_1 q \right] d\Omega \\ &= \sum_j \left( \int_{\Omega^e} \frac{dN_1}{dx} kGA \frac{dN_j}{dx} d\Omega \right) v_j - \sum_j \left( \int_{\Omega^e} \frac{dN_1}{dx} kGA N_j d\Omega \right) \theta_j - \int_{\Omega^e} N_1 q\, d\Omega \\ &= kGA \int_{\Omega^e} \frac{dN_1}{dx} \frac{dN_1}{dx} d\Omega\, v_1 + kGA \int_{\Omega^e} \frac{dN_1}{dx} \frac{dN_2}{dx} d\Omega\, v_2 \\ &\quad - kGA \int_{\Omega^e} \frac{dN_1}{dx} N_1 d\Omega\, \theta_1 - kGA \int_{\Omega^e} \frac{dN_1}{dx} N_2 d\Omega\, \theta_2 - q \int_{\Omega^e} N_1 d\Omega \end{aligned} \tag{f}$$

In the last step constant $kGA$ and $q$ have been assumed. The other steps should be obvious for example following the manipulations explained in Section 2.3.1. Evaluating the integrals making again use of the formulas in Section F.1.1 gives

$$\begin{aligned} F_1 &= kGA \left( \frac{1}{h} v_1 - \frac{1}{h} v_2 \right) - kGA \left( -\frac{1}{2}\theta_1 - \frac{1}{2}\theta_2 \right) - q\frac{h}{2} \\ &= \frac{kGA}{h} v_1 + \frac{kGA}{2}\theta_1 - \frac{kGA}{h} v_2 + \frac{kGA}{2}\theta_2 - \frac{qh}{2} \end{aligned} \tag{g}$$

Performing the rest of the calculations we arrive at the element contributions

$$\underset{4\times1}{\{F\}^e} = \underset{4\times4}{[K]^e}\ \underset{4\times1}{\{a\}^e} - \underset{4\times1}{\{b\}^e} \tag{h}$$

with

$$\underset{4\times4}{[K]^e} = \begin{bmatrix} \overset{1}{\frac{kGA}{h}} & \overset{2}{\frac{kGA}{2}} & \overset{3}{-\frac{kGA}{h}} & \overset{4}{\frac{kGA}{2}} \\ \frac{kGA}{2} & \frac{kGAh}{3} + \frac{EI}{h} & -\frac{kGA}{2} & \frac{kGAh}{6} - \frac{EI}{h} \\ -\frac{kGA}{h} & -\frac{kGA}{2} & \frac{kGA}{h} & -\frac{kGA}{2} \\ \frac{kGA}{2} & \frac{kGAh}{6} - \frac{EI}{h} & -\frac{kGA}{2} & \frac{kGAh}{3} + \frac{EI}{h} \end{bmatrix} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} \tag{i}$$

$$\underset{4\times1}{\{b\}^e} = qh\begin{Bmatrix}1/2\\0\\1/2\\0\end{Bmatrix}\begin{matrix}1\\2\\3\\4\end{matrix}$$

The dimensionless ratio

$$\varepsilon = \frac{EI}{kGAL^2} \tag{j}$$

is a measure of the relation between the overall bending stiffness and the shearing stiffness. When the beam gets thinner this ratio gets smaller. Similarly, the dimensionless ratio

$$\varepsilon_h = \frac{EI}{kGAh^2} \tag{k}$$

is a convenient shorthand notation to be used on the element level. With (k), the stiffness matrix (i) can be written as

$$[K] = \frac{kGA}{h}\begin{bmatrix} 1 & h/2 & -1 & h/2 \\ h/2 & h^2(1/3+\varepsilon_h) & -h/2 & h^2(1/6-\varepsilon_h) \\ -1 & -h/2 & 1 & -h/2 \\ h/2 & h^2(1/6-\varepsilon_h) & -h/2 & h^2(1/3+\varepsilon_h) \end{bmatrix}\begin{matrix}1\\2\\3\\4\end{matrix} \qquad \begin{matrix}1&2&3&4\end{matrix} \tag{l}$$

From Figure (b) the data concerning the local and global nodal parameter numbering is given in the following table (see Remark 3.9). The symbols with superscript "star" refer here to the local nodal parameter numbers to discern them from the local node numbering symbols.

| | (1) | | (2) | |
|---|---|---|---|---|
| | $(1)^*$ | $(2)^*$ | $(3)^*$ | $(4)^*$ |
| ① | 1 | 2 | 3 | 4 |
| ② | 3 | 4 | 5 | 6 |

(m)

The assembly rules (2.3.39) — taking Remark 3.9 into account — give

$$\underset{6\times6}{[K]} = \begin{bmatrix} K_{11}^1 & K_{12}^1 & K_{13}^1 & K_{14}^1 & 0 & 0 \\ K_{21}^1 & K_{22}^1 & K_{23}^1 & K_{24}^1 & 0 & 0 \\ K_{31}^1 & K_{32}^1 & K_{33}^1+K_{11}^2 & K_{34}^1+K_{12}^2 & K_{13}^2 & K_{14}^2 \\ K_{41}^1 & K_{42}^1 & K_{43}^1+K_{21}^2 & K_{44}^1+K_{22}^2 & K_{23}^2 & K_{24}^2 \\ 0 & 0 & K_{31}^2 & K_{32}^2 & K_{33}^2 & K_{34}^2 \\ 0 & 0 & K_{41}^2 & K_{42}^2 & K_{43}^2 & K_{44}^2 \end{bmatrix}\begin{matrix}1\\2\\3\\4\\5\\6\end{matrix} \qquad \begin{matrix}1&2&3&4&5&6\end{matrix}$$



(n)

$$\underset{6\times1}{\{b\}} = \begin{Bmatrix} b_1^1 \\ b_2^1 \\ b_3^1+b_1^2 \\ b_4^1+b_2^2 \\ b_3^2 \\ b_4^2 \end{Bmatrix}\begin{matrix}1\\2\\3\\4\\5\\6\end{matrix}$$

The first two nodal variables are fixed by (b). The remaining active equations are thus

$$\begin{aligned} K_{33}v_2 + K_{34}\theta_2 + K_{35}v_3 + K_{36}\theta_3 &= b_3 \\ K_{43}v_2 + K_{44}\theta_2 + K_{45}v_3 + K_{46}\theta_3 &= b_4 \\ K_{53}v_2 + K_{54}\theta_2 + K_{55}v_3 + K_{56}\theta_3 &= b_5 \\ K_{63}v_2 + K_{64}\theta_2 + K_{65}v_3 + K_{66}\theta_3 &= b_6 \end{aligned} \tag{o}$$

Collecting the terms from (l) according to (n) gives in detail the set

$$\begin{aligned} &\frac{kGA}{h}(2\cdot v_2 + 0\cdot h\theta_2 - 1\cdot v_3 + 1/2\cdot h\theta_3) = qh \\ &\frac{kGA}{h}h\left[0\cdot v_2 + (2/3+2\varepsilon_h)h\theta_2 - 1/2\cdot v_3 + (1/6-\varepsilon_h)h\theta_3\right] = 0 \\ &\frac{kGA}{h}(1\cdot v_2 - 1/2\cdot h\theta_2 + 1\cdot v_3 - 1/2\cdot h\theta_3) = \frac{qh}{2} \\ &\frac{kGA}{h}h\left[1/2\cdot v_2 + (1/6-\varepsilon_h)h\theta_2 - 1/2\cdot v_3 + (1/3+\varepsilon_h)h\theta_3\right] = 0 \end{aligned} \tag{p}$$

Here, for a rectangular cross section,

$$\begin{aligned} kGA &= \frac{5}{6}\frac{E}{2(1+1/3)}bt = \frac{5}{16}Ebt \\ EI &= E\frac{bt^3}{12} = \frac{1}{12}Ebt^3 \end{aligned} \tag{q}$$

We continue with the case $t = L/10$ and obtain ($h = L/2$)

$$\varepsilon_h = \frac{Ebt^3\cdot 16}{12\cdot 5\cdot Ebth^2} = \frac{4t^2}{15(L/2)^2} = \frac{4t^2}{15(5t)^2} = \frac{4}{375} \tag{r}$$

With this data set (p) is found to be

$$\begin{aligned} &\frac{kGA}{h}\left(2\cdot v_2 + 0\cdot h\theta_2 - 1\cdot v_3 + \frac{1}{2}h\theta_3\right) = qh \\ &\frac{kGA}{h}h\left(0\cdot v_2 + \frac{258}{375}h\theta_2 - \frac{1}{2}v_3 + \frac{117}{750}h\theta_3\right) = 0 \end{aligned}$$

$$\frac{kGA}{h}\left(1\cdot v_2-\frac{1}{2}h\theta_2+1\cdot v_3-\frac{1}{2}h\theta_3\right)=\frac{qh}{2} \qquad (s)$$

$$\frac{kGA}{h}h\left(\frac{1}{2}v_2+\frac{117}{750}h\theta_2-\frac{1}{2}v_3+\frac{129}{375}h\theta_3\right)=0$$

The solution is

$$v_2=\frac{383}{47}\frac{qh^2}{kGA}=\frac{383}{47}\frac{q\varepsilon_h h^4}{EI}=0.005\,432\,6\cdot\frac{qL^4}{EI}$$

$$\theta_2=\frac{625}{47}\frac{qh}{kGA}=\frac{625}{47}\frac{q\varepsilon_h h^3}{EI}=0.017\,731\cdot\frac{qL^3}{EI}$$

$$v_3=\frac{1094}{47}\frac{qh^2}{kGA}=\frac{1094}{47}\frac{q\varepsilon_h h^4}{EI}=0.015\,518\cdot\frac{qL^4}{EI} \qquad (t)$$

$$\theta_3=\frac{750}{47}\frac{qh}{kGA}=\frac{750}{47}\frac{q\varepsilon_h h^3}{EI}=0.021\,277\cdot\frac{qL^3}{EI}$$

Consistent stress resultant expressions in an element are found from (4) and (d) to be

$$\bar{Q}_c=kGA\left[\frac{v_2-v_1}{h}-\theta_1-(\theta_2-\theta_1)\xi\right]$$

$$\bar{M}_c=-EI\left(\frac{\theta_2-\theta_1}{h}\right) \qquad (u)$$

Thus, with assumed constant element properties, the shearing force varies linearly and the bending moment is constant in an element.

The exact solutions for the beam problem are found to be here

$$v=\frac{qL^4}{EI}\left\{\frac{1}{4}\left(\frac{x}{L}\right)^2-\frac{1}{6}\left(\frac{x}{L}\right)^3+\frac{1}{24}\left(\frac{x}{L}\right)^4+\frac{1}{375}\left[\frac{x}{L}-\frac{1}{2}\left(\frac{x}{L}\right)^2\right]\right\}$$

$$\theta=\frac{qL^4}{EI}\left[\frac{1}{2}\frac{x}{L}-\frac{1}{2}\left(\frac{x}{L}\right)^2+\frac{1}{6}\left(\frac{x}{L}\right)^3\right]$$

$$Q=qL\left(1-\frac{x}{L}\right) \qquad (v)$$

$$M=qL^2\left[-\frac{1}{2}+\frac{x}{L}-\frac{1}{2}\left(\frac{x}{L}\right)^2\right]$$

The exact results and the results by the finite element method are shown in Figures (d) to (g) for the deflection, rotation, shearing force and bending moment, respectively.



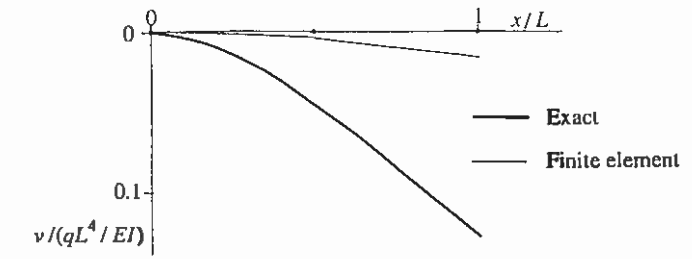



**Figure (d)**

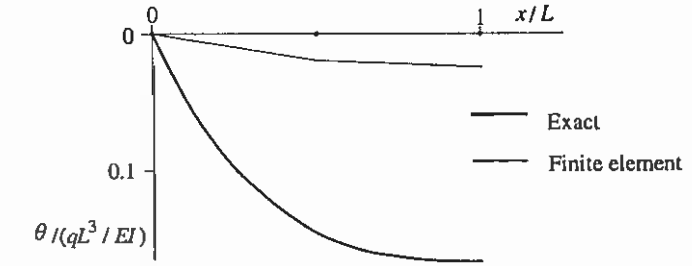



**Figure (e)**

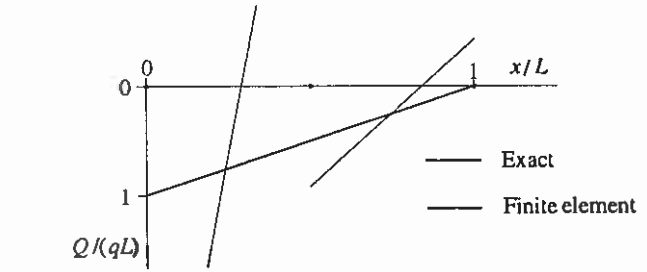



**Figure (f)**

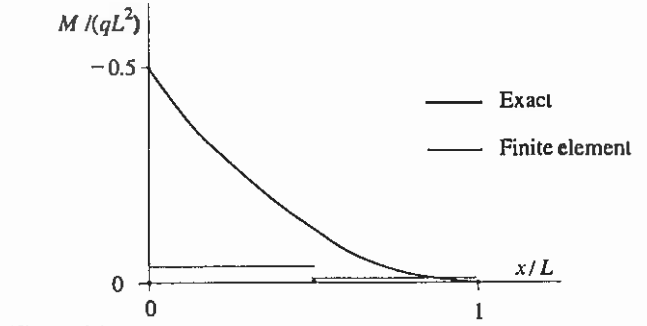



**Figure (g)**

The results by the standard Galerkin finite element method are seen to be very inaccurate and the locking phenomenon thus appears strongly here.

**Remark 5.7.** The mesh in Example 5.1 is crude but even taking this into account the results are still intolerably poor. However, according to the theory presented in Chapter 4 the displacement solution obtained is the best one in the energy norm! This clearly shows that the energy norm is not necessarily a good practical measure in engineering work. In Chapter 4 we had the case where only one function appears in the energy norm. The energy norm concerns two functions here:

$$\| v,\theta \|_a = \left\{ \int_\Omega \left[ kGA\left( \frac{\mathrm{d}v}{\mathrm{d}x} - \theta \right)^2 + EI\left( \frac{\mathrm{d}\theta}{\mathrm{d}x} \right)^2 \right] \mathrm{d}\Omega \right\}^{1/2} \tag{27}$$

It is defined as the square root of the strain energy (the quadratic part) appearing in the potential energy expression (15) (usually without the factor 1/2). Using (4), we see that the right-hand side of (27) can be expressed also as

$$\left[ \int_\Omega \left( \frac{1}{kGA} Q^2 + \frac{1}{EI} M^2 \right) \mathrm{d}\Omega \right]^{1/2} = \left[ \int_\Omega \frac{1}{EI} \left( \frac{EI}{kGAL^2} L^2 Q^2 + M^2 \right) \mathrm{d}\Omega \right]^{1/2}$$
$$= \left[ \int_\Omega \frac{1}{EI} \left( \varepsilon L^2 Q^2 + M^2 \right) \mathrm{d}\Omega \right]^{1/2} \tag{28}$$

These manipulations have been performed to compare the terms due to shearing and bending. The dimensionless number $\varepsilon$ introduced in Example 5.1 by formula (j) gets smaller when the beam gets thinner. However, the actual shearing force and bending moment values in a problem are usually found not to change strongly when the beam gets thinner. In fact, in the cantilever problem considered in Example 5.1 these values do no depend at all on the beam properties. This is because we have then a so-called *statically determinate* (staattisesti määrätty) case (see formulas (v)). Based on this, it is seen from (28) that the contribution from shearing gets smaller and smaller compared to the contribution from bending when the beam gets thinner. From (27) it is then seen that the term $\mathrm{d}v/\mathrm{d}x - \theta = 0$ must remain small for the whole term to remain small. We thus again arrive towards the Bernoulli case discussed from a slightly different point of view in Remark 5.5. The standard finite element solution tries to approximate as well as possible the exact value of (28). So it also tries to keep the term $\mathrm{d}\bar{v}/\mathrm{d}x - \bar{\theta}$ small. Actually the solution is found to put too much emphasis on the limit condition $\mathrm{d}v/\mathrm{d}x - \theta = 0$. Demanding this condition to be exactly valid in a two-noded element gives

$$\frac{\mathrm{d}\bar{v}}{\mathrm{d}x} - \bar{\theta} = \frac{v_2 - v_1}{h} - (1-\xi)\theta_1 - \xi\theta_2 = \frac{v_2 - v_1}{h} - \theta_1 + (\theta_1 - \theta_2)\xi = 0 \tag{29}$$

from which follows two conditions per element:

$$\begin{aligned} &\frac{v_2 - v_1}{h} - \theta_1 = 0 \\ &\theta_1 - \theta_2 = 0 \end{aligned} \tag{30}$$



Applying these relations sequentially element by element, say, for the cantilever beam finite element model starting from the clamped end, it is easily found that no displacements can take place: the locking is complete. It should be finally noted that even when the shearing force $\bar{Q}$ obtains high (unrealistic) values (Figure (f), Example 5.1), the value of $\mathrm{d}\bar{v}/\mathrm{d}x - \bar{\theta}$ in an element can still be very small as the shearing force is obtained by a multiplication with a very large number $kGA$. This is a case where reduced (cf. Section 3.3.3) integration (one point numerical integration applied to the term associated with shearing) makes the otherwise standard Galerkin formulation with two-noded elements to work quite well. In fact, equation (29) is then replaced by

$$\frac{v_2 - v_1}{h} - \theta_1 + (\theta_1 - \theta_2)\frac{1}{2} = 0 \tag{31}$$

which represents now only one kinematical condition per element. This changes the element "softer". □

### 5.1.5 Sensitized potential energy

The quotations given in Section 5.1.1 show that the sensitizing idea of Courant consists of appending a standard functional with a least squares type expression obtained from the governing differential equations (which can be found also from the stationarity condition of the functional as the Euler-Lagrange equations).

Let us now proceed with the Timoshenko beam problem in the way suggested by Courant. The functional is given by (15) and the field equations by (9). We write a *sensitized functional* (sensitoitu funktionaali)

$$\begin{aligned} \Pi_s(v,\theta) &\equiv \Pi(v,\theta) + \Pi^{(0)}(v,\theta) + \Pi^{(1)}(v,\theta) + \Pi^{(2)}(v,\theta) + \cdots \\ &\equiv \Pi(v,\theta) + \frac{1}{2}\int_\Omega \begin{Bmatrix} R_v \\ R_\theta \end{Bmatrix}^{\mathrm{T}} \begin{bmatrix} \tau_{vv} & \tau_{v\theta} \\ \tau_{\theta v} & \tau_{\theta\theta} \end{bmatrix}^{(0)} \begin{Bmatrix} R_v \\ R_\theta \end{Bmatrix} \mathrm{d}\Omega \\ &\quad + \frac{1}{2}\int_\Omega \frac{\mathrm{d}}{\mathrm{d}x}\begin{Bmatrix} R_v \\ R_\theta \end{Bmatrix}^{\mathrm{T}} \begin{bmatrix} \tau_{vv} & \tau_{v\theta} \\ \tau_{\theta v} & \tau_{\theta\theta} \end{bmatrix}^{(1)} \frac{\mathrm{d}}{\mathrm{d}x}\begin{Bmatrix} R_v \\ R_\theta \end{Bmatrix} \mathrm{d}\Omega \\ &\quad + \frac{1}{2}\int_\Omega \frac{\mathrm{d}^2}{\mathrm{d}x^2}\begin{Bmatrix} R_v \\ R_\theta \end{Bmatrix}^{\mathrm{T}} \begin{bmatrix} \tau_{vv} & \tau_{v\theta} \\ \tau_{\theta v} & \tau_{\theta\theta} \end{bmatrix}^{(2)} \frac{\mathrm{d}^2}{\mathrm{d}x^2}\begin{Bmatrix} R_v \\ R_\theta \end{Bmatrix} \mathrm{d}\Omega + \cdots \end{aligned} \tag{32}$$

In the words of Courant we have modified the variational problem without changing the solution as clearly the sensitizing terms disappear for the exact solution. First, the multipliers 1/2 have been introduced just to avoid the final system equations to contain the factors 2 (see Remark 2.6). Second, the *sensitizing parameter* (sensitointiparametri) matrices $[\tau]^{(0)}$, $[\tau]^{(1)}$, $\cdots$ in the corresponding quadratic forms can be taken symmetric without loss of generality (see Remark D.4). In connection with the least squares method in

Appendix D we called the corresponding matrices as *weight factor* (painotekijä) matrices. As the least squares functionals appear here in a somewhat different role, we consequently also employ a different terminology for the matrices. Thirdly, the elements of the sensitizing parameter matrices must naturally have such physical dimensions that the whole expression remains dimensionally homogeneous. Fourth, we must have some criteria to determine suitable values for the parameters. This is considered in the following. In any case, we are no more just "at the mercy" of the pure conventional variational principle. We have now available the possibility to try to steer the discrete solution in a more advantageous direction by a suitable selection of the values of the sensitizing parameters.

In what follows we retain only the first sensitizing integral in (32) so we have the functional

$$\begin{aligned}\Pi_s(v,\theta) &\equiv \Pi(v,\theta)+\Pi^{(0)}(v,\theta)\\ &\equiv \int_\Omega\left[\frac{1}{2}kGA\left(\frac{\mathrm{d}v}{\mathrm{d}x}-\theta\right)^2+\frac{1}{2}EI\left(\frac{\mathrm{d}\theta}{\mathrm{d}x}\right)^2-qv\right]\mathrm{d}\Omega+\mathrm{bt}\\ &+\frac{1}{2}\int_\Omega\begin{Bmatrix}R_v\\R_\theta\end{Bmatrix}^{\mathrm{T}}\begin{bmatrix}\tau_{vv}&\tau_{v\theta}\\\tau_{\theta v}&\tau_{\theta\theta}\end{bmatrix}^{(0)}\begin{Bmatrix}R_v\\R_\theta\end{Bmatrix}\mathrm{d}\Omega\end{aligned}\qquad(33)$$

The problem is thus to determine the three optimal values $\tau_{vv}$, $\tau_{\theta\theta}$, $\tau_{v\theta}=\tau_{\theta v}$. This is effected by making use of certain reference solutions and of a special kind of patch test to be explained in Section 5.2.

**Remark 5.8.** It will be found that the optimal *sensitizing parameter values depend on the mesh* used when the finite element method is applied which should be kept in mind when looking at expressions like (32) and (33). In fact, it is found that the parameter values tend to zero when the mesh size parameter gets to zero. Thus in the theoretical limit considered in convergence studies no sensitizing is needed. But in practice we naturally have always to live with finite meshes. □

### 5.1.6 Sensitized finite element expressions

We proceed here to generate the system equations corresponding to the sensitized functional (33) similarly as in Section 5.1.3 by first generating the equation $\delta\Pi_s=0$. The variation $\delta\Pi$ of the original functional has been performed already in Section 5.1.3 so it remains to study the sensitizing part:

$$\delta\Pi^{(0)}=\delta\left(\frac{1}{2}\int_\Omega\begin{Bmatrix}R_v\\R_\theta\end{Bmatrix}^{\mathrm{T}}\begin{bmatrix}\tau_{vv}&\tau_{v\theta}\\\tau_{\theta v}&\tau_{\theta\theta}\end{bmatrix}^{(0)}\begin{Bmatrix}R_v\\R_\theta\end{Bmatrix}\mathrm{d}\Omega\right)$$



$$\begin{aligned}&=\frac{1}{2}\int_\Omega\delta\begin{Bmatrix}R_v\\R_\theta\end{Bmatrix}^{\mathrm{T}}\begin{bmatrix}\tau_{vv}&\tau_{v\theta}\\\tau_{\theta v}&\tau_{\theta\theta}\end{bmatrix}^{(0)}\begin{Bmatrix}R_v\\R_\theta\end{Bmatrix}\mathrm{d}\Omega\\ &+\frac{1}{2}\int_\Omega\begin{Bmatrix}R_v\\R_\theta\end{Bmatrix}^{\mathrm{T}}\begin{bmatrix}\tau_{vv}&\tau_{v\theta}\\\tau_{\theta v}&\tau_{\theta\theta}\end{bmatrix}^{(0)}\delta\begin{Bmatrix}R_v\\R_\theta\end{Bmatrix}\mathrm{d}\Omega\\ &=\int_\Omega\begin{Bmatrix}L_v(\delta v,\delta\theta)\\L_\theta(\delta v,\delta\theta)\end{Bmatrix}^{\mathrm{T}}\begin{bmatrix}\tau_{vv}&\tau_{v\theta}\\\tau_{\theta v}&\tau_{\theta\theta}\end{bmatrix}^{(0)}\begin{Bmatrix}R_v(v,\theta)\\R_\theta(v,\theta)\end{Bmatrix}\mathrm{d}\Omega\end{aligned}\qquad(34)$$

The steps used should be rather obvious and are explained also in Remark D.5. The calculation rules of Table D.1 are again applied. The two scalars on the second and third row are seen to be equal as transposing the term $\{R\}^{\mathrm{T}}[\tau]^{(0)}\delta\{R\}$ gives (transposition of a scalar does not change its value)

$$\left(\{R\}^{\mathrm{T}}[\tau]^{(0)}\delta\{R\}\right)^{\mathrm{T}}=\delta\{R\}^{\mathrm{T}}\left([\tau]^{(0)}\right)^{\mathrm{T}}\left(\{R\}^{\mathrm{T}}\right)^{\mathrm{T}}=\delta\{R\}^{\mathrm{T}}[\tau]^{(0)}\{R\}\qquad(35)$$

Finally, for instance (see equation (9)),

$$\delta R_v(v,\theta)=\frac{\mathrm{d}}{\mathrm{d}x}\left[kGA\left(\frac{\mathrm{d}\delta v}{\mathrm{d}x}-\delta\theta\right)\right]=L_v(\delta v,\delta\theta)\qquad(36)$$

We again make the interpretations (17): $\delta v=w_v$, $\delta\theta=w_\theta$ to obtain

$$\delta\Pi^{(0)}=\int_\Omega\begin{Bmatrix}L_v(w_v,w_\theta)\\L_\theta(w_v,w_\theta)\end{Bmatrix}^{\mathrm{T}}\begin{bmatrix}\tau_{vv}&\tau_{v\theta}\\\tau_{\theta v}&\tau_{\theta\theta}\end{bmatrix}^{(0)}\begin{Bmatrix}R_v(v,\theta)\\R_\theta(v,\theta)\end{Bmatrix}\mathrm{d}\Omega\qquad(37)$$

The weak form obtained from the sensitized functional (33) is thus now

$$F_s\equiv\delta\Pi_s\equiv\delta\Pi+\delta\Pi^{(0)}\equiv F+F^{(0)}=0\qquad(38)$$

or in full (compare to (24))

$$F_s\equiv\boxed{\begin{aligned}&\int_\Omega\left[\left(\frac{\mathrm{d}w_v}{\mathrm{d}x}-w_\theta\right)kGA\left(\frac{\mathrm{d}v}{\mathrm{d}x}-\theta\right)+\frac{\mathrm{d}w_\theta}{\mathrm{d}x}EI\frac{\mathrm{d}\theta}{\mathrm{d}x}-w_v q\right]\mathrm{d}\Omega\\ &+\int_\Omega\begin{Bmatrix}L_v(w_v,w_\theta)\\L_\theta(w_v,w_\theta)\end{Bmatrix}^{\mathrm{T}}\begin{bmatrix}\tau_{vv}&\tau_{v\theta}\\\tau_{\theta v}&\tau_{\theta\theta}\end{bmatrix}^{(0)}\begin{Bmatrix}R_v(v,\theta)\\R_\theta(v,\theta)\end{Bmatrix}\mathrm{d}\Omega=0\end{aligned}}\qquad(39)$$

Taking the approximation (25), the system equations for node $i$ using the Galerkin method are thus

$$\int_\Omega \left[ \frac{dN_i}{dx} kGA \left( \frac{d\tilde{v}}{dx} - \tilde{\theta} \right) - N_i q \right] d\Omega$$
$$+ \int_\Omega \begin{Bmatrix} L_v(N_i,0) \\ L_\theta(N_i,0) \end{Bmatrix}^T \begin{bmatrix} \tau_{vv} & \tau_{v\theta} \\ \tau_{\theta v} & \tau_{\theta\theta} \end{bmatrix}^{(0)} \begin{Bmatrix} \tilde{R}_v \\ \tilde{R}_\theta \end{Bmatrix} d\Omega = 0$$
$$\int_\Omega \left[ -N_i kGA \left( \frac{d\tilde{v}}{dx} - \tilde{\theta} \right) + \frac{dN_i}{dx} EI \frac{d\tilde{\theta}}{dx} \right] d\Omega \qquad (40)$$
$$+ \int_\Omega \begin{Bmatrix} L_v(0,N_i) \\ L_\theta(0,N_i) \end{Bmatrix}^T \begin{bmatrix} \tau_{vv} & \tau_{v\theta} \\ \tau_{\theta v} & \tau_{\theta\theta} \end{bmatrix}^{(0)} \begin{Bmatrix} \tilde{R}_v \\ \tilde{R}_\theta \end{Bmatrix} d\Omega = 0$$

We shall consider them in more detail in Example 5.2.

## 5.2 DETERMINATION OF SENSITIZING PARAMETER VALUES

### 5.2.1 Reference solutions

Weighted residual methods in a way give up the study of the detailed field equations and consider them only in an average, integrated sense. *We now try to inject information about the actual local solution behavior into the formulation.* Let us consider a generic point in the domain of the solution. To simplify the treatment we assume constant operator data in the differential equations (5.1.9):

$$R_v(v,\theta) \equiv L_v(v,\theta) + q \approx kGA \frac{d^2 v}{dx^2} - kGA \frac{d\theta}{dx} + q = 0$$
$$R_\theta(v,\theta) \equiv L_\theta(v,\theta) \approx kGA \frac{dv}{dx} - kGA\theta + EI \frac{d^2\theta}{dx^2} = 0 \qquad (1)$$

Here $kGA$ and $EI$ are some local constant representative values around the generic point under study. It should be emphasized that the intention is just to make the process of the determination of some roughly suitable sensitizing parameter values simple enough. No error with respect to convergence is introduced in the possible approximations included in (1). This is understood by considering Section 5.3.2.

Since set (1) is linear and has constant coefficients, the solution is found by standard methods of mathematics as the sum of the general solution of the homogeneous system ($q = 0$) and of a particular solution of the full system to be (Through differentiations and eliminations first a fourth order differential equation for $v$ can be derived and solved. After that $\theta$ can be solved from another differential equation.)



$$v = A + Bx + Cx^2 + Dx^3 + q_0 \left( -\frac{1}{2kGA} x^2 + \frac{1}{24EI} x^4 \right) + \cdots$$
$$\theta = A \cdot 0 + B \cdot 1 + C \cdot 2x + D \left( \frac{6EI}{kGA} + 3x^2 \right) + q_0 \frac{1}{6EI} x^3 + \cdots \qquad (2)$$

Here $A$, $B$, $C$, $D$ are integration constants. The loading has been developed into a Taylor series about the generic point:

$$q = q_0 + (q_x)_0 x + \cdots \qquad (3)$$

We have taken for convenience here and in the following without loss of generality the local origin $x = 0$ at the generic point under study. Only the constant part $q_0$ of the loading has been included in the solution shown in (2). It will be found that ending at this suffices for the determination of the sensitizing parameter values.

Now the exact solution around the generic point under study must be approximately — or if the data happens to be constant — exactly according to (2). How can we make use of this information? At a first glance one could speculate that some kind of iterative procedure might be needed, as the values of the integration constants would have to be found at the point under question from a preliminary numerical solution. Fortunately this is not the case. It will be presently seen that the values of the integration constants are not needed at all.

We collect $v$, $\theta$ and $q$ in a column matrix and obtain the presentation

$$\begin{Bmatrix} v \\ \theta \\ q \end{Bmatrix} = A \begin{Bmatrix} 1 \\ 0 \\ 0 \end{Bmatrix} + B \begin{Bmatrix} x \\ 1 \\ 0 \end{Bmatrix} + C \begin{Bmatrix} x^2 \\ 2x \\ 0 \end{Bmatrix} + D \begin{Bmatrix} x^3 \\ 6EI/(kGA) + 3x^2 \\ 0 \end{Bmatrix}$$
$$+ q_0 \begin{Bmatrix} -1/(2kGA) \cdot x^2 + 1/(24EI) \cdot x^4 \\ 1/(6EI) \cdot x^3 \\ 1 \end{Bmatrix} + \cdots \qquad (4)$$

We call this combination of $v$, $\theta$ and $q$ as the *reference solution* (referenssiratkaisu). The values $A$, $B$, $C$, $D$, $q_0$, $\cdots$ fix the solution. We obtain specific reference solutions taking consecutively only $A \neq 0$, only $B \neq 0$ only $C \neq 0$, only $D \neq 0$ only $q_0 \neq 0$, $\cdots$ These give the consecutive reference solutions (see Remark 5.9)

$$
\begin{aligned}
&v=1, \quad \theta=0, \quad q=0,\\
&v=x, \quad \theta=1, \quad q=0,\\
&v=x^2, \quad \theta=2x, \quad q=0,\\
&\cdots
\end{aligned}
\tag{5}
$$

As a check one can take the solutions consecutively from each line in (5) (or from each column matrix on the right-hand side in (4)) and find that equations (1) are always satisfied.

**Remark 5.9.** It will be found that $A$, $B$, $\cdots$ cancel in the equations used in the patch test in Section 5.2.3 so that we can simply take $A=1$, $B=1$, $\cdots$. □

### 5.2.2 Series form reference solutions

The reference solutions found in the previous section were obtained using the mathematics theory for *ordinary* linear differential equations. In two or more dimensions, the field equations are *partial* differential equations and a corresponding simple mathematical theory does not exist. For these situations we can try to use a series type approach, Freund and Salonen (1998). The ideas are introduced here to be later extended into multidimensional cases.

The starting point is again the simplified field equations (1):

$$
\begin{aligned}
&kGA\,v_{,x}-kGA\theta_x+q=0\\
&kGA\,v_x-kGA\theta+EI\,\theta_{,x}=0
\end{aligned}
\tag{6}
$$

We employ the obvious new notation for the derivatives to simplify the formulas. We develop the unknowns into Taylor series about the generic point ($x=0$):

$$
\begin{aligned}
v(x)&=v_0+(v_x)_0\,x+\frac{1}{2}(v_{,x})_0\,x^2+\frac{1}{6}(v_{,xx})_0\,x^3+\frac{1}{24}(v_{,xxx})_0\,x^4+\cdots\\
\theta(x)&=\theta_0+(\theta_x)_0\,x+\frac{1}{2}(\theta_{,x})_0\,x^2+\frac{1}{6}(\theta_{,xx})_0\,x^3+\frac{1}{24}(\theta_{,xxx})_0\,x^4+\cdots
\end{aligned}
\tag{7}
$$

and the same also for the loading:

$$
q(x)=q_0+(q_x)_0\,x+\frac{1}{2}(q_{,x})_0\,x^2+\cdots \tag{8}
$$

The subscript 0 refers to a quantity evaluated at the local origin. Evaluating (6) and its differentiated forms at the origin gives



$$
\begin{aligned}
&kGA(v_{,x})_0-kGA(\theta_x)_0+q_0=0\\
&kGA(v_{,xx})_0-kGA(\theta_{,x})_0+(q_x)_0=0\\
&kGA(v_{,xxx})_0-kGA(\theta_{,xx})_0+(q_{,x})_0=0\\
&kGA(v_{,xxxx})_0-kGA(\theta_{,xxx})_0+(q_{,xx})_0=0\\
&\cdots\\
&kGA(v_x)_0-kGA(\theta)_0+EI(\theta_{,x})_0=0\\
&kGA(v_{,x})_0-kGA(\theta_x)_0+EI(\theta_{,x})_0=0\\
&kGA(v_{,xx})_0-kGA(\theta_{,x})_0+EI(\theta_{,xx})_0=0\\
&kGA(v_{,xxx})_0-kGA(\theta_{,xx})_0+EI(\theta_{,xxx})_0=0\\
&\cdots
\end{aligned}
\tag{9}
$$

*Equations* (9) *contain information about the governing field equations.* If we end as shown, there are nine unknown quantities $(v_x)_0$, $(v_{,x})_0$, $(v_{,xx})_0$, $(v_{,xxx})_0$, $\theta$, $(\theta_x)_0$, $(\theta_{,x})_0$, $(\theta_{,xx})_0$, $(\theta_{,xxx})_0$ in the six equations (9). Thus we can try solve for six of the quantities and express them in the rest. Here it seems logical to consider the lowest order derivatives $(v_x)_0$, $\theta_0$, $(\theta_x)_0$, as given (compare to initial or boundary conditions in general which are of lower order than the highest order terms in the differential equations) and to solve the rest from (9). After that the solutions are substituted in (7). As the calculations would be rather tedious by hand, we perform them below using Mathematica:

```
eqs = {kGA v'' - kGA θx + q == 0,
       kGA v''' - kGA θ'' + qx == 0,
       kGA v'''' - kGA θ''' + qx' == 0,
       kGA vx - kGA θ + EI θ'' == 0,
       kGA v'' - kGA θx + EI θ''' == 0,
       kGA v''' - kGA θ'' + EI θ'''' == 0};

sol = Solve[eqs, {v'', v''', v'''', θx, θ'', θ'''}]
```

$$
\left\{\left\{\theta''''\to\frac{q_x}{EI},\ v''''\to-\frac{-kGA\,q+EI\,q_{x'}}{EI\,kGA},\ v'''\to-\frac{q_x}{kGA}-\frac{-kGA\,\theta+kGA\,v_x}{EI},\right.\right.
\left.\left.\theta'''\to\frac{q}{EI},\ v''\to-\frac{q-kGA\,\theta_x}{kGA},\ \theta''\to-\frac{-kGA\,\theta+kGA\,v_x}{EI}\right\}\right\}
$$

```
v[x_] := v + vx x + 1/2 v'' x^2 + 1/6 v''' x^3 + 1/24 v'''' x^4

Collect[v[x] /. sol, {v, vx, θ, θx, q, qx, qx'}]
```

$$
\left\{v+q\left(-\frac{x^2}{2\,kGA}+\frac{x^4}{24\,EI}\right)+\frac{kGA\,x^3\,\theta}{6\,EI}-\frac{x^3\,q_x}{6\,kGA}-\frac{x^4\,q_{x'}}{24\,kGA}+\left(x-\frac{kGA\,x^3}{6\,EI}\right)v_x+\frac{x^2\,\theta_x}{2}\right\}
$$

```
θ[x_] := θ + θx x + 1/2 θ'' x^2 + 1/6 θ''' x^3 + 1/24 θ'''' x^4
```

Collect[θ[x] /. sol, { $v_x$, θ, $\theta_x$, $q_0$, $q_x$, $q_{xx}$ }]

$$\left\{\frac{q\,x^3}{6\,EI}+\left\{1+\frac{kGA\,x^2}{2\,EI}\right\}\theta+\frac{x^4\,q_x}{24\,EI}-\frac{kGA\,x^2\,v_x}{2\,EI}+x\,\theta_x\right\}$$

We have obtained the tentative reference solution

$$\begin{Bmatrix} v \\ \theta \\ q \end{Bmatrix} = v_0 \begin{Bmatrix} 1 \\ 0 \\ 0 \end{Bmatrix} + (v_x)_0 \begin{Bmatrix} x - kGA/(6EI)\cdot x^3 \\ -kGA/(2EI)\cdot x^2 \\ 0 \end{Bmatrix} + \theta_0 \begin{Bmatrix} kGA/(6EI)\cdot x^3 \\ 1 + kGA/(2EI)\cdot x^2 \\ 0 \end{Bmatrix}$$
$$+ (\theta_x)_0 \begin{Bmatrix} 1/2\cdot x^2 \\ x \\ 0 \end{Bmatrix} + q_0 \begin{Bmatrix} -1/(2kGA)\cdot x^2 + 1/(24EI)\cdot x^4 \\ 1/(6EI)\cdot x^3 \\ 1 \end{Bmatrix}$$
$$+ (q_x)_0 \begin{Bmatrix} -1/(6kGA)\cdot x^3 \\ 1/(24EI)\cdot x^4 \\ x \end{Bmatrix} + (q_{xx})_0 \begin{Bmatrix} -1/(24kGA)\cdot x^4 \\ 0 \\ 1/2\cdot x^2 \end{Bmatrix} + \cdots \qquad (10)$$

The terms in the first five column matrices on the right-hand side are again reference solutions: each of them is found to satisfy separately equations (1). The multipliers $v_0$, $(v_x)_0$, $\theta_0$, $(\theta_x)_0$ have now the role of the integration constants. Comparison with (4) shows that only the first of the four first reference solutions are identical. However, certain linear combinations of the new reference solutions are seen produce the old solutions (4) so the presentations are equivalent. The last two solutions in (10) are found not to be any more exact reference solutions. This is obviously due to the truncation in the series representation of the unknowns. However, the five separate exact reference solutions are here enough for the determination of the sensitizing parameter values.

The series type approach to reference solutions can be extended to two or more dimensions. We will give a demonstration of this in Chapter 12.

### 5.2.3 Sensitizing patch test

The obvious criterion for selection of the sensitizing parameter values is *the goal of achieving the nodally exact solution*. This has been emphasized already in Remark 4.1 and we continue here to follow this goal.



An alternative version of the patch test, described in its conventional form in Section 4.1, is now employed for the determination of the sensitizing parameter values.

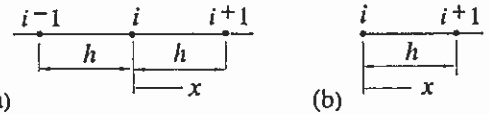



**Figure 5.2 (a)** Two-element patch. **(b)** One-element patch.

We consider here the simplest case of two-noded elements similarly as in Example 5.1. A generic element is "cloned" to form a uniform mesh and a typical two-element patch is taken as the system (Figure 5.2 (a)). Using the sensitized formulation, the two general finite element system equations corresponding to node $i$ are formed. The details are given in Example 5.2. Constant parameter values are assumed in an element. The nodal values are taken according to the reference solutions (5):

$$\begin{aligned} &v_{i-1}=1, \quad v_i=1, \quad v_{i+1}=1 \\ &\theta_{i-1}=0, \quad \theta_i=0, \quad \theta_{i+1}=0 \end{aligned} \qquad (11)$$

$$\begin{aligned} &v_{i-1}=-h, \quad v_i=0, \quad v_{i+1}=h \\ &\theta_{i-1}=1, \quad \theta_i=1, \quad \theta_{i+1}=1 \\ &\ldots \end{aligned}$$

In the fifth reference solution used here, the constant loading $q_0=1$ is already included. The nodal values (11) are substituted consecutively. Correspondingly, there are finally obtained the following five sets of two-equation systems:

$$\begin{aligned} &0=0 \\ &\tau_{v\theta}-\tau_{\theta v}=0 \end{aligned} \qquad (12)$$

$$\begin{aligned} &-\tau_{v\theta}+\tau_{\theta v}=0 \\ &0=0 \end{aligned} \qquad (13)$$

$$\begin{aligned} &0=0 \\ &\tau_{\theta v}=0 \end{aligned} \qquad (14)$$

$$\begin{aligned} &-EI/(kGA)\tau_{v\theta}+\left[h^2-EI/(kGA)\right]\tau_{\theta v}=0 \\ &-\tau_{vv}+EI/(kGA)\cdot\tau_{\theta\theta}=0 \end{aligned} \qquad (15)$$

$$\begin{aligned}&\left(kGA+12EI/h^2\right)\tau_{\theta\theta}+1=0\\&-\tau_{v\theta}+2\tau_{\theta v}-12EI/\left(kGAh^2\right)\cdot(\tau_{v\theta}+\tau_{\theta v})=0\end{aligned}\tag{16}$$

Following the comments in Section 5.3.2, we have for generality developed the expressions without assuming a symmetric sensitizing parameter matrix. Immediately, sets (12) and (13) demand symmetry, $\tau_{v\theta}=\tau_{\theta v}$, and from (14),

$$\tau_{v\theta}=\tau_{\theta v}=0\tag{17}$$

and from (15) and (16),

$$\begin{aligned}\tau_{\theta\theta}&=-\frac{1}{kGA+12EI/h^2}=-\frac{1}{1+12\varepsilon_h}\frac{1}{kGA}\\\tau_{vv}&=\frac{EI}{kGA}\tau_{\theta\theta}=-\frac{EI/kGA}{kGA+12EI/h^2}=-\frac{\varepsilon_h}{1+12\varepsilon_h}\frac{h^2}{kGA}\end{aligned}\tag{18}$$

where the dimensionless number

$$\varepsilon_h=\frac{EI}{kGAh^2}\tag{19}$$

Thus the sensitizing parameter matrix is found to be diagonal with negative elements and this means physically that sensitizing makes the solution more soft counteracting locking.

**Remark 5.10.** In the conventional patch test an *irregular patch* is taken and the test is performed to determine the nodal values at the internal node: are they according to the polynomial expressions? In the sensitizing patch test a *regular patch* is taken and also the internal nodal values are fixed according to the reference solutions. The test is performed to determine the sensitizing parameter values. The intuitive idea is that if the element as cloned to form a regular mesh produces a good response, the element will probably behave reasonably well with the generic parameter values obtained also as an individual element in an irregular mesh. Numerical results have confirmed this behavior. □

**Remark 5.11.** As the finite element system equations are linear with respect to the nodal values and with respect to the source term, using the optimal values of the sensitizing parameters found above, *the patch test is now seen to be passed for the full expression* (4) *with arbitrary values of A, B,* ⋯ . Thus in the case of constant beam properties, uniform mesh and constant source term, the finite element solution will be nodally exact at least with essential boundary conditions. With essential boundary conditions the active system equations consist of equations for only the internal nodes, which explains the reservation in the sentence above. In the case of natural boundary conditions we can perform a study of a "boundary patch" around a boundary node which would be in one dimension a one-element patch (see



Figure 5.2(b)). This is not considered to be a very important feature in the design of a sensitized method. It will be studied in some detail in Section 6.2.3. □

**Remark 5.12.** In the above example case we obtained the four (if the parameter matrix is assumed originally as non-symmetric) unknown parameters in principle from ten system equations. Thus the system from which the parameters were determined was seemingly *overdetermined* (ylimääräytyvä) in the sense that there were more equations than unknowns. Here, however, the system was consistent so that four linearly independent equations could be found which determine the solution uniquely. In the general case, it may well happen that using several reference solutions we may really obtain an overdetermined system. Then the conventional least squares solution method employed in connection of overdetermined systems could be used to try solve the set. □

**Example 5.2.** We evaluate first the element contributions for a two-noded element (Figure (a)) using sensitized formulation. Constant element properties and constant sensitizing parameter values in an element are assumed.

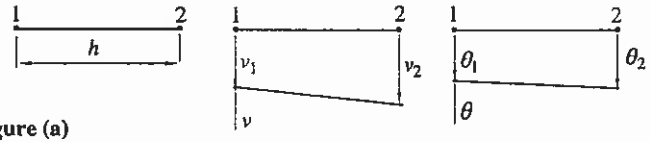



**Figure (a)**

The numbering of the nodal parameters is done similarly as in Example 5.1. Making use of expressions (5.1.39) on the element level gives

$$\begin{aligned}F_1^s\equiv F_1+F_1^{(0)}&\equiv\int_{\Omega^e}\left[\frac{dN_1}{dx}kGA\left(\frac{d\bar v}{dx}-\bar\theta\right)-N_1q\right]d\Omega\\&+\int_{\Omega^e}\begin{Bmatrix}L_v(N_1,0)\\L_\theta(N_1,0)\end{Bmatrix}^{\mathrm T}\begin{bmatrix}\tau_{vv}&\tau_{v\theta}\\\tau_{\theta v}&\tau_{\theta\theta}\end{bmatrix}^{(0)}\begin{Bmatrix}\bar R_v\\\bar R_\theta\end{Bmatrix}d\Omega\\F_2^s\equiv F_2+F_2^{(0)}&\equiv\int_{\Omega^e}\left[-N_1kGA\left(\frac{d\bar v}{dx}-\bar\theta\right)+\frac{dN_1}{dx}EI\frac{d\bar\theta}{dx}\right]d\Omega\\&+\int_{\Omega^e}\begin{Bmatrix}L_v(0,N_1)\\L_\theta(0,N_1)\end{Bmatrix}^{\mathrm T}\begin{bmatrix}\tau_{vv}&\tau_{v\theta}\\\tau_{\theta v}&\tau_{\theta\theta}\end{bmatrix}^{(0)}\begin{Bmatrix}\bar R_v\\\bar R_\theta\end{Bmatrix}d\Omega\\F_3^s\equiv F_3+F_3^{(0)}&\equiv\int_{\Omega^e}\left[\frac{dN_2}{dx}kGA\left(\frac{d\bar v}{dx}-\bar\theta\right)-N_2q\right]d\Omega\\&+\int_{\Omega^e}\begin{Bmatrix}L_v(N_2,0)\\L_\theta(N_2,0)\end{Bmatrix}^{\mathrm T}\begin{bmatrix}\tau_{vv}&\tau_{v\theta}\\\tau_{\theta v}&\tau_{\theta\theta}\end{bmatrix}^{(0)}\begin{Bmatrix}\bar R_v\\\bar R_\theta\end{Bmatrix}d\Omega\\F_4^s\equiv F_4+F_4^{(0)}&\equiv\int_{\Omega^e}\left[-N_2kGA\left(\frac{d\bar v}{dx}-\bar\theta\right)+\frac{dN_2}{dx}EI\frac{d\bar\theta}{dx}\right]d\Omega\\&+\int_{\Omega^e}\begin{Bmatrix}L_v(0,N_2)\\L_\theta(0,N_2)\end{Bmatrix}^{\mathrm T}\begin{bmatrix}\tau_{vv}&\tau_{v\theta}\\\tau_{\theta v}&\tau_{\theta\theta}\end{bmatrix}^{(0)}\begin{Bmatrix}\bar R_v\\\bar R_\theta\end{Bmatrix}d\Omega\end{aligned}\tag{a}$$

The standard parts $F_1$, $F_2$, $F_3$, $F_4$ have been evaluated already in Example 5.1 (see formulas (i)). For the sensitizing terms $F_1^{(0)}$, $F_2^{(0)}$, $F_3^{(0)}$, $F_4^{(0)}$, the simplified residual expressions (1):

$$R_v(v,\theta)=L_v(v,\theta)+q=kGA\frac{d^2v}{dx^2}-kGA\frac{d\theta}{dx}+q$$
$$R_\theta(v,\theta)=L_0(v,\theta)=kGA\frac{dv}{dx}-kGA\theta+EI\frac{d^2\theta}{dx^2} \tag{b}$$

and the corresponding approximations

$$\tilde{R}_v=0-kGA\frac{d\tilde{\theta}}{dx}+q=0\cdot v_1+\frac{kGA}{h}\theta_1+0\cdot v_2-\frac{kGA}{h}\theta_2+q$$
$$\tilde{R}_\theta=kGA\frac{d\tilde{v}}{dx}-kGA\tilde{\theta}+0=-\frac{kGA}{h}v_1-kGAN_1\theta_1+\frac{kGA}{h}v_2-kGAN_2\theta_2 \tag{c}$$

are used. It is seen that for this low order approximation the bending behavior — expressed through the bending stiffness — disappears. From (b),

$$L_v(N_1,0)=0, \qquad L_\theta(N_1,0)=-\frac{kGA}{h}$$
$$L_v(0,N_1)=\frac{kGA}{h}, \qquad L_\theta(0,N_1)=-kGAN_1$$
$$L_v(N_2,0)=0, \qquad L_\theta(N_2,0)=\frac{kGA}{h} \tag{d}$$
$$L_v(0,N_2)=-\frac{kGA}{h}, \qquad L_\theta(0,N_2)=-kGAN_2$$

For instance, writing

$$\underset{1\times1}{F_1^{(0)}}=\underset{1\times4}{[K_1]^{(0)}}\underset{4\times1}{\{a\}}-\underset{1\times1}{b_1^{(0)}} \tag{e}$$

we obtain

$$[K_1]^{(0)}\{a\}=\int_{\Omega^e}\begin{Bmatrix}L_v(N_1,0)\\L_\theta(N_1,0)\end{Bmatrix}^{\mathrm{T}}\begin{bmatrix}\tau_{vv}&\tau_{v\theta}\\\tau_{\theta v}&\tau_{\theta\theta}\end{bmatrix}^{(0)}\begin{Bmatrix}\tilde{R}_v\\\tilde{R}_\theta\end{Bmatrix}d\Omega$$
$$=\int_{\Omega^e}\begin{bmatrix}0&-\frac{kGA}{h}\end{bmatrix}\begin{bmatrix}\tau_{vv}&\tau_{v\theta}\\\tau_{\theta v}&\tau_{\theta\theta}\end{bmatrix}^{(0)}\cdot$$
$$\cdot\begin{bmatrix}0&\frac{kGA}{h}&0&-\frac{kGA}{h}\\-\frac{kGA}{h}&-kGAN_1&\frac{kGA}{h}&-kGAN_2\end{bmatrix}d\Omega\begin{Bmatrix}v_1\\\theta_1\\v_2\\\theta_2\end{Bmatrix}$$



$$=-\frac{kGA}{h}[\tau_{\theta v}\quad\tau_{\theta\theta}]\int_{\Omega^e}\begin{bmatrix}0&\frac{kGA}{h}&0&-\frac{kGA}{h}\\-\frac{kGA}{h}&-kGAN_1&\frac{kGA}{h}&-kGAN_2\end{bmatrix}d\Omega\begin{Bmatrix}v_1\\\theta_1\\v_2\\\theta_2\end{Bmatrix}$$
$$=-\frac{kGA}{h}[\tau_{\theta v}\quad\tau_{\theta\theta}]\begin{bmatrix}0&kGA&0&-kGA\\-kGA&-kGAh/2&kGA&-kGAh/2\end{bmatrix}\begin{Bmatrix}v_1\\\theta_1\\v_2\\\theta_2\end{Bmatrix}$$
$$=\frac{(kGA)^2}{h}\left[\tau_{\theta\theta}\cdot v_1+(-\tau_{\theta v}+h\tau_{\theta\theta}/2)\theta_1-\tau_{\theta\theta}\cdot v_2+(\tau_{\theta v}+h\tau_{\theta\theta}/2)\theta_2\right] \tag{f}$$

and

$$b_1^{(0)}=-\int_{\Omega^e}\begin{Bmatrix}L_v(N_1,0)\\L_\theta(N_1,0)\end{Bmatrix}^{\mathrm{T}}\begin{bmatrix}\tau_{vv}&\tau_{v\theta}\\\tau_{\theta v}&\tau_{\theta\theta}\end{bmatrix}^{(0)}\begin{Bmatrix}q\\0\end{Bmatrix}d\Omega$$
$$=-\int_{\Omega^e}\begin{bmatrix}0&-\frac{kGA}{h}\end{bmatrix}\begin{bmatrix}\tau_{vv}&\tau_{v\theta}\\\tau_{\theta v}&\tau_{\theta\theta}\end{bmatrix}^{(0)}\begin{Bmatrix}q\\0\end{Bmatrix}d\Omega$$
$$=\frac{kGA}{h}[\tau_{\theta v}\quad\tau_{\theta\theta}]\int_{\Omega^e}\begin{Bmatrix}q\\0\end{Bmatrix}d\Omega$$
$$=kGA\tau_{\theta v}q_0 \tag{g}$$

The last form of (g) is obtained assuming a constant loading $q_0$. The other contributions are arrived at similarly. Denoting

$$\underset{4\times1}{\{F\}^{(0)}}=\underset{4\times4}{[K]^{(0)}}\underset{4\times1}{\{a\}}-\underset{4\times1}{\{b\}^{(0)}} \tag{h}$$

we obtain the sensitizing stiffness matrix

$$[K]^{(0)}=\frac{(kGA)^2}{h}\begin{bmatrix}\overset{1}{\tau_{\theta\theta}}&\overset{2}{-\tau_{\theta v}+h\tau_{\theta\theta}/2}\\-\tau_{v\theta}+h\tau_{\theta\theta}/2&\tau_{vv}-h\tau_{v\theta}/2-h\tau_{\theta v}/2+h^2\tau_{\theta\theta}/3\\-\tau_{\theta\theta}&\tau_{\theta v}-h\tau_{\theta\theta}/2\\\tau_{v\theta}+h\tau_{\theta\theta}/2&-\tau_{vv}-h\tau_{v0}/2+h\tau_{\theta v}/2+h^2\tau_{\theta\theta}/6\end{bmatrix}$$
$$\begin{matrix}\overset{3}{-\tau_{\theta\theta}}&\overset{4}{\tau_{\theta v}+h\tau_{\theta\theta}/2}\\\tau_{v\theta}-h\tau_{\theta\theta}/2&-\tau_{vv}-h\tau_{v0}/2+h\tau_{\theta v}/2+h^2\tau_{\theta\theta}/6\\\tau_{\theta\theta}&-\tau_{\theta v}-h\tau_{\theta\theta}/2\\-\tau_{v0}-h\tau_{\theta\theta}/2&\tau_{vv}+h\tau_{v0}/2+h\tau_{0v}/2+h^2\tau_{00}/3\end{matrix}\begin{matrix}1\\2\\3\\4\end{matrix} \tag{i}$$

and the column vector

$$\{b\}^{(0)} = kGAq_0 \begin{Bmatrix} \tau_{0v} \\ -\tau_{vv} + h\tau_{0v}/2 \\ -\tau_{0v} \\ \tau_{vv} + h\tau_{0v}/2 \end{Bmatrix} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} \qquad \text{(j)}$$

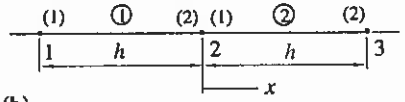



**Figure (b)**

We proceed now to form the system equations corresponding to node $i$ of the patch shown in Figure 5.2(a). However, to have simple indexing we consider the mesh shown in Figure (b) and its node 2. Similarly as in Example 5.1, we arrive at the system equations

$$\begin{aligned} &K_{31}v_1 + K_{32}\theta_1 + K_{33}v_2 + K_{34}\theta_2 + K_{35}v_3 + K_{36}\theta_3 - b_3 = 0 \\ &K_{41}v_1 + K_{42}\theta_1 + K_{43}v_2 + K_{44}\theta_2 + K_{45}v_3 + K_{46}\theta_3 - b_4 = 0 \end{aligned} \qquad \text{(k)}$$

or in more detail

$$\begin{aligned} &K_{31}^1 v_1 + K_{32}^1\theta_1 + \left(K_{33}^1 + K_{11}^2\right)v_2 + \left(K_{34}^1 + K_{12}^2\right)\theta_2 + K_{13}^2 v_3 + K_{14}^2\theta_3 - \left(b_3^1 + b_1^2\right) = 0 \\ &K_{41}^1 v_1 + K_{42}^1\theta_1 + \left(K_{43}^1 + K_{21}^2\right)v_2 + \left(K_{44}^1 + K_{22}^2\right)\theta_2 + K_{23}^2 v_3 + K_{24}^2\theta_3 - \left(b_4^1 + b_2^2\right) = 0 \end{aligned} \qquad \text{(l)}$$

Although we apply the sensitized formulation, we for simplicity neglect the index s. We collect the terms from formulas (i) of Example 5.1 and from formulas (i) and (j) here:

$$\begin{aligned} &K_{31} = -\frac{kGA}{h} + \frac{(kGA)^2}{h}(-\tau_{00}), \quad K_{32} = -\frac{kGA}{2} + \frac{(kGA)^2}{h}(\tau_{0v} - h\tau_{00}/2) \\ &K_{33} = \frac{2kGA}{h} + \frac{(kGA)^2}{h}2\tau_{00}, \quad K_{34} = \frac{(kGA)^2}{h}(-\tau_{v0} - \tau_{0v}) \\ &K_{35} = -\frac{kGA}{h} + \frac{(kGA)^2}{h}(-\tau_{00}), \quad K_{36} = \frac{kGA}{2} + \frac{(kGA)^2}{h}(\tau_{0v} + h\tau_{00}/2) \\ &b_3 = q_0 h \end{aligned} \qquad \text{(m)}$$

$$\begin{aligned} &K_{41} = \frac{kGA}{2} + \frac{(kGA)^2}{h}(\tau_{v0} + h\tau_{00}/2) \\ &K_{42} = \frac{kGAh}{6} - \frac{EI}{h} + \frac{(kGA)^2}{h}\left(-\tau_{vv} - h\tau_{0v}/2 + h\tau_{v0}/2 + h^2\tau_{00}/6\right) \\ &K_{43} = \frac{(kGA)^2}{h}(-\tau_{v0} - \tau_{0v}) \\ &K_{44} = \frac{2kGAh}{3} + \frac{2EI}{h} + \frac{(kGA)^2}{h}\left(2\tau_{vv} + 2h^2\tau_{00}/3\right) \end{aligned} \qquad \text{(n)}$$



$$\begin{aligned} &K_{45} = -\frac{kGA}{2} + \frac{(kGA)^2}{h}(\tau_{v0} - h\tau_{00}/2) \\ &K_{46} = \frac{kGAh}{6} - \frac{EI}{h} + \frac{(kGA)^2}{h}\left(-\tau_{vv} - h\tau_{v0}/2 + h\tau_{0v}/2 + h^2\tau_{00}/6\right) \\ &b_4 = kGAq_0 h\tau_{0v} \end{aligned}$$

The first reference solution (5) gives the nodal values ($i \triangleq 2$)

$$\begin{aligned} &v_1 = 1, \quad v_2 = 1, \quad v_3 = 1 \\ &\theta_1 = 0, \quad \theta_2 = 0, \quad \theta_3 = 0 \end{aligned} \qquad \text{(o)}$$

with no loading. The corresponding system equations are thus

$$\begin{aligned} &\left[-\frac{kGA}{h} + \frac{(kGA)^2}{h}(-\tau_{00})\right]\cdot 1 \\ &+\left[\frac{2kGA}{h} + \frac{(kGA)^2}{h}2\tau_{00}\right]\cdot 1 \\ &+\left[-\frac{kGA}{h} + \frac{(kGA)^2}{h}(-\tau_{00})\right]\cdot 1 = 0 \end{aligned} \qquad \text{(p)}$$

$$\begin{aligned} &\left[\frac{kGA}{2} + \frac{(kGA)^2}{h}(\tau_{v0} + h\tau_{00}/2)\right]\cdot 1 \\ &+\left[\frac{(kGA)^2}{h}(-\tau_{v0} - \tau_{0v})\right]\cdot 1 \\ &+\left[-\frac{kGA}{2} + \frac{(kGA)^2}{h}(\tau_{v0} - h\tau_{00}/2)\right] = 0 \end{aligned}$$

Further development gives

$$\begin{aligned} &0 = 0 \\ &\tau_{v\theta} - \tau_{\theta v} = 0 \end{aligned} \qquad \text{(q)}$$

This is set (12). Equations (13) - (16) are obtained similarly.

### 5.2.3 Refined stress resultant expressions

Before recalculating with sensitizing the problem of Example 5.1, we consider the evaluation of the *stress resultants* (jännitysresultantti) — shearing forces and bending moments — from the sensitized solution. The displacement assumptions of a linearly varying deflection and rotation are clearly very unrealistic. On the basis of the optimal design of the element properties by

sensitizing, the element is found, however, to be able to give nodal values of good accuracy. As the element is very simple, so are the resulting finite element expressions and the element is thus convenient for practical calculations. The final overall results must in any case be interpreted simply just as kind of sample data obtained for selected points (nodes) of the structure. From this raw material one can try to extract more refined results by rejecting the original displacement assumptions and considering locally some better alternatives.

We recall first the consistent shearing force and bending moment expressions for the two-noded element (see expressions (u), Example 5.1):

$$\begin{aligned} \tilde{Q}_c &= kGA\left[\frac{v_2-v_1}{h}-\theta_1-(\theta_2-\theta_1)\xi\right] \\ \tilde{M}_c &= -EI\left(\frac{\theta_2-\theta_1}{h}\right) \end{aligned} \tag{20}$$

Subscript c refers to "consistent".

The actual analytical solution for the deflection $v$ and for the cross section rotation $\theta$ of a uniform Timoshenko beam (we associate again some local constant representative data for an element) is given by (2) with $A$, $B$, $C$, $D$ as the integration constants. We now consider the deflections $v_1$, $v_2$ and the rotations $\theta_1$, $\theta_2$ at the beam element ends as given. Using these four boundary conditions, we can solve the integration constants. The expression for the deflection is found to be (no distributed loading)

$$\begin{aligned} v &= H_1 v_1 + H_2(\theta_1+\gamma) + H_3 v_2 + H_4(\theta_2+\gamma) \\ &= H_1 v_1 + H_2\theta_1 + H_3 v_2 + H_4\theta_2 + (H_2+H_4)\gamma \end{aligned} \tag{21}$$

where

$$\begin{aligned} H_1 &= 1-3\xi^2+2\xi^3 \\ H_2 &= \left(\xi-2\xi^2+\xi^3\right)h \\ H_3 &= 3\xi^2-2\xi^3 \\ H_4 &= \left(-\xi^2+\xi^3\right)h \end{aligned} \tag{22}$$

and where $h$ is the element length. Also, the shearing strain (5.1.6):

$$\gamma = \frac{dv}{dx}-\theta \tag{23}$$



is here a constant. The functions $H$ happen to be the so-called cubic *Hermitian* shape functions ($\xi \in [0,1]$) used in some situations in the finite element method, e.g., Zienkiewicz and Taylor (2000, p. 36, 435).

The shape functions are sketched in Figure 5.3. An approximation in a Hermitian element for a generic function $\phi(x)$ is ($\xi$ can be expressed in $x$)

$$\phi(x) \approx H_1(x)\phi_1 + H_2(x)\phi_1' + H_3(x)\phi_2 + H_2(x)\phi_2' \tag{24}$$

where

$$\phi_1' = \left(\frac{d\phi}{dx}\right)_1, \qquad \phi_2' = \left(\frac{d\phi}{dx}\right)_2 \tag{25}$$

This is a case — not considered in this text further — where some of the nodal parameters in a finite element method can be derivatives of the function under study. The Hermitian element has been used in problems where a $C^1$ continuous approximation is necessary; for instance in structural mechanics especially in connection with the Bernoulli beam theory.

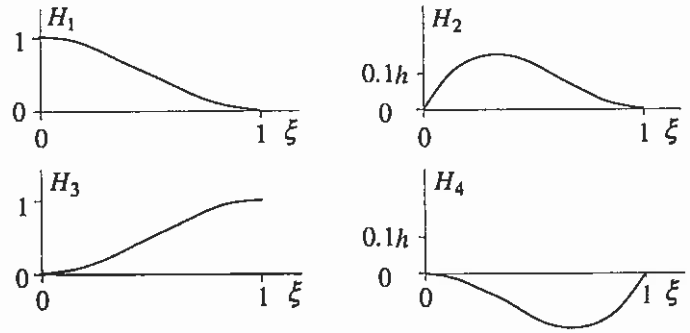



**Figure 5.3** Cubic Hermitian shape functions.

It seems quite natural to try to make use of the rather realistic deflection expression (21) in post-processing. (We could have included here also the effect of a distributed loading but this idea seems to be difficult to generalize for plate problems so we reject this possibility.)

The constant shearing strain is associated via (5.1.4) and (5.1.6) with a constant shearing force:

$$\gamma = \frac{Q}{kGA} \tag{26}$$

From (23),

$$\theta = \frac{\mathrm{d}v}{\mathrm{d}x} - \gamma \tag{27}$$

and the *curvature* (käyristymä) (This term is actually the curvature of the beam axis only in the Bernoulli theory; here it is a measure connected to the bending moment via (5.1.4).)

$$\begin{aligned}
-\frac{\mathrm{d}\theta}{\mathrm{d}x} &= -\frac{\mathrm{d}^2 v}{\mathrm{d}x^2} + 0 = -\frac{\mathrm{d}^2 v}{\mathrm{d}x^2} \\
&= -\frac{\mathrm{d}^2 H_1}{\mathrm{d}x^2} v_1 - \frac{\mathrm{d}^2 H_2}{\mathrm{d}x^2}\theta_1 - \frac{\mathrm{d}^2 H_3}{\mathrm{d}x^2} v_2 - \frac{\mathrm{d}^2 H_4}{\mathrm{d}x^2}\theta_2 - \frac{\mathrm{d}^2 (H_2 + H_4)}{\mathrm{d}x^2}\gamma \\
&= -\frac{1}{h^2}(-6+12\xi)v_1 - \frac{1}{h}(-4+6\xi)\theta_1 \\
&\quad -\frac{1}{h^2}(6-12\xi)v_2 - \frac{1}{h}(-2+6\xi)\theta_2 - \frac{1}{h}(-6+12\xi)\gamma
\end{aligned} \tag{28}$$

The curvature is seen to vary linearly along the beam. The values at the ends are

$$\begin{aligned}
\left(-\frac{\mathrm{d}\theta}{\mathrm{d}x}\right)_1 &= \frac{6}{h^2}v_1 + \frac{4}{h}\theta_1 - \frac{6}{h^2}v_2 + \frac{2}{h}\theta_2 + \frac{6}{h}\gamma \\
\left(-\frac{\mathrm{d}\theta}{\mathrm{d}x}\right)_2 &= -\frac{6}{h^2}v_1 - \frac{2}{h}\theta_1 + \frac{6}{h^2}v_2 - \frac{4}{h}\theta_2 - \frac{6}{h}\gamma
\end{aligned} \tag{29}$$

The bending moments at the ends are obtained thus from

$$\begin{aligned}
M_1 &= EI\left(\frac{6}{h^2}v_1 + \frac{4}{h}\theta_1 - \frac{6}{h^2}v_2 + \frac{2}{h}\theta_2\right) + 6h\varepsilon_h Q \\
M_2 &= EI\left(-\frac{6}{h^2}v_1 - \frac{2}{h}\theta_1 + \frac{6}{h^2}v_2 - \frac{4}{h}\theta_2\right) - 6h\varepsilon_h Q
\end{aligned} \tag{30}$$

Expression (26) and the notation (19) have been introduced. The bending moment varies linearly between these values. From (5.1.3b), the shearing force

$$Q = \frac{M_2 - M_1}{h} \tag{31}$$

In post-processing, we are given the nodal displacements and we want to know the three quantities $Q$, $M_1$, $M_2$. Solving from (30) and (31), we obtain



$$\begin{aligned}
\tilde{Q}_\mathrm{r} &= \frac{EI}{(1+12\varepsilon_h)h}\left(-\frac{12}{h^2}v_1 - \frac{6}{h}\theta_1 + \frac{12}{h^2}v_2 - \frac{6}{h}\theta_2\right) \\
\left(\tilde{M}_\mathrm{r}\right)_1 &= \frac{EI}{(1+12\varepsilon_h)}\left[\frac{6}{h^2}v_1 + \frac{4}{h}\theta_1 - \frac{6}{h^2}v_2 + \frac{2}{h}\theta_2 + \frac{12\varepsilon_h}{h}(\theta_1 - \theta_2)\right] \\
\left(\tilde{M}_\mathrm{r}\right)_2 &= \frac{EI}{(1+12\varepsilon_h)}\left[-\frac{6}{h^2}v_1 - \frac{2}{h}\theta_1 + \frac{6}{h^2}v_2 - \frac{4}{h}\theta_2 + \frac{12\varepsilon_h}{h}(\theta_1 - \theta_2)\right]
\end{aligned} \tag{32}$$

The notations have been now changed to indicate that these are approximate results. Subscript r refers to "refined". For slender elements the terms containing $\varepsilon_h$ become small. Bending moment inside the element is assumed to vary linearly between the nodes:

$$\tilde{M}_\mathrm{r} = N_1\left(\tilde{M}_\mathrm{r}\right)_1 + N_2\left(\tilde{M}_\mathrm{r}\right)_2 = (1-\xi)\left(\tilde{M}_\mathrm{r}\right)_1 + \xi\left(\tilde{M}_\mathrm{r}\right)_2 \tag{33}$$

The shearing force $\tilde{Q}_\mathrm{r}$ is constant in an element. It may be noted that (33) coincides with the consistent expresssion (20b) at the element midpoint.

**Example 5.3.** We repeat the analysis of the cantilever beam presented in Example 5.1 now using sensitized formulation.

From (17) and (18), $\tau_{v0} = \tau_{0v} = 0$ and

$$\begin{aligned}
\tau_{vv} &= -\frac{\varepsilon_h}{1+12\varepsilon_h}\frac{h^2}{kGA} \\
\tau_{0\theta} &= -\frac{1}{1+12\varepsilon_h}\frac{1}{kGA}
\end{aligned} \tag{a}$$

When these values are substituted in formulas (i) and (j) of Example 5.2, we obtain

$$[K]^{(0)} = -\frac{1}{1+12\varepsilon_h}\frac{kGA}{h}\begin{bmatrix} 1 & h/2 & -1 & h/2 \\ h/2 & h^2(1/3+\varepsilon_h) & -h/2 & h^2(1/6-\varepsilon_h) \\ -1 & -h/2 & 1 & -h/2 \\ h/2 & h^2(1/6-\varepsilon_h) & -h/2 & h^2(1/3+\varepsilon_h) \end{bmatrix} \tag{b}$$

and

$$\{b\}^{(0)} = \frac{\varepsilon_h}{1+12\varepsilon_h} q h^2 \begin{Bmatrix} 0 \\ 1 \\ 0 \\ -1 \end{Bmatrix} \tag{c}$$

It is interesting to note that the stiffness matrix (b) is just the original stiffness matrix (l) of Example 5.1 multiplied by the factor $-1/(1+12\varepsilon_h)$. Thus the final element stiffness matrix of the sensitized formulation is simply

$$[K]^s = [K] + [K]^{(0)}$$

$$= \frac{12\varepsilon_h}{1+12\varepsilon_h}\frac{kGA}{h}\begin{bmatrix} 1 & h/2 & -1 & h/2 \\ h/2 & h^2(1/3+\varepsilon_h) & -h/2 & h^2(1/6-\varepsilon_h) \\ -1 & -h/2 & 1 & -h/2 \\ h/2 & h^2(1/6-\varepsilon_h) & -h/2 & h^2(1/3+\varepsilon_h) \end{bmatrix} \quad \text{(d)}$$

The mesh and the data is the same as in Example 5.1. So $\varepsilon_h = 4/375$ and

$$\frac{12\varepsilon_h}{1+12\varepsilon_h} = \frac{16}{141} \quad \text{(e)}$$

Sensitizing is thus seen to change the element stiffness matrix here substantially. The sensitizing loading terms are

$$\begin{aligned} b_3^{(0)} &= \left(b_3^1\right)^{(0)} + \left(b_1^2\right)^{(0)} = 0+0=0 \\ b_4^{(0)} &= \left(b_4^1\right)^{(0)} + \left(b_2^2\right)^{(0)} = \frac{\varepsilon_h}{1+12\varepsilon_h} qh^2(-1+1) = 0 \\ b_5^{(0)} &= \left(b_3^2\right)^{(0)} = 0+0=0 \\ b_4^{(0)} &= \left(b_3^1\right)^{(0)} + \left(b_1^2\right)^{(0)} = \frac{\varepsilon_h}{1+12\varepsilon_h} qh^2(-1) = -\frac{4}{423} qh\cdot h \end{aligned} \quad \text{(f)}$$

Instead of set (s) of Example 5.1 we have now

$$\begin{aligned} &\frac{16}{141}\frac{kGA}{h}\left(2\cdot v_2 + 0\cdot h\theta_2 - 1\cdot v_3 + \frac{1}{2}h\theta_3\right) = qh \\ &\frac{16}{141}\frac{kGA}{h}h\left(0\cdot v_2 + \frac{258}{375}h\theta_2 - \frac{1}{2}v_3 + \frac{117}{750}h\theta_3\right) = 0 \\ &\frac{16}{141}\frac{kGA}{h}\left(1\cdot v_2 - \frac{1}{2}h\theta_2 + 1\cdot v_3 - \frac{1}{2}h\theta_3\right) = \frac{qh}{2} \\ &\frac{16}{141}\frac{kGA}{h}h\left(\frac{1}{2}v_2 + \frac{117}{750}h\theta_2 - \frac{1}{2}v_3 + \frac{129}{375}h\theta_3\right) = -\frac{4}{423}qh\cdot h \end{aligned} \quad \text{(g)}$$

The solution is



$$\begin{aligned} v_2 &= \frac{161009}{2256}\frac{qh^2}{kGA} = \frac{161009}{2256}\frac{q\varepsilon_h h^4}{EI} = 0.047579\cdot\frac{qL^4}{EI} \\ \theta_2 &= \frac{262375}{2256}\frac{qh}{kGA} = \frac{262375}{2256}\frac{q\varepsilon_h h^3}{EI} = 0.15507\cdot\frac{qL^3}{EI} \\ v_3 &= \frac{229381}{1128}\frac{qh^2}{kGA} = \frac{229381}{1128}\frac{q\varepsilon_h h^4}{EI} = 0.13557\cdot\frac{qL^4}{EI} \\ \theta_3 &= \frac{156625}{1128}\frac{qh}{kGA} = \frac{156625}{1128}\frac{q\varepsilon_h h^3}{EI} = 0.18514\cdot\frac{qL^3}{EI} \end{aligned} \quad \text{(h)}$$

The exact results and the results by the sensitized finite element method are shown in Figures (a) to (d) for the deflection, rotation, shearing force and bending moment, respectively.

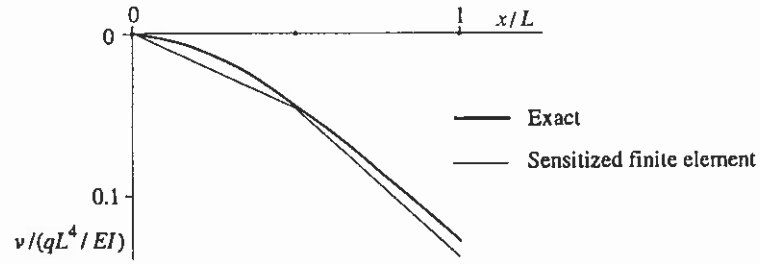



**Figure (a)**

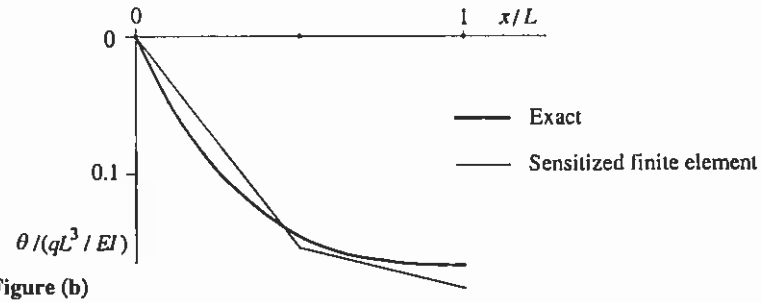



**Figure (b)**

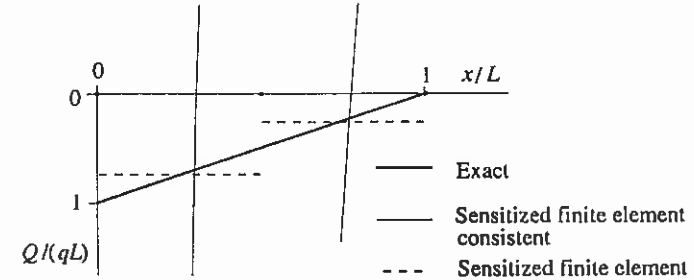



**Figure (c)**

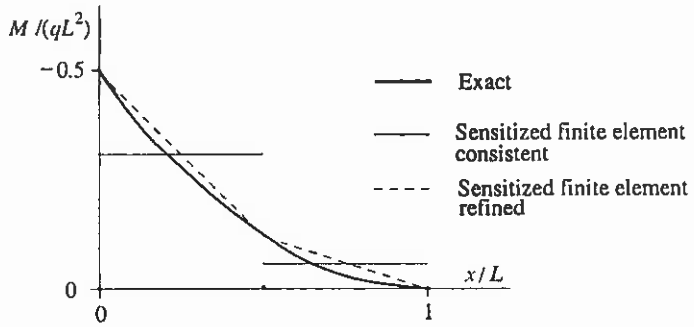



**Figure (d)**

Compared to the results by the standard version, a dramatic increase in accuracy is detected. Similarly, the refined stress resultant distributions are clearly to be preferred over the consistent ones. The displacement are not here nodally exact, but as the boundary conditions on the free end are natural, this result is according to Remark 5.11 not in contradiction with the theory.

**Remark 5.13.** The analogue of the Timoshenko model for plates is the so-called Reissner-Mindlin model. Sensitizing can be applied also there, Freund and Salonen (1998). It is found, however, that some de-stabilization (roughly, use of reduced integration for the shearing terms in the standard Galerkin method) must be first performed and only then sensitizing to have an accurate enough formulation. □

## 5.3 WEAK FORMS AND SENSITIZING

### 5.3.1 Explanation starting from the variational form

It seems that at the time Courant wrote the important articles about sensitizing the main emphasis was on variational formulations and weak formulations were hardly mentioned in engineering applications. Starting from the eighties specially Hughes and his associates have introduced numerous important contributions where weak forms have been "stabilized" by appending them with



additional terms, e.g., Hughes and Franca (1987), Hughes et al. (1989). It seems that these formulations have evolved gradually without any knowledge about the similar kind of formulation by Courant. Comparing with Courant, the main difference is of course that the starting point is more general as it is a weak form and not a variational principle but the approaches are similar in the respect that certain additional terms are appended to the standard expressions. Starting from the sensitized presentation of Courant, we can now rather easily shed some light on the emergence of these so-called stabilized terms in weak formulations.

Let us consider as an example again the sensitized functional (5.1.31). Demanding its stationarity gives the equation

$$\delta\Pi_s = \delta\Pi + \delta\frac{1}{2}\int_\Omega \{R\}^{\mathrm{T}}[\tau]^{(0)}\{R\}\mathrm{d}\Omega + \\ + \delta\frac{1}{2}\int_\Omega \frac{\mathrm{d}}{\mathrm{d}x}\{R\}^{\mathrm{T}}[\tau]^{(1)}\frac{\mathrm{d}}{\mathrm{d}x}\{R\}\mathrm{d}\Omega + \cdots = 0 \tag{1}$$

The terms involved are in more detail

$$\delta\Pi = \int_\Omega \left[ kGA\left(\frac{\mathrm{d}v}{\mathrm{d}x}-\theta\right)\left(\frac{\mathrm{d}\delta v}{\mathrm{d}x}-\delta\theta\right) + EI\frac{\mathrm{d}\theta}{\mathrm{d}x}\frac{\mathrm{d}\delta\theta}{\mathrm{d}x} - q\delta v\right]\mathrm{d}\Omega + \delta\,\mathrm{bt}$$

$$\delta\frac{1}{2}\int_\Omega \{R\}^{\mathrm{T}}[\tau]^{(0)}\{R\}\mathrm{d}\Omega = \int_\Omega \begin{Bmatrix} L_v(\delta v,\delta\theta) \\ L_\theta(\delta v,\delta\theta)\end{Bmatrix}^{\mathrm{T}}[\tau]^{(0)}\begin{Bmatrix} R_v(v,\theta) \\ R_\theta(v,\theta)\end{Bmatrix}\mathrm{d}\Omega \tag{2}$$

$$\delta\frac{1}{2}\int_\Omega \frac{\mathrm{d}}{\mathrm{d}x}\{R\}^{\mathrm{T}}[\tau]^{(1)}\frac{\mathrm{d}}{\mathrm{d}x}\{R\}\mathrm{d}\Omega \\ = \int_\Omega \frac{\mathrm{d}}{\mathrm{d}x}\begin{Bmatrix} L_v(\delta v,\delta\theta) \\ L_\theta(\delta v,\delta\theta)\end{Bmatrix}^{\mathrm{T}}[\tau]^{(1)}\frac{\mathrm{d}}{\mathrm{d}x}\begin{Bmatrix} R_v(v,\theta) \\ R_\theta(v,\theta)\end{Bmatrix}\mathrm{d}\Omega$$

The manipulations needed are explained already in connection with formulas (5.1.16) and (5.1.33). Making again the interpretations $\delta v = w_v$ and $\delta\theta = w_\theta$, where $w_v$ and $w_\theta$ are weighting functions, we can write a *sensitized weak form* (sensitoitu heikko muoto)

$$\boxed{F + F^{(0)} + F^{(1)} + \cdots = 0} \tag{3}$$

where

$$F \equiv \int_\Omega \left[\left(\frac{\mathrm{d}w_v}{\mathrm{d}x}-w_\theta\right)kGA\left(\frac{\mathrm{d}v}{\mathrm{d}x}-\theta\right) + \frac{\mathrm{d}w_\theta}{\mathrm{d}x}EI\frac{\mathrm{d}\theta}{\mathrm{d}x} - qw_v\right]\mathrm{d}\Omega + \mathrm{bt} \tag{4}$$

$$F^{(0)} \equiv \int_\Omega \begin{Bmatrix} L_v(w_v, w_\theta) \\ L_\theta(w_v, w_\theta) \end{Bmatrix}^{\mathrm{T}} [\tau]^{(0)} \begin{Bmatrix} R_v(v,\theta) \\ R_\theta(v,\theta) \end{Bmatrix} \mathrm{d}\Omega \tag{5}$$

$$F^{(1)} \equiv \int_\Omega \frac{\mathrm{d}}{\mathrm{d}x} \begin{Bmatrix} L_v(\delta v, \delta\theta) \\ L_\theta(\delta v, \delta\theta) \end{Bmatrix}^{\mathrm{T}} [\tau]^{(1)} \frac{\mathrm{d}}{\mathrm{d}x} \begin{Bmatrix} R_v(v,\theta) \\ R_\theta(v,\theta) \end{Bmatrix} \mathrm{d}\Omega \tag{6}$$

...

In (4) the notation bt means again some terms from the boundary differing from the use above. *Let us now forget the variational principle* and consider that a standard weak form $F = 0$ has been arrived at from the governing differential equations by multiplying them with the weighting functions, integrating over the domain, integrating by parts in the usual way etc., as explained in connection with formulas (5.1.20) - (5.1.23). A *least squares weak form* (pienimmän neliön heikko muoto) $F^{(0)} = 0$ is seen to be arrived at directly from the corresponding least squares functional. Similarly, the *gradient least squares weak form* (gradientti pienimmän neliön heikko muoto) $F^{(1)} = 0$ follows from the corresponding gradient least squares functional etc. The sensitized weak form (3) can thus be interpreted as a *linear combination of several weak forms*. It contains free parameters by which we can again try to steer the discrete solution in the direction we want using the patch test similarly as before.

The gradient least squares type appended terms in connection with weak forms have been presented for the first time in Franca and Dutra Do Carmo (1989).

Here the weak form $F = 0$ with $F$ according to (4) is the principle of virtual work applied to the Timoshenko beam problem and weak form (3) may be thus called as *sensitized principle of virtual work* (sensitoitu virtuaalisen työn pariaate). It is obvious how a sensitized principle of virtual work can be generated for other structural problems.

### 5.3.2 Concluding comments

At a quick glance the sensitizing terms used above seem to have a serious defect with respect to the low continuity of approximation. For instance, the residual (5.2.1a) contains the second derivative term $\mathrm{d}^2v/\mathrm{d}x^2$. According to the continuity condition of Section 4.1 this would demand a $C^1$ continuous approximation for convergence. However, it must be remembered that when studying convergence, the element sizes tend by definition towards zero. Thus roughly speaking, *any kind of extra terms producing beneficial behavior (or not) can be used if these terms vanish fast enough when the element size gets to zero*. This last behavior is found to be normally the case. We need no



sensitizing in the theoretical limit never reached in practice. Let us as an example consider the sensitizing parameter expressions (5.2.18). Written as

$$\tau_{vv} = -\frac{EI/kGA}{kGA + \dfrac{12EI}{kGAh^2}}, \qquad \tau_{\theta\theta} = -\frac{1}{kGA + \dfrac{12EI}{kGAh^2}} \tag{7}$$

we see that for a given $kGA$ and $EI$ the denominators in (7) get larger and larger with a diminishing $h$ and the expressions itself thus tend clearly to zero. □

The sensitizing terms are written in the literature usually in the form

$$\sum_e \int_{\Omega^e} \{L\}^{\mathrm{T}} [\tau]^{(0)} \{R\} \mathrm{d}\Omega \tag{8}$$

etc., to emphasize that they are to be evaluated *only over the element interiors*. This kind of representation is found sometimes to be used to motivate why we can violate the conventional continuity requirements in the approximation. To be honest, the elementwise representation is finally used also for the standard part of the weak form and the explanation seems not to be watertight. But the logic of the next paragraph explains how we can well violate the conventional continuity rules with respect to the sensitizing terms. As indicated already in Remark 5.8, the parameter values depend on the mesh and an elementwise representation like (8) reminds us about that. We will, however, usually avoid the elementwise form for simplicity of presentation.

We arrived above at a sensitized weak formulation via a variational principle mainly to see the connections with the older literature on the theme. As weak formulations are more general than variational formulations, we can now broaden the possibilities. In a sensitized weak form the sensitizing terms are residuals of the field equations (or their derivatives) multiplied by some factors and added to the standard weak form expression. *If the exact solution is considered, just zeros are added.* The sensitized formulation remains thus *consistent* (konsistenttti formulaatio). By this concept is meant in the finite element method that the exact solution satisfies the weak form. This property is usually considered important. Obviously we can modify the sensitizing terms now in many ways. First, the modeling possibilities may increase if we let the sensitizing parameter matrices be non-symmetric. Second, the multiplying factors $L$ may be changed, say to simplify the resulting expressions, and the formulation is still consistent. Third, based on the previous paragraph, we can further simplify also the residual expressions — as has been done already in Section 5.2.1 — and violate somewhat consistency because the sensitizing terms in any case vanish when the element size gets to zero.

In some situations the sensitizing terms can be given a transparent physical interpretation. This is found to be the case especially in connection with the diffusion-convection problem dealt with in Chapter 6. In general, however, no deep interpretations are necessary. We may be simply satisfied by the fact that sensitizing brings into a formulation additional quantities, sensitizing parameters — or so-called *tuning parameters* (viritysparametri) — which give more freedom for the discrete solution to simulate a problem.

The long neglected sensitizing idea of Courant is due to its simplicity conceptually very appealing. In a way one could say that the idea takes the best of two worlds. The Galerkin method (or the variational method if available) needs low continuity for the approximation but does not always work well with reasonable meshes. The least squares method leads to nice symmetric system equations but is awkward to apply as such due to the high demands on continuity. By combining the two methods free parameters emerge for optimizing the discrete solution and the least squares terms no more need high continuity. The idea is naturally not restricted just to the finite element method. However, combined there with the logic of determining the sensitizing parameter values by a special type of patch test it seems to offer a powerful tool to further enhance the applicability of the finite element method.

## REFERENCES


Courant, R. (1923). *Über ein konvergenzerzeugendes Princip in der Variatiionsrechnung*, Nachrichten von der Könichlichen Gesellschaft der Wissenscahten zu Göttingen, Matematisch-Physikalische Klasse aus dem Jahre 1922. 144 - 150.

Courant, R. (1943). Variational Methods for the Solution of Problems of Equilibrium and Vibrations, *Bull. Amer. Math. Soc.*, Vol. 49, 1 - 23.

Dym, C. L. and Shames, I. H. (1973). *Solid Mechanics: A Variational Approach*, McGraw-Hill, Tokyo, ISBN 0-07-018556-5.

Franca, L. P. and Dutra Do Carmo, E. D. (1989). The Galerkin gradient least-squares method, *Comput. Methods Appl. Mech. Engrg.* 74, 41-54.

Freund, J. and Salonen, E.-M. (1998). Sensitizing the Timoshenko beam and the Reissner-Mindlin plate finite element solution, Report no. 50, *Laboratory of Theoretical and Applied Mechanics, Helsinki University of Technology*.

Freund, J. and Salonen, E.-M. (2000). Sensitizing according to Courant the Timoshenko beam finite element solution, accepted for publication in *International Journal for Numerical Methods in Engineering*.

Hughes, T. J. R. and Franca, L. P. (1987). A new finite element formulation for computational fluid dynamics: VII. The Stokes problem with various well-posed boundary conditions: symmetric formulations that converge for all velocity/pressure spaces, *Comput. Methods Appl. Mech. Engrg.* 65, 85-96.

Hughes, T. J. R., Franca, L. P. and Hulbert, G. M. (1989). A new finite element formulation for computational fluid dynamics: VIII. The Galerkin/least-squares method for advective-diffusive equations, *Comput. Methods Appl. Mech. Engrg.* 73, 173-189.

Washizu, K. (1975). *Variational Methods in Elasticity and Plasticity*, 2nd ed. Pergamon Press, Oxford.





Zienkiewicz, O. C. and Taylor, R. L. (2000). *The Finite Element Method*, 5th ed., Butterworth-Heineman, Oxford. Vol. 1: *The Basis*, ISBN 0 7506 5049 4.


## PROBLEMS

# 6 DIFFUSION-CONVECTION

## 6.1 INTRODUCTION

In this chapter the effect of convection is considered. The numerical treatment of convection has been a major problem and this text concentrates strongly on it. The standard Galerkin method does not work well in convection dominated cases. Wildly oscillating solutions (in space) are obtained with reasonable meshes. This behavior can be remedied by sensitizing.

### 6.1.1 Energy equation completed

Before dealing with the diffusion-convection problem we introduce a rather general form of the energy equation. This far it has consisted of the field equation (3.1.13):

$$\frac{\partial q_\alpha}{\partial x_\alpha} - s = 0 \tag{1}$$

and of the Fourier law (3.1.50):

$$q_\alpha = -k_{\alpha\beta}\frac{\partial T}{\partial x_\beta} \tag{2}$$

We employ here and from now on mainly the index notation in Cartesian coordinates and summation convention similarly as in Appendix A. Instead of conventional typical Latin indices such as $i, j$, we have used here Greek symbols such as $\alpha, \beta$ and reserved the former for finite element shape function and nodal parameter indexing.

Equation (1) is valid in the *steady* case in a continuum at *rest*. The general local energy equation following from the principle of balance of energy is, Malvern (1969), Ziegler (1983),

$$\boxed{\nabla \bullet \mathbf{q} + \rho \dot{e} - \boldsymbol{\sigma} : \mathbf{d} - s = 0} \tag{3a}$$

or

$$\frac{\partial q_\alpha}{\partial x_\alpha} + \rho \dot{e} - \sigma_{\alpha\beta} d_{\alpha\beta} - s = 0 \tag{3b}$$

Here $\rho$ is the *density* (tiheys) ($[\rho] = \mathrm{kg/m^3}$) of the continuum, $e$ the *specific internal energy* (ominaissisäenergia) ($[e] = \mathrm{J/kg}$), $\boldsymbol{\sigma}$ the *stress tensor* (jännitys-





tensori) ($[\mathbf{s}] = \mathrm{N/m^2}$), $\mathbf{d}$ the *deformation rate tensor* (deformaationopeustensori) ($[\mathbf{d}] = 1/\mathrm{s}$) and $\mathbf{q}$ ($[\mathbf{q}] = \mathrm{W/m^2}$) and $s$ ($[s] = \mathrm{W/m^3}$) have the same meaning as before.

There are two main ways to describe a continuous medium in motion; the *Lagrangian* and *Eulerian description* (Lagrangen ja Eulerin esitystapa), Malvern (1969, p.138). The former is usually employed for solids and the latter for fluids. As our applications here with moving bodies will be only for fluids, we will use the *Eulerian description*. (In the applications for heat transfer in Chapter 2 and 3 the motion of the continuum was assumed to be negligible. Therefore the description used there can be considered equally well to be either Lagrangian or Eulerian.)

The deformation rate is connected to the *velocity* (nopeus) $\mathbf{v}$ ($[\mathbf{v}] = \mathrm{m/s}$) of the medium by the kinematic relation

$$d_{\alpha\beta} = \frac{1}{2}\left(\frac{\partial v_\alpha}{\partial x_\beta} + \frac{\partial v_\beta}{\partial x_\alpha}\right) \tag{4}$$

The quantity

$$d_{\alpha\alpha} = \nabla \bullet \mathbf{v} = \frac{\partial v_\alpha}{\partial x_\alpha} \tag{5}$$

is called *dilatation rate* (dilataationopeus) and it describes the relative volume time rate of a continuum element.

One local form of the principle of the conservation of mass is

$$\boxed{\dot{\rho} + \rho \nabla \bullet \mathbf{v} = 0} \qquad \dot{\rho} + \rho \frac{\partial v_\alpha}{\partial x_\alpha} = 0 \tag{6}$$

This is often called the *continuity equation* (jatkuvuusyhtälö).

With (5) and (6), the dilatation rate can be expressed also as

$$d_{\alpha\alpha} = -\frac{\dot{\rho}}{\rho} \tag{7}$$

The relations above are generally valid. Depending on the specific further assumptions, a confusing number of versions of the energy equation exists. We now make some constitutive assumptions to proceed to a certain version.

First, the medium is assumed to be a *Newtonian fluid* (Newtonin fluidi), Ziegler (1983, p. 78):

$$\sigma_{\alpha\beta} = -p\delta_{\alpha\beta} + 2\mu d_{\alpha\beta} + \lambda d_{\gamma\gamma}\delta_{\alpha\beta} \tag{8}$$

Here $p$ is the *pressure* (paine) ($[p] = \mathrm{N/m^2}$), $\mu$ the *viscosity* (viskositeetti) of the fluid ($[\mu] = \mathrm{Ns/m^2}$) and $\lambda = -2\mu/3$.

The so-called stress power $\sigma_{\alpha\beta} d_{\alpha\beta}$ in (3b) obtains the form

$$\begin{aligned}\sigma_{\alpha\beta} d_{\alpha\beta} &= -p\delta_{\alpha\beta} d_{\alpha\beta} + 2\mu d_{\alpha\beta} d_{\alpha\beta} + \lambda d_{\gamma\gamma}\delta_{\alpha\beta} d_{\alpha\beta} \\ &= -p d_{\alpha\alpha} + 2\mu d_{\alpha\beta} d_{\alpha\beta} + \lambda d_{\alpha\alpha} d_{\beta\beta} \\ &= -p d_{\alpha\alpha} + \Phi\end{aligned} \tag{9}$$

where

$$\Phi = 2\mu d_{\alpha\beta} d_{\alpha\beta} + \lambda d_{\alpha\alpha} d_{\beta\beta} \tag{10}$$

is called the *dissipation function* (dissipaatiofunktio).

Second, we assume a *mechanically incompressible fluid* (mekaanisesti kokoonpuristumaton fluidi):

$$\mathrm{d}\rho = -\gamma_p \rho \mathrm{d}T \tag{11}$$

The coefficient $\gamma_p$ ($[\gamma_p] = 1/\mathrm{K}$) is called *isobaric cubic expansion coefficient* (isobaarinen tilavuuden lämpötilakerroin).

**Remark 6.1.** The term "mechanically incompressible fluid" is not in wide use. If we consider a general constitutive relationship $\rho = \rho(p, T)$ and differentiate it, we get

$$\mathrm{d}\rho = \frac{\partial \rho}{\partial p}\mathrm{d}p + \frac{\partial \rho}{\partial T}\mathrm{d}T \tag{12}$$

or using standard notation

$$\frac{\mathrm{d}\rho}{\rho} = \kappa_T \mathrm{d}p - \gamma_p \mathrm{d}T \tag{13}$$

Coefficient $\kappa_T$ ($[\kappa_T] = 1/\mathrm{Pa}$) is called *isothermal compressibility* (isoterminen kokoonpuristuvuus). With the mechanically incompressible fluid assumption we effectively set $\kappa_T = 0$, that is, we assume that no volume changes follow from pressure changes. This means that the pressure is a (generalized) constraint force and it has no constitutive relation. The fluid can, however, respond to temperature changes by volume changes. For instance the



so-called *natural* or *free convection* (luonnollinen eli vapaa konvektio) can still be realistically modelled but not for instance pressure wave phenomena.

The differential form (11) is often replaced by an approximate finite form

$$\rho - \rho^\circ = -\gamma_p \rho^\circ (T - T^\circ) \tag{14}$$

or

$$\rho = \rho^\circ + \gamma_p \rho^\circ T^\circ - \gamma_p \rho^\circ T \tag{15}$$

where $\rho^\circ$ and $T^\circ$ refer to a certain reference state of the fluid. □

Third, it can be shown that the differential of the specific internal energy for a mechanically incompressible fluid has the form

$$\mathrm{d}e = c_p \mathrm{d}T + \frac{p}{\rho^2}\mathrm{d}\rho \tag{16}$$

where $c_p$ ($[c_p] = \mathrm{J/(kg\,K)}$) is the *specific heat capacity at constant pressure* (ominaislämpökapasiteetti vakiopaineessa). Division of (16) by the time differential $\mathrm{d}t$ and taking result (7) into account gives

$$\dot{e} = c_p \dot{T} + \frac{p}{\rho^2}\dot{\rho} = c_p \dot{T} - \frac{p}{\rho} d_{\alpha\alpha} \tag{17}$$

and

$$\rho \dot{e} = \rho c_p \dot{T} - p d_{\alpha\alpha} \tag{18}$$

Substitution of expressions (9) and (18) into (3b) gives a specialized energy equation

$$\frac{\partial q_\alpha}{\partial x_\alpha} + \rho c_p \dot{T} - \Phi - s = 0 \tag{19}$$

We have used the notation $(\cdot)^{\cdot}$ for the *material* (*time*) *derivative, substantial* (*time*) *derivative* (aineellinen aikaderivaatta, ainederivaatta, substantiaalinen derivaatta). In the Eulerian representation

$$\boxed{\dot{f}(\mathbf{x}, t) \equiv \frac{\mathrm{D}f}{\mathrm{D}t} = \frac{\partial f}{\partial t} + \mathbf{v} \bullet \nabla f} = \frac{\partial f}{\partial t} + v_\alpha \frac{\partial f}{\partial x_\alpha} \tag{20}$$

Using this notation and employing the Fourier constitutive law (2) gives the energy equation

$$\boxed{\rho c_p \frac{\partial T}{\partial t} + \frac{\partial}{\partial x_\alpha}(-k_{\alpha\beta}\frac{\partial T}{\partial x_\beta}) + \rho c_p v_\alpha \frac{\partial T}{\partial x_\alpha} - s - \Phi = 0} \tag{21}$$

If the term $\rho c_p$ is assumed to be constant in space we can express it also as

$$\frac{\partial T}{\partial t} + \frac{\partial}{\partial x_\alpha}\left(-\frac{k_{\alpha\beta}}{\rho c_p}\frac{\partial T}{\partial x_\beta}\right) + v_\alpha \frac{\partial T}{\partial x_\alpha} - \frac{s+\Phi}{\rho c_p} = 0 \tag{22}$$

**Remark 6.2.** The heat flux vector **q** consists in the general case addition to heat conduction also of a contribution due to thermal radiation. In the flow of mixtures some additional terms further emerge. Here we assume that these additional effects have been buried in the source term. It should be added that with fluids the conductivity tensor is normally assumed to be isotropic ($k_{\alpha\beta} = k\delta_{\alpha\beta}$), so that the diffusion term in (21) is in fact

$$\frac{\partial}{\partial x_\alpha}\left(-k\frac{\partial T}{\partial x_\alpha}\right) \tag{23}$$ □

**Remark 6.3.** The energy equation contains the velocity field explicitly in the convection term and implicitly in the dissipation function. At this phase we assume that the *velocity field is given* so that the only unknown is the temperature. (See also remark A.4.) In reality the velocity field can be determined in some cases with sufficient accuracy uncoupled from the energy equation (say in forced convection) but in many cases not (say in free convection). □

### 6.1.2 General D-C-R model problem

As the energy equation in the previous section has a somewhat complicated form we will present the basic mathematical properties from now on using the notation introduced in Appendix A for the general diffusion-convection-reaction equation. When applying the results for the energy equation, it is then easy to make the notational changes necessary.

The general model problem consists of the field equation

$$\boxed{\frac{\partial \phi}{\partial t} + \frac{\partial j_\alpha^{\mathrm{d}}}{\partial x_\alpha} + \frac{\partial}{\partial x_\alpha}(v_\alpha \phi) + c\phi - f = 0} \quad \text{in } \Omega^t \tag{24}$$

where the diffusion flux vector

$$j_\alpha^{\mathrm{d}} = -D_{\alpha\beta}\frac{\partial \phi}{\partial x_\beta} \tag{25}$$



of the boundary conditions

$$\boxed{\begin{array}{ll} \phi = \bar{\phi} & \text{on } \Gamma_{\mathrm{D}}^t \\ j^{\mathrm{d}} = \bar{j}^{\mathrm{d}} & \text{on } \Gamma_{\mathrm{N}}^t \\ j^{\mathrm{d}} = a\phi + b & \text{on } \Gamma_{\mathrm{R}}^t \end{array}} \tag{26}$$

where the diffusion flux density

$$j^{\mathrm{d}} \equiv n_\alpha j_\alpha^{\mathrm{d}} = -n_\alpha D_{\alpha\beta}\frac{\partial \phi}{\partial x_\beta} \tag{27}$$

and of the initial condition

$$\boxed{\phi(\mathbf{x},t) = \bar{\phi}_0(\mathbf{x})} \quad \text{in } \Omega \text{ at } t = 0 \tag{28}$$

The superscript $t$ in $\Omega$ and $\Gamma$ refers to the fact that the domains in question are in space and time. These notations are explained in more detail in Chapter 9. Similarly as in Chapter 3 we have not yet introduced the constitutive relation (25) for the diffusion flux vector into the governing equations to keep them as basic as possible. The above equations have been explained in detail in Appendix A.

## 6.2 ONE DIMENSION

### 6.2.1 Standard Galerkin method

A one-dimensional steady diffusion-convection problem is described by the *diffusion-convection equation* (later D-C equation)

$$R(\phi)\equiv \boxed{\frac{\mathrm{d}}{\mathrm{d}x}\left(-D\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)+\frac{\mathrm{d}}{\mathrm{d}x}(u\phi)-f=0} \quad \text{in } \Omega=]a,b[ \tag{1}$$

and for example by the boundary conditions

$$\boxed{\phi=\bar{\phi}} \qquad \text{on } \Gamma_{\mathrm{D}}=\{a\} \tag{2}$$

$$\boxed{-D\frac{\mathrm{d}\phi}{\mathrm{d}x}=\bar{j}^{\mathrm{d}}} \qquad \text{on } \Gamma_{\mathrm{N}}=\{b\} \tag{3}$$

This is a special case of the formulation in Section 6.1.2. We have employed the same type of example model problem as in Section 2.1.1 just with new notation and extended by the convection term. Quantity $u$ is the given convection velocity positive when directed into the positive $x$-axis direction.

The standard weak form corresponding to (1), (2) and (3) is

$$\int_\Omega \frac{\mathrm{d}w}{\mathrm{d}x}D\frac{\mathrm{d}\phi}{\mathrm{d}x}\mathrm{d}\Omega+\int_\Omega w\frac{\mathrm{d}}{\mathrm{d}x}(u\phi)\mathrm{d}\Omega-\int_\Omega wf\,\mathrm{d}\Omega+w\bar{j}^{\mathrm{d}}\Big|_{\Gamma_{\mathrm{N}}}=0 \tag{4}$$

It is obtained the way explained in Remark 2.5. The only difference with respect to (2.1.28) in addition to notation is the convection term.

**Remark 6.4.** In deriving (4), the term $w\mathrm{d}(-D\mathrm{d}\phi/\mathrm{d}x)/\mathrm{d}x$ from the field equation due to diffusion has been integrated by parts similarly as before to lower the order of the derivative on $\phi$. One can do the same for the convection term:

$$\int_\Omega w\frac{\mathrm{d}}{\mathrm{d}x}(u\phi)\mathrm{d}\Omega=-\int_\Omega \frac{\mathrm{d}w}{\mathrm{d}x}u\phi\,\mathrm{d}\Omega+\Big|_a^b wu\phi \tag{5}$$

This manipulation, however, helps little here. (For a non-constant $u$ some simplification could be claimed as no derivative acts on $u$ on the right-hand side of (5).) If $u$ is a $C^0$ function in (see Section B.1), as is assumed here, the finite element system equations are found not to change. This is because with the conventional continuous finite element approximation even $\tilde{w}$ and $\tilde{\phi}$ are $C^0$ functions and equation (5) is exactly valid also when $w$ and $\phi$ are replaced with $\tilde{w}$ and $\tilde{\phi}$. The convection term appears in the differential equation often alternatively as



$u\,\mathrm{d}\phi/\mathrm{d}x$ and thus in the weak form integral as $wu\,\mathrm{d}\phi/\mathrm{d}x$ but even here no simplification is achieved with integration by parts. □

Taking the finite element approximation

$$\tilde{\phi}(x)=\sum_j N_j(x)\phi_j \tag{6}$$

and applying the Galerkin method in (4) similarly as in Chapter 2 gives the system equations

$$[K]\{a\}=\{b\} \tag{7}$$

with

$$\begin{aligned} K_{ij}&=\int_\Omega \frac{\mathrm{d}N_i}{\mathrm{d}x}D\frac{\mathrm{d}N_j}{\mathrm{d}x}\mathrm{d}\Omega+\int_\Omega N_i\frac{\mathrm{d}}{\mathrm{d}x}(uN_j)\mathrm{d}\Omega \\ b_i&=\int_\Omega N_i f\,\mathrm{d}\Omega-N_i\bar{j}^{\mathrm{d}}\Big|_{\Gamma_{\mathrm{N}}} \end{aligned} \tag{8}$$

The convection term is now seen to make the system coefficient matrix non-symmetric.

The simple special case

$$-D\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2}+u\frac{\mathrm{d}\phi}{\mathrm{d}x}=0 \qquad \text{in } \Omega=]0,L[ \tag{9}$$

$$\phi(0)=0, \qquad \phi(L)=\bar{\phi} \tag{10}$$

is used below to explain certain solution behavior. This is a case with zero source term, constant diffusivity $D$, constant velocity $u$, and Dirichlet boundary conditions.

Equation (9) is a second order linear differential equation with constant coefficients and the exact solution is easy to find by standard procedures. There is obtained

$$\phi(x)=\frac{\exp(\mathrm{Pe}\,x/L)-1}{\exp(\mathrm{Pe})-1}\bar{\phi} \tag{11}$$

where

$$\mathrm{Pe} = \frac{u\,L}{D} \tag{12}$$

is a global Peclet number (see (A.2.8)). If Pe is small, diffusion dominates and the solution is nearly linear between the values determined by the boundary data (10). If $|\mathrm{Pe}|$ is large, convection dominates. If we take for instance the case $u > 0$, point $x = 0$ represents the inflow boundary. As described in Section A.3 and in fact directly found from (11), due to the condition $\phi(0) = 0$, the solution must then be nearly zero almost everywhere in the domain except at the neighborhood of the outflow boundary $x = L$ where a boundary layer is to be expected due to the condition $\phi(L) = \bar{\phi}$.

The weak form (4) simplifies to

$$\int_\Omega \left( \frac{\mathrm{d}w}{\mathrm{d}x} D \frac{\mathrm{d}\phi}{\mathrm{d}x} + wu \frac{\mathrm{d}\phi}{\mathrm{d}x} \right) \mathrm{d}\Omega = 0 \tag{13}$$

The discrete equations are obtained correspondingly from

$$\int_\Omega \left( \frac{\mathrm{d}\bar{w}}{\mathrm{d}x} D \frac{\mathrm{d}\bar{\phi}}{\mathrm{d}x} + \underline{\bar{w} u \frac{\mathrm{d}\bar{\phi}}{\mathrm{d}x}} \right) \mathrm{d}\Omega = 0 \tag{14}$$

With large convection the discrete diffusion term practically disappears compared with the underlined term due to convection. We would again like to obtain the nodally exact solution. We can now draw some conclusions without any actual calculations. Let us consider Figure 6.1. A uniform mesh of two-noded line elements (length $= h$) is used. The nodes and the elements are numbered from left to right. The exact solution is practically zero except for the thin right-hand side boundary layer, so the interpolant to the exact solution is essentially non-zero only in the last element (Figure (a)). Figure (b) shows the corresponding residual $u\,\mathrm{d}\bar{\phi}/\mathrm{d}x$. Figure (c) shows the weighting function $\bar{w} = N_{n-1}$ used to generate the system equation corresponding to node $n-1$. These two terms are positive. Thus multiplying them and performing the integration gives a positive left-hand side in (14) and *the equation cannot be satisfied for the assumed interpolant solution*. The Galerkin method must have negative residual in the second from right element to satisfy the discrete equation. What happens is shown in the figure. The Galerkin solution alternates between the values $\approx 0$ and $\approx \bar{\phi}$ at the nodes. This is in the case where the number of elements is odd. When the number of elements is even, a detailed study shows that every second of the nodal values tend to minus infinity as the Peclet number grows without limit and the results are thus still more unsatisfactory than in Figure (a).



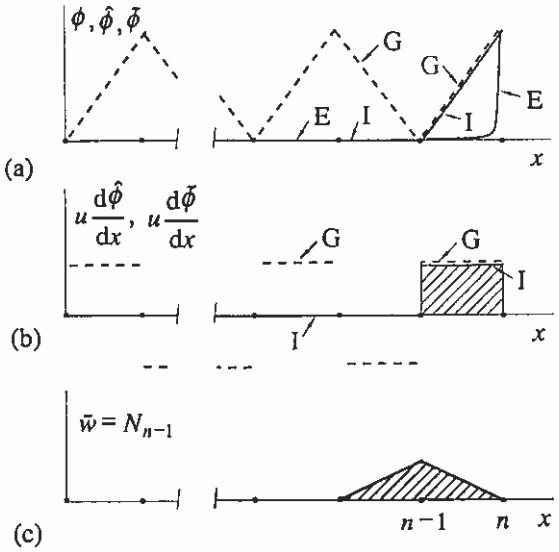



**Figure 6.1** Convection dominated case, **(a)** Exact solution $\phi$ ($\hat{=}$ E), interpolant to the exact solution $\hat{\phi}$ ($\hat{=}$ I), Galerkin solution $\bar{\phi}$ ($\hat{=}$ G). **(b)** Residual for the interpolant and for the Galerkin solution. **(c)** Shape function $N_{n-1}$.

We now look in more detail at the typical discrete equation for the uniform mesh. Using the notations of Figure 6.2, the system equation for node $i$ is

$$K_{i,i-1}\phi_{i-1} + K_{ii}\phi_i + K_{i,i+1}\phi_{i+1} = b_i \tag{15}$$

with

$$\begin{aligned} &K_{i,i-1} = K_{21}^{i-1}, \quad K_{ii} = K_{22}^{i-1} + K_{11}^{i}, \quad K_{i,i+1} = K_{12}^{i} \\ &b_i = b_2^{i-1} + b_1^{i} \end{aligned} \tag{16}$$

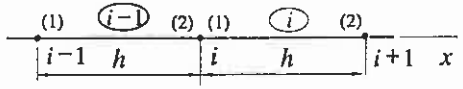



**Figure 6.2** Part of a uniform mesh.

The element contributions are (see formulas (F.1.1) or Example 4.1)

$$[K]^e = \frac{D}{h}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \frac{u}{2}\begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}$$
$$\{b\}^e = \int_{\Omega^e} \begin{Bmatrix} N_1^e f \\ N_2^e f \end{Bmatrix} \mathrm{d}\Omega \tag{17}$$

Here $f = 0$ so we do not need the latter expressions. We obtain

$$K_{i,i-1} = -\frac{D}{h} - \frac{u}{2}$$
$$K_{ii} = \frac{D}{h} + \frac{u}{2} + \frac{D}{h} - \frac{u}{2} = \frac{2D}{h} \tag{18}$$
$$K_{i,i+1} = -\frac{D}{h} + \frac{u}{2}$$

and the system equation is

$$\frac{D}{h}(-\phi_{i-1} + 2\phi_i - \phi_{i+1}) + \frac{u}{2}(-\phi_{i-1} + \phi_{i+1}) = 0 \tag{19}$$

Alternatively, if the finite difference method is applied directly to the differential equation (9) using well-known central difference formulas, the following typical discrete equation

$$-\frac{D}{h^2}(\phi_{i-1} - 2\phi_i + \phi_{i+1}) + \frac{u}{2h}(-\phi_{i-1} + \phi_{i+1}) = 0 \tag{20}$$

is obtained. This is seen to be equivalent to (19).

Equation (19) or (20) can be considered as a difference equation with constant coefficients and it can be solved in closed form for any number of nodes. A study of the solution shows that it starts to give unphysical "wiggles" when the value of $|\mathrm{Pe}_h|$ exceeds 2. Here $\mathrm{Pe}_h$ is a local elementwise Peclet number defined as

$$\mathrm{Pe}_h = \frac{uh}{D} \tag{21}$$

In Figure 6.3 the solution for a uniform mesh of five elements by the Galerkin method is compared with the interpolant to the exact solution for two values of the Peclet number. The Galerkin method solution can be obtained either by solving directly the finite element system equations or by employing the finite



difference closed form solution. The solution for $\mathrm{Pe}_h = 2.5$ is found to be already quite useless due to the oscillations.

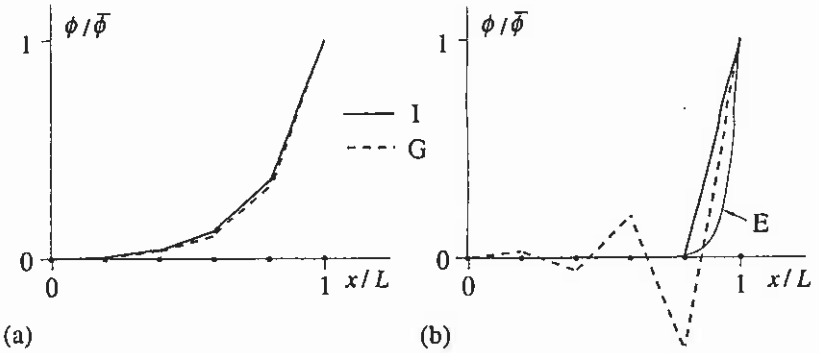



**Figure 6.3** The solution **(a)** $\mathrm{Pe} = 5$, $\mathrm{Pe}_h = 1$. **(b)** $\mathrm{Pe} = 25$, $\mathrm{Pe}_h = 5$. Interpolant to the exact solution (I), Galerkin solution (G).

All what has been seen indicates that the Galerkin method would demand a very dense mesh for avoiding the wiggles. The wiggles are generated by the convection term. It is often said in a somewhat unscientific way that in problems with convection the *information* on the value of the dependent variable from the point of view of a fixed spatial point *is more important on the upwind side than on the downwind side.* This is discussed also in Section A.3. The concept of *upwinding* or *upwind scheme* or *upstream scheme* (ylävirtapainotus) is employed in many numerical methods to somehow take this feature into account. The central difference method and the Galerkin method clearly do not have any directional preferencies and thus they do not contain any upwinding.

In the model problem under study a simple upwinding method — which might be called *full upwinding* — is achieved with the finite difference method in the case $u > 0$ by replacing (20) with

$$-\frac{D}{h^2}(\phi_{i-1} - 2\phi_i + \phi_{i+1}) + \frac{u}{h}(-\phi_{i-1} + \phi_i) = 0 \tag{22}$$

The derivative $(\mathrm{d}\phi/\mathrm{d}x)_i$ is approximated thus by a unilateral backward difference formula; information is taken here only from the upwind side. If $u < 0$, the approximation is similarly

$$-\frac{D}{h^2}(\phi_{i-1} - 2\phi_i + \phi_{i+1}) + \frac{u}{h}(-\phi_i + \phi_{i+1}) = 0 \tag{23}$$

By studying the general solution of the difference equation (22) it is found that no wiggles appear any more for any $\mathrm{Pe}_h > 0$.

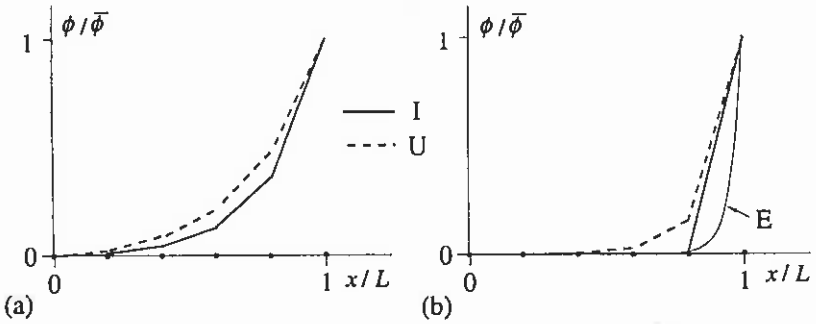



**Figure 6.4** The solution **(a)** $\mathrm{Pe}=5$, $\mathrm{Pe}_h=1$. **(b)** $\mathrm{Pe}=25$, $\mathrm{Pe}_h=5$. Interpolant to the exact solution (I), finite difference solution with full upwinding (U).

Figure 6.4 is the counterpart of Figure 6.3 but the numerical results are obtained from (22). It is seen that indeed even for a large value of the Peclet number no wiggles appear, however, the accuracy for a small value is worse than without upwinding.

Full upwinding can be given the following interpretation. Let us replace the true diffusivity in (9) with

$$D := D + D^* \tag{24}$$

where

$$D^* = \frac{uh}{2} \tag{25}$$

and let us generate the system equations with central differences or, which is the same, using the Galerkin method. We obtain using equation (20),

$$-\frac{1}{h^2}\left(D+\frac{uh}{2}\right)(\phi_{i-1}-2\phi_i+\phi_{i+1})+\frac{u}{2h}(-\phi_{i-1}+\phi_{i+1})=0 \tag{26}$$

or

$$-\frac{D}{h^2}(\phi_{i-1}-2\phi_i+\phi_{i+1})+\frac{u}{h}(-\phi_{i-1}+\phi_i)=0 \tag{27}$$

which is nothing but equation (22).



$D^*$ is called the coefficient of *artificial diffusion* or balancing diffusion (keinotekoinen diffuusio). This term sometimes also means the case where the use of a certain numerical method can be interpreted as a procedure where the real physical diffusivity of the problem is altered by some additional computational diffusion.

As is discussed in Section A.3, diffusion tends to smooth the solution and especially in the model problem (9) and (10) the solution for large diffusion is nearly a straight line determined by the boundary values. By studying Figures 6.3 and 6.4 one can now say that in fact the Galerkin method or the use of central differences means that the method is *underdiffuse* (alidiffuusi) and that on the other hand the use of full upwinding leads to an *overdiffuse* (ylidiffuusi) method. In the model problem it is presumably possible to select such an optimal value for $D^*$ that the Galerkin method gives exact values for $\phi$ at the nodes; in other words the finite element interpolant to the exact solution is achieved. The optimal value is, Brooks and Hughes (1982),

$$D^* = uh\left(\frac{1}{2}\coth\frac{\mathrm{Pe}_h}{2}-\frac{1}{\mathrm{Pe}_h}\right) \tag{28}$$

In the limit cases $\mathrm{Pe}_h = 0$ and $\mathrm{Pe}_h \to \infty$ the values $D^* = 0$ and $D^* = uh/2$, respectively, are obtained where the latter means full upwinding.

The difficulties associated with convection in the finite element method are based on the fact that non-zero convection means that the problem is *no more self-adjoint* (see Appendix D). Let us consider the weak form (4). The first two integrals on the left-hand side are together a bilinear form $a(w,\phi)$ in $w$ and $\phi$. (see Appendix C). Unfortunately, it is not a symmetric bilinear form — $a(w,\phi) \neq a(\phi,w)$ — because of the convection term. (Integration by parts of the convection term does not help the situation.) This means that we cannot generate here an inner product and an energy norm from the bilinear form as was done in Section 4.2.2. Similarly, such results as the best approximation property of the Galerkin method are lost. A reasonable engineering approach in such a situation is to strive for formulations giving accurate nodal values in the spirit of Remark 4.1. This means that the pure Galerkin method has to be abandoned and in C-D problems it means that some kind of upwinding must be included.

Upwinding was applied obviously for the first time in the finite difference method in Courant et al. (1952). The first application of it in the finite element method is in Christie et al. (1967). The idea was to simulate the procedures found to work in finite differences. One very simple way was to add suitable amount of artificial diffusion and then use the standard Galerkin method as was

explained in connection with equation (9). This procedure was found to produce, however, in two-, and three-dimensional problems an intolerably large amount of false diffusion (see Remark A.6). It was realized that the artificial diffusion must operate only in the flow direction and thus the corresponding diffusivity tensor must be anisotropic. With a non-zero source terms and in unsteady cases even this version did not give satisfactory results.

The final systematic breakthrough was the invention of the so-called *streamline upwind/Petrov-Galerkin method* (SUPG-method). Reference Brooks and Hughes (1982) gives an understandable explanation of the formulation. The main idea is to write the weighting function in the form

$$w^* = w + p \tag{29}$$

where $w$ is a continuous Galerkin method type weighting function (shape function) and $p$ a discontinuous perturbation weighting function. Function $p$ acts inside each element on the full field equation residual, no integration by parts is applied there on the diffusion term. Prescriptions on the selection of proper $p$ is given in the literature.

**Remark 6.5.** The SUPG-method could be translated in Finnish as "virtaviiva-ylävirta / Petrov-Galerkinin menetelmä". The terms in the name of the method refer to the following. The word "upwind" emphasizes that the weighting functions have from the point of view of a given point stronger weighting on the upwind than on the downwind side. The word "streamline" indicates that this directional weighting must operate expressly in the direction given by the streamline through the point in question. Finally, the term "Petrov-Galerkin" is nowadays often employed in cases where the trial functions and the weighting functions are from different function sets. □

A later generalization the SUPG-method is the so-called *Galerkin/least squares method* (GLS-method), Shakib and Hughes (1991). Even this formulation has been extended. As suitable names start easily to get out of hand, we will rely on the concepts introduced in Chapter 5 and mainly speak about *sensitized weak forms*. Describing in each case the structure of the sensitizing terms fixes the formulation.

### 6.2.2 Sensitized Galerkin method

The general ideas behind sensitizing have been explained in Chapter 5. We will apply them here first on the steady one-dimensional D-C equation

$$R(\phi) \equiv L(\phi) - f \equiv \boxed{\frac{\mathrm{d}}{\mathrm{d}x}\left(-D\frac{\mathrm{d}\phi}{\mathrm{d}x}\right) + \frac{\mathrm{d}}{\mathrm{d}x}(u\phi) - f = 0} \tag{30}$$

with some boundary conditions. We write a preliminary sensitized weak form



$$\int_\Omega \frac{\mathrm{d}w}{\mathrm{d}x} D \frac{\mathrm{d}\phi}{\mathrm{d}x}\mathrm{d}\Omega + \int_\Omega w \frac{\mathrm{d}}{\mathrm{d}x}(u\phi)\mathrm{d}\Omega - \int_\Omega w f \,\mathrm{d}\Omega + \mathrm{bt}$$
$$+ \int_\Omega L(w)\tau^{\mathrm{c}} R(\phi)\mathrm{d}\Omega = 0 \tag{31}$$

The first line contains the standard weak form left-hand side. Only the equation residual (no gradient terms) has been included in the sensitizing term. There is just one sensitizing parameter $\tau^{\mathrm{c}}$ to be determined.

**Remark 6.6.** As has been done already in Chapter 5, we often use the symbol bt in an expression to indicate that this term comes from the boundary of the domain under consideration. In this manner general relations can be discussed without complicating too much the formulas. With the specific boundary conditions used, the "bt-terms" obtain specific expressions, which can be easily deduced by considering the derivation of the weak form under question. It should be emphasized that appending sensitizing terms do not change the boundary conditions and not the bt-terms appearing in the standard forms. For instance, if boundary conditions (2) and (3) are used, we have in (31)

$$\mathrm{bt} = w\bar{j}^{\mathrm{d}}\Big|_{\Gamma_{\mathrm{N}}} \tag{32}$$ □

**Remark 6.7.** In this text we are going to use the sensitized formulations in connection with the simplest type of elements and similarly as in Chapter 5 *we assume constant values for the sensitizing parameters in an element*; the values can of course vary from element to element. For more complicated elements the values of the parameters should obviously vary also inside an element for optimal results. As the theory in this respect seems not to be quite fully developed, we do not treat this theme here. □

**Remark 6.8.** By expanding the derivatives in (30) we obtain

$$R = -D\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2} - \frac{\mathrm{d}D}{\mathrm{d}x}\frac{\mathrm{d}\phi}{\mathrm{d}x} + u\frac{\mathrm{d}\phi}{\mathrm{d}x} + \frac{\mathrm{d}u}{\mathrm{d}x}\phi - f \tag{33}$$

or

$$R = -D\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2} + \bar{u}\frac{\mathrm{d}\phi}{\mathrm{d}x} + \bar{c}\phi - f \tag{34}$$

where

$$\bar{u} = u - \frac{\mathrm{d}D}{\mathrm{d}x}, \qquad \bar{c} = \frac{\mathrm{d}u}{\mathrm{d}x} \tag{35}$$

Form (34) is a full D-C-R equation and this could be used for added accuracy. However, we again here and in the following always simplify and assume constant operator data in the differential operators in the sensitizing terms. Based on the comments in Section 5.3.2, no error is introduced with respect to convergence. The formulation is then not any more strictly consistent if $D$ or $u$ depend on position. If the error is considered too large we can always if necessary employ the more complicated expressions (and replace the data finally in practice with some representative values). However, as in two or three dimensions we usually cannot

in any case quite achieve the goal of exact nodal values, we should perhaps not exaggerate too much in the evaluation of the contributions. □

Following the remark above, *when sensitizing*, we replace (30) by (superscrift c is used just to indicate the difference with (30))

$$R^{c}(\phi) \equiv L^{c}(\phi) - f \equiv \boxed{-D\frac{d^2\phi}{dx^2} + u\frac{d\phi}{dx} - f = 0} \tag{36}$$

where $D$ and $u$ are some constant local representative values (we do not introduce new notation for this) and the final sensitized weak form becomes

$$\boxed{\begin{aligned}&\int_\Omega \frac{dw}{dx} D\frac{d\phi}{dx}d\Omega + \int_\Omega w\frac{d}{dx}(u\phi)d\Omega - \int_\Omega wf\,d\Omega + \text{bt} + \\ &+\int_\Omega L^{c}(w)\tau^{c}R^{c}(\phi)d\Omega = 0\end{aligned}} \tag{37}$$

or written in full

$$\begin{aligned}&\int_\Omega \frac{dw}{dx} D\frac{d\phi}{dx}d\Omega + \int_\Omega w\frac{d}{dx}(u\phi)d\Omega - \int_\Omega wf\,d\Omega + \text{bt} \\ &+\int_\Omega \left(-D\frac{d^2w}{dx^2} + u\underline{\frac{dw}{dx}}\right)\tau^{c}\left(-D\frac{d^2\phi}{dx^2} + u\underline{\frac{d\phi}{dx}} - f\right)d\Omega = 0\end{aligned} \tag{38}$$

Although we can consider sensitizing just as a mathematical device by which some tuning parameters (cf. Section 5.3.2) are introduced into a formulation, here the underlined terms give a physical interpretation for the beneficial behaviour as explained in Section D.4.1; the term

$$\frac{dw}{dx}\tau^{c}u^2\frac{d\phi}{dx} \tag{39}$$

can be interpreted as an additional diffusion term damping the oscillations. However, differing from the artificial diffusion concept the formulation here is now *consistent* (at least for constant operator data).

We next determine the reference solutions following Section 5.2.1. The governing simplified field equation according to (36) is

$$-D\frac{d^2\phi}{dx^2} + u\frac{d\phi}{dx} - f = 0 \tag{40}$$



This is a second order linear differential equation with constant coefficients. Its solution is of the well-known form

$$\phi(x) = Ae^{r_1 x} + Be^{r_2 x} + \phi_p(x) \tag{41}$$

where $r_1$ and $r_2$ are the roots

$$r_1 = 0, \qquad r_2 = \frac{u}{D} \tag{42}$$

of the characteristic equation

$$-Dr^2 + ur = 0 \tag{43}$$

and $\phi_p$ is a particular solution for the non-homogeneous equation. The source term is developed into a Taylor series

$$f = f_0 + (f_x)_0 x + \frac{1}{2}(f_{xx})_0 x^2 + \cdots \tag{44}$$

and the local origin of $x$ has been taken at the generic point under study. We obtain in detail

$$\phi(x) = A + Be^{ux/D} + f_0\frac{1}{u}x + (f_x)_0\left(\frac{D}{u^2}x + \frac{1}{2u}x^2\right) + \cdots \tag{45}$$

Using similar representation as in formula (5.2.4), we have thus the reference solution

$$\begin{aligned}\begin{Bmatrix}\phi \\ f\end{Bmatrix} &= A\begin{Bmatrix}1 \\ 0\end{Bmatrix} + B\begin{Bmatrix}e^{ux/D} \\ 0\end{Bmatrix} \\ &+ f_0\begin{Bmatrix}1/u\cdot x \\ 1\end{Bmatrix} + (f_x)_0\begin{Bmatrix}D/u^2\cdot x + 1/(2u)\cdot x^2 \\ x\end{Bmatrix} + \cdots\end{aligned} \tag{46}$$

The patch test is performed in Example 6.1 for the two-noded line element. It gives the optimal value

$$\boxed{\tau^{c} = \frac{h}{u}\left(\frac{1}{2}\coth\frac{\text{Pe}_h}{2} - \frac{1}{\text{Pe}_h}\right)} \tag{47}$$

where $\mathrm{Pe}_h$ is the elementwise Peclet number (21). The study performed in Example 6.1 shows that with this sensitizing parameter value, nodally exact results are obtained up to a linear source term if the mesh is uniform and the operator data is constant at least with essential boundary conditions. With variable data and mesh, $\tau^c$ is evaluated for each element from (47) using some representative values.

We define a dimensionless sensitizing parameter $\hat{\tau}^c$ by

$$\hat{\tau}^c \equiv \frac{\tau^c}{h/u} = \frac{1}{2}\coth\frac{\mathrm{Pe}_h}{2} - \frac{1}{\mathrm{Pe}_h} \tag{48}$$

Figure 6.5 shows the graph of this parameter.

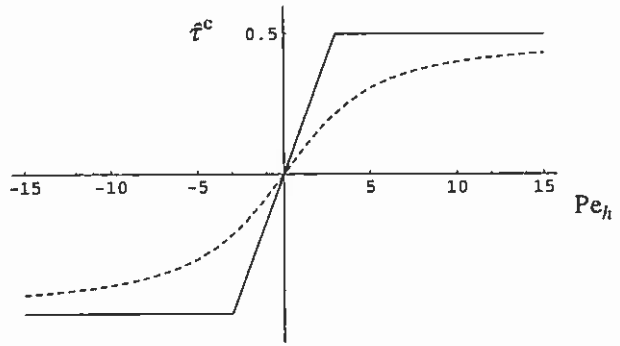



**Figure 6.5** Dimensionless sensitizing parameter $\hat{\tau}^c$ as a function of $\mathrm{Pe}_h$.

It maybe noted that for a given $u$ and $D$, the sensitizing parameter evaluated from (47) indeed approaches zero when the mesh size $h$ goes to zero as in addition $\hat{\tau}^c$ also approaches zero with vanishing $\mathrm{Pe}_h$. The diffusivity $\tau^c u^2$ in (39) obtains the forms

$$D^c \equiv \tau^c u^2 = \hat{\tau}^c u h = \hat{\tau}^c \mathrm{Pe}_h D \tag{49}$$

It is seen that this remains always (with non-zero $u$) positive. If $u$ is negative, so is also $\hat{\tau}^c$ (see Figure 6.5). Thus positive artificial diffusivity and damping is always introduced by sensitizing. (We will call $D^c$ from this on *damping diffusivity* and not artificial diffusivity to discern it from (28) in general as here the concept is based on a consistent formulation.)



The evaluation of $\hat{\tau}^c$ for a small value of $\mathrm{Pe}_h$ by computer is somewhat awkward as both $\coth(\mathrm{Pe}_h/2)$ and $1/\mathrm{Pe}_h$ become separately unbounded. For efficiency of calculations, a doubly asymptotic approximation, Brooks and Hughes (1982),

$$\hat{\tau}^c = \begin{cases} \mathrm{Pe}_h/12, & |\mathrm{Pe}_h| \le 6 \\ 1/2 \cdot \operatorname{sgn}\mathrm{Pe}_h, & |\mathrm{Pe}_h| > 6 \end{cases} \tag{50}$$

can be used. It is indicated by the dashed line in Figure 6.5. It is obtained from the tangent at the origin and from the tangents at $\pm$ infinity.

As a comment on the order of magnitude we may notice that using (50) we obtain corresponding say to the cases of $\mathrm{Pe}_h = 6$ and $\mathrm{Pe}_h = 12$ the damping diffusivities are $D^c = 3D$ and $D^c = 6D$ respectively. So the amount of damping needed in these cases is considerable.

**Remark 6.9.** Following the logic discussed in Section 5.3.2, the sensitized weak form (38) can be simplified so that the sensitizing integral is just

$$\int_\Omega u\frac{dw}{dx}\tau^c\left(-D\frac{d^2\phi}{dx^2} + u\frac{d\phi}{dx} - f\right)d\Omega \tag{51}$$

that is, the second derivative term has been dropped from the weighting. Then the formulation is no more purely of the least squares type but the important term (39) is still included. In fact, for two-noded elements the discrete equations become the same as with the least squares form as the second derivatives vanish in any case. □

Figure 6.6 is connected to formula (29). We see from (37) that when a typical original finite dimensional continuous weighting function $\tilde{w} = N_i$ is acting on the standard part of the weak form, the weighting in the sensitizing integral takes the form (for two-noded elements)

$$L^c(\tilde{w}) = u\frac{d\tilde{w}}{dx}\tau^c = \tau^c u\frac{d\tilde{w}}{dx} = \tau^c u\frac{dN_i}{dx} \tag{52}$$

The interpretation is that the original continuous weighting function $\tilde{w} = N_i$ has been amended to the form $\tilde{w}^* = \tilde{w} + p$; where $p$ is given by (52). Figures (c) to (f) have been drawn assuming full upwinding ($\tau^c u = h/2 \cdot \operatorname{sgn}\mathrm{Pe}_h$). The amended weighting function clearly has a directional preference and provides upwinding. This interpretation is different from the additional diffusion explanation described in Section D.4.1 and gives another point of view.

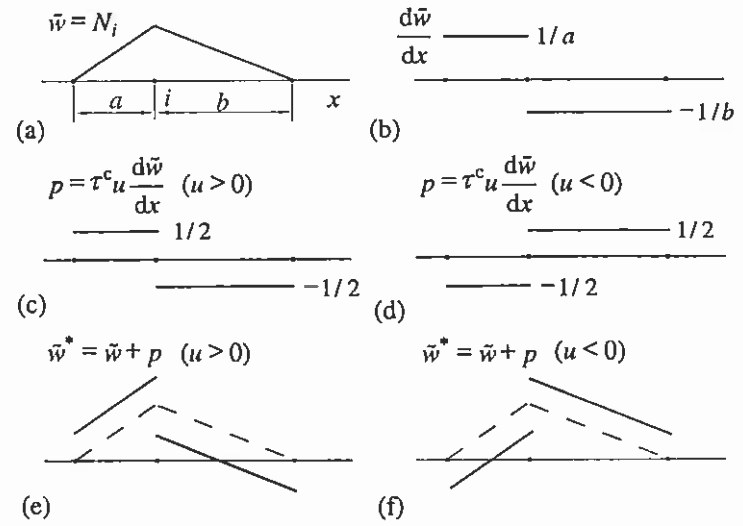



**Figure 6.6** **(a)** Continuous (Galerkin) weighting function. **(b)** Its derivative. **(c)** Discontinuous perturbation weighting function ($u>0$). **(d)** Discontinuous perturbation weighting function ($u<0$). **(e)** Total weighting function ($u>0$). **(f)** Total weighting function ($u<0$).

**Remark 6.10.** As mentioned already in Remark 6.5, the term *Petrov-Galerkin method* is sometimes used when the weighting functions are not taken from the set of trial basis functions. Similarly the Galerkin method is sometimes called the *Bubnov-Galerkin method* to emphasize the difference. But if we agree to call the symbol $w$ or the finite dimensional symbol $\tilde{w}$ here as the weighting function, we are in this text actually always using the Galerkin method or to make the terminology more specific we can speak about the *sensitized Galerkin method.* □

When the Galerkin method is applied in (38) (using two-noded elements which means that the second derivatives vanish both in the weighting and in the residual), we obtain the system equations

$$[K]\{a\} = \{b\} \tag{53}$$

with

$$K_{ij} = \int_\Omega \frac{dN_i}{dx} D \frac{dN_j}{dx} d\Omega + \int_\Omega N_i \frac{d}{dx}\left(uN_j\right) d\Omega + \int_\Omega \frac{dN_i}{dx} \tau^c u^2 \frac{dN_j}{dx} d\Omega + \text{bt} \tag{54}$$



$$b_i = \int_\Omega N_i f \, d\Omega + \int_\Omega \frac{dN_i}{dx} \tau^c u f \, d\Omega + \text{bt}$$

These can be compared with the corresponding expressions (8) without sensitizing. The two bt-terms in (54) have naturally different interpretations. (If the boundary conditions are for instance (2) and (3), the bt is the one appearing at the end of the last expression in (8). A possible Robin boundary condition gives a contribution bt for the coefficient matrix term.)

As $\tau^c$ is assumed to be elementwise constant (and $u$ in the sensitizing terms), the element contributions are

$$K_{ij}^e = \int_{\Omega^e} \frac{dN_i^e}{dx} D \frac{dN_j^e}{dx} d\Omega + \int_{\Omega^e} N_i^e \frac{d}{dx}\left(uN_j^e\right) d\Omega + \tau^c u^2 \int_{\Omega^e} \frac{dN_i^e}{dx} \frac{dN_j^e}{dx} d\Omega + \text{bt} \tag{55}$$

$$b_i^e = \int_{\Omega^e} N_i^e f \, d\Omega + \tau^c u \int_{\Omega^e} \frac{dN_i^e}{dx} f \, d\Omega + \text{bt}$$

It is again easy to see that the sensitizing is injecting diffusion into the system equations.

**Example 6.1.** We derive the formula for $\tau^c$ using the patch test as explained in Section 5.2.2. The notations for the patch (Figure (a) are the same as in Example 5.2.

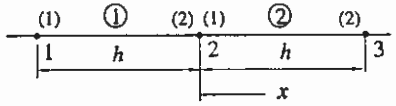



**Figure (a)**

The element contributions are according to (55) (constant data)

$$[K]^e = \frac{D}{h}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \frac{u}{2}\begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix} + \frac{\tau^c u^2}{h}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

$$\{b\}^e = \int_{\Omega^e} \begin{Bmatrix} N_1^e f \\ N_2^e f \end{Bmatrix} d\Omega + \tau^c u \int_{\Omega^e} \begin{Bmatrix} dN_1^e/dx \, f \\ dN_2^e/dx \, f \end{Bmatrix} d\Omega \tag{a}$$

In the system equation for node 2:

$$K_{21}\phi_1 + K_{22}\phi_2 + K_{23}\phi_3 - b_2 = 0 \tag{b}$$

$$K_{21} = K_{21}^1 = -\frac{D}{h} - \frac{u}{2} - \frac{\tau^c u^2}{h}$$

$$K_{22} = K_{22}^1 + K_{11}^2 = \frac{D}{h} + \frac{u}{2} + \frac{\tau^c u^2}{h} + \frac{D}{h} - \frac{u}{2} + \frac{\tau^c u^2}{h} = \frac{2D}{h} + \frac{2\tau^c u^2}{h} \tag{c}$$

$$K_{23} = K_{12}^2 = -\frac{D}{h} + \frac{u}{2} - \frac{\tau^c u^2}{h}$$

$$b_2 = b_2^1 + b_1^2 = \int_{\Omega^1} N_2^1 f \,\mathrm{d}\Omega + \int_{\Omega^2} N_1^2 f \,\mathrm{d}\Omega + \tau^c u \int_{\Omega^1} \frac{\mathrm{d}N_2^1}{\mathrm{d}x} f \,\mathrm{d}\Omega + \tau^c u \int_{\Omega^2} \frac{\mathrm{d}N_1^2}{\mathrm{d}x} f \,\mathrm{d}\Omega$$

Equation (b) is thus in detail

$$\left(-\frac{D}{h} - \frac{u}{2} - \frac{\tau^c u^2}{h}\right)\phi_1 + \left(\frac{2D}{h} + \frac{2\tau^c u^2}{h}\right)\phi_2 + \left(-\frac{D}{h} + \frac{u}{2} - \frac{\tau^c u^2}{h}\right)\phi_3 - b_2 = 0 \tag{d}$$

The first specific reference solution in (46) ( $A = 1$ ) gives the nodal values

$$\phi_1 = 1, \qquad \phi_2 = 1, \qquad \phi_3 = 1 \tag{e}$$

with zero source term. Equation (d) is

$$\left(-\frac{D}{h} - \frac{u}{2} - \frac{\tau^c u^2}{h}\right)1 + \left(\frac{2D}{h} + \frac{2\tau^c u^2}{h}\right)1 + \left(-\frac{D}{h} + \frac{u}{2} - \frac{\tau^c u^2}{h}\right)1 = 0 \tag{f}$$

or

$$0 = 0 \tag{g}$$

so the patch test is passed automatically.

The second specific reference solution ( $B = 1$ ) gives the nodal values

$$\phi_1 = \mathrm{e}^{-uh/D}, \qquad \phi_2 = 1, \qquad \phi_3 = \mathrm{e}^{uh/D} \tag{h}$$

with zero source term. Some manipulation of (d) ( $b_2 = 0$ ) gives first

$$\frac{\tau^c u^2}{h}(-\phi_1 + 2\phi_2 - \phi_3) = -\frac{D}{h}(-\phi_1 + 2\phi_2 - \phi_3) + \frac{u}{2}(\phi_1 - \phi_3) \tag{i}$$

and further

$$\tau^c = \frac{h}{2u}\,\frac{\phi_1 - \phi_3}{-\phi_1 + 2\phi_2 - \phi_3} - \frac{D}{u^2} \tag{j}$$

Substitution of the values (h) gives finally

$$\tau^c = \frac{h}{2u}\,\frac{\mathrm{e}^{-uh/D} - \mathrm{e}^{uh/D}}{-\mathrm{e}^{-uh/D} + 2 - \mathrm{e}^{uh/D}} - \frac{D}{u^2} \tag{k}$$

This can be brought into a cleaner form by using the local Peclet number

$$\mathrm{Pe}_h \equiv \mathrm{P} = \frac{uh}{D} \tag{l}$$



which produces

$$\begin{aligned}\tau^c &= \frac{h}{2u}\,\frac{\mathrm{e}^{-\mathrm{P}} - \mathrm{e}^{\mathrm{P}}}{-\mathrm{e}^{-\mathrm{P}} + 2 - \mathrm{e}^{\mathrm{P}}} - \frac{D}{u^2} = \frac{h}{2u}\,\frac{-2\sinh \mathrm{P}}{2 - 2\cosh \mathrm{P}} - \frac{D}{u^2} \\ &= \frac{h}{2u}\,\frac{-4\sinh(\mathrm{P}/2)\cdot\cosh(\mathrm{P}/2)}{2 - 2 - 4\sinh^2(\mathrm{P}/2)} - \frac{D}{u^2} = \frac{h}{2u}\,\frac{\cosh(\mathrm{P}/2)}{\sinh(\mathrm{P}/2)} - \frac{D}{u^2} \\ &= \frac{h}{2u}\coth(\mathrm{P}/2) - \frac{D}{u^2}\end{aligned} \tag{m}$$

or

$$\tau^c = \frac{h}{u}\left(\frac{1}{2}\coth\frac{\mathrm{Pe}_h}{2} - \frac{1}{\mathrm{Pe}_h}\right) \tag{n}$$

Some use of certain standard formulas for hyperbolic functions are needed in the manipulations.

The third reference solution ( $f_0 = 1$ ) gives the nodal values

$$\phi_1 = -h/u, \qquad \phi_2 = 0, \qquad \phi_3 = h/u \tag{o}$$

and the source term $f = 1$. The term

$$\begin{aligned}b_2 &= \int_{\Omega^1} N_2^1 \mathrm{d}\Omega + \int_{\Omega^2} N_1^2 \mathrm{d}\Omega + \tau^c u \int_{\Omega^1} \frac{\mathrm{d}N_2^1}{\mathrm{d}x}\mathrm{d}\Omega + \tau^c u \int_{\Omega^2} \frac{\mathrm{d}N_1^2}{\mathrm{d}x}\mathrm{d}\Omega \\ &= \frac{h}{2} + \frac{h}{2} + \tau^c u \cdot 1 + \tau^c u \cdot (-1) = h\end{aligned} \tag{p}$$

Equation (d) becomes

$$\left(-\frac{D}{h} - \frac{u}{2} - \frac{\tau^c u^2}{h}\right)\left(-\frac{h}{u}\right) + \left(-\frac{D}{h} + \frac{u}{2} - \frac{\tau^c u^2}{h}\right)\frac{h}{u} - h \overset{?}{=} 0 \tag{q}$$

This is also seen to be satisfied automatically. Continuing similarly, it is found that even in the case ( $(f_x)_0 = 1$ ) the patch test is passed but no more in the case ( $(f_{xx})_0 = 1$ ). Actually, the cases ( $A = 1$ ) and ( $f_0 = 1$ ) are of the type to be used in the standard patch test of Section 4.1 to verify convergence.

Some numerical results are shown in Figure 6.7 for a problem

$$-\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2} + 25\frac{\mathrm{d}\phi}{\mathrm{d}x} - 50 = 0 \quad \text{in} \quad \Omega = ]0,1[ \tag{56}$$

$$\phi(0) = 0, \qquad \phi(1) = 1 \tag{57}$$

presented in dimensionless form.

The behavior is according to the theory. Accurate nodal values are obtained with the optimal sensitizing parameter value even with the irregular mesh. The global Peclet number is $\mathrm{Pe}=25$ and the elementwise Peclet number for the regular element mesh elements is $\mathrm{Pe}_h=5$. According to Figure 6.5, for this value of $\mathrm{Pe}_h$, the approximate $\hat{\tau}^{\mathrm{c}}$ overestimates the optimal $\hat{\tau}^{\mathrm{c}}$ rather much leading to some overdamping. However, the sensitivity of the results on the value of the parameter seems not to be very strong.

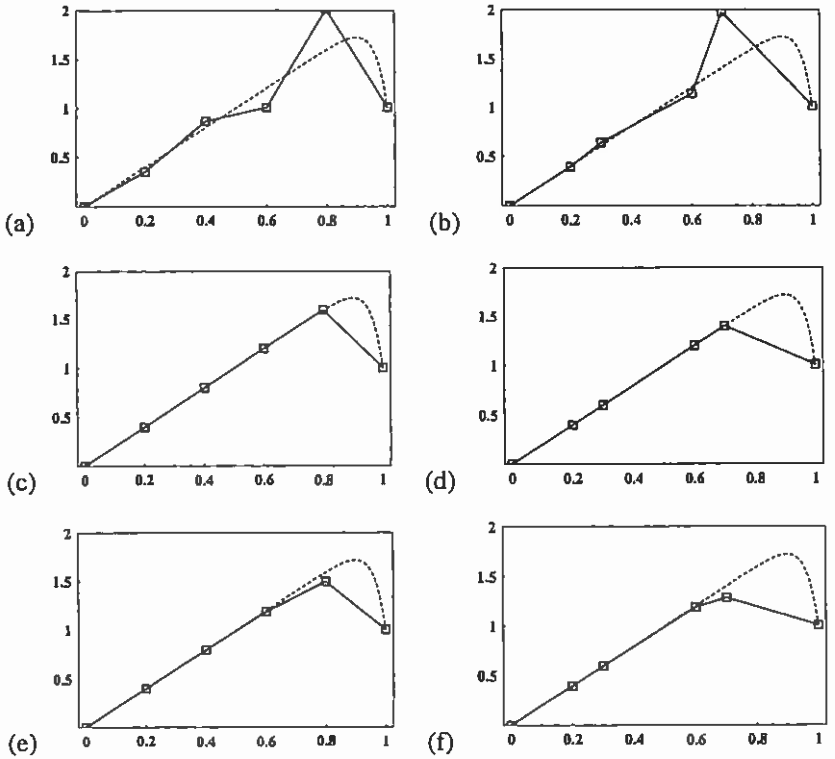



**Figure 6.7** On the left-hand side regular five element mesh, on the right-hand side irregular five element mesh. **(a)** and **(b)** Standard Galerkin method solution. **(c)** and **(d)** Sensitized Galerkin method solution with $\hat{\tau}^{\mathrm{c}}$ according to formula (48). **(e)** and **(f)** Sensitized Galerkin method solution with $\hat{\tau}^{\mathrm{c}}$ according to the approximate formula (50).

### 6.2.3 Boundary patch considerations



The possibility to study sensitizing in connection with natural boundary conditions was mentioned in Remark 5.11. We continue on this theme here. Let us consider in one dimension the left-hand side of a domain, say, with the notation used in Figure 5.2 (b) and with a given flux boundary condition

$$-n_x D\left(\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)_0=\bar{j}^{\mathrm{d}} \tag{58}$$

or

$$D\left(\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)_0=\bar{j}^{\mathrm{d}} \tag{59}$$

Here the value of the $x$-component of the unit outward normal vector is $-1$ and the meaning of the notations are understood from (6.1.27). In addition to the general solution (41):

$$\phi(x)=A+Be^{ux/D}+\phi_{\mathrm{p}}(x) \tag{60}$$

we must obviously introduce the boundary condition from (59):

$$\left(\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)_0=\frac{1}{D}\bar{j}^{\mathrm{d}} \tag{61}$$

Taking this into account in (60) gives the reference solution

$$\phi=A+\left[\frac{1}{u}\bar{j}^{\mathrm{d}}-\frac{D}{u}\left(\frac{\mathrm{d}\phi_{\mathrm{p}}}{\mathrm{d}x}\right)_0\right]e^{ux/D}+\phi_{\mathrm{p}} \tag{62}$$

This is applied in some detail in Example 6.2. The optimal $\tau^{\mathrm{c}}$ is found still to be according to expression (47). However, the constant source case $f_0=1$ does not any more pass the test.

The calculations can be repeated for the Robin condition

$$-n_x D\left(\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)_0=a\phi_0+b \tag{63}$$

Again it will be found that expression (47) is valid.

**Example 6.2.** We use again instead of the notation of Figure 5.2(b) that shown in Figure (a) for simplify the presentation. The case of given flux discussed above is considered and the task is to find the optimal sensitizing parameter value $\tau^c$.

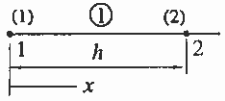



**Figure (a)**

The element contributions are still according to (55) but we must now include in detail the term due to the flux. Consideration of, say, the weak form (6.2.4) shows that the left hand contains the term

$$w\bar{j}^{\mathrm{d}}\Big|_{\Gamma_N} = w\bar{j}^{\mathrm{d}}\Big|_{x=0} \tag{a}$$

and thus the left hand side of the first (and here the only) system equation obtains from this the contribution

$$N_1\bar{j}^{\mathrm{d}}\Big|_{x=0} = 1\cdot\bar{j}^{\mathrm{d}} \tag{b}$$

The element contributions of the first element are thus (see formulas (a) of Example 6.1)

$$\begin{aligned}[K]^1 &= \frac{D}{h}\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix} + \frac{u}{2}\begin{bmatrix}-1 & 1\\ -1 & 1\end{bmatrix} + \frac{\tau^c u^2}{h}\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}\\ \{b\}^1 &= \int_{\Omega^1}\begin{Bmatrix}N_1^1 f\\ N_2^1 f\end{Bmatrix}\mathrm{d}\Omega + \tau^c u\int_{\Omega^1}\begin{Bmatrix}\mathrm{d}N_1^1/\mathrm{d}x\, f\\ \mathrm{d}N_2^1/\mathrm{d}x\, f\end{Bmatrix}\mathrm{d}\Omega - \begin{Bmatrix}\bar{j}^{\mathrm{d}}\\ 0\end{Bmatrix}\end{aligned} \tag{c}$$

In the system equation for node 1:

$$K_{11}\phi_1 + K_{12}\phi_2 - b_2 = 0 \tag{d}$$

$$\begin{aligned}K_{11} &= K_{11}^1 = \frac{D}{h} - \frac{u}{2} + \frac{\tau^c u^2}{h}\\ K_{12} &= K_{12}^1 = -\frac{D}{h} + \frac{u}{2} - \frac{\tau^c u^2}{h}\\ b_1 &= b_1^1 = \int_{\Omega^1} N_1^1 f\,\mathrm{d}\Omega + \tau^c u\int_{\Omega^1}\frac{\mathrm{d}N_1^1}{\mathrm{d}x} f\,\mathrm{d}\Omega - \bar{j}^{\mathrm{d}}\end{aligned} \tag{e}$$

Equation (d) is thus

$$\left(\frac{D}{h} - \frac{u}{2} + \frac{\tau^c u^2}{h}\right)\phi_1 + \left(-\frac{D}{h} + \frac{u}{2} - \frac{\tau^c u^2}{h}\right)\phi_2 - b_1 = 0 \tag{f}$$

The reference solution (62) can be written in more detail as

$$\begin{Bmatrix}\phi\\ f\end{Bmatrix} = A\begin{Bmatrix}1\\ 0\end{Bmatrix} + \bar{j}^{\mathrm{d}}\begin{Bmatrix}1/u\cdot\mathrm{e}^{ux/D}\\ 0\end{Bmatrix} + f_0\begin{Bmatrix}x/u - D/u^2\cdot\mathrm{e}^{ux/D}\\ 1\end{Bmatrix} + \cdots \tag{g}$$



The source term has been developed again in Taylor series.

The first reference solution ( $A = 1$ ) gives the nodal values

$$\phi_1 = 1, \qquad \phi_2 = 1, \qquad \phi_3 = 1 \tag{h}$$

with zero source term and zero flux. (Here the given flux is a "forcing" term similarly as the source function and their influence comes through the second and following specific reference solutions in (g)). Equation (d) is

$$\left(\frac{D}{h} - \frac{u}{2} + \frac{\tau^c u^2}{h}\right)1 + \left(-\frac{D}{h} + \frac{u}{2} - \frac{\tau^c u^2}{h}\right)1 = 0 \tag{i}$$

or

$$0 = 0 \tag{j}$$

so the patch test is passed automatically.

The second reference solution ( $\bar{j}^{\mathrm{d}} = 1$ ) gives the nodal values

$$\phi_1 = 1/u, \qquad \phi_2 = 1/u\cdot\mathrm{e}^{uh/D} \tag{k}$$

with zero source term and with a unit flux. Equation (f) is

$$\left(\frac{D}{h} - \frac{u}{2} + \frac{\tau^c u^2}{h}\right)\frac{1}{u} + \left(-\frac{D}{h} + \frac{u}{2} - \frac{\tau^c u^2}{h}\right)\frac{1}{u}\mathrm{e}^{uh/D} + 1 = 0 \tag{l}$$

We obtain

$$\tau^c = \frac{h}{u}\left(\frac{1}{2}\frac{\mathrm{e}^{\mathrm{Pe}_h}+1}{\mathrm{e}^{\mathrm{Pe}_h}-1} - \frac{1}{\mathrm{Pe}_h}\right) \tag{m}$$

where $\mathrm{Pe}_h$ is the element Peclet number. Some further manipulation shows that this is actually just equal to the optimal expression (47) foumd by the two-element patch test.

The third reference solution ( ( $f_0 = 1$ ) ) gives the nodal values

$$\phi_1 = -D/u^2 \qquad \phi_2 = h/u - D/u^2\cdot\mathrm{e}^{uh/D} \tag{n}$$

with the source term $f = 1$ and with zero flux. The term

$$b_1 = \int_{\Omega^1} N_1^1\mathrm{d}\Omega + \tau^c u\int_{\Omega^1}\frac{\mathrm{d}N_1^1}{\mathrm{d}x}\mathrm{d}\Omega = \frac{h}{2} - \tau^c u \tag{o}$$

Equation (f) becomes

$$\left(\frac{D}{h} - \frac{u}{2} + \frac{\tau^c u^2}{h}\right)\left(-\frac{D}{u^2}\right) + \left(-\frac{D}{h} + \frac{u}{2} - \frac{\tau^c u^2}{h}\right)\left(\frac{h}{u} - \frac{D}{u^2}\mathrm{e}^{uh/D}\right) - \frac{h}{2} + \tau^c u \overset{?}{=} 0 \tag{p}$$

Further manipulation shows that the patch test is no more passed.

## 6.3 TWO DIMENSIONS (unfinished)

### 6.3.1 Sensitized weak form; general considerations

**Introduction.** The governing field equation is

$$R(\phi) \neq \boxed{\frac{\partial}{\partial x_\alpha}\left(-D_{\alpha\beta}\frac{\partial \phi}{\partial x_\beta}\right)+\frac{\partial}{\partial x_\alpha}(v_\alpha \phi)-f=0} \tag{1}$$

in $\Omega$ with appropriate Dirichlet, Neumann and Robin boundary conditions as given in Section 6.1.2. Similarly as in Section 6.2.2, we employ here for sensitizing purposes the simplified equation

$$R^c(\phi) \equiv L^c(\phi)-f = \boxed{-D_{\alpha\beta}\frac{\partial^2 \phi}{\partial x_\alpha \partial x_\beta}+v_\alpha \frac{\partial \phi}{\partial x_\alpha}-f=0} \tag{2}$$

The sensitized weak form is thus

$$\boxed{\begin{aligned}&\int_\Omega \frac{\partial w}{\partial x_\alpha} D_{\alpha\beta}\frac{\partial \phi}{\partial x_\beta}\,d\Omega+\int_\Omega w\frac{\partial}{\partial x_\alpha}(v_\alpha \phi)\,d\Omega-\int_\Omega wf\,d\Omega+\text{bt}\\ &+\int_\Omega L^c(w)\tau^c R^c(\phi)\,d\Omega=0\end{aligned}} \tag{3}$$

The steps needed to obtain (3) should be obvious from the earlier derivations.

**Remark 6.11.** The sensitizing integrand in (3) is in detail (to avoid erroneous application of the summation convention, different indexing is used here in the weighting term and in the residual)

$$\begin{aligned}&L^c(w)\tau^c R^c(\phi)=\\ &\left(-D_{\alpha\beta}\frac{\partial^2 w}{\partial x_\alpha \partial x_\beta}+\underline{v_\alpha \frac{\partial w}{\partial x_\alpha}}\right)\tau^c\left(-D_{\gamma\delta}\frac{\partial^2 \phi}{\partial x_\gamma \partial x_\delta}+\underline{v_\gamma \frac{\partial \phi}{\partial x_\gamma}}-f\right)\end{aligned} \tag{4}$$

The important term from the point of view of convection comes from the underlined terms:

$$\frac{\partial w}{\partial x_\alpha}\tau^c v_\alpha v_\gamma \frac{\partial \phi}{\partial x_\gamma} \tag{5}$$

or using matrix notation in two dimensions:

$$\begin{Bmatrix}\partial w/\partial x\\ \partial w/\partial y\end{Bmatrix}\begin{bmatrix}\tau^c uu & \tau^c uv\\ \tau^c vu & \tau^c vv\end{bmatrix}\begin{Bmatrix}\partial \phi/\partial x\\ \partial \phi/\partial y\end{Bmatrix} \tag{6}$$



This term has been discussed in Section D.4.2 (without the parameter $\tau^c$). Comparison with the first integrand in (3) shows that sensitizing can be interpreted as injection of *anisotropic damping diffusion* into the formulation. The damping diffusivity tensor is $\tau^c v_\alpha v_\beta$. Firstly, the real diffusivity tensor $D_{\alpha\beta}$ is normally for physical reasons positive definite, the tensor $\tau^c v_\alpha v_\beta$ is, however, only positive semidefinite (the determinant is zero). According to Crandall (1956, p.355), the corresponding pure diffusion problems would be elliptic and parabolic, respectively. Second, if we momentarily take for example the $x$-axis to coincide with the local flow direction, the damping diffusivity matrix becomes

$$\begin{bmatrix}\tau^c uu & 0\\ 0 & 0\end{bmatrix} \tag{7}$$

as in this coordinate system $v=0$. This can interpreted physically, say, in connection with heat conduction so that the conductivity is zero perpendicular to streamlines and the information can proceed only along the streamlines. (To make the case more concrete, we could imagine an isotropic bulk material consisting of separate fibers of highly conducting material embedded in a highly isolating matrix. Temperature measurements are performed in the conducting material only.) Altogether, the least squares sensitizing is seen to mimic in an admirable way the pure convection behavior. □

**Remark 6.12.** In the discrete equations to follow, we always further simplify by neglecting the second order derivatives possibly appearing in $R^c(\tilde{\phi})$ and $L^c(\tilde{w})$. This is correct for three-noded triangular elements but not in general for four-noded quadrilateral elements. It can be shown, Freund (1996), that the resulting consistency error does not affect the rate of convergence. □

**Remark 6.13.** Recalling expressions (5) and (6) in Remark 6.11 we will use the following notation for the damping diffusivity tensor

$$D^c_{\alpha\beta} \equiv \tau^c v_\alpha v_\beta \tag{8}$$

and in two dimensions for the damping diffusivity matrix

$$\left[D^c\right] \equiv \begin{bmatrix}\tau^c uu & \tau^c uv\\ \tau^c vu & \tau^c vv\end{bmatrix} \tag{9}$$

It seems that in numerical determination of the appropriate damping, it is more straightforward to determine directly the damping terms than first the parameter $\tau^c$. In fact, it is the damping diffusivities, which are needed in the final element contribution calculations. We therefore write (8) and (9) in the forms

$$D^c_{\alpha\beta} = D^c \frac{v_\alpha}{|v|}\frac{v_\beta}{|v|} \tag{10}$$

and

$$\left[D^c\right]=\begin{bmatrix} D^c\dfrac{u}{|\mathbf{v}|}\dfrac{u}{|\mathbf{v}|} & D^c\dfrac{u}{|\mathbf{v}|}\dfrac{v}{|\mathbf{v}|} \\ D^c\dfrac{v}{|\mathbf{v}|}\dfrac{u}{|\mathbf{v}|} & D^c\dfrac{v}{|\mathbf{v}|}\dfrac{v}{|\mathbf{v}|} \end{bmatrix} \tag{11}$$

where the scalar damping diffusivity

$$D^c \equiv \tau^c |\mathbf{v}|^2 \tag{12}$$

The terms $v_\alpha/|\mathbf{v}|$ or $u/|\mathbf{v}|$ etc. are direction cosines of the velocity vector. We remember from Remark 6.11 that the damping diffusivity tensor "acts only in streamline direction". This means in other words that in a coordinate system with one coordinate axis in the velocity direction, there is only one non-zero tensor component in that system and it has the double indices corresponding to the axis in question. This component is here the scalar $D^c$. This can be seen in detail by transforming this very simple tensor to the present coordinate system by well-known tensor transformation formulas. Formulas (11) and (12) are found to follow. In sensitizing patch test calculations we therefore concentrate on determining directly the scalar $D^c$ and then finally use (10) or (11). $D^c$ is more transparent than $\tau^c$ as we can compare $D^c$ directly in values with $D$ or in the general case with $D_{\alpha\beta}$. □

Application of the Galerkin method in (3) gives the set of system equations (take Remarks 6.11 to 6.13 into account)

$$[K]\{a\}=\{b\} \tag{13}$$

with

$$\begin{aligned} K_{ij} &= \int_\Omega \frac{\partial N_i}{\partial x_\alpha} D_{\alpha\beta} \frac{\partial N_j}{\partial x_\beta} \mathrm{d}\Omega + \int_\Omega N_i \frac{\partial}{\partial x_\alpha}\left(v_\alpha N_j\right)\mathrm{d}\Omega \\ &\quad + \int_\Omega \frac{\partial N_i}{\partial x_\alpha} D^c \frac{v_\alpha}{|\mathbf{v}|}\frac{v_\beta}{|\mathbf{v}|}\frac{\partial N_j}{\partial x_\beta}\mathrm{d}\Omega + \mathrm{bt} \\ b_i &= \int_\Omega N_i f \,\mathrm{d}\Omega + \int_\Omega \frac{\partial N_i}{\partial x_\alpha}\frac{D^c}{|\mathbf{v}|}\frac{v_\alpha}{|\mathbf{v}|} f \,\mathrm{d}\Omega + \mathrm{bt} \end{aligned} \tag{14}$$

The element contribution expressions are usually no more given in what follows as they should be obvious on the basis of Remark 2.11.

We have now to determine the sensitizing parameter $\tau^c$ (or $D^c$) in a considerable more complicated situation than in the one-dimensional case. New features to be dealt with are connected to reference solutions, to element cloning for the sensitizing patch test and to the directional property of the convection term. We will discuss them each separately.



**Reference solutions.** The series form reference solution strategy described in Section 5.2.2 could be applied. However, a more straightforward procedure seems to be here to make direct use of the one-dimensional reference solutions of Section 6.2.2. We assume what we call here *cylindrical solutions* (or just one-dimensional solutions) in different directions described by the line $s$ in Figure 6.8 (a). That is, we take

$$\phi = \phi(s) \tag{15}$$

where

$$s = \mathbf{e}\cdot\mathbf{x} = e_\alpha x_\alpha = e_1 x_1 + e_2 x_2 = \cos\psi_1 \cdot x_1 + \cos\psi_2 \cdot x_2 \tag{16}$$

Here $\mathbf{e}$ is the unit vector along $s$ and the rest of the notations are understood from Figure 6.8 (a). This means that $\phi$ is assumed to be constant on any line perpendicular to $\mathbf{e}$. (We employ here still mainly indexed notations $x_1$ and $x_2$ instead of $x$ and $y$ so that a possible extension to three dimensions becomes more transparent.)

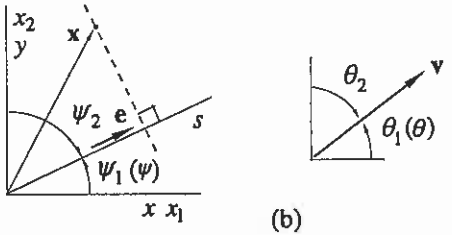



**Figure 6.8 (a)** Cylindrical solution direction and some notations. **(b)** Flow velocity vector.

Chain differentiation gives

$$\frac{\partial\phi}{\partial x_\alpha} = \frac{\mathrm{d}\phi}{\mathrm{d}s}\frac{\partial s}{\partial x_\alpha} = e_\alpha \frac{\mathrm{d}\phi}{\mathrm{d}s} \tag{17}$$

and similarly

$$\frac{\partial^2\phi}{\partial x_\alpha \partial x_\beta} = e_\alpha e_\beta \frac{\mathrm{d}^2\phi}{\mathrm{d}s^2} \tag{18}$$

Field equation (2) becomes

$$\boxed{-\bar{D}\frac{d^2\phi}{ds^2}+\bar{u}\frac{d\phi}{ds}-f=0} \tag{19}$$

with

$$\begin{aligned}\bar{D}=e_\alpha e_\beta D_{\alpha\beta} &= \cos\psi_1\cos\psi_1\cdot D_{11}+\cos\psi_1\cos\psi_2\cdot D_{12}\\ &+\cos\psi_2\cos\psi_2\cdot D_{21}+\cos\psi_2\cos\psi_2\cdot D_{22}\end{aligned} \tag{20}$$

$$\bar{u}=e_\alpha v_\alpha=\cos\psi_1\cdot v_1+\cos\psi_2\cdot v_2 \tag{21}$$

In the normal case with fluids we have isotropic diffusivity, i.e., $D_{11}=D_{22}=D$, $D_{12}=D_{21}=0$, and we obtain from (21) simply $\bar{D}=D$. Using the notations of Figure 6.8 (b), we can further write

$$v_1=\cos\theta_1\,|\,\mathbf{v}\,|,\qquad v_2=\cos\theta_2\,|\,\mathbf{v}\,| \tag{22}$$

and we arrive at a more transparent formula for $\bar{u}$ ($\psi_1\to\psi$, $\theta_1\to\theta$):

$$\begin{aligned}\bar{u}&=\cos\psi_1\cdot\cos\theta_1\,|\,\mathbf{v}\,|+\cos\psi_2\cdot\cos\theta_2\,|\,\mathbf{v}\,|\\ &=\big(\cos\psi\cdot\cos\theta+\sin\psi\cdot\sin\theta\big)\,|\,\mathbf{v}\,|=\cos\big(\psi-\theta\big)\,|\,\mathbf{v}\,|\end{aligned} \tag{23}$$

Equation (19) is exactly of the type we have dealt with in one dimension. Thus the corresponding reference solution is (cf. (6.2.40) and (6.2.46))

$$\exp\frac{\bar{u}s}{\bar{D}}=\exp\frac{\bar{u}e_\alpha x_\alpha}{\bar{D}}=\exp\frac{\bar{u}\big(\cos\psi_1\cdot x_1+\cos\psi_2\cdot x_2\big)}{\bar{D}} \tag{24}$$

How we should select the direction of $s$ to obtain the equation needed for the determination of $D^c$ by the sensitizing patch test is discussed somewhat later on.

Let us consider as an example the four-noded quadrilateral element shown in Figure 6.9 (a) in the $xy$-plane. Quite a lot of data is needed to describe the element: the overall orientation, the size, the aspect ratio, the skewness, etc. In the element's natural coordinate system (Figure (b)) the situation is greatly simplified: the orientation is aligned along the coordinate axes, the size is 1×1, etc. This leads to the obvious thought that the properties of the element should studied and the possible patch test be performed first in the natural coordinate system. After that the results can be transformed back to the physical plane. In some of our earlier efforts, efforts, e.g., Freund and Salonen (1995), we have indeed proceeded in this manner to obtain as far as possible closed form expression for the sensitizing parameter. This has, however, meant in practice



some additional assumptions during the manipulations. Also, the transformation formulas become very unwieldy and complicated to explain. If we decide to abandon the analytical approach and try to determine the parameter values only numerically, the situation becomes much simpler to follow and new practical advantages are likely to emerge. The price one has to pay is that a number of equation sets have to be generated and solved with the sensitizing patch test for each element to find the parameter values. In the diffusion-convection problem under study, only one equation is needed.

**Element cloning.** In one dimension the cloning of an element to produce a patch for the sensitizing patch test contains no problems. Contrary to this, let us consider again the four-noded quadrilateral element (Figure 6.9 (a)). If we replace the element by a *substitute parallelogram* element (say by replacing the isoparametric mapping from the reference element by a linearized mapping in $\xi$ and $\eta$ applying a Taylor expansion at the midpoint of the element) indicated by the dashed line in the figure we can produce a patch of four elements which has no gaps or overlaps.

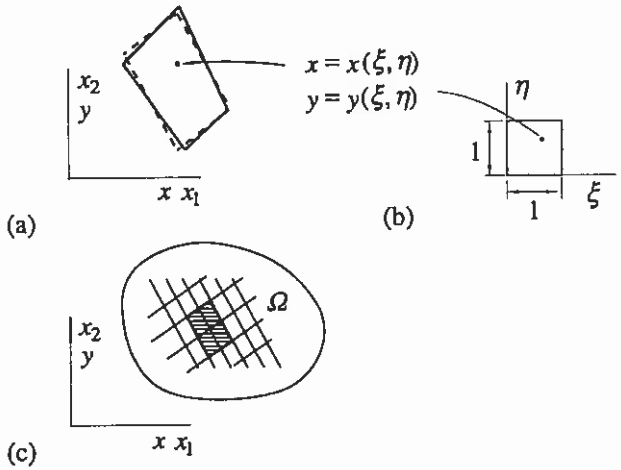



**Figure 6.9 (a)** An element in the $xy$-plane. **(b)** The element in the natural $\xi\eta$-plane. **(c)** $xy$-plane and a patch.

It should be emphasized that this replacement is done only for obtaining the sensitizing parameter value for the element. The actual form of the element should naturally be used in the final calculations. However, in Section 6.3.2 we suggest a further modified version of the patch test for the four-noded quadrilateral element, which is in more accordance with the procedure used for the triangle element.

**Directional property of the convection term.** The two- or three-dimensional diffusion-convection problem *is different in nature from the one-dimensional case in one important respect*: the strength of the convection term depends now also on the angle between the flow velocity direction and the (initially unknown) solution gradient direction. In one dimension these directions coincide. Let us look in more detail the convection term in (1) (for simplicity of presentation in the incompressible case $\partial v_\alpha / \partial x_\alpha = 0$):

$$\frac{\partial}{\partial x_\alpha}(v_\alpha \phi) = v_\alpha \frac{\partial \phi}{\partial x_\alpha} = \mathbf{v} \cdot \nabla \phi \tag{25}$$

The convection term is seen to consist of the scalar product of the velocity vector $\mathbf{v}$ and the solution gradient vector $\nabla \phi$. For instance, if the angle between $\mathbf{v}$ and $\nabla \phi$ happens to be 90 degrees, the convection term vanishes and we in principle need at such a point no sensitizing as we have there a pure diffusion case. This feature automatically means that to obtain really good sensitizing, *an iterative method is necessary* to feed in information about the originally unknown gradient direction. (This is now a case in contrary to the discussion in Section 5.2.1 about the possible need for a preliminary solution.) The natural procedure here seems to take the cylindrical reference solution in the gradient direction. Then the actual solution and the reference solution are as far as possible similar in the sense that they both have the same gradient direction. The local data for the element (and the new gradient direction) are evaluated from the values at the element midpoint coordinates. When numerical solutions are used, the specific cases appearing with the local Peclet number small or large etc. need a careful study. This is considered next.

**Computational aspects.** In the numerical determination of the scalar damping diffusivity $D^c$ using the sensitizing patch test, certain points must be taken into account to avoid ill-conditioning. Some of these points have been found from experience with the simple case considered in Example 6.3, where a "semianalytical" approach sheds some light on solution behavior. Two obvious difficult situations appear: the velocity field is very weak or it is very strong.

We consider first the weak velocity case. We define here for each element an element Peclet number

$$\frac{|\mathbf{v}| h_m}{D_m} \tag{26}$$

Here $h_m$ is a linear measure of the element, in two dimensions, say $h_m = \sqrt{A}$, where $A$ is the area of the element. With anisotropic diffusivity, $D_m$ could be in



two dimensions, say $(D_{xx} + D_{yy})/2$. In the isotropic case this is $D$. If (26) becomes small, we have numerical difficulties. The reason is that the velocity $\bar{u}$ in (24) is then also very small and the reference solution is nearly a constant (or more generally nearly linear in $x$ and $y$) and we then know that the multiplier of $D^c$ in the discrete equation obtained by the patch test nearly disappears. (The patch test is passed for convergence reasons irrespective of the value of the sensitizing parameter.) Thus if

$$\frac{|\mathbf{v}| h_m}{D_m} < \varepsilon_1 \tag{27}$$

where $\varepsilon_1$ is a small positive number obtained by numerical experiments by a computer, we put simply

$$D^c = 0 \tag{28}$$

as the situation is in practice of the pure diffusion type.

We consider next the strong velocity case. We define the quantity

$$\frac{\bar{u}\bar{h}}{\bar{D}} \tag{29}$$

where $\bar{h}$ is in principle the utmost measure in the patch in the $\psi$ angle direction. If (29) becomes large, we encounter numerical difficulties, as some nodal values in the patch test from the reference solution can become then extremely large. If

$$\frac{\bar{u}\bar{h}}{\bar{D}} > \varepsilon_2 \tag{30}$$

where $\varepsilon_2$ is determined again by numerical experiments, we reckon as follows. In the one-dimensional case we see from Figure 6.5 and from formula (6.2.49) that the quantity

$$\frac{D^c}{uh} \tag{31}$$

remains nearly constant with large element Peclet numbers. Let us say that we have obtained for a reduced rather large $u = \hat{u}$ the corresponding $D^c = \hat{D}^c$. Then we know the ratio $\hat{D}^c/(\bar{u}h)$ and putting

$$\frac{D^c}{uh}=\frac{\hat{D}^c}{\hat{u}h} \tag{32}$$

we obtain

$$D^c=\frac{u}{\hat{u}}\hat{D}^c \tag{33}$$

Here $\hat{u}$ represents a reduced velocity by which we can still safely evaluate numerically $\hat{D}^c$. This reasoning described above for the one-dimensional case is applied now also here. We reduce the velocity $\overline{u}$ in (29) to a smaller value $\hat{\overline{u}}$ (in magnitude) so that

$$\frac{\hat{\overline{u}}\overline{h}}{\overline{D}}=\varepsilon_2 \tag{34}$$

We then determine the corresponding damping diffusivity $D^c=\hat{D}^c$ using the sensitizing patch test. The final damping diffusivity is taken to be

$$D^c=\frac{\overline{u}}{\hat{\overline{u}}}\hat{D}^c \tag{35}$$

An additional numerical difficulty appears when the gradient direction angle $\psi$ becomes nearly perpendicular to the velocity vector direction angle $\theta$, that is, when $|\psi-\theta|\approx 90^\circ$ (see Remark 6.14). Contrary to what one would expect from the continuum case discussed earlier, rather suprisingly, the following was found in most cases studied in Example 6.3. When $|\psi-\theta|\approx 90^\circ$, the damping diffusivity $D^c$ does not usually tend to zero, but in fact obtains larger values than when the directions are parallel. Now if $|\psi-\theta|\approx 90^\circ$, the velocity $\overline{u}$ becomes nearly zero (see (23)) and the numerical solution does not succeed due to ill-conditioning. However, here we cannot any more proceed realistically using (28). Thus we change the gradient direction slightly to $\psi:=\psi+\Delta\psi$, where $\Delta\psi$ is a small angle making $|\psi-\theta|$ to differ more from $90^\circ$ so that $D^c$ can still be evaluated with reasonable accuracy. Of course, for this to work, we again need some limits for practical calculations.

**Remark 6.14.** It can be seen from formulas (16) and (23) that if the direction of $s$ is changed opposite, that is, if $\psi$ is changed to $\psi+\pi$, the reference solution (24) does not change. We can therefore select the direction of the reference solution from the two possibilities always so that $|\psi-\theta|\le 90^\circ$ when applying the sensitizing patch test. □



The obvious starting direction for $\psi$ for lack of any further information is to take simply $\psi=\theta$. The new directions for $\psi$ in each element are taken according to gradient directions obtained from the solutions to follow. If the magnitude of the gradient is rather small (compared to a predefined reference value), there seems to be no strong point to try to update the directions. Otherwise even small "stochastic" changes would demand new irrelevant changes.

### 6.3.2 Quadrilateral elements

Figure 6.10 (a) shows a generic quadrilateral element and Figure 6.10 (b) the corresponding cloned patch. Now we not even care to use the substitute parallelogram concept discussed in connection with Figure 6.9. Each of the four nodes of the element is connected to the central patch node so the situation is in this sense impartial with respect to the element nodes. The patch is no more conventional as there are gaps and some parts overlap. We are ready to accept this kind of ad hoc procedure to simplify the treatment. The remarks in Section 5.3.2 justify all kind of "crimes" in connection with sensitizing parameters as far as their values tend to zero with vanishing element size.

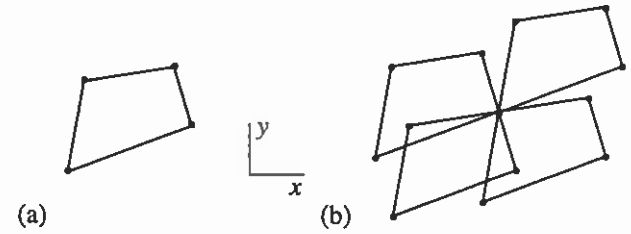



**Figure 6.10 (a)** Quadrilateral element. **(b)** Sensitizing patch.

**Example 6.3.** We try to obtain some knowledge about the behavior of the damping diffusivity as a function of the cylindrical solution direction in a simplest possible situation. We consider the case of square element shown in Figure (a). The corresponding sensitizing patch is shown in Figure (b).

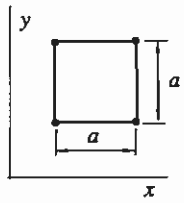



Figure (a)

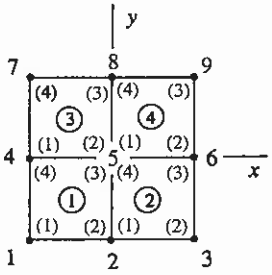



Figure (b)

The element contributions are from (14) applied at the element level

$$K_{ij}^e = D\int_{\Omega^e}\left(\frac{\partial N_i}{\partial x}\frac{\partial N_j}{\partial x}+\frac{\partial N_i}{\partial y}\frac{\partial N_j}{\partial y}\right)\mathrm{d}\Omega + u\int_{\Omega^e} N_i\frac{\partial N_j}{\partial x}\mathrm{d}\Omega + v\int_{\Omega^e} N_i\frac{\partial N_j}{\partial y}\mathrm{d}\Omega$$
$$+ D^c\frac{u}{|v|}\frac{u}{|v|}\int_{\Omega^e}\frac{\partial N_i}{\partial x}\frac{\partial N_j}{\partial x}\mathrm{d}\Omega + D^c\frac{u}{|v|}\frac{v}{|v|}\int_{\Omega^e}\frac{\partial N_i}{\partial x}\frac{\partial N_j}{\partial y}\mathrm{d}\Omega$$
$$+ D^c\frac{v}{|v|}\frac{u}{|v|}\int_{\Omega^e}\frac{\partial N_i}{\partial y}\frac{\partial N_j}{\partial x}\mathrm{d}\Omega + D^c\frac{v}{|v|}\frac{v}{|v|}\int_{\Omega^e}\frac{\partial N_i}{\partial y}\frac{\partial N_j}{\partial y}\mathrm{d}\Omega \qquad (a)$$
$$b_i^e = \int_{\Omega^e} N_i f\,\mathrm{d}\Omega + \frac{D^c}{|v|}\frac{u}{|v|}\int_{\Omega^e}\frac{\partial N_i}{\partial x} f\,\mathrm{d}\Omega + \frac{D^c}{|v|}\frac{v}{|v|}\int_{\Omega^e}\frac{\partial N_i}{\partial y} f\,\mathrm{d}\Omega$$

We have resorted to the conventional notation $x_1 \to x$, $x_2 \to y$, $v_1 \to u$, $v_2 \to v$. Further, isotropic diffusivity and constant $D$, $u$, $v$ have been assumed.

With rectangular four-noded elements aligned along the coordinate axes it is not necessary to make use of isoparametric mappings etc. as we can find closed form expressions directly using formulas (F.2.3). We obtain in detail

$$K_{11}^e = D\frac{4}{6} - ua\frac{2}{12} - va\frac{2}{12} + D^c\left(\frac{u}{|v|}\frac{u}{|v|}\frac{2}{6}+\frac{u}{|v|}\frac{v}{|v|}\frac{1}{4}+\frac{v}{|v|}\frac{u}{|v|}\frac{1}{4}+\frac{v}{|v|}\frac{v}{|v|}\frac{2}{6}\right)$$
$$K_{12}^e = -D\frac{1}{6} + ua\frac{2}{12} - va\frac{1}{12} + D^c\left(-\frac{u}{|v|}\frac{u}{|v|}\frac{2}{6}+\frac{u}{|v|}\frac{v}{|v|}\frac{1}{4}-\frac{v}{|v|}\frac{u}{|v|}\frac{1}{4}+\frac{v}{|v|}\frac{v}{|v|}\frac{1}{6}\right)$$
$$K_{13}^e = -D\frac{2}{6} + ua\frac{1}{12} + va\frac{1}{12} + D^c\left(-\frac{u}{|v|}\frac{u}{|v|}\frac{1}{6}-\frac{u}{|v|}\frac{v}{|v|}\frac{1}{4}-\frac{v}{|v|}\frac{u}{|v|}\frac{1}{4}-\frac{v}{|v|}\frac{v}{|v|}\frac{1}{6}\right)$$
$$K_{14}^e = -D\frac{1}{6} - ua\frac{1}{12} + va\frac{2}{12} + D^c\left(\frac{u}{|v|}\frac{u}{|v|}\frac{1}{6}-\frac{u}{|v|}\frac{v}{|v|}\frac{1}{4}+\frac{v}{|v|}\frac{u}{|v|}\frac{1}{4}-\frac{v}{|v|}\frac{v}{|v|}\frac{2}{6}\right)$$



$$K_{21}^e = -D\frac{1}{6} - ua\frac{2}{12} - va\frac{1}{12} + D^c\left(-\frac{u}{|v|}\frac{u}{|v|}\frac{2}{6}-\frac{u}{|v|}\frac{v}{|v|}\frac{1}{4}+\frac{v}{|v|}\frac{u}{|v|}\frac{1}{4}+\frac{v}{|v|}\frac{v}{|v|}\frac{1}{6}\right)$$
$$K_{22}^e = D\frac{4}{6} + ua\frac{2}{12} - va\frac{2}{12} + D^c\left(\frac{u}{|v|}\frac{u}{|v|}\frac{2}{6}-\frac{u}{|v|}\frac{v}{|v|}\frac{1}{4}-\frac{v}{|v|}\frac{u}{|v|}\frac{1}{4}+\frac{v}{|v|}\frac{v}{|v|}\frac{2}{6}\right)$$
$$K_{23}^e = -D\frac{1}{6} + ua\frac{1}{12} + va\frac{2}{12} + D^c\left(\frac{u}{|v|}\frac{u}{|v|}\frac{1}{6}+\frac{u}{|v|}\frac{v}{|v|}\frac{1}{4}-\frac{v}{|v|}\frac{u}{|v|}\frac{1}{4}-\frac{v}{|v|}\frac{v}{|v|}\frac{2}{6}\right)$$
$$K_{24}^e = -D\frac{2}{6} - ua\frac{1}{12} + va\frac{1}{12} + D^c\left(-\frac{u}{|v|}\frac{u}{|v|}\frac{1}{6}+\frac{u}{|v|}\frac{v}{|v|}\frac{1}{4}+\frac{v}{|v|}\frac{u}{|v|}\frac{1}{4}-\frac{v}{|v|}\frac{v}{|v|}\frac{1}{6}\right)$$

$$K_{31}^e = -D\frac{2}{6} - ua\frac{1}{12} - va\frac{1}{12} + D^c\left(-\frac{u}{|v|}\frac{u}{|v|}\frac{1}{6}-\frac{u}{|v|}\frac{v}{|v|}\frac{1}{4}-\frac{v}{|v|}\frac{u}{|v|}\frac{1}{4}-\frac{v}{|v|}\frac{v}{|v|}\frac{1}{6}\right)$$
$$K_{32}^e = -D\frac{1}{6} + ua\frac{1}{12} - va\frac{2}{12} + D^c\left(\frac{u}{|v|}\frac{u}{|v|}\frac{1}{6}-\frac{u}{|v|}\frac{v}{|v|}\frac{1}{4}+\frac{v}{|v|}\frac{u}{|v|}\frac{1}{4}-\frac{v}{|v|}\frac{v}{|v|}\frac{2}{6}\right)$$
$$K_{33}^e = D\frac{4}{6} + ua\frac{2}{12} + va\frac{2}{12} + D^c\left(\frac{u}{|v|}\frac{u}{|v|}\frac{2}{6}+\frac{u}{|v|}\frac{v}{|v|}\frac{1}{4}+\frac{v}{|v|}\frac{u}{|v|}\frac{1}{4}+\frac{v}{|v|}\frac{v}{|v|}\frac{2}{6}\right)$$
$$K_{34}^e = -D\frac{1}{6} - ua\frac{2}{12} + va\frac{2}{12} + D^c\left(-\frac{u}{|v|}\frac{u}{|v|}\frac{2}{6}+\frac{u}{|v|}\frac{v}{|v|}\frac{1}{4}-\frac{v}{|v|}\frac{u}{|v|}\frac{1}{4}+\frac{v}{|v|}\frac{v}{|v|}\frac{1}{6}\right)$$ (b)

$$K_{41}^e = -D\frac{1}{6} - ua\frac{1}{12} - va\frac{2}{12} + D^c\left(\frac{u}{|v|}\frac{u}{|v|}\frac{1}{6}+\frac{u}{|v|}\frac{v}{|v|}\frac{1}{4}-\frac{v}{|v|}\frac{u}{|v|}\frac{1}{4}-\frac{v}{|v|}\frac{v}{|v|}\frac{2}{6}\right)$$
$$K_{42}^e = -D\frac{2}{6} + ua\frac{1}{12} - va\frac{1}{12} + D^c\left(-\frac{u}{|v|}\frac{u}{|v|}\frac{1}{6}+\frac{u}{|v|}\frac{v}{|v|}\frac{1}{4}+\frac{v}{|v|}\frac{u}{|v|}\frac{1}{4}-\frac{v}{|v|}\frac{v}{|v|}\frac{1}{6}\right)$$
$$K_{43}^e = -D\frac{1}{6} + ua\frac{2}{12} + va\frac{1}{12} + D^c\left(-\frac{u}{|v|}\frac{u}{|v|}\frac{2}{6}-\frac{u}{|v|}\frac{v}{|v|}\frac{1}{4}+\frac{v}{|v|}\frac{u}{|v|}\frac{1}{4}+\frac{v}{|v|}\frac{v}{|v|}\frac{1}{6}\right)$$
$$K_{44}^e = D\frac{4}{6} - ua\frac{2}{12} + va\frac{2}{12} + D^c\left(\frac{u}{|v|}\frac{u}{|v|}\frac{2}{6}-\frac{u}{|v|}\frac{v}{|v|}\frac{1}{4}-\frac{v}{|v|}\frac{u}{|v|}\frac{1}{4}+\frac{v}{|v|}\frac{v}{|v|}\frac{2}{6}\right)$$

The system equation for node 5 ( $f = 0$ ) of the mesh in Figure (b) is

$$K_{51}\phi_1 + K_{52}\phi_2 + K_{53}\phi_3 + K_{54}\phi_4 + K_{55}\phi_5 + K_{56}\phi_6 + K_{57}\phi_7 + K_{58}\phi_8 + K_{59}\phi_9 = 0 \qquad (c)$$

with

$$K_{51} = K_{31}^1 = -\frac{1}{3}D - \frac{1}{12}\frac{u}{|v|}|v|a - \frac{1}{12}va + \left(-\frac{1}{6}\frac{u}{|v|}\frac{u}{|v|}-\frac{1}{2}\frac{u}{|v|}\frac{v}{|v|}-\frac{1}{6}\frac{v}{|v|}\frac{v}{|v|}\right)D^c$$
$$K_{52} = K_{32}^1 + K_{41}^2 = -\frac{1}{3}D - \frac{4}{12}\frac{v}{|v|}|v|a + \left(\frac{2}{6}\frac{u}{|v|}\frac{u}{|v|}-\frac{4}{6}\frac{v}{|v|}\frac{v}{|v|}\right)D^c$$

$$K_{53} = K_{42}^2 = -\frac{1}{3}D + \frac{1}{12}\frac{u}{|v|}|v|a - \frac{1}{12}\frac{v}{|v|}|v|a + \left(-\frac{1}{6}\frac{u}{|v|}\frac{u}{|v|} + \frac{1}{2}\frac{u}{|v|}\frac{v}{|v|} - \frac{1}{6}\frac{v}{|v|}\frac{v}{|v|}\right)D^c$$

$$K_{54} = K_{34}^1 + K_{21}^3 = -\frac{1}{3}D - \frac{4}{12}\frac{u}{|v|}|v|a + \left(-\frac{4}{6}\frac{u}{|v|}\frac{u}{|v|} + \frac{2}{6}\frac{v}{|v|}\frac{v}{|v|}\right)D^c$$

$$K_{55} = K_{33}^1 + K_{44}^2 + K_{22}^3 + K_{11}^4 = \frac{8}{3}D + \left(\frac{8}{6}\frac{u}{|v|}\frac{u}{|v|} + \frac{8}{6}\frac{v}{|v|}\frac{v}{|v|}\right)D^c \qquad \text{(d)}$$

$$K_{56} = K_{43}^2 + K_{12}^4 = -\frac{1}{3}D + \frac{4}{12}\frac{u}{|v|}|v|a + \left(-\frac{4}{6}\frac{u}{|v|}\frac{u}{|v|} + \frac{2}{6}\frac{v}{|v|}\frac{v}{|v|}\right)D^c$$

$$K_{57} = K_{24}^3 = -\frac{1}{3}D - \frac{1}{12}\frac{u}{|v|}|v|a + \frac{1}{12}\frac{v}{|v|}|v|a + \left(-\frac{1}{6}\frac{u}{|v|}\frac{u}{|v|} + \frac{1}{2}\frac{u}{|v|}\frac{v}{|v|} - \frac{1}{6}\frac{v}{|v|}\frac{v}{|v|}\right)D^c$$

$$K_{58} = K_{23}^3 + K_{14}^4 = -\frac{1}{3}D + \frac{4}{12}\frac{v}{|v|}|v|a + \left(\frac{2}{6}\frac{u}{|v|}\frac{u}{|v|} - \frac{4}{6}\frac{v}{|v|}\frac{v}{|v|}\right)D^c$$

$$K_{59} = K_{13}^4 = -\frac{1}{3}D + \frac{1}{12}\frac{u}{|v|}|v|a + \frac{1}{12}\frac{v}{|v|}|v|a + \left(-\frac{1}{6}\frac{u}{|v|}\frac{u}{|v|} - \frac{1}{2}\frac{u}{|v|}\frac{v}{|v|} - \frac{1}{6}\frac{v}{|v|}\frac{v}{|v|}\right)D^c$$

Equation (c) is presented in Figure (c) using self-evident "mathematical molecules", Salvadori and Baron (1961). The field equation corresponding to the weak form (3) after all the simplifications mentioned earlier is given also in the figure. Readers familiar with central difference formulas can detect the connections between the field equation and the molecules.

$$\frac{1}{3}\begin{matrix}-1&-1&-1\\-1&8&-1\\-1&-1&-1\end{matrix}D + \left[\frac{1}{12}\begin{matrix}-1&0&1\\-4&0&4\\-1&0&1\end{matrix}\frac{u}{|v|} + \frac{1}{12}\begin{matrix}1&4&1\\0&0&0\\-1&-4&-1\end{matrix}\frac{v}{|v|}\right]|v|a$$

$$+\left[\frac{1}{6}\begin{matrix}-1&2&-1\\-4&8&-4\\-1&2&-1\end{matrix}\frac{u}{|v|}\frac{u}{|v|} + \frac{1}{2}\begin{matrix}1&0&-1\\0&0&0\\-1&0&1\end{matrix}\frac{u}{|v|}\frac{v}{|v|} + \frac{1}{6}\begin{matrix}-1&-4&-1\\2&8&2\\-1&-4&-1\end{matrix}\frac{v}{|v|}\frac{v}{|v|}\right]D^c = 0$$

$$-D\frac{\partial^2\phi}{\partial x^2} - D\frac{\partial^2\phi}{\partial y^2} + u\frac{\partial\phi}{\partial x} + v\frac{\partial\phi}{\partial y} - D^c\left(\frac{u}{|v|}\frac{u}{|v|}\frac{\partial^2\phi}{\partial x^2} + 2\frac{u}{|v|}\frac{v}{|v|}\frac{\partial^2\phi}{\partial x\partial y} + \frac{v}{|v|}\frac{v}{|v|}\frac{\partial^2\phi}{\partial y^2}\right) = 0$$

**Figure (c)**

It is first easily seen from the mathematical molecule of Figure (c) that in the cases $\phi = 1$, $\phi = x$ with $f = u$, $\phi = y$ with $f = v$, that the patch test is satisfied for any value of $D^c$ (when $f \neq 0$, a corresponding term $b_5$ must be naturally included in equation (c)) as demanded by convergence according to Section 4.1.



The reference solution (24) is

$$\phi(x,y) = e^{\bar{u}(\cos\psi\cdot x + \sin\psi\cdot y)/\bar{D}} \qquad \text{(e)}$$

For this solution the source term $f = 0$. The coordinates of the nodes in the patch in Figure (b) are easy to read and we obtain thus the nodal values

$$\phi_1 = \phi(-a,-a) = e^{\bar{u}(-\cos\psi\cdot a - \sin\psi\cdot a)/\bar{D}}$$

$$\phi_2 = \phi(0,-a) = e^{\bar{u}(-\sin\psi\cdot a)/\bar{D}} \qquad \text{(f)}$$

...

Using Mathematica, these are substituted in (c) and the damping diffusivity $D^c$ is determined from the resulting equation. For a given velocity direction $\theta$, all this is repeated for a number of values of the cylindrical solution direction $\psi$.

We show here some results in three cases. In case (1) the flow velocity is in the positive $x$-axis direction: $\theta = 0$. In case (2) $\theta = 15°$ and in case (3) the flow velocity is in the diagonal direction: $\theta = 45°$. The data has been selected so that the Peclet number $|v|a/D = 5$.

Figures (d), (e) and (f) show the distribution of $D^c/D$. for the cases (1), (2) and (3), respectively. In more detail, the length of the segment from the origin to the curves gives $D^c/D$ in the assumed cylindrical solution direction.

The results obtained in cases (1) and (2) are in contradiction with the discussion in connection with the continuum case where it was speculated that no damping diffusion is needed when the gradient direction is perpendicular to the flow velocity direction. In fact, experimentation with several flow directions gave similar type of behavior as shown in Figures (d) and (e). Thus actually, the damping diffusivity is usually higher when $\nabla\phi \perp v$ than when $\nabla\phi \parallel v$. Case (3) happens to "obey the theory". A heuristic explanation for the at the first sight odd behavior could be that the mesh introduces some kind of anisotropy into the discrete model not present in the (here isotropic) continuum.

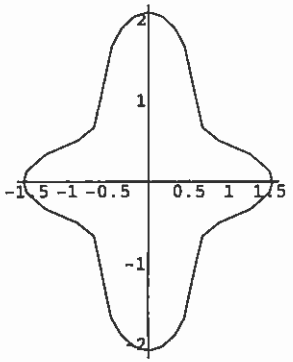


**Figure (d)** ($\theta = 0$)

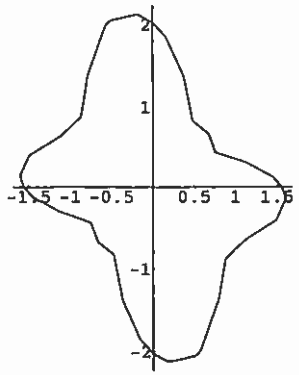



**Figure (e)** ($\theta = 15^\circ$)

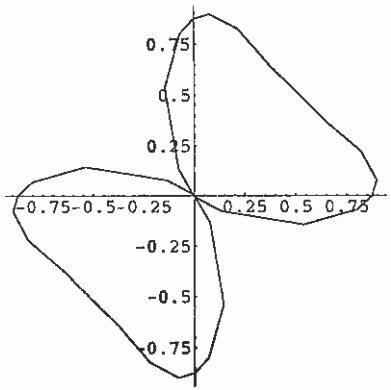



**Figure (f)** ($\theta = 45^\circ$)

It is realized from the figures — as is seen on the basis of Remark 6.14 — that the distributions are symmetrical with respect to the origin, that is, $D^c$ is the same for $\psi$ and for $\psi + \pi$.

### 6.3.3 Triangular elements

Figure 6.11 (a) shows a generic triangular element and Figure 6.11 (b) the corresponding cloned patch.



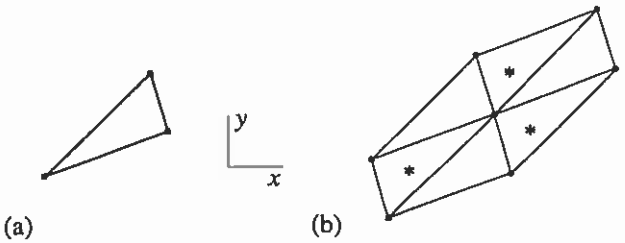
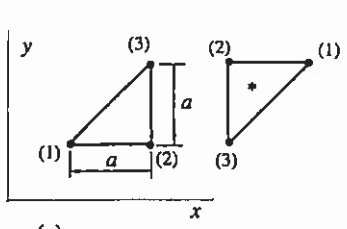



**Figure 6.11 (a)** Triangular element. **(b)** Sensitizing patch.

The three elements (without the "star" mark) are clones of the original element. Each of the three nodes of the element is connected to the central patch node so the situation is in this sense impartial with respect to the element nodes. However, large gaps remain in the patch. To fill the gaps we employ in lack of anything better the original element rotated first 180 degrees three times. The rotated elements (with the star mark) are now not quite correct clones of the original element as they have a different orientation. When forming the system equations (equation) associated with the central node we have roughly two choices. First, we can equip also the rotated elements with the same unknown parameter values (value) as assumed for the original element. The parameter values are then determined from the system equations associated with the central node and obtained using the reference solutions. Second, we may assume from the beginning that the total contribution to the system equations from the rotated elements is approximately the same as from the original cloned elements. Thus by multiplying by two the contributions from only the cloned elements and equating this to zero we get the approximate system equations. Finally, dividing these system equations by two, we realize that we can form the final system equations for the patch equally well just by using the cloned elements. Numerical experiments performed in Example 6.4 showed that the second alternative speculated on did not work well and produced ???. Using the first alternative means that the same sensitizing parameter value a element

**Example 6.4.** We continue with similar experiments as recorded in Example 6.3 now with a simple right-angled equilateral triangular element (type 1) shown in Figure (a) (left). The corresponding three element patch formed by elements 1, 2, 3 is named here patch 1 (Figure (b)). The element (type 2) is obtained from the original element by a rotation of 180 degrees and is shown on right in Figure (b) and equipped with a star for easy recognition. The corresponding three element patch formed by the elements 4, 5, 6 is named here patch 2 (Figure (b)). Finally, a six element "full" patch consisting of elements 1,2,3, 4, 5, 6 and named patch 3 is considered. The local nodal numbering to be used has been indicated only in Figure (a) for the two element types to avoid the patches becoming too filled with data. We perform some sensitizing patch calculations with these configurations to get some information about of the damping diffusivity behavior.



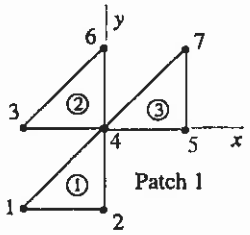
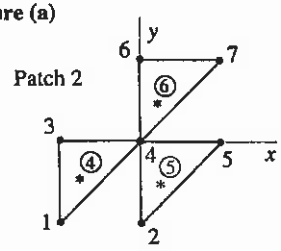




**Figure (a)**



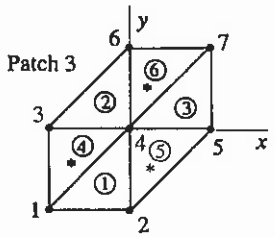




**Figure (b)**

The element contribution expressions are repeated here for completeness from Example 6.3:

$$K_{ij}^e = D\int_{\Omega^e}\left(\frac{\partial N_i}{\partial x}\frac{\partial N_j}{\partial x}+\frac{\partial N_i}{\partial y}\frac{\partial N_j}{\partial y}\right)\mathrm{d}\Omega + u\int_{\Omega^e} N_i\frac{\partial N_j}{\partial x}\mathrm{d}\Omega + v\int_{\Omega^e} N_i\frac{\partial N_j}{\partial y}\mathrm{d}\Omega$$
$$+ D^c\frac{u}{|\mathbf{v}|}\frac{u}{|\mathbf{v}|}\int_{\Omega^e}\frac{\partial N_i}{\partial x}\frac{\partial N_j}{\partial x}\mathrm{d}\Omega + D^c\frac{u}{|\mathbf{v}|}\frac{v}{|\mathbf{v}|}\int_{\Omega^e}\frac{\partial N_i}{\partial x}\frac{\partial N_j}{\partial y}\mathrm{d}\Omega$$
$$+ D^c\frac{v}{|\mathbf{v}|}\frac{u}{|\mathbf{v}|}\int_{\Omega^e}\frac{\partial N_i}{\partial y}\frac{\partial N_j}{\partial x}\mathrm{d}\Omega + D^c\frac{v}{|\mathbf{v}|}\frac{v}{|\mathbf{v}|}\int_{\Omega^e}\frac{\partial N_i}{\partial y}\frac{\partial N_j}{\partial y}\mathrm{d}\Omega \qquad (a)$$
$$b_i^e = \int_{\Omega^e} N_i f\,\mathrm{d}\Omega + \frac{D^c}{|\mathbf{v}|}\frac{u}{|\mathbf{v}|}\int_{\Omega^e}\frac{\partial N_i}{\partial x} f\,\mathrm{d}\Omega + \frac{D^c}{|\mathbf{v}|}\frac{v}{|\mathbf{v}|}\int_{\Omega^e}\frac{\partial N_i}{\partial y} f\,\mathrm{d}\Omega$$

With three-noded triangular elements it is again not necessary to make use of isoparametric mappings etc. as we can find closed form expressions directly using formulas (F.2.2). We obtain in detail first

$$K_{ij}^e = D\frac{1}{4A}\left(b_ib_j + c_ic_j\right) + u\frac{1}{6}b_j + v\frac{1}{6}c_j$$
$$+ D^c\frac{1}{4A}\left(\frac{u}{|\mathbf{v}|}\frac{u}{|\mathbf{v}|}b_ib_j + \frac{u}{|\mathbf{v}|}\frac{v}{|\mathbf{v}|}b_ic_j + \frac{v}{|\mathbf{v}|}\frac{u}{|\mathbf{v}|}c_ib_j + \frac{v}{|\mathbf{v}|}\frac{v}{|\mathbf{v}|}c_ic_j\right) \qquad (b)$$
$$b_i^e = \int_A N_i f\,\mathrm{d}A + \frac{D^c}{|\mathbf{v}|}\frac{u}{|\mathbf{v}|}\frac{b_i}{2A}\int_A f\,\mathrm{d}A + \frac{D^c}{|\mathbf{v}|}\frac{v}{|\mathbf{v}|}\frac{c_i}{2A}\int_A f\,\mathrm{d}A$$

If the source term $f$ is constant, we have

$$b_i^e = \frac{1}{3}Af + \frac{D^c}{|\mathbf{v}|}\frac{u}{|\mathbf{v}|}\frac{b_i}{2}f + \frac{D^c}{|\mathbf{v}|}\frac{v}{|\mathbf{v}|}\frac{c_i}{2}f \qquad (c)$$



When applying formulas of (F.2.1), we associate the symbols $k$, $l$, $m$ in Figure F.3 now consecutively with internal nodes 1, 2, 3, and 2, 3, 1 and 3, 1, 2 of an element:

$$b_1 = y_2 - y_3, \quad b_2 = y_3 - y_1, \quad b_3 = y_1 - y_2$$
$$c_1 = x_3 - x_2, \quad c_2 = x_1 - x_3, \quad c_3 = x_2 - x_1 \qquad (d)$$

**Type 1 element.** We consider first the original (type 1) element. The coordinates of the internal nodes 1, 2, 3 of the element are in general ($\bar{x}$ and $\bar{y}$ are arbitrary values)

$$x_1 = \bar{x}, \quad x_2 = \bar{x} + a, \quad x_3 = \bar{x} + a$$
$$y_1 = \bar{y}, \quad y_2 = \bar{y}, \quad y_3 = \bar{y} + a \qquad (e)$$

Formulas (d) give

$$b_1 = -a, \quad b_2 = a, \quad b_3 = 0$$
$$c_1 = 0, \quad c_2 = -a, \quad c_3 = a \qquad (f)$$

The area of the element is (see formula (d), Example E.2 as a check)

$$A = \frac{1}{2}\begin{vmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{vmatrix} = \frac{1}{2}a^2 \qquad (g)$$

We obtain by Mathematica

$$K_{11}^e = \frac{1}{2}D - \frac{1}{6}\frac{u}{|\mathbf{v}|}|\mathbf{v}|a + \frac{1}{2}\frac{u}{|\mathbf{v}|}\frac{u}{|\mathbf{v}|}D^c$$
$$K_{12}^e = -\frac{1}{2}D + \frac{1}{6}\left(-\frac{u}{|\mathbf{v}|}+\frac{v}{|\mathbf{v}|}\right)|\mathbf{v}|a + \frac{1}{2}\left(-\frac{u}{|\mathbf{v}|}\frac{u}{|\mathbf{v}|}+\frac{u}{|\mathbf{v}|}\frac{v}{|\mathbf{v}|}\right)D^c$$
$$K_{13}^e = \frac{1}{6}\frac{v}{|\mathbf{v}|}|\mathbf{v}| - \frac{1}{2}\frac{u}{|\mathbf{v}|}\frac{v}{|\mathbf{v}|}D^c$$
$$K_{21}^e = -\frac{1}{2}D - \frac{1}{6}\frac{u}{|\mathbf{v}|}|\mathbf{v}|a + \frac{1}{2}\left(-\frac{u}{|\mathbf{v}|}\frac{u}{|\mathbf{v}|}+\frac{u}{|\mathbf{v}|}\frac{v}{|\mathbf{v}|}\right)D^c$$
$$K_{22}^e = D + \frac{1}{6}\left(\frac{u}{|\mathbf{v}|}-\frac{v}{|\mathbf{v}|}\right)|\mathbf{v}|a + \frac{1}{2}\left(\frac{u}{|\mathbf{v}|}\frac{u}{|\mathbf{v}|} - 2\frac{u}{|\mathbf{v}|}\frac{v}{|\mathbf{v}|}+\frac{v}{|\mathbf{v}|}\frac{v}{|\mathbf{v}|}\right)D^c \qquad (h)$$
$$K_{23}^e = -\frac{1}{2}D + \frac{1}{6}\frac{v}{|\mathbf{v}|}|\mathbf{v}|a + \frac{1}{2}\left(\frac{u}{|\mathbf{v}|}\frac{v}{|\mathbf{v}|}-\frac{v}{|\mathbf{v}|}\frac{v}{|\mathbf{v}|}\right)D^c$$

$$K^e_{31} = -\frac{1}{6}\frac{u}{|v|}|v| - \frac{1}{2}\frac{u}{|v|}\frac{v}{|v|}D^c$$

$$K^e_{32} = -\frac{1}{2}D + \frac{1}{6}\left(\frac{u}{|v|} - \frac{v}{|v|}\right)|v|a + \frac{1}{2}\left(\frac{u}{|v|}\frac{v}{|v|} - \frac{v}{|v|}\frac{v}{|v|}\right)D^c$$

$$K^e_{33} = \frac{1}{2}D + \frac{1}{6}\frac{v}{|v|}|v|a + \frac{1}{2}\frac{v}{|v|}\frac{v}{|v|}D^c$$

**Type 2 element.** We consider next the rotated (type 2) element. The coordinates of the internal nodes 1, 2, 3 of the element are in general ($\overline{x}$ and $\overline{y}$ are arbitrary values)

$$\begin{aligned} x_1 &= \overline{x}, \quad x_2 = \overline{x} - a, \quad x_3 = \overline{x} - a \\ y_1 &= \overline{y}, \quad y_2 = \overline{y}, \quad\quad y_3 = \overline{y} - a \end{aligned} \tag{i}$$

Formulas (d) give (from this on unfinished)

? (j)

The element area is naturally again (g). We equip the element contribution symbols by an overbar to discern them from the contributions of the original element. Mathematica gives

$$\begin{aligned} \overline{K}^e_{11} &= \\ \overline{K}^e_{12} &= \\ \overline{K}^e_{13} &= \\ \\ \overline{K}^e_{21} &= \\ \overline{K}^e_{22} &= \\ \overline{K}^e_{23} & \\ \\ \overline{K}^e_{31} &= \\ \overline{K}^e_{32} &= \\ \overline{K}^e_{33} &= \end{aligned} \tag{k}$$

The system equation for node 4 ( $f = 0$ ) of the mesh in Figure (b) is

$$K_{41}\phi_1 + K_{42}\phi_2 + K_{43}\phi_3 + K_{44}\phi_4 + K_{45}\phi_5 + K_{46}\phi_6 + K_{47}\phi_7 = 0 \tag{l}$$

With **patch 1**:



$$K_{41} = K^1_{31} = -\frac{1}{6}\frac{u}{|v|}|v| - \frac{1}{2}\frac{u}{|v|}\frac{v}{|v|}D^c$$

$$K_{42} = K^1_{32} = -\frac{1}{2}D + \frac{1}{6}\left(\frac{u}{|v|} - \frac{v}{|v|}\right)|v|a + \frac{1}{2}\left(\frac{u}{|v|}\frac{v}{|v|} - \frac{v}{|v|}\frac{v}{|v|}\right)D^c$$

$$K_{43} = K^2_{21} = -\frac{1}{2}D - \frac{1}{6}\frac{u}{|v|}|v|a + \frac{1}{2}\left(-\frac{u}{|v|}\frac{u}{|v|} + \frac{u}{|v|}\frac{v}{|v|}\right)D^c$$

$$K_{44} = K^1_{33} + K^2_{22} + K^3_{11} = 2D + \left(\frac{u}{|v|}\frac{u}{|v|} - \frac{u}{|v|}\frac{v}{|v|} + \frac{v}{|v|}\frac{v}{|v|}\right)D^c \tag{j}$$

$$K_{45} = K^3_{12} = -\frac{1}{2}D - \frac{1}{6}\left(\frac{u}{|v|} - \frac{v}{|v|}\right)|v|a + \frac{1}{2}\left(-\frac{u}{|v|}\frac{u}{|v|} + \frac{u}{|v|}\frac{v}{|v|}\right)D^c$$

$$K_{46} = K^2_{23} = -\frac{1}{2}D + \frac{1}{6}\frac{v}{|v|}|v|a + \frac{1}{2}\left(\frac{u}{|v|}\frac{v}{|v|} - \frac{v}{|v|}\frac{v}{|v|}\right)D^c$$

$$K_{47} = K^2_{13} = \frac{1}{6}\frac{v}{|v|}|v| - \frac{1}{2}\frac{u}{|v|}\frac{v}{|v|}D^c$$

With **patch 2**:

$$\begin{aligned} K_{41} &= \overline{K}^4_{13} = \\ K_{42} &= \overline{K}^5_{23} = \\ K_{43} &= \overline{K}^4_{12} = \\ K_{44} &= \overline{K}^4_{11} + \overline{K}^5_{22} + \overline{K}^6_{33} = \\ K_{45} &= \overline{K}^5_{21} = \\ K_{46} &= \overline{K}^6_{32} = \\ K_{47} &= \overline{K}^e_{31} = \end{aligned}$$

With **patch 3**:

$$\begin{aligned} K_{41} &= K^1_{31} + \overline{K}^4_{13} = \\ K_{42} &= K^1_{32} + \overline{K}^5_{23} = \\ K_{43} &= K^2_{21} + \overline{K}^4_{12} = \\ \\ K_{44} &= K^1_{33} + K^2_{22} + K^3_{11} + \overline{K}^4_{11} + \overline{K}^5_{22} + \overline{K}^6_{33} = \\ \\ K_{45} &= K^3_{12} + \overline{K}^5_{21} = \\ K_{46} &= K^2_{23} + \overline{K}^6_{32} = \\ K_{47} &= K^2_{13} + \overline{K}^e_{31} = \end{aligned}$$

The coordinates of the nodes in the patches in Figure (b) are

$$x_1=-a,\quad y_1=-a,\quad x_2=0,\quad y_2=-a,\quad x_3=-a,\quad y_3=0$$
$$x_4=0,\quad y_4=0 \qquad ()$$
$$x_5=a,\quad y_5=,\quad x_6=0,\quad y_6=a,\quad x_7=a,\quad y_7=a$$

The standard tests cases $\phi=1$, $\phi=x$ with $f=u$, $\phi=y$ with $f=v$, are found again to be passed for any value of $D^c$ (when $f\neq 0$, a corresponding term $b_5$ must be included) for all the three patches.

The reference solution (24) is

$$\phi(x,y)=e^{\bar{u}(\cos\psi\cdot x+\sin\psi\cdot y)/\bar{D}} \qquad (?)$$

For this solution the source term $f=0$.

The data is selected so that the local Peclet number $|\bar{v}|a/D=5$.

We record some results obtained. We start with the case $\theta=0$, that is, $u=|v|$ and $v=0$. Table (a) gives values of the ratio $D^c/D$ versus the cylindrical solution direction angle $\psi$.

Table (a) $D^c/D$

The odd behavior near the troublesome case $\psi\approx 90^\circ$ for patches 1 and 2 can be understood on the basis of formulas () and (). Exactly at $\psi=90^\circ$, $\phi_1=\phi_2$, $\phi_3=\phi_4=\phi_5$, $\phi_6=\phi_7$ and we see from () and () that the coefficient of $D^c$ in () disappears leading in the limit to a division by zero. With patch 3 no such happens. The results of this simple case definitely and in the rest of this example we thus continue only with patch 3.

Table (b) gives values of the ratio $D^c/D$ versus the cylindrical solution direction angle $\psi$ in the cases $\theta=15^\circ$, $\theta=45^\circ$, $\theta=90^\circ$. The singular cylindrical solution direction angles corresponding to these are, $\psi=105^\circ$, $\psi=135^\circ$, $\psi=180^\circ$. Solutions around these directions are recorded to detect the detailed behavior.

Table (b) $D^c/D$

### 6.3.4 Numerical results

Some numerical results are shown in in the following obtained for the equation

$$-D\frac{\partial^2\phi}{\partial x^2}-D\frac{\partial^2\phi}{\partial y^2}+u\frac{\partial\phi}{\partial x}+v\frac{\partial\phi}{\partial y}=0 \qquad (36)$$



represented in dimensionless form and valid in the $1\times1$ domain of Figure 6.12. Dirichlet boundary conditions are used with zero data except the value one on part $0.2\le y\le 1$ of the side $x=0$. The velocity components $u$ and $v$ are constants. The sides $x=0$ and $y=0$ form the inflow boundaries.

?

**Figure 6 .12** Solution domain and an irregular triangular mesh.

?

**Figure 6.13** Medium convection: $D=10^{-2}$, $u=2$, $v=1$ (left), $u=1$, $v=2$ (right). Quadrilateral elements. On the first row standard Galerkin method solution. On the second row sensitized Galerkin method solution.

The Standard Galerkin method solution shows oscillations already with these moderate convection values.

?

**Figure 6.14** Medium convection: $D=10^{-2}$, $u=2$, $v=1$ (left), $u=1$, $v=2$ (right). Triangular elements with irregular orientation. On the first row standard Galerkin method solution. On the second row sensitized Galerkin method solution.

?

**Figure 6.15** Large convection: $D=10^{-6}$, $u=2$, $v=1$ (left), $u=1$, $v=2$ (right). Quadrilateral elements. On the first row sensitized Galerkin method solution. On the second row sensitized Galerkin method solution with gradient direction correction.

In this case of large convection the exact solution consists nearly of two level surfaces with the values one and zero due to the inflow data (as explained in Section A.3) divided by the streamline starting at $x=0$, $y=0.2$. Thus a strong internal boundary layer and some strong boundary layers are present. The standard Galerkin method results cannot be drawn any more due to the wild oscillations. The gradient direction correction procedure does not change the solutions very much. Rather strong crosswind diffusion (see Remark A.6) is present with this crude mesh.

?

**Figure 6.16** Large convection: $D=10^{-6}$, $u=2$, $v=1$ (left), $u=1$, $v=2$ (right). Triangular elements with regular orientation. On the first row sensitized Galerkin method solution. On the second row sensitized Galerkin method solution with gradient direction correction.

**Remark 6.15.** The numerical results have been obtained using roughly the procedures explained in the theory part of this text. The triangular element details and the gradient direction correction procedure are still under development and experimentation and the results presented should be considered as preliminary. Especially the behaviour with elongated elements must be studied before any reliable conclusions can be drawn. □

## REFERENCES


Brooks, N. and Hughes, T.J. R (1982). Streamline upwind Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations, *Comput. Methods Appl. Mech. Engrg.*, Vol. 32, pp. 199 - 259.

Christie, I., D., F., Griffiths, F., Mitchell, A. R. and Zienkiewicz, O. C. (1976). Finite element methods for second order differential equations with significant first derivatives, *Int. J. num. Meth. Engng.*, Vol. 10, pp. 1389 - 1396.

Courant, R., Isaacson, E. and Rees, M.(1952). On the Solution of Non-Linear Hyperbolic Differential Equations by Finite Differencies, *Comm. Pure Appl. Math.*, Vol. 5, p. 243.

Crandall, S. (1956). *Engineering Analysis*, McGraw-Hill, New York.

Freund, J. (1996): "On Space-Time FEM for Second Order Problems; an Algorithmic Approach," D. Sc. Dissertation.

Freund, J. and Salonen, E.-M. (1995): "Diffusion-Convection-Reaction Problems and the Patch Test," *Finite Elements in Fluids*, Eds. M. Morandi, K. Morgan, J. Perioux, B. A. Schrefler, Universita di Padova, pp. 215 - 224.

Malvern, L. E. (1969). *Introduction to the Mechanics of a Continuous Medium*, Prentice hall, Englewood Cliffs, New Jersey.

Salvadori, M. G. and Baron, M. L (1961). *Numerical Methods in Engineering*, 2nd ed., Prentice Hall, Englewood Cliffs, N. J.

Shakib, F., and Hughes, T. J. R. (1991) A new finite element formulation for computational fluid dynamics: IX. Fourier analysis of space-time Galerkin/least squares algorithms, *Comput. Methods Appl. Mech. Engrg.*, Vol. 87, pp. 35...58.

Ziegler, H. (1983). *An Introduction to Thermomechanics*, 2nd ed., North-Holland, Amsterdam.


## PROBLEMS

# 7 DIFFUSION-REACTION

## 7.1 ONE DIMENSION

Strong reaction produces also unrealistic oscillations attached to internal and boundary layers if the conventional Galerkin method is used although this problem is not so severe as the one with large convection.

### 7.1.1 Standard Galerkin method

We repeat here the beginning of Section 6.2.1 now for the one-dimensional steady diffusion-reaction problem. It is described by the *diffusion-reaction equation* (later D-R equation)

$$R(\phi)\equiv \boxed{\frac{\mathrm{d}}{\mathrm{d}x}\left(-D\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)+c\phi-f=0} \quad \text{in } \Omega=]a,b[ \tag{1}$$

and for example by the boundary conditions

$$\boxed{\phi=\bar{\phi}} \qquad \text{on } \Gamma_{\mathrm{D}}=\{a\} \tag{2}$$

$$\boxed{-D\frac{\mathrm{d}\phi}{\mathrm{d}x}=\bar{j}^{\mathrm{d}}} \quad \text{on } \Gamma_{\mathrm{N}}=\{b\} \tag{3}$$

The convection term is missing and the reaction term $c\phi$ is now present. The standard weak form corresponding to (1), (2) and (3) is

$$\int_\Omega \frac{\mathrm{d}w}{\mathrm{d}x}D\frac{\mathrm{d}\phi}{\mathrm{d}x}\mathrm{d}\Omega+\int_\Omega wc\phi\,\mathrm{d}\Omega-\int_\Omega wf\,\mathrm{d}\Omega+w\bar{j}^{\mathrm{d}}\Big|_{\Gamma_{\mathrm{N}}}=0 \tag{4}$$

Taking again the finite element approximation

$$\tilde{\phi}(x)=\sum_j N_j(x)\phi_j \tag{5}$$

and employing the Galerkin method in (4) gives the system equations

$$[K]\{a\}=\{b\} \tag{6}$$

with





$$\begin{aligned} K_{ij}&=\int_\Omega \frac{\mathrm{d}N_i}{\mathrm{d}x}D\frac{\mathrm{d}N_j}{\mathrm{d}x}\mathrm{d}\Omega+\int_\Omega N_icN_j\,\mathrm{d}\Omega \\ b_i&=\int_\Omega N_if\,\mathrm{d}\Omega-N_i\bar{j}^{\mathrm{d}}\Big|_{\Gamma_{\mathrm{N}}} \end{aligned} \tag{7}$$

The coefficient matrix is thus found to remain symmetric with the inclusion of reaction.

Similarly as in Section 6.2.1 a simple special case

$$-D\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2}+c\phi=0 \qquad \text{in } \Omega=]0,L[ \tag{8}$$

$$\phi(0)=0, \qquad \phi(L)=\bar{\phi} \tag{9}$$

is used to explain certain solution behavior. This is a steady D-R problem with zero source term, constant diffusivity $D$, constant positive sink factor $c$, and Dirichlet boundary conditions.

Equation (8) is a second order linear differential equation with constant coefficients and the exact solution is found to be

$$\phi(x)=\frac{\sinh(\sqrt{\mathrm{Ce}}\,x/L)}{\sinh\sqrt{\mathrm{Ce}}}\bar{\phi} \tag{10}$$

where

$$\mathrm{Ce}=\frac{cL^2}{D} \tag{11}$$

is a global "Ceclet" number (see (A.3.18)). If Ce is small, diffusion dominates and the solution is nearly linear between the values determined by the boundary data (9). If Ce is large, reaction dominates. From (10) or directly according to (A.3.17), the solution tends to zero except at the neighborhood of the right-hand boundary $x=L$ where a boundary layer is developed due to the condition $\phi(L)=\bar{\phi}$.

The weak form (4) simplifies to

$$\int_\Omega\left(\frac{\mathrm{d}w}{\mathrm{d}x}D\frac{\mathrm{d}\phi}{\mathrm{d}x}+wc\phi\right)\mathrm{d}\Omega=0 \tag{12}$$

The discrete equations are obtained correspondingly from

$$\int_\Omega \left( \frac{d\tilde{w}}{dx} D \frac{d\tilde{\phi}}{dx} + \tilde{w} c \tilde{\phi} \right) d\Omega = 0 \tag{13}$$

With large reaction the diffusion term practically disappears compared with the underlined term due to reaction. We would again like to obtain a nodally exact solution. We can draw some conclusions without any actual calculations. Let us consider Figure 7.1. A uniform mesh of two-noded line elements (length $= h$) is used. The nodes and the elements are numbered from left to right. The exact solution is practically zero except for the thin right-hand side boundary layer so the interpolant to the exact solution is essentially non-zero only in the last element (Figure (a)). Figure (b) shows the corresponding residual $c\tilde{\phi}$. Figure (c) shows the weighting function $\tilde{w} = N_{n-1}$ used to generate the system equation corresponding to node $n-1$. These two terms are positive. Thus multiplying them and performing the integration gives a positive left-hand side in (13) and *the equation cannot be satisfied for the assumed interpolant solution*. The Galerkin method must have negative residual in the second from right element to satisfy the discrete equation. What happens is shown in the figure. The Galerkin solution oscillates with diminishing amplitude so the behaviour is unsatisfactory but not so bad as in the convection dominated case. The ratio $2-\sqrt{3} \approx 0.268$ corresponds to the limiting case of an infinite Ceclet number.

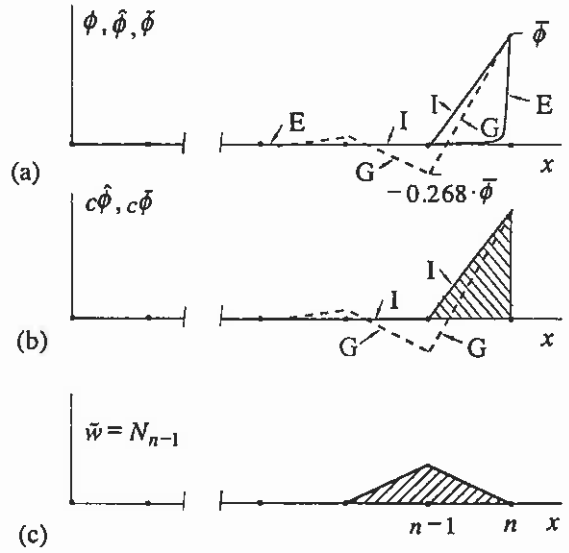





**Figure 7.1** Reaction dominated case, **(a)** Exact solution $\phi$ ($\triangleq$ E), interpolant to the exact solution $\hat{\phi}$ ($\triangleq$ I), Galerkin solution $\tilde{\phi}$ ($\triangleq$ G). **(b)** Residual for the interpolant and for the Galerkin solution. **(c)** Shape function $N_{n-1}$.

As the problem is now self-adjoint, the Galerkin solution is according to Section 4.2.2 the best one in the energy norm, which is here

$$\| u \|_a = a(u,u)^{1/2} = \left[ \int_\Omega \left( \frac{du}{dx} D \frac{du}{dx} + ucu \right) d\Omega \right]^{1/2} \tag{14}$$

From an engineering point of view this result is however not satisfactory since we like to consider the nodally exact solution as the ideal one.

**Remark 7.1.** In the pure reaction case the finite element solution is in fact the least squares fit (weighted by $c$) to the exact solution, that is, the Galerkin method solution minimizes the expression

$$\Pi(\{a\}) = \frac{1}{2} \int_\Omega c \left( \phi - \tilde{\phi} \right)^2 d\Omega \tag{15}$$

with respect to the nodal parameters. The discrete equations corresponding to this condition are

$$F_i \equiv \frac{\partial \Pi}{\partial a_i} = \frac{1}{2} \int_\Omega c\, 2(\phi - \tilde{\phi}) \left( -\frac{\partial \tilde{\phi}}{\partial a_i} \right) d\Omega = -\int_\Omega N_i c \left( \phi - \tilde{\phi} \right) d\Omega = 0 \tag{16}$$

Using the corresponding weak formulation, we have a typical discrete equation

$$\int_\Omega N_i \left( c\tilde{\phi} - f \right) d\Omega = 0 \tag{17}$$

For the exact solution, similarly:

$$\int_\Omega N_i \left( c\phi - f \right) d\Omega = 0 \tag{18}$$

By subtracting these last two equations from each other gives (16) which shows the connection. The behaviour of the Galerkin method solution in Figure 7.1 (a) can be roughly understood also in the light of expression (15): the square of the difference $\phi - \tilde{\phi}$ should be small in an average sense.

Similarly as with (15), it can be shown that in the pure diffusion case the finite element solution is the least squares fit (weighted by $D$) between the derivatives, that is, the Galerkin method solution minimizes the expression

$$\Pi(\{a\}) = \frac{1}{2} \int_\Omega D \left( \frac{d\phi}{dx} - \frac{d\tilde{\phi}}{dx} \right)^2 d\Omega \tag{19}$$

This is one way of explaining why accurate flux values are obtained by the weak formulation. □

### 7.1.2 Sensitized Galerkin method

Appending the least squares weak form of the field equation is readily found not to alleviate the oscillatory behaviour of the standard Galerkin method. As explained in Section D.5.1, we have now to use the gradient least squares weak form. Similarly, as explained in Section 6.2.2, *when sensitizing*, we employ the simplified field equation

$$R^r(\phi) \equiv L^r(\phi) - f \equiv \boxed{-D\frac{d^2\phi}{dx^2} + c\phi - f = 0} \tag{20}$$

and its differentiated form

$$\frac{dR^r(\phi)}{dx} \equiv \frac{dL^r(\phi)}{dx} - \frac{df}{dx} \equiv -D\frac{d^3\phi}{dx^3} + c\frac{d\phi}{dx} - \frac{df}{dx} = 0 \tag{21}$$

where $D$ and $c$ are some local representative constant values. The corresponding sensitized weak form becomes thus

$$\boxed{\begin{aligned}&\int_\Omega \frac{dw}{dx} D\frac{d\phi}{dx} d\Omega + \int_\Omega wc\phi d\Omega - \int_\Omega wf d\Omega + \text{bt} + \\ &+\int_\Omega \frac{dL^r(w)}{dx}\tau^r \frac{dR^r(\phi)}{dx} d\Omega = 0\end{aligned}} \tag{22}$$

where $\tau^r$ is the sensitizing parameter. Written in full, this is

$$\begin{aligned}&\int_\Omega \frac{dw}{dx} D\frac{d\phi}{dx} d\Omega + \int_\Omega wc\phi d\Omega - \int_\Omega wf d\Omega + \text{bt} \\ &+\int_\Omega \left(-D\frac{d^3w}{dx^3} + c\underline{\frac{dw}{dx}}\right)\tau^r\left(-D\frac{d^3\phi}{dx^3} + c\underline{\frac{d\phi}{dx}} - \frac{df}{dx}\right) d\Omega = 0\end{aligned} \tag{23}$$

The underlined terms, multiplied together, give the contribution

$$\frac{dw}{dx}\tau^r c^2 \frac{d\phi}{dx} \tag{24}$$

This explains similarly as in connection with formula (6.2.39) how oscillations can now be damped.



We determine next the reference solutions following Section 5.2.1 or 6.2.2. The governing simplified field equation according to (20) is

$$-D\frac{d^2\phi}{dx^2} + c\phi - f = 0 \tag{25}$$

Its solution is

$$\phi(x) = Ae^{r_1 x} + Be^{r_2 x} + \phi_p(x) \tag{26}$$

where $r_1$ and $r_2$ are the roots

$$r_1 = \sqrt{\frac{c}{D}}, \qquad r_2 = -\sqrt{\frac{c}{D}} \tag{27}$$

of the characteristic equation

$$-Dr^2 + c = 0 \tag{28}$$

and $\phi_p$ is a particular solution for the non-homogeneous equation. The source term has been developed again into a Taylor series

$$f = f_0 + (f_x)_0 x + \frac{1}{2}(f_{xx})_0 x^2 + \cdots \tag{29}$$

and the local origin of $x$ has been taken at the generic point under study.

We obtain in detail

$$\begin{aligned}\phi(x) = &Ae^{\sqrt{c/D}\cdot x} + Be^{-\sqrt{c/D}\cdot x} \\ &+ f_0\frac{1}{c} + (f_x)_0\frac{1}{c}x + (f_{xx})_0\left(\frac{D}{c^2} + \frac{1}{2c}x^2\right) + \cdots\end{aligned} \tag{30}$$

We can write the reference solution thus as

$$\begin{Bmatrix}\phi \\ f\end{Bmatrix} = A\begin{Bmatrix}e^{\sqrt{c/D}\cdot x} \\ 0\end{Bmatrix} + B\begin{Bmatrix}e^{-\sqrt{c/D}\cdot x} \\ 0\end{Bmatrix} + f_0\begin{Bmatrix}1/c \\ 1\end{Bmatrix} + (f_x)_0\begin{Bmatrix}1/c\cdot x \\ x\end{Bmatrix} + \cdots \tag{31}$$

The sensitizing patch test for a two-noded linear element is performed in Example 7.1. It gives the optimal value

$$\boxed{\tau^r = \frac{h^2}{c}\left(\frac{1}{6}\frac{\cosh\sqrt{\mathrm{Ce}_h}+2}{\cosh\sqrt{\mathrm{Ce}_h}-1} - \frac{1}{\mathrm{Ce}_h}\right)} \tag{32}$$

where $\mathrm{Ce}_h$ is an elementwise Ceclet number:

$$\mathrm{Ce}_h = \frac{ch^2}{D} \tag{33}$$

The study performed in Example 7.1 shows that with this sensitizing parameter value nodally exact results are obtained up to a linear source term if the mesh is uniform and the operator data is constant at least with essential boundary conditions. With variable data and mesh, $\tau^r$ is evaluated for each element from (32) using some representative values.

We define a dimensionless sensitizing parameter $\hat{\tau}^r$ by

$$\hat{\tau}^r \equiv \frac{\tau^r}{h^2/c} = \frac{1}{6}\frac{\cosh\sqrt{\mathrm{Ce}_h}+2}{\cosh\sqrt{\mathrm{Ce}_h}-1} - \frac{1}{\mathrm{Ce}_h} \tag{34}$$

Figure 7.2 shows the graph of this parameter.

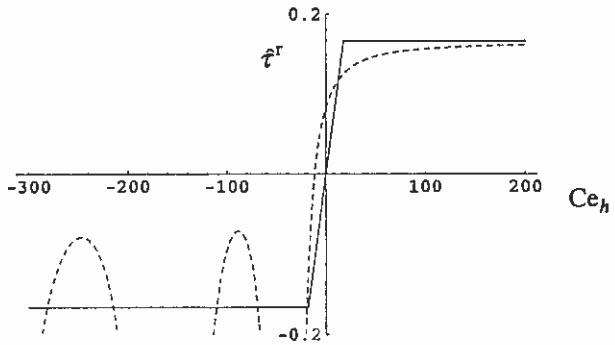



**Figure 7.2** Dimensionless sensitizing parameter $\hat{\tau}^r$ as a function of $\mathrm{Ce}_h$.

It is interesting that the value of $\hat{\tau}^r$ at the origin — meaning the pure diffusion case — does not quite vanish. As in this case we obtain according to Section 4.2.5 exact nodal values for *any* source distribution by the standard Galerkin method, the inclusion of sensitizing for very small values of $\mathrm{Ce}_h$ may thus in fact make the results more inaccurate in general.



Again we note that for a given $c$ and $D$, the sensitizing parameter evaluated from (32) approaches zero when the mesh size goes to zero as in addition $\hat{\tau}^r$ remains finite with vanishing $\mathrm{Ce}_h$. The diffusivity $\tau^r c^2$ in (24) obtains the forms

$$D^r \equiv \tau^r c^2 = \hat{\tau}^r c\, h^2 = \hat{\tau}^r \mathrm{Ce}_h D \tag{35}$$

Similarly as in Chapter 6, it is more illuminating to operate with the *damping diffusivity* (here $D^r$) than with the sensitizing parameter alone. Then we may directly compare the magnitude of $D^r$ with the real diffusivity $D$ as is seen from (35). For instance, for a large $\mathrm{Ce}_h$, we obtain using (37) the magnitude $D^r \approx 1/6 \cdot \mathrm{Ce}_h D$.

**Remark 7.2.** We have this far considered the case where $c$ is positive (we always take $D$ as positive). If $c$ is negative, the D-R equation is of the type familiar say from one degree of freedom linear oscillator: $\phi$ corresponds to displacement, $x$ to time, $D$ to mass, $|c|$ to spring constant. The solution of this type of equation is quite different from what we have considered earlier, consisting of harmonic oscillations in space. A detailed study shows that expression (34) is still valid. As $\mathrm{Ce}_h$ is now negative, the term $\sqrt{\mathrm{Ce}_h}$ is complex and use of well known formulas between hyperbolic and trigonometric functions shows that (34) can be presented in the form

$$\hat{\tau}^r = \frac{1}{6}\frac{\cos\sqrt{|\mathrm{Ce}_h|}+2}{\cos\sqrt{|\mathrm{Ce}_h|}-1} - \frac{1}{\mathrm{Ce}_h} \tag{36}$$

when $\mathrm{Ce}_h$ is negative. Part of the graph of this is shown in Figure 7.2. Use of this expression can give nodally exact solutions by construction but the problem remains that the denominator in the first term becomes zero at $|\mathrm{Ce}_h| = n^2(2\pi)^2$, $(n = 0, 1, \cdots)$. □

A computationally convenient approximation to (34) and very roughly to (36) is

$$\hat{\tau}^r = \begin{cases} \mathrm{Ce}_h/108, & |\mathrm{Ce}_h| \le 18 \\ 1/6 \cdot \mathrm{sgn}\mathrm{Ce}_h, & |\mathrm{Ce}_h| > 18 \end{cases} \tag{37}$$

The approximation is indicated by the dashed line in Figure 7.2. A somewhat different approximation is given in Franca and Dutra do Carmo (1989). As mentioned in Section 5.3.1, the idea of the *Galerkin Gradient Least Squares method* (GGLS-method) taking care of the reaction type term was apparently presented for the first time in this reference.

When the Galerkin method is applied in (23) (using two-noded elements which means that that the third order derivatives vanish both in the weighting and in the residual), we obtain the system equations

$$[K]\{a\}=\{b\} \tag{38}$$

with

$$\begin{aligned} K_{ij} &= \int_\Omega \frac{dN_i}{dx} D \frac{dN_j}{dx} d\Omega + \int_\Omega N_i c N_j \, d\Omega + \int_\Omega \frac{dN_i}{dx} \tau^r c^2 \frac{dN_j}{dx} d\Omega + bt \\ b_i &= \int_\Omega N_i f \, d\Omega + \int_\Omega \frac{dN_i}{dx} \tau^r c \frac{df}{dx} d\Omega + bt \end{aligned} \tag{39}$$

As $\tau^r$ is assumed to be elementwise constant (and $c$ in the sensitizing terms), the element contributions are

$$\begin{aligned} K^e_{ij} &= \int_{\Omega^e} \frac{dN^e_i}{dx} D \frac{dN^e_j}{dx} d\Omega + \int_{\Omega^e} N^e_i c N^e_j d\Omega + \tau^r c^2 \int_{\Omega^e} \frac{dN^e_i}{dx} \frac{dN^e_j}{dx} d\Omega + bt \\ b^e_i &= \int_{\Omega^e} N^e_i f \, d\Omega + \tau^r c \int_{\Omega^e} \frac{dN^e_i}{dx} \frac{df}{dx} d\Omega + bt \end{aligned} \tag{40}$$

Here the upwinding interpretation described in connection with Figure 6.6 is not relevant due to the symmetry of the formulation. The explanation of additional diffusion to damp the wiggles can, however, still be used.

We repeat in the following the counterpart of Example 6.1 to find out the optimum value for $\tau^r$.

**Example 7.1.** We derive the formula for $\tau^r$ using the patch shown in Figure (a).

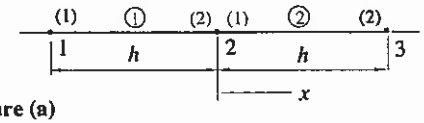



**Figure (a)**

The element contributions are according to (40) (constant data)

$$\begin{aligned} [K]^e &= \frac{D}{h}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \frac{ch}{6}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} + \frac{\tau^r c^2}{h}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \\ \{b\}^e &= \int_{\Omega^e} \begin{Bmatrix} N^e_1 f \\ N^e_2 f \end{Bmatrix} d\Omega + \tau^r c \int_{\Omega^e} \begin{Bmatrix} dN^e_1/dx \cdot df/dx \\ dN^e_2/dx \cdot df/dx \end{Bmatrix} d\Omega \end{aligned} \tag{a}$$

In the system equation for node 2:

$$K_{21}\phi_1 + K_{22}\phi_2 + K_{23}\phi_3 - b_2 = 0 \tag{b}$$

$$K_{21} = K^1_{21} = -\frac{D}{h} + \frac{ch}{6} - \frac{\tau^r c^2}{h}$$



$$K_{22} = K^1_{22} + K^2_{11} = \frac{D}{h} + \frac{2ch}{6} + \frac{\tau^r c^2}{h} + \frac{D}{h} + \frac{2ch}{6} + \frac{\tau^r c^2}{h} = \frac{2D}{h} + \frac{4ch}{6} + \frac{2\tau^r c^2}{h}$$

$$K_{23} = K^2_{12} = -\frac{D}{h} + \frac{ch}{6} - \frac{\tau^r c^2}{h} \tag{c}$$

$$b_2 = b^1_2 + b^2_1 = \int_{\Omega^1} N^1_2 f d\Omega + \int_{\Omega^2} N^2_1 f d\Omega + \tau^r c \int_{\Omega^1} \frac{dN^1_2}{dx}\frac{df}{dx} d\Omega + \tau^r c \int_{\Omega^2} \frac{dN^2_1}{dx}\frac{df}{dx} d\Omega$$

Equation (b) is thus in detail

$$\left(-\frac{D}{h} + \frac{ch}{6} - \frac{\tau^r c^2}{h}\right)\phi_1 + \left(\frac{2D}{h} + \frac{4ch}{6} + \frac{2\tau^r c^2}{h}\right)\phi_2 + \left(-\frac{D}{h} + \frac{ch}{6} - \frac{\tau^r c^2}{h}\right)\phi_3 - b_2 = 0 \tag{d}$$

The first reference solution in (31) ( $A = 1$ ) gives the nodal values

$$\phi_1 = e^{-\sqrt{c/D}\cdot h}, \qquad \phi_2 = 1, \qquad \phi_3 = e^{\sqrt{c/D}\cdot h} \tag{e}$$

with zero source term. Some manipulation of (d) ( $b_2 = 0$ ) gives first

$$\frac{\tau^r c^2}{h}(-\phi_1 + 2\phi_2 - \phi_3) = -\frac{D}{h}(-\phi_1 + 2\phi_2 - \phi_3) - \frac{ch}{6}(\phi_1 + 4\phi_2 + \phi_3) \tag{f}$$

and further

$$\tau^r = \frac{h^2}{6c}\frac{\phi_1 + 4\phi_2 + \phi_3}{\phi_1 - 2\phi_2 + \phi_3} - \frac{D}{c^2} \tag{g}$$

Substitution of the values (e) gives finally

$$\tau^r = \frac{h^2}{6c}\frac{e^{-\sqrt{c/D}\cdot h} + 4 + e^{\sqrt{c/D}\cdot h}}{e^{-\sqrt{c/D}\cdot h}_1 - 2 + e^{\sqrt{c/D}\cdot h}} - \frac{D}{c^2} \tag{h}$$

This can be brought into a cleaner form by using the local Ceclet number

$$Ce_h \equiv C = \frac{ch^2}{D} \tag{i}$$

which produces

$$\begin{aligned} \tau^r &= \frac{h^2}{6c}\frac{e^{-\sqrt{C}} + 4 + e^{\sqrt{C}}}{e^{-\sqrt{C}} - 2 + e^{\sqrt{C}}} - \frac{D}{c^2} = \frac{h^2}{6c}\frac{2\cosh\sqrt{C} + 4}{2\cosh\sqrt{C} - 2} - \frac{D}{c^2} \\ &= \frac{h^2}{c}\left(\frac{1}{6}\frac{\cosh\sqrt{C} + 2}{\cosh\sqrt{C} - 1} - \frac{D}{ch^2}\right) \end{aligned} \tag{j}$$

or

$$\tau^r = \frac{h^2}{c}\left(\frac{1}{6}\frac{\cosh\sqrt{C} + 2}{\cosh\sqrt{C} - 1} - \frac{1}{C}\right) \tag{k}$$

The second reference solution ( $B=1$ ) gives the nodal values

$$\phi_1 = e^{\sqrt{c/D}\cdot h}, \qquad \phi_2 = 1, \qquad \phi_3 = e^{-\sqrt{c/D}\cdot h} \tag{l}$$

which are the nodal values (e) just in the opposite order and the same value for $\tau^r$ is arrived at from the patch test.

The third specific reference solution ( $f_0 = 1$ ) gives the nodal values

$$\phi_1 = 1/c, \qquad \phi_2 = 1/c, \qquad \phi_3 = 1/c \tag{m}$$

and the source term $f=1$. The term

$$b_2 = \int_{\Omega^1} N_2^1 \,d\Omega + \int_{\Omega^2} N_1^2 \,d\Omega + \tau^r c \int_{\Omega^1} \frac{dN_2^1}{dx} 0 \,d\Omega + \tau^r c \int_{\Omega^2} \frac{dN_1^2}{dx} 0 \,d\Omega$$
$$= \frac{h}{2} + \frac{h}{2} = h \tag{n}$$

Equation (d) becomes

$$\left(-\frac{D}{h} + \frac{ch}{6} - \frac{\tau^r c^2}{h}\right)\frac{1}{c} + \left(\frac{2D}{h} + \frac{4ch}{6} + \frac{2\tau^r c^2}{h}\right)\frac{1}{c} + \left(-\frac{D}{h} + \frac{ch}{6} - \frac{\tau^r c^2}{h}\right)\frac{1}{c} - h \overset{?}{=} 0 \tag{o}$$

This is found to be satisfied automatically. Continuing similarly, it is found that even in the case ( $(f_x)_0 = 1$ ) the patch test is passed but no more in the case ( $(f_{xx})_0 = 1$ ). Again, passing of the test in the cases ( $f_0 = 1$ ) and ( $(f_x)_0 = 1$ ) (meaning constant and linear solutions) are seen to correspond to the standard patch test requirement of Section 4.1 for achieving convergenge.

Some numerical results are shown in Figure 7.3 for a problem

$$-\frac{d^2\phi}{dx^2} + 500\phi - 250 = 0 \quad \text{in} \quad \Omega = ]0,1[ \tag{41}$$

$$\phi(0) = 0, \qquad \phi(1) = 1 \tag{42}$$

presented in dimensionless form. The global Ceclet number value is $Ce = 500$ and the elementwise Ceclet number for the regular element mesh elements is $Ce_h = 20$ thus implying according to Figure 7.2 again a rather large difference between the optimal and the approximate $\hat{\tau}^r$.

Figure 7.4 shows the results for a problem

$$-\frac{d^2\phi}{dx^2} - 400\phi = 0 \quad \text{in} \quad \Omega = ]0,1[ \tag{43}$$



$$\phi(0) = 0, \qquad \phi(1) = 1 \tag{44}$$

presented in dimensionless form and having a negative sink factor discussed in Remark 7.2.

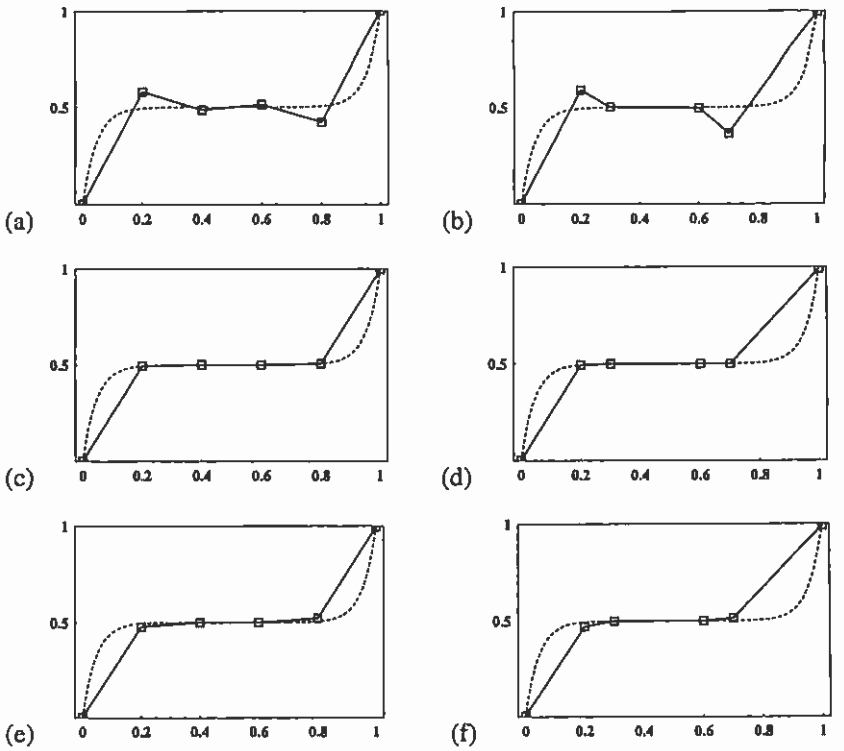



**Figure 7.3** On the left-hand side regular five element mesh, on the right-hand side irregular five element mesh. **(a)** and **(b)** Standard Galerkin method solution. **(c)** and **(d)** Sensitized Galerkin method solution with $\hat{\tau}^r$ according to formula (34). **(e)** and **(f)** Sensitized Galerkin method solution with $\hat{\tau}^r$ according to formula (37).

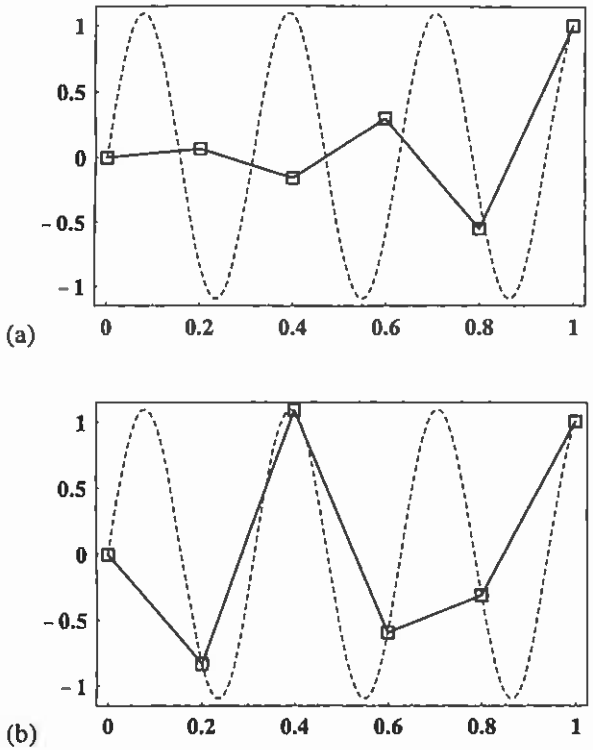



**Figure 7.4** Regular five element mesh. **(a)** Standard Galerkin method solution. **(b)** Sensitized Galerkin method solution with $\hat{\tau}^r$ according to formula (36).

Both solutions are for obvious reasons far from the truth but the sensitized solution again rather miraculously hits on the exact solution at the nodes.

## 7.2 TWO DIMENSIONS (unfinished)

### 7.2.1 Sensitized weak form; general considerations

**Introduction.** The governing field equation is

$$R(\phi)\equiv\boxed{\frac{\partial}{\partial x_\alpha}\left(-D_{\alpha\beta}\frac{\partial\phi}{\partial x_\beta}\right)+c\phi-f=0} \tag{1}$$

in $\Omega$ with approriate Dirichlet, Neumann and Robin boundary conditions as given in Section 6.1.2. Again, for sensitizing purposes we employ the simplified equation

$$R^{\mathrm{r}}(\phi)\equiv L^{\mathrm{r}}(\phi)-f=\boxed{-D_{\alpha\beta}\frac{\partial^2\phi}{\partial x_\alpha\partial x_\beta}+c\phi-f=0} \tag{2}$$

and its differentiated forms

$$\frac{\partial R^{\mathrm{r}}(\phi)}{\partial x_\gamma}\equiv\frac{\partial L^{\mathrm{r}}(\phi)}{\partial x_\gamma}-\frac{\partial f}{\partial x_\gamma}=-D_{\alpha\beta}\frac{\partial^3\phi}{\partial x_\alpha\partial x_\beta\partial x_\gamma}+c\frac{\partial\phi}{\partial x_\gamma}-\frac{\partial f}{\partial x_\gamma}=0 \tag{3}$$

The sensitized weak form is thus

$$\boxed{\begin{aligned}&\int_\Omega\frac{\partial w}{\partial x_\alpha}D_{\alpha\beta}\frac{\partial\phi}{\partial x_\beta}\mathrm{d}\Omega+\int_\Omega wc\phi\,\mathrm{d}\Omega-\int_\Omega wf\,\mathrm{d}\Omega+\mathrm{bt}+\\&+\int_\Omega\frac{\partial L^{\mathrm{r}}(w)}{\partial x_\gamma}\tau^{\mathrm{r}}_{\gamma\delta}\frac{\partial R^{\mathrm{r}}(\phi)}{\partial x_\delta}\mathrm{d}\Omega=0\end{aligned}} \tag{4}$$

The steps needed to obtain (4) should be obvious from earlier derivations and from the discussion in Section D.5.2.

**Remark 7.3.** This comment is the counterpart of Remark 6.11. The sensitizing integrand in (4) is in detail

$$\frac{\partial L^{\mathrm{r}}(w)}{\partial x_\gamma}\tau^{\mathrm{r}}_{\gamma\delta}\frac{\partial R^{\mathrm{r}}(\phi)}{\partial x_\delta}=$$

$$\left(-D_{\varepsilon\lambda}\frac{\partial^3 w}{\partial x_\varepsilon\partial x_\lambda\partial x_\gamma}+c\underline{\frac{\partial w}{\partial x_\gamma}}\right)\tau^{\mathrm{r}}_{\gamma\delta}\left(-D_{\alpha\beta}\frac{\partial^3\phi}{\partial x_\alpha\partial x_\beta\partial x_\delta}+c\underline{\frac{\partial\phi}{\partial x_\delta}}-\frac{\partial f}{\partial x_\delta}\right) \tag{5}$$

The important term from the point of view of reaction is



$$\frac{\partial w}{\partial x_\gamma}\tau^{\mathrm{r}}_{\gamma\delta}c^2\frac{\partial\phi}{\partial x_\delta} \tag{6}$$

or using matrix notation, the term

$$\begin{Bmatrix}\partial w/\partial x\\ \partial w/\partial y\end{Bmatrix}\begin{bmatrix}\tau^{\mathrm{r}}_{xx}c^2 & \tau^{\mathrm{r}}_{xy}c^2\\ \tau^{\mathrm{r}}_{yx}c^2 & \tau^{\mathrm{r}}_{yy}c^2\end{bmatrix}\begin{Bmatrix}\partial\phi/\partial x\\ \partial\phi/\partial y\end{Bmatrix} \tag{7}$$

Comparison with the first integrand in (4) again shows that sensitizing can be interpreted as injection of anisotropic damping diffusion into the formulation. Here, however, no specific physical properties can be easily associated to the sensitizing term. □

**Remark 7.4.** Similarly as commented on in Remark 6.12, in the discrete equations to follow, we always further simplify by neglecting the second or higher order derivatives possibly appearing in the sensitizing terms. □

**Remark 7.5.** Recalling expressions (6) and (7) in Remark 7.3 we will use the following notation for the damping diffusivity tensor

$$D^{\mathrm{r}}_{\alpha\beta}\equiv\tau^{\mathrm{r}}_{\alpha\beta}c^2 \tag{8}$$

and in two dimensions for the damping diffusivity matrix

$$\left[D^{\mathrm{r}}\right]=\begin{bmatrix}D^{\mathrm{r}}_{xx} & D^{\mathrm{r}}_{xy}\\ D^{\mathrm{r}}_{yx} & D^{\mathrm{r}}_{yy}\end{bmatrix}\equiv\begin{bmatrix}\tau^{\mathrm{r}}_{xx}c^2 & \tau^{\mathrm{r}}_{xy}c^2\\ \tau^{\mathrm{r}}_{yx}c^2 & \tau^{\mathrm{r}}_{yy}c^2\end{bmatrix} \tag{9}$$

As the damping diffusivity components are more illuminating than the sensitizing parameter alone, we will introduce the components $D^{\mathrm{r}}_{\alpha\beta}$ from now on to be used in the numerical calculations in connection with the sensitizing patch test. □

The system equations are (see expression (5) and take Remarks 7.3 to 7.5 into account)

$$[K]\{a\}=\{b\} \tag{10}$$

with

$$\begin{aligned}K_{ij}&=\int_\Omega\frac{\partial N_i}{\partial x_\alpha}D_{\alpha\beta}\frac{\partial N_j}{\partial x_\beta}\mathrm{d}\Omega+\int_\Omega N_i cN_j\,\mathrm{d}\Omega\\&\quad+\int_\Omega\frac{\partial N_i}{\partial x_\alpha}D^{\mathrm{r}}_{\alpha\beta}\frac{\partial N_j}{\partial x_\beta}\mathrm{d}\Omega+\mathrm{bt}\\b_i&=\int_\Omega N_i f\,\mathrm{d}\Omega+\int_\Omega\frac{\partial N_i}{\partial x_\alpha}\frac{D^{\mathrm{r}}_{\alpha\beta}}{c}\frac{\partial f}{\partial x_\beta}\mathrm{d}\Omega+\mathrm{bt}\end{aligned} \tag{11}$$

**Reference solutions.** We employ cylindrical solutions similarly as in Section 6.3.1. We make the assumption

$$\phi = \phi(s) \tag{12}$$

with

$$s = e_\alpha x_\alpha = \cos\psi_1 \cdot x_1 + \cos\psi_2 \cdot x_2 \tag{13}$$

Field equation (2) becomes

$$\boxed{-\bar{D}\frac{\mathrm{d}^2\phi}{\mathrm{d}s^2} + c\phi - f = 0} \tag{14}$$

with

$$\begin{aligned}\bar{D} = e_\alpha e_\beta D_{\alpha\beta} &= \cos\psi_1 \cos\psi_1 \cdot D_{11} + \cos\psi_1 \cos\psi_2 \cdot D_{12} \\ &+ \cos\psi_2 \cos\psi_2 \cdot D_{21} + \cos\psi_2 \cos\psi_2 \cdot D_{22}\end{aligned} \tag{15}$$

Equation (14) is exactly of the type we have dealt with in one dimension. Thus the corresponding reference solution is (cf. (7.1.25) and (7.1.31))

$$\exp\sqrt{\frac{c}{\bar{D}}}s = \exp\left[\sqrt{\frac{c}{\bar{D}}}\left(\cos\psi_1 \cdot x_1 + \cos\psi_2 \cdot x_2\right)\right] \tag{16}$$

The situation differs here in nature from the convection case. No directional character can be associated with the reaction term and now in two dimensions four sensitizing parameters have to be determined. We simplify the treatment from this on assuming symmetry both in the real diffusivity and in the damping diffusivity; thus $D_{\alpha\beta} = D_{\beta\alpha}$ and $D^r_{\alpha\beta} = D^r_{\beta\alpha}$. The former relation is the common one due to physics and the latter relation makes it simpler to try to determine appropriate values for the damping tensor components. So in two dimensions we have to determine the three constants $D^r_{xx}$, $D^r_{yy}$, $D^r_{xy} = D^r_{yx}$ for each element. Some numerical experience obtained especially in connection with semianalytical studies performed in Example 7.2 indicate the following. For safety of achieving a solution and to have a reasonable direction independent solution, we should take at least four different cylindrical solution directions in the sensitizing patch test as shown in Figure 7.5. We then obtain four discrete equations and determine the diffusivities using the least squares method for solving the overdetermined system as commented on in Remark 5.12 (see Remark 7.7). The at first sight apparent bias of the directions in the figure is justified by the fact that a direction obtained from a given direction by adding angle $\pi$ (showed by the dashed lines in the figure) gives an equivalent reference solution. Our experience this far has been mainly with isotropic diffusivity. Usually it is then natural for lack of any better knowledge to take one cylindrical direction in a coordinate direction (say direction 1 in Figure 7.5 in the $x$-axis direction). See, however, Remarks 7.6 and 7.7.



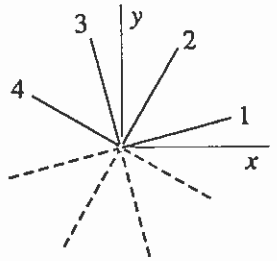



**Figure 7.5** Four cylindrical solution directions following each other with angle difference $\pi/4$.

**Remark 7.6.** For very elongated elements there might be reasons for some preferencial directions to be used in the patch test. For instance, the principal directions of the element inertia tensor connected to the element geometry could have some meaning. Similarly, the principal directions of the real diffusivity tensor in the anisotropic case is another possibility. □

**Remark 7.7.** Although we stated above that the diffusion term has no directional character as such, use of one-dimensional reference solutions introduces some directional effects in connection with the sensitizing patch test. It seems obvious similarly as in diffusion-convection problems that the one-dimensional reference solution (16) is nearest to the actual solution if its direction coincides with the gradient direction of the actual solution. If we are prepared also here to use an iterative procedure, we could put more emphasis on the directions near the gradient direction. One possibility to proceed could be to take the directions differently from that shown in Figure 7.5 so that they form more or less a "fan" around the gradient direction. Another possibility, considered in more detail here, is to weigh the equations obtained in the patch test differently depending on the directions associated with them. Let us consider as an example the case shown in Figure 7.5. We have obtained from the sensitizing patch test the discrete equations

$$lhs1 = 0, \qquad lhs2 = 0, \qquad lhs3 = 0, \qquad lhs4 = 0 \tag{17}$$

where the meaning of the notations is obvious. In the least squares solution method we form the expression

$$E\left(D^r_{xx}, D^r_{yy}, D^r_{xy}\right) = \alpha_1 \left(lhs1\right)^2 + \alpha_2 \left(lhs2\right)^2 + \alpha_3 \left(lhs3\right)^2 + \alpha_4 \left(lhs4\right)^2 \tag{18}$$

and determine the damping diffusivity components from the system

$$\frac{\partial E}{\partial D^{\mathrm{r}}_{xx}}=0,\quad \frac{\partial E}{\partial D^{\mathrm{r}}_{yy}}=0,\quad \frac{\partial E}{\partial D^{\mathrm{r}}_{xy}}=0 \tag{19}$$

The weight factors $\alpha$ are selected so that the gradient direction is favored. If we assume that direction 1 is put in the gradient direction, we may experiment, say, with the selection

$$\alpha_1=100,\quad \alpha_2=\alpha_3=\alpha_4=1 \tag{20}$$ □

**Computational aspects.** Similar points as discussed in Section 6.3.1 need to be taken into account in the numerical determination of the damping diffusivity components using the sensitizing patch test. The two obvious difficult situations are: the reaction term is very weak or it is very large.

We consider first the weak reaction case. We define an element based Ceclet number (cf. (6.3.26))

$$\frac{ch_{\mathrm{m}}^2}{D_{\mathrm{m}}} \tag{21}$$

If this is small, we put damping diffusivities to zero (yksityiskohdat kesken).

We consider next the large reaction case. We evaluate (cf. (6.3.29))

$$\frac{c\bar{h}^2}{\bar{D}} \tag{22}$$

If this is large for any of the four directions, we take a reduced $c=\hat{c}$ and determine the diffusivities $D^{\mathrm{r}}_{xx}=\hat{D}^{\mathrm{r}}_{xx}$, $D^{\mathrm{r}}_{yy}=\hat{D}^{\mathrm{r}}_{yy}$, $D^{\mathrm{r}}_{xy}=\hat{D}^{\mathrm{r}}_{xy}$ using this. Then the final diffusivities to be used are obtained from

$$D^{\mathrm{r}}_{xx}=\frac{c}{\hat{c}}\hat{D}^{\mathrm{r}}_{xx},\quad D^{\mathrm{r}}_{yy}=\frac{c}{\hat{c}}\hat{D}^{\mathrm{r}}_{yy},\quad D^{\mathrm{r}}_{xy}=\frac{c}{\hat{c}}\hat{D}^{\mathrm{r}}_{xy} \tag{23}$$

The logic behind this can be obtained by considering first the one-dimensional case similarly as was explained in Section 6.3.1.

### 7.2.2 Quadrilateral elements

We try to obtain some understanding of the behavior of the damping diffusivity components from Example 7.2. It is a counterpart of Example 6.3.

**Example 7.2.** We consider the case of square element shown in Figure (a). The corresponding sensitizing patch is shown in Figure (b).

With convection missing, the applications might also be in solid mechanics and then the real diffusivity tensor can well be anisotropic. To be ready for such situations in full



generality, we therefore initially write the expressions here assuming the (unusual) general case $D_{xy}\neq D_{yx}$ and also similarly $D^{\mathrm{r}}_{xy}\neq D^{\mathrm{r}}_{yx}$. It may be noted that if $D_{xy}\neq D_{yx}$, the corresponding field equation contains the term

$$-D_{xy}\frac{\partial^2\phi}{\partial x\partial y}-D_{yx}\frac{\partial^2\phi}{\partial y\partial x}=-\left(D_{xy}+D_{yx}\right)\frac{\partial^2\phi}{\partial x\partial y} \tag{a}$$

which is different from

$$-2D_{xy}\left(=-2D_{yx}\right)\frac{\partial^2\phi}{\partial x\partial y} \tag{b}$$

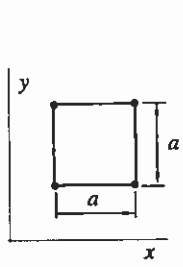



**Figure (a)**

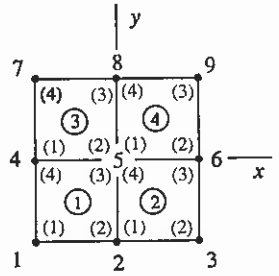



**Figure (b)**

The element contributions are from (11) applied at the element level

$$\begin{aligned}K^e_{ij}=&D_{xx}\int_{\Omega^e}\frac{\partial N_i}{\partial x}\frac{\partial N_j}{\partial x}\,\mathrm{d}\Omega+D_{xy}\int_{\Omega^e}\frac{\partial N_i}{\partial x}\frac{\partial N_j}{\partial y}\,\mathrm{d}\Omega\\&+D_{yx}\int_{\Omega^e}\frac{\partial N_i}{\partial y}\frac{\partial N_j}{\partial x}\,\mathrm{d}\Omega+D_{yy}\int_{\Omega^e}\frac{\partial N_i}{\partial y}\frac{\partial N_j}{\partial y}\,\mathrm{d}\Omega\\&+c\int_{\Omega^e}N_iN_j\,\mathrm{d}\Omega\\&+D^{\mathrm{r}}_{xx}\int_{\Omega^e}\frac{\partial N_i}{\partial x}\frac{\partial N_j}{\partial x}\,\mathrm{d}\Omega+D^{\mathrm{r}}_{xy}\int_{\Omega^e}\frac{\partial N_i}{\partial x}\frac{\partial N_j}{\partial y}\,\mathrm{d}\Omega\\&+D^{\mathrm{r}}_{yx}\int_{\Omega^e}\frac{\partial N_i}{\partial y}\frac{\partial N_j}{\partial x}\,\mathrm{d}\Omega+D^{\mathrm{r}}_{yy}\int_{\Omega^e}\frac{\partial N_i}{\partial y}\frac{\partial N_j}{\partial y}\,\mathrm{d}\Omega\end{aligned} \tag{c}$$

$$\begin{aligned}b^e_i=&\int_{\Omega^e}N_i f\,\mathrm{d}\Omega+\frac{D^{\mathrm{r}}_{xx}}{c}\int_{\Omega^e}\frac{\partial N_i}{\partial x}\frac{\partial f}{\partial x}\,\mathrm{d}\Omega+\frac{D^{\mathrm{r}}_{xy}}{c}\int_{\Omega^e}\frac{\partial N_i}{\partial x}\frac{\partial f}{\partial y}\,\mathrm{d}\Omega\\&+\frac{D^{\mathrm{r}}_{yx}}{c}\int_{\Omega^e}\frac{\partial N_i}{\partial y}\frac{\partial f}{\partial y}\,\mathrm{d}\Omega+\frac{D^{\mathrm{r}}_{yy}}{c}\int_{\Omega^e}\frac{\partial N_i}{\partial y}\frac{\partial f}{\partial y}\,\mathrm{d}\Omega\end{aligned}$$

We obtain in detail (cf. formulas (F.2.3))

$$K_{11}^e = D_{xx}\frac{2}{6} + D_{xy}\frac{1}{4} + D_{yx}\frac{1}{4} + D_{yy}\frac{2}{6} + ca^2\frac{4}{36}$$
$$+ D_{xx}^r\frac{2}{6} + D_{xy}^r\frac{1}{4} + D_{xy}^r\frac{1}{4} + D_{yy}^r\frac{2}{6}$$

$$K_{12}^e = -D_{xx}\frac{2}{6} + D_{xy}\frac{1}{4} - D_{yx}\frac{1}{4} + D_{yy}\frac{1}{6} + ca^2\frac{2}{36}$$
$$- D_{xx}^r\frac{2}{6} + D_{xy}^r\frac{1}{4} - D_{xy}^r\frac{1}{4} + D_{yy}^r\frac{1}{6}$$

$$K_{13}^e = -D_{xx}\frac{1}{6} - D_{xy}\frac{1}{4} - D_{yx}\frac{1}{4} - D_{yy}\frac{1}{6} + ca^2\frac{1}{36}$$
$$- D_{xx}^r\frac{1}{6} - D_{xy}^r\frac{1}{4} + D_{xy}^r\frac{1}{4} - D_{yy}^r\frac{1}{6}$$

$$K_{14}^e = D_{xx}\frac{1}{6} - D_{xy}\frac{1}{4} - D_{yx}\frac{1}{4} - D_{yy}\frac{2}{6} + ca^2\frac{2}{36}$$
$$+ D_{xx}^r\frac{1}{6} - D_{xy}^r\frac{1}{4} - D_{xy}^r\frac{1}{4} - D_{yy}^r\frac{2}{6}$$

$$K_{21}^e = -D_{xx}\frac{2}{6} - D_{xy}\frac{1}{4} + D_{yx}\frac{1}{4} + D_{yy}\frac{1}{6} + ca^2\frac{2}{36}$$
$$- D_{xx}^r\frac{2}{6} - D_{xy}^r\frac{1}{4} + D_{xy}^r\frac{1}{4} + D_{yy}^r\frac{1}{6}$$

$$K_{22}^e = D_{xx}\frac{2}{6} - D_{xy}\frac{1}{4} - D_{yx}\frac{1}{4} + D_{yy}\frac{2}{6} + ca^2\frac{4}{36}$$
$$+ D_{xx}^r\frac{2}{6} - D_{xy}^r\frac{1}{4} - D_{xy}^r\frac{1}{4} + D_{yy}^r\frac{2}{6}$$

$$K_{23}^e = D_{xx}\frac{1}{6} + D_{xy}\frac{1}{4} - D_{yx}\frac{1}{4} - D_{yy}\frac{2}{6} + ca^2\frac{2}{36}$$
$$+ D_{xx}^r\frac{1}{6} + D_{xy}^r\frac{1}{4} - D_{xy}^r\frac{1}{4} - D_{yy}^r\frac{2}{6}$$

$$K_{24}^e = -D_{xx}\frac{1}{6} + D_{xy}\frac{1}{4} + D_{yx}\frac{1}{4} - D_{yy}\frac{1}{6} + ca^2\frac{1}{36}$$
$$- D_{xx}^r\frac{1}{6} + D_{xy}^r\frac{1}{4} + D_{xy}^r\frac{1}{4} - \tau_{yy}^r c^2\frac{1}{6}$$

$$K_{31}^e = -D_{xx}\frac{1}{6} - D_{xy}\frac{1}{4} - D_{yx}\frac{1}{4} - D_{yy}\frac{1}{6} + ca^2\frac{1}{36}$$
$$- D_{xx}^r\frac{1}{6} - D_{xy}^r\frac{1}{4} - D_{yx}^r\frac{1}{4} - D_{yy}^r\frac{1}{6}$$

(d)



$$K_{32}^e = D_{xx}\frac{1}{6} - D_{xy}\frac{1}{4} + D_{yx}\frac{1}{4} - D_{yy}\frac{2}{6} + ca^2\frac{2}{36}$$
$$+ D_{.xx}^r\frac{1}{6} - D_{xy}^r\frac{1}{4} + D_{yx}^r\frac{1}{4} - D_{yy}^r\frac{2}{6}$$

$$K_{33}^e = D_{xx}\frac{2}{6} + D_{xy}\frac{1}{4} + D_{yx}\frac{1}{4} + D_{yy}\frac{2}{6} + ca^2\frac{4}{36}$$
$$+ D_{xx}^r\frac{2}{6} + D_{xy}^r\frac{1}{4} + D_{yx}^r\frac{1}{4} + D_{yy}^r\frac{2}{6}$$

$$K_{34}^e = -D_{xx}\frac{2}{6} + D_{xy}\frac{1}{4} - D_{yx}\frac{1}{4} + D_{yy}\frac{1}{6} + ca^2\frac{2}{36}$$
$$- D_{xx}^r\frac{2}{6} + D_{xy}^r\frac{1}{4} - D_{yx}^r\frac{1}{4} + D_{yy}^r\frac{1}{6}$$

$$K_{41}^e = D_{xx}\frac{1}{6} + D_{xy}\frac{1}{4} - D_{yx}\frac{1}{4} - D_{yy}\frac{2}{6} + ca^2\frac{2}{36}$$
$$+ D_{xx}^r\frac{1}{6} + D_{xy}^r\frac{1}{4} - D_{yx}^r\frac{1}{4} - D_{yy}^r\frac{2}{6}$$

$$K_{42}^e = -D_{xx}\frac{1}{6} + D_{xy}\frac{1}{4} + D_{yx}\frac{1}{4} - D_{yy}\frac{1}{6} + ca^2\frac{1}{36}$$
$$- D_{xx}^r\frac{1}{6} + D_{xy}^r\frac{1}{4} + D_{yx}^r\frac{1}{4} - D_{yy}^r\frac{1}{6}$$

$$K_{43}^e = -D_{xx}\frac{2}{6} - D_{xy}\frac{1}{4} + D_{yx}\frac{1}{4} + D_{yy}\frac{1}{6} + ca^2\frac{2}{36}$$
$$- D_{xx}^r\frac{2}{6} - D_{xy}^r\frac{1}{4} + D_{yx}^r\frac{1}{4} + D_{yy}^r\frac{1}{6}$$

$$K_{44}^e = D_{xx}\frac{2}{6} - D_{xy}\frac{1}{4} - D_{yx}\frac{1}{4} + D_{yy}\frac{2}{6} + ca^2\frac{4}{36}$$
$$+ D_{xx}^r\frac{2}{6} - D_{xy}^r\frac{1}{4} - D_{yx}^r\frac{1}{4} + D_{yy}^r\frac{2}{6}$$

The system equation for node 5 ( $f = 0$ ) from the mesh of Figure (b) is

$$K_{51}\phi_1 + K_{52}\phi_2 + K_{53}\phi_3 + K_{54}\phi_4 + K_{55}\phi_5 + K_{56}\phi_6 + K_{57}\phi_7 + K_{58}\phi_8 + K_{59}\phi_9 = 0 \quad \text{(e)}$$

with

$$K_{51} = K_{31}^1 = -\frac{1}{6}D_{.xx} - \frac{1}{4}D_{xy} - \frac{1}{4}D_{yx} - \frac{1}{6}D_{yy} + \frac{1}{36}ca^2$$
$$-\frac{1}{6}D_{xx}^r - \frac{1}{4}D_{xy}^r - \frac{1}{4}D_{yx}^r - \frac{1}{6}D_{yy}^r$$

$$K_{52} = K_{32}^1 + K_{41}^2 = \frac{2}{6}D_{xx} + 0\cdot D_{xy} + 0\cdot D_{yx} - \frac{4}{6}D_{yy} + \frac{4}{36}ca^2$$
$$+\frac{2}{6}D_{xx}^r + 0\cdot D_{xy}^r + 0\cdot D_{yx}^r - \frac{4}{6}D_{yy}^r$$

$$K_{53} = K_{42}^2 = -\frac{1}{6}D_{xx} + \frac{1}{4}D_{xy} + \frac{1}{4}D_{yx} - \frac{1}{6}D_{yy} + \frac{1}{36}ca^2$$
$$-\frac{1}{6}D_{xx}^r + \frac{1}{4}D_{xy}^r + \frac{1}{4}D_{yx}^r - \frac{1}{6}D_{yy}^r$$

$$K_{54} = K_{34}^1 + K_{21}^3 = -\frac{4}{6}D_{xx} + 0\cdot D_{xy} + 0\cdot D_{yx} + \frac{2}{6}D_{yy} + \frac{4}{36}ca^2$$
$$-\frac{4}{6}D_{xx}^r + 0\cdot D_{xy}^r + 0\cdot D_{yx}^r + \frac{2}{6}D_{yy}^r$$

$$K_{55} = K_{33}^1 + K_{44}^2 + K_{22}^3 + K_{11}^4 = \frac{8}{6}D_{xx} + 0\cdot D_{xy} + 0\cdot D_{yx} + \frac{8}{6}D_{yy} + \frac{16}{36}ca^2$$
$$+\frac{8}{6}D_{xx}^r + 0\cdot D_{xy}^r + 0\cdot D_{yx}^r + \frac{8}{6}D_{yy}^r$$

$$K_{56} = K_{43}^2 + K_{12}^4 = -\frac{4}{6}D_{xx} + 0\cdot D_{xy} + 0\cdot D_{yx} + \frac{2}{6}D_{yy} + \frac{4}{36}ca^2 \qquad (f)$$
$$-\frac{4}{6}D_{xx}^r + 0\cdot D_{xy}^r + 0\cdot D_{yx}^r + \frac{2}{6}D_{yy}^r$$

$$K_{57} = K_{24}^3 = -\frac{1}{6}D_{xx} + \frac{1}{4}D_{xy} + \frac{1}{4}D_{yx} - \frac{1}{6}D_{yy} + \frac{1}{36}ca^2$$
$$-\frac{1}{6}D_{xx}^r + \frac{1}{4}D_{xy}^r + \frac{1}{4}D_{yx}^r - \frac{1}{6}D_{yy}^r$$

$$K_{58} = K_{23}^3 + K_{14}^4 = \frac{2}{6}D_{xx} + 0\cdot D_{xy} + 0\cdot D_{yx} - \frac{4}{6}D_{yy} + \frac{4}{36}ca^2$$
$$+\frac{2}{6}D_{xx}^r + 0\cdot D_{xy}^r + 0\cdot D_{yx}^r - \frac{4}{6}D_{yy}^r$$

$$K_{59} = K_{13}^4 = -\frac{1}{6}D_{xx} - \frac{1}{4}D_{xy} - \frac{1}{4}D_{yx} - \frac{1}{6}D_{yy} + \frac{1}{36}ca^2$$
$$-\frac{1}{6}D_{xx}^r - \frac{1}{4}D_{xy}^r - \frac{1}{4}D_{yx}^r - \frac{1}{6}D_{yy}^r$$

Equation (e) is presented in Figure (c) using mathematical molecules. The corresponding field equation is also given.



$$\frac{1}{6}\begin{matrix}-1 & 2 & -1\\ -4 & 8 & -4\\ -1 & 2 & -1\end{matrix} D_{xx} + \frac{1}{4}\begin{matrix}1 & 0 & -1\\ 0 & 0 & 0\\ -1 & 0 & 1\end{matrix} D_{xy}$$
$$+\frac{1}{4}\begin{matrix}1 & 0 & -1\\ 0 & 0 & 0\\ -1 & 0 & 1\end{matrix} D_{yx} + \frac{1}{6}\begin{matrix}-1 & -4 & -1\\ 2 & 8 & 2\\ -1 & -4 & -1\end{matrix} D_{yy} + \frac{1}{36}\begin{matrix}1 & 4 & 1\\ 4 & 16 & 4\\ 1 & 4 & 1\end{matrix} ca^2$$
$$+\frac{1}{6}\begin{matrix}-1 & 2 & -1\\ -4 & 8 & -4\\ -1 & 2 & -1\end{matrix} D_{xx}^r + \frac{1}{4}\begin{matrix}1 & 0 & -1\\ 0 & 0 & 0\\ -1 & 0 & 1\end{matrix} D_{xy}^r$$
$$+\frac{1}{4}\begin{matrix}1 & 0 & -1\\ 0 & 0 & 0\\ -1 & 0 & 1\end{matrix} D_{yx}^r + \frac{1}{6}\begin{matrix}-1 & -4 & -1\\ 2 & 8 & 2\\ -1 & -4 & -1\end{matrix} D_{yy}^r = 0$$

$$-D_{xx}\frac{\partial^2\phi}{\partial x^2} - D_{xy}\frac{\partial^2\phi}{\partial x\partial y} - D_{yx}\frac{\partial^2\phi}{\partial y\partial x} - D_{yy}\frac{\partial^2\phi}{\partial y^2} + c\phi$$
$$-D_{xx}^r\frac{\partial^2\phi}{\partial x^2} - D_{xy}^r\frac{\partial^2\phi}{\partial x\partial y} - D_{xy}^r\frac{\partial^2\phi}{\partial y\partial x} - D_{yy}^r\frac{\partial^2\phi}{\partial y^2} = 0$$

**Figure (c)**

Again, it is rather easily seen from the mathematical molecules of Figure (c) that in the cases $\phi = 1$ with $f = c$, $\phi = x$ with $f = cx$, $\phi = y$ with $f = cy$, the patch test is satisfied for any values of the damping diffusivity components (when $f \neq 0$, a corresponding term $b_5$ must naturally be included in equation (e)).

From this on we have assumed in general isotropic real diffusivity, that is $D_{xx} = D_{yy} = D$ and $D_{xy} = D_{yx} = 0$.In addition we always take $D_{xy}^r = D_{yx}^r$. The data is selected so that the Peclet number $ca^2/D = 25$. The calculations are performed by Mathematica.

(1) We experiment first taking just two perpendicular directions at a time and assumu correspondingly $D_{xy}^r = D_{yx}^r = 0$ so that we have to determine only the two unknowns: $D_{xx}^r$ and $D_{yy}^r$. Taking first $\psi = 0$ and $\psi = 90^\circ$ gives

$$D^r_{xx} \approx 3.34\,D, \qquad D^r_{yy} \approx 3.34\,D \tag{g}$$

Second, taking $\psi = 45°$ and $\psi = 135°$, we obtain two identical equations and thus no unique solution is found. So here the end result depends strongly on the directions used.

(2) We take three directions thus obtaining three equations to determine $D^r_{xx}$, $D^r_{yy}$, $D^r_{xy}$. Taking first $\psi = 0$, $\psi = 60°$, $\psi = 120°$ gives

$$D^r_{xx} \approx 3.34\,D, \qquad D^r_{yy} \approx 0.55\,D, \qquad D^r_{xy} = 0 \tag{h}$$

Second, taking $\psi = 45°$, $\psi = 105°$, $\psi = 165°$ gives

$$D^r_{xx} \approx 2.06\,D, \qquad D^r_{yy} \approx 2.06\,D, \qquad D^r_{xy} = -0.41\,D \tag{i}$$

These results are still not satisfactory as the values are seen again to depend strongly on the directions used.

(3) We take according to Figure 7.5 four directions thus obtaining four equations to determine $D^r_{xx}$, $D^r_{yy}$, $D^r_{xy}$. The equation system is now ovedetermined and it is solved by least squares using (17) to (19) with equal weight factors. The solution is obtained conveniently by Mathematica by a minimizing command. Taking first $\psi = 0$, $\psi = 45°$, $\psi = 90°$, $\psi = 135°$ gives

$$D^r_{xx} \approx 1.68\,D, \qquad D^r_{yy} \approx 1.68\,D, \qquad D^r_{xy} = 0 \tag{j}$$

Second, taking taking $\psi = 22{,}5°$, $\psi = 67{,}5°$, $\psi = 112{,}5°$, $\psi = 157{,}5°$ gives

$$D^r_{xx} \approx 1.99\,D, \qquad D^r_{yy} \approx 1.99\,D, \qquad D^r_{xy} = 0 \tag{k}$$

Third, taking taking $\psi = 45°$, $\psi = 90°$, $\psi = 135°$, $\psi = 180°$ gives again (j) as was to be expected due to symmetry. As results (j) and (k) are roughly of the same magnitude we conclude that the procedure suggested in connection with Figure 7.5 is reasonable at least as a working hypothesis.

(4) We repeat case (3) now with the weightings (20). Taking first $\psi_1 = 0$, $\psi_2 = 45°$, $\psi_3 = 90°$, $\psi_4 = 135°$ gives

$$D^r_{xx} \approx 3.31\,D, \qquad D^r_{yy} \approx 0.38\,D, \qquad D^r_{xy} = 0 \tag{l}$$

Second, taking $\psi_1 = 22{,}5°$, $\psi_2 = 67{,}5°$, $\psi_3 = 112{,}5°$, $\psi_4 = 157{,}5°$ gives

$$D^r_{xx} \approx 1.99\,D, \qquad D^r_{yy} \approx 1.99\,D, \qquad D^r_{xy} = 0 \tag{m}$$



This is the same as in (k). Here it happens that all the four equations are satisfied by (m) so different weightings do not change the solution. Third, taking $\psi_1 = 45°$, $\psi_2 = 90°$, $\psi_3 = 135°$, $\psi_4 = 180°$ gives

$$D^r_{xx} \approx 1.58\,D, \qquad D^r_{yy} \approx 1.58\,D, \qquad D^r_{xy} = -0.8\,D \tag{n}$$

The weighting does thus in general have an effect on the solution.

(5) In this final case we experiment with an isotropic case and with four directions without different weightings in least squares. The data is $D_{xx} = D$, $D_{yy} = 0.5\,D$, $D_{xy} = D_{yx} = 0.25\,D$ and the Peclet number $ca^2 / D_{xx} = 25$. We take the same directions as in cases (3) and (4) above. First, directions $\psi = 0$, $\psi = 45°$, $\psi = 90°$, $\psi = 135°$ give

$$D^r_{xx} \approx -0.05\,D, \qquad D^r_{yy} \approx 3.64\,D, \qquad D^r_{xy} = 0.00\,D \tag{o}$$

Second, directions $\psi_1 = 22{,}5°$, $\psi_2 = 67{,}5°$, $\psi_3 = 112{,}5°$, $\psi_4 = 157{,}5°$ give

$$D^r_{xx} \approx 2.53\,D, \qquad D^r_{yy} \approx 1.43\,D, \qquad D^r_{xy} \approx 0.07\,D \tag{p}$$

Third, directions $\psi = 45°$, $\psi = 90°$, $\psi = 135°$, $\psi = 180°$ give again (o).

?

### 7.2.3 Triangular elements.

?

### 7.2.4 Numerical results

Some numerical results are shown in Figure 7.6 for the equation

$$-10^{-6}\frac{\partial^2 \phi}{\partial x^2} - 10^{-6}\frac{\partial^2 \phi}{\partial y^2} + \phi - f = 0 \tag{24}$$

with

$$\begin{aligned} f &= 1, \quad y > 0.2 + 0.5x \\ f &= 0, \quad y < 0.2 + 0.5x \end{aligned} \tag{25}$$

?

**Figure 7.5.** Large reaction. Quadrilateral elements (left). Triangular elements (right). On the first row standard Galerkin method solution. On the second row sensitized Galerkin method solution. On the third row sensitized Galerkin method solution with gradient direction correction.

The domain and the boundary conditions are the same as in the numerical examples of Section 6.3.4. In this large reaction case the exact solution behavior can be understood again from the discussion in Section A.3. Sensitizing is clearly seen to improve the solution behavior.

## REFERENCE

Franca, L. P., and Dutra Do Carmo, E. D. (1989). The Galerkin Gradient Least-Squares Method, *Comput. Methods Appl. Mech. Engrg.*, Vol. 74, pp. 41 - 54.

## PROBLEMS

# 8 DIFFUSION-CONVECTION-REACTION

The sensitized weak forms needed in connection with complete D-C-R problems are obvious from the preceding chapters. It is enough here to write down the relevant expressions and to give just some comments on the determination of the sensitizing parameter values.

## 8.1 ONE DIMENSION

The complete governing field equation is now

$$R(\phi)\equiv \boxed{\frac{\mathrm{d}}{\mathrm{d}x}\left(-D\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)+\frac{\mathrm{d}}{\mathrm{d}x}(u\phi)+c\phi-f=0} \tag{1}$$

and the equation forms used for sensitizing are

$$R^{\mathrm{cr}}(\phi)\equiv L^{\mathrm{cr}}(\phi)-f\equiv \boxed{-D\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2}+u\frac{\mathrm{d}\phi}{\mathrm{d}x}+c\phi-f=0} \tag{2}$$

and

$$\frac{\mathrm{d}R^{\mathrm{cr}}(\phi)}{\mathrm{d}x}\equiv\frac{\mathrm{d}L^{\mathrm{cr}}(\phi)}{\mathrm{d}x}-\frac{\mathrm{d}f}{\mathrm{d}x}\equiv -D\frac{\mathrm{d}^3\phi}{\mathrm{d}x^3}+u\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2}+c\frac{\mathrm{d}\phi}{\mathrm{d}x}-\frac{\mathrm{d}f}{\mathrm{d}x}=0 \tag{3}$$

The corresponding sensitized weak form is thus

$$\boxed{\begin{aligned}&\int_\Omega \frac{\mathrm{d}w}{\mathrm{d}x}D\frac{\mathrm{d}\phi}{\mathrm{d}x}\mathrm{d}\Omega+\int_\Omega w\frac{\mathrm{d}}{\mathrm{d}x}(u\phi)\mathrm{d}\Omega+\int_\Omega wc\phi\mathrm{d}\Omega-\int_\Omega wf\,\mathrm{d}\Omega+\mathrm{bt}\\&+\int_\Omega L^{\mathrm{cr}}(w)\tau^{\mathrm{c}}R^{\mathrm{cr}}(\phi)\mathrm{d}\Omega+\int_\Omega\frac{\mathrm{d}L^{\mathrm{cr}}(w)}{\mathrm{d}x}\tau^{\mathrm{r}}\frac{\mathrm{d}R^{\mathrm{cr}}(\phi)}{\mathrm{d}x}\mathrm{d}\Omega\end{aligned}} \tag{4}$$

This version might be called *Galerkin / least-squares / gradient least-squares-method* (GLSGLS-method), Harari and Hughes (1994).

The system equations and the element contributions are also obvious from the earlier chapters.

The reference solutions are obtained from the solution of

$$-D\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2}+u\frac{\mathrm{d}\phi}{\mathrm{d}x}+c\phi-f=0 \tag{5}$$





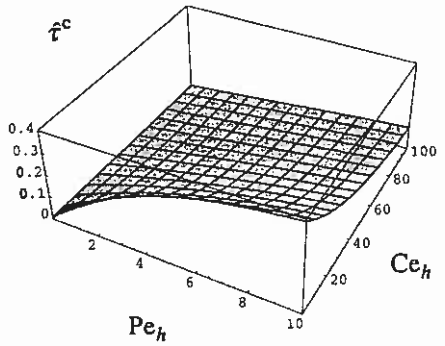
This is similarly as before

$$\phi(x)=Ae^{r_1x}+Be^{r_2x}+\phi_{\mathrm{p}}(x) \tag{6}$$

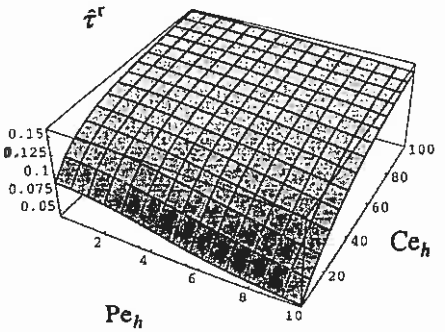
where $r_1$ and $r_2$ are now the roots

$$r_{1,2}=\frac{u\mp\sqrt{u^2+4Dc}}{2D}\ , \tag{7}$$

The one-dimensional patch test to determine $\tau^{\mathrm{c}}$ and $\tau^{\mathrm{r}}$ for a two-noded element described earlier, is too tedious to be performed by hand. General formulas have been generated by the Mathematica program. The resulting graphs for $\hat{\tau}^{\mathrm{c}}$ and $\hat{\tau}^{\mathrm{r}}$ are shown in Figure 8.1. The curves with $\mathrm{Ce}_h=0$ in Figure (a) and with $\mathrm{Pe}_h=0$ in Figure (b) have appeared earlier in Figures 6.5 and 7.2, respectively.




(a)




(b)

**Figure 8.1** (a) Parameter $\hat{\tau}^c$ as a function of $\mathrm{Pe}_h = uh/D$ and $\mathrm{Ce}_h = ch^2/D$. (b) Parameter $\hat{\tau}^r$ as a function of $\mathrm{Pe}_h = uh/D$ and $\mathrm{Ce}_h = ch^2/D$.

**Remark 8.1.** The graphs in Figure 8.1 have been obtained in fact by simplifying the presentation by neglecting the term $cw$ in $L^{cr}(w)$. This is again allowable as discussed in Section 5.3.2. □

Figure 8.2 shows some results obtained for the problem

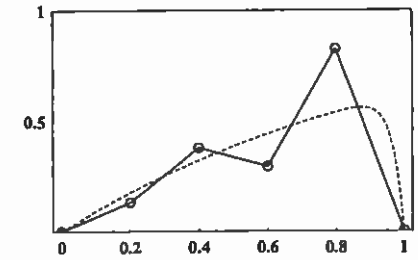
$$-\frac{d^2\phi}{dx^2}+30\frac{d\phi}{dx}+30\phi-30=0 \quad \text{in} \quad \Omega=]0,1[ \tag{8}$$

$$\phi(0)=0, \qquad \phi(1)=0 \tag{9}$$

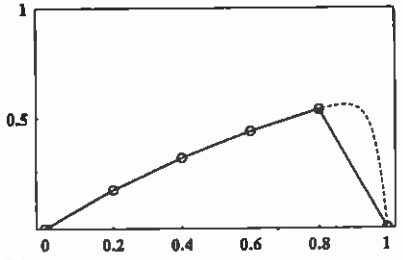
The term "simplified" in the text of Figure 8.2 means that $\hat{\tau}^c$ is evaluated with $\mathrm{Ce}_h = 0$ and $\hat{\tau}^r$ with $\mathrm{Pe}_h = 0$. In other words, *we evaluate $\hat{\tau}^c$ as if there would not be no reaction included and $\hat{\tau}^r$ as if there would be no convection included.* This practical procedure makes the evaluations considerably cheaper and at least in this example the accurary is not much affected.

The term "simplified/simplified" means that in addition to the above simplification, the asymptotic approximations described in Chapters 6 and 7 are employed. This step has more effect on the accuracy than the previous one.

The calculations with an irregular mesh gave quite similar results as with the regular one and they are not reproduced here.

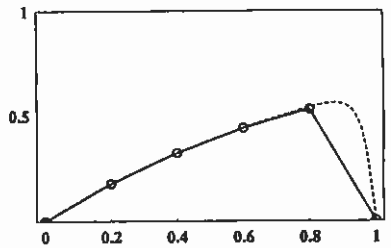


(a)

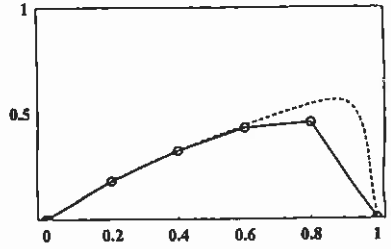


(b)





(c)



(d)

**Figure 8.2** Regular five element mesh. **(a)** Standard Galerkin method solution. **(b)** sensitized Galerkin method solution with the parameters according to optimal values of Figure 8.1. **(c)** Sensitized Galerkin method solution with simplified parameter values. **(d)** Sensitized Galerkin method solution with simplified/simplified parameter values.

.

## 8.2 TWO DIMENSIONS

For completeness, we write down the essential formulas. The differential equation is

$$R(\phi)\equiv \boxed{\frac{\partial}{\partial x_\alpha}\left(-D_{\alpha\beta}\frac{\partial\phi}{\partial x_\beta}\right)+\frac{\partial}{\partial x_\alpha}(v_\alpha\phi)+c\phi-f=0} \tag{1}$$

and the equations for sensitizing purposes are

$$R^{cr}(\phi)\equiv L^{cr}(\phi)-f=\boxed{-D_{\alpha\beta}\frac{\partial^2\phi}{\partial x_\alpha\partial x_\beta}+v_\alpha\frac{\partial\phi}{\partial x_\alpha}+c\phi-f=0} \tag{2}$$

and

$$\frac{\partial R^{cr}(\phi)}{\partial x_\gamma}\equiv\frac{\partial L^{cr}(\phi)}{\partial x_\gamma}-\frac{\partial f}{\partial x_\gamma}=-D_{\alpha\beta}\frac{\partial^3\phi}{\partial x_\alpha\partial x_\beta\partial x_\gamma}+v_\alpha\frac{\partial^2\phi}{\partial x_\alpha\partial x_\gamma}+c\frac{\partial\phi}{\partial x_\gamma}-\frac{\partial f}{\partial x_\gamma}=0 \tag{3}$$

The sensitized is weak form is thus

$$\int_\Omega \frac{\partial w}{\partial x_\alpha} D_{\alpha\beta} \frac{\partial \phi}{\partial x_\beta} d\Omega + \int_\Omega w \frac{\partial}{\partial x_\alpha}(v_\alpha \phi)\, d\Omega$$
$$+ \int_\Omega wc\phi\, d\Omega - \int_\Omega wf\, d\Omega + bt \qquad (4)$$
$$+ \int_\Omega L^{cr}(w)\tau^c R^{cr}(\phi) d\Omega + \int_\Omega \frac{\partial L^r(w)}{\partial x_\gamma} \tau^r_{\gamma\delta} \frac{\partial R^r(\phi)}{\partial x_\delta} d\Omega = 0$$

Based on the experience of the one-dimensional behavior, the parameter values $\tau^c$ and $\tau^r$ are selected according to the simplified or simplified/simplified procedure explained in the previous chapters.

## REFERENCE

Harari, I. and Hughes, T. J. R. (1994). Stabilized Finite Element Methods for Steady Advection-Diffusion with Production, *Comput. Methods Appl. Mech. Engrg.*, Vol. 115, pp. 165 - 191.

## PROBLEMS

# 9 TIME DEPENDENCE

## 9.1 INTRODUCTION

In this chapter some features of the effect of time dependence in connection with the finite element method are discussed. The presentation deals mainly with one-dimensional problems in space.

### 9.1.1 Some notations and a model problem

If a problem is in space one- or two-dimensional and if it depends additionally on time we can have in principle the solution domains indicated in Figure 9.1.

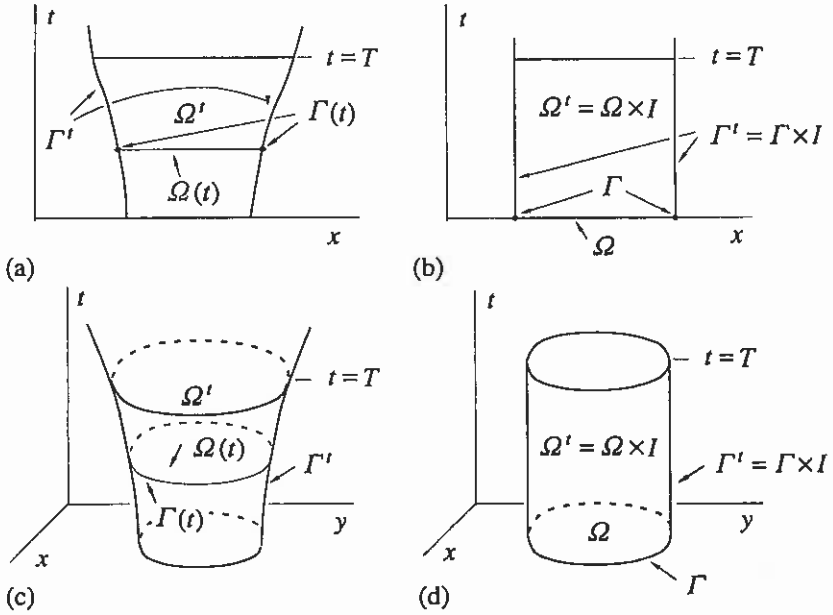



**Figure 9.1** Space-time domains and boundaries and some notations.

The left-hand side figures (a) and (c) describe the general case where the spatial domain $\Omega(t)$ and the spatial boundary $\Gamma(t)$ change with time. Examples could be for instance one-dimensional gas flow in a tube controlled by a moving piston or two-dimensional unsteady fluid flow with a free surface. Quite often the simpler cases of the right-hand side figures (b) and (d) with non-varying spatial domain $\Omega$ and boundary $\Gamma$ appear. We deal in this text with the latter simpler case. In three space dimensions we can no more draw figures like 9.1 but the ideas remain the same.





The origin of the time coordinate can be selected at will and here it is taken as the instant of time we start to consider a certain phenomenon. The time $t = T$ refers to that instant of time we end our study of the phenomenon.

**Remark 9.1.** As discussed in Patankar (1980), time is a *one-way coordinate* (yksisuuntainen koordinaatti). "A one-way coordinate is such that the conditions at a given location in that coordinate are influenced by changes in conditions on only one side of that location." Time clearly satisfies this definition as it is common belief — principle of causality — that only past events and not future ones can influence the present. In some cases also a space coordinate can be a one-way coordinate; conventional approximations used to obtain the boundary layer equations of fluid flow, lead, for example, to such a situation. □

The boundary conditions on the space-time domain boundary at $t = 0$ are usually called *initial conditions* (alkuehto). It is clear from Remark 9.1 that no boundary conditions can be given on the space-time domain boundary at $t = T$.

**Remark 9.2.** A space-time problem could be solved in principle by covering the whole solution domain $\Omega^t$ at once by a mesh and by proceeding similarly as before in space problems. But because time is a one-way coordinate this is not computationally feasible. The space-time domain can be divided by lines or planes or hyperplanes of constant time in two, or three or four dimensions, respectively, into as thin as we want slices or *space-time slabs* (paikka-aika kaistale). The solution is advanced through the first slab using the initial conditions. Then the solution is advanced through the second slab employing the end results from the first slab as the initial conditions, etc. □

To make the presentation concrete, we consider the following model problem. The field equation is the *unsteady one-dimensional diffusion equation*

$$\boxed{\frac{\partial \phi}{\partial t} + \frac{\partial}{\partial x}\left(-D\frac{\partial \phi}{\partial x}\right) - f = 0} \quad \text{in } \Omega^t = \Omega \times I = ]a,b[\times]0,T[ \tag{1}$$

with the boundary conditions

$$\boxed{\phi = \bar{\phi}(t)} \quad \text{on } \Gamma_D^t = \Gamma_D \times I = \{a\}\times]0,T[ \tag{2}$$

$$\boxed{-D\frac{\partial \phi}{\partial x} = \bar{j}^d(t)} \quad \text{on } \Gamma_N^t = \Gamma_N \times I = \{b\}\times]0,T[ \tag{3}$$

and with the initial condition

$$\boxed{\phi = \bar{\phi}_0(x)} \quad \text{in } \Omega = ]a,b[ \text{ at } t = 0 \tag{4}$$

The unknown function to be determined is $\phi(x,t)$ in $\overline{\Omega}^t$.

Figure 9.2 shows the problem in pictorial form. Space-time slabs $S_0$, $S_1$, $\cdots$, $S_n$ are also indicated.

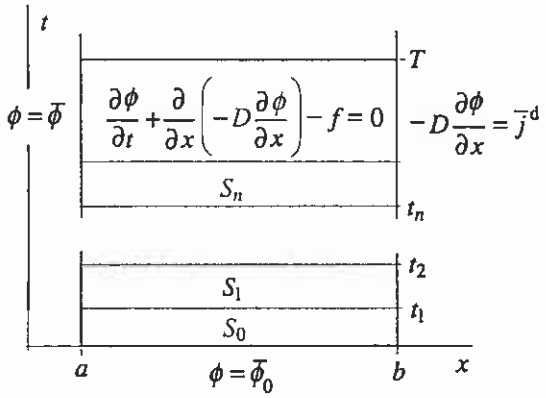



**Figure 9.2** The governing equations and the solution domain.

**Remark 9.3.** In this text we consider cases where only the first order time derivative appears in the field equations. This is typically the case in heat conduction and in fluid mechanics problems in general when the Eulerian description is used. In these situations the initial conditions are of the Dirichlet type: the function value is given. In dynamic solid mechanics problems treated by the Lagrangian description, also the first order time derivative appears in the initial conditions. This is familiar from particle mechanics where the equations of motion demand as initial conditions both the position and velocity of the particle. □

**Remark 9.4.** The model problem is essentially the one-dimensional heat conduction problem of Section 2.1.1 enlarged to the unsteady case. According to (6.1.21), the corresponding heat conduction field equation is

$$\rho c_p\frac{\partial T}{\partial t}+\frac{\partial}{\partial x}\left(-k\frac{\partial T}{\partial x}\right)-s=0 \tag{5}$$

We derive below the finite element formulas employing the model problem (1) $\cdots$ (4) but the changes necessary to consider heat conduction are obvious. □

There are two main ways to deal with time dependent problems; *semidiscretization* or partial discretization (osittainen diskretointi) and *full discretization* (täydellinen diskretointi). These are explained in the following two sections. A method, called *time-discontinuous Galerkin method* (ajan suhteen epäjatkuva Galerkinin menetelmä) has gained recently much popularity. This belongs to the full discretization category but is explained separately in Section 9.3.

### 9.1.2 Semidiscretization



The starting point is the weak form

$$\int_\Omega w(x)\left[\frac{\partial\phi}{\partial t}+\frac{\partial}{\partial x}\left(-D\frac{\partial\phi}{\partial x}\right)-f\right]\mathrm{d}\Omega=0 \tag{6}$$

obtained from the field equation (1) in the usual way. It should be noted that here we select the *weighting function to depend only on* $x$ and integrate only over the space domain (the $x$-axis).

As earlier, the diffusion term is integrated by parts to lower the order of derivatives on $\phi$. The Neumann condition (3) is made use of on the right-hand boundary. On the left-hand boundary we put again $w(a)=0$ and obtain the final weak form

$$\boxed{\int_\Omega\left(w\frac{\partial\phi}{\partial t}+\frac{\partial w}{\partial x}D\frac{\partial\phi}{\partial x}\right)\mathrm{d}\Omega-\int_\Omega wf\,\mathrm{d}\Omega+w\bar{j}^{\mathrm{d}}\Big|_{\Gamma_{\mathrm{N}}}=0} \tag{7}$$

We have assumed that the admissible $\phi$ in (7) satisfies in advance in addition to the Dirichlet boundary condition also the initial condition. Comparison say with (2.1.28) shows that the only change (apart from notation) due to time dependence is the appearance of the term $w\partial\phi/\partial t$ in the integral.

The discrete analogue of (7) is

$$\int_\Omega\left(\tilde{w}\frac{\partial\tilde{\phi}}{\partial t}+\frac{\partial\tilde{w}}{\partial x}D\frac{\partial\tilde{\phi}}{\partial x}\right)\mathrm{d}\Omega-\int_\Omega\tilde{w}f\,\mathrm{d}\Omega+\tilde{w}\bar{j}^{\mathrm{d}}\Big|_{\Gamma_{\mathrm{N}}}=0 \tag{8}$$

The following form of the finite element approximation

$$\boxed{\tilde{\phi}(x,t)=\sum_j N_j(x)\phi_j(t)} \tag{9}$$

is essential in semidiscretization. That is, the given trial basis functions $N_j(x)$ depend only on space coordinates (here on one space coordinate) and the nodal parameters $\phi_j(t)$ are *unknown functions of time* to be determined. The term "semidiscretization" is obvious from the form of (9).

**Remark 9.5.** Approximation (9) resembles the separation of variables method or the product form often used in analytical efforts to solve partial differential equations where representations of the type $\phi(x,t)=F(x)G(t)$ are employed. Here, however, form (9) implies no restriction on the ability to approximate any reasonable continuous function $\phi(x,t)$ with an arbitrary accuracy. To see this, let us consider any fixed value of time. The

approximation in space would be $\sum N_j(x)\phi_j$ where the nodal parameters would be suitable constants and with mesh dense enough an arbitrary accuracy can be achieved at this instant of time. When the instant of time is changed, the approriate values for the nodal parameters also change and we in fact end with the representation (9). □

**Remark 9.6.** In classical mechanics with originally finite degrees of freedom systems — say in connection with rigid body mechanisms — one usually operates with *generalized coordinates* (or with generalized displacements) $q$, see e.g. Lanczos (1970). In dynamics these become functions of time: $q = q(t)$. When solid continuum mechanics problems are discretized by the finite element method using a displacement formulation, the nodal parameters have in fact the role of generalized coordinates. We have had an application of this in statics in connection with the Timoshenko beam in Chapter 5. The nodal deflections and nodal cross-sectional rotations can clearly be interpreted as generalized coordinates associated with the model having now a finite number of degrees of freedom. If we extend the applications to dynamics — which we did not do in Chapter 5 — it is then obviously quite natural just to assume the time dependence to take place in the nodal parameters, that is, to assume a semidiscetization. In heat transfer and fluid mechanics one does not have normally any originally obvious finite degree of systems and the corresponding terminology. Therefore in these areas one may be more ready to think also about the full discretization to be discussed in Section 9.1.3. □

The weak form (7) and its analogue (8) have been generated keeping the partial form (9) in mind. Even if the field equation is valid in a domain in the $x,t$-plane, we strive to "kill" the residual only on a line along the $x$-axis.

From (9),

$$\frac{\partial\phi}{\partial t}\approx\frac{\partial\bar{\phi}}{\partial t}=\sum_j N_j(x)\frac{\mathrm{d}\phi_j(t)}{\mathrm{d}t}\equiv\sum_j N_j(x)\dot{\phi}_j \tag{10}$$

$$\frac{\partial\phi}{\partial x}\approx\frac{\partial\bar{\phi}}{\partial x}=\sum_j\frac{\mathrm{d}N_j(x)}{\mathrm{d}x}\phi_j(t)\equiv\sum_j N_j'(x)\phi_j(t) \tag{11}$$

where some new notation is introduced.

The Galerkin method is now again applied. Here this obviously means that the discrete weighting functions are taken from the set of trial basis functions in space, that is, they are the functions $N_i(x)$. Substituting (10) and (11) into (8) and using the Galerkin method gives the system equations (the reader might go through the details similarly as in Section 2.3.1)

$$\boxed{[M]\{\dot{a}\}+[K]\{a\}=\{b\}} \tag{12}$$

where



$$\begin{aligned}M_{ij}&=\int_\Omega N_iN_j\,\mathrm{d}\Omega\\ K_{ij}&=\int_\Omega\frac{\mathrm{d}N_i}{\mathrm{d}x}D\frac{\mathrm{d}N_j}{\mathrm{d}x}\,\mathrm{d}\Omega\\ b_i&=\int_\Omega N_i f\,\mathrm{d}\Omega-N_i\bar{j}^{\mathrm{d}}\big|_{\Gamma_{\mathrm{N}}}\end{aligned} \tag{13}$$

Equations (12) are a *linear system of first order ordinary differential equations* with time as the independent variable. This set has been referred to in Section 1.1 as formula (1.1.9).

System (12) must be completed with the discrete initial conditions

$$\{a(t)\}=\{a\}^0 \quad \text{at } t=0 \tag{14}$$

The components $\phi_i^0$ of $\{a\}^0$ are obtained simply from the values of function $\bar{\phi}_0(x)$ in (4) at the nodes or alternatively say by determining the spatial finite element least squares fit to $\bar{\phi}_0(x)$.

The coefficient matrix $[M]$ multiplying the column matrix of nodal parameter time derivatives $\{\dot{a}\}$ is sometimes called the *mass matrix* (massamatriisi) as similar structure is obtained say in fluid flow problems in connection of inertia forces. The same type of matrix arises also from the reaction term, see for instance formula (7.1.7). In heat conduction, the mass matrix obtains (see Remark 9.4) a form with entries

$$C_{ij}=\int_\Omega\rho c_p N_iN_j\,\mathrm{d}\Omega \tag{15}$$

The corresponding matrix $[C]$ is sometimes called the *capacity matrix* (kapasiteettimatriisi).

The assembly of the system equations from the element contributions

$$\begin{aligned}M_{ij}^e&=\int_{\Omega^e}N_i^eN_j^e\,\mathrm{d}\Omega\\ K_{ij}^e&=\int_{\Omega^e}\frac{\mathrm{d}N_i^e}{\mathrm{d}x}D\frac{\mathrm{d}N_j^e}{\mathrm{d}x}\,\mathrm{d}\Omega\\ b_i^e&=\int_{\Omega^e}N_i^e f\,\mathrm{d}\Omega-N_i^e\bar{j}^{\mathrm{d}}\big|_{\Gamma_{\mathrm{N}}}\end{aligned} \tag{16}$$

goes similarly as before also with respect to the mass matrix.

The solution of the system equations (12) demands normally a new separate discretization; numerical time integration by some method. This is considered in Section 9.2.

**Remark 9.7.** The formulas needed in time dependent problems are derived sometimes in the literature by basing the derivation on formulas obtained in the corresponding steady cases. For instance, the governing field equation (1) can be generated from the steady field equation

$$\frac{\partial}{\partial x}\left(-D\frac{\partial \phi}{\partial x}\right)-f=0 \tag{17}$$

using the substitution

$$f := f-\frac{\partial \phi}{\partial t} \approx f-\sum_j N_j \dot{\phi}_j \tag{18}$$

where the notation ":= "means: replace the left-hand side expression with the right-hand side expression. Substitution of (18) into the term

$$b_i = \int_\Omega N_i f \,\mathrm{d}\Omega - N_i \bar{j}^{\mathrm{d}}\big|_{\Gamma_{\mathrm{N}}} \tag{19}$$

valid in the steady case, is seen to produce the entries of the mass matrix. The procedure described is familiar in nature from basic dynamics in connection of the inertia force idea: the equation of motion can be obtained from the equilibrium equation by the replacement $\mathbf{F} := \mathbf{F} - m\mathbf{a}$. □

**Example 9.1.** The system equations (12) corresponding to the model problem (1)··· (4) are developed in more detail. We employ a crude mesh of three uniform two-noded line elements and four nodes in the $x$- axis direction shown in Figure (a).

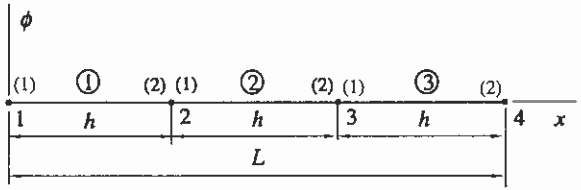



**Figure (a)**

Diffusivity $D$ is assumed to be constant in space and time and $f=0$. The boundary conditions at $a=0$ and also at $b=L$ are taken to be of the Dirichlet type and to be simply

$$\phi(0,t)=\bar{\phi}(0,t)=0, \qquad \phi(L,t)=\bar{\phi}(L,t)=0 \tag{a}$$

This means that at the nodes 1 and 4 "moving" parallel to the $t$- axis,

$$\begin{aligned}\phi_1(t)&=0\\ \phi_4(t)&=0\end{aligned} \tag{b}$$



The element contributions are according to (16)

$$M_{ij}^e = \int_{\Omega^e} N_i^e N_j^e \,\mathrm{d}\Omega, \qquad K_{ij}^e = D\int_{\Omega^e} N_i'^e N_j'^e \,\mathrm{d}\Omega \tag{c}$$

As the elements are identical, we obtain for all of them using formulas (F.1.1) and matrix forms:

$$[M]^e = \frac{h}{6}\begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix}, \qquad [K]^e = \frac{D}{h}\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix} \tag{d}$$

Assembling the system equations and taking into account the given nodal data (b) gives

$$\frac{h}{6}\begin{bmatrix}4 & 1\\ 1 & 4\end{bmatrix}\begin{Bmatrix}\dot{\phi}_2\\ \dot{\phi}_3\end{Bmatrix}+\frac{D}{h}\begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}\begin{Bmatrix}\phi_2\\ \phi_3\end{Bmatrix}=\begin{Bmatrix}0\\ 0\end{Bmatrix} \tag{e}$$

This is a simple illustration of (12).

If the initial condition (4) is for example

$$\phi(x,0)=\bar{\phi}_0=\text{constant} \tag{f}$$

the corresponding initial conditions (14) for the nodal variables are

$$\begin{Bmatrix}\phi_2\\ \phi_3\end{Bmatrix}=\begin{Bmatrix}\bar{\phi}_0\\ \bar{\phi}_0\end{Bmatrix} \quad \text{at } t=0 \tag{g}$$

The analytical solution corresponding to boundary and initial conditions (a) and (f) — obtainable by the separation of variables method — is

$$\phi(x,t)=\sum_{n=1}^{\infty} B_n \sin\frac{n\pi x}{L}\exp(-\lambda_n^2 t) \tag{h}$$

where

$$\lambda_n=\sqrt{D}\,\frac{n\pi}{L} \tag{i}$$

and

$$\begin{aligned}B_n &= \frac{4}{n\pi}\bar{\phi}_0, \quad n \text{ is odd}\\ &= 0, \qquad\quad n \text{ is even}\end{aligned} \tag{j}$$

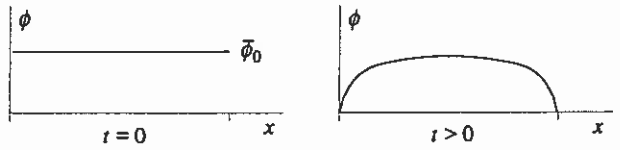



**Figure (b)**

The solution decays monotonicly in time since the exponents $-\lambda_n^2 t$ are negative. The terms in the series where $n$ is large ($\Rightarrow \lambda_n$ is large) decay fastest. Figure (b) shows schematically the distribution of $\phi$ along the $x$-axis at the initial time $t=0$ and at a later time $t>0$. It is realized that the solution is due to the boundary and initial conditions discontinuous at points $(0,0)$ and $(L,0)$. This obviously leads at the beginning near these points to large values of $\partial^2\phi/\partial x^2$ and thus through the field equation (1) to large values of $\partial\phi/\partial t$, that is, the solution must change at the beginning near these points also rapidly with respect to time. These features are demanding on a numerical method.

### 9.1.3 Full discretization

Partial discretization has been the most common way to deal with time dependent problems. Lately it has been replaced, however, more and more by full discretization. The main idea is to conceive time just as one additional space coordinate; but not in every respect as time in any case has the special property of being a one-way coordinate.

We consider again the model problem (1) ⋯ (4) and concentrate on a typical space-time slab $S \equiv S_n = \Omega \times I_n = ]a,b[\times]t_n, t_{n+1}[$ in $\Omega^t$ (Figure 9.2). We take as the starting point the weak form

$$\int_S w(x,t)\left[\frac{\partial\phi}{\partial t}+\frac{\partial}{\partial x}\left(-D\frac{\partial\phi}{\partial x}\right)-f\right]\mathrm{d}S=0 \tag{20a}$$

or using a somewhat different notation, the form

$$\int_\Omega\int_{I_n} w(x,t)\left[\frac{\partial\phi}{\partial t}+\frac{\partial}{\partial x}\left(-D\frac{\partial\phi}{\partial x}\right)-f\right]\mathrm{d}t\mathrm{d}\Omega=0 \tag{20b}$$

This can be compared with the starting point (6) for semidiscretization. It has again been assumed that the admissible $\phi$ satisfies in advance the Dirichlet boundary condition and the initial condition at $t=t_n$ produced from the previous slab.

Integration by parts is applied now in two dimensions for the diffusion term. In detail

$$\int_S w\frac{\partial}{\partial x}\left(-D\frac{\partial\phi}{\partial x}\right)\mathrm{d}S=\int_S \frac{\partial w}{\partial x}D\frac{\partial\phi}{\partial x}\mathrm{d}S-\int_{\partial S} wD\frac{\partial\phi}{\partial x}n_x\mathrm{d}(\partial S) \tag{21}$$



We have applied the first formula (B.2.1a) with $x \triangleq x$, $y \triangleq t$, $g \triangleq w$, $h \triangleq -D\partial\phi/\partial x$ in the space-time slab $S$ with boundary $\partial S$. As seen from Figure 9.3, the component $n_x$ of the outward unit normal vector $\mathbf{n}$ disappears at the sides $t=t_n$ and $t=t_{n+1}$ and has the values $-1$ and $+1$ on the sides $x=a$ and $x=b$, respectively. As $w$ is taken to vanish on the Dirichlet boundary and making use of the given data on the Neumann boundary, we obtain the final weak form

$$\boxed{\int_S\left(w\frac{\partial\phi}{\partial t}+\frac{\partial w}{\partial x}D\frac{\partial\phi}{\partial x}\right)\mathrm{d}S-\int_S wf\,\mathrm{d}S+\int_{(\partial S)_N} w\bar{j}^{\mathrm{d}}\mathrm{d}(\partial S)=0} \tag{22}$$

The notation $(\partial S)_N$ refers here to the right-hand side edge of the slab, that is, to $\{b\}\times I_n$. This weak form looks like form (7), obtained through semi-discretization, but there the domain was in space and here it is in space-time.

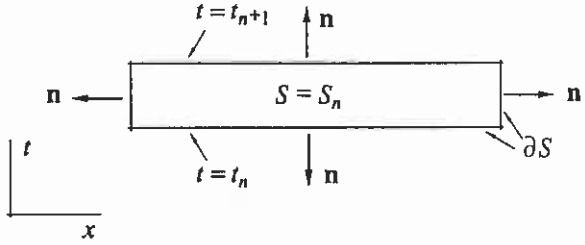



**Figure 9.3** Space-time slab $S$ and its boundary $\partial S$.

The discrete analogue of (22) is

$$\int_S\left(\tilde{w}\frac{\partial\tilde{\phi}}{\partial t}+\frac{\partial\tilde{w}}{\partial x}D\frac{\partial\tilde{\phi}}{\partial x}\right)\mathrm{d}S-\int_S \tilde{w}f\,\mathrm{d}S+\int_{(\partial S)_N}\tilde{w}\bar{j}^{\mathrm{d}}\mathrm{d}(\partial S)=0 \tag{23}$$

Instead of (9), the finite element approximation is

$$\boxed{\tilde{\phi}(x,t)=\sum_j N_j(x,t)\phi_j} \tag{24}$$

The nodal parameters $\phi_j$ are now *unknown constants* and the two-dimensional shape functions $N_j(x,t)$ are at simplest those of triangular or bilinear elements.

Application of the Galerkin method in (23) with $N_i(x,t)$ as weighting functions generates the system equations

$$\boxed{[K]\{a\}=\{b\}} \tag{25}$$

with

$$K_{ij}=\int_S\left(N_i\frac{\partial N_i}{\partial t}+\frac{\partial N_i}{\partial x}D\frac{\partial N_j}{\partial x}\right)\mathrm{d}S$$
$$b_i=\int_S N_i f\,\mathrm{d}S-\int_{(\partial S)_N}N_i\bar{j}^{\mathrm{d}}\mathrm{d}(\partial S) \tag{26}$$

The system equations are now again a linear system of algebraic equations. Figure 9.4 could illustrate the situation.

The length of the time step $\Delta t=t_{n+1}-t_n$ should be so small that one layer of elements is enough to give the accuracy needed. It is seen that many known nodal values (roughly half of the total number) appear in the mesh. After the unknown nodal values have been determined from (25), they appear as given values for the next slab.

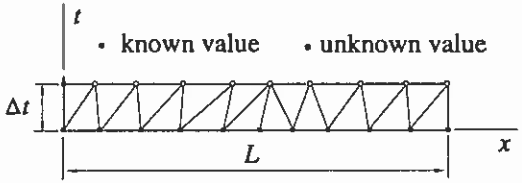



**Figure 9.4** One layer of elements in a slab.

**Remark 9.8.** It can be realized from Figure 9.4 that full discretization contains in principle more flexibility than semidiscretization. As example consider a sharp internal layer advancing with time along the $x$-axis. In space-time this means a skew discontinuity line and if a two element interface is arranged to coincide with this line, good accuracy is to be expected. □

**Remark 9.9.** Full and semidiscretization have been considered above only in the case of one space dimension and without convection and reaction. The extensions of the formulations to more general situations are, however, rather obvious and are not treated here. With three space dimensions full discretization means the use of four-dimensional elements which is difficult to visualize and one must rely on pure mathematical extensions of the steps familiar from the transfer between two and three dimensions. □

**Example 9.2.** We treat the problem of Example 9.1 now with full discretization. The extremely crude mesh used for slab $S_0$ is shown in Figure (a). It consists of three identical four-noded rectangular elements. Figure (b) gives the local nodal numbering of the element.



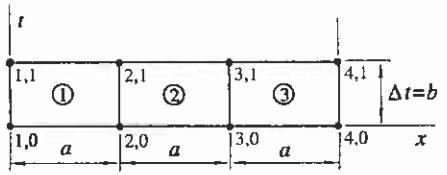



**Figure (a)**

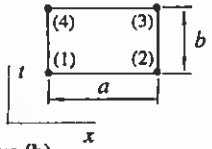



**Figure (b)**

The element contributions are from (26)

$$K_{ij}^e=\int_{S^e}N_i^e\dot{N}_j^e\,\mathrm{d}S+D\int_{S^e}N_i'^eN_j'^e\,\mathrm{d}S \tag{a}$$

This is in matrix form

$$[K]^e=\frac{a}{12}\begin{bmatrix}-2&-1&1&2\\-1&-2&2&1\\-1&-2&2&1\\-2&-1&1&2\end{bmatrix}+\frac{D}{6}\frac{b}{a}\begin{bmatrix}2&-2&-1&1\\-2&2&1&-1\\-1&1&2&-2\\1&-1&-2&2\end{bmatrix} \tag{b}$$

Formulas (F.2.3) have been applied with appropriate interpretations.

The nodes of the mesh of Figure (a) have been numbered here with a double index system much used in the finite difference method so that the first index increases with position and the second with time. The nodal values are denoted similarly in the finite difference fashion as $\phi_j^n$ where the subscript refers to position and the superscript to time.

Assembly gives the system equations

$$\left(\frac{a}{12}\begin{bmatrix}4&1\\1&4\end{bmatrix}+\frac{D}{6}\frac{b}{a}\begin{bmatrix}4&-2\\-2&4\end{bmatrix}\right)\begin{Bmatrix}\phi_2^1\\\phi_3^1\end{Bmatrix}$$
$$+\left(\frac{a}{12}\begin{bmatrix}-4&-1\\-1&-4\end{bmatrix}+\frac{D}{6}\frac{b}{a}\begin{bmatrix}2&-1\\-1&2\end{bmatrix}\right)\begin{Bmatrix}\phi_2^0\\\phi_3^0\end{Bmatrix}=\begin{Bmatrix}0\\0\end{Bmatrix} \tag{c}$$

The information

$$\phi_1^0=\phi_4^0=\phi_1^1=\phi_4^1=0 \tag{d}$$

from the zero boundary conditions have been made use of.

System equations (c) are of form (25). Here we have only grouped the nodal variables in the "past" and "future" sets and taken the boundary conditions already into account. The representation (c) would be in general of the form

$$[A]\{a\}^{n+1}+[B]\{a\}^n+\{C\}=\{0\} \tag{e}$$

where $\{a\}$ contains on each time level only those nodal parametrs which are not prescibed directly from the boundary conditions. The formal solution would thus be

$$\{a\}^{n+1} = -[A]^{-1}\left([B]\{a\}^n + \{C\}\right) \tag{f}$$

This gives the algorithm: value of $\{a\}^0$ is known, we obtain $\{a\}^1$ from (f), etc. A similar situation is considered in more detail in Section 9.2.

In time dependent problems it is important to try define some relevant characteristic time interval to be in general able to say that a time step is small or large. Here we take a reference time

$$t_r = \frac{1}{\lambda_1^2} = \frac{L^2}{D\pi^2} \tag{g}$$

The fysical meaning of this is as follows. The slowest decaying term in formula (h) of Example 9.1 has decreased to the value $e^{-1} \approx 0.37$ times its original value.

The initial condition $\bar{\phi}_0 = \text{constant}$ gives the initial nodal values

$$\phi_2^0 = \phi_3^0 = \bar{\phi}_0 \tag{h}$$

We take the step size $\Delta t = b = 0.5 t_r$. Then the coefficient

$$\frac{Db}{a} = \frac{D \cdot 0.5 (3a)^2}{D\pi^2 a} = \frac{9a}{2\pi^2} \tag{i}$$

Because of symmetry, $\phi_3^n = \phi_2^n$. It is then enough to consider either of equations (c). Generalizing, there is obtained first

$$\left(\frac{a}{12}5 + \frac{D}{6}\frac{b}{a}2\right)\phi_2^{n+1} + \left(-\frac{a}{12}5 + \frac{D}{6}\frac{b}{a}1\right)\phi_2^n = 0 \tag{j}$$

and finally

$$\phi_2^{n+1} \approx 0.599 \cdot \phi_2^n \tag{k}$$

This is a simple special case of formula (f).

Results evaluated by (k) are shown in Figure (c). To shorten the time step does not change the results apppreciably because the crudeness of the mesh in space generates errors which do not disappear if $\Delta t$ tends separately to zero. The analytical solution is evaluated from formula (h) of Example 9.1 at $x = L/3$.

$\phi / \bar{\phi}_0$



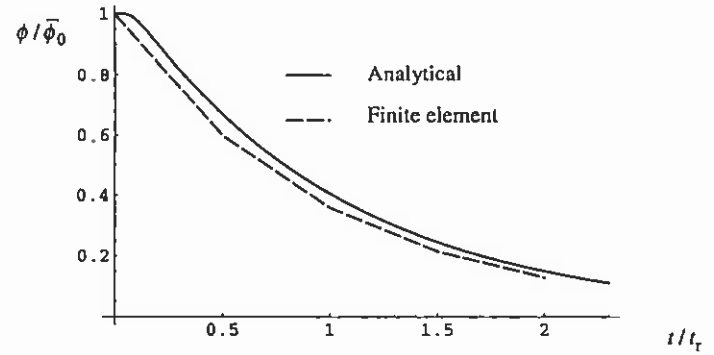



**Figure (c)**

## 9.2 TIME INTEGRATION

### 9.2.1 General

We consider the linear constant coefficient system of ordinary first order differential equations written in the form

$$\boxed{[M]\{\dot{a}\} + [K]\{a\} = \{b\}} \tag{1}$$

As initial conditions we have

$$\boxed{\{a(t)\} = \{a\}^0} \quad \text{at } t = 0 \tag{2}$$

This setting has been described already in Section 9.1.2 where it emerged as a result of partial discretization of a continuum problem in space and time. But this kind of problem arises in many physical systems directly without any discretization.

The solution of system (1) ⋯ (2) means the determination of the column matrix $\{a(t)\}$ as a function of time. Column matrix $\{b\}$ is a given function of time and matrices $[M]$ and $[K]$ can in general also depend on time. If they depend in addition on $\{a\}$, the system is non-linear. The numerical treatment of problem (1) ⋯ (2) is often called *time integration* (aikaintegrointi). Due to the early appearance and importance of this kind of problems in physics — movement of heavenly bodies, ballistic problems — the literature contains a confusing number of algorithms dealing with time integration; see e.g. Mäkelä et al. (1982).

The finite element method can be applied to system (1) ... (2) to generate different time integration algorithms, Zienkiewicz and Morgan (1983). As here the solution domain boundary consists simply of two points: $t = 0$ and $t = T$, the geometric flexibility of the finite element method probably has not much new to add to the older algorithms. Many old versions can in fact be generated by the finite element method. We therefore do not present the application of the finite element method to this problem as we concentrate later on the time-discontinuous Galerkin method applied to space-time continuum problems.

Some main concepts on time integration are now touched upon. We denote different discrete values of time or *time levels* (aikataso) as follows

$$\begin{aligned} t_0 &= 0 \\ t_1 &= t_0 + \Delta t \\ t_2 &= t_1 + \Delta t = t_0 + 2\Delta t \\ &\cdots \\ t_n &= t_{n-1} + \Delta t = t_0 + n\Delta t \end{aligned} \tag{3}$$

The *time step* (aika-askel) $\Delta t$ between the time levels need not necessarily be constant as has been indicated for notational convenience in (3). In fact, in adaptive time stepping, the step length is constantly monitored for optimum efficiency.

Corresponding to (3), we denote

$$\begin{aligned} \{a(t_0)\} &= \{a\}^0 \\ \{a(t_1)\} &= \{a\}^1 \\ &\cdots \\ \{a(t_n)\} &= \{a\}^n \end{aligned} \tag{4}$$

The terms *single-step method* (yksiaskelmenetelmä) and *multistep-method* (moniaskelmenetelmä) are used according to if the algorithm determining quantity $\{a\}^{n+1}$ contains only the value of $\{a\}^n$ or more levels, respectively. Alternatively, the terms two-level method, three-level method, etc. are used. Here we only consider single-step (two-level) schemes.

An important concept is the stability of a time integration algorithm. In some methods the errors can under certain conditions tend to grow exponentially with time. If this is the case, the method is called *unstable* (epästabiili), otherwise it is *stable* (stabiili). In certain algorithms there exists a so-called critical time step. The method stays stable if the time step remains smaller than the critical time step. The method is then called *conditionally stable* (ehdollisesti stabiili). If no



conditions are necessary to guarantee stability, the method is called *unconditionally stable* (ehdoitta stabiili). We refer to Zienkiewicz and Morgan (1983) for details.

A classification connected closely with the text above is the following. A method is *explicit* (eksplisiittinen, avoin) if the unknows of the new time level are obtained directly as certain expressions without solution of a system of equations. A method is *implicit* (implisiittinen, suljettu) if this is not the case. Explicit methods are usually conditionally stable and implicit methods are usually unconditionally stable. Thus even if the computational work needed per time step is much less in a former method compared to the latter, the limitation of the critical time step length may make the number of time steps so large that the total computational effort may become larger. On the other hand, the time step length of an unconditionally stable method cannot be taken arbitrarily large because the discretization errors naturally grow correspondingly. Bathe and Wilson (1976) is a classic text dealing with these themes from the point of view of the finite element method.

**Remark 9.10.** In mathematical texts a more general setting than (1) is often given in the so-called explicit form

$$\{\dot{a}\} = \{f(t, \{a(t)\})\} \tag{5}$$

A corresponding linear or linearized form would be

$$\{\dot{a}\} = [A]\{a\} + \{g\} \tag{6}$$

where $[A]$ and $\{g\}$ are mostly functions of time. If matrix $[M]$ in (1) is regular, we arrive at (5) in principle by operating with $[M]^{-1}$ in (1). Because application of the finite element method (in space) produces normally form of type (1), it is natural to start the time stepping directly from it and not from the form conventional in numerical mathematics textbooks. □

### 9.2.2 $\theta$-method

**Introduction.** We only consider here one rather widely used difference method type procedure, which might be called the $\theta$-method to discretize in time.

Equation (1) written at the instant of time $t = t_n + \theta\,\Delta t$ where $0 \le \theta \le 1$ gives a yet exact result

$$[M]\{\dot{a}\}^{n+\theta} + [K]\{a\}^{n+\theta} = \{b\}^{n+\theta} \tag{7}$$

We apply now the finite difference approximations

$$\{\dot{a}\}^{n+\theta} = \frac{1}{\Delta t}\left(\{a\}^{n+1} - \{a\}^n\right) + \mathrm{O}(\Delta t) \tag{8}$$

and

$$\{a\}^{n+\theta} = (1-\theta)\{a\}^n + \theta\{a\}^{n+1} + O\left(\Delta t^2\right) \tag{9}$$

This means that for each component of $\{\dot{a}\}$ and $\{a\}$, the approximations of the type shown in Figure 9.5 are used. The true function is simply replaced by linear interpolation between the end values and the derivative and the function value is evaluated from it.

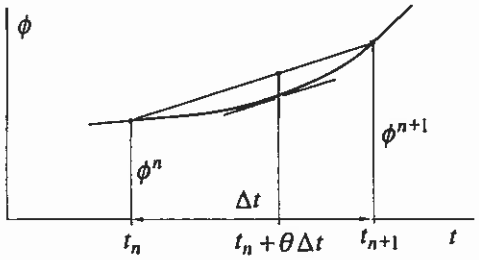



**Figure 9.5** Finite difference approximations.

It should be mentioned that when $\theta = 1/2$, the order of the error in (8) is $O(\Delta t^2)$, Zienkiewicz and Morgan (1983, p. 288).

Introducing (8) and (9) into (7) gives

$$\boxed{\left(\frac{1}{\Delta t}[M]+\theta[K]\right)\{a\}^{n+1} = \left(\frac{1}{\Delta t}[M]-(1-\theta)[K]\right)\{a\}^n + \{b\}^{n+\theta}} \tag{10}$$

Thus the discretization has again transformed a differential equation system into an algebraic system. The unknowns $\{a\}^{n+1}$ are solved from this set. Formally

$$\{a\}^{n+1} = \left(\frac{1}{\Delta t}[M]+\theta[K]\right)^{-1} \cdot \left(\left(\frac{1}{\Delta t}[M]-(1-\theta)[K]\right)\{a\}^n + \{b\}^{n+\theta}\right) \tag{11}$$

The resulting algorithm is of very simple nature. When $n = 0$, $\{a\}^0$ is known from the inititial condition (2) and formula (11) gives $\{a\}^1$. Next step gives $\{a\}^2$ and so on. If the time step is kept constant the coefficient matrix $[M]/\Delta t + \theta[K]$ is also constant and new solutions can be obtained relatively cheaply. Depending on the value of $\theta$ different versions of the $\theta$-method are obtained.



**Forward difference scheme, forward Euler scheme** (etudifferenssimenetelmä, Eulerin menetelmä, eksplisiittinen Eulerin menetelmä). Here we put $\theta = 0$ to obtain

$$\frac{1}{\Delta t}[M]\{a\}^{n+1} = \left(\frac{1}{\Delta t}[M]-[K]\right)\{a\}^n + \{b\}^n \tag{12}$$

If the $[M]$-matrix is diagonal, the scheme is really explicit since no solution of a system of equations is needed. In addition, it is not necessary to assemble the $[M]$- and $[K]$-matrices as the contributions $[K]\{a\}^n$, etc. can be evaluated at the element level. In connection with the finite element method the $[M]$-matrix is not in general diagonal (see Example 9.1) — contrary to what is obtained when the finite difference method is applied to the field equation. To ease the calculations, the $[M]$-matrix is however often transformed artificially into a diagonal form by lumping (keskittäminen, "möykyttäminen") the matrix elements on the diagonal by taking for instance

$$M_{ii} = \sum_j M_{ij} \tag{13}$$

Reference Hughes (1987) contains material on lumping.

**Crank-Nicolson method, central difference method** (Crank-Nicolsonin menetelmä, keskeisdifferenssimenetelmä). We put $\theta = 1/2$ to obtain

$$\left(\frac{1}{\Delta t}[M]+\frac{1}{2}[K]\right)\{a\}^{n+1} = \left(\frac{1}{\Delta t}[M]-\frac{1}{2}[K]\right)\{a\}^n + \{b\}^{n+1/2} \tag{14}$$

This is an implicit method, as the $[K]$-matrix cannot any more be lumped into a diagonal form in a reasonable way. The method is very popular and has good accuracy. This is partly explained by the text following Figure 9.5.

**Backward difference scheme, backward Euler scheme** (takadifferenssimenetelmä, implisiittinen Eulerin menetelmä). We put $\theta = 1$ to obtain a strongly implicit method

$$\left(\frac{1}{\Delta t}[M]+[K]\right)\{a\}^{n+1} = \frac{1}{\Delta t}[M]\{a\}^n + \{b\}^{n+1} \tag{15}$$

**Example 9.3.** We consider the problem

$$\dot{\phi} + c\phi = 0 \tag{a}$$

with the initial condition

$$\phi(0)=\bar{\phi} \tag{b}$$

The analytical solution is

$$\phi(t)=e^{-ct}\,\bar{\phi} \tag{c}$$

When $c$ is positive — as is assumed here — the solution tends to xero the faster the larger $c$.

The problem may be considered as problem (9.1.1) with no diffusion and no source term and reaction included (unsteady reaction problem). Alternatively, we may consider it as equation (1) with only one component. This latter interpretation with $[M]\triangleq 1$, $[K]\triangleq c$, $\{b\}\triangleq 0$, transforms (10) into the form

$$\left(\frac{1}{\Delta t}+\theta c\right)\phi^{n+1}=\left(\frac{1}{\Delta t}-(1-\theta)c\right)\phi^{n} \tag{d}$$

The algorithm is thus

$$\begin{aligned}&\phi^{0}=\bar{\phi}\\ &\phi^{n+1}=\frac{1/\Delta t-(1-\theta)c}{1/\Delta t+\theta c}\phi^{n}\end{aligned} \tag{e}$$

Figures (a) and (b) show some results obtained by the three discrete schemes described above. In Figure (a) the time step length $\Delta t=0.5/c$ and in Figure (b) $\Delta t=2.2/c$. Crank-Nicolson method is clearly the most accurate in Figure (a). With the longer time step the forward difference scheme exhibits already unstable behavior so the critical time step length has been surpassed. Especially the results by the Crank-Nicolson method clearly show that even in implicit methods the step length cannot be too large for accuracy reasons.

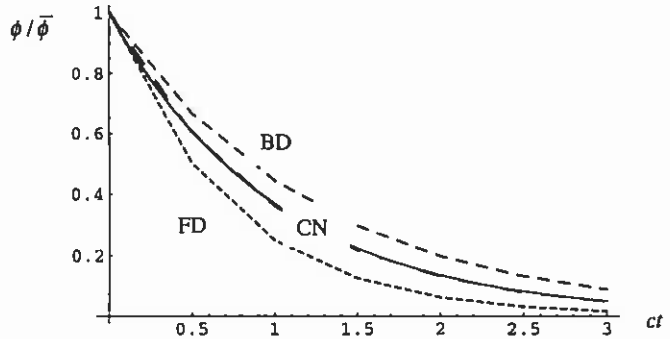





**Figure (a)** $\phi/\bar{\phi}$ as a function of $ct$. $\Delta t=0.5/c$. FD = forward difference, CN = Crank-Nicolson, BD = backward difference.

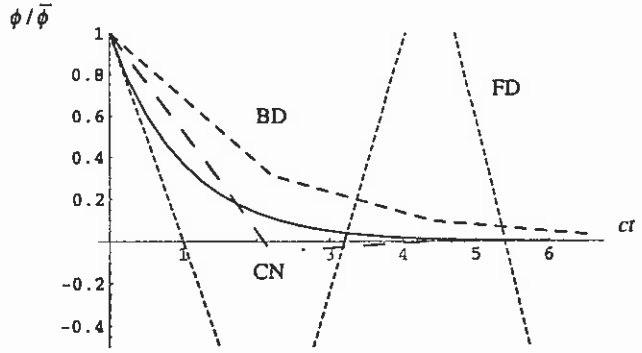



**Figure (b)** $\phi/\bar{\phi}$ as a function of $ct$. $\Delta t=2.2/c$. FD = forward difference, CN = Crank-Nicolson, BD = backward difference.

**Example 9.4.** We obtained by semidiscretization in Example 9.1 the matrix system

$$\frac{h}{6}\begin{bmatrix}4&1\\1&4\end{bmatrix}\begin{Bmatrix}\dot{\phi}_2\\\dot{\phi}_3\end{Bmatrix}+\frac{D}{h}\begin{bmatrix}2&-1\\-1&2\end{bmatrix}\begin{Bmatrix}\phi_2\\\phi_3\end{Bmatrix}=\begin{Bmatrix}0\\0\end{Bmatrix} \tag{a}$$

with the initial data

$$\begin{Bmatrix}\phi_2\\\phi_3\end{Bmatrix}=\begin{Bmatrix}\bar{\phi}_0\\\bar{\phi}_0\end{Bmatrix}\quad\text{at } t=0 \tag{b}$$

We solve this now by the Crank-Nicolson method.

As here from symmetry, $\phi_3=\phi_2$ and $\dot{\phi}_3=\dot{\phi}_2$, we could get from (a) a single equation

$$\frac{5h}{6}\dot{\phi}_2+\frac{D}{h}\phi_2=0 \tag{c}$$

for $\phi_2$ and similarly for $\phi_3$. However, to demonstrate the matrix formulas we start from (a). Formula (14) gives

$$\left(\frac{h}{6\Delta t}\begin{bmatrix}4&1\\1&4\end{bmatrix}+\frac{1}{2}\frac{D}{h}\begin{bmatrix}2&-1\\-1&2\end{bmatrix}\right)\begin{Bmatrix}\phi_2\\\phi_3\end{Bmatrix}^{n+1}=\left(\frac{h}{6\Delta t}\begin{bmatrix}4&1\\1&4\end{bmatrix}-\frac{1}{2}\frac{D}{h}\begin{bmatrix}2&-1\\-1&2\end{bmatrix}\right)\begin{Bmatrix}\phi_2\\\phi_3\end{Bmatrix}^{n} \tag{d}$$

The initial values are

$$\begin{Bmatrix}\phi_2\\ \phi_3\end{Bmatrix}^0 = \begin{Bmatrix}\bar{\phi}_0\\ \bar{\phi}_0\end{Bmatrix} \tag{e}$$

We use the same characteristic time interval as in Example 9.2:

$$t_r = \frac{L^2}{D\pi^2} \tag{f}$$

and also the same time step size: $\Delta t = 0.5 t_r$. The term

$$\frac{h}{\Delta t} = \frac{hD\pi^2}{0.5(3h)^2} = \frac{2\pi^2}{9}\frac{D}{h} \tag{g}$$

System (d) simplifies to

$$\left(\frac{\pi^2}{27}\begin{bmatrix}4 & 1\\ 1 & 4\end{bmatrix} + \frac{1}{2}\begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}\right)\begin{Bmatrix}\phi_2\\ \phi_3\end{Bmatrix}^{n+1} = \left(\frac{\pi^2}{27}\begin{bmatrix}4 & 1\\ 1 & 4\end{bmatrix} - \frac{1}{2}\begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}\right)\begin{Bmatrix}\phi_2\\ \phi_3\end{Bmatrix}^n \tag{h}$$

Using Mathematica, we manipulate this by inverting the coefficient matrix to

$$\begin{Bmatrix}\phi_2\\ \phi_3\end{Bmatrix}^{n+1} = \begin{bmatrix}0.207523 & 0.36287\\ 0.36287 & 0.207523\end{bmatrix}\begin{Bmatrix}\phi_2\\ \phi_3\end{Bmatrix}^n \tag{i}$$

Taking the symmetry into account gives further the relationship

$$\phi_2^{n+1} \approx 0.570\,\phi_2^n \tag{j}$$

Comparison with formula (k) of Example 9.2 shows that a similar but a little lower running graph than that in Figure (c) of Example 9.2 is obtained.



## 9.3 TIME-DISCONTINUOUS GALERKIN METHOD

### 9.3.1 Introduction

To use discontinuous approximation for quantities, which are known to be in reality continuous may seem at first sight rather odd. This idea has, however, proved to be very useful, Johnson (1987).

One main ingredient in the time-discontinuous Galerkin method is that the initial Dirichlet conditions are not satisfied in advance (in strong sense) — as we have been accustomed to deal in connection with Dirichlet conditions — but only in a weak sense.

We try to explain the nature of the weak form appearing in the time-discontinuous Galerkin method using a simple example. We consider the differential equation

$$\frac{d\phi}{dt} - f(t) = 0 \tag{1}$$

with the initial condition

$$\phi = \bar{\phi} \equiv \phi^0 \quad \text{at } t = 0 \tag{2}$$

Equation (1) is seen to be a very simple special case of the D-C-R equation considered in Appendix A. Only the unsteady term and the source term are included.

The exact solution is

$$\phi(t) = \bar{\phi} + \int_0^t f(t)\,dt \tag{3}$$

Figure 9.6 shows a schematic solution and some notations. The time axis is divided as before into time intervals or time slabs

$$I_n = ]t_n, t_{n+1}[ \tag{4}$$

We are prepared for possible discontinuous behaviour at the time level "interfaces" by equipping $\phi$ and the weighting function $w$ with plus- and minus-subscripts at the time levels $t_0$, $t_1$, etc.

## 9.3 TIME-DISCONTINUOUS GALERKIN METHOD

### 9.3.1 Introduction

To use discontinuous approximation for quantities, which are known to be in reality continuous may seem at first sight rather odd. This idea has, however, proved to be very useful, Johnson (1987).

One main ingredient in the time-discontinuous Galerkin method is that the initial Dirichlet conditions are not satisfied in advance (in strong sense) — as we have been accustomed to deal in connection with Dirichlet conditions — but only in a weak sense.

We try to explain the nature of the weak form appearing in the time-discontinuous Galerkin method using a simple example. We consider the differential equation

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} - f(t) = 0 \tag{1}$$

with the initial condition

$$\phi = \overline{\phi} \equiv \phi^0 \quad \text{at } t = 0 \tag{2}$$

Equation (1) is seen to be a very simple special case of the D-C-R equation considered in Appendix A. Only the unsteady term and the source term are included.

The exact solution is

$$\phi(t) = \overline{\phi} + \int_0^t f(t)\,\mathrm{d}t \tag{3}$$

Figure 9.6 shows a schematic solution and some notations. The time axis is divided as before into time intervals or time slabs

$$I_n = ]t_n, t_{n+1}[ \tag{4}$$

We are prepared for possible discontinuous behavior at the time level "interfaces" by equipping $\phi$ and the weighting function $w$ with plus- and minus-subscripts at the time levels $t_0$, $t_1$, etc.



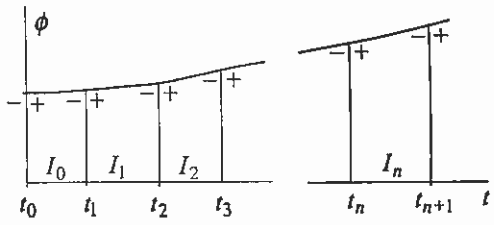



**Figure 9.6** Some notations.

We consider one typical time interval $I_n$ and obtain general expressions, which can then be applied consecutively for each interval. The starting point is the obvious weak form.

$$\int_{I_n} w\left(\frac{\mathrm{d}\phi}{\mathrm{d}t} - f\right)\mathrm{d}t = 0 \tag{5}$$

We integrate by parts:

$$\int_{I_n} w\frac{\mathrm{d}\phi}{\mathrm{d}t}\mathrm{d}t = -\int_{I_n}\frac{\mathrm{d}w}{\mathrm{d}t}\phi\,\mathrm{d}t + \Big|_{t_n}^{t_{n+1}} w\phi = -\int_{I_n}\frac{\mathrm{d}w}{\mathrm{d}t}\phi\,\mathrm{d}t + w_-^{n+1}\phi_-^{n+1} - w_+^n\phi_+^n \tag{6}$$

It is seen that the limiting values approached from inside of the interval are used in the notations. However, as the exact solution is continuous, $\phi_+^n = \phi_-^n$, which is taken into account in (6) to give

$$\int_{I_n} w\frac{\mathrm{d}\phi}{\mathrm{d}t}\mathrm{d}t = -\int_{I_n}\frac{\mathrm{d}w}{\mathrm{d}t}\phi\,\mathrm{d}t + w_-^{n+1}\phi_-^{n+1} - w_+^n\phi_-^n \tag{7}$$

Now we integrate back by parts on the right-hand side of (7) to give

$$\begin{aligned}\int_{I_n} w\frac{\mathrm{d}\phi}{\mathrm{d}t}\mathrm{d}t &= \int_{I_n} w\frac{\mathrm{d}\phi}{\mathrm{d}t}\mathrm{d}t - w_-^{n+1}\phi_-^{n+1} + w_+^n\phi_+^n + w_-^{n+1}\phi_-^{n+1} - w_+^n\phi_-^n \\ &= \int_{I_n} w\frac{\mathrm{d}\phi}{\mathrm{d}t}\mathrm{d}t + w_+^n\left(\phi_+^n - \phi_-^n\right)\end{aligned} \tag{8}$$

When this is introduced in (5) we end up with the weak form

$$\boxed{\int_{I_n} w\left(\frac{\mathrm{d}\phi}{\mathrm{d}t} - f\right)\mathrm{d}t + w_+^n\left(\phi_+^n - \phi_-^n\right) = 0} \tag{9}$$

Knowing the one-way nature of the time coordinate, we have succeeded in introducing through the manipulations an additional term to (5), which concerns quantities at the initial instant of the time interval. At $t = t_0 = 0$ we associate $\bar{\phi} \equiv \phi^0$ in (2) as $\phi^0_-$.

**Remark 9.11.** We could have written down a weak form like (9) directly from (1) and (2) say as

$$\int_I w\left(\frac{\mathrm{d}\phi}{\mathrm{d}t} - f\right)\mathrm{d}t + w_{\mathrm{D}}\left(\phi - \bar{\phi}\right)\big|_{t=0} = 0 \tag{10}$$

A similar starting point was described in Section 2.1.2. However, it is immediately not obvious which is the appropriate relationship between the weighting function $w$ and the weighting constant $w_{\mathrm{D}}$ to achieve accurate results. The forward and backward integration by parts manipulation generates here the rule: $w_{\mathrm{D}} = w(0)$. □

The discrete analogue of (9) is

$$\int_{I_n} \tilde{w}\left(\frac{\mathrm{d}\tilde{\phi}}{\mathrm{d}t} - f\right)\mathrm{d}t + \tilde{w}^n_+\left(\tilde{\phi}^n_+ - \tilde{\phi}^n_-\right) = 0 \tag{11}$$

Let the approximation in $I_n$ be say the beginning of a Taylor series at $t = t_n$

$$\tilde{\phi}(t) = \alpha + \beta(t - t_n) = \alpha \cdot 1 + \beta \cdot (t - t_n) \tag{12}$$

where $\alpha$ and $\beta$ are unknown parameters to be determined. This is the first case in this text where we encounter so-called nodeless discrete unknowns in the finite element method mentioned in Remark 3.9. The trial basis functions are indicated in the last form of (11) to consist of functions 1 and $t - t_n$ (Figure 9.7).


**Figure 9.7 (a)** First trial basis function. **(b)** Second trial basis function.

(1) We consider first just the case $\tilde{\phi}(t) = \alpha$. Then

$$\frac{\mathrm{d}\tilde{\phi}}{\mathrm{d}t} = 0, \qquad \tilde{\phi}^n_+ = \alpha \tag{13}$$

and the discrete equation when using the Galerkin method is obtained by taking as the weighting function the trial basis function, that is, $\tilde{w} = 1$. This gives



$$\int_{I_n} 1 \cdot (0 - f)\mathrm{d}t + 1 \cdot \left(\alpha - \tilde{\phi}^n_-\right) = 0 \tag{14}$$

or

$$\alpha = \tilde{\phi}^n_- + \int_{I_n} f\,\mathrm{d}t \tag{15}$$

(2) If we use the full approximation (12),

$$\frac{\mathrm{d}\tilde{\phi}}{\mathrm{d}t} = \beta, \qquad \tilde{\phi}^n_+ = \alpha \tag{16}$$

In the Galerkin method the discrete weightings are now $\tilde{w} = 1$ and $\tilde{w} = t - t_n$ giving the system equations

$$\begin{aligned} &\int_{I_n} 1 \cdot (\beta - f)\mathrm{d}t + 1 \cdot \left(\alpha - \tilde{\phi}^n_-\right) = 0 \\ &\int_{I_n} (t - t_n)(\beta - f)\mathrm{d}t + 0 \cdot \left(\alpha - \tilde{\phi}^n_-\right) = 0 \end{aligned} \tag{17}$$

or

$$\begin{aligned} &\alpha + \Delta t \cdot \beta = \tilde{\phi}^n_- + \int_{I_n} f\,\mathrm{d}t \\ &\frac{1}{2}(\Delta t)^2 \cdot \beta = \int_{I_n} (t - t_n) f\,\mathrm{d}t \end{aligned} \tag{18}$$

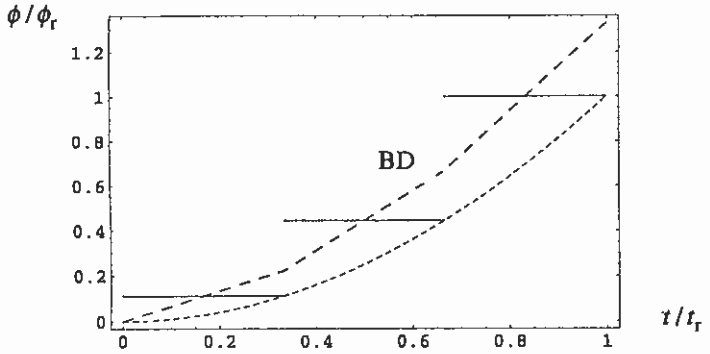
For the first interval $I_0$, we take $\tilde{\phi}^0_- = \bar{\phi}$. It is readily seen that the value $\tilde{\phi}^1_-$ obtained at $t = t^1_-$ happens to coincide with the exact one for both formulations in this simple example case. This value is used as a new initial condition for the second interval $I_1$ and again exact value is obtained at right-hand endpoint, etc.

It is a general property of the time-discontinuous Galerkin method that in general accurate values are obtained at the "future" end of the time interval. This is naturally advantageous for the next time interval, as good initial values are available.

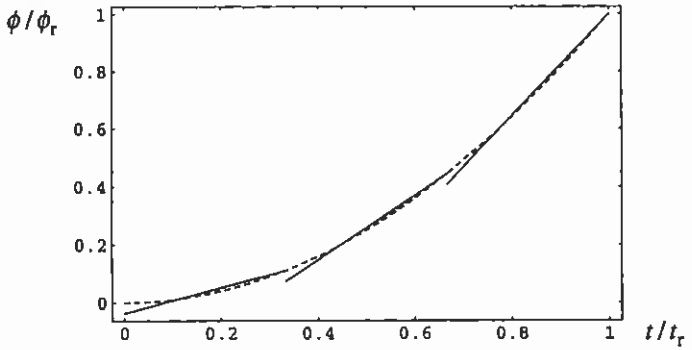
Figures 9.8 (a) and (b) show results with three time steps for the case $f = 2t\phi_{\mathrm{r}}/t^2_{\mathrm{r}}$, $\bar{\phi} = 0$ where the exact solution is according to (3)

$$\phi(t) = \frac{t^2}{t^2_{\mathrm{r}}}\phi_{\mathrm{r}} \tag{19}$$

where $\phi_r$ is a reference value for $\phi$ and $t_r$ similarly a reference value for $t$. Especially from Figure (a) it is clearly evident that the initial conditions for each time interval are satisfied only in a weak sense.




(a)




(b)

**Figure 9.8** **(a)** Constant approximation. **(b)** Linear approximation.

Figure (a) shows also the result obtained by the backward difference (BD) scheme, which has in this case the form

$$\phi^{n+1} = \phi^n + \Delta t f^{n+1} \tag{20}$$

### 9.3.2 Space-time application

We consider again the model problem of Section 9.1.1:

$$\frac{\partial \phi}{\partial t} + \frac{\partial}{\partial x}\left(-D\frac{\partial \phi}{\partial x}\right) - f = 0 \quad \text{in } \Omega^t \tag{21}$$



$$\phi = \bar{\phi}(t) \quad \text{on } \Gamma_D^t \tag{22}$$

$$-D\frac{\partial \phi}{\partial x} = \bar{j}^d(t) \quad \text{on } \Gamma_N^t \tag{23}$$

$$\phi = \bar{\phi}_0(x) \quad \text{in } \Omega \text{ at } t = 0 \tag{24}$$

The appropriate weak form can be generated as a combination of the manipulations explained in Sections 9.1.3 and 9.3.1. We obtain for a slab

$$\boxed{\begin{aligned}&\int_{S_n}\left(w\frac{\partial \phi}{\partial t} + \frac{\partial w}{\partial x}D\frac{\partial \phi}{\partial x}\right)\mathrm{d}S - \int_{S_n} wf\,\mathrm{d}S + \int_{(\partial S_n)_N} w\bar{j}^d\,\mathrm{d}(\partial S_n)\\ &+\int_\Omega w_+^n\left(\phi_+^n - \phi_-^n\right)\mathrm{d}\Omega = 0\end{aligned}} \tag{25}$$

where the admissible $\phi$ has now only to satisfy the Dirichlet boundary condition. The change compared to (9.1.22) is the addition of the integral over $\Omega$ at the initial time $t_n$ of the time interval $I_n = ]t_n, t_{n+1}[$. It may be noted that in the forwards and backwards integration by parts manipulation of term $w\partial\phi/\partial t$ over the slab depicted in Figure 9.3, the component $n_t$ of the outward unit normal vector $\mathbf{n}$ disappears at the sides $x = a$ and $x = b$ and has the values $-1$ and $+1$ at the sides $t = t_n$ and $t = t_{n+1}$, respectively.

The discrete analogue of (25) is

$$\begin{aligned}&\int_{S_n}\left(\bar{w}\frac{\partial \tilde{\phi}}{\partial t} + \frac{\partial \bar{w}}{\partial x}D\frac{\partial \tilde{\phi}}{\partial x}\right)\mathrm{d}S - \int_{S_n} \bar{w}f\,\mathrm{d}S + \int_{(\partial S_n)_N} \bar{w}\bar{j}^d\,\mathrm{d}(\partial S_n)\\ &+\int_\Omega \bar{w}_+^n\left(\tilde{\phi}_+^n - \tilde{\phi}_-^n\right)\mathrm{d}\Omega = 0\end{aligned} \tag{26}$$

In the time-discontinuous Galerkin method the simplest finite element representation is achieved by using a *product form* somewhat similarly as in Section 3.2.2 where two-dimensional shape functions were obtained for quadrilateral elements by multiplying together one-dimensional shape functions. Here the multiplying functions in the space "direction" are simply conventional space shape functions but in the time direction a power series approximation is employed. Thus we may write (we denote $t_n = t_+$)

$$\tilde{\phi}(x,t) = F(x)G(t) \approx \left[\sum_k N_k(x)\hat{\phi}_k\right]\left[\alpha + \beta(t - t_+) + \gamma\frac{1}{2}(t - t_+)^2 + \cdots\right] \tag{27}$$

This may be written also as

$$\tilde{\phi}(x,t)=\sum_k N_k(x)\phi_k^{(0)}+\sum_k N_k(x)(t-t_+)\phi_k^{(1)} + \sum_k N_k(x)\frac{1}{2}(t-t_+)^2\phi_k^{(2)}+\cdots \tag{28}$$

where the resulting unknown parameters $a_j$ are $\phi_k^{(0)}=\alpha\hat{\phi}_k$, $\phi_k^{(1)}=\beta\hat{\phi}_k$, $\phi_k^{(2)}=\gamma\hat{\phi}_k$. (We have used the symbol $\hat{\phi}_k$ in (27) so that the term $\phi_k^{(0)}$ has now the physical interpretation of being the value of $\tilde{\phi}$ at a spatial node $k$ at $t=t_+$; consider (28) at $t=t_+$. The terms $\phi_k^{(1)}$ and $\phi_k^{(2)}$ have no more any transparent interpretations as they are of different physical dimension than $\phi$.) We can thus finally write

$$\boxed{\tilde{\phi}(x,t)=\sum_j N_j(x,t)a_j} \tag{29}$$

where the trial basis functions are of the type

$$N_j(x,t)=N_k(x)\frac{1}{l!}(t-t_+)^l \tag{30}$$

and where the meaning of the notations should be obvious. (We do not give the detailed connections between the different indices in this rather cursory introduction.) It may be further remarked that the product type basis functions (30) are rather easy to comprehend even in the case of three space dimensions. The three first global trial basis functions (30) corresponding to a node in connection with the use of two-noded line elements in space are sketched in Figure 9.9.

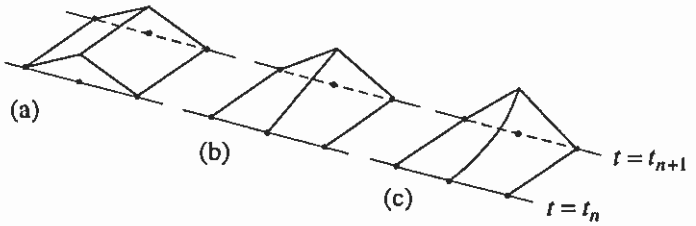



**Figure 9.9** Trial basis function **(a)** constant in time, **(b)** linear in time. **(c)** quadratic in time.

**Remark 9.12.** The finite element approximation presented above can be approached from an alternative point of view. As the space-time slabs can be taken arbitrarily thin in the time direction, the possibility of *dimension reduction* (dimensioreduktio) is always available. Dimension reduction means that in a continuum problem the dependence of a quantity on one (or more) coordinate is reduced into a finite dimensional dependence. A typical example in structural mechanics in two dimensions is beam analysis when the beam is thin, that is, the thickness is small compared to the length of the beam. The displacement field is expanded in power series in the thickness direction and usually only the constant and (maybe) the linear term are included. This means that the continuum problem is reduced to the determination of two or more functions of only one coordinate: the beam length coordinate. These functions can then be determined by some further discretization process. (The dimension reduction itself is clearly a discretization procedure but of a somewhat different nature than the conventional finite element method.) Here we may refer to the kinematical assumptions (5.1.2):

$$\begin{aligned} u(x,y)&=y\theta(x)\\ v(x,y)&=v(x)\end{aligned} \tag{31}$$

used in the Timoshenko beam theory. These can be considered to have been obtained from the Taylor expansions in the $y$- direction:

$$\begin{aligned} u(x,y)&=u(x,y)\big|_{y=0}+\frac{\partial u(x,y)}{\partial y}\Big|_{y=0}y+\cdots\equiv u(x)+\theta(x)y+\cdots\\ v(x,y)&=v(x,y)\big|_{y=0}+\cdots\equiv v(x)+\cdots\end{aligned} \tag{32}$$

with the further assumption $u(x,y)\big|_{y=0}\equiv u(x)=0$ based on certain physics of the problem.

□

In the time-discontinuous Galerkin method we can start similarly as explained in Remark 9.12 by expanding $\phi(x,t)$ into a Taylor expansion in the time direction to give

$$\begin{aligned}\phi(x,t)&=\phi(x,t)\big|_{t=t_+}+\frac{\partial\phi(x,t)}{\partial t}\Big|_{t=t_+}(t-t_+)+\frac{1}{2}\frac{\partial^2(x,t)}{\partial t^2}\Big|_{t=t_+}(t-t_+)^2+\cdots\\ &\equiv\phi^{(0)}(x)+\phi^{(1)}(x)(t-t_+)+\frac{1}{2}\phi^{(2)}(x)(t-t_+)^2+\cdots\end{aligned} \tag{33}$$

If we stop with the terms shown, we have three unknown functions *depending only on x* to be determined. Now we continue with the approximations

$$\begin{aligned}\phi^{(0)}(x)&\approx\sum_k N_k(x)\phi_k^{(0)}\\ \phi^{(1)}(x)&\approx\sum_k N_k(x)\phi_k^{(1)}\\ \phi^{(2)}(x)&\approx\sum_k N_k(x)\phi_k^{(2)}\end{aligned} \tag{34}$$

Substitution of these into (33) gives again form (29).

Using approximation (29) and applying the Galerkin method in (26) with $N_i(x,t)$ as weighting functions generates the system equations

$$\boxed{[K]\{a\}=\{b\}} \tag{35}$$

for the slab with

$$\begin{aligned} K_{ij} &= \int_{S_n}\left( N_i\frac{\partial N_j}{\partial t}+\frac{\partial N_i}{\partial x}D\frac{\partial N_j}{\partial x}\right)\mathrm{d}S+\int_{\Omega}(N_i)_+^n(N_j)_+^n\,\mathrm{d}\Omega \\ b_i &= \int_{S_n}N_i f\,\mathrm{d}S-\int_{(\partial S_n)_N}N_i\bar{j}^{\mathrm{d}}\mathrm{d}(\partial S)+\int_{\Omega}(N_i)_+^n\bar{\phi}_-^n\,\mathrm{d}\Omega \end{aligned} \tag{36}$$

Johnson (1987) advocates the usefulness of the time-discontinuous Galerkin method especially with respect to the good possibilities to apply adaptive strategies.

### 9.3.3 Space-time applications with constant in time approximation

**Introduction.** We employ now the theory presented in the previous section in more detail taking the simplest case: only the constant term in the power series in time is included. Although this option is admittedly crude and may sometimes demand rather short time steps to achieve the accuracy needed, it leads to a very simple and pleasant final formulation, which contains the earlier steady case weak form just appended with some additional boundary terms.

The approximation (28) in a time slab is now simply

$$\tilde{\phi}(x,t)=\sum_j N(x)\phi_j \tag{37}$$

We include here also the possibility of convection and extend the field equation (21) to consist of (to simplify the following manipulations, we write in the the convection term here so that the convection velocity is outside the derivative operator)

$$R(\phi)\equiv L(\phi)-f\equiv\frac{\partial\phi}{\partial t}+\frac{\partial}{\partial x}\left(-D\frac{\partial\phi}{\partial x}\right)+u\frac{\partial\phi}{\partial x}-f=0 \tag{38}$$

The weak form (25) enhances first to ($a=0$, $b=L$)

$$\int_0^L\int_I\left( w\frac{\partial\phi}{\partial t}+\frac{\partial w}{\partial x}D\frac{\partial\phi}{\partial x}+wu\frac{\partial\phi}{\partial x}-wf\right)\mathrm{d}t\,\mathrm{d}x+\int_I w\bar{j}^{\mathrm{d}}\,\mathrm{d}t$$



$$+\int_0^L w^+\left(\phi^+-\phi^-\right)\mathrm{d}x+\int_0^L\int_I L(w)\tau^{\mathrm{c}}R(\phi)\,\mathrm{d}t\,\mathrm{d}x=0 \tag{39}$$

We have changed the notation from the more general to the more specific to have a more transparent formula. Further, the index $n$ referring to the time slab has been dropped for simplicity; we are considering a generic time slab. Also, we have now included the sensitizing term. The discrete analogue of (39) is

$$\begin{aligned} &\int_0^L\int_I\left(\tilde{w}\frac{\partial\tilde{\phi}}{\partial t}+\frac{\partial\tilde{w}}{\partial x}D\frac{\partial\tilde{\phi}}{\partial x}+\tilde{w}u\frac{\partial\tilde{\phi}}{\partial x}-\tilde{w}f\right)\mathrm{d}t\mathrm{d}x+\int_I\tilde{w}\bar{j}^{\mathrm{d}}\,\mathrm{d}t \\ &+\int_0^L\tilde{w}^+\left(\tilde{\phi}^+-\tilde{\phi}^-\right)\mathrm{d}x+\int_0^L\int_I L(\tilde{w})\tau^{\mathrm{c}}R(\tilde{\phi})\,\mathrm{d}t\,\mathrm{d}x=0 \end{aligned} \tag{40}$$

With (37), we can write here using our new notation in the whole slab

$$\tilde{\phi}(x,t)\equiv\tilde{\phi}^+(x,t)=\sum_j N(x)\phi_j^+ \tag{41}$$

and on the slab interface

$$\tilde{\phi}^-(x,t)=\sum_j N(x)\phi_j^- \tag{42}$$

The approximation $\tilde{\phi}$ and thus the finite dimensional weighting $\tilde{w}$ in the Galerkin method do not depend on time. The term $\partial\tilde{\phi}/\partial t$ vanishes and also $\partial\tilde{w}/\partial t$ in the sensitizing term. Further, the inner integrals

$$\begin{aligned} &\int_I\frac{\partial\tilde{w}}{\partial x}D\frac{\partial\tilde{\phi}}{\partial x}\mathrm{d}t=\frac{\partial\tilde{w}}{\partial x}\frac{\partial\tilde{\phi}}{\partial x}\int_I D(x,t)\mathrm{d}t\equiv\frac{\partial\tilde{w}}{\partial x}\frac{\partial\tilde{\phi}}{\partial x}D_{\mathrm{m}}(x)\Delta t \\ &\int_I\tilde{w}u\frac{\partial\tilde{\phi}}{\partial x}\mathrm{d}t\approx\tilde{w}\frac{\partial\tilde{\phi}}{\partial x}\int_I u(x,t)\mathrm{d}t\equiv\frac{\partial\tilde{w}}{\partial x}\frac{\partial\tilde{\phi}}{\partial x}u_{\mathrm{m}}(x)\Delta t \\ &\int_I\tilde{w}f\,\mathrm{d}t=\tilde{w}\int_I f(x,t)\mathrm{d}t\equiv\tilde{w}f_{\mathrm{m}}(x)\Delta t \\ &\int_I\tilde{w}\bar{j}^{\mathrm{d}}\mathrm{d}t=\tilde{w}\int_I\bar{j}^{\mathrm{d}}(t)\mathrm{d}t\equiv\tilde{w}\bar{j}_{\mathrm{m}}^{\mathrm{d}}\Delta t \\ &\int_I L(\tilde{w})\tau^{\mathrm{c}}R(\tilde{\phi})\mathrm{d}t\approx\int_I u\frac{\partial\tilde{w}}{\partial x}\tau^{\mathrm{c}}\left(u\frac{\partial\tilde{\phi}}{\partial x}-f(x,t)\right)\mathrm{d}t \\ &=\tau^{\mathrm{c}}u^2\frac{\partial\tilde{w}}{\partial x}\frac{\partial\tilde{\phi}}{\partial x}\int_I\mathrm{d}t-\tau^{\mathrm{c}}u\frac{\partial\tilde{w}}{\partial x}\int_I f(x,t)\mathrm{d}t\equiv\tau^{\mathrm{c}}u^2\frac{\partial\tilde{w}}{\partial x}\frac{\partial\tilde{\phi}}{\partial x}\Delta t-\tau^{\mathrm{c}}u\frac{\partial\tilde{w}}{\partial x}f_{\mathrm{m}}\Delta t \end{aligned} \tag{43}$$

The given data $D$, $u$, $f$ depend in general on space and time and $\bar{j}^{\mathrm{d}}$ on time. We have defined in (43) average values for them with respect to time. In connection with the sensitizing term similar simplifications as described earlier have been applied. Using (43) and dropping for simplicity the subscript m referring to the average values transforms (40) after division by $\Delta t$ to

$$\int_0^L \left( \frac{\partial \tilde{w}}{\partial x} D \frac{\partial \tilde{\phi}}{\partial x} + \tilde{w} u \frac{\partial \tilde{\phi}}{\partial x} - \tilde{w} f \right) \mathrm{d}x + \tilde{w} \bar{j}^{\mathrm{d}} \Big|_{x=L}$$
$$+ \int_0^L \frac{\partial \tilde{w}}{\partial x} \tau^{\mathrm{c}} u^2 \frac{\partial \tilde{\phi}}{\partial x} \mathrm{d}x - \int_0^L \frac{\partial \tilde{w}}{\partial x} \tau^{\mathrm{c}} u f \, \mathrm{d}x$$
$$+ \frac{1}{\Delta t} \int_0^L \tilde{w}^+ \left( \tilde{\phi}^+ - \tilde{\phi}^- \right) \mathrm{d}x = 0 \tag{44}$$

**Remark 9.13.** The weak form (44) is interesting. The time dimension has disappeared and we have a discrete weak form only in space. Further, this weak form is in the first and second row completely analogous to the weak form in the steady case as presented in Section 6.2. (We could replace finally the operators $\partial/\partial x$ in (44) to $\mathrm{d}/\mathrm{d}x$ if wanted.). It is obvious that also when we have a time dependent problem in two or three space dimensions, the constant in time approximation leads again back formally to a steady formulation. One could easily conclude from this that the same sensitizing parameter values found good in the steady case could work also here. Numerical experiments show, however, that this is not the case. The last line in (44) changes strongly the situation. The writers have had this far only limited experience on determination of the sensitizing parameter values in time dependent problems. One result is described below. □

The discrete system corresponding to (44) becomes

$$\sum_j \left( K_{ij}^{\mathrm{s}} + K_{ij}^{\mathrm{t}} \right) \phi_j^+ = b_i^{\mathrm{s}} + \sum_j K_{ij}^{\mathrm{t}} \phi_j^- \tag{45}$$

with

$$K_{ij}^{\mathrm{s}} = \int_0^L \frac{\mathrm{d}N_i}{\mathrm{d}x} D \frac{\mathrm{d}N_j}{\mathrm{d}x} \mathrm{d}x + \int_0^L N_i u \frac{\mathrm{d}N_j}{\mathrm{d}x} \mathrm{d}x + \int_0^L \frac{\mathrm{d}N_i}{\mathrm{d}x} \tau^{\mathrm{c}} u^2 \frac{\mathrm{d}N_j}{\mathrm{d}x} \mathrm{d}x$$
$$K_{ij}^{\mathrm{t}} = \frac{1}{\Delta t} \int_0^L N_i N_j \, \mathrm{d}x \tag{46}$$
$$b_i = \int_0^L N_i f \, \mathrm{d}x - N_i \bar{j}^{\mathrm{d}} \Big|_{x=L} + \int_0^L \frac{\mathrm{d}N_i}{\mathrm{d}x} \tau^{\mathrm{c}} u f \, \mathrm{d}x$$

**Remark 9.14.** It is realized from (45) that the time-discontinuous Galerkin method produces an implicit scheme as coupled unknowns appear at the future end of the space-time slab. In a purely hyperbolic problem — here the case (38) with $D = 0$ — this may be a disadvantage as explicit schemes are usually considered more appropriate in connection with hyperbolic problems, e.g., Anderson et al. (1984, p. 110). Example 9.6 is concerned with this hyperbolic



case. On the other hand, the time-discontinuous Galerkin method generates systematically the discrete equations over the full range of values of $D$ and $u$. Further, we have not tried here to make use of all the possibilities in the time-discontinuous Galerkin method: mesh adaptively changing from slab to slab, elements adaptively oriented in the slabs, etc., e.g. Hansbo (1994). □

**Reference solution.** We consider the simplified field equation (38):

$$\phi_t - D\phi_{xx} + u\phi_x - f = 0 \tag{47}$$

We employ the series form approach described in Section 5.2.2. A short Mathematica code is attached:

```
eqs = { φt - D φxx + u φx - f = 0,
        φxt - D φxxx + u φxx - fx = 0,
        φtt - D φxxt + u φxt - ft = 0 };

sol = Solve [eqs, {φt, φxx, φtt}]

{{φt → f - u φx - (D (- fx + φxt - D φxxx)) / u,
  φtt → ft - u φxt + D φxxt, φxx → - (- fx + φxt - D φxxx) / u }}

ϕ[x_, t_] := ϕ + φx x + φt t + 1/2 φxx x^2 + φxt x t + 1/2 φtt t^2

Collect [ϕ[x, t] /. sol, {ϕ, φx, φxt, φxxx, φxxt, f, fx, ft}]

{f t + ϕ + t^2 ft / 2 + (D t / u + x^2 / (2 u)) fx + (- t u + x) φx +
  (- D t / u - t^2 u / 2 + t x - x^2 / (2 u)) φxt + 1/2 D t^2 φxxt + (D^2 t / u + D x^2 / (2 u)) φxxx}
```

Ending as seen from the code, we have obtained the reference solution

$$\begin{Bmatrix} \phi \\ f \end{Bmatrix} = \phi_0 \begin{Bmatrix} 1 \\ 0 \end{Bmatrix} + (\phi_x)_0 \begin{Bmatrix} x - ut \\ 0 \end{Bmatrix} + (\phi_{xt})_0 \begin{Bmatrix} -\frac{D}{u}t - \frac{1}{2u}x^2 + xt - \frac{u}{2}t^2 \\ 0 \end{Bmatrix}$$
$$+ f_0 \begin{Bmatrix} t \\ 1 \end{Bmatrix} + (f_x)_0 \begin{Bmatrix} \frac{D}{u}t + \frac{1}{2u}x^2 \\ x \end{Bmatrix} + (f_t)_0 \begin{Bmatrix} \frac{1}{2}t^2 \\ t \end{Bmatrix}$$

$$+(\phi_{xxx})_0\begin{Bmatrix}\frac{D^2}{u}t+\frac{D}{2u}x^2\\0\end{Bmatrix}+(\phi_{xxt})_0\begin{Bmatrix}\frac{D}{2}t^2\\0\end{Bmatrix} \tag{48}$$

Actually, the last two specific solutions are no more exact reference solutions.

A study — not repeated here — shows that for two-noded linear elements the "damping diffusivity" becomes

$$\boxed{D^c \equiv \tau^c u^2 = -\frac{1}{2}\Delta t\, u^2} \tag{49}$$

This is obtained by using the third specific reference solution in (48) in a patch test with two elements. The result, giving a negative value for the damping diffusivity is completely different from what was found in Chapter 6 in the steady case.

An important non-dimensional quantity in numerical fluid mechanics is the *Courant number*

$$\boxed{\nu = \frac{|\mathbf{v}|\Delta t}{h}} \tag{50}$$

Here $|\mathbf{v}|$ is the speed of the fluid flow and $\Delta t$ and $h$ are characteristic time step and mesh size respectively. In many explicit schemes the Courant number around value 1 gives good results and further, this value cannot be overstepped for stability reasons.

**Example 9.5.** We consider once again the unsteady pure diffusion problem of Example 9.1. We have the field equation

$$\frac{\partial\phi}{\partial t} - D\frac{\partial^2\phi}{\partial x^2} = 0, \qquad 0<x<L, \quad 0<t<T \tag{a}$$

the boundary conditions

$$\phi(0,t) = \bar{\phi}(0,t) = 0, \qquad \phi(L,t) = \bar{\phi}(L,t) = 0, \qquad t>0 \tag{b}$$

and the initial condition

$$\phi(x,0) = \bar{\phi}_0 = \text{constant}, \qquad 0<x<L \tag{c}$$



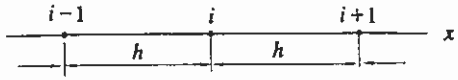



**Figure (a)**

We have assumed a constant $D$ and put $u=0$, $f=0$ in (46). We use two-noded line elements. Formula (49) gives no sensitizing. A typical system equation inside a uniform mesh (Figure (a)) becomes

$$\left(-\frac{D}{h}+\frac{1}{6}\frac{h}{\Delta t}\right)\phi_{i-1}^{+}+\left(2\frac{D}{h}+\frac{4}{6}\frac{h}{\Delta t}\right)\phi_i^{+}+\left(-\frac{D}{h}+\frac{1}{6}\frac{h}{\Delta t}\right)\phi_{i+1}^{+} = \frac{1}{6}\frac{h}{\Delta t}\phi_{i-1}^{-}+\frac{4}{6}\frac{h}{\Delta t}\phi_i^{-}+\frac{1}{6}\frac{h}{\Delta t}\phi_{i+1}^{-} \tag{d}$$

If we consider the + and - values to situate the distance $\Delta t$ apart in the time direction and use the "lumping type" replacements

$$\begin{aligned}\frac{1}{6}\frac{h}{\Delta t}\phi_{i-1}^{+}+\frac{4}{6}\frac{h}{\Delta t}\phi_i^{+}+\frac{1}{6}\frac{h}{\Delta t}\phi_{i+1}^{+} &:= \frac{h}{\Delta t}\phi_i^{+}\\ \frac{1}{6}\frac{h}{\Delta t}\phi_{i-1}^{-}+\frac{4}{6}\frac{h}{\Delta t}\phi_i^{-}+\frac{1}{6}\frac{h}{\Delta t}\phi_{i+1}^{-} &:= \frac{h}{\Delta t}\phi_i^{-}\end{aligned} \tag{e}$$

we have rederived the so-called *simple implicit (Laasonen) method*, Anderson et al. (1984, p. 111).

Returning to the case described by equations (a), (b) and (c) and using the spatial mesh shown in Figure (a) of Example 9.1, system equations (d) look

$$\begin{aligned}\left(2\frac{D}{h}+\frac{4}{6}\frac{h}{\Delta t}\right)\phi_2^{+}+\left(-\frac{D}{h}+\frac{1}{6}\frac{h}{\Delta t}\right)\phi_3^{+} &= \frac{4}{6}\frac{h}{\Delta t}\phi_2^{-}+\frac{1}{6}\frac{h}{\Delta t}\phi_3^{-}\\ \left(-\frac{D}{h}+\frac{1}{6}\frac{h}{\Delta t}\right)\phi_2^{+}+\left(2\frac{D}{h}+\frac{4}{6}\frac{h}{\Delta t}\right)\phi_3^{+} &= \frac{1}{6}\frac{h}{\Delta t}\phi_2^{-}+\frac{4}{6}\frac{h}{\Delta t}\phi_3^{-}\end{aligned} \tag{f}$$

with $h = L/3$.

A more realistic case with uniform mesh of ten elements in space and with the time step $\Delta t = t_r/50$ is calculated by MATHFEM. The code reads

```
d = 1; ϕb = 1; L = 1; tr = L^2/(d*Pi^2);
mm = 50; dt = tr/mm;
nn = 10; dom = {{0}, {1}};
msh = MSH[dom, {nn}, 2];
apr = APR[msh, {ϕb} &];
prb = PRB[
apr, {0, w[1]*d* ϕ[1] + (w[0]*(ϕ[0] -
ϕ̃[0]))/dt}];
prb = FIX[prb, dom, {{0}, {0}}];
SHOW1D[PLOT[NONSTATIONARY[prb, mm]]];
```

Figure (b) gives the distribution of the exact and finite element solution at $t = 0.5t_r$. As the exact solution is here very smooth, it is no wonder that the finite element solution is very accurate even with a rather crude spatial mesh.

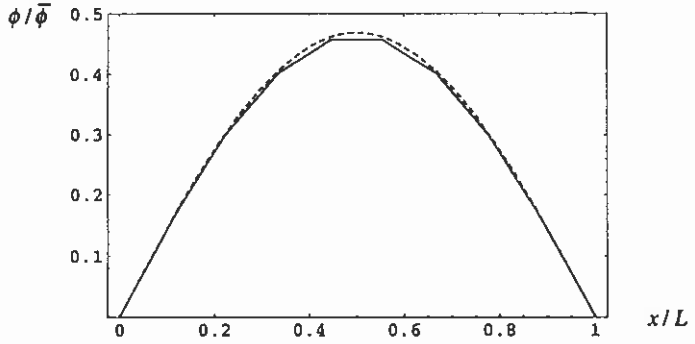



**Figure (b)**

**Example 9.6.** We consider an unsteady diffusion-convection problem with constant diffusivity and flow velocity and no source term. We have the field equation

$$\frac{\partial \phi}{\partial t} - D\frac{\partial^2 \phi}{\partial x^2} + u\frac{\partial \phi}{\partial x} = 0, \quad 0 < x < L, \quad 0 < t < T \tag{a}$$

the boundary conditions

$$\begin{aligned} &\phi(0,t) = (t/\Delta T)\bar{\phi}, \quad 0 < t \le \Delta T \\ &\phi(0,t) = \bar{\phi} = \text{constant}, \quad t > \Delta T \\ &j^{\mathrm{d}}(t)\big|_{x=L} = \bar{j}^{\mathrm{d}}(t)\big|_{x=L} = 0, \quad t > 0 \end{aligned} \tag{b}$$

and the initial condition

$$\phi(x,0) = 0, \quad 0 < x < L \tag{c}$$

This means that a linearly in time increasing Dirichlet data appears on the left-hand boundary up to time $\Delta T$ after which the data is constant. The solution should thus finally approach a steady state.

We put $f = 0$ in (46). We use two-noded line elements and assume in the following that $u$ is positive.

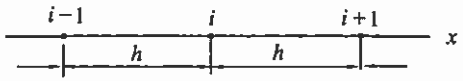



**Figure (a)**

A typical system equation inside a uniform mesh (Figure (a)) becomes



$$\begin{aligned} &\left(-\frac{D+D^{\mathrm{c}}}{h} - \frac{1}{2}u + \frac{1}{6}\frac{h}{\Delta t}\right)\phi_{i-1}^{+} + \left(2\frac{D+D^{\mathrm{c}}}{h} + \frac{4}{6}\frac{h}{\Delta t}\right)\phi_i^{+} \\ &+\left(-\frac{D+D^{\mathrm{c}}}{h} + \frac{1}{2}u + \frac{1}{6}\frac{h}{\Delta t}\right)\phi_{i+1}^{+} = \frac{1}{6}\frac{h}{\Delta t}\phi_{i-1}^{-} + \frac{4}{6}\frac{h}{\Delta t}\phi_i^{-} + \frac{1}{6}\frac{h}{\Delta t}\phi_{i+1}^{-} \end{aligned} \tag{d}$$

We further consider here the case where the Peclet number is in practice infinite and put $D = 0$. We have thus an unsteady pure convection case (hyperbolic case). Equation (d) becomes

$$\begin{aligned} &\left(-\frac{D^{\mathrm{c}}}{h} - \frac{1}{2}u + \frac{1}{6}\frac{h}{\Delta t}\right)\phi_{i-1}^{+} + \left(2\frac{D^{\mathrm{c}}}{h} + \frac{4}{6}\frac{h}{\Delta t}\right)\phi_i^{+} \\ &+\left(-\frac{D^{\mathrm{c}}}{h} + \frac{1}{2}u + \frac{1}{6}\frac{h}{\Delta t}\right)\phi_{i+1}^{+} = \frac{1}{6}\frac{h}{\Delta t}\phi_{i-1}^{-} + \frac{4}{6}\frac{h}{\Delta t}\phi_i^{-} + \frac{1}{6}\frac{h}{\Delta t}\phi_{i+1}^{-} \end{aligned} \tag{e}$$

Using (49) for the value of $D^{\mathrm{c}}$, we have further the detailed system equation

$$\begin{aligned} &\left(\frac{u^2\Delta t}{2h} - \frac{1}{2}u + \frac{1}{6}\frac{h}{\Delta t}\right)\phi_{i-1}^{+} + \left(-\frac{u^2\Delta t}{h} + \frac{4}{6}\frac{h}{\Delta t}\right)\phi_i^{+} \\ &+\left(\frac{u^2\Delta t}{2h} + \frac{1}{2}u + \frac{1}{6}\frac{h}{\Delta t}\right)\phi_{i+1}^{+} = \frac{1}{6}\frac{h}{\Delta t}\phi_{i-1}^{-} + \frac{4}{6}\frac{h}{\Delta t}\phi_i^{-} + \frac{1}{6}\frac{h}{\Delta t}\phi_{i+1}^{-} \end{aligned} \tag{f}$$

If we apply lumping the way described in formula (e) of Example (9.5), we obtain from formula (f) a scheme having much similarity with the *Lax-Wendroff method*, Anderson et al. (1984, p. 101). However, the Lax-Wendroff method is an explicit scheme and (f) is implicit.

For comparison, it is interesting to write down the system equation, if we use the damping diffusivity $D^{\mathrm{c}}$ obtained in the steady case. With a large Peclet number, $\hat{\tau}^{\mathrm{c}} = 1/2$, and

$$\frac{D^{\mathrm{c}}}{h} = \frac{\hat{\tau}^{\mathrm{c}}\mathrm{Pe}_h D}{h} = \frac{\frac{1}{2}\frac{uh}{D}D}{h} = \frac{1}{2}u \tag{g}$$

This corresponds to the full upwinding value discussed in Section 6.2.1. The system equation becomes

$$\begin{aligned} &\left(-u + \frac{1}{6}\frac{h}{\Delta t}\right)\phi_{i-1}^{+} + \left(u + \frac{4}{6}\frac{h}{\Delta t}\right)\phi_i^{+} + \left(\frac{1}{6}\frac{h}{\Delta t}\right)\phi_{i+1}^{+} = \\ &\frac{1}{6}\frac{h}{\Delta t}\phi_{i-1}^{-} + \frac{4}{6}\frac{h}{\Delta t}\phi_i^{-} + \frac{1}{6}\frac{h}{\Delta t}\phi_{i+1}^{-} \end{aligned} \tag{h}$$

Finally, a kind of compromise is obtained by putting simply $D^{\mathrm{c}} = 0$. This produces the system equation

$$\left(-\frac{1}{2}u+\frac{1}{6}\frac{h}{\Delta t}\right)\phi_{i-1}^{+}+\left(\frac{4}{6}\frac{h}{\Delta t}\right)\phi_{i}^{+}+\left(\frac{1}{2}u+\frac{1}{6}\frac{h}{\Delta t}\right)\phi_{i+1}^{+}$$
$$=\frac{1}{6}\frac{h}{\Delta t}\phi_{i-1}^{-}+\frac{4}{6}\frac{h}{\Delta t}\phi_{i}^{-}+\frac{1}{6}\frac{h}{\Delta t}\phi_{i+1}^{-} \qquad \text{(i)}$$

With lumping, this becomes exactly the *Euler implicit method*, Anderson et al. (1984, p. 98).

In flow problems one obvious characteristic time interval $t_r$ is the time needed for a fluid particle to cross a typical length measure of the problem with a typical flow speed. Here we put

$$t_r=\frac{L}{u} \qquad \text{(j)}$$

The exact solution behaves as sketched in Figure (b). Here the comments at the end of Section A.3 help to see the character of the exact solution. As the "velocity components" $u$ and 1 in the $xt$-plane are constants, the "streamlines" are straight lines. The inflow boundary consists of the lines $t=0, 0\le x\le L$, $x=0, 0\le t$. As there is no diffusion and no source term, the boundary data is transferred unchanged into the domain as explained in two space dimensions in connection with Figure A.2. To simulate such a behavior with a discrete model is rather demanding. We also see qualitatively, why an implicit scheme does not describe the physics very truly. A system equation associated with a nodal point with value $\phi_i^+$ contains terms $\phi_{i-1}^+$ and $\phi_{i+1}^+$. These latter terms introduce effects, which should strictly not be present as the value of $\phi_i^+$ is in principle totally determined just from the information on its own streamline.

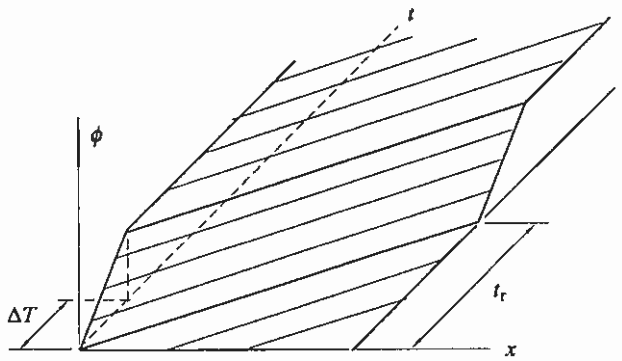



**Figure (b)**

This problem has been solved by MATHFEM taking $\Delta T=t_r/2$ and by using a uniform spatial mesh of 40 two-noded elements. The time step length $\Delta t=t_r/320$ has been employed. This gives the element Courant numbert $\nu=1/8$. Some solutions are shown in Figures (c), (d) and (e) at the times $t=t_r/4$, $t=2t_r/4$, $t=3t_r/4$, $t=t_r$, $t=5t_r/4$, $t=6t_r/4$ for the cases $D^c=1/2(uh)$, $D^c=0$ and $D^c=-1/2(\Delta t\,u^2)$, respectively, discussed above.



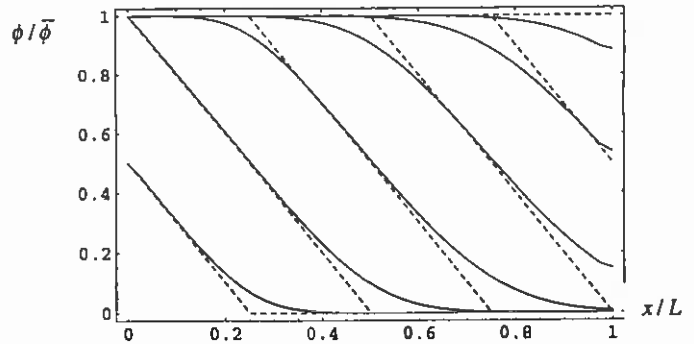



**Figure (c)**

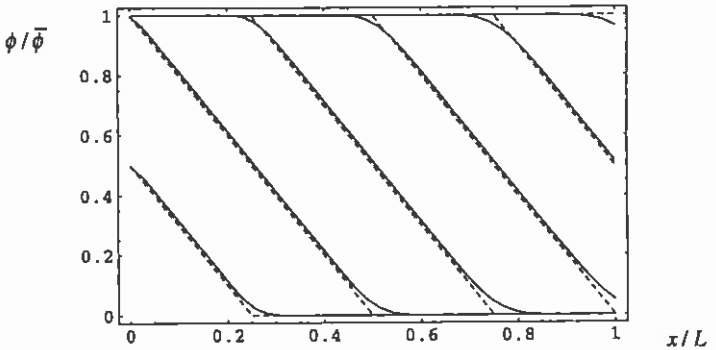



**Figure (d)**

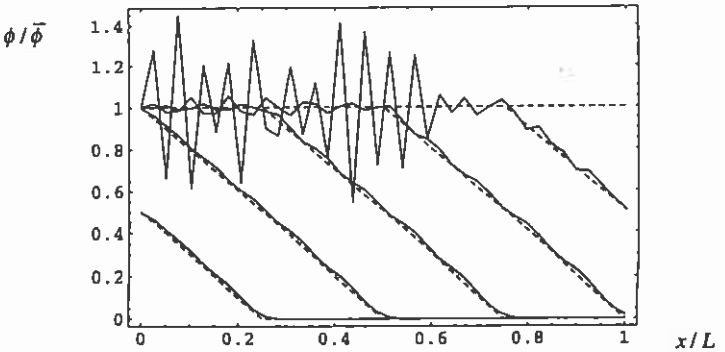



**Figure (e)**

It is seen from Figure (c) that the positive diffusive damping corresponding to the steady optimal value damps the solution too strongly. On the other hand, the negative"optimal" value obtained above for the unsteady case leads to unacceptable oscillations. The case of no diffusive damping works here clearly best. The problem needs more study especially with respect to possible Dirichlet outflow boundary conditions.

**Truncation error.** In the finite difference method it is usual to speak about the *truncation error* (katkaisuvirhe) and of the *order* (kertaluku) of a specific scheme. The idea is to develop the unknown function (functions) into Taylor series about a convenient point and to substitute this evaluated at the gridpoints into the governing finite difference molecule. For instance, doing this for equation (f) of Example 9.6 gives the end result

$$(\phi_t)_0 + u(\phi_x)_0 + \frac{u}{2}(\phi_{xt})_0 \Delta t + \frac{1}{2}(\phi_{xt})_0 h + \cdots = 0 \tag{51}$$

The terms with lowest powers of $\Delta t$ and $h$ emerging are shown. The truncation error is all on the left-hand side following the two first terms, which give the left-hand side of the governing field equation at the expansion point. The order of the scheme is $0(\Delta t)$ with respect to time and $0(h)$ with respect to space. This means that the leading terms in the truncation error are proportional to the first powers of $\Delta t$ and $h$. In obtaining (51), use has been made of the differentiated field equations similarly as in obtaining the reference solution (48). For the schemes (h) and (i) of Example 9.6, we obtain respectively the results

$$(\phi_t)_0 + u(\phi_x)_0 + \frac{u^2}{6}(\phi_{xxt})_0 (\Delta t)^2 + \frac{u}{24}(\phi_{xxxt})_0 h^2 \Delta t + \cdots = 0 \tag{52}$$

and

$$(\phi_t)_0 + u(\phi_x)_0 + \frac{u}{2}(\phi_{xt})_0 \Delta t + \frac{u}{12}(\phi_{xxxt})_0 h^2 \Delta t + \cdots = 0 \tag{53}$$

Again the terms with lowest powers of $\Delta t$ and $h$ are shown. The orders of the leading truncation terms indicate how the errors depend on the mesh size in space and time. In principle, the higher the exponents of $\Delta t$ and $h$, the better the scheme as the errors diminish the faster with the refinement of the mesh. However, the mesh must be "sufficiently" dense until the behavior suggested by the truncation terms really work. With practical meshes the situation may be such that the order considerations do not mean much. For instance, result (52) corresponding to the Lax-Wendroff type scheme would seem to be preferred because it has the highest order truncation errors.

## REFERENCES



Anderson, D. A., Tannehill, J. C. and Pletcher, R. H. (1984) *Computational Fluid Mechanics and Heat Transfer*, Hemisphere, Washington, ISBN 0-07-050328-1.

Bathe, K.-J. and Wilson, E. L. (1976) *Numerical Methods in Finite Element Analysis*, Prentice-Hall, Englewood Cliffs, New Jersey.

Hansbo, P. (1994). Space-time oriented streamline diffusion methods for non-linear conservation laws in one dimension, *Communications in Numerical Methods in Engineering*, Vol. 10, 203 – 215.

Hughes, T. J. R. (1987). *The Finite Element Method — Linear Static and Dynamic Finite Element Analysis*, Prentice-Hall, Englewood Cliffs, New Jersey, ISBN 0-13-317017-9.

Johnson, C. (1987). *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Studentlitteratur, Lund.

Lanczos, C. (1970). *The Variational Principles of Mechanics*, 4th ed., University of Toronto Press, Toronto.

Mäkelä, M., Nevanlinna, O. and Virkkunen, J. (1982). *Numeerinen Matematiikka*, Gaudeamus, Helsinki, ISBN 0-951-662-326-3.

Patankar, S. H. (1980). *Numerical Heat Transfer and Fluid Flow*, Mc-Graw-Hill, New York, ISBN 0-07-048740-5.

Zienkiewicz, O. C. and Morgan, K. (1983). *Finite Elements and Approximation*, Wiley, Chichester, ISBN 0-471-89089-8.

## PROBLEMS

# 11 NON-LINEARITY

In reality the governing equations in heat transfer and especially in fluid flow are often more or less non-linear. In attacking these problems linearization and iteration are the conventional tools. The deltaform discussed in Remark 2.14 is the basic expression to deal with the linearization. Similarities in the procedures discussed in Section 13.1.2 can be found.

## 11.1 STEADY CASE

### 11.1.1 Introduction

Let us represent a differential equation formally as, e.g., Shih (1984),

$$F(\phi,\phi',\phi'')=0 \tag{1}$$

This just means that the left-hand side is an expression containing an unknown function $\phi$ and its first and second derivatives with respect to a space coordinate. We have written (1) in the one-dimensional case but the extension to more dimensions is obvious. Further, (1) can also describe a boundary condition and if the derivatives are missing it becomes an algebraic equation.

Assume that we have obtained by guesswork or by iteration a preliminary solution $\bar{\phi}$, $\Rightarrow \bar{\phi}'$, $\bar{\phi}''$ which does not satisfy (1) and using the deltaform we try to get a better solution

$$\phi=\bar{\phi}+\Delta\phi \tag{2}$$

A truncated Taylor series representation of (1) about the preliminary solution gives

$$F\approx\bar{F}+\frac{\partial F}{\partial\phi}\Delta\phi+\frac{\partial F}{\partial\phi'}(\Delta\phi)'+\frac{\partial F}{\partial\phi''}(\Delta\phi)''=0 \tag{3}$$

where the partial derivatives are to be evaluated at the preliminary solution values. Also, $(\Delta\phi)'=\Delta\phi'$ and $(\Delta\phi)''=\Delta\phi''$. Equation (3) is a linear differential equation for $\Delta\phi$. After $\Delta\phi$ has been determined, we obtain from the left-hand side of (2) a new preliminary solution. Equation (3) is then applied again etc. until the solution changes hopefully stay under certain tolerances.

The step contained in (3) is called *linearization* (linearisointi) of the problem. In practice the determination of $\Delta\phi$ is of course done numerically; here by the finite element method. Three example cases where non-linearities are present are discussed in the following.





Approach like (3) may be called the *Newton-Raphson method* although this terminology is usually employed in connection with a similar approach in the solution of non-linear algebraic equations (see Section 13.1.2). An alternative method not employing the deltaform is called in the literature the *Picard method*. Also the names fixed point iteration, direct iteration, successive approximation are sometimes used. This approach is explained below in connection with the applications.

### 11.1.2 Variable diffusivity

Let us consider the diffusion equation

$$\left[-D(\phi)\phi'\right]'-f=0 \tag{4}$$

Here the diffusivity is not constant as has been assumed earlier but depends (in addition to possibly on position) on the solution itself. This is common in heat conduction where the dependence $k=k(T)$ may be quite strong. Reference Stelzer (1984) contains large material data for heat transfer in readily programmable form.

To obtain a more specific case, we take as an example the form

$$D=D_0(1+\beta\phi) \tag{5}$$

where $D_0$ and $\beta$ are constants even with respect to position. Developing (4) gives

$$\begin{aligned}F(\phi,\phi',\phi'')&\equiv-D'\phi'-D\phi''-f=-D_0\beta\phi'\phi'-D\phi''-f\\&=-D_0\beta(\phi')^2-D_0(1+\beta\phi)\phi''-f\end{aligned} \tag{6}$$

and

$$\frac{\partial F}{\partial\phi}=-D_0\beta\phi'', \qquad \frac{\partial F}{\partial\phi'}=-2D_0\beta\phi', \qquad \frac{\partial F}{\partial\phi''}=-D \tag{7}$$

Application of (3) gives thus

$$\left[-D(\bar{\phi})\bar{\phi}'\right]'-f-D_0\beta\bar{\phi}''\Delta\phi-2D_0\beta\bar{\phi}'(\Delta\phi)'-D(\bar{\phi})(\Delta\phi)''=0 \tag{8}$$

or

$$-D(\bar{\phi})(\Delta\phi)''-2D_0\beta\bar{\phi}'(\Delta\phi)'-D_0\beta\bar{\phi}''\Delta\phi-\bar{f}=0 \tag{9}$$

with

$$\bar{f} \equiv f + \left[ D(\bar{\phi}) \bar{\phi}' \right]' \tag{10}$$

Equation (9) is a D-C-R type equation for $\Delta\phi$. The second derivative term is, however, not quite in the form suitable for obtaining the weak form so we manipulate the equation into the final form

$$\left[ -D(\bar{\phi})(\Delta\phi)' \right]' - D_0 \beta \bar{\phi}' (\Delta\phi)' - D_0 \beta \bar{\phi}'' \Delta\phi - \bar{f} = 0 \tag{11}$$

The iteration can be started, say, with a $\bar{\phi}$, a "suitable" extension of the Dirichlet boundary data as discussed in Remark 2.15. However, see Remark 11.2. After discretizing, we arrive at a system

$$\left[ K(\{\bar{a}\}) \right] \{\Delta a\} = \{ b(\{\bar{a}\}) \} \tag{12}$$

with obvious meaning of the notations. A rather complicated updating is needed in each iteration especially if sensitizing is applied.

An alternative self-evident linearization is obtained by writing equation (4) just as

$$\left[ -D(\bar{\phi}) \phi' \right]' - f = 0 \tag{13}$$

Iteration is started with a suitable $\bar{\phi}$ and the $\phi$ obtained from (13) is used as an updated $\bar{\phi}$ etc. This is an example of the Picard method and it is here considerably simpler than the Newton-Raphson method. When discretized, this leads to a system

$$\left[ K(\{\bar{a}\}) \right] \{a\} = \{b\} \tag{13}$$

**Remark 11.1.** The modified source term (9) contains a diffusion type part. In obtaining the corresponding weak form the second derivative terms are integrated by parts. This manipulation should be applied also on the source term part. □

**Remark 11.2.** In a linear problem the initial selection of the finite element representation of the extension $\bar{\phi}$ of the Dirichlet data does not have an effect on the final solution. In non-linear cases the situation is different as the possibility of achieving convergence at all may depend on a good initial guess. Thus say an initial solution obtained using some constant average diffusion data or in a convection dominated case a smooth initial solution obtained without convection may be of value. □

**Remark 11.3.** A natural approximation for a variable diffusivity could be



$$D = D_0 + \alpha(\phi - \phi_0) \tag{14}$$

where $D_0 = D(\phi_0)$ and $\alpha = (\mathrm{d}D/\mathrm{d}\phi)_0$. However, Example 11.1 deals with an application from heat transfer literature where expression type (5) is used and this explains the selection used above. □

**Remark 11.4.** The MATHFEM program performs the linearization by itself. Thus, even when hand calculations to produce the linearized equations are given in the main text and in the three examples to follow to make the presentation more concrete, samples of the corresponding MATHFEM programs are included to show how the program in fact operates. □

**Example 11.1.** We consider a one-dimensional heat conduction problem given by the diferential equation

$$\left[ -k(T) T' \right]' = 0, \quad 0 < x < L \tag{a}$$

and by the Dirichlet boundary conditions

$$T(0) = T_1, \qquad T(L) = T_2 \tag{b}$$

The temperature dependent thermal conductivity is

$$k = k_0 (1 + \beta T) \tag{c}$$

The term $\beta$ is called the *temperature coefficient of thermal conductivity*. The exact solution can be found in this case rather easily:

$$T = -\frac{1}{\beta} \pm \sqrt{\frac{1}{\beta^2} - \frac{2k_{\text{ave}}}{\beta k_0} \frac{x}{L} (T_1 - T_2) + T_1^2 + \frac{2}{\beta} T_1} \tag{d}$$

The shorthand notation $k_{\text{ave}}$ means the following:

$$k_{\text{ave}} = k_0 \left( 1 + \beta \frac{T_2 + T_1}{2} \right) \tag{e}$$

According to Çengel (1998, p. 107), the proper sign of the square root term in (d) is determined from the requirement that the temperature at any point within the medium must remain between $T_1$ and $T_2$.

The differential equation corresponding to (11) can be written here as

$$\left[ -\bar{k} (\Delta T)' \right]' + \bar{u} (\Delta T)' + \bar{c} \Delta T - \bar{f} = 0 \tag{f}$$

with

$$\bar{k} = k_0 (1 + \beta \bar{T}), \qquad \bar{u} = -k_0 \beta \bar{T}', \qquad \bar{c} = -k_0 \beta \bar{T}'', \qquad \bar{f} = \left[ k_0 (1 + \beta \bar{T}) \bar{T}' \right]' \tag{g}$$

Thus the corresponding weak form with the Dirichlet boundary conditions is

$$\int_\Omega w'\bar{k}\,(\Delta T)'\,d\Omega + \int_\Omega w\bar{u}\,(\Delta T)'\,d\Omega + \int_\Omega w\bar{c}\,\Delta T\,d\Omega - \int_\Omega w\bar{f}\,d\Omega = 0 \tag{h}$$

See Remark 11.1 with respect to the source term.

The weak form corresponding to (13) is here simply

$$\int_\Omega w'\bar{k}T'\,d\Omega = 0 \tag{i}$$

again with

$$\bar{k} = k_0\left(1+\beta\bar{T}\right) \tag{j}$$

Following Çengel (1998), we take the example data

$$\begin{aligned} &L = 0.1\,\mathrm{m},\quad k_0 = 38\,\mathrm{W/(m\cdot K)},\quad \beta = 9.21\cdot 10^{-4}\,\mathrm{K^{-1}},\\ &T_1 = 600\,\mathrm{K},\quad T_2 = 400\,\mathrm{K} \end{aligned} \tag{k}$$

Figures (a) to (c) give some results given by MATHFEM with a regular mesh of four two-noded elements. No sensitizing has been applied. Results on the left (Figures (a) and (c)) are obtained by the Newton-Raphson method and results on the right (Figures (b) and (d)) by the Picard method. Different initial guesses (solid lines) are used in the upper and and lower figures. The solutions for the initial guess and after the first and the final iterationare shown by solid lines. The dashed line (hard to see) is the exact solution.

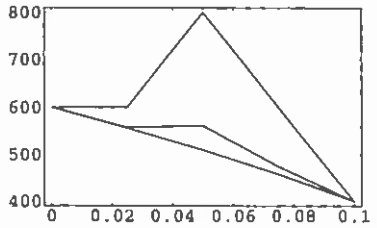



**Figure (a)**

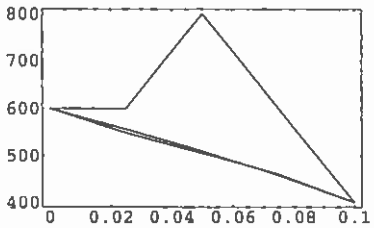



**Figure (b)**

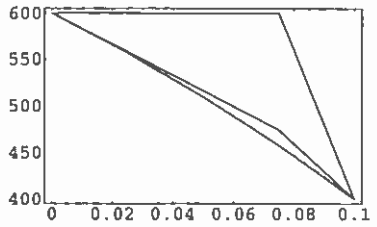



**Figure (c)**

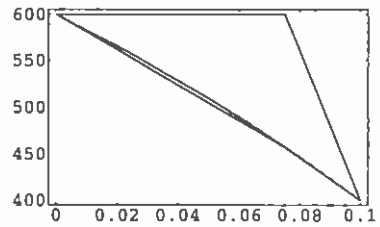



**Figure (d)**

The non-linearity is rather mild and convergence is reached very rapidly. Table (a) gives the results for the quantity



$$e = \frac{\left\|\tilde{T}^{(n)} - \tilde{T}^{(\infty)}\right\|_2}{\left\|\tilde{T}^{(\infty)}\right\|_2} \tag{l}$$

This is the relative rms-error *in the nodal values* between the exact solution to the discrete problem and the iterated solution after $n$ iterations. The initial guess is according to the case of the upper row (Figures (a) and (b)).

**Table (a)**

| Relative rms-error in the solution | | |
|---|---|---|
| Iteration | Newton | Picard |
| 1 | $2.9\ 10^{-1}$ | $2.9\ 10^{-1}$ |
| 2 | $4.8\ 10^{-2}$ | $8.4\ 10^{-3}$ |
| 3 | $2.1\ 10^{-3}$ | $5.8\ 10^{-4}$ |
| 4 | $4.6\ 10^{-6}$ | $2.1\ 10^{-5}$ |
| 5 | $2.4\ 10^{-11}$ | $1.3\ 10^{-6}$ |

A sample of the MATHFEM program follows:

```
<< mathfem.m;
nn = 5; L = 0.1; k0 = 38; β = 9.21 * 10^-2; ϕ1 = 600; ϕ2 = 400;
ϕ = ϕ1 + Cos[2 * Pi * #/ L] * (ϕ2 - ϕ1) &;
dom = {{0}, {L}};
msh = MSH[dom, {nn}, 2];
apr = APR[msh, ϕ];
k = k0 * (1 + β * φ[0]);
prb = PRB[apr, {0, w[1] * k * ϕ[1]}];
prb = FIX[prb, dom, {{ϕ1}, {ϕ2}}];
SHOWID[PLOT[NONLINEAR[prb, 0.0001]]];
```

### 11.1.3 Fluid flow momentum equation

A one-dimensional form of the steady momentum equation described Chapter 12 is

$$\left(-\mu u'\right)' + \rho u u' + p' - \rho b = 0 \tag{15}$$

The convection term $\rho u u'$ is the cause of the non-linearity. We consider a simplified non-dimensional case, Shih (1984),

$$F\left(\phi,\phi',\phi''\right) \equiv -\phi'' - 2\phi\phi' = 0 \tag{16}$$

using here the boundary conditions

$$\phi(-0.9)=10, \qquad \phi(3.1)=1/4.1 \tag{17}$$

Equation (16) contains a convection type term and serves as a demonstration case. The exact solution to the problem is found to be

$$\phi=\frac{1}{1+x} \tag{18}$$

Linearization of (16) gives

$$-\bar{\phi}''-2\bar{\phi}\bar{\phi}'-2\bar{\phi}'\Delta\phi-2\bar{\phi}(\Delta\phi)'-1\cdot(\Delta\phi)''=0 \tag{19}$$

or

$$-(\Delta\phi)''-2\bar{\phi}(\Delta\phi)'-2\bar{\phi}'\Delta\phi-\bar{f}=0 \tag{20}$$

with

$$\bar{f}\equiv\bar{\phi}''+2\bar{\phi}\bar{\phi}' \tag{21}$$

Equation (20) is again a linear D-C-R equation for $\Delta\phi$. The boundary conditions are

$$\bar{\phi}(-0.9)=10, \qquad \bar{\phi}(3.1)=1/4.1 \tag{22}$$

and

$$\Delta\phi(-0.9)=0, \qquad \Delta\phi(3.1)=0 \tag{23}$$

The Picard type approach is not always apparent or unique. We could write (16) here linearly for example as a D-R equation

$$-\phi''-2\bar{\phi}'\phi=0 \tag{24}$$

or as a D-C equation

$$-\bar{\phi}''-2\bar{\phi}\phi'=0 \tag{25}$$

The latter choice seems to be the governing one used in the literature.

**Example 11.2.** The differential equation (20) can be written as

$$-\bar{D}(\Delta\phi)''+\bar{u}(\Delta\phi)'+\bar{c}\Delta\phi-\bar{f}=0 \tag{a}$$

with



$$\bar{D}=1, \qquad \bar{u}=-2\bar{\phi}, \qquad \bar{c}=-2\bar{\phi}', \qquad \bar{f}=\bar{\phi}''+2\bar{\phi}\bar{\phi}' \tag{b}$$

The corresponding weak form is without sensitizing

$$\int_\Omega w'\bar{D}(\Delta\phi)'\,d\Omega+\int_\Omega w\bar{u}(\Delta\phi)'\,d\Omega+\int_\Omega w\bar{c}\Delta\phi\,d\Omega-\int_\Omega w\bar{f}\,d\Omega=0 \tag{c}$$

Again, Remark 11.1 should be taken into account when the contributions from the source term are evaluated.

The sensitizing terms are using two-noded elements

$$\int_\Omega(\bar{u}w'+\bar{c}w)\tau^c\left[\bar{u}(\Delta\phi)'+\bar{c}\Delta\phi-\bar{f}\right]d\Omega+\int_\Omega\bar{c}w'\tau^r\left[\bar{c}(\Delta\phi)'-\bar{f}'\right]d\Omega \tag{d}$$

For instance, equation (25) can be written as

$$-\bar{D}\phi''+\bar{u}\phi'=0 \tag{e}$$

with

$$\bar{D}=1, \qquad \bar{u}=-2\bar{\phi} \tag{f}$$

The corresponding weak form is without sensitizing

$$\int_\Omega w'\bar{D}(\Delta\phi)'\,d\Omega+\int_\Omega w\bar{u}(\Delta\phi)'\,d\Omega=0 \tag{g}$$

The sensitizing term is using two-noded elements

$$\int_\Omega\bar{u}w'\tau^c\bar{u}(\Delta\phi)'\,d\Omega \tag{h}$$

Figures (a) to (d) give some results given by MATHFEM with a regular mesh of four (upper row) and nineteen (lower row) two-noded elements. Results on the left (Figures (a) and (c)) are obtained by the Newton-Raphson method and sensitizing only with $\tau^c$ (GLS) and on the right (Figures (b) and (d)) sensitizing with both $\tau^c$ and $\tau^r$ (GGLS). The solid lines illustrate the initial guess and solutions after the first, second and final iteration. The dashed line gives the exact solution. Because the sensitizing parameter expressions used are not smooth, they have been treated in a "Picard fashion" in the program.

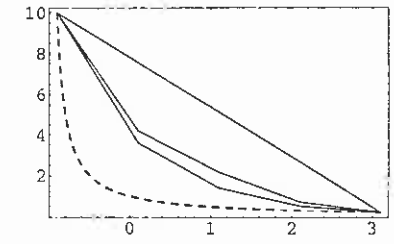



**Figure (a)**

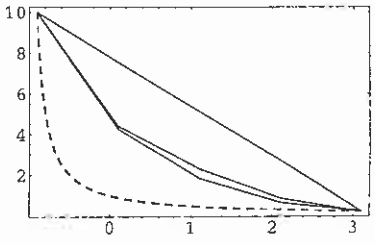



**Figure (b)**

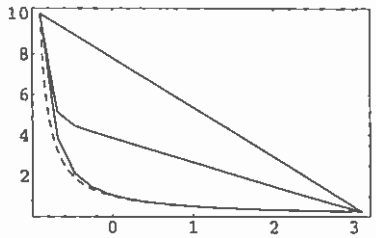



Figure (c)

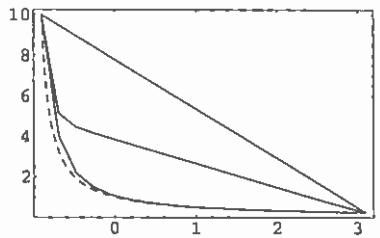



Figure (d)

Table (a) gives some results for the error quantity defined as in Example 11.1, formula (l), for the case with four elements. Use of the GGLS-version increases the rate of convergence. The difference in the table values with respect to the initial guess ($n = 0$) depends on the difference in the final discrete solutions.

**Table (a)**

| Relative rms-error in the solution | | |
|---|---|---|
| Iteration | $\tau^c \neq 0, \tau^r = 0$ | $\tau^c \neq 0, \tau^r \neq 0$ |
| 1 | 0.54272 | 0.46027 |
| 2 | 0.21721 | 0.16041 |
| 3 | 0.07968 | 0.04835 |
| 4 | 0.02461 | 0.00826 |
| 5 | 0.00685 | 0.00069 |
| 6 | 0.00183 | 0.00012 |

A sample of the MATHFEM program used is given:

```
<< mathfem.m;
nn = 20;
x1 = -0.9; x2 = 3.1; φ1 = 10; φ2 = 1/4.1;
d = 1.; u = -2*φ[0]; v = -2*φ[0]; c = -2*φ[1];
h = (x2 - x1) / (nn - 1);
tc = 1/Max[12*d/h^2, 2*Abs[v]/h];
tr = h^2/Max[108*d/h^2, 6*Abs[c]];
exact[x_] := 1/(1+x);
φ̂ = φ1 + (#-x1) / (x2 - x1) * (φ2 - φ1) &;
dom = {{x1}, {x2}};
msh = MSH[dom, {nn}, 2];
apr = APR[msh, φ̂];
prb =
  PRB[apr, {0, w[1] * φ[1] + w[0] *u*φ[1] + tc* (v*w[1]) * (v*φ[1]) + tr* (c*w[1]) * (c*φ[1])
prb = FIX[prb, dom, {{φ1}, {φ2}}];
SHOWID[PLOT[NONLINEAR[prb, 0.0001]]];
```



### 11.1.4 Radiation boundary condition

Thermal radiation heat transfer between communicating surfaces and participating media is an extremely complicated phenomenon to analyze and massive textbooks have been written on the subject, e.g., Siegel and Howell (1996). We here consider one simplified example case. The heat flow rate density $q_n^r$ by radiation from a surface of a body is given by the expression, e.g., Incropera and DeWitt (1996, p. 10),

$$q^r = \varepsilon\sigma\left(T^4 - T_{sur}^4\right) \tag{26}$$

Here $T$ and $T_{sur}$ are the termodynamic temperatures of the body surface and the surrounding surface, respectively, $\sigma = 5.67 \cdot 10^{-8}\,\mathrm{W/(m^2K^4)}$ is the *Stefan-Boltzmann constant* and $\varepsilon$ ($[\varepsilon] = -$) is the *emissivity* (emissiivisyys) of body surface. This expression is valid if the surrounding surface is large compared to the body surface the and the medium between the surfaces does not participate in radiation.

Using the deltaform and writing the right hand side of (26) as $F(T)$, we obtain by truncated Taylor series

$$F \approx \overline{F} + \frac{dF}{dT}\Delta T \tag{27}$$

or

$$q^r = \varepsilon\sigma\left(\overline{T}^4 - T_{sur}^4\right) + 4\varepsilon\sigma\overline{T}^3\Delta T \tag{28}$$

The right-hand side corresponds to a Robin boundary condition term for $\Delta T$.

A rather common Picard type approach is to use the relation

$$\left(T^4 - T_{sur}^4\right) = \left(T + T_{sur}\right)\left(T^2 + T_{sur}^2\right)\left(T - T_{sur}\right) \tag{29}$$

and write (26) as

$$q^r = h_r\left(T - T_{sur}\right) \tag{30}$$

where

$$h_r = \varepsilon\sigma\left(T + T_{sur}\right)\left(T^2 + T_{sur}^2\right) \tag{31}$$

The term $h_r$ ($[h_r] = \mathrm{W/(m^2K)}$) is called *radiation heat transfer coefficient* (säteilyn lämmönsiirtymiskerroin). Using form (30), we have obtained formally the familiar convection type boundary condition term; see formula (3.1.7).

**Example 11.3.** We use here again an example case of heat transfer taken from Çengel (1998, p. 93). The problem can be described by the differential equation

$$-kT'' = 0, \qquad 0 < x < L \tag{a}$$

and by the boundary conditions

$$-kT'(L) = \varepsilon\sigma\left[T(L)^4 - T_{\mathrm{sur}}^4\right] \tag{b}$$

On the right-hand boundary, heat flux by radiation is taking place which makes the determination of the temperature distribution non-linear.

Using the deltaform and the Newton-Raphson method the weak form in an iteration step is

$$\int_\Omega w'k\left(\Delta T\right)'\mathrm{d}\Omega + \int_\Omega w'k\bar{T}'\,\mathrm{d}\Omega + w\varepsilon\sigma\left[\bar{T}^4 - T_{\mathrm{sur}}^4\right]\Big|_{x=L} + w4\varepsilon\sigma\bar{T}^3\Delta T\big|_{x=L} = 0 \tag{c}$$

In the Picard method the weak form in an iteration step is

$$\int_\Omega w'kT'\mathrm{d}\Omega + wh_{\mathrm{r}}T\big|_{x=L} - wh_{\mathrm{r}}T_{\mathrm{sur}}\big|_{x=L} = 0 \tag{d}$$

with

$$h_{\mathrm{r}} = \varepsilon\sigma\left(\bar{T} + T_{\mathrm{sur}}\right)\left(\bar{T}^2 + T_{\mathrm{sur}}^2\right) \tag{e}$$

We take the example data

$$\begin{aligned} &L = 0.06\,\mathrm{m}, \quad k = 1.2\,\mathrm{W/(m\cdot K)}, \quad \varepsilon = 0.85, \\ &T_1 = 300\,\mathrm{K}, \quad T_{\mathrm{sur}} = 0\,\mathrm{K} \end{aligned} \tag{f}$$

Figure (a) shows results obtained by the Newton-Raphson method and Figure (b) by the Picard method given by MATHFEM. A regural mesh of four two-noded elements has been used. The solid lines illustrate the initial guessed solution (constant) and the solution after the first iteration.

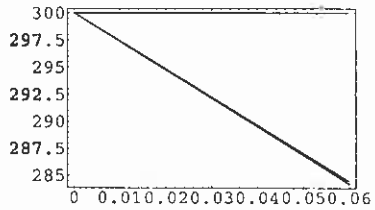



**Figure (a)**

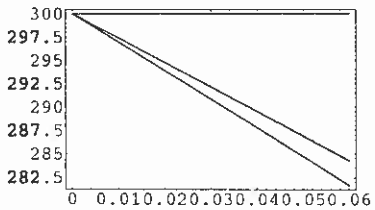



**Figure (b)**



Table (a) contains similar information as the corresponding table of Example 11.1.

**Table (a)**

| Relative rms-error in the solution | | |
|---|---|---|
| Iteration | Newton | Picard |
| 1 | $3.3\ 10^{-2}$ | $3.3\ 10^{-2}$ |
| 2 | $5.2\ 10^{-4}$ | $5.4\ 10^{-3}$ |
| 3 | $1.2\ 10^{-7}$ | $8.4\ 10^{-4}$ |
| 4 | $6.8\ 10^{-15}$ | $1.3\ 10^{-4}$ |
| 5 | - | $2.1\ 10^{-5}$ |
| 6 | - | $3.3\ 10^{-6}$ |

Finally, a sample of the MATHFEM program used is again given:

```
<< mathfem.m;
nn = 10;
L = 0.06; x1 = 0; x2 = L; φ1 = 300; φ2 = 0; d = 1.2; ε = 0.85; σ = 5.67*10^-8;
φ = φ1 + 0*#&;
dom = {{x1}, {x2}};
msh = MSH[dom, {nn}, 2] ;
apr = APR[msh, φ] ;
prb = PRB[apr, {0, w[1]*d*φ[1], w[0]*ε*σ*(φ[0]^4 - φ2^4)}];
prb = FIX[prb, {x1}, {{φ1}}];
prb[[2, 2, nn - 1, 3]] = 3;
SHOWLD[PLOT[NONLINEAR[prb, 0.0001]]];
```

## 11.1.5 Some comments

The above example cases indicate that the way to apply the Picard method is not unique. In non-linear cases the convergence of a procedure is never sure. Combinations of the Picard and the Newton-Raphson methods are sometimes advocated in the literature.

## 11.2 TRANSIENT CASE (missing)

### 11.2.1 Introduction

### 11.2.2 Applications

## REFERENCES

Çengel, Y. A. (1998). *Heat Transfer, A Practical Approach*, McGraw-Hill, Boston, ISBN 0-07-115223-7.

Incropera, F. P. and DeWitt, D. P. (1996). *Fundamentals of Heat Transfer*, 4th ed.,Wiley, New York, ISBN 0-471-30460-3.

Shih, T.-M. (1984). *Numerical Heat Transfer*, Hemisphere Publishing Corporation, Washington, ISBN 0-89116-257-7.

# 12 FLUID FLOW

## 12.1 MOMENTUM EQUATIONS

This far we have dealt mainly — except in Chapter 5 — with problems concerning only one field equation and one unknown function to be determined. Problems with fluids in motion usually mean that a few (say at least three or more) unknown functions with coupled field equations have to be considered simultaneously.

The general local form of the principle of balance of momentum (liikemäärän taseen periaate), e.g., Malvern (1969), is

$$\boxed{\nabla \cdot \boldsymbol{\sigma} + \rho \mathbf{b} = \rho \mathbf{a}} \qquad \frac{\partial \sigma_{\beta\alpha}}{\partial x_\beta} + \rho b_\alpha = \rho a_\alpha \tag{1}$$

where $\boldsymbol{\sigma}$ is the *stress tensor*, $\rho$ the *density*, $\mathbf{b}$ the *specific body force* (massavoiman intensiteetti) ($[\mathbf{b}] = \mathrm{N/kg}$), and $\mathbf{a}$ the *acceleration* (kiihtyvyys) ($[\mathbf{a}] = \mathrm{m/s^2}$) of the medium. Equation (1) is usually called the *momentum equation* (or equations) or the *equations of motion* (liikemääräyhtälö, liikeyhtälö). If the continuum is at rest, $\mathbf{a} = \mathbf{0}$, the *equilibrium equations* (tasapainoyhtälö) are arrived at.

The stress vector or traction $\mathbf{t}$ (jännitysvektori, traktio) ($[\mathbf{t}] = \mathrm{N/m^2}$), acting on a differential surface element with the unit outward normal vector $\mathbf{n}$, is connected to the stress tensor through

$$\boxed{\mathbf{t} = \mathbf{n} \cdot \boldsymbol{\sigma}} \qquad t_\alpha = n_\beta \sigma_{\beta\alpha} \tag{2}$$

This relationship also follows from the principle of balance of momentum. As mentioned in Section 3.1.1, the relationship is analogous to (3.1.3).

From the principle of balance of moment of momentum (liikemäärämomentin taseen periaate) follows the result

$$\boxed{\boldsymbol{\sigma}^{\mathrm{T}} = \boldsymbol{\sigma}} \qquad \sigma_{\beta\alpha} = \sigma_{\alpha\beta} \tag{3}$$

or the stress tensor is *symmetric*.

In the Eulerian description the kinematic relation

$$\boxed{\mathbf{a} = \frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v}} \qquad a_\alpha = \frac{\partial v_\alpha}{\partial t} + v_\beta \frac{\partial v_\alpha}{\partial x_\beta} \tag{4}$$





is valid, where $\mathbf{v}$ is the *velocity* of the medium. The latter term in (4) — the *convective acceleration* — is the main cause of complexity in fluid mechanics generating non-linearity into the equations.

In fluid mechanics the stress tensor is usually decomposed into an isotropic part and into a deviatoric part:

$$\boldsymbol{\sigma} = -p\mathbf{I} + \boldsymbol{\sigma}^*, \qquad \sigma_{\alpha\beta} = -p\delta_{\alpha\beta} + \sigma^*_{\alpha\beta} \tag{5}$$

where the multiplier in the isotropic part is called the *pressure* (paine):

$$p = -\frac{1}{3}\sigma_{\gamma\gamma} \tag{6}$$

The deviatoric stress is often called the *viscous stress* (viskoosi jännitys) as in fluid at rest it vanishes so it is caused by friction or viscosity when the fluid is in motion. In the ideal fluid flow model it is assumed to vanish also when the fluid is in motion.

Substitution of (4) and (5) into (1) gives the form

$$\boxed{\rho \frac{\partial \mathbf{v}}{\partial t} - \nabla \cdot \boldsymbol{\sigma}^* + \rho \mathbf{v} \cdot \nabla \mathbf{v} + \nabla p - \rho \mathbf{b} = \mathbf{0}} \tag{7a}$$

or

$$\rho \frac{\partial v_\alpha}{\partial t} - \frac{\partial \sigma^*_{\beta\alpha}}{\partial x_\beta} + \rho v_\beta \frac{\partial v_\alpha}{\partial x_\beta} + \frac{\partial p}{\partial x_\alpha} - \rho b_\alpha = 0 \tag{7b}$$

These differ somewhat from equations (A.2.6). They can be arrived at from (A.2.6) by some manipulation with the continuity equation (A.2.1).

The equations above are exact. To proceed, we have to make constitutive assumptions.

The constitutive law for the pressure is generally of the type $p = p(\rho, T)$ such as the ideal gas law $p = R\rho T$. Here we assume the mechanically incompressible fluid model discussed in Remark 6.1 so the pressure is a constraint force and it has no constitutive expression.

For the viscous stress we employ the conventional *Stokes' viscosity law* (Stokesin kitkalaki)

$$\sigma^*_{\alpha\beta} = 2\mu d_{\alpha\beta} + \lambda d_{\gamma\gamma}\delta_{\alpha\beta} \tag{8}$$

where the meaning of the notations has been explained in connection with equation (6.1.8) and where the deformation rate

$$d_{\alpha\beta} = \frac{1}{2}\left( \frac{\partial v_\alpha}{\partial x_\beta} + \frac{\partial v_\beta}{\partial x_\alpha} \right) \tag{9}$$

In the fully incompressible case the dilatation rate $d_{\gamma\gamma}$ vanishes. In our slightly incompressible fluid model we ignore the dilatation rate in (8). Substitution of expressions (8) and (9) into (7) gives the momentum equations

$$\boxed{\rho \frac{\partial v_\alpha}{\partial t} - \frac{\partial}{\partial x_\beta}\mu\left( \frac{\partial v_\alpha}{\partial x_\beta} + \frac{\partial v_\beta}{\partial x_\alpha} \right) + \rho v_\beta \frac{\partial v_\alpha}{\partial x_\beta} + \frac{\partial p}{\partial x_\alpha} - \rho b_\alpha = 0} \tag{10}$$

These are the famous *Navier-Stokes equations* (Navier-Stokesin yhtälöt). They are written here using the incompressibility condition in the Stokes' viscosity law. The Stokes' viscosity law is known to be accurately valid for many common fluids such as water and air.

In the *fully incompressible case* with a homogeneous fluid the density $\rho$ *is a given constant for the fluid.* This is an often-used model for liquids in forced convection.

To be able to treat natural convection, see Remark 6.1, the density changes generated via temperature changes must be accounted for. In this connection the so-called *Boussinesq-approximation* is often used:

> The density of the fluid is assumed to be constant in all the governing equations except in the body force term in the momentum equations.

The validity of this approximation can be shown by making the equations dimensionless and by studying the order of magnitude of the different terms.

Equation (6.1.15) gives the constitutive relation

$$\rho = \rho^\circ - \gamma_p \rho^\circ \Delta T \tag{11}$$

which can be used here. When this is substituted in (10) we obtain (The superscript $\circ$ is removed and $\rho$ now denotes the constant reference density associated with the reference temperature.)



$$\rho \frac{\partial v_\alpha}{\partial t} - \frac{\partial}{\partial x_\beta}\mu\left( \frac{\partial v_\alpha}{\partial x_\beta} + \frac{\partial v_\beta}{\partial x_\alpha} \right) + \rho v_\beta \frac{\partial v_\alpha}{\partial x_\beta} + \frac{\partial p}{\partial x_\alpha} + \gamma_p \rho \Delta T b_\alpha - \rho b_\alpha = 0 \tag{12}$$

The part $\gamma_p \rho \Delta T b_\alpha$ is often called the *buoyancy term* (nostetermi) especially if the body force is due to gravity in which case $\mathbf{b} = \mathbf{g}$. The buoyancy term is seen to couple the momentum equations through the temperature with the energy equation; see Remark 5.3. Further, if the dependence of viscosity on temperature is taken into account, additional coupling is introduced.

The mechanical boundary conditions associated with fluid flow consist usually of given velocities:

$$v_\alpha = \bar{v}_\alpha \quad \text{on } \Gamma_v \tag{13}$$

and of given tractions

$$t_\alpha = \bar{t}_\alpha \quad \text{on } \Gamma_\sigma \tag{14}$$

The *velocity boundary* $\Gamma_v$ and the *traction boundary* $\Gamma_\sigma$ form together the whole space boundary $\Gamma$. Using equations (2) and (5), the latter condition obtains the form

$$n_\beta \sigma^*_{\beta\alpha} - n_\alpha p = \bar{t}_\alpha \tag{15}$$

Finally, employing the constitutive law (8) in the incompressible case, the condition transforms to

$$n_\beta \mu\left( \frac{\partial v_\beta}{\partial x_\alpha} + \frac{\partial v_\alpha}{\partial x_\beta} \right) - n_\alpha p = \bar{t}_\alpha \tag{16}$$

The conditions (13) and (16) are clearly analogous to the Dirichlet and Neumann conditions, respectively, used in problems with only one unknown function.

The boundary conditions (13) and (14) are the conventional ones but many alternatives have been used. The theory seems not to be quite complete concerning all the possibilities. Especially the inflow and outflow boundaries (see Section A.3) are usually synthetic surfaces decided on by the applier dividing the fluid domain rather arbitrarily without any real physical basis. To guess realistic boundary conditions for them is often difficult.

**Remark 12.1.** The velocity and traction boundary conditions (13) and (14) have been presented above in a simplified form. The more general forms are for any point on the boundary

$$\begin{aligned} v_1' &= \overline{v}_1' \quad \text{or} \quad t_1' = \overline{t}_1' \\ v_2' &= \overline{v}_2' \quad \text{or} \quad t_2' = \overline{t}_2' \\ v_3' &= \overline{v}_3' \quad \text{or} \quad t_3' = \overline{t}_3' \end{aligned} \tag{17}$$

From each row either of the conditions (but not both) is taken. The notations refer to a local rectangular coordinate system with one coordinate axis usually in the normal direction to the boundary. The application of these forms demands use of certain transformation formulas between the local and global coordinate systems. □

## 12.2 GOVERNING EQUATIONS FOR FLUID FLOW

We collect here a fairly general set of governing equations for slightly compressible fluid flow.

*Field equations*:

Momentum equations

$$\rho\frac{\partial v_\alpha}{\partial t}-\frac{\partial}{\partial x_\beta}\mu\left(\frac{\partial v_\alpha}{\partial x_\beta}+\frac{\partial v_\beta}{\partial x_\alpha}\right)+\rho v_\beta\frac{\partial v_\alpha}{\partial x_\beta}+\frac{\partial p}{\partial x_\alpha}+\gamma_p\rho\Delta T b_\alpha-\rho b_\alpha=0 \tag{1}$$

Continuity equation

$$\frac{\partial v_\alpha}{\partial x_\alpha}=0 \tag{2}$$

Energy equation

$$\rho c_p\frac{\partial T}{\partial t}+\frac{\partial}{\partial x_\alpha}\left(-k_{\alpha\beta}\frac{\partial T}{\partial x_\beta}\right)+\rho c_p v_\alpha\frac{\partial T}{\partial x_\alpha}-s-\Phi=0 \tag{3}$$

*Mechanical boundary conditions*:

Given velocity

$$v_\alpha=\overline{v}_\alpha \quad \text{on } \Gamma_v \tag{4}$$

Given traction

$$n_\beta\mu\left(\frac{\partial v_\beta}{\partial x_\alpha}+\frac{\partial v_\alpha}{\partial x_\beta}\right)-n_\alpha p=\overline{t}_\alpha \quad \text{on } \Gamma_\sigma \tag{5}$$

*Thermal boundary conditions*:

Dirichlet



$$T=\overline{T} \quad \text{on } \Gamma_D \tag{6}$$

Neumann

$$-n_\alpha k_{\alpha\beta}\frac{\partial T}{\partial x_\beta}=\overline{q} \quad \text{on } \Gamma_N \tag{7}$$

Robin

$$-n_\alpha k_{\alpha\beta}\frac{\partial T}{\partial x_\beta}=h\left(T-T_\infty\right) \quad \text{on } \Gamma_R \tag{8}$$

*Initial conditions*:

Given velocity

$$v_\alpha=\left(\overline{v}_\alpha\right)_0 \quad \text{in } \Omega \text{ at } t=0 \tag{9}$$

Given temperature

$$T=\overline{T}_0 \quad \text{in } \Omega \text{ at } t=0 \tag{10}$$

The equations have been presented in such a form that the basic unknown functions appearing are $v_\alpha$, $p$, $T$. In three dimensions there are thus $3+1+1=5$ unknowns. Most of the equations have appeared earlier in the text. The Boussinesq approximation for slightly compressible fluids has been used so that for example the continuity equation is in fact the incompressibility condition. These equations must be solved in the general case coupled, that is, simultaneously.

The equations are valid for laminar flows (and also for turbulent flows but the minute details of these latter flows cannot yet in practice be simulated with present computers). In turbulent flows turbulent modeling is practised which means that typically at least two additional diffusion-convection-reaction type field equations emerge, e.g., Wilcox (1994). Flow of mixtures also demand additional equations describing concentration of species.

In fully compressible fluid flow as in gas dynamics a constitutive law $p=p(\rho,T)$ such as the ideal gas law must be introduced and the density is no more a given constant but one of the unknowns.

## 12.3 STOKES PROBLEM

### 12.3.1 General considerations

By the Stokes problem is usually meant the case where the inertia forces due to fluid acceleration can be neglected in comparison to the viscous forces, that is, a very low Reynolds number flow (see Section A.2). This case is a suitable simple starting point for looking at the formulations needed.

We have from Sections 12.1 and 12.2 here the field equations

$$R_\alpha(v_\alpha,p)\equiv L_\alpha(v_\alpha,p)-\rho b_\alpha\equiv\boxed{-\frac{\partial\sigma^*_{\beta\alpha}}{\partial x_\beta}+\frac{\partial p}{\partial x_\alpha}-\rho b_\alpha=0}\quad\text{in }\Omega \tag{1a}$$

$$R(v_\alpha)\equiv L(v_\alpha)\equiv\boxed{-\frac{\partial v_\alpha}{\partial x_\alpha}=0}\quad\text{in }\Omega \tag{2a}$$

(cf. Remark 12.2) and the boundary conditions

$$\boxed{v_\alpha=\bar v_\alpha}\quad\text{on }\Gamma_v \tag{3a}$$

$$\boxed{n_\beta\sigma^*_{\beta\alpha}-n_\alpha p=\bar t_\alpha}\quad\text{on }\Gamma_\sigma \tag{4a}$$

It is again convenient to introduce the constitutive relation

$$\sigma^*_{\beta\alpha}=\mu\left(\frac{\partial v_\beta}{\partial x_\alpha}+\frac{\partial v_\alpha}{\partial x_\beta}\right) \tag{5a}$$

later into the formulation. The unknown functions to be determined are $v_\alpha(\mathbf{x})$ and $p(\mathbf{x})$.

**Remark 12.2.** From this on we write the continuity equation often with a negative sign on the left-hand side. By this choice it is found that the resulting discrete equations following from the weak form become symmetric without later changes of signs for the corresponding weighting function; see Remark 12.5. □

To make the formulas perhaps more familiar we further write down the two-dimensional version with the notational changes $x_1\to x$, $x_2\to y$, $v_1\to u$, $v_2\to v$, similarly as in Section A.3:

$$\begin{aligned}R_x(u,v,p)&\equiv L_x(u,v,p)-\rho b_x\equiv-\frac{\partial\sigma^*_{xx}}{\partial x}-\frac{\partial\sigma^*_{yx}}{\partial y}+\frac{\partial p}{\partial x}-\rho b_x=0\\ R_y(u,v,p)&\equiv L_y(u,v,p)-\rho b_y\equiv-\frac{\partial\sigma^*_{xy}}{\partial x}-\frac{\partial\sigma^*_{yy}}{\partial y}+\frac{\partial p}{\partial y}-\rho b_y=0\end{aligned} \tag{1b}$$



$$R(u,v)\equiv L(u,v)\equiv-\frac{\partial u}{\partial x}-\frac{\partial v}{\partial y}=0 \tag{2b}$$

$$\begin{aligned}u&=\bar u\\ v&=\bar v\end{aligned} \tag{3b}$$

$$\begin{aligned}n_x\sigma^*_{xx}+n_y\sigma^*_{yx}-n_x p&=\bar t_x\\ n_y\sigma^*_{yx}+n_y\sigma^*_{yy}-n_y p&=\bar t_y\end{aligned} \tag{4b}$$

the unknown functions to be determined are $u(x,y)$, $v(x,y)$, and $p(x,y)$. The analogues of (5a) are

$$\begin{aligned}\sigma^*_{xx}&=\mu\left(\frac{\partial u}{\partial x}+\frac{\partial u}{\partial x}\right),\quad&\sigma^*_{xy}&=\mu\left(\frac{\partial u}{\partial y}+\frac{\partial v}{\partial x}\right)\\ \sigma^*_{yx}&=\mu\left(\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}\right),\quad&\sigma^*_{yy}&=\mu\left(\frac{\partial v}{\partial y}+\frac{\partial v}{\partial y}\right)\end{aligned} \tag{5b}$$

These must be considered to be introduced in (1) when the arguments in the residuals are listed. We will consider in the following even when using the index notation only the two-dimensional case. There are three field equations and three residual expressions. The ideal situation would obviously be the case where in a problem with certain unknown functions, each field equation would the populated from terms of only one of the unknowns, that is, the equations would be uncoupled. Here, the situation is far from this ideal one. In any case, the $x$- and $y$-component momentum equations can be rather clearly considered to have as their "main variables" the $x$- and $y$-axis velocity components $u$ and $v$, respectively, and we would thus want to associate the pressure with the continuity equation. However, the continuity equation is seen not to contain the pressure at all. In compressible flow the continuity equation contains the pressure when the constitutive law $p=p(\rho,T)$ or $\rho=\rho(p,T)$ is made use of. The fact that the pressure is missing in the incompressibility equation has caused much trouble in solving incompressible or nearly incompressible flow problems numerically and rather complicated formulations have been developed to somehow take care of this feature.

**Remark 12.3.** In the finite element method it has been necessary until quite recently, Hughes et al. (1986), to select the type of approximation between the velocity and pressure very carefully to obtain a working formulation; the approximation for the pressure has had to be of lower order than that for the velocity. By introducing certain sensitizing terms this problem no more exists and equal order approximation can be used. □

**Remark 12.4** One way to deal with the problem of $p$ missing from the incompressibility condition is the so-called *penalty formulation* (sakkoformulaatio). It is based on replacing (2) with a *perturbed* (häiritty) form

$$\frac{1}{\lambda}p+\frac{\partial v_\alpha}{\partial x_\alpha}=0 \tag{6}$$

where $\lambda$ is a given large number with appropriate dimension, the *penalty parameter* (sakkoparametri). The pressure has now been introduced artificially into the formulation. (In fact, to see this result to be valid, a study based on a variational formulation of the problem is needed.) As the pressure is for physical reasons bounded and if $\lambda$ is taken to be very large, equation (6) is nearly the same as equation (2). From (6) follows a fictitious constitutive relation

$$p=-\lambda\frac{\partial v_\alpha}{\partial x_\alpha} \tag{7}$$

with $\lambda$ as kind of dilatational viscosity. By substituting (7) into the momentum equations the pressure is eliminated and only the velocity components remain to be determined. The penalty formulation has been employed quite widely in connection with the finite element method. However, although the formulation is very simple, it has some disadvantages and is not considered further in this text. □

We now proceed to derive a weak form corresponding to the Stokes problem. The procedure is a rather obvious generalization from what has been done earlier; see Remark 5.6. The first field equation is multiplied by an arbitrary weighting function $w_1(x_1,x_2)=w_x(x,y)$, the second by $w_2(x_1,x_2)=w_y(x,y)$, and the third by $w(x_1,x_2)=w(x,y)$, the resulting equations are integrated over the domain $\Omega$ and added together to produce a weak form

$$F\equiv\int_\Omega\left(w_\alpha R_\alpha+wR\right)\mathrm{d}\Omega=0 \tag{8a}$$

or

$$F\equiv\int_\Omega\left(w_xR_x+w_yR_y+wR\right)\mathrm{d}\Omega=0 \tag{8b}$$

As the weighting functions are arbitrary, this one scalar equation is equivalent to the three field equations.

**Remark 12.5.** Of course, the integrand in (8b) can be written equally well as $w_xR_x+w_yR_y-wR$ or $w_xR_x-w_yR_y-wR$ etc., that is, we can take the signs arbitrarily. This is because the weighting functions are arbitrary and the weak form still remains equivalent to the field equations. Further, we can write the original field equations with changed signs, which would apparently again change the outlook of the detailed weak form. We have here wanted to leave the weak form to have the clean standard outlook with only plus signs. Similarly, we finally want to use the finite dimensional weightings corresponding to (15) in the forms $\tilde{w}_x=N_i$, $\tilde{w}_y=N_i$, $w=N_i$, without any possible changes in signs. In this way it is found that the discrete equations become symmetric, which explains the choice discussed in Remark 12.2. The reason for the symmetry is based on the fact that the Stokes problem can actually be presented also as an equivalent variational principle. □



The next step similarly as with diffusion problems is to integrate by parts the terms containing through the constitutive relation second derivatives (see formula (B.3.1b):

$$-\int_\Omega\left(w_\alpha\frac{\partial\sigma^*_{\beta\alpha}}{\partial x_\beta}\right)\mathrm{d}\Omega=\int_\Omega\left(\frac{\partial w_\alpha}{\partial x_\beta}\sigma^*_{\beta\alpha}\right)\mathrm{d}\Omega-\int_\Gamma w_\alpha\sigma^*_{\beta\alpha}n_\beta\mathrm{d}\Gamma \tag{9a}$$

or

$$\begin{aligned}&-\int_\Omega\left[w_x\left(\frac{\partial\sigma^*_{xx}}{\partial x}+\frac{\partial\sigma^*_{yx}}{\partial y}\right)+w_y\left(\frac{\partial\sigma^*_{xy}}{\partial x}+\frac{\partial\sigma^*_{yy}}{\partial y}\right)\right]\mathrm{d}\Omega=\\&+\int_\Omega\left(\frac{\partial w_x}{\partial x}\sigma^*_{xx}+\frac{\partial w_x}{\partial y}\sigma^*_{yx}+\frac{\partial w_y}{\partial x}\sigma^*_{xy}+\frac{\partial w_y}{\partial y}\sigma^*_{yy}\right)\mathrm{d}\Omega\\&-\int_\Gamma\left(w_x\sigma^*_{xx}n_x+w_x\sigma^*_{yx}n_y+w_y\sigma^*_{xy}n_x+w_y\sigma^*_{yy}n_y\right)\mathrm{d}\Gamma\end{aligned} \tag{9b}$$

In this kind of manipulations the index notation and the use of the summation convention soon becomes the more attractive alternative — even if the formulas may seem at first sight too concise — as the manipulations become otherwise extremely painful to write down.

Here it proves useful to integrate by parts also the terms containing the pressure:

$$\int_\Omega w_\alpha\frac{\partial p}{\partial x_\alpha}\mathrm{d}\Omega=-\int_\Omega\frac{\partial w_\alpha}{\partial x_\alpha}p\,\mathrm{d}\Omega+\int_\Gamma w_\alpha p\,n_\alpha\mathrm{d}\Gamma \tag{10}$$

This is because the pressure now appears also on the boundary and this can be made use of on the traction boundary $\Gamma_\sigma$.

The weak form is transformed into

$$\begin{aligned}F\equiv&\int_\Omega\left(\frac{\partial w_\alpha}{\partial x_\beta}\sigma^*_{\beta\alpha}-\frac{\partial w_\alpha}{\partial x_\alpha}p+w\frac{\partial v_\alpha}{\partial x_\alpha}-w_\alpha\rho b_\alpha\right)\mathrm{d}\Omega\\&-\int_\Gamma w_\alpha\left(n_\beta\sigma^*_{\beta\alpha}-n_\alpha p\right)\mathrm{d}\Gamma=0\end{aligned} \tag{11}$$

The boundary conditions have not yet been discussed. The velocity boundary conditions (3) are assumed to be satisfied in advance. Similarly as in connection with the diffusion problem we then demand that the admissible "velocity" weighting functions satisfy

$$w_\alpha = 0 \quad \text{on } \Gamma_v \tag{12}$$

Then there remains in the boundary integral in (11) only the part over the traction boundary $\Gamma_\sigma$. But there the traction boundary condition (4) shows that the term inside the parentheses is $\bar{t}_\alpha$ and this given value can be substituted there. Introducing further the constitutive expression (5), we obtain finally

$$F \equiv \boxed{\begin{aligned} &\int_\Omega \left[ \frac{\partial w_\alpha}{\partial x_\beta} \mu \left( \frac{\partial v_\beta}{\partial x_\alpha} + \frac{\partial v_\alpha}{\partial x_\beta} \right) - \frac{\partial w_\alpha}{\partial x_\alpha} p - w \frac{\partial v_\alpha}{\partial x_\alpha} \right] \mathrm{d}\Omega \\ &- \int_\Omega w_\alpha \rho b_\alpha \mathrm{d}\Omega - \int_{\Gamma_\sigma} w_\alpha \bar{t}_\alpha \, \mathrm{d}\Gamma = 0 \end{aligned}} \tag{13}$$

### 12.3.2 Sensitized form

**Introduction.** We now develop the expressions needed in two dimensions in more detail. We also change the notation somewhat. For instance, $w_x \to w_u$, $R_x \to R_u$ etc. The left-hand side in (13) is

$$\begin{aligned} F = \int_\Omega \Bigg[ & \frac{\partial w_u}{\partial x} \mu \left( \frac{\partial u}{\partial x} + \frac{\partial u}{\partial x} \right) + \frac{\partial w_u}{\partial y} \mu \left( \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \right) \\ & + \frac{\partial w_v}{\partial x} \mu \left( \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \right) + \frac{\partial w_v}{\partial y} \mu \left( \frac{\partial v}{\partial y} + \frac{\partial v}{\partial y} \right) \\ & - \frac{\partial w_u}{\partial x} p - \frac{\partial w_v}{\partial y} p - w_p \left( \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right) \Bigg] \mathrm{d}\Omega \\ & - \int_\Omega \left( w_u \rho b_x + w_v \rho b_y \right) \mathrm{d}\Omega - \int_{\Gamma_\sigma} \left( w_u \bar{t}_x + w_v \bar{t}_y \right) \mathrm{d}\Gamma \end{aligned} \tag{14}$$

We use the same type finite element approximations for all the unknowns:

$$\begin{aligned} \tilde{u}(x,y) &= \sum_j N_j(x,y) u_j \\ \tilde{v}(x,y) &= \sum_j N_j(x,y) v_j \\ \tilde{p}(x,y) &= \sum_j N_j(x,y) p_j \end{aligned} \tag{15}$$

Using the Galerkin method, the contributions to the three system equations corresponding to node $i$ are thus (put first $\bar{w}_u = N_i$, $\bar{w}_v = 0$, $\bar{w}_p = 0$, second $\bar{w}_u = 0$, $\bar{w}_v = N_i$, $\bar{w}_p = 0$, third $\bar{w}_u = 0$, $\bar{w}_v = 0$, $\bar{w}_p = N_i$)



$$\begin{aligned} \left(F_u\right)_i &= \int_\Omega \left[ \frac{\partial N_i}{\partial x} \mu 2 \frac{\partial \tilde{u}}{\partial x} + \frac{\partial N_i}{\partial y} \mu \left( \frac{\partial \tilde{v}}{\partial x} + \frac{\partial \tilde{u}}{\partial y} \right) - \frac{\partial N_i}{\partial x} \tilde{p} \right] \mathrm{d}\Omega \\ &\quad - \int_\Omega N_i \rho b_x \, \mathrm{d}\Omega - \int_{\Gamma_\sigma} N_i \bar{t}_x \, \mathrm{d}\Gamma \\ \left(F_v\right)_i &= \int_\Omega \left[ \frac{\partial N_i}{\partial x} \mu \left( \frac{\partial \tilde{u}}{\partial y} + \frac{\partial \tilde{v}}{\partial x} \right) + \frac{\partial N_i}{\partial y} \mu 2 \frac{\partial \tilde{v}}{\partial y} - \frac{\partial N_i}{\partial y} \tilde{p} \right] \mathrm{d}\Omega \\ &\quad - \int_\Omega N_i \rho b_y \, \mathrm{d}\Omega - \int_{\Gamma_\sigma} N_i \bar{t}_y \, \mathrm{d}\Gamma \\ \left(F_p\right)_i &= -\int_\Omega N_i \left( \frac{\partial \tilde{u}}{\partial x} + \frac{\partial \tilde{v}}{\partial y} \right) \mathrm{d}\Omega \end{aligned} \tag{16}$$

As mentioned in Remark 12.3, some sensitizing terms are needed to make the discrete version work with equal order approximation for velocity and pressure. Proceeding similarly as in Section 5.3, we start from the additional least squares functional

$$\Pi(u,v,p) = \frac{1}{2} \int_\Omega \begin{Bmatrix} R_u \\ R_v \\ R_p \end{Bmatrix}^{\mathrm{T}} \begin{bmatrix} \tau_{uu} & \tau_{uv} & \tau_{up} \\ \tau_{vu} & \tau_{vv} & \tau_{vp} \\ \tau_{pu} & \tau_{pv} & \tau_{pp} \end{bmatrix} \begin{Bmatrix} R_u \\ R_v \\ R_p \end{Bmatrix} \mathrm{d}\Omega \tag{17}$$

Taking the variation of (17) and introducing the interpretations $\delta u = w_u$, $\delta v = w_v$, $\delta p = w_p$, gives the sensitizing term

$$F^{(0)} = \int_\Omega \begin{Bmatrix} L_u(w_u, w_v, w_p) \\ L_v(w_u, w_v, w_p) \\ L_p(w_u, w_v) \end{Bmatrix}^{\mathrm{T}} \begin{bmatrix} \tau_{uu} & \tau_{uv} & \tau_{up} \\ \tau_{vu} & \tau_{vv} & \tau_{vp} \\ \tau_{pu} & \tau_{pv} & \tau_{pp} \end{bmatrix} \begin{Bmatrix} R_u(u,v,p) \\ R_v(u,v,p) \\ R_p(u,v) \end{Bmatrix} \mathrm{d}\Omega \tag{18}$$

The field equation residuals are according (1b) and (2b) with the simplification of assumed constant $\mu$ in (5b)

$$\begin{aligned} R_u(u,v,p) &\equiv L_u(u,v,p) - \rho b_x \equiv -\mu \left( \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) + \frac{\partial p}{\partial x} - \rho b_x = 0 \\ R_u(u,v,p) &\equiv L_u(u,v,p) - \rho b_y \equiv -\mu \left( \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} \right) + \frac{\partial p}{\partial y} - \rho b_y = 0 \\ R_p(u,v) &\equiv L_p(u,v) \equiv -\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} = 0 \end{aligned} \tag{19}$$

In the formulas for the momentum equations, the continuity equation has in fact been made use of. In what follows we assume that it is enough to use a diagonal sensitizing matrix. Further, we neglect the second order derivative terms in the residuals and weightings (cf. Remark 6.12). Expression (18) becomes thus

$$F^{(0)} = \int_\Omega \left[ \frac{\partial w_p}{\partial x} \tau_{uu} \left( \frac{\partial p}{\partial x} - \rho b_x \right) + \frac{\partial w_p}{\partial y} \tau_{vv} \left( \frac{\partial p}{\partial y} - \rho b_y \right) + \left( \frac{\partial w_u}{\partial x} + \frac{\partial w_v}{\partial y} \right) \tau_{pp} \left( \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right) \right] \mathrm{d}\Omega \tag{20}$$

The contributions to the three system equations corresponding to node $i$ are thus

$$\begin{aligned} \left(F_u\right)_i^{(0)} &= \int_\Omega \frac{\partial N_i}{\partial x} \tau_{pp} \left( \frac{\partial \bar{u}}{\partial x} + \frac{\partial \bar{v}}{\partial y} \right) \mathrm{d}\Omega \\ \left(F_v\right)_i^{(0)} &= \int_\Omega \frac{\partial N_i}{\partial y} \tau_{pp} \left( \frac{\partial \bar{u}}{\partial x} + \frac{\partial \bar{v}}{\partial y} \right) \mathrm{d}\Omega \\ \left(F_p\right)_i^{(0)} &= \int_\Omega \left[ \frac{\partial N_i}{\partial x} \tau_{uu} \left( \frac{\partial \tilde{p}}{\partial x} - \rho b_x \right) + \frac{\partial N_i}{\partial y} \tau_{vv} \left( \frac{\partial \tilde{p}}{\partial y} - \rho b_y \right) \right] \mathrm{d}\Omega \end{aligned} \tag{21}$$

**Remark 12.6.** It is of some interest to consider the additional terms emerging into the Euler-Lagrange equations from (20) using the sensitized weak form $F + F^{(0)} = 0$. Assuming constant sensitizing parameter values $\tau_{uu}$ and $\tau_{vv}$ we obtain by integration by parts respectively the terms

$$\begin{aligned} & -\frac{\partial}{\partial x} \tau_{pp} \left( \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right) \\ & -\frac{\partial}{\partial y} \tau_{pp} \left( \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right) \\ & -\tau_{uu} \left( \frac{\partial^2 p}{\partial x^2} - \rho \frac{\partial b_x}{\partial x} \right) - \tau_{vv} \left( \frac{\partial^2 p}{\partial y^2} - \rho \frac{\partial b_y}{\partial y} \right) \end{aligned} \tag{22}$$

on the left-hand sides of (19). Parameter $\tau_{pp}$ has now actually the role of some kind of additional dilatational viscosity; see the similarity with expression (7) when introduced into the momentum equations. Further, the last expression (22) shows that the parameters $\tau_{uu}$ and $\tau_{vv}$ introduce desired missing pressure terms into the continuity equation. In fact, in fluid mechanics a *Poisson equation for pressure* is often introduced by differentiating the first momentum equation (19) with respect to $x$ and the second momentum equation with respect to $y$ and summing the results. The velocity components are found to vanish in the manipulation. Especially, if $\tau_{uu} = \tau_{vv}$, the last term (22) represents exactly (with a minus sign) the left-hand side obtained by the manipulation. □



**Reference solutions.** We try to find some reference solutions by the Taylor series approach introduced in Section 5.2.2. The governing field equations are (assuming for reference solution purposes constant viscosity, (see (19))

$$\begin{aligned} & -\mu u_{xx} - \mu u_{yy} + p_x - \rho b^x = 0 \\ & -\mu v_{xx} - \mu v_{yy} + p_y - \rho b^y = 0 \\ & u_x + v_y = 0 \end{aligned} \tag{23}$$

(The body force components are indicated here using superscripts to avoid confusion with respect to differentiation.) We develop the unknowns and the body force components into Taylor series about a generic point ($x = 0$, $y = 0$):

$$\begin{aligned} u(x,y) &= u_0 + \left(u_x\right)_0 x + \left(u_y\right)_0 y + \frac{1}{2}\left(u_{xx}\right)_0 x^2 + \left(u_{xy}\right)_0 xy + \frac{1}{2}\left(u_{yy}\right)_0 y^2 \\ &\quad + \frac{1}{6}\left(u_{xxx}\right)_0 x^3 + \frac{1}{2}\left(u_{xxy}\right)_0 x^2 y + \frac{1}{2}\left(u_{xyy}\right)_0 xy^2 + \frac{1}{6}\left(u_{yyy}\right)_0 y^3 + \cdots \\ v(x,y) &= v_0 + \left(u_x\right)_0 x + \left(v_y\right)_0 y + \cdots \\ p(x,y) &= p_0 + \left(p_x\right)_0 x + \left(p_y\right)_0 y + \cdots \\ b^x(x,y) &= b^x{}_0 + \left(b^x{}_x\right)_0 x + \left(b^x{}_y\right)_0 y + \cdots \\ b^y(x,y) &= b^y{}_0 + \left(b^y{}_x\right)_0 x + \left(b^y{}_y\right)_0 y + \cdots \end{aligned} \tag{24}$$

Evaluating (23) and its differentiated forms at the origin gives

$$\begin{aligned} & -\mu\left(u_{xx}\right)_0 - \mu\left(u_{yy}\right)_0 + \left(p_x\right)_0 - \rho b^x{}_0 = 0 \\ & -\mu\left(v_{xx}\right)_0 - \mu\left(v_{yy}\right)_0 + \left(p_y\right)_0 - \rho b^y{}_0 = 0 \\ & \left(u_x\right)_0 + \left(v_y\right)_0 = 0 \end{aligned} \tag{25}$$

$$\begin{aligned} & -\mu\left(u_{xxx}\right)_0 - \mu\left(u_{xyy}\right)_0 + \left(p_{xx}\right)_0 - \rho\left(b^x{}_x\right)_0 = 0 \\ & -\mu\left(u_{xxy}\right)_0 - \mu\left(u_{yyy}\right)_0 + \left(p_{xy}\right)_0 - \rho\left(b^x{}_y\right)_0 = 0 \\ & \cdots \end{aligned} \tag{26}$$

$$\begin{aligned} & -\mu\left(v_{xxx}\right)_0 - \mu\left(v_{xyy}\right)_0 + \left(p_{xy}\right)_0 - \rho\left(b^y{}_x\right)_0 = 0 \\ & -\mu\left(v_{xxy}\right)_0 - \mu\left(v_{yyy}\right)_0 + \left(p_{yy}\right)_0 - \rho\left(b^y{}_y\right)_0 = 0 \\ & \cdots \end{aligned} \tag{27}$$

$$
\begin{aligned}
&(u_{xx})_0 + (v_{xy})_0 = 0\\
&(u_{xy})_0 + (v_{yy})_0 = 0\\
&(u_{xxx})_0 + (v_{xxy})_0 = 0\\
&(u_{xxy})_0 + (v_{xyy})_0 = 0\\
&(u_{xyy})_0 + (v_{yyy})_0 = 0\\
&\ldots
\end{aligned}
\tag{28}
$$

Ending as shown, we have obtained 12 equations (25) to (28) containing 21 unknowns:

$$
\begin{aligned}
&(u_x)_0,\ (u_{xx})_0,\ (u_{xy})_0,\ (u_{yy})_0,\ (u_{xxx})_0,\ (u_{xxy})_0,\ (u_{xyy})_0,\ (u_{yyy})_0\\
&(v_y)_0,\ (v_{xx})_0,\ (v_{xy})_0,\ (v_{yy})_0,\ (v_{xxx})_0,\ (v_{xxy})_0,\ (v_{xyy})_0,\ (v_{yyy})_0\\
&(p_x)_0,\ (p_y)_0,\ (p_{xx})_0,\ (p_{xy})_0,\ (p_{yy})_0
\end{aligned}
\tag{29}
$$

We can hope at best to solve 12 of the unknowns (if the corresponding determinant is non-zero) in terms of the 9 remaining. Experimentation with Mathematica showed that at least the selection

$$
\begin{aligned}
&(p_x)_0,\ (p_y)_0,\ (p_{xx})_0,\ (p_{xy})_0,\ (p_{yy})_0\\
&(v_y)_0,\ (v_{xy})_0,\ (v_{yy})_0,\ (v_{xxx})_0,\ (v_{xxy})_0,\ (v_{xyy})_0,\ (v_{yyy})_0
\end{aligned}
\tag{30}
$$

worked. The solution was

$$
\begin{aligned}
(p_x)_0 &= \mu(u_{xx})_0 + \mu(u_{yy})_0 + \rho b^x{}_0\\
(p_y)_0 &= \mu(u_{xy})_0 + \mu(v_{xx})_0 + \rho b^y{}_0\\
(p_{xx})_0 &= \mu(u_{xxx})_0 + \mu(u_{xyy})_0 + \rho(b^x{}_x)_0\\
(p_{xy})_0 &= \mu(u_{xxy})_0 + \mu(u_{yyy})_0 + \rho(b^x{}_y)_0\\
(p_{yy})_0 &= -\mu(u_{xxx})_0 - \mu(u_{xyy})_0 + \rho(b^y{}_y)_0\\
(v_y)_0 &= -(u_x)_0
\end{aligned}
$$



$$
\begin{aligned}
(v_{xy})_0 &= -(u_{xx})_0\\
(v_{yy})_0 &= -(u_{xy})_0\\
(v_{xxx})_0 &= -(u_{yyy})_0 + \frac{1}{\mu}(b^x{}_y)_0 - \frac{1}{\mu}(b^y{}_x)_0\\
(v_{xxy})_0 &= -(u_{xxx})_0\\
(v_{xyy})_0 &= -(u_{xxy})_0\\
(v_{yyy})_0 &= -(u_{xyy})_0
\end{aligned}
\tag{31}
$$

Substitution of these results into the Taylor series expressions gives finally the reference solution

$$
\begin{Bmatrix} u\\ v\\ p\\ \rho b^x\\ \rho b^y \end{Bmatrix} = u_0 \begin{Bmatrix} 1\\0\\0\\0\\0 \end{Bmatrix} + v_0 \begin{Bmatrix} 0\\1\\0\\0\\0 \end{Bmatrix} + p_0 \begin{Bmatrix} 0\\0\\1\\0\\0 \end{Bmatrix} + (u_x)_0 \begin{Bmatrix} x\\-y\\0\\0\\0 \end{Bmatrix} + (u_y)_0 \begin{Bmatrix} y\\0\\0\\0\\0 \end{Bmatrix} + (v_x)_0 \begin{Bmatrix} 0\\x\\0\\0\\0 \end{Bmatrix}
$$

$$
+(u_{xx})_0 \begin{Bmatrix} 1/2\cdot x^2\\ -xy\\ \mu x\\ 0\\ 0 \end{Bmatrix} + (u_{xy})_0 \begin{Bmatrix} xy\\ -1/2\cdot y^2\\ -\mu y\\ 0\\ 0 \end{Bmatrix} + (u_{yy})_0 \begin{Bmatrix} 1/2\cdot y^2\\ 0\\ \mu x\\ 0\\ 0 \end{Bmatrix} + (v_{xx})_0 \begin{Bmatrix} 0\\ 1/2\cdot x^2\\ \mu y\\ 0\\ 0 \end{Bmatrix}
$$

$$
+(u_{xxx})_0 \begin{Bmatrix} 1/6\cdot x^3\\ -1/2\cdot x^2 y\\ 1/2\cdot \mu x^2 - 1/2\cdot \mu y^2\\ 0\\ 0 \end{Bmatrix} + (u_{xxy})_0 \begin{Bmatrix} 1/2\cdot x^2 y\\ 1/2\cdot xy^2\\ \mu xy\\ 0\\ 0 \end{Bmatrix}
$$

$$+\left(u_{xyy}\right)_0\begin{Bmatrix}1/2\cdot xy^2\\-1/6\cdot y^3\\1/2\cdot\mu x^2-1/2\cdot\mu y^2\\0\\0\end{Bmatrix}+\left(u_{yyy}\right)_0\begin{Bmatrix}1/6\cdot y^3\\1/6\cdot x^3\\\mu xy\\0\\0\end{Bmatrix}+\rho b^x{}_0\begin{Bmatrix}0\\0\\x\\-1\\0\end{Bmatrix}+\cdots \qquad (32)$$

These are applied in Example 12.1 in an effort to try to determine the sensitizing parameter values.

**Example 12.1.** We develop some formulas in more detail. We consider again the square bilinear element (Figure (a)) and the corresponding four-element patch (Figure (b)).

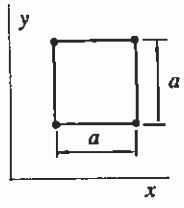



**Figure (a)**

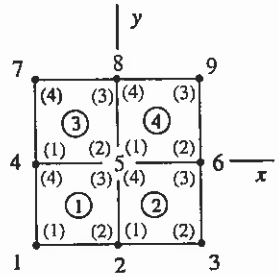



**Figure (b)**

The element approximations are

$$\begin{aligned}u&=\sum_j N_j u_j = N_1u_1+N_2u_2+N_3u_3+N_4u_4\\ v&=\sum_j N_j v_j = N_1v_1+N_2v_2+N_3v_3+N_4v_4\\ p&=\sum_j N_j p_j = N_1p_1+N_2p_2+N_3p_3+N_4p_4\end{aligned} \qquad (a)$$

with the shape functions

$$\begin{aligned}N_1&=1-\xi-\eta+\xi\eta\\ N_2&=\xi-\xi\eta\\ N_3&=\xi\eta\\ N_4&=\eta-\xi\eta\end{aligned} \qquad (b)$$

There are 12 nodal parameters associated with the element. We list the parameters in the order

$$\begin{aligned}\underset{12\times1}{\{a\}}&=[u_1\ v_1\ p_1\ u_2\ v_2\ p_2\ u_3\ v_3\ p_3\ u_4\ v_4\ p_4]^{\mathrm{T}}\\ &\equiv[a_1\ a_2\ a_3\ a_4\ a_5\ a_6\ a_7\ a_8\ a_9\ a_{10}\ a_{11}\ a_{12}]^{\mathrm{T}}\end{aligned} \qquad (c)$$



We evaluate first the element contributions from the standard Galerkin part. Making use of formulas (16) at the element level gives assuming constant viscosity

$$\begin{aligned}F_1&=(F_u)_1=\mu\int_{\Omega^e}\left[\frac{\partial N_1}{\partial x}2\frac{\partial\tilde{u}}{\partial x}+\frac{\partial N_1}{\partial y}\left(\frac{\partial\tilde{v}}{\partial x}+\frac{\partial\tilde{u}}{\partial y}\right)\right]\mathrm{d}\Omega-\int_{\Omega^e}\frac{\partial N_1}{\partial x}\tilde{p}\,\mathrm{d}\Omega\\ &\quad-\int_{\Omega^e}N_1\rho b_x\mathrm{d}\Omega-\int_{\Gamma_\sigma^e}N_1\bar{t}_x\mathrm{d}\Gamma\\ F_2&=(F_v)_1=\mu\int_{\Omega^e}\left[\frac{\partial N_1}{\partial x}\left(\frac{\partial\tilde{u}}{\partial y}+\frac{\partial\tilde{v}}{\partial x}\right)+\frac{\partial N_1}{\partial y}2\frac{\partial\tilde{v}}{\partial y}\right]\mathrm{d}\Omega-\int_{\Omega^e}\frac{\partial N_1}{\partial y}\tilde{p}\,\mathrm{d}\Omega\\ &\quad-\int_{\Omega^e}N_1\rho b_y\mathrm{d}\Omega-\int_{\Gamma_\sigma^e}N_1\bar{t}_y\mathrm{d}\Gamma\\ F_3&=(F_p)_1=-\int_{\Omega^e}N_1\left(\frac{\partial\tilde{u}}{\partial x}+\frac{\partial\tilde{v}}{\partial y}\right)\mathrm{d}\Omega\\ F_4&=(F_u)_2=\cdots\end{aligned} \qquad (d)$$

Further,

$$\begin{aligned}F_1&=\mu\sum_j\left[\int_{\Omega^e}\left(2\frac{\partial N_1}{\partial x}\frac{\partial N_j}{\partial x}+\frac{\partial N_1}{\partial y}\frac{\partial N_j}{\partial y}\right)\mathrm{d}\Omega\right]u_j+\mu\sum_j\left(\int_{\Omega^e}\frac{\partial N_1}{\partial y}\frac{\partial N_j}{\partial x}\mathrm{d}\Omega\right)v_j\\ &\quad-\sum_j\left(\int_{\Omega^e}\frac{\partial N_1}{\partial x}N_j\,\mathrm{d}\Omega\right)p_j-\int_{\Omega^e}N_1\rho b_x\mathrm{d}\Omega-\int_{\Gamma_\sigma^e}N_1\bar{t}_x\mathrm{d}\Gamma\\ F_2&=\mu\sum_j\left(\int_{\Omega^e}\frac{\partial N_1}{\partial x}\frac{\partial N_j}{\partial y}\mathrm{d}\Omega\right)u_j+\mu\sum_j\left[\int_{\Omega^e}\left(\frac{\partial N_1}{\partial x}\frac{\partial N_j}{\partial x}+2\frac{\partial N_1}{\partial y}\frac{\partial N_j}{\partial y}\right)\mathrm{d}\Omega\right]v_j\\ &\quad-\sum_j\left(\int_{\Omega^e}\frac{\partial N_1}{\partial y}N_j\,\mathrm{d}\Omega\right)p_j-\int_{\Omega^e}N_1\rho b_y\mathrm{d}\Omega-\int_{\Gamma_\sigma^e}N_1\bar{t}_y\mathrm{d}\Gamma\\ F_3&=-\sum_j\left(\int_{\Omega^e}N_1\frac{\partial N_j}{\partial x}\mathrm{d}\Omega\right)u_j-\sum_j\left(\int_{\Omega^e}N_1\frac{\partial N_j}{\partial y}\mathrm{d}\Omega\right)v_j\\ F_4&=\cdots\end{aligned} \qquad (e)$$

Finally,

$$\begin{aligned}F_1&=\mu\frac{1}{6}(6u_1-3u_2-u_3+0\cdot u_4)+\mu\frac{1}{4}(v_1-v_2-v_3+v_4)\\ &\quad+\frac{a}{12}(2p_1+2p_2+p_3+p_4)-\int_{\Omega^e}N_1\rho b_x\mathrm{d}\Omega-\int_{\Gamma_\sigma^e}N_1\bar{t}_x\mathrm{d}\Gamma\\ F_2&=\mu\frac{1}{4}(6u_1+0\cdot u_2-3u_3-u_4)+\mu\frac{1}{6}(v_1-v_2-v_3+v_4)\\ &\quad+\frac{a}{12}(2p_1+p_2+p_3+2p_4)-\int_{\Omega^e}N_1\rho b_y\mathrm{d}\Omega-\int_{\Gamma_\sigma^e}N_1\bar{t}_y\mathrm{d}\Gamma\\ F_3&=-\frac{a}{12}(-2u_1+2u_2+u_3-u_4)-\frac{a}{12}(-2v_1-v_2+v_3+2v_4)\\ F_4&=\cdots\end{aligned} \qquad (f)$$

The sensitizing terms from (21) applied at the element level give assuming constant parameter values

$$F_1^{(0)} = \left(F_u\right)_1^{(0)} = \tau_{pp}\int_{\Omega^e}\frac{\partial N_1}{\partial x}\left(\frac{\partial \bar{u}}{\partial x}+\frac{\partial \bar{v}}{\partial y}\right)\mathrm{d}\Omega$$

$$F_2^{(0)} = \left(F_v\right)_1^{(0)} = \tau_{pp}\int_{\Omega^e}\frac{\partial N_1}{\partial y}\left(\frac{\partial \bar{u}}{\partial x}+\frac{\partial \bar{v}}{\partial y}\right)\mathrm{d}\Omega$$

$$F_3^{(0)} = \left(F_p\right)_1^{(0)} = \tau_{uu}\int_{\Omega^e}\frac{\partial N_1}{\partial x}\frac{\partial \tilde{p}}{\partial x}\mathrm{d}\Omega + \tau_{vv}\int_{\Omega^e}\frac{\partial N_i}{\partial y}\frac{\partial \tilde{p}}{\partial y}\mathrm{d}\Omega$$
$$-\tau_{uu}\int_{\Omega^e}\frac{\partial N_1}{\partial y}\rho b_x\mathrm{d}\Omega - \tau_{vv}\int_{\Omega^e}\frac{\partial N_1}{\partial y}\rho b_y\mathrm{d}\Omega \qquad \text{(g)}$$

$$F_4^{(0)} = \left(F_u\right)_2^{(0)} = \cdots$$

Further,

$$F_1^{(0)} = \tau_{pp}\sum_j\left(\int_{\Omega^e}\frac{\partial N_1}{\partial x}\frac{\partial N_j}{\partial x}\mathrm{d}\Omega\right)u_j + \tau_{pp}\sum_j\left(\int_{\Omega^e}\frac{\partial N_1}{\partial x}\frac{\partial N_j}{\partial y}\mathrm{d}\Omega\right)v_j$$

$$F_2^{(0)} = \tau_{pp}\sum_j\left(\int_{\Omega^e}\frac{\partial N_1}{\partial y}\frac{\partial N_j}{\partial x}\mathrm{d}\Omega\right)u_j + \tau_{pp}\sum_j\left(\int_{\Omega^e}\frac{\partial N_1}{\partial y}\frac{\partial N_j}{\partial y}\mathrm{d}\Omega\right)v_j$$

$$F_3^{(0)} = \tau_{uu}\sum_j\left(\int_{\Omega^e}\frac{\partial N_1}{\partial x}\frac{\partial N_j}{\partial x}\mathrm{d}\Omega\right)p_j + \tau_{vv}\sum_j\left(\int_{\Omega^e}\frac{\partial N_1}{\partial y}\frac{\partial N_j}{\partial y}\mathrm{d}\Omega\right)p_j \qquad \text{(h)}$$
$$-\tau_{uu}\int_{\Omega^e}\frac{\partial N_1}{\partial x}\rho b_x\mathrm{d}\Omega - \tau_{vv}\int_{\Omega^e}\frac{\partial N_1}{\partial y}\rho b_y\mathrm{d}\Omega$$

$$F_4^{(0)} = \cdots$$

Again, still further development gives

$$F_1^{(0)} = \tau_{pp}\frac{1}{6}\left(2u_1+u_2-u_3-2u_4\right)+\tau_{pp}\frac{1}{4}\left(v_1+v_2-v_3-v_4\right)$$

$$F_2^{(0)} = \tau_{pp}\frac{1}{4}\left(u_1-u_2-u_3+u_4\right)+\tau_{pp}\frac{1}{6}\left(2v_1+v_2-v_3-2v_4\right)$$

$$F_3^{(0)} = \tau_{uu}\frac{1}{6}\left(2p_1+p_2-p_3-2p_4\right)+\tau_{vv}\frac{1}{6}\left(2p_1+p_2-p_3-2p_4\right) \qquad \text{(i)}$$
$$-\tau_{uu}\int_{\Omega^e}\frac{\partial N_1}{\partial y}\rho b_x\mathrm{d}\Omega - \tau_{vv}\int_{\Omega^e}\frac{\partial N_1}{\partial y}\rho b_y\mathrm{d}\Omega$$

$$F_4^{(0)} = \cdots$$

The data for system equations assembly is obtained from Figure (b) and is given in the following table:



| | (1) | | | (2) | | | (3) | | | (4) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1)* | (2)* | (3)* | (4)* | (5)* | (6)* | (7)* | (8)* | (9)* | (10)* | (11)* | (12)* |
| ① | 1 | 2 | 3 | 4 | 5 | 6 | 13 | 14 | 15 | 10 | 11 | 12 |
| ② | 4 | 5 | 6 | 7 | 8 | 9 | 16 | 17 | 18 | 13 | 14 | 15 |
| ③ | 10 | 11 | 12 | 13 | 14 | 15 | 22 | 23 | 24 | 19 | 20 | 21 |
| ④ | 13 | 14 | 15 | 16 | 17 | 18 | 25 | 26 | 27 | 22 | 23 | 24 |

The nodal parameter numbering has followed the global nodal numbering in the order first $u$, then $v$, then $p$.

The three system equations corresponding to the central node 5 are

$$\sum_{j=1}^{27}K_{13,j}a_j - b_{13} = 0$$

$$\sum_{j=1}^{27}K_{14,j}a_j - b_{14} = 0 \qquad \text{(j)}$$

$$\sum_{j=1}^{27}K_{15,j}a_j - b_{15} = 0$$

The assembly happens as explained earlier. For instance,

$$K_{13,1} = K_{7,1}^{1} \qquad \text{(k)}$$

$$K_{13,10} = K_{7,10}^{1} + K_{4,1}^{3}$$

The detailed calculations are performed by Mathematica. There is obtained

$$\left(-\frac{\mu}{2}-\frac{\tau_{pp}}{6}\right)u[1]+\left(-\frac{\mu}{4}-\frac{\tau_{pp}}{4}\right)v[1]-\frac{1}{12}a_p[1]+\frac{1}{3}\tau_{pp}u[2]+$$
$$\left(-\frac{\mu}{2}-\frac{\tau_{pp}}{6}\right)u[3]+\left(\frac{\mu}{4}+\frac{\tau_{pp}}{4}\right)v[3]+\frac{1}{12}a_p[3]+\left(-\mu-\frac{2\tau_{pp}}{3}\right)u[4]-$$
$$\frac{1}{3}a_p[4]+\left(4\mu+\frac{4\tau_{pp}}{3}\right)u[5]+\left(-\mu-\frac{2\tau_{pp}}{3}\right)u[7]+\frac{1}{3}a_p[6]+$$
$$\left(-\frac{\mu}{2}-\frac{\tau_{pp}}{6}\right)u[7]+\left(\frac{\mu}{4}+\frac{\tau_{pp}}{4}\right)v[7]-\frac{1}{12}a_p[7]+\frac{1}{3}\tau_{pp}u[8]+$$
$$\left(-\frac{\mu}{2}-\frac{\tau_{pp}}{6}\right)u[9]+\left(-\frac{\mu}{4}-\frac{\tau_{pp}}{4}\right)v[9]+\frac{1}{12}a_p[9]-b[13]=0 \qquad \text{(l)}$$

$$\left(-\frac{\mu}{4}-\frac{\tau_{pp}}{4}\right)u[1]+\left(-\frac{\mu}{2}-\frac{\tau_{pp}}{6}\right)v[1]-\frac{1}{12}a_p[1]+\left(-\mu-\frac{2\tau_{pp}}{3}\right)v[2]-$$
$$\frac{1}{3}a_p[2]+\left(\frac{\mu}{4}+\frac{\tau_{pp}}{4}\right)u[3]+\left(-\frac{\mu}{2}-\frac{\tau_{pp}}{6}\right)v[3]-\frac{1}{12}a_p[3]+\frac{1}{3}\tau_{pp}v[4]+$$
$$\left(4\mu+\frac{4\tau_{pp}}{3}\right)v[5]+\frac{1}{3}\tau_{pp}v[6]+\left(\frac{\mu}{4}+\frac{\tau_{pp}}{4}\right)u[7]+\left(-\frac{\mu}{2}-\frac{\tau_{pp}}{6}\right)v[7]+$$
$$\frac{1}{12}a_p[7]+\left(-\mu-\frac{2\tau_{pp}}{3}\right)v[8]+\frac{1}{3}a_p[8]+\left(-\frac{\mu}{4}-\frac{\tau_{pp}}{4}\right)u[9]+$$
$$\left(-\frac{\mu}{2}-\frac{\tau_{pp}}{6}\right)v[9]+\frac{1}{12}a_p[9]-b[14]=0 \qquad \text{(m)}$$

$$\frac{1}{12}\,au[1] + \frac{1}{12}\,av[1] + \left(-\frac{\tau uu}{6} - \frac{\tau vv}{6}\right) p[1] + \frac{1}{3}\,av[2] + \left(\frac{\tau uu}{3} - \frac{2\,\tau vv}{3}\right) p[2] -$$
$$\frac{1}{12}\,au[3] + \frac{1}{12}\,av[3] + \left(-\frac{\tau uu}{6} - \frac{\tau vv}{6}\right) p[3] + \frac{1}{3}\,au[4] +$$
$$\left(-\frac{2\,\tau uu}{3} + \frac{\tau vv}{3}\right) p[4] + \left(\frac{4\,\tau uu}{3} + \frac{4\,\tau vv}{3}\right) p[5] - \frac{1}{3}\,au[6] +$$
$$\left(-\frac{2\,\tau uu}{3} + \frac{\tau vv}{3}\right) p[6] + \frac{1}{12}\,au[7] - \frac{1}{12}\,av[7] + \left(-\frac{\tau uu}{6} - \frac{\tau vv}{6}\right) p[7] -$$
$$\frac{1}{3}\,av[8] + \left(\frac{\tau uu}{3} - \frac{2\,\tau vv}{3}\right) p[8] - \frac{1}{12}\,au[9] - \frac{1}{12}\,av[9] + \left(-\frac{\tau uu}{6} - \frac{\tau vv}{6}\right) p[9] -$$
$$b[15] = 0 \qquad \text{(n)}$$

The constant terms $b$ depend on the detailed source terms used in the reference solutions.

It is rather obvious that in equation (n) corresponding to the nodal pressure parameter $p_5$, nonzero sensitizing parameter values $\tau_{uu}$ and $\tau_{vv}$ are needed to introduce the pressure nodal parameters into the formulation.

It is of some interest to notice that the terms of the form $x^m y^n$ $(m>0,\ n>0)$ appearing in the reference solutions for the velocity components and the pressure disappear on the $x$- and $y$-axes. Thus, it is only the nodal values at the nodes 1, 3, 7 and 9 that can participate in the possible determination of the sensitizing parameter values in the patch test. The conclusion obtained with the reference solutions given above proved to be a disappointment as the only result obtained was the obvious one $\tau_{uu}=\tau_{vv}$. Usually the patch test was passed irrespective of the parameter values. Thus further study is needed.

**Application.** In an earlier effort some parameter values were obtained in Freund and Salonen (1998) and we here just give some end results. For example, the selection

$$\begin{bmatrix} \tau_{uu} & \tau_{uv} & \tau_{up} \\ \tau_{vu} & \tau_{vv} & \tau_{vp} \\ \tau_{pu} & \tau_{pv} & \tau_{pp} \end{bmatrix} = \frac{1}{12}\begin{bmatrix} h^2/\mu & 0 & 0 \\ 0 & h^2/\mu & 0 \\ 0 & 0 & \mu \end{bmatrix} \qquad (33)$$

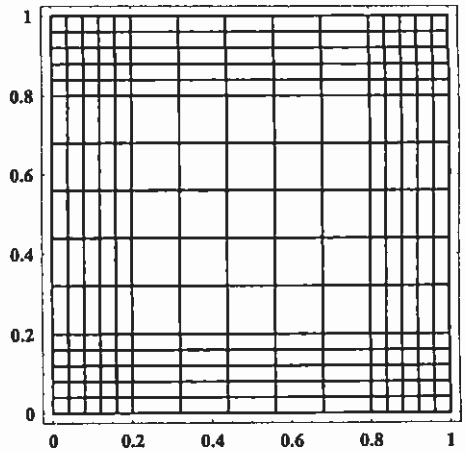
where $h$ is a measure of the element size, was found appropriate for bilinear elements.

**Remark 12.7.** Contrary to what has been the case with all sensitizing parameters encountered this far, the parameter $\tau_{pp}$ in (33) does not vanish with a vanishing mesh size. However, as there are only first derivatives present in the corresponding residual $R_p$ in (19), no crimes concerning the continuity requirements for convergence are performed here. □

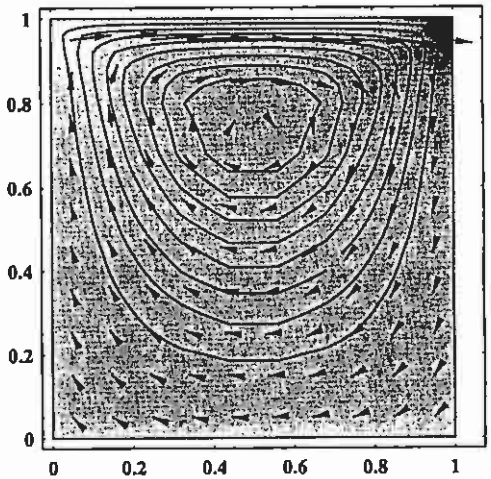
Figure 12.1 shows some results for the so-called moving lid problem presented in dimensionless form in a unit square domain with a non-regular crude 15×15 mesh of rectangular bilinear elements. The boundary conditions consist of velocity boundary conditions with zero data except on the edge $y=1$ where $u=1$. With prescribed velocity on the whole boundary, the pressure is



determined only up to an additive constant. The pressure level is fixed here by setting $p=0$ at point (0.0,0.0). The light (dark) color means high (low) pressure. (In the black and white version here the pressure distribution cannot be well discerned.)




(a)




(b)

**Figure 12.1** (a) Mesh. (b) Some velocity vectors, streamlines and pressure distribution.

## 12.4 NAVIER-STOKES PROBLEM

When the inertia forces are no more negligible, the field equations (1a) have in the steady case the additional term

$$\rho v_\beta \frac{\partial v_\alpha}{\partial x_\beta} \tag{1}$$

on the left-hand side. The integrand in the standard weak form (13) is modified with the additional term

$$w_\alpha \rho v_\beta \frac{\partial v_\alpha}{\partial x_\beta} \tag{2}$$

Because of this non-linear term, the problem must be solved iteratively. In practice we use the deltaform: $v_\alpha = \bar{v}_\alpha + \Delta v_\alpha$, in the way described in Chapter 11. Here $\bar{v}_\alpha$ is the current updated solution satisfying the velocity boundary conditions.

The sensitizing matrix in the sensitizing term (17) is found to have the form

$$\begin{bmatrix} \tau_{uu} & \tau_{uv} & \tau_{up} \\ \tau_{vu} & \tau_{vv} & \tau_{vp} \\ \tau_{pu} & \tau_{pv} & \tau_{pp} \end{bmatrix} = \begin{bmatrix} \alpha & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & \beta \end{bmatrix} \tag{3}$$

with

$$\alpha = \frac{1}{2\rho|\mathbf{v}|/h + 12\mu/h^2}, \qquad \beta = \frac{\mu}{12} \tag{4}$$

where $|\mathbf{v}|$ is the flow speed. Compared to the Stokes case the diagonal elements except the last one, contain an additional term that is essential in the convection dominated case, i.e., when the Reynolds number is large.

Figure 12.2 gives some results for a case similar to that explained in connection with Figure 12.1; however, the Reynolds number is not zero but has the value 2000. The solution is now unsymmetrical and secondary vortices appear at the bottom corners (when using a denser mesh).



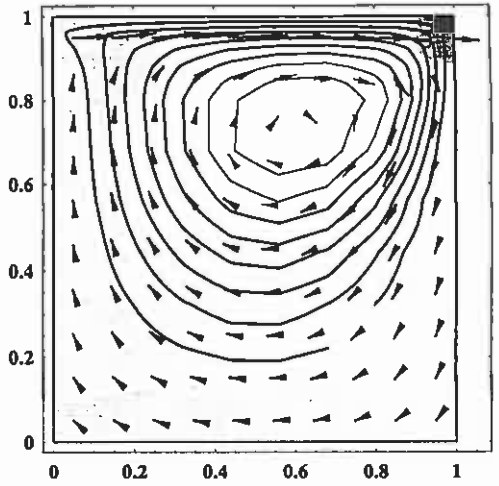



**Figure 12.2** Some velocity vectors, streamlines and pressure distribution.

## REFERENCES

Freund, J. and Salonen, E.-M. (1998). Stability parameters in connection with fluid flow problems, Report no. 47, *Laboratory of Theoretical and Applied Mechanics, Helsinki University of Technology.*

Hughes, T. J. R., Franca, L. P. and Balestra, M. (1986). A new finite element formulation for computational fluid dynamics: V. Circumventing the Babuska-Brezzi condition: A stable Petrov-Galerkin formulation of the Stokes problem accommodating equal order interpolations, *Comput. Methods Appl. Mech. Engrg.*, Vol. 59, pp. 85...99.

Malvern, L. E. (1969). *Introduction to the Mechanics of a Continuous Medium*, Prentice hall, Englewood Cliffs, New Jersey.

Wilcox, D. C, (1994). *Turbulence Modelling for CFD*, DCW Industries, Inc. La Canada, California, ISBN 0-9636051-0-0.

## PROBLEMS

Siegel, R. and Howell, J. R. (1981). *Thermal Radiation Heat Transfer*, 2nd ed., Hemisphere Publishing Corporation, New York, ISBN 0-89116-506-1.
Stelzer, J. F. (1984). *Physical Property Algorithms*. Karl Thiemig.

## PROBLEMS

# 13 SOLUTION OF SYSTEM EQUATIONS

Discretization by the finite element method produces what we call system equations. They are usually algebraic equations, eigenvalue equations or ordinary differential equations and especially their linear forms. We shall deal here with algebraic system equations and briefly with eigenvalue problems. Ordinary differential system equations have been considered in some extent in Section 9.2. These themes are classical and belong to the contents of textbooks on numerical mathematics, e.g., Mäkelä et al. (1982). Efficient solution of system equations is naturally very important in practice as often a large part of the computations is spent in this phase.

## 13.1 ALGEBRAIC EQUATIONS

### 13.1.1 Linear equations

**Introduction.** Let us consider the linear algebraic system of equations

$$\{F(\{a\})\} \equiv [K]\{a\} - \{b\} = \{0\} \tag{1}$$

or in simpler notation the system

$$\underset{n\times n}{[K]}\,\underset{n\times 1}{\{a\}} = \underset{n\times 1}{\{b\}} \tag{2}$$

The task is to determine the column vector $\{a\}$. In the pioneering days of the finite element method the problem was considered large when the number of unknowns $n$ was of the order of one hundred. Nowadays $n$ can easily be in practice of the order of tens of thousands, sometimes even of the order of millions.

There are in principle two different ways to find the solution: *direct methods* and *iterative methods*. In direct methods the solution is obtained in theory exactly — if roundoff is not considered — by executing a finite number of certain calculation steps. The direct methods are in general versions of the *Gauss algorithm*, e.g., Kivelä (1980). In iterative methods — such as the *Gauss-Seidel method* — the solution is improved by consecutive steps until it is accurate enough but fully accurate solution cannot be achieved in a finite number of steps. Both methods have their own merits. Lately the popularity of iterative methods has been growing. For instance adaptive methods and multigrid methods have increased this trend.





To be efficient, the solution methods must take into account the most important features of the system equations generated by the finite element method. These are:

(1) The coefficient matrix is *sparse* (harva), that is, the relative number of non-zero entries is small.

(2) The coefficient matrix is — when a suitable numbering order of the nodal parameters is used — often *banded* (nauhamainen), that is, the non-zero entries situate near the main diagonal.

(3) If the discretization can be based on a quadratic functional, the coefficient matrix is *symmetric* and often also *positive-definite*.

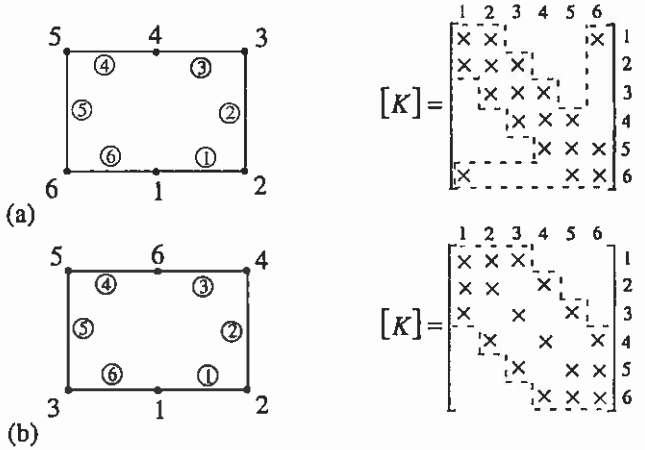



**Figure 13.1** Two identical meshes of one-dimensional two-noded elements numbered in different ways.

Figures 13.1 (a) and (b) show the pattern of two system matrices resulting from two different numbering order of two identical meshes. The meshes could represent for example a pipe network loop. The case concerns a situation with one nodal parameter per node and the nodal parameters are numbered according to the nodal numbering. Those entries denoted by × in the matrices can be non-zero. The dotted line gives the *profile* or *skyline* (profiili, ääriviiva) of the matrix. It is the line, which reaches vertically in each column up to the last non-zero entry and similarly horizontally to the left.

It is of some interest to note that even in the case of a non-symmetric coefficient matrix where in general $K_{ji} \neq K_{ij}$ the profile is still symmetric meaning that the non-zero entries situate symmetrically when the system is generated by the

finite element method. To see this we have to recall the assembly process explained in Section 2.3.2. Let the nodal parameters be numbered in a certain way and we have ended up with the quantities $a_1, a_2, \cdots, a_n$. Each parameter has been associated with a certain node (be it actual or bookkeeping one) which is further connected to one or several elements of the mesh. The system equations are generated normally so that the first equation is associated with the first nodal parameter, the second equation with the second nodal parameter and so on. In a variational formulation the equation associated with the parameter $a_i$ is clearly $\partial \tilde{\Pi} / \partial a_i = 0$. In the residual formulation this concept is actually not self-evident. We have in principle first to agree on the weighting functions associated with the nodal parameter in question, after which the system equation is generated from the relevant weak form. If the weighting functions are selected so that they are non-zero at most in the elements connected to the nodal parameter (as is the case especially in the Galerkin method) the corresponding equation is seen to contain coupling only from the nodal parameters connected just to those elements. The assembly rules given in Section 2.3.2 in fact have been based on this assumption. Considering the assembly rules, we realize that if the element gives an entry $K_{ij}$ when assembling the system equation associated with parameter $a_i$, (the element thus has the global nodal parameter values $i$ and $j$), when generating the system equation associated with parameter $a_j$, it also gives a possible contribution $K_{ji}$.

The example cases of Figure 13.1 show that the structure of the system matrix depends on the numbering order of the nodal parameters. Numbers $m_i$, $i = 1,2,\cdots,n$, determine the profile of the matrix and the numbers $(i - m_i)$, $i = 1,2,\cdots,n$, the so-called *column heights* (sarakekorkeus), Bathe and Wilson (1987). For example, in the case of Figure 13.1 (a), $m_1 = 1$, $m_2 = 1$, $m_3 = 2$, $m_4 = 3$, $m_5 = 4$, $m_6 = 1$ and the corresponding column heights are 0,1,1,1,1,5. The maximum value of the column height is called the *half-bandwidth*, also *semi-bandwidth* (puolinauhanleveys; usein myös nauhanleveys). In the cases of Figure 13.1 (a) and (b) the values of the half-bandwidth are 5 and 2, respectively. The half-bandwidth is determined by the formula

$$B = \max_e \left( \max |i - j| \right) \tag{3}$$

where $i$ and $j$ refer to the global nodal parameter numbers of element $e$ of the mesh. The maximum of the absolute value is searched for from the whole mesh. To keep $B$ as small as possible, the nodal parameter numbering should be executed so that no large differences in the numbers should appear in an element. Algorithms for keeping $B$ small have been developed, as the solution by a banded solver is the cheaper the smaller $B$.



**Gauss algorithm.** The Gauss algorithm or the Gauss elimination method is based on systematic modification of the equations by linear combinations so that the number of unknowns gets smaller in the new equations until the last equation contains only one unknown. After this the determination of the unknowns takes place through consecutive substitutions.

**Example 13.1.** Let us consider the solution by the Gauss elimination method of the equation system

$$\begin{bmatrix} 4 & 0 & -4 \\ 2 & 4 & 1 \\ 1 & -4 & 6 \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \\ a_3 \end{Bmatrix} = \begin{Bmatrix} 4 \\ 0 \\ 1 \end{Bmatrix} \Big) -\frac{1}{2} \Big) -\frac{1}{4} \tag{a}$$

having three unknowns. The first phase of the Gauss elimination method has been already indicated. The notations mean that the first equation has been added multiplied by the factors $-1/2$ and $-1/4$, respectively, to the second and third equation. In this way we obtain the next system

$$\begin{bmatrix} 4 & 0 & -4 \\ 0 & 4 & 3 \\ 0 & -4 & 7 \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \\ a_3 \end{Bmatrix} = \begin{Bmatrix} 4 \\ -2 \\ 10 \end{Bmatrix} \Big) 1 \tag{b}$$

The unknown $a_1$ has been thus eliminated from the last two equations. The essential point for succeeding in this is that the corresponding element of the matrix, the so-called *pivot* (tukialkio) — here the member $()_{11}$ — is non-zero. If this is not the case, some rearrangement of the order of the equations or the unknowns has to be done before we can proceed. The arrow in (b) shows the next step in the Gauss elimination, which produces the set

$$\begin{bmatrix} 4 & 0 & -4 \\ 0 & 4 & 3 \\ 0 & 0 & 10 \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \\ a_3 \end{Bmatrix} = \begin{Bmatrix} 4 \\ -2 \\ 8 \end{Bmatrix} \tag{c}$$

We have produced a so-called *upper triangular set* (yläkolmioryhmä). The steps used can be called *triangularization* (kolmiointi). The *back substitution* (takaisinsijoitus) or the final determination of the values of the unknowns takes place using set (c) with the following obvious formulas:

$$\begin{aligned} a_3 &= \frac{1}{10}(8) = 0.8 \\ a_2 &= \frac{1}{4}(-2 - 3 \cdot a_3) = -1.1 \\ a_1 &= \frac{1}{4}(4 + 4 \cdot a_3 - 0 \cdot a_2) = 1.8 \end{aligned} \tag{d}$$

Let us further write down the matrices

$$[L]=\begin{bmatrix}1 & 0 & 0\\ 1/2 & 1 & 0\\ 1/4 & -1 & 1\end{bmatrix},\qquad [U]=\begin{bmatrix}4 & 0 & -4\\ 0 & 4 & 3\\ 0 & 0 & 10\end{bmatrix} \tag{e}$$

The former is a lower triangular matrix whose diagonal elements are ones and the rest of the elements are the pivots with negative signs used in the triangularization and put at obvious positions. The latter matrix is the matrix of the upper triangular system (c). For reasons explained for instance in Kivelä (1980), we have

$$[K]=[L][U] \tag{f}$$

It is said that (f) is the LU-*decomposition* (LU-hajotelma) of the original matrix.

By performing the calculations we see that the product

$$[L][U]=\begin{bmatrix}1 & 0 & 0\\ 1/2 & 1 & 0\\ 1/4 & -1 & 1\end{bmatrix}\begin{bmatrix}4 & 0 & -4\\ 2 & 4 & 1\\ 1 & -4 & 6\end{bmatrix} \tag{g}$$

$$\begin{bmatrix}4 & 0 & -4\\ 0 & 4 & 3\\ 0 & 0 & 10\end{bmatrix}$$

so that the coefficient matrix of (a) is indeed obtained.

Using the notations of the above example, the Gauss algorithm is written often in the form

$$\begin{aligned}[L]\{x\}&=\{b\}\\ [U]\{a\}&=\{x\}\end{aligned} \tag{4}$$

If the LU-decomposition is available, it is easy to determine $\{x\}$ from the first set (4) and then $\{a\}$ from the second set (4).

The most popular solution versions based on the Gauss elimination method used in connection of the finite element method are the *band-*, *profile-* and *frontal solvers* (nauha-, profiili- ja rintamaratkaisija).

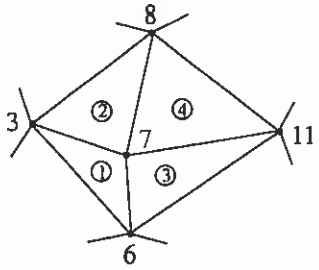
When using the band solver, the bookkeeping is based on the domain determined by the half-bandwidth so that considerable economy is achieved in connection with banded matrices, as the zero elements outside the banded domain need not be processed.

As the name implies, the profile solver differs from the band solver in the respect that the bookkeeping includes only the domain bordered by the profile. The algorithm becomes more complicated but it is much more efficient

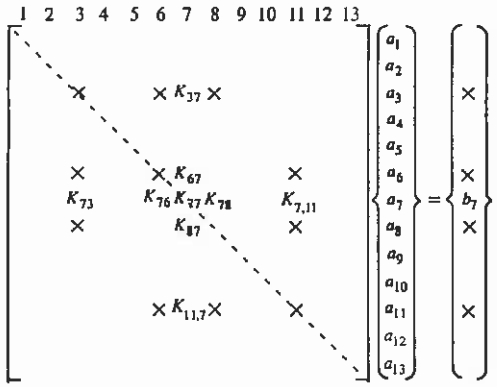


especially if the matrix contains only few columns having high column heights, which thus determine the half-bandwidth (cf. Figure 13.1).

The frontal solver differs from the two versions above in an essential manner. The elimination is started already during the assembly process and for efficiency the numbering order of the elements, not the nodes, is of importance. To see the main idea, let us consider Figure 13.2. Figure (a) shows the elements 1, 2, 3, 4 and the global nodes 3, 6, 7, 8, 11. One nodal parameter is associated with each node. We assume that the contributions from elements 1, 2, 3, 4 have been assembled into the system matrix and into the column matrix. We obtain the situation shown in Figure (b). The entries denoted by the letter symbols are clearly fully assembled. The entries denoted by × have also obtained non-zero contributions but have not yet been fully assembled. The equation associated with the nodal parameter $a_7$ is here the only one fully ready. This means that by using it, $a_7$ can be eliminated from all the equations — the 3rd, 6th, 8th, 11th equation — in which this variable is going to appear. Although these equations are not yet ready assembled, the factors $K_{37}/K_{77}$, $K_{67}/K_{77}$, $K_{87}/K_{77}$, $K_{11,7}/K_{77}$ needed in the elimination, are known. The elimination thus means the addition of certain terms on the rows 3, 6, 8, 11. The addition of the contributions of the next element (element 5) can mean that the equation associated to a new nodal parameter has been fully assembled and again this variable can be eliminated from the system. It should be noted that after the first eliminated variable, the equations are in general no more the original system equations but some linear combinations of them because the assembly and the elimination (= formation of linear combinations) are mixed. The set of variables present at each stage of the procedure is called sometimes the *front* or *wave front* (rintama, aaltorintama), which gives the name for the solution method. In Figure 13.2 (a) the front is after the elimination of $a_7$ the set $a_3, a_6, a_8, a_{11}$. The bookkeeping takes naturally place differing from the presentation of Figure 13.2 (b) in a condensed form, which makes the algorithm rather complicated.




(a)




(b)

**Figure 13.2 (a)** Part of a mesh. **(b)** Contributions from elements 1, 2, 3, 4 into the system equations.

**Remark 13.1.** Above, a very brief description of the versions of the Gauss elimination method used in connection with the finite element method has been given. For instance the references Bathe and Wilson (1976), Hughes (1987), Zienkiewicz and Taylor (2000) give much important information on this theme. It should be noted that the triangulation or thus the formation of the LU-decomposition is the most expensive part of the solution. After this has been obtained, the solution for a possible new right hand side $\{b\}$, is using substitutions (4) rather cheap. This can made use of for instance in the time integration described in Section 9.2. □

**Remark 13.2.** If the coefficient matrix is symmetric, all the versions discussed above can be designed so that the symmetry is taken into account, which means that the algorithms get more efficient. If the coefficient matrix is positive-definite, it can be shown, Bathe and Wilson (1976, p. 37), that the elimination can be performed without pivot search in a fixed order without danger of division by zero. With indefinite matrices the elimination in a fixed order can lead to a failure even when the matrix is non-singular. □

**Jacobi and Gauss-Seidel iteration.** These are the simplest iterative methods. Nowadays they are not often used as such alone but for instance as a part of a multigrid method. In iterative methods a sequence of consecutive solutions $\{a\}^{(0)}$, $\{a\}^{(1)}$, $\{a\}^{(2)}$, $\cdots$ is generated which hopefully tends to exact solution $\{a\}$. To measure the convergence we must have a suitable norm. The most usual one is the *Euclidian vector norm*

$$\left\|\{a\}\right\| = \left(a_1^2 + a_1^2 + \cdots + a_n^2\right)^{1/2} \tag{5}$$

Instead of the mere error it is usually more sensible to monitor the relative dimensionless error using in theory the expression



$$\varepsilon = \frac{\left\|\{a\}^{(i)} - \{a\}\right\|}{\left\|\{a\}\right\|} \tag{6}$$

However, as the exact solution is in general unknown, in practice often the measure

$$\varepsilon = \frac{\left\|\{a\}^{(i)} - \{a\}^{(i-1)}\right\|}{\left\|\{a\}^{(i-1)}\right\|} \tag{7}$$

is used.

**Example 13.2.** We consider the system of equations of Example 13.1:

$$\begin{aligned} 4a_1 + 0\cdot a_2 - 4a_3 &= 4 \\ 2a_1 + 4a_2 + 1a_3 &= 0 \\ 1a_1 - 4a_2 + 6a_3 &= 11 \end{aligned} \tag{a}$$

whose exact solution is

$$\{a\} = [1.8 \quad -1.1 \quad 0.8]^{\mathrm{T}} \tag{b}$$

We assume the reader to know the content of the Jacobi and Gauss-Seidel iteration. We just describe their use in this example.

In the Jacobi iteration the obvious iteration procedure is

$$\begin{aligned} a_1^{(k+1)} &= \frac{1}{4}\left( \quad -0\cdot a_2^{(k)} + 4a_3^{(k)} + 4\right) \\ a_2^{(k+1)} &= \frac{1}{4}\left(-2a_1^{(k)} \quad -1a_3^{(k)} + 0\right) \\ a_3^{(k+1)} &= \frac{1}{6}\left(-1a_1^{(k)} + 4a_2^{(k)} \quad +11\right) \end{aligned} \tag{c}$$

In the Gauss-Seidel iteration

$$\begin{aligned} a_1^{(k+1)} &= \frac{1}{4}\left( \quad -0\cdot a_2^{(k)} + 4a_3^{(k)} + 4\right) \\ a_2^{(k+1)} &= \frac{1}{4}\left(-2a_1^{(k+1)} \quad -1a_3^{(k)} + 0\right) \\ a_3^{(k+1)} &= \frac{1}{6}\left(-1a_1^{(k+1)} + 4a_2^{(k+1)} \quad +11\right) \end{aligned} \tag{d}$$

The difference with the former version is just that new information is used as soon as it is available.

Tables (a) and (b) show some results from calculations when the starting vector $\{a\}^{(0)}$ has been the zero vector. In the evaluation of the relative error expression (6) has been used where

$$\|\{a\}\| = \left(1.8^2 + 1.1^2 + 0.8^2\right)^{1/2} \approx 2.256 \tag{e}$$

The Gauss-Seidel iteration seems to converge; the Jacobi iteration seems not to.

**Table (a)** Jacobi iteration

| $k$ | $a_1^{(k)}$ | $a_2^{(k)}$ | $a_3^{(k)}$ | $\varepsilon$ |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 1 |
| 1 | 1.00 | 0.00 | 1.83 | 0.76 |
| 2 | 2.83 | -1.87 | 0.12 | 0.65 |
| 5 | 2.25 | -1.44 | 0.50 | 0.28 |
| 10 | 1.92 | -1.23 | 0.75 | 0.082 |
| 15 | 1.90 | - 0.98 | 0.90 | 0.082 |
| * | 1.8 | -1.1 | 0.8 | 0 |

* Exact

**Table (b)** Gauss-Seidel iteration

| $k$ | $a_1^{(k)}$ | $a_2^{(k)}$ | $a_3^{(k)}$ | $\varepsilon$ |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 1 |
| 1 | 1.00 | - 0.50 | 1.33 | 0.50 |
| 2 | 2.33 | -1.50 | 0.45 | 0.33 |
| 5 | 1.980 | -1.235 | 0.680 | 0.11 |
| 10 | 1.776 | -1.082 | 0.816 | 0.015 |
| 15 | 1.803 | -1.102 | 0.796 | 0.0018 |
| * | 1.8 | -1.1 | 0.8 | 0 |

* Exact

A sufficient condition for convergence of the Jacobi and Gauss-Seidel iteration is that the coefficient matrix is *diagonally dominant* (lävistäjävaltainen), that is, on each row



$$\sum_{j=1,\, j\neq i}^{n} |K_{ij}| < |K_{ii}| \tag{8}$$

In the case of Example 13.2 this condition is satisfied only in the equality form in the first equation. However, in the Gauss-Seidel iteration it is enough for convergence if (8) is given with the equality sign if at least one row satisfies (8) with the inequality sign as is the case in Example 13.2. The Gauss-Seidel iteration converges also if the coefficient matrix is positive definite.

In connection with the finite element method the Gauss-Seidel iteration has the advantage that it is not necessary to assemble the whole system matrix and column matrix at all as the evaluation of the terms $K_{ij}\,a^{(k)}$ can be performed by summation at the element level. A disadvantage is the uncertainty about the convergence rate. Also, with a new right-hand side $\{b\}$, the previous solution is of no help (see Remark 13.1). Convergence rate can be improved by using so-called *overrelaxation* (ylirelaksaatio), e.g., Kivelä (1980, p. 228).

**Remark 13.3.** When expressions like (5) are used, one has to take care so that the terms have the same physical dimension. (For instance, one cannot perform summation of nodal parameter values consisting of say of velocity and pressure.) This can be taken into account for example by writing instead of (5) say

$$\|\{a\}\| = \left(c_1 a_1^2 + c_2 a_1^2 + \cdots + c_n a_n^2\right)^{1/2} \tag{9}$$

where the positive factors $c$ make the expression dimensionally homogeneous. □

**Miscellaneous.** Important solution methods, not discussed here, are *projection methods* (projektiomenetelmä), *multigrid methods* (moniverkkomenetelmä), *element-by-element methods* (elementti elementiltä-menetelmä) and *conjugate-gradient methods* (kojugaattigradienttimenetelmä). Reference Pitkäranta (1986) gives a clear explanation of the multigrid method. Reference Reddy and Gartling (2001) contains a detailed exposition of solution of system equations in general.

### 13.1.2 Non-linear equations

**Introduction.** System (1) remains in the non-linear case in the form

$$\boxed{\{F(\{a\})\} = \{0\}} \tag{10}$$

which is often written also as

$$[K(\{a\})]\{a\} = \{b\} \tag{11}$$

The latter form — in which the non-linearity has been buried in the coefficient matrix — is however not unique as is found in Example 13.3.

**Remark 13.4.** In this text non-linear continuum cases are linearized usually already at the differential equation phase as described in Chapter 11. In this case the resulting algebraic system equations are always linear. It is however instructive to consider the Picard method and the Newton-Raphson method also here to see the similarities in the formulations with those discussed in Chapter 11. □

**Example 13.3.** Let us consider the non-linear system, Faux and Pratt (1979, p. 299),

$$\begin{aligned} F_1 &\equiv a_1^2 + a_2^2 - 4 = 0 \\ F_2 &\equiv a_1 a_2 - 1 = 0 \end{aligned} \tag{a}$$

which is quadratic with respect to its unknowns.

Form (11) can be represented for example as

$$\begin{bmatrix} a_1 & a_2 \\ 0 & a_1 \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} 4 \\ 1 \end{Bmatrix} \tag{b}$$

or as

$$\begin{bmatrix} a_1 & a_2 \\ a_2 & 0 \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} 4 \\ 1 \end{Bmatrix} \tag{c}$$

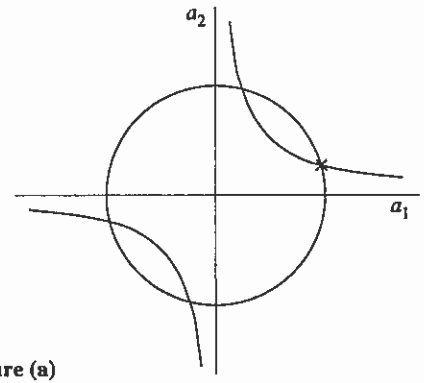



**Figure (a)**

The graphs of the equations in the $a_1 a_2$ - plane are shown in Figure (a). The first graph is a circle with radius 2 centered at the origin. The second graph is a hyperbola whose asymptotes are the $a_1$ - and $a_2$ -axes. According to the figure, the system has four solutions. The point indicated by × is the solution

$$\{a\} = \left[ (2+\sqrt{3})^{1/2}, (2-\sqrt{3})^{1/2} \right]^T \approx [1.932\ ,\ 0.518]^T \tag{d}$$

obtainable analytically.



The example shows clearly that non-linear systems can have several — or no — solutions. Which solution is possibly obtained depends often on the selected starting vector in the iteration. The solution methods for non-linear systems are in practice always iterative.

**Picard method.** Here the iteration takes place based on (11) in the form

$$\left[ K\left( \{a\}^{(k)} \right) \right] \{a\}^{(k+1)} = \{b\} \tag{12}$$

Also the names fixed point iteration, direct iteration, successive approximation are used in this connection. In certain cases — as in heat conduction with temperature dependent conductivity — this kind of simple solution version emerges almost by itself.

**Example 13.4.** We consider the solution of the system of Example 13.3 by the Picard method.

We try to arrive at the solution (d) of Example 13.3:

$$\{a\} \approx [1.932\ ,\ 0.518]^T \tag{a}$$

by taking as the starting vector for example

$$\{a\}^{(0)} = [2\ ,\ 0]^T \tag{b}$$

Forms (b) and (c) of Example 13.3 give the starting sets

$$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \end{Bmatrix}^{(1)} = \begin{Bmatrix} 4 \\ 1 \end{Bmatrix} \tag{c}$$

and

$$\begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \begin{Bmatrix} a_1 \\ a_2 \end{Bmatrix}^{(1)} = \begin{Bmatrix} 4 \\ 1 \end{Bmatrix} \tag{d}$$

respectively.

The coefficient matrix of the latter set is singular and the calculations cannot proceed. The calculations based on the former set proceed as follows:

$$
\begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix}\begin{Bmatrix}a_1\\ a_2\end{Bmatrix}^{(1)} = \begin{Bmatrix}4\\ 1\end{Bmatrix} \Rightarrow \begin{Bmatrix}a_1\\ a_2\end{Bmatrix}^{(1)} = \begin{Bmatrix}2\\ 0.5\end{Bmatrix}
$$

$$
\begin{bmatrix}2 & 0.5\\ 0 & 2\end{bmatrix}\begin{Bmatrix}a_1\\ a_2\end{Bmatrix}^{(2)} = \begin{Bmatrix}4\\ 1\end{Bmatrix} \Rightarrow \begin{Bmatrix}a_1\\ a_2\end{Bmatrix}^{(2)} = \begin{Bmatrix}1.875\\ 0.5\end{Bmatrix}
$$

$$
\begin{bmatrix}1.875 & 0.5\\ 0 & 1.875\end{bmatrix}\begin{Bmatrix}a_1\\ a_2\end{Bmatrix}^{(3)} = \begin{Bmatrix}4\\ 1\end{Bmatrix} \Rightarrow \begin{Bmatrix}a_1\\ a_2\end{Bmatrix}^{(3)} = \begin{Bmatrix}1.991\\ 0.533\end{Bmatrix}
$$

$$
\begin{bmatrix}1.991 & 0.533\\ 0 & 1.991\end{bmatrix}\begin{Bmatrix}a_1\\ a_2\end{Bmatrix}^{(4)} = \begin{Bmatrix}4\\ 1\end{Bmatrix} \Rightarrow \begin{Bmatrix}a_1\\ a_2\end{Bmatrix}^{(4)} = \begin{Bmatrix}1.875\\ 0.502\end{Bmatrix} \tag{e}
$$

$$
\begin{bmatrix}1.875 & 0.533\\ 0 & 1.875\end{bmatrix}\begin{Bmatrix}a_1\\ a_2\end{Bmatrix}^{(5)} = \begin{Bmatrix}4\\ 1\end{Bmatrix} \Rightarrow \begin{Bmatrix}a_1\\ a_2\end{Bmatrix}^{(5)} = \begin{Bmatrix}1.991\\ 0.533\end{Bmatrix}
$$

...

The solutions seem to accumulate at two different slightly erroneous points so the iteration does not converge. (The mean values $a_1 = 1.933$, $a_2 = 0.518$, however, are here near the exact solution.)

The pessimistic end result found in this example should not be feared too much. The problem here is such that the coefficient matrix depends very strongly on the updated solution vector value; for instance at zero vector the coefficient matrix elements vanish. In the applications of the finite element method the behavior is usually much milder.

**Newton-Raphson method.** This is the standard solution method in connection with non-linear systems. Usually the name Newton method is used if only one unknown appears; otherwise the name Newton-Raphson method is used.

Let us consider the illustrative case of only two unknowns:

$$
\begin{aligned}
F_1(a_1, a_2) &= 0\\
F_2(a_1, a_2) &= 0
\end{aligned} \tag{13}
$$

Let the current approximate solution obtained by iteration be $a_1^{(k)}$, $a_2^{(k)}$ which should be made more accurate. Equations (13) are thus not satisfied and we obtain the non-zero residuals

$$
\begin{aligned}
F_1^{(k)} &\equiv F_1\left(a_1^{(k)}, a_2^{(k)}\right)\\
F_2^{(k)} &\equiv F_2\left(a_1^{(k)}, a_2^{(k)}\right)
\end{aligned} \tag{14}
$$

We develop functions $F_1$ and $F_2$ into truncated Taylor series and demand that the resulting expressions vanish:



$$
\begin{aligned}
F_1^{(k+1)} &\approx F_1^{(k)} + \left(\frac{\partial F_1}{\partial a_1}\right)^{(k)} \Delta a_1 + \left(\frac{\partial F_1}{\partial a_2}\right)^{(k)} \Delta a_2 = 0\\
F_2^{(k+1)} &\approx F_2^{(k)} + \left(\frac{\partial F_2}{\partial a_1}\right)^{(k)} \Delta a_1 + \left(\frac{\partial F_2}{\partial a_2}\right)^{(k)} \Delta a_2 = 0
\end{aligned} \tag{15}
$$

In this way again a linear system now for the changes of the values of the unknowns is obtained:

$$
\begin{bmatrix}\partial F_1/\partial a_1 & \partial F_1/\partial a_2\\ \partial F_2/\partial a_1 & \partial F_2/\partial a_2\end{bmatrix}^{(k)} \begin{Bmatrix}\Delta a_1\\ \Delta a_2\end{Bmatrix}^{(k)} = -\begin{Bmatrix}F_1\\ F_2\end{Bmatrix}^{(k)} \tag{16}
$$

The notations for the coefficient matrix and for the right-hand side mean that they are evaluated with the current variable values $a_1^{(k)}$, $a_2^{(k)}$. After the solution of (16) has been determined, the updated variable values are obtained from

$$
\{a\}^{(k+1)} = \{a\}^{(k)} + \Delta\{a\}^{(k)} \tag{17}
$$

and the iteration can be continued until the error in some norm falls under a given tolerance.

The generalization for several unknowns is obvious. The analogue of (16) can be written as

$$
[J]^{(k)} \Delta\{a\}^{(k)} = -\{F\}^{(k)} \tag{18}
$$

where the coefficient matrix with entries

$$
J_{ij} = \partial F_i / \partial a_j \tag{19}
$$

is called in this connection again as the Jacobian matrix.

**Example 13.5.** We consider again the case of Example 13.3 now using the Newton-Raphson method.

Here

$$
\begin{aligned}
F_1 &\equiv a_1^2 + a_2^2 - 4 = 0\\
F_2 &\equiv a_1 a_2 - 1 = 0
\end{aligned} \tag{a}
$$

so

$$
[J] = \begin{bmatrix}\partial F_1/\partial a_1 & \partial F_1/\partial a_2\\ \partial F_2/\partial a_1 & \partial F_2/\partial a_2\end{bmatrix} = \begin{bmatrix}2a_1 & 2a_2\\ a_2 & a_1\end{bmatrix} \tag{b}
$$

Taking the starting vector again as

$$\{a\}^{(0)} = [2\ ,\ 0]^T \tag{c}$$

the calculations proceed as follows:

$$\begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix} \Delta \begin{Bmatrix} a_1 \\ a_2 \end{Bmatrix}^{(0)} = -\begin{Bmatrix} 0 \\ -1 \end{Bmatrix} \Rightarrow \Delta \begin{Bmatrix} a_1 \\ a_2 \end{Bmatrix}^{(0)} = \begin{Bmatrix} 0 \\ 0.5 \end{Bmatrix}$$

$$\{a\}^{(1)} = [2\ ,\ 0.5]^T$$

$$\begin{bmatrix} 4 & 1 \\ 0.5 & 2 \end{bmatrix} \Delta \begin{Bmatrix} a_1 \\ a_2 \end{Bmatrix}^{(1)} = -\begin{Bmatrix} 0.25 \\ 0 \end{Bmatrix} \Rightarrow \Delta \begin{Bmatrix} a_1 \\ a_2 \end{Bmatrix}^{(1)} = \begin{Bmatrix} -1/15 \\ 1/60 \end{Bmatrix} \tag{d}$$

$$\{a\}^{(2)} = [29/15\ ,\ 31/60]^T \approx [1.933\ ,\ 0.517]^T$$

...

The solutions converge very fast. The residuals are after two iterations

$$F_1^{(2)} = 0.004\,72, \quad F_2^{(2)} = -0.001\,11 \tag{e}$$

whereas in the initial guess

$$F_1^{(0)} = 2, \qquad F_2^{(0)} = 0 \tag{f}$$

**Miscellaneous.** Application of the Newton-Raphson iteration means that several consecutive linear systems with different coefficient matrices have to be solved. Because the execution of the LU-decomposition is expensive, the work burden is often lessened by applying the so-called *modified Newton-Raphson* method also called the secant method, chord method. This simply means that the coefficient matrix is updated only now and then and not on each iteration.

A good starting vector is important for attaining convergence. Often a problem contains a parameter on which the non-linearity depends and grows with the parameter value. (For instance the Reynolds number in fluid flow.) One can then proceed so that first an approximate solution for a reasonable small value of the parameter is determined. The solution obtained acts then as a starting vector for a new problem with a higher value for the parameter etc.

## 13.2 EIGENVALUE PROBLEMS

The linear algebraic eigenvalue problem can be presented in the form

$$\underset{n\times n}{[K]}\underset{n\times 1}{\{\hat{a}\}} = \lambda \underset{n\times n}{[M]}\underset{n\times 1}{\{\hat{a}\}} \tag{1}$$



The task is to determine the values $\lambda_i$ of the unknown scalar $\lambda$ — the so-called *eigenvalues* (ominaisarvo) or characteristic numbers or latent roots — and the corresponding values $\{\hat{a}\}_i$ of the column vector $\{\hat{a}\}$ — so-called *eigenvectors* (ominaisvektori) or modes — for which the set (1) is satisfied. We have used the notation $\{\hat{a}\}$ instead of the conventional $\{a\}$ as often when using the finite element method the components of $\{\hat{a}\}$ are not directly the nodal parameters of the unknown function of the original problem.

This kind of problem emerges for instance in vibration and wave phenomena after discretization. Reference Bathe and Wilson (1976) is a well-written important textbook on the theme.

More specifically, problem (1) is called *generalized eigenvalue problem* (yleistetty ominaisarvotehtävä). If matrix $[M]$ is a unit matrix, we obtain a so-called *standard eigenvalue problem* (tavallinen ominaisarvotehtävä):

$$[K]\{a\} = \lambda\{a\} \tag{2}$$

If matrix $[M]$ is non-singular, one can transform the generalized eigenvalue problem (1) into a standard form in theory by multiplication from the left by $[M]^{-1}$ and the algorithms available for the standard problems can be used. However, if $[M]$ is not diagonal, its inverse is no more sparse and one has to operate with full matrices.

The solution of eigenvalue problems leads in practice always to an iterative procedure. In principle we have first to determine the $n$ roots $\lambda_i$ of the characteristic equation

$$\det\left([K] - \lambda[M]\right) = 0 \tag{3}$$

This is a polynomial equation of degree $n$. (Closed form expressions are available only for $n \le 4$.) For each $\lambda_i$ we then have to solve the linear system (1) for the corresponding eigenvector $\{\hat{a}\}_i$. The practical algorithms in use are however based on different lines of thought.

Eigenvalues and eigenvectors have an important role in many theoretical and practical considerations, e.g., Crandall (1956).

## REFERENCES

Bathe, K.-J. and Wilson, E.L. (1967). *Numerical Methods in Finite Element Analysis*. Prentice-Hall.

Crandall, S. (1956). *Engineering Analysis*, McGraw-Hill, New York.

Faux, I. D. and Pratt, M. J. Computational Geometry for Design and Manufacture. Ellis Horwood, Chichester, ISBN 0-470-27069-1.

Hughes, T. J. R. (1987). *The Finite Element Method — Linear Static and Dynamic Finite Element Analysis*, Prentice-Hall, Englewood Cliffs, New Jersey, ISBN 0-13-317017-9.

Kivelä, S., K. (1980). *Matriisilasku ja lineaarialgebra*, Otakustantamo, ISBN 951-671-266-5.

Mäkelä, M., Nevanlinna, O. and Virkkunen, J. (1982). *Numeerinen matematiikka*, Gaudeamus, ISBN 951-662-326-3.

Pitkäranta, J. (1986). Reuna-arvotehtävien moniverkkoratkaisijat — numeronmurskauksesta numeroiden pehmittelyyn. *Arkhimedes*, Vol. 38, pp. 55-56.

Reddy, J. N. and Gartling, D. K. (2001). *The Finite Element Method in Heat Transfer and Fluid Dynamics*,2nd ed., CRC Press, Boca Raton, ISBN 0-8493-2355-X.

Zienkiewicz, O. C. and Taylor, R. L. (2000). *The Finite Element Method*, 5th ed., Butterworth-Heineman, Oxford. Vol. 1: *The Basis*, ISBN 0 7506 5049 4. Vol 2: *Solid Mechanics*, ISBN 0 7506 5055 9. Vol 3: *Fluid Dynamics*, ISBN 0 7506 5050 8.

## PROBLEMS

# NOMENCLATURE

The most important notations used in the text are explained here. The symbols used in the MATHFEM program are not included.

## Sets

| | |
|---|---|
| $\{\}$ | set |
| $\{u$:P$\}$ | set whose typical member is $u$; statement P contains further explanation |
| $R$ | real number |
| $\cup$ | union (yhdiste) |
| $\cap$ | intersection (leikkaus) |
| $\varnothing$ | empty set |
| $\in$ | is a member of |
| $\subset$ | is a subset of |
| $\forall$ | for each |
| $]a,b[$ | open interval |
| $[a,b]$ | closed interval |

## Various integers

| | |
|---|---|
| $n_e$ | number of elements |
| $n_n$ | number of nodes |

## Matrices

| | |
|---|---|
| $[]$ | square or rectangular matrix |
| $\{\}$ | column vector (pystyvektori) |
| $[]^T$ | matrix transpose |
| $[]^{-1}$ | matrix inverse |
| $[]^{-T}$ | $([]^{-1})^T = ([]^T)^{-1}$ |
| $\{a\}$ | column matrix of nodal parameters |
| $\{b\}$ | column matrix of given quantities |
| $[K]$ | coefficient matrix |
| $[J]$ | Jacobian matrix |
| $[M]$ | mass matrix |

## Error study

| | |
|---|---|
| $\lvert\cdot\rvert$ | absolute value, seminorm |
| $\lVert\cdot\rVert$ | norm |
| $(\cdot,\cdot)$ | inner product |
| $a(\cdot,\cdot)$ | bilinear form |
| $b(\cdot)$ | linear form |



| | |
|---|---|
| $C^m$ | continuity up to the $m$th derivative included |

## Latin symbols

| | |
|---|---|
| $a_i$ | nodal parameter, nodal value |
| $A$ | plane surface |
| $c$ | sink factor, heat capacity |
| $D$ | diffusivity |
| $f$ | source term |
| $h$ | heat transfer coefficient, mesh parameter |
| $\mathbf{j}^c$ | convection flux vector |
| $\mathbf{j}^d$ | diffusion flux vector |
| $k$ | thermal conductivity |
| $L$ | linear operator, length |
| $L_i$ | Lagrange interpolation polynomial, length, area, volume coordinate |
| $M$ | bending moment, magnification factor |
| $\mathbf{n}$ | outward unit normal vector |
| $N_i$ | shape function |
| $\mathbf{q}$ | heat flux vector |
| $q$ | heat flow rate density |
| $Q$ | shearing force |
| $r$ | radius |
| $R$ | residual |
| $s$ | curve length, heat source rate per volume |
| $S$ | curved surface |
| $\mathbf{t}$ | stress vector (traction) |
| $t$ | time, thickness |
| $T$ | temperature |
| $u,v,w$ | Cartesian velocity components |
| $\mathbf{v}$ | velocity |
| $V$ | volume |
| $w$ | weighting function |
| $W_i$ | weight coefficient |
| $x,y,z$ | Cartesian coordinates |

## Greek symbols

| | |
|---|---|
| $\alpha$ | weight factor |
| $\alpha,\beta,\gamma$ | coefficients |
| $\delta$ | variation symbol, Kronecker delta |
| $\theta$ | polar angle |
| $\Pi$ | functional |
| $\boldsymbol{\sigma}$ | stress tensor |

| | |
|---|---|
| $\tau$ | sensitizing parameter |
| $\phi$ | typical unknown function |
| $\varphi$ | trial function |
| $\Omega$ | domain (alue) |
| $\bar{\Omega}$ | closure of $\Omega$ = domain and its boundary |
| $\Gamma$ | boundary of $\Omega$ (alueen reuna) |
| $\xi, \eta, \zeta$ | natural coordinates |

## Superscripts

| | |
|---|---|
| $()^c$ | convection |
| $()^e$ | quantity connected to $e$th element |
| $()^r$ | reaction |
| $\dot{()}$ | time derivative |
| $\dot{()}$ | material time derivative |
| $\mathring{()}$ | time rate of change of a non-state quantity |
| $\tilde{()}$ | approximation, finite dimensional |
| $\bar{()}$ | given quantity, extension of Dirichlet data, with sets: closure (sulkeuma) |
| $\hat{()}$ | finite element interpolant to the exact solution, dimensionless quantity |
| $()^+$ | value on the +side |
| $()^-$ | value on the -side |

## Subscripts

| | |
|---|---|
| $()_D$ | Dirichlet |
| $()_L$ | left |
| $()_M$ | middle |
| $()_n$ | outward normal |
| $()_N$ | Neumann |
| $()_p$ | at constant pressure |
| $()_R$ | Robin, right |
| $()_T$ | at constant temperature |
| $()_v$ | velocity |
| $()_\sigma$ | traction |
| $()_\infty$ | free stream value |

## Miscellaneous

| | |
|---|---|
| bt | terms arising from boundary |
| Pe | Peclet number |
| Re | Reynolds number |

# 10 THREE DIMENSIONS

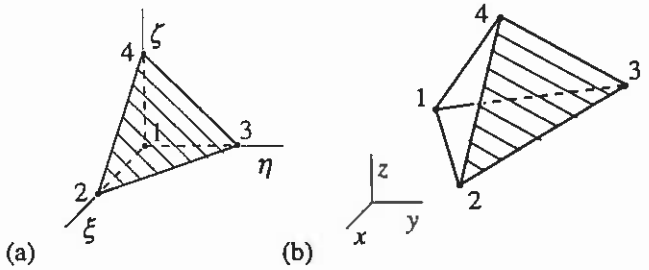
Nothing new is needed on the theoretical level to extend finite element procedures from two space dimensions to three. The computational burden to solve practical problems naturally increases considerably.

## 10.1 SOME ELEMENTS

The elements in two dimensions extend in a natural way to three dimensions. A triangle becomes a *tetraehedron* (tetraedri, nelitahokas) and a quadrilateral a *hexahedron* (heksaedri, kahdeksantahokas). In fact, the elements and shape functions can be extended analogously formally to four or more dimensions. In all cases considered here, isoparametric mapping is used again to generate the elements in the global space. The shape function expressions can be usually detected rather easily directly by inspection using the logic explained in Section 3.2.1.

### 10.1.1 Tetrahedral elements

**Four-noded element.** Figure 10.1 shows a *four-noded* or *linear tetrahedral element* (nelisolmuinen tai lineaarinen tetraedrielementti). It is the extension of the triangular element of Figure 3.7.




**10.1 (a)** Linear reference element. **(b)** Linear element in global space.

The independent natural coordinates are $\xi \in [0,1]$, $\eta \in [0,1]$, $\zeta \in [0,1]$. The shape function expressions are

$$\begin{aligned} N_1 &= L_1 = 1-\xi-\eta-\zeta \\ N_2 &= L_2 = \xi \\ N_3 &= L_3 = \eta \\ N_4 &= L_4 = \zeta \end{aligned} \tag{1}$$

The formulas include the alternative forms in *volume coordinates* (tilavuus-koordinaatti)





$$L_1 = \frac{V_1}{V}, \quad L_2 = \frac{V_2}{V}, \quad L_3 = \frac{V_3}{V}, \quad L_4 = \frac{V_4}{V} \tag{2}$$

These are defined quite analogously as in two dimensions: $V$ is the volume of the tetrahedron, $V_1$ is the volume of the subtetrahedron defined by the vertex points P, 2, 3, 4 where P : $(x,y,z)$ is the generic point inside the tetrahedron, etc. $L_1$ can thus also be interpreted as a dimensionless distance of P from the face 234 etc., (see Figure 3.8 for a corresponding interpretation). The volume coordinates are not independent as they must satisfy the obvious condition

$$L_1 + L_2 + L_3 + L_4 = 1 \tag{3}$$

The isoparametric mapping

$$\begin{aligned} x &= \sum_i N_i x_i \\ y &= \sum_i N_i y_i \\ z &= \sum_i N_i z_i \end{aligned} \tag{4}$$

is used to obtain the element in the global space (Figure 10.1 (b)).

**Ten-noded element.** Figure 10.2 shows a *ten-noded or quadratic tetrahedral element* (kymmensolmuinen tai kvadraattinen tetraedrielementti). It is the extension of the triangular element of Figure 3.9. One possible systematic node numbering order differing from the logic of Figure 3.9 is shown in Figure 10.2.

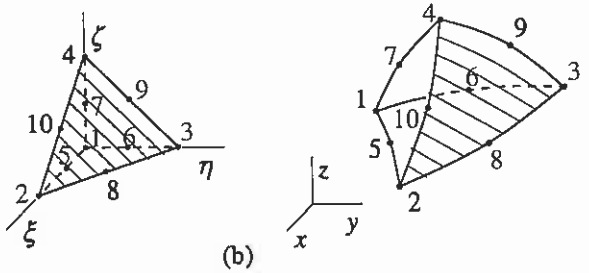



**Figure 10.2 (a)** Quadratic reference element. **(b)** Quadratic element in global space.

Shape functions are for a typical vertex node

$$N_1 = (2L_1 - 1)L_1 \tag{5}$$

and for a midside node

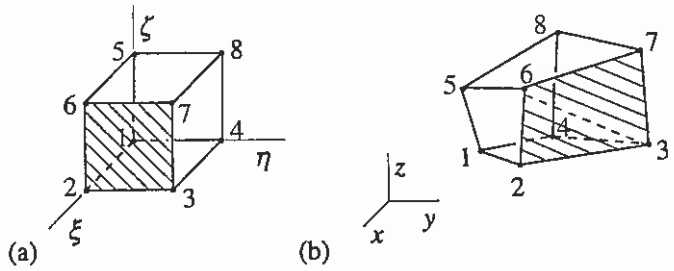
$$N_5 = 4L_1L_2 \tag{6}$$

The element faces in the global space are no more planes in general.

**Remark 10.1.** We have not cared to number the element *faces* (tahko) here although this would be naturally needed in actual calculations using a MATHFEM program type formulation. □

### 10.1.2 Hexahedral elements

**Eight-noded element.** Figure 10.3 shows an *eight-noded or trilinear hexahedral element* (kuusisolmuinen eli trilineaarinen heksaedrielementti). It is the extension of the four-noded quadrilateral element of Figure 3.11. Often the telling name "brick element" is used for hexahedral elements.




**10.3 (a)** Trilinear reference element. **(b)** Trilinear element in global space.

The independent natural coordinates are $\xi \in [0,1]$, $\eta \in [0,1]$, $\zeta \in [0,1]$. Using the nodal numbering order of Figure 10.3, the shape function expressions are

$$\begin{aligned}
N_1 &= (1-\xi)(1-\eta)(1-\zeta) \\
N_2 &= \xi(1-\eta)(1-\zeta) \\
N_3 &= \xi\eta(1-\zeta) \\
N_4 &= (1-\xi)\eta(1-\zeta) \\
N_5 &= (1-\xi)(1-\eta)\zeta \\
N_6 &= \xi(1-\eta)\zeta \\
N_7 &= \xi\eta\zeta \\
N_8 &= (1-\xi)\eta\zeta
\end{aligned} \tag{7}$$

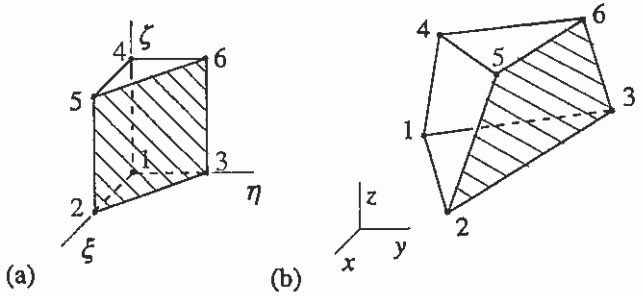
Again, the faces of the element in the global space no more remain planes in general.



**27- noded element.** The extension of the nine-noded quadrilateral element of Figure 3.13 produces a *27-noded or triquadratic hexahedral element* (27-solmuinen eli trikvadraattinen heksaedrielementti). Extension of the eight-noded Serendipity element of Remark 3.5 produces a rather popular 20-noded hexahedral element. We do not consider these elements here in more detail.

### 10.1.3 Wedge elements

In general, a great variety of different elements due to shape and node arrangements are possible in three dimensions. Figure 10.4 shows as a further example a simple *six-noded wedge* or *triangular prism element* (kuusisolmuinen kiila-eli kolmioprismaelementti) .




**10.4 (a)** Six-noded wedge reference element. **(b)** Six-noded wedge element in global space.

Again, the independent natural coordinates are $\xi \in [0,1]$, $\eta \in [0,1]$, $\zeta \in [0,1]$. The shape function expressions are

$$\begin{aligned}
N_1 &= L_1(1-\zeta) = (1-\xi-\eta)(1-\zeta) \\
N_2 &= L_2(1-\zeta) = \xi(1-\zeta) \\
N_3 &= L_3(1-\zeta) = \eta(1-\zeta) \\
N_4 &= L_1\zeta = (1-\xi-\eta)\zeta \\
N_5 &= L_2\zeta = \xi\zeta \\
N_6 &= L_3\zeta = \eta\zeta
\end{aligned} \tag{8}$$

where now the notations $L_1$, $L_2$, $L_3$ refer naturally to the area coordinates (3.2.2).

Reference Zienkiewicz and Taylor (2000), for instance, contains additional collections of three-dimensional elements.

## 10.2 APPLICATION (unfinished)

We consider the corner region formed by three perpendicular walls (Figure 10. 5). The walls are all of constant thickness $t$ and of homogeneous material with a constant thermal conductivity $k$. Convective heat transfer is taking place with equal heat transfer coefficient $h = h^+ = h^-$ on the inside and outside surfaces. The inside and outside air temperatures are $T_\infty^+$ and $T_\infty^-$.

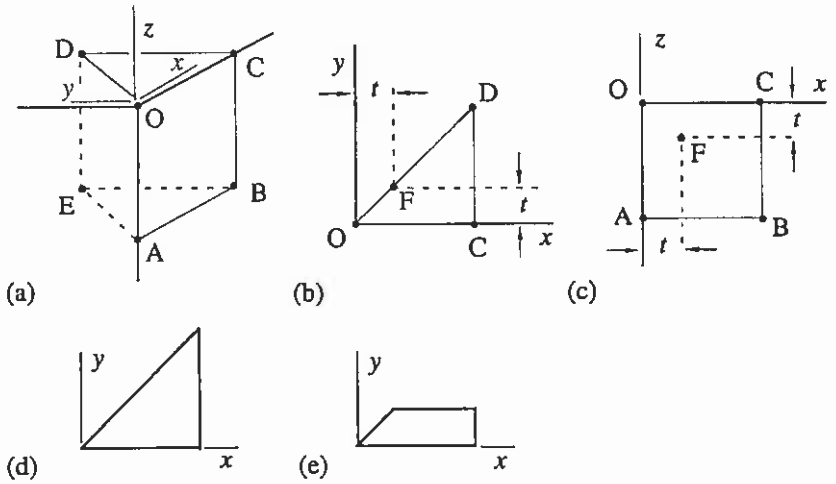



**Figure 10.5** **(a)** Corner region. **(b)** View from the positive $z$-axis direction. **(c)** View from the negative $y$-axis direction. **(d)** Section $z = \text{constant}$, $0 \geq z \geq -t$. **(e)** Section $z = \text{constant}$, $-t > z \geq -3t$.

It is assumed that the walls extend from the corner to a considerably length without no change in the conditions. Based on the resulting symmetry, the wegdelike computational domain with triangular bottom ABE and a triangular top OCD is considered. (This domain contains some volume without material). In fact, due to symmetry, a smaller domain could have been taken but as this is difficult to visualize, we are satisfied with the domain described. To explain the situation in more detail, we give the coordinates of certain points indicated in the figure in the following Table 10.1.

The mesh consisting of ??? elements is shown in Figure 10.6 corresponding to sections shown in Figures 10.5 (d) and (e).



**Table 10.1** Coordinates of some points

| Point | $x$ | $y$ | $z$ |
|---|---|---|---|
| O | 0 | 0 | 0 |
| A | 0 | 0 | $-3t$ |
| B | $3t$ | 0 | $-3t$ |
| C | $3t$ | 0 | 0 |
| D | $3t$ | $3t$ | 0 |
| E | $3t$ | $3t$ | $-3t$ |
| F | $t$ | $t$ | $-t$ |

??

**Figure 10.6** **(a)** ???. **(b)** ???

The standard weak form to be used here is (isotropic thermal conductivity and zero source term)

$$\int_V \left( \frac{\partial w}{\partial x} k \frac{\partial T}{\partial x} + \frac{\partial w}{\partial y} k \frac{\partial T}{\partial y} + \frac{\partial w}{\partial z} k \frac{\partial T}{\partial z} \right) \mathrm{d}V + \int_{S_R} wh(T - T_\infty) \mathrm{d}S = 0 \qquad (1)$$

Surfaces consisting of parts of planes $x = 3t$, $z = -3t$, $y = x$ are Neumann boundaries with $\overline{q} = 0$. On plane $y = x$ this follows from symmetry and on the two other planes it is assumed that the heat flow in the wall is already mainly one-dimensional, that is, perpendicular to the wall. Thus the surface integral over the Neumann boundary in the standard weak form is missing in (1). The inside surfaces $y = t$, $z = -t$ and the outside surfaces $y = 0$, $z = 0$ are Robin boundaries. There is no Dirichlet boundary in this problem.

In the one-dimensional case far from the corner, the temperature distribution is schematically according to Figure 10.7.

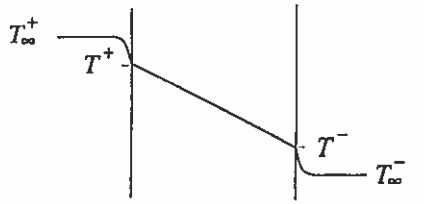



**Figure 10.7** One-dimensional temperature distribution in the wall.

Analytical calculation gives the corresponding wall surface temperatures

$$T^+ = T_\infty^+ - \frac{1}{2+ht/k}\left(T_\infty^+ - T_\infty^-\right)$$
$$T^- = T_\infty^- + \frac{1}{2+ht/k}\left(T_\infty^+ - T_\infty^-\right) \tag{2}$$

We take the following data: $t = 0.20\,\text{m}$, $k = 0.5\,\text{W/(m}\cdot\text{K)}$, $h = 10\,\text{W/(m}^2\cdot\text{K)}$, $T_\infty^+ = 20^\circ\text{C}$, $T_\infty^- = -10^\circ\text{C}$. From formulas (2):

$$T^+ = 15^\circ\text{C}$$
$$T^- = -5^\circ\text{C} \tag{3}$$

The temperature distribution obtained by the finite element method on the outside surface part ABCD in Figure 10.5 (c) is shown in Figure 10.8. As expected, the surface temperature is lowest at the corner. Some comments on the results could be made. From symmetry, the exact solution, say at points D and B should be equal: $T_\text{D} = T_\text{B}$. Finite elements gave $\tilde{T}_\text{D} = ???$, $\tilde{T}_\text{B} = ???$. Also, for the Neumann boundary condition $\bar{q} = 0$ to be realistic on faces $x = 3t$, $z = -3t$, we should obtain $\tilde{T}_\text{D} \approx \tilde{T}_\text{B} \approx T^- = -5^\circ\text{C}$. This is seen to be rather well satisfied ??.

??

**Figure 10.8** Temperature distribution on the outside surface part ABCD by the finite element method.

## REFERENCE

Zienkiewicz, O. C. and Taylor, R. L. (2000). *The Finite Element Method*, 5th ed., Butterworth-Heineman, Oxford. Vol. 1: *The Basis*, ISBN 0 7506 5049 4. Vol 2: *Solid Mechanics*, ISBN 0 7506 5055 9. Vol 3: *Fluid Dynamics*, ISBN 0 7506 5050 8.

## PROBLEMS

# APPENDIX A GENERAL DIFFUSION-CONVECTION-REACTION EQUATION

## A.1 SOME DEFINITIONS

The equation to be studied is of the form

$$\boxed{\overset{(1)}{\frac{\partial \phi}{\partial t}} + \overset{(2)}{\nabla \cdot (-\mathbf{D} \cdot \nabla \phi)} + \overset{(3)}{\nabla \cdot (\mathbf{v}\phi)} + \overset{(4)}{c\phi} - \overset{(5)}{f} = 0} \tag{1a}$$

or using Cartesian coordinates and the summation convention

$$\frac{\partial \phi}{\partial t} + \frac{\partial}{\partial x_i}\left(-D_{ij}\frac{\partial \phi}{\partial x_j}\right) + \frac{\partial}{\partial x_i}(v_i \phi) + c\phi - f = 0 \tag{1b}$$

Function $\phi(\mathbf{x},t)$ or $\phi(x_i,t)$ is the unknown to be determined for position $\mathbf{x} \in \bar{\Omega} = \Omega \cup \Gamma$ and for time $t \in [0,T]$. This kind of equation — or at least a closely similar one — is found to be present in various applications of continuum mechanics when the Eulerian description is employed, as is usual in fluid mechanics. The equation is called here the general *diffusion-convection-reaction equation* (diffuusio-konvektio-reaktioyhtälö) (later D-C-R equation). We will discuss first separately the five terms indicated in equation (1). The formulas are given both in symbolic form (to help to reference to literature) and in index notation in Cartesian coordinates (for hopefully increasing the readability).

(1) *Unstationary term* (epästationaarisuustermi) $\partial\phi/\partial t$. Another name is the *unsteady term* or the *time derivative* term. If this term is zero (and if further the boundary conditions do not depend on time) the solution $\phi(\mathbf{x})$ does not depend on time and we have the so-called *stationary* (stationaarinen) or *steady* (pysyvä) diffusion-convection-reaction problem.

(2) *Diffusion term* (diffuusiotermi) $\nabla \cdot (-\mathbf{D} \cdot \nabla \phi)$. $\mathbf{D}$ is the so-called generalized *diffusivity tensor* (diffusiivisuustensori). It is usually symmetric. Sometimes the notation

$$\mathbf{j}^{\mathrm{d}} = -\mathbf{D} \cdot \nabla \phi, \qquad j_i^{\mathrm{d}} = -D_{ij}\frac{\partial \phi}{\partial x_j} \tag{2}$$

is used. $\mathbf{j}^{\mathrm{d}}$ is called the *diffusion flux vector* (diffuusiovuovektori). Actually expression (2) is the most usual type of constitutive relationship *assumed* for the diffusion flux, c.f. for instance the Fourier law of heat conduction. If $\mathbf{D}$ is





isotropic, that is, $\mathbf{D} = D\mathbf{I}$ or $D_{ij} = D\delta_{ij}$, where $D$ is a scalar, $\mathbf{I}$ the identity tensor and $\delta_{ij}$ the Kronecker delta,

$$\mathbf{j}^{\mathrm{d}} = -D\nabla\phi, \qquad j_i^{\mathrm{d}} = -D\frac{\partial \phi}{\partial x_i} \tag{3}$$

The diffusion term gives the transfer of a quantity $\phi$ due to microscopic processes through a medium even if it is at rest, say heat transfer due to conduction.

(3) *Convection term* (konvektiotermi) $\nabla \cdot (\mathbf{v}\phi)$. This is also called the *advection term* (advektiotermi). Quantity $\mathbf{v}$ is the flow velocity vector of the medium. Sometimes the notation

$$\mathbf{j}^{\mathrm{c}} = \mathbf{v}\phi, \qquad j_i^{\mathrm{c}} = v_i\phi \tag{4}$$

is used. $\mathbf{j}^{\mathrm{c}}$ is called the *convection flux vector* (konvektiovuovektori). The convection term is generated as the continuum transfers with velocity $\mathbf{v}$ the quantity $\phi$ bounded to its particles. The convection type term is typical in fluid mechanics but appears seldom in solid mechanics when the Lagrangian description is used in the latter.

(4) *Reaction term* (reaktiotermi) $c\phi$. This name has it basis in chemical reactions producing this type of term. Often the source term is not strictly a given quantity, but may contain the unknown function. A linear expansion around the current value gives a reaction term. It may be mentioned that the part "reaction" is rather seldom used in the name of the D-C-R equation even if it is contained in (1). Sometimes the name *generalized sink factor* (yleistetty nielutekijä) or *absorption coefficient* is used for $c$. Turbulence model equations, the Coriolis force in momentum equations written in a rotating frame and equilibrium equations for elastic materials on elastic supports are further examples of cases containing the reaction type term.

(5) *Source term* (lähdetermi) $f$. As indicated above, in addition to being a given forcing function, some "inconvenient terms are often buried" in the source term which is then updated iteratively during the solution process. The source term is often called also the *forcing term* (herätetermi).

Using notations (3) and (4), we can now represent the D-C-R equation for later purposes simply as

$$\boxed{\frac{\partial \phi}{\partial t} + \nabla \cdot \mathbf{j}^{\mathrm{d}} + \nabla \cdot \mathbf{j}^{\mathrm{c}} + c\phi - f = 0} \qquad \frac{\partial \phi}{\partial t} + \frac{\partial j_i^{\mathrm{d}}}{\partial x_i} + \frac{\partial j_i^{\mathrm{c}}}{\partial x_i} + c\phi - f = 0 \tag{5}$$

The conventional *boundary conditions* (reunaehto) in connection with the D-C-R equation are

$$\begin{aligned} \phi &= \bar{\phi} && \text{on } \Gamma_D \\ j^d &= \bar{j}^d && \text{on } \Gamma_N \\ j^d &= a\phi + b && \text{on } \Gamma_R \end{aligned} \tag{6}$$

where $\bar{\phi}$, $\bar{j}^d$, $a$, $b$ are given functions of position and time. The term

$$j^d \equiv \mathbf{n} \cdot \mathbf{j}^d = -\mathbf{n} \cdot \mathbf{D} \cdot \nabla\phi = n_i j_i^d = -n_i D_{ij} \frac{\partial \phi}{\partial x_j} \tag{7}$$

where $\mathbf{n}$ is the unit outward normal vector to the boundary, may be called the *diffusion flux density* (diffuusiovuotiheys). Parts $\Gamma_D$, $\Gamma_N$, $\Gamma_R$ form together without gaps or overlaps the whole space boundary $\Gamma$. The notations are similar to those used in Chapter 3 in connection with heat conduction. The Dirichlet and the Neumann boundary conditions are seen to be obtainable as special cases of the Robin condition. To keep some of the formulas as basic as possible, the constitutive relation (2) should preferably be substituted as late as possible.

In the unsteady case the distribution of $\phi$ in $\Omega$ at the initial instant of time must also be given:

$$\phi(\mathbf{x}, t) = \phi_0(\mathbf{x}) \quad \text{at } t = 0 \tag{8}$$

This is called the *initial condition* (alkuehto).

**Remark A.1.** The D-C-R equation obtains many variations in outlook if the *material (time) derivative* (ainederivaatta, aineellinen aikaderivaatta, substantiaalinen derivaatta) expression in the Eulerian description

$$\frac{Df(\mathbf{x}, t)}{Dt} = \frac{\partial f}{\partial t} + \mathbf{v} \cdot \nabla f, \qquad \frac{Df(x_i, t)}{Dt} = \frac{\partial f}{\partial t} + v_i \frac{\partial f}{\partial x_i} \tag{9}$$

for a general function $f$ of position and time and if the continuity equation (see (A.2.1)), obtainable from the mass conservation principle, are made use of. □

**Remark A.2.** The convection term appears often in a modified form, which follows from the vector calculus identity

$$\nabla \cdot (\mathbf{v}\phi) = \mathbf{v} \cdot \nabla\phi + (\nabla \cdot \mathbf{v})\phi, \qquad \frac{\partial}{\partial x_i}(v_i \phi) = v_i \frac{\partial \phi}{\partial x_i} + \frac{\partial v_i}{\partial x_i}\phi \tag{10}$$

The D-C-R equation can thus be written equally well as



$$\frac{\partial \phi}{\partial t} + \nabla \cdot (-\mathbf{D} \cdot \nabla\phi) + \mathbf{v} \cdot \nabla\phi + \bar{c}\phi - f = 0 \tag{11}$$

where $\bar{c}$ now may contain the term $\nabla \cdot \mathbf{v}$. Especially in an *incompressible flow* (kokoonpuristumaton virtaus) situation, where $\nabla \cdot \mathbf{v} = 0$, the convection term can be written thus also as $\mathbf{v} \cdot \nabla\phi$ or as $v_i \partial\phi / \partial x_i$ also in (1). □

**Remark A.3.** Equation (1) appears in the literature frequently in such a form, that $\phi$ is replaced in the unstationary term and in the convection term by $\rho\phi$:

$$\frac{\partial}{\partial t}(\rho\psi) + \nabla \cdot (-\mathbf{D} \cdot \nabla\psi) + \nabla \cdot (\mathbf{v}\rho\psi) + c\psi - f = 0 \tag{12}$$

We have employed here the notation $\psi$ for the unknown function to emphasize the difference in form. This increases the number of different versions still more. □

**Remark A.4.** Fluid mechanics problems usually consist of several simultaneous D-C-R type equations where in each quantity $\phi$ has a different physical meaning and the solution must be found from a system of coupled partial differental equations. One solution strategy is, however, at least as a thought experiment to try to solve each $\phi$ from its "own" equation considering then the other unknowns temporarily as given. As the equation system is usually nonlinear especially due to the momentum equations, one has in any case to use an iterative solution method, and thus this way of thought is quite natural. In any case one can learn much by studying qualitatively the behavior of the solution of a typical scalar D-C-R equation. It is obvious that the nature of the solution must depend on the relative magnitudes of the convection, diffusion, and reaction terms. □

## A.2 SPECIAL CASES

Next the continuity equation, the energy equation, and the momentum equations are considered as special cases of the D-C-R equation.

*Continuity equation* (jatkuvuusyhtälö) is

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0, \qquad \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x_i}(\rho v_i) = 0 \tag{1}$$

Comparing with (A.1.1) we have

$$\phi \triangleq \rho, \quad \mathbf{j}^d = \mathbf{0}, \quad c = 0, \quad f = 0 \tag{2}$$

*Energy equation* (energiayhtälö) is for example in the mechanically incompressible case (see Section 6.1.1)

$$\rho c_p \left( \frac{\partial T}{\partial t} + \mathbf{v} \cdot \nabla T \right) = \nabla \cdot (\mathbf{k} \cdot \nabla T) + s + \Phi \tag{3a}$$

or

$$\rho c_p\left(\frac{\partial T}{\partial t}+v_i\frac{\partial T}{\partial x_i}\right)=\frac{\partial}{\partial x_i}\left(k_{ij}\frac{\partial \phi}{\partial x_j}\right)+s+\Phi \tag{3b}$$

where $T$ is the temperature, $c_p$ the specific heat capacity at constant pressure, $\mathbf{k}$ the heat conductivity tensor, $s$ the heat source rate per volume, and $\Phi$ the dissipation function. If the term $\rho c_p$ is assumed to be constant in space (if not, the source term obtains some extra terms), we obtain further

$$\frac{\partial T}{\partial t}+\nabla\cdot\left(-\frac{\mathbf{k}}{\rho c_p}\cdot\nabla T\right)+\mathbf{v}\cdot\nabla T-\frac{s+\Phi}{\rho c_p}=0 \tag{4}$$

Comparison with (A.1.11) gives the interpretations

$$\phi\triangleq T,\quad \mathbf{D}\triangleq\frac{\mathbf{k}}{\rho c_p},\quad \bar{c}=0,\quad f\triangleq\frac{s+\Phi}{\rho c_p} \tag{5}$$

The dissipation function depends on the velocity field, and thus increases the coupling between the temperature and velocity.

*Momentum equations* (liikemääräyhtälö, liikeyhtälö) are

$$\frac{\partial}{\partial t}(\rho\mathbf{v})-\nabla\cdot\boldsymbol{\sigma}^*+\nabla\cdot(\rho\,\mathbf{v}\mathbf{v})+\nabla p-\rho\mathbf{b}=0 \tag{6a}$$

or

$$\frac{\partial}{\partial t}(\rho v_i)-\frac{\partial\sigma^*_{ji}}{\partial x_j}+\frac{\partial}{\partial x_j}(\rho v_j v_i)+\frac{\partial p}{\partial x_i}-\rho b_i=0 \tag{6b}$$

where $\boldsymbol{\sigma}^*$ is the deviatoric stress tensor, $p$ the pressure and $\mathbf{b}$ the specific body force. Comparison with (A.1.12) gives the interpretations

$$\psi\triangleq\mathbf{v},\quad \mathbf{j}^{\mathrm{d}}\triangleq-\boldsymbol{\sigma}^*,\quad c=0,\quad f\triangleq-\nabla p+\rho\mathbf{b} \tag{7}$$

The deviatoric stress tensor is not exactly of the form $\mathbf{D}\cdot\nabla\mathbf{v}$, but in any case it depends on the space derivatives of $\mathbf{v}$ contrary to the convection flux $\mathbf{j}^{\mathrm{c}}=\rho\,\mathbf{v}\mathbf{v}$ thus giving rise to second order derivatives in the field equation. It is realized that to consider equation (6) linear with respect to $\mathbf{v}$, we have to represent the tensor product in $\mathbf{j}^{\mathrm{c}}$, say, in the form $\bar{\mathbf{v}}\mathbf{v}$ where $\bar{\mathbf{v}}$ is assumed to be known from a previous iteration.



Comparison of the terms in the D-C-R equation with form (A.1.1) or (A.1.12) shows that for example the quantities $D_{11}\partial\phi/\partial x_1$ and $v_1\phi$ or $D_{11}\partial\psi/\partial x_1$ and $\rho v_1\psi$, correspondingly, must have the same dimension. Thus we can infer that

$$\boxed{\mathrm{Pe}=\frac{vL}{D}}\quad\text{or}\quad\boxed{\mathrm{Pe}=\frac{\rho vL}{D}} \tag{8}$$

the *Peclet number* ( Peclet'n luku), is a dimensionless quantity, which measures in some sense the relative magnitude of convection with respect to diffusion. Quantities $v$, $L$, and $D$ are agreed characteristic speed of flow, linear measure of the domain and diffusivity of the medium.

No diffusion is present in the continuity equation (1) and thus $\mathrm{Pe}=\infty$. In the energy equation (4), $\mathrm{Pe}=vL\rho c_p/k$, where $k$ is a characteristic heat conductivity. In the momentum equations (6) the diffusivity is represented by the coefficient of viscosity $\mu$ and the Peclet number is seen to represent the Reynolds number $\mathrm{Re}=\rho vL/\mu$.

It is illuminating to consider the extreme cases $\mathrm{Pe}=\infty$ and $\mathrm{Pe}=0$. (For simplicity we start in the following in the steady case.) In the former case diffusion disappears completely in comparison with convection and we obtain the *pure convection equation* (puhdas konvektioyhtälö). In the latter case convection disappears completely in comparison with diffusion and we obtain the *pure diffusion equation* (puhdas diffuusioyhtälö). The order of the D-C-R equation drops by one when the limit $\mathrm{Pe}=\infty$ is reached. This fact has its effect also on the boundary conditions available.

## A.3 QUALITATIVE BEHAVIOUR

The direction of the flow field with respect to the boundary is significant from the boundary condition point of view especially at the limit $\mathrm{Pe}=\infty$. The boundary is divided (Figure A.1) into the *inflow boundary* (sisäänvirtausreuna) $\Gamma_-$, the *neutral boundary* (neutraali reuna) $\Gamma_\mp$ and the *outflow boundary* (ulosvirtausreuna) $\Gamma_+$ according to

$$\begin{aligned}\Gamma_-&=\{\mathbf{x}:\mathbf{x}\in\Gamma,\ \mathbf{n}\cdot\mathbf{v}<0\}\\ \Gamma_\mp&=\{\mathbf{x}:\mathbf{x}\in\Gamma,\ \mathbf{n}\cdot\mathbf{v}=0\}\\ \Gamma_+&=\{\mathbf{x}:\mathbf{x}\in\Gamma,\ \mathbf{n}\cdot\mathbf{v}>0\}\end{aligned} \tag{1}$$

Quantity $\mathbf{n}\cdot\mathbf{v}$ is the velocity component in the normal direction to the boundary positive in the outward direction. The meaning of the terminology used is thus understandable. The neutral boundary appears for instance with

real fluids for stationary solid walls (the velocity vanishes) and with ideal fluids also for stationary solid walls (the velocity is parallel to the wall).

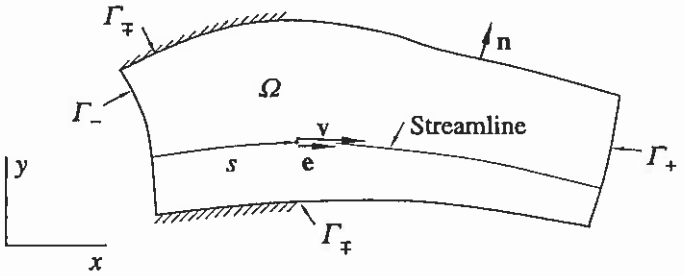



**Figure A.1** Some notations.

For easy graphical visualization we shall consider next the steady D-C-R plane case (Figure A.1) with $x_1 \to x$, $x_2 \to y$, $v_1 \to u$, $v_2 \to v$. Equation (1b) obtains the form

$$-\frac{\partial}{\partial x}\left(D_{xx}\frac{\partial \phi}{\partial x}+D_{xy}\frac{\partial \phi}{\partial y}\right)-\frac{\partial}{\partial y}\left(D_{yx}\frac{\partial \phi}{\partial x}+D_{yy}\frac{\partial \phi}{\partial y}\right)$$
$$+\frac{\partial}{\partial x}(u\phi)+\frac{\partial}{\partial y}(v\phi)+c\phi-f=0 \qquad (2)$$

The pure convection equation corresponding to this is

$$\frac{\partial}{\partial x}(u\phi)+\frac{\partial}{\partial y}(v\phi)-f=0 \qquad (3)$$

or by expanding the derivatives,

$$u\frac{\partial \phi}{\partial x}+v\frac{\partial \phi}{\partial y}+\left(\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}\right)\phi-f=0 \qquad (4)$$

Let us consider the form this equation obtains on a certain streamline. Let the unit vector in the direction of the selected positive arclength $s$ direction on the streamline be e (Figure A.1). Thus

$$u=e_x\bar{v}, \quad v=e_y\bar{v} \qquad (5)$$

where $\bar{v}$ is the scalar flow velocity positive in the e-direction. On the other hand, the derivative of $\phi$ in the direction of e is by vector calculus



$$\frac{\partial \phi}{\partial e}=\mathbf{e}\cdot\nabla\phi=e_x\frac{\partial \phi}{\partial x}+e_y\frac{\partial \phi}{\partial y} \qquad (6)$$

so that the term

$$u\frac{\partial \phi}{\partial x}+v\frac{\partial \phi}{\partial y}=\bar{v}\left(e_x\frac{\partial \phi}{\partial x}+e_y\frac{\partial \phi}{\partial y}\right)=\bar{v}\frac{\partial \phi}{\partial e} \qquad (7)$$

The pure convection equation on a certain streamline has thus the form (we denote $\partial\phi/\partial e=\mathrm{d}\phi/\mathrm{d}s$)

$$\boxed{\bar{v}\frac{\mathrm{d}\phi}{\mathrm{d}s}+\left(\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}\right)\phi-f=0} \qquad (8)$$

When we consider — as agreed — the velocity field and the source term as given in $\Omega$, we in principle know them on the streamline as functions of $s$. Thus the *steady pure convection equation is an ordinary first order differential equation on a streamline.* The independent variable is the arclength $s$ along the streamline (or some other suitable curve parameter). This result is of course valid also in three space dimensions. In one dimension the, say, $x$-axis is the streamline and $x$ is the arclength. It may be mentioned that the inclusion of the reaction term $c\phi$ in (8) would not clearly change the conclusions obtained.

The pure convection equation is classified in mathematics texts as a hyperbolic equation and the streamlines are called characteristic curves. The quantity $\phi$ is determined on a streamline in principle from the solution of the differential equation after the value of $\phi$ is given at one point somewhere on the streamline. For physical reasons this point is on the inflow boundary $\Gamma_-$. Thus the boundary condition for the pure convection equation is simply

$$\phi=\bar{\phi} \qquad \text{on } \Gamma_- \qquad (9)$$

*On the rest of the boundary no boundary condition is needed or in fact no condition can be given.*

On the basis of equation (8) we can say that in the pure convection equation "effect or information is carried only in the streamline direction". This is easiest to see in the incompressible case $\partial u/\partial x+\partial v/\partial y=0$ and when $f=0$. Equation (8) obtains the form $\mathrm{d}\phi/\mathrm{d}s=0$ or simply $\phi=$constant on a streamline. Figure A.2 gives a typical qualitative solution for the geometry and flow field of Figure A.1. If $\bar{\phi}$ would be completely discontinuous, the corresponding jump would also propagate into the domain without any smoothening.

**Remark A.5.** If there are contrary to Figure A.1 closed streamlines inside the domain, an imaginary cut crossing each streamline once must be drawn and the value of $\phi$ given on one side of the cut for $\phi$ to be determinate in the whole domain. □

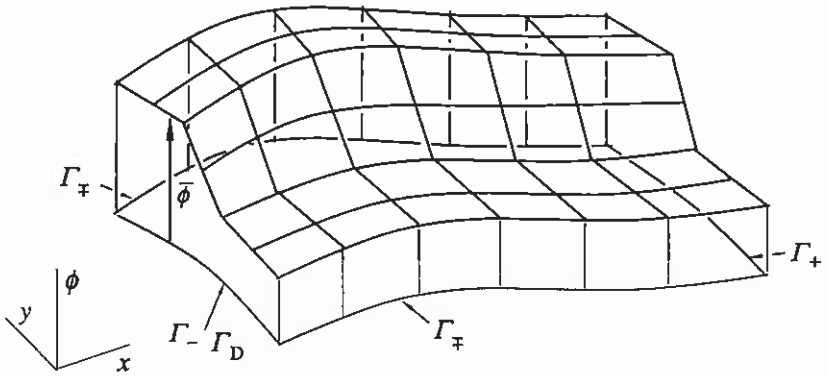



**Figure A.2** A qualitative solution for the pure convection equation.

The pure diffusion equation corresponding to (2) is after a change of sign

$$\frac{\partial}{\partial x}\left(D_{xx}\frac{\partial\phi}{\partial x}+D_{xy}\frac{\partial\phi}{\partial y}\right)+\frac{\partial}{\partial y}\left(D_{yx}\frac{\partial\phi}{\partial x}+D_{yy}\frac{\partial\phi}{\partial y}\right)+f=0 \tag{10}$$

Further development gives ($D_{yx}=D_{xy}$)

$$\boxed{D_{xx}\frac{\partial^2\phi}{\partial x^2}+2D_{xy}\frac{\partial^2\phi}{\partial x\partial y}+D_{yy}\frac{\partial^2\phi}{\partial y^2}+f^*=0} \tag{11}$$

where some terms have been buried in

$$f^*=f+\left(\frac{\partial D_{xx}}{\partial x}+\frac{\partial D_{yx}}{\partial y}\right)\frac{\partial\phi}{\partial x}+\left(\frac{\partial D_{yy}}{\partial y}+\frac{\partial D_{xy}}{\partial x}\right)\frac{\partial\phi}{\partial y} \tag{12}$$

It is seen that if the diffusivities vary with position, in fact some convection type terms emerge. According to mathematics texts, (11) is an elliptic differential equation if, Crandall (1956, p. 355)

$$\left(2D_{xy}\right)^2-4D_{xx}D_{yy}=4\left(D_{xy}^2-D_{xx}D_{yy}\right)<0 \tag{13}$$

For physical reasons the diffusivity matrix $[D]$ is usually positive definite so that, Crandall (1956, p. 15)



$$\det[D]=D_{xx}D_{yy}-D_{xy}^2>0 \tag{14}$$

The inequality (13) is then valid and thus the *steady pure diffusion equation is a second order elliptic partial differential equation. Boundary conditions must be given on the whole boundary* $\Gamma$. The inclusion of the reaction term in (10) would not change the conclusions obtained. Boundary conditions (A.1.6) can all be used. The steady heat conduction problem is a standard example leading to a typical pure diffusion equation.

Let us consider as an illustrative special case the *Poisson's equation*

$$\frac{\partial^2\phi}{\partial x^2}+\frac{\partial^2\phi}{\partial y^2}+\frac{f}{D}=0 \tag{15}$$

which is obtained with an isotropic and homogeneous diffusivity. This equation is known to describe for instance the behavior of a stretched membrane where $\phi$ is the (small) transverse displacement of the membrane due to the transverse loading per unit surface $f$ (and due to the given boundary displacements) and $D$ the uniform tension per unit membrane length.

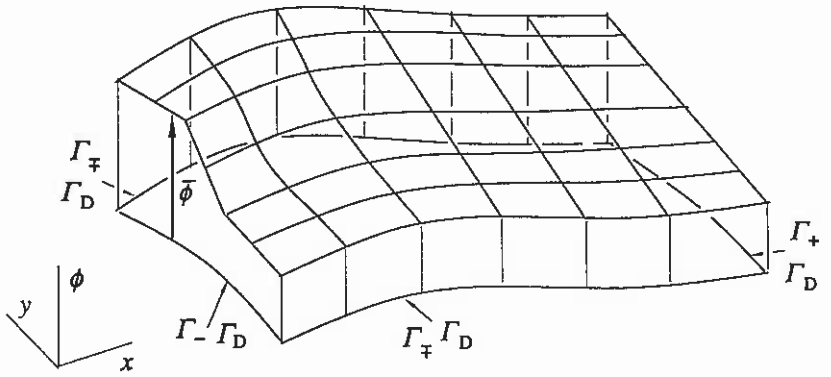



**Figure A.3** A qualitative solution for the pure diffusion equation.

Figure A.3 shows a rough sketch of the type of solution to be expected for the pure diffusion equation with Dirichlet boundary conditions on the whole boundary. The same non-smooth distribution has been taken for the boundary $\Gamma_-$ as in Figure A.2. At least with the interpretations in connection with (15), it is easy to see intuitively that the solution smoothes all irregularities possibly present in the boundary data. It is also obvious that in a longish geometry as in the figure the boundary conditions at $\Gamma_-$ cannot have much influence on the solution far from $\Gamma_-$. In other words, the solution is determined more or less by

the boundary data nearest to the point in question. This behavior is completely different in nature from that corresponding to the pure convection equation where the inflow data is transferred undiminished into the solution domain.

On the basis of the discussion above it is apparent that the solution for a general D-C equation must be roughly some kind of weighted average of the solution for the pure convection and for the pure diffusion equation. The actual value of the Peclet number has an essential effect on the end result. When $\mathrm{Pe} \neq \infty$, boundary conditions (A.1.6) must be given on the whole boundary.

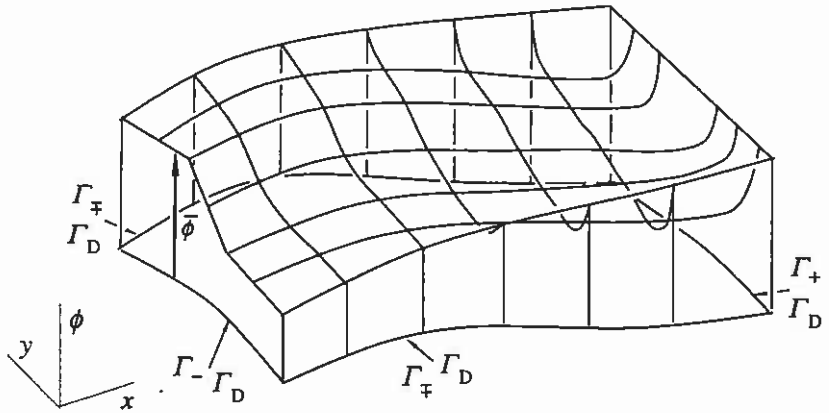



**Figure A.4** A qualitative solution for the diffusion-convection equation.

A possible type of solution for a moderate value of the Peclet number has been sketched in Figure A.4. Dirichlet boundary conditions have been assumed on the whole boundary and the data on the inflow boundary is the same as in Figures A.2 and A.3 and on the rest of the boundary the data is the same as in Figure A.3. The nature of the solution can be understood by a simultaneous study of Figures A.2, A.3, and A.4. The solution in the domain must resemble the more the solution of the pure convection equation the higher the value of Pe. The solution, however, has to satisfy the boundary conditions on $\Gamma_{\mp}$ and $\Gamma_+$ not present in the pure convection problem. This means that $\phi$ must alter its values with large gradients in the neighborhood of these boundaries; a *boundary layer* (rajakerros) is generated. The boundary layer concept is not confined to the velocity components, for which the phenomenon is perhaps best known, but in principle any of the dependent quantities such as temperature or concentration of a species can show a similar behavior. As the gradients perpendicular to the boundary layer are large so is also the diffusion (the second derivatives must also obtain large values). Thus for instance in connection with the flow field the diffusion flux in the boundary layer — essentially the same as



the shearing stress — cannot be neglected if realistic results are to be expected even when the layer is very thin.

When numerical methods are employed, one must be prepared to model boundary layers either with dense meshes or by basing the model to some theories taking in advance into account the properties of the boundary layer.

**Remark A.6.** *Diffusion tends to smooth possible irregularities in the solution.* In longish geometries in the flow direction the boundary layers have time to get thick due to diffusion and they can finally fill the whole flow field. The term *crosswind diffusion* or *false diffusion* (poikkivirtadiffuusio, valediffuusio) is often used in connection with numerical methods. This means shortly a harmful phenomenon in which diffusion type results are obtained due to the numerical procedure even in pure convection problems. False diffusion takes place when the flow direction is oblique to the grid lines and when simultaneously the dependent quantity has a non-zero gradient component perpendicular to the flow direction, Patankar (1980, p. 108). □

**Remark A.7.** The Dirichlet type boundary condition on the outflow boundary $\Gamma_+$ — as used for demonstration purposes in connection with Figure A.4 — is for physical reasons not very sensible when the Peclet number is relative large. In fact the flow field carries the unknown with it and we usually have no realistic basis to prescribe meaningful Dirichlet data on the outflow boundary. A rather standard way is to assume that the diffusion flux is small compared to the convection flux and to employ the Neumann condition (A.1.6) in the form $j^{\mathrm{d}} = 0$. This kind "soft" boundary condition is not so demanding on numerical methods as the "hard" Dirichlet condition as the exact solution behaves then usually smoothly in the neighborhood of $\Gamma_+$. □

The steady *pure reaction equation* (puhdas reaktioyhtälö)

$$\boxed{c\phi - f = 0} \tag{16}$$

is no more a differential equation and we can solve it directly for $\phi$:

$$\phi = \frac{f}{c} \tag{17}$$

If boundary data is given, it has no connection with the field equation. If $f$ is discontinuous, so is $\phi$. In practice there is always some amount of diffusion present making the solution continuous. Figures A.5(b) and (c) show the corresponding solution behavior in one dimension. Boundary layer and *internal layer* (sisäkerros) type solutions can thus again emerge and difficulties with the numerical solution are to be expected.

A study of equation (A.1.1) shows that the dimensionless quantity

$$\boxed{\mathrm{Ce} = \frac{cL^2}{D}} \tag{18}$$

measures in some sense the relative magnitude of reaction with respect to diffusion. Quantities $c$, $L$, and $D$ are agreed characteristic sink factor value, linear measure of the domain and diffusivity of the medium. It seems that quantity (18) has no settled name in the literature, so we will call it the *Ceclet number* (Ceclet'n luku) by inventing an imaginary person with a name similar to Peclet.

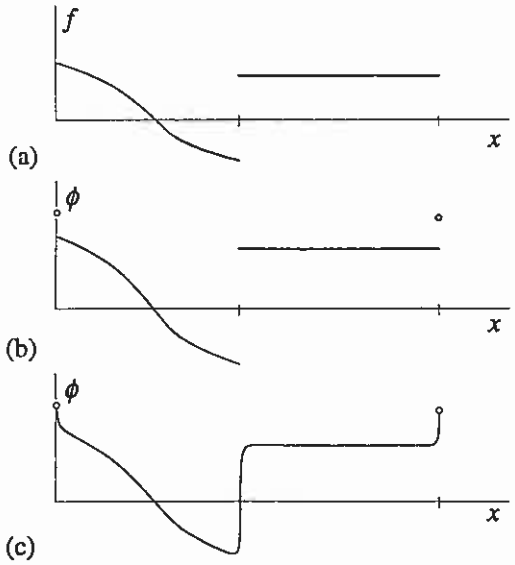



**Figure A.5(a)** Discontinuous source term. **(b)** A qualitative solution for the pure reaction equation. **(c)** A qualitative solution for the diffusion-reaction equation with large reaction.

We have this far considered the D-C-R equation without the unsteady term. Completing equation (A.2.2) in this respect, we obtain

$$\frac{\partial \phi}{\partial t}+\frac{\partial j_x^{\mathrm{d}}}{\partial x}+\frac{\partial j_y^{\mathrm{d}}}{\partial y}+\frac{\partial}{\partial x}(u\phi)+\frac{\partial}{\partial y}(v\phi)+c\phi-f=0 \qquad (19)$$

where the diffusion terms are expressed using shortly the flux vector notations. Employing a little bit of imagination this new situation can be represented alternatively as follows:

$$\frac{\partial j_x^{\mathrm{d}}}{\partial x}+\frac{\partial j_y^{\mathrm{d}}}{\partial y}+\frac{\partial\left(j_t^{\mathrm{d}}\equiv 0\right)}{\partial t}+\frac{\partial}{\partial x}(u\phi)+\frac{\partial}{\partial y}(v\phi)+\frac{\partial}{\partial t}(1\phi)+c\phi-f=0 \qquad (20)$$



The idea is to try to make use of what we have learned earlier in connection with the steady two-dimensional D-C-R equation. We now have a three-dimensional "steady" D-C-R equation having "streamlines" in the $xyt$-space with the velocity field $(u,v,1)$. The diffusion flux component is zero and the velocity component is 1 in the time direction. This interpretation can be useful say for instance as follows. Maybe we have devised a well behaving numerical method for the true steady case. We can immediately try to use the same kind of method also in the unsteady case by just considering time as an additional space coordinate.

## REFERENCES

Crandall, S. (1956). *Engineering Analysis*, McGraw-Hill, New York.

Patankar, S. H. (1980). *Numerical Heat Transfer and Fluid Flow*, Mc-Graw-Hill, New York, ISBN 0-07-048740-5.

# APPENDIX B INTEGRATION BY PARTS

## B.1 ONE DIMENSION

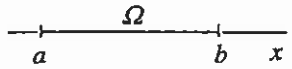



**Figure B.1** One-dimensional domain $\Omega = ]a,b[$ and its boundary $\Gamma = \{a,b\}$.

The one-dimensional integration by parts formula can be written as (Figure B.1))

$$\int_a^b g\frac{dh}{dx}dx = -\int_a^b \frac{dg}{dx}h dx + \Big|_a^b gh \tag{1a}$$

or using more general notation as

$$\boxed{\int_\Omega g\frac{dh}{dx}d\Omega = -\int_\Omega \frac{dg}{dx}h\,d\Omega + \Big|_a^b gh} \tag{1b}$$

Functions $g(x)$ and $h(x)$ must be $C^0$ functions or smoother in $\bar{\Omega} = [a,b]$ (cf. Remark B.1).

**Remark B.1.** Following roughly the definitions in Belytschko et al. (2000, p. 27), a $C^0$ function (or shortly $C$ function) is continuous and its derivative is at least piecewisely continuous. In one dimension the derivatives of $C^0$ functions can have discontinuities or jumps in their values at separate discontinuity points, in two dimensions at separate discontinuity lines and in three dimensions at separate discontinuity surfaces. Between the discontinuities, however, we assume that the functions are smooth enough to posses as high order derivatives as we like. Similarly, a $C^{-1}$ function is itself only piecewisely continuous. A $C^{-1}$ function can have in one dimension jumps in its value at separate discontinuity points, in two dimensions at separate discontinuity lines and in three dimensions at separate discontinuity surfaces. Again, between the discontinuities, however, we assume that the functions are smooth enough to posses as high order derivatives as we like. Further, we define a $C^m$ function similarly as above so that it together with its first $m$ derivatives is continuous and the derivatives of order $m-1$ can have jumps. Finally, notation like $f \in C$ or in more detail $f \in C(\Omega)$ or $f \in C(\bar{\Omega})$ means that $f$ is a $C$ function. In the latter notations the domains of definition for $f$ are indicated. □

Formula (1) is based on the basic formula — called the fundamental theorem of calculus —

$$\int_a^b \frac{df}{dx}dx = \Big|_a^b f = f(b) - f(a) \tag{2}$$





$f(x)$ must be here a $C^0$ function or smoother. Integration of the identity (product rule of differentiation)

$$\frac{d}{dx}(gh) = \frac{dg}{dx}h + g\frac{dh}{dx} \tag{3}$$

over $\Omega$ gives

$$\int_a^b \frac{d}{dx}(gh)dx = \int_a^b \frac{dg}{dx}h dx + \int_a^b g\frac{dh}{dx}dx \tag{4}$$

Application of rule (2) for the integral on the left-hand side leads to formula (1).

## B.2 TWO DIMENSIONS

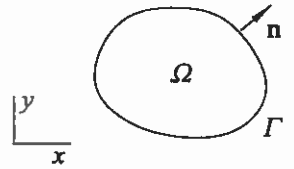



**Figure B.2** Two-dimensional domain $\Omega = A$ and its boundary $\Gamma = s$.

The two-dimensional integration by parts formulas can be expressed as (Figure B. 2)

$$\begin{aligned}\int_A g\frac{\partial h}{\partial x}dA &= -\int_A \frac{\partial g}{\partial x}h\,dA + \int_s ghn_x ds\\ \int_A g\frac{\partial h}{\partial y}dA &= -\int_A \frac{\partial g}{\partial y}h\,dA + \int_s ghn_y ds\end{aligned} \tag{1a}$$

or as

$$\boxed{\begin{aligned}\int_\Omega g\frac{\partial h}{\partial x}d\Omega &= -\int_\Omega \frac{\partial g}{\partial x}h\,d\Omega + \int_\Gamma ghn_x d\Gamma\\ \int_\Omega g\frac{\partial h}{\partial y}d\Omega &= -\int_\Omega \frac{\partial g}{\partial y}h\,d\Omega + \int_\Gamma ghn_y d\Gamma\end{aligned}} \tag{1b}$$

Functions $g(x,y)$ and $h(x,y)$ must be again $C^0$ functions or smoother.

Formula (1) is based on the so-called divergence theorem or Gauss's theorem in the plane:

$$\int_\Omega \frac{\partial f}{\partial x}\mathrm{d}\Omega = \int_\Gamma f n_x \mathrm{d}\Gamma$$
$$\int_\Omega \frac{\partial f}{\partial y}\mathrm{d}\Omega = \int_\Gamma f n_y \mathrm{d}\Gamma \qquad (2)$$

derived in mathematics texts. Integration of the identity

$$\frac{\partial}{\partial x}(g\,h) = \frac{\partial g}{\partial x}h + g\frac{\partial h}{\partial x} \qquad (3)$$

over $\Omega$ gives

$$\int_\Omega \frac{\partial}{\partial x}(g\,h)\mathrm{d}\Omega = \int_\Omega \frac{\partial g}{\partial x}h\,\mathrm{d}\Omega + \int_\Omega g\frac{\partial h}{\partial x}\mathrm{d}\Omega \qquad (4)$$

Application of the first formula (2) to the integral on the left-hand side gives the first formula (1) and the second formula can be obtained similarly.

## B.3 THREE DIMENSIONS

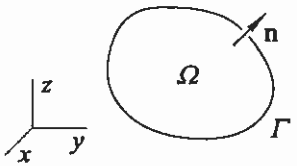



**Figure B.3** Three-dimensional domain $\Omega = V$ and its boundary $\Gamma = S$.

For shortness we employ here the index notation. The formulas are direct generalizations from the two-dimensional case. The integration by parts formulas are (Figure B.3)

$$\int_V g\frac{\partial h}{\partial x_i}\mathrm{d}V = -\int_V \frac{\partial g}{\partial x_i}h\,\mathrm{d}V + \int_S g h n_i \mathrm{d}S \qquad (1a)$$

or

$$\boxed{\int_\Omega g\frac{\partial h}{\partial x_i}\mathrm{d}\Omega = -\int_\Omega \frac{\partial g}{\partial x_i}h\,\mathrm{d}\Omega + \int_\Gamma g h n_i \mathrm{d}\Gamma} \qquad (1b)$$

The Gauss's theorem is





$$\int_\Omega \frac{\partial f}{\partial x_i}\mathrm{d}\Omega = \int_\Gamma f n_i \mathrm{d}\Gamma \qquad (2)$$

**Remark B.2.** Actually (2) represents the Gauss's formula for one component of a vector. Thus by replacing $f$ with $f_i$ and by letting the summation convention be valid we obtain

$$\int_\Omega \frac{\partial f_i}{\partial x_i}\mathrm{d}\Omega = \int_\Gamma f_i n_i \mathrm{d}\Gamma \qquad (3a)$$

or using symbolic notation

$$\int_\Omega \nabla \cdot \mathbf{f}\,\mathrm{d}\Omega = \int_\Gamma \mathbf{n}\cdot\mathbf{f}\,\mathrm{d}\Gamma \qquad (3b)$$

This form is usually called the divergence theorem. □

**Remark B.3.** In the one-dimensional case the formulas look a little bit untidy because there appears plus and minus signs in the boundary terms; see for instance formula (B.1.2). However, in this case the boundary consists of just two separate points and we can consider the unit outward normal vector $\mathbf{n}$ to have the component $+1$ at the right-hand boundary and the component $-1$ at left-hand boundary. With this interpretation we see that also the one-dimensional formulas are special cases of the general formulas (1) and (2). □

# APPENDIX C SOME CONCEPTS OF FUNCTIONAL ANALYSIS

## C.1 INTRODUCTION

*Functional analysis* (funktionaalianalyysi) is a part of mathematics dealing with "spaces". The basic idea is to consider functions as points or vectors in infinite dimensional spaces.

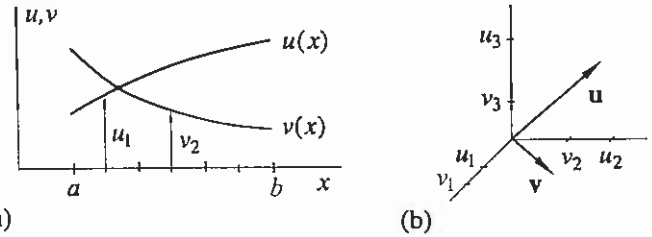



**Figure C.1(a)** Two functions. **(b)** Two vectors.

Let us consider the functions $u(x)$ and $v(x)$ in Figure C.1(a). A very rough discrete information of $u$ and $v$ is obtained, say by dividing the interval $[a,b]$ into three equal length subintervals and by measuring the function values at their midpoints. A function is represented then by three values; $u$ by $u_1$, $u_2$, $u_3$ and $v$ by $v_1$, $v_2$, $v_3$. We now associate a Cartesian coordinate for each subinterval and put the function values along these coordinate lines. Two points or vectors $\mathbf{u}=(u_1,u_2,u_3)$, $\mathbf{v}=(v_1,v_2,v_3)$ represent these functions in some crude manner in this new three-dimensional space (Figure C.1(b)). If we increase the number of subintervals and proceed similarly, we cannot any more draw a picture of the space generated, but we can still speak about the set of discrete numbers as a vector in a *finite dimensional space* (äärellisdimensioinen avaruus). Continuing without limit leads us to speak about a function as a vector in an *infinite dimensional space* (ääretöndimensioinen avaruus) or shortly *function space*. The set of discrete nodal values of a finite element simulation of a function can clearly also be given an interpretation of the type described in Figure C.1.

Many concepts in functional analysis can be considered as generalizations of familiar concepts of ordinary geometry.

In what follows we borrow heavily from Reddy (1986) and Hughes (1987).





## C.2 LINEAR SPACE

A set $V$ is a *linear space* (lineaariavaruus) if it has an operation called addition, an operation called multiplication by a scalar (real number) and it satisfies certain axioms, Reddy (1986, p. 45).

Here we do not present the axioms. The basic property is that if $u$ and $v$ are two members of a linear space $V$, that is, $u,v\in V$, then the quantity $\alpha u+\beta v$, where $\alpha,\beta\in R$, is also a member of the set, that is, $\alpha u+\beta v\in V$.

An example of linear space is the set $V=\{u:u(\mathbf{x})\in C(\Omega),\mathbf{x}\in\Omega\}$. Alternatively, this set is often denoted $C(\Omega)$ and called $C(\Omega)$ space. Similarly, if the domain is closed, we have $C(\bar{\Omega})$ space. In one dimension the corresponding spaces could be notated, say by $C]a,b[$ and by $C[a,b]$. As an example, function $1/x\in C]0,1[$ but $1/x\notin C[0,1]$ as $1/x$ is not continuous at $x=0$.

The set $V=\{u:u(\mathbf{x})\in C(\Omega),\ u(\mathbf{x})\geq 0,\mathbf{x}\in\Omega\}$ is not a linear space. For instance, $\alpha u(\mathbf{x})$ is not a member of $V$ for negative values of $\alpha$.

## C.3 INNER PRODUCT

Let $V$ be a linear space. The *inner product* or *scalar product* (sisätulo, skalaaritulo) $(u,v)$ of $u,v\in V$ is a mapping $V\times V\rightarrow R$, that is, it associates a real number with any two members of $V$, satisfying the following properties

$$\begin{aligned}&(u,v)=(v,u) && \text{(symmetry)}\\ &(\alpha u+\beta v,w)=\alpha(u,w)+\beta(v,w) && \text{(linearity)}\\ &(u,u)\geq 0 \ \text{ and } \ (u,u)=0 \ \text{ iff } \ u=0 && \text{(positive-definiteness)}\end{aligned} \tag{1}$$

for all $u,v,w\in V$ and $\alpha,\beta\in R$.

An example in $V=R^3$ satisfying these properties is the conventional dot or scalar product

$$(\mathbf{x},\mathbf{y})=\mathbf{x}\cdot\mathbf{y}=x_1y_1+x_2y_2+x_3y_3 \tag{2}$$

of two vectors $\mathbf{x}=(x_1,x_2,x_3)$ and $\mathbf{y}=(y_1,y_2,y_3)$.

A linear space $V$ on which an inner product has been defined is called an *inner product space* (sisätuloavaruus).

If

$$(u,v)=0 \tag{3}$$

the two members $u$ and $v$ of an inner product space $V$ are said to be *orthogonal* (ortogonaalinen).

For an inner product the following important result

$$\boxed{|(u,v)|\le(u,u)^{1/2}(v,v)^{1/2}} \tag{4}$$

called *the Schwarz inequality* (Schwarzin epäyhtälö), can be shown to be valid. The Schwarz inequality is a generalization of the familiar result $|\mathbf{x}\cdot\mathbf{y}|\le|\mathbf{x}||\mathbf{y}|$ from vector analysis.

## C.4 NORM

Let $V$ be a linear space. The *norm* (normi) $\|\cdot\|$ on the space $V$ is defined to be a mapping $V\to R$, that is, it associates a real number with any member of $V$ satisfying the following properties

$$\begin{aligned}
&\|u\|\ge 0 \quad\text{and}\quad \|u\|=0 \ \text{iff}\ u=0 && \text{(positive-definiteness)}\\
&\|\alpha u\|=|\alpha|\|u\| && \text{(linearity)}\\
&\|u+v\|\le\|u\|+\|v\| && \text{(triangle inequality)}
\end{aligned} \tag{1}$$

for all $u,v\in V$ and $\alpha\in R$.

A norm describes in some abstract manner the magnitude of a function. It is a generalization of the concept "length of a vector" familiar from vector analysis. Namely, let $V=R^3$, then

$$\|\mathbf{x}\|=(x_1^2+x_2^2+x_3^2)^{1/2}=|\mathbf{x}| \tag{2}$$

for $\mathbf{x}=(x_1,x_2,x_3)$.

The above interpretation explains why $\|u\text{-}v\|$ is often called the distance between $u$ and $v$.



A linear space $V$ on which a norm has been defined is called a *normed space* (normiavaruus).

It should be noticed that an inner product $(u,u)$ generates automatically a norm $\|u\|\equiv(u,u)^{1/2}$. This can be checked by applying formulas (C.3.1) and the Schwarz inequality to see that conditions (1) are satisfied. In this case the Schwartz inequality can be written alternatively as

$$|(u,v)|\le\|u\|\|v\| \tag{3}$$

One usual measure is the so-called $L_2$*-norm*

$$\|u\|_{L_2}\equiv\left(\int_\Omega u^2\,d\Omega\right)^{1/2} \tag{4}$$

Function with the right-hand side of (4) finite is called *square integrable* (neliöintegroituva). The set of square integrable functions is said to form $L_2(\Omega)$ space.

A convenient problem dependent measure for certain problem class is the so-called *energy norm* (energianormi) which in one-dimensional heat conduction is simply

$$\|u\|_a\equiv\left[\int_\Omega k\left(\frac{du}{dx}\right)^2 dx\right]^{1/2} \tag{5}$$

It should be noted that a norm can be taken of a quantity even when it is not a member of a linear space; "the norm does not see the difference".

In the mathematical analysis of the finite element method, norms consisting of integrals over the domain under study like (4) or (5) are in common use. This is understandable because error analyses can start in a natural way from weak forms which themselves consist of integrals.

The *seminorm* (seminormi) $|\cdot|$ is defined similarly as the norm; the only difference being that the first condition of (1) is replaced with

$$|u|\ge 0 \qquad \text{(positive-semidefiniteness)} \tag{6}$$

The seminorm symbol is usually equipped with some subscript to discern it from the absolute value symbol.

An example:

$$|u|_s^2 \equiv \sum_{i+j\cdots=s} \int_\Omega \left( \frac{\partial^s u}{\partial x^i \partial y^j \cdots} \right)^2 \mathrm{d}x\mathrm{d}y\cdots \tag{7}$$

In fact, (5) defines a seminorm unless $u$ is demanded to satisfy a homogeneous Dirichlet condition. (Otherwise a member $u(x) = \text{constant} \neq 0$ gives the zero value.)

## C.5 LINEAR FORM AND BILINEAR FORM

*Linear form* (lineaarimuoto) $b(u)$ of $u \in V$ ($V$ linear space) is a mapping $V \to R$, that is, it associates a real number with any member of $V$ with the property

$$b(\alpha u + \beta v) = \alpha b(u) + \beta b(v) \tag{1}$$

for all $u, v \in V$ and $\alpha, \beta \in R$.

An example. Let $V = C[a,b]$. The mapping

$$b(u) = \int_a^b us\,\mathrm{d}x \tag{2}$$

where $s \in L_2[a,b]$ is given, is a linear form.

*Bilinear form* (bilineaarimuoto) $a(u,v)$ of $u \in U$ and $v \in V$ ($U$ and $V$ linear spaces) is a mapping $U \times V \to R$, that is, it associates a real number with any pair of members of $U$ and $V$, satisfying the following properties

$$\begin{aligned} &a(\alpha u + \beta w, v) = \alpha a(u,v) + \beta a(w,v) \quad u,w \in U,\ v \in V,\ \alpha,\beta \in R \\ &a(u, \alpha v + \beta w) = \alpha a(u,v) + \beta a(u,w) \quad u \in U,\ v,w \in V,\ \alpha,\beta \in R \end{aligned} \tag{3}$$

Often in applications $U = V$.

An example. Let $U = V = C^1[a,b]$. The mapping

$$a(u,v) \equiv \int_a^b \left( uv + \frac{\mathrm{d}u}{\mathrm{d}x}\frac{\mathrm{d}v}{\mathrm{d}x} \right) \mathrm{d}x \tag{4}$$

is a bilinear form. It is also an inner product.

It should be noticed that expressions like (4) appearing in mathematics text make usually no sense in physics unless a dimensionless formulation is used or unless we define instead of (4) say



$$a(u,v) \equiv \int_a^b \left( uv + c_1 \frac{\mathrm{d}u}{\mathrm{d}x}\frac{\mathrm{d}v}{\mathrm{d}x} \right) \mathrm{d}x \tag{5}$$

where $c_1$ is a positive constant making the expression dimensionally homogeneous.

Finally, again, linear and bilinear forms can be generated from quantities that are not members of linear spaces.

## REFERENCES

Hughes, T. J. R. (1987). *The Finite Element Method — Linear Static and Dynamic Finite Element Analysis*, Prentice-Hall, Englewood Cliffs, New Jersey, ISBN 0-13-317017-9.

Reddy, B. D. (1986). *Functional Analysis and Boundary Value Problems: an Introductory Treatment*, Longman, New York, 0-582-98826-8.

# APPENDIX D VARIATIONAL CALCULUS

As mentioned in Chapter 1, the earliest applications of the finite element method were based on the variational formulation which uses a variational principle (see Remark 2.3), i.e., the stationarity of a functional is the starting point for discretization. As the residual formulation is more general and includes the variational formulation as a special case, we strictly do not need to discuss the latter formulation and the concept of a functional at all. However, sensitizing principles can be introduced (see Chapter 5) rather understandably by making some use of functionals and variational calculus. In addition, a mild knowledge of variational calculus is useful for following the finite element literature in general.

## D.1 FUNCTIONAL

*Functional* (funktionaali) is an operation associating a real number $\Pi$ for each member $\phi$ of a set $S$ the members consisting of functions of agreed kind, that is, it is a mapping $\Pi : S \rightarrow R$.

Usually the mapping is effected via a definite integral. An example:

$$\Pi(\phi) = \int_a^b \sqrt{1 + \left(\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)^2}\,\mathrm{d}x \tag{1}$$

Here $S$ is the set of functions $\phi(x)$ defined on the $x$-axis interval $[a,b]$. The functions must satisfy the conditions $\phi(a) = \alpha$ and $\phi(b) = \beta$ were $\alpha$ and $\beta$ are given and to be so smooth that the integral can be evaluated.

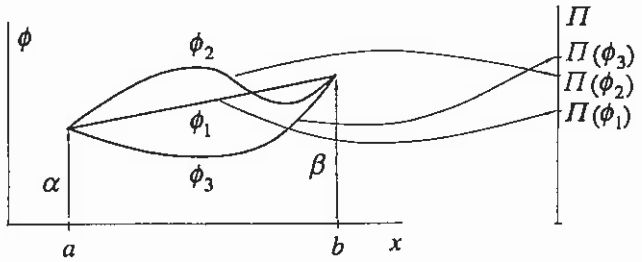



**Figure D.1** Mappings of three functions.

The integral clearly represents the length of a curve passing through the points $(a,\alpha)$ and $(b,\beta)$. Figure D.1 shows schematically the mapping of three functions $\phi(x)$ to real numbers $\Pi$. Because of the geometric interpretation, it is clear that $\Pi$ obtains the minimum value as $\phi(x)$ describes a straight line.





Functional (1) is a special case of a more general situation

$$\Pi(\phi) = \int_a^b f\left(x, \phi, \frac{\mathrm{d}\phi}{\mathrm{d}x}\right)\mathrm{d}x \tag{2}$$

where the integrand depends on the independent variable $x$, on the *argument function* (argumenttifunktio) $\phi(x)$ and on its derivative $\mathrm{d}\phi/\mathrm{d}x$. In the example case (1) $x$ and $\phi$ are missing.

A functional can contain higher order derivatives than the first, several argument functions and several independent variables and terms from the boundary of the domain.

When we pick a certain argument function, its derivatives can be evaluated and it can be fed in the functional expression to give as the output a certain number. The main task of *variational calculus* (variaatiolaskenta) is to determine that argument function giving the functional an extremal value. The necessary condition for this is that the functional obtains a *stationary value* (stationaarinen arvo). This means that for "small" changes of the argument function the changes of the functional are zero. The corresponding argument function is called the *stationary function* (stationaarinen funktio).

From the stationarity condition so-called *Euler differential equation(s)* or Euler-Lagrange differential equation(s) (Eulerin differentiaaliyhtälö(t)) with their boundary conditions can be derived.

If we are able to find a functional, for which the stationarity condition gives the differential equations and boundary conditions we want to solve, we have obtained a convenient way to perform the discretization; it can be based on the functional.

For any differential equation set a corresponding functional unfortunately does not exist. A notably example are the Navier-Stokes momentum equations. This fact constraints the usefulness of the variational formulation.

To study the changes of functionals and functions when the argument function experiences changes we need a way of thinking different from conventional differential calculus. Certain notations and calculation rules necessary to deal with functionals are considered next. In the following we borrow strongly from the fine text by Lanczos (1970).

## D.2 VARIATIONAL NOTATION

We again consider just the case of one independent variable $x$ and one argument function $\phi(x)$ (Figure D.2). The function can experience two kinds of changes.

The infinitesimal change $d\phi$ is due to the infinitesimal chance $dx$ of $x$. The infinitesimal change $\delta\phi(x)$, however, is effected by the change of moving from the curve $\phi(x)$ to an infinitesimally near neighboring curve $\phi^*(x) = \phi(x) + \delta\phi(x)$. In variational calculus this latter type of change is considered. The *variational symbol* (variaatiomerkki) $\delta$ is customarily used instead of the symbol d to tell the difference. The quantity $\delta\phi$ is called the *variation* (variaatio) — in more detail the first variation — of function $\phi$ and the new function $\phi^*$ is called the *varied function* or modified function or comparison function (varioitu funktio).

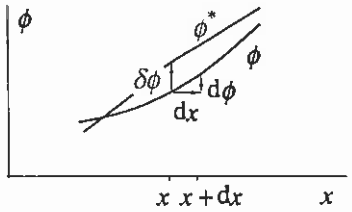



**Figure D.2** Differential and variation.

If $\phi(x)$ is the actual function representing some quantity, the generation of a varied function means that some kind of mathematical thought experiment is performed to obtain comparison results: what would be the outcome if instead of ⋯ ? In mechanics the most common example of the variation of a function is probably the concept of virtual displacement. It is usually defined to be an infinitely small imagined displacement, which is thought to take place with time "freezed". This definition is seen to equivalent to the concept of the variation of a function. ($x$ means now the time and $\phi$ is one space coordinate of a particle. Freezing time means that we move in vertical and not in horizontal direction in Figure D.2.)

The expressions containing the argument function and the functional in particular obtain changes due the variation of the argument function. These changes are also called variations and the $\delta$-symbol is again used.

**Table D.1** Some rules of variational calculus

| | |
|---|---|
| $\delta(f_1 + f_2) = \delta f_1 + \delta f_2$ | Variation of sum |
| $\delta(kf) = k\delta f$ | Transfer rule for a constant |
| $\delta(f_1 \cdot f_2) = \delta f_1 \cdot f_2 + f_1 \cdot \delta f_2$ | Variation of product |



| | |
|---|---|
| $\delta(f^n) = nf^{n-1}\delta f$ | Variation of power function |
| $\delta(d\phi/dx) = d(\delta\phi)/dx$ | Variation of derivative |
| $\delta\int f\,dx = \int \delta f\,dx$ | Variation of definite integral |

Table D.1 contains some calculation rules of variational calculus. These rules are valid also in the case of several independent variables. The formulas are quite analogous to the corresponding differentiation expressions.and they are not especially difficult to derive.

The stationarity condition of a functional $\Pi$ is represented in the form

$$\boxed{\delta\Pi = 0} \tag{1}$$

or the variation of the functional must be zero with respect to arbitrary admissible variation of the argument function(s). The content of condition (1) is called *variational principle* (variaatioperiaate). Perhaps the most well-known variational principle of mechanics is the principle of stationary (or minimum) potential energy: when the potential energy of a conservative system obtains a stationary value, the corresponding configuration of the system is the equilibrium position.

The argument functions competing in a functional must obey some smoothness conditions so that the functional can be evaluated and in general some boundary conditions. In this way defined argument functions are called *admissible functions* (luvallinen funktio). The boundary conditions demanded to be satisfied in advance from the admissible functions are called *essential boundary conditions* (oleellinen reunaehto). The stationarity condition gives as consequences the Euler differential equations and the so-called *natural* or free or additional *boundary conditions* (luonnollinen reunaehto). (The concepts of an admissible function, essential and natural boundary conditions are used in a similar meaning also in connection with weak forms; see Section 4.2.2.) These features are described concisely in the following with some simple applications.

## D.3 HEAT CONDUCTION

### D.3.1 One dimension

Let us consider the functional

$$\boxed{\Pi(T)=\int_a^b\left[\frac{1}{2}k\left(\frac{\mathrm{d}T}{\mathrm{d}x}\right)^2-sT\right]\mathrm{d}x+\overline{q}T\big|_{x=b}} \tag{1}$$

in which the admissible argument function $T(x)$ must satisfy the essential boundary condition

$$\boxed{T=\overline{T}} \quad \text{on } \Gamma_{\mathrm{D}}=\{a\} \tag{2}$$

The meaning of the notations is the same as in Section 2.1.1. We state the variational principle

$$\delta\Pi=0 \tag{3}$$

and find out what follows from it.

Variation gives first

$$\delta\Pi=\int_a^b\left[\frac{1}{2}k\delta\left(\frac{\mathrm{d}T}{\mathrm{d}x}\right)^2-s\delta T\right]\mathrm{d}x+\overline{q}\delta T\big|_{x=b} \tag{4}$$

Rules (6), (1) and (2) of Table D.1 have been employed. It should be noted that the term "constant" means here that the quantity in question does not depend on the argument function. By further applying rules (4) and (5) we get

$$\delta\left(\frac{\mathrm{d}T}{\mathrm{d}x}\right)^2=2\frac{\mathrm{d}T}{\mathrm{d}x}\delta\frac{\mathrm{d}T}{\mathrm{d}x}=2\frac{\mathrm{d}T}{\mathrm{d}x}\frac{\mathrm{d}\delta T}{\mathrm{d}x} \tag{5}$$

and

$$\delta\Pi=\int_a^b\left[k\frac{\mathrm{d}T}{\mathrm{d}x}\frac{\mathrm{d}\delta T}{\mathrm{d}x}-s\delta T\right]\mathrm{d}x+\overline{q}\delta T\big|_{x=b} \tag{6}$$

The next step is based on the fact that variation $\delta T$ is arbitrary. However, no conclusions can be yet drawn from expression (6) as it contains in addition to $\delta T$ its derivative. The derivative must first be removed by applying integration by parts. Similarly as in generating weak forms, integration by parts manipulation is always needed in variational calculus to produce the Euler equations. Formula (B.1.1) with $g\triangleq k\,\mathrm{d}T/\mathrm{d}x$ and $h\triangleq\delta T$ gives

$$\int_a^b k\frac{\mathrm{d}T}{\mathrm{d}x}\frac{\mathrm{d}\delta T}{\mathrm{d}x}\mathrm{d}x=-\int_a^b\frac{\mathrm{d}}{\mathrm{d}x}\left(k\frac{\mathrm{d}T}{\mathrm{d}x}\right)\delta T\,\mathrm{d}x+\left|_a^b k\frac{\mathrm{d}T}{\mathrm{d}x}\delta T\right. \tag{7}$$



The variation of the functional looks now

$$\delta\Pi=\int_a^b\left[\frac{\mathrm{d}}{\mathrm{d}x}\left(-k\frac{\mathrm{d}T}{\mathrm{d}x}\right)-s\right]\delta T\mathrm{d}x+\left(k\frac{\mathrm{d}T}{\mathrm{d}x}+\overline{q}\right)\delta T\big|_{x=b}-k\frac{\mathrm{d}T}{\mathrm{d}x}\delta T\Big|_{x=a} \tag{8}$$

Because the admissible $T$ must satisfy the essential boundary condition (2), the variation $\delta T=T^*-T$ must vanish on the Dirichlet boundary $x=a$ or

$$\delta T\big|_{x=a}=0 \tag{9}$$

Thus we are left with the condition

$$\int_a^b\left[\frac{\mathrm{d}}{\mathrm{d}x}\left(-k\frac{\mathrm{d}T}{\mathrm{d}x}\right)-s\right]\delta T\mathrm{d}x+\left(k\frac{\mathrm{d}T}{\mathrm{d}x}+\overline{q}\right)\delta T\big|_{x=b}=0 \tag{10}$$

This gives the Euler differential equation

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}x}\left(-k\frac{\mathrm{d}T}{\mathrm{d}x}\right)-s=0} \quad \text{in} \quad \Omega=]a,b[ \tag{11}$$

and the natural boundary condition

$$\boxed{k\frac{\mathrm{d}T}{\mathrm{d}x}+\overline{q}=0} \quad \text{on} \quad \Gamma_{\mathrm{N}}=\{b\} \tag{12}$$

Thus the variational principle (3) is equivalent to the problem posed by the differential equation (11) and the boundary conditions (2) and (12). These are the governing equations of Section 2.1.1 describing one-dimensional heat conduction.

Let us return in more detail on the logic of obtaining the Euler equation and the natural boundary condition from the stationarity condition (10). Use is made of the so-called *fundamental lemma of variational calculus* (variaatiolaskennan peruslemma):

If the relation

$$\int_a^b f(x)\eta(x)\mathrm{d}x=0 \tag{13}$$

where $f(x)$ is a continuous function is valid for all continuous functions $\eta(x)$,

$$f(x)=0 \quad \text{in } \Omega=]a,b[ \tag{14}$$

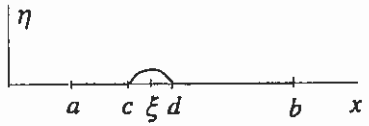



**Figure D.3** Bubble function.

This is proved roughly as follows. Let us consider a point $x=\xi$ in the open interval $]a,b[$ and assume that contrary to what has been stated, $f$ is there non-zero and say positive. Because of the continuity of $f$ there exists a neighborhood $c<x<d$ of $\xi$ where $f$ is positive. We now select $\eta$ as a "bubble" (Figure D.3) say of the type

$$\eta=\begin{cases}(x-c)(d-x), & c<x<d\\ 0 \quad \text{elsewhere}\end{cases} \tag{15}$$

The first expression in (15) is then positive and thus

$$\int_a^b f(x)\eta(x)\mathrm{d}x=\int_c^d f(x)\eta(x)\mathrm{d}x>0 \tag{16}$$

which is against (13).

In our application the variation $\delta T$ has the role of $\eta$. In the weak forms treated earlier the weighting function $w$ has had the role of $\eta$. The lemma can be extended in an obvious way to cases with several independent variables.

Thus making use of the arbitrariness of $\delta T$ in (10) we obtain the result (11) and (10) is simplified into the form

$$\left(k\frac{\mathrm{d}T}{\mathrm{d}x}+\bar{q}\right)\delta T\Big|_{x=b}=0 \tag{17}$$

As $\delta T$ at $x=b$ can taken arbitrarily, there finally follows (12).

The discretization by the finite element method based on the variational formulation proceeds as follows. When the approximation

$$\tilde{T}(x)=\sum_j N_j(x)T_j \tag{18}$$



is substituted into the functional $\Pi(T)$ and the integration with respect to $x$ has been thought to be performed — it is not yet actually necessary to do this — it becomes and ordinary function of the nodal parameters $T_1$, $T_2$, $\cdots$ or shortly $\tilde{\Pi}(\{a\})$. Instead of the condition $\delta\Pi=0$ we now demand this function to have a stationary value with respect to the nodal parameters $\{a\}$. System equations are thus obtained as ordinary stationarity conditions (see (2.1.41))

$$\boxed{F_i\equiv\frac{\partial\tilde{\Pi}}{\partial T_i}=0} \qquad i=1,2,\cdots \tag{19}$$

This is considered in more detail in Example D.1.

**Remark D.1.** In the finite element method based on the variational formulation, it is enough to select a reasonable approximation such as (18). Given the functional, the discrete equations (19) follow automatically "by turning the handle". Nothing more is demanded from the applier. This is convenient but it also means that the wider possibilities of the residual formulation where one can make selections in addition with respect to the approximation also with respect to the weighting are lost. In fact the weighting is produced automatically by the trial basis functions and this means that the formulation can be interpreted as the use of the Galerkin method. The Petrov-Galerkin type method possibility (see Remark 6.10) is not available. □

**Remark D.2.** Comparison of say equation (10) for example with equation (2.1.6) shows that we have in fact here a *weak formulation corresponding to the functional* and that *the variation of the argument function has the role of the weighting function*:

$$\delta T=w \tag{20}$$

This interpretation has been made use in Chapter 5 and it will be used later also in this appendix. It is quite obvious that by manipulating any correct (meaning that it corresponds to a differential equation system with certain boundary conditions = strong form) functional by taking the variation we can finally arrive at an expression like (10). Now this result can alternatively be generated also from the strong form as a weak formulaton. This explains that the variational formulation cannot produce anything, which is not obtainable from a suitably selected weak formulation. □

**Remark D.3.** One may wonder how we can make interpretation (20) as the quantity on the left-hand side was defined to be of infinitesimal size and we have not stated this kind of restriction for the quantity on the right-hand side. It is true that the variation must be infinitesimal for formulas like rule (4) in Table D.1 to be valid. But after the variational manipulations have been performed we always end up with expressions like (6) or (10) where *the variation appears linearly* (in the first power; also the possible derivatives). The same concerns the weighting function: *the weighting function always appears linearly in a weak form*. This feature is fundamental in connection with weak forms; if the governing differential equation happens to be non-linear, this does not change the situation. After obtaining an equation like (10), we can scale the variation by an arbitrary multiplier without changing the conclusions to be drawn (Euler equations and natural boundary conditions). So at this phase we can even consider the multiplier to be unbounded and thus we can equally well consider the variation to be here a finite quantity (or the weighting function an infinitesimal quantity).

This line of thought is seen to be valid also for finite dimensional weighting used to obtain the system equations: we usually say that in the Galerkin method we take $\bar{w}$ to be the shape function $N_i$. However, an equivalent equation is seen to be obtained by using the weighting function $kN_i$ where $k$ is an arbitrary constant. □

**Example D.1.** We derive the discrete finite element system equations corresponding to functional (1):

$$\Pi(T)=\int_a^b\left[\frac{1}{2}k\left(\frac{\mathrm{d}T}{\mathrm{d}x}\right)^2-sT\right]\mathrm{d}x+\bar{q}T\Big|_{x=b} \tag{a}$$

Substitution of the finite element approximation (18):

$$\tilde{T}(x)=\sum_j N_j(x)T_j \tag{b}$$

transforms the functional to a function

$$\begin{aligned}\tilde{\Pi}([a]) &= \int_a^b\left[\frac{1}{2}k\left(\frac{\mathrm{d}\tilde{T}}{\mathrm{d}x}\right)^2-s\tilde{T}\right]\mathrm{d}x+\bar{q}\tilde{T}\Big|_{x=b}\\ &= \int_a^b\left[\frac{1}{2}k\left(\sum_j\frac{\mathrm{d}N_j}{\mathrm{d}x}T_j\right)^2-s\sum_j N_jT_j\right]\mathrm{d}x+\bar{q}\sum_j N_jT_j\Big|_{x=b}\end{aligned} \tag{c}$$

A typical system equation is obtained by differentiation this with respect to a nodal parameter $a_i=T_i$ and by setting the result equal to zero. Differentation gives first

$$\begin{aligned}F_i\equiv\frac{\partial\tilde{\Pi}}{\partial T_i} &= \int_a^b\left[\frac{1}{2}k\,2\frac{\mathrm{d}\tilde{T}}{\mathrm{d}x}\frac{\partial}{\partial T_i}\frac{\mathrm{d}\tilde{T}}{\mathrm{d}x}-s\frac{\partial\tilde{T}}{\partial T_i}\right]\mathrm{d}x+\bar{q}\frac{\partial\tilde{T}}{\partial T_i}\Big|_{x=b}\\ &= \int_a^b\left(k\frac{\mathrm{d}\tilde{T}}{\mathrm{d}x}\frac{\mathrm{d}N_i}{\mathrm{d}x}-sN_i\right)\mathrm{d}x+\bar{q}N_i\Big|_{x=b}\end{aligned} \tag{d}$$

Differentiation has been brought inside the integral and use have been made of the formulas

$$\frac{\partial\tilde{T}}{\partial T_i}=N_i,\qquad\frac{\partial}{\partial T_i}\frac{\mathrm{d}\tilde{T}}{\mathrm{d}x}=\frac{\mathrm{d}N_i}{\mathrm{d}x} \tag{e}$$

obtainable from (b). The system equations are thus

$$F_i\equiv\int_a^b\frac{\mathrm{d}N_i}{\mathrm{d}x}k\frac{\mathrm{d}\tilde{T}}{\mathrm{d}x}\mathrm{d}x-\int_a^b N_i s\mathrm{d}x+N_i\bar{q}\Big|_{x=b}=0\quad i=1,2,\cdots,n_n \tag{f}$$

These are exactly the same obtained by the Galerkin method (see (2.3.5)) and thus there is no need to develop them again in more detail.

### D.3.2 Two dimensions



Let us consider the functional

$$\begin{aligned}\Pi(T)=&\frac{1}{2}\int_\Omega\begin{Bmatrix}\partial T/\partial x\\ \partial T/\partial y\end{Bmatrix}^{\mathrm{T}}\begin{bmatrix}k_{xx}&k_{xy}\\ k_{yx}&k_{yy}\end{bmatrix}\begin{Bmatrix}\partial T/\partial x\\ \partial T/\partial y\end{Bmatrix}\mathrm{d}\Omega+\frac{1}{2}\int_\Omega cT^2\mathrm{d}\Omega\\ &+\frac{1}{2}\int_{\Gamma_{\mathrm{R}}}hT^2\mathrm{d}\Gamma-\int_\Omega sT\mathrm{d}\Omega+\int_{\Gamma_{\mathrm{N}}}\bar{q}T\mathrm{d}\Gamma-\int_{\Gamma_{\mathrm{R}}}hT_\infty T\mathrm{d}\Gamma\end{aligned} \tag{21}$$

in which the admissible argument function $T(x,y)$ must satisfy the essential boundary condition

$$T=\bar{T}\quad\text{on}\quad\Gamma_{\mathrm{D}} \tag{22}$$

This is roughly the most general formulation in connection with linear heat conduction we can devise. The notation is the same as in Chapter 3.

**Remark D.4.** In a *quadratic form* (neliömuoto) like

$$\begin{aligned}\begin{Bmatrix}\partial T/\partial x\\ \partial T/\partial y\end{Bmatrix}^{\mathrm{T}}\begin{bmatrix}k_{xx}&k_{xy}\\ k_{yx}&k_{yy}\end{bmatrix}\begin{Bmatrix}\partial T/\partial x\\ \partial T/\partial y\end{Bmatrix}=&k_{xx}\frac{\partial T}{\partial x}\frac{\partial T}{\partial x}+k_{xy}\frac{\partial T}{\partial x}\frac{\partial T}{\partial y}+\\ &+k_{yx}\frac{\partial T}{\partial y}\frac{\partial T}{\partial x}+k_{yy}\frac{\partial T}{\partial y}\frac{\partial T}{\partial y}\end{aligned} \tag{23}$$

the symmetry relation $k_{yx}=k_{xy}$ can be introduced without loss of generality. If originally $k_{yx}\neq k_{xy}$, we can put

$$k_{xy}^{\mathrm{new}}=k_{yx}^{\mathrm{new}}=(k_{xy}^{\mathrm{old}}+k_{yx}^{\mathrm{old}})/2 \tag{24}$$

where the meaning of the notations is obvious. The value of the quadratic form is seen to remain unchanged. □

The variational principle

$$\delta\Pi=0 \tag{25}$$

is found to give the Euler equation

$$\frac{\partial}{\partial x}\left(-k_{xx}\frac{\partial T}{\partial x}-k_{xy}\frac{\partial T}{\partial y}\right)+\frac{\partial}{\partial y}\left(-k_{yx}\frac{\partial T}{\partial x}-k_{yy}\frac{\partial T}{\partial y}\right)+cT-s=0\quad\text{in}\quad\Omega \tag{26}$$

and the natural boundary conditions

$$n_x\left(k_{xx}\frac{\partial T}{\partial x}+k_{xy}\frac{\partial T}{\partial y}\right)+n_y\left(k_{yx}\frac{\partial T}{\partial x}+k_{yy}\frac{\partial T}{\partial y}\right)+\bar{q}=0\quad\text{on}\quad\Gamma_{\mathrm{N}} \tag{27}$$

$$n_x\left(k_{xx}\frac{\partial T}{\partial x}+k_{xy}\frac{\partial T}{\partial y}\right)+n_y\left(k_{yx}\frac{\partial T}{\partial x}+k_{yy}\frac{\partial T}{\partial y}\right)+h(T-T_\infty)=0 \quad \text{on } \Gamma_R \tag{28}$$

Equations (26), (22), (27) and (28) describe correctly heat conduction in an anisotropic medium with the reaction term included. However, *convection cannot be introduced via the variational formulation*. The reader may experiment say by adding a superficially promising looking term like $Tu\partial T/\partial x$ in the domain integrand to see that nothing useful is achieved. Literature contains rules telling when a corresponding variational principle exists for a problem described by a differential equation formulation. For linear problems the system must be *self-adjoint* (itseadjungoitu), Crandall (1956, p. 210). See also Section D.4.1.

## D.4 LEAST SQUARES FUNCTIONAL, D-C-R EQUATION

### D.4.1 One dimension

We will consider the one-dimensional steady D-C-R-equation (see Appendix A) in the form

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}x}\left(-D\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)+\frac{\mathrm{d}}{\mathrm{d}x}(u\phi)+c\phi-f=0} \quad \text{in } \Omega=]a,b[ \tag{1}$$

with the Dirichlet boundary conditions

$$\begin{aligned}&\boxed{\phi=\bar{\phi}_a} \quad \text{at } x=a\\ &\boxed{\phi=\bar{\phi}_b} \quad \text{at } x=b\end{aligned} \tag{2}$$

We form a least squares functional similarly as explained in Section 2.1.2. To shorten the expression we define

$$R(\phi)\equiv L(\phi)-f\equiv\frac{\mathrm{d}}{\mathrm{d}x}\left(-D\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)+\frac{\mathrm{d}}{\mathrm{d}x}(u\phi)+c\phi-f=0 \tag{3}$$

The least squares functional is (cf. (2.1.39))

$$\boxed{\Pi(\phi)=\frac{1}{2}\int_\Omega R^2\mathrm{d}\Omega} \tag{4}$$

We consider boundary conditions (2) as essential so no contribution from them is needed in (4). Similarly, no weight factor is needed as there is now only one term in the least squares expression (see Remark 2.6).



We will derive the Euler equation and the natural boundary conditions due to the variational principle $\delta\Pi=0$. The variation is with the help of Table D.1

$$\delta\Pi=\frac{1}{2}\int_\Omega 2R\delta R\mathrm{d}\Omega=\int_\Omega R\delta R\mathrm{d}\Omega \tag{5}$$

Further

$$\delta R=\frac{\mathrm{d}}{\mathrm{d}x}\left(-D\frac{\mathrm{d}\delta\phi}{\mathrm{d}x}\right)+\frac{\mathrm{d}}{\mathrm{d}x}(u\delta\phi)+c\delta\phi=L(\delta\phi) \tag{6}$$

and thus

$$\delta\Pi=\int_\Omega RL(\delta\phi)\mathrm{d}\Omega \tag{7}$$

To deduce the Euler equation we have to integrate by parts. First,

$$\begin{aligned}&\int_\Omega R\frac{\mathrm{d}}{\mathrm{d}x}\left(-D\frac{\mathrm{d}\delta\phi}{\mathrm{d}x}\right)\mathrm{d}\Omega=\int_\Omega\frac{\mathrm{d}R}{\mathrm{d}x}D\frac{\mathrm{d}\delta\phi}{\mathrm{d}x}\mathrm{d}\Omega-\Big|_a^b RD\frac{\mathrm{d}\delta\phi}{\mathrm{d}x}\\ &\int_\Omega R\frac{\mathrm{d}}{\mathrm{d}x}(u\delta\phi)\mathrm{d}\Omega=-\int_\Omega\frac{\mathrm{d}R}{\mathrm{d}x}u\delta\phi\,\mathrm{d}\Omega+\Big|_a^b Ru\delta\phi\end{aligned} \tag{8}$$

Second,

$$\int_\Omega\frac{\mathrm{d}R}{\mathrm{d}x}D\frac{\mathrm{d}\delta\phi}{\mathrm{d}x}\mathrm{d}\Omega=\int_\Omega\frac{\mathrm{d}}{\mathrm{d}x}\left(-D\frac{\mathrm{d}R}{\mathrm{d}x}\right)\delta\phi\,\mathrm{d}\Omega+\Big|_a^b D\frac{\mathrm{dR}}{\mathrm{d}x}\delta\phi \tag{9}$$

Collecting all terms gives

$$\begin{aligned}\delta\Pi=&\int_\Omega\left[\frac{\mathrm{d}}{\mathrm{d}x}\left(-D\frac{\mathrm{d}R}{\mathrm{d}x}\right)-u\frac{\mathrm{d}R}{\mathrm{d}x}+cR\right]\delta\phi\,\mathrm{d}\Omega\\ &+\Big|_a^b\left[-RD\frac{\mathrm{d}\delta\phi}{\mathrm{d}x}+D\frac{\mathrm{dR}}{\mathrm{d}x}\delta\phi+Ru\delta\phi\right]\end{aligned} \tag{10}$$

We denote

$$L^*(\phi)=\frac{\mathrm{d}}{\mathrm{d}x}\left(-D\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)-u\frac{\mathrm{d}\phi}{\mathrm{d}x}+c\phi \tag{11}$$

$L^*$ is called the *adjoint operator* (adjungoitu operaattori) of $L$. (If $L^*=L$, the operator $L$ is called self-adjoint. Here convection makes the operator non-self-adjoint.)

As $\delta\phi = 0$ at $x = a$ and $x = b$ due to the essential boundary conditions (2), the stationarity condition has obtained the form

$$\int_\Omega L^*(R)\delta\phi \, \mathrm{d}\Omega - \Big|_a^b DR\frac{\mathrm{d}\delta\phi}{\mathrm{d}x} = 0 \tag{12}$$

This gives the Euler equation

$$L^*(R) = L^*\big(L(\phi) - f\big) = 0 \quad \text{in } \Omega \tag{13}$$

and the natural boundary condition

$$R = L(\phi) - f = 0 \quad \text{on } \Gamma = \{a, b\} \tag{14}$$

It is thus seen — as mentioned in Remark 2.7 — that the least squares method produces equations, which are not directly those of the differential equation formulation although the exact solution $\phi(x)$ is clearly seen to satisfy the set generated.

For our purposes form (7) is the most useful. If we make the interpretation $\delta\phi = w$ we obtain

$$\boxed{\int_\Omega L(w)R \, \mathrm{d}\Omega = 0} \tag{15}$$

which we will call the *least squares weak form* (pienimmän neliön heikko muoto). This is in detail

$$\int_\Omega \left[\frac{\mathrm{d}}{\mathrm{d}x}\left(-D\frac{\mathrm{d}w}{\mathrm{d}x}\right) + \underline{\frac{\mathrm{d}}{\mathrm{d}x}(uw)} + cw\right] \cdot \left[\frac{\mathrm{d}}{\mathrm{d}x}\left(-D\frac{\mathrm{d}\phi}{\mathrm{d}x}\right) + \underline{\frac{\mathrm{d}}{\mathrm{d}x}(u\phi)} + c\phi - f\right]\mathrm{d}\Omega = 0 \tag{16}$$

The underlined terms are essential in sensitizing convection dominated flows. A preliminary explanation is shortly as follows. Let us assume for simplicity that $u$ is constant. The underlined terms multiplied together produce then the term

$$\frac{\mathrm{d}w}{\mathrm{d}x} u^2 \frac{\mathrm{d}\phi}{\mathrm{d}x} \tag{17}$$

in the integrand. This is mathematically of the same form as the integrand term

$$\frac{\mathrm{d}w}{\mathrm{d}x} k \frac{\mathrm{d}T}{\mathrm{d}x} \tag{18}$$



in the pure diffusion problem weak form (see for instance (2.1.28)). Quantity $u^2$ can thus be interpreted as a positive diffusivity (or in particular as a thermal conductivity). Qualitatively speaking we can through the least squares weak form inject diffusion into the formulation which can be employed to damp the oscillations found in the standard weak formulation.

### D.4.2 Two dimensions

We consider the two-dimensional steady D-C-R equation (see Appendix A) in the form

$$R(\phi) \equiv L(\phi) - f \equiv$$

$$\equiv \boxed{\frac{\partial}{\partial x}\left(-D\frac{\partial\phi}{\partial x}\right) + \frac{\partial}{\partial y}\left(-D\frac{\partial\phi}{\partial y}\right) + \frac{\partial}{\partial x}(u\phi) + \frac{\partial}{\partial y}(v\phi) + c\phi - f = 0} \quad \text{in } \Omega \tag{19}$$

Isotropic diffusivity has been taken as the idea can be seen already in this case. No attention is paid on the boundary conditions as they are not present in the sensitizing terms.

The least squares functional

$$\Pi(\phi) = \frac{1}{2}\int_\Omega R^2 \mathrm{d}\Omega \tag{20}$$

leads again to the least squares weak form (see equations (4) and (15))

$$\int_\Omega L(w)R \, \mathrm{d}\Omega = 0 \tag{21}$$

This is in detail

$$\int_\Omega \left[\frac{\partial}{\partial x}\left(-D\frac{\partial w}{\partial x}\right) + \frac{\partial}{\partial y}\left(-D\frac{\partial w}{\partial y}\right) + \underline{\frac{\partial}{\partial x}(uw)} + \underline{\frac{\partial}{\partial y}(vw)} + cw\right] \cdot \left[\frac{\partial}{\partial x}\left(-D\frac{\partial\phi}{\partial x}\right) + \frac{\partial}{\partial y}\left(-D\frac{\partial\phi}{\partial y}\right) + \underline{\frac{\partial}{\partial x}(u\phi)} + \underline{\frac{\partial}{\partial y}(v\phi)} + c\phi - f\right]\mathrm{d}\Omega = 0 \tag{22}$$

The underlined terms multiplied together give the term ($u$ and $v$ have been assumed for simplicity to be constants) essential in convection dominated flows:

$$\frac{\partial w}{\partial x}\left(uu\frac{\partial \phi}{\partial x}+uv\frac{\partial \phi}{\partial y}\right)+\frac{\partial w}{\partial y}\left(vu\frac{\partial \phi}{\partial x}+vv\frac{\partial \phi}{\partial y}\right)$$
$$=\begin{Bmatrix}\partial w/\partial x\\ \partial w/\partial y\end{Bmatrix}^{\mathrm{T}}\begin{bmatrix}uu & uv\\ vu & vv\end{bmatrix}\begin{Bmatrix}\partial \phi/\partial x\\ \partial \phi/\partial y\end{Bmatrix} \tag{23}$$

In the pure diffusion weak form (say equations (3.3.1) and (3.3.2) with the latter replaced by an anisotropic thermal conductivity) the integrand corresponding to diffusion is

$$\frac{\partial w}{\partial x}\left(k_{xx}\frac{\partial T}{\partial x}+k_{xy}\frac{\partial T}{\partial y}\right)+\frac{\partial w}{\partial y}\left(k_{yx}\frac{\partial T}{\partial x}+k_{yy}\frac{\partial T}{\partial y}\right)$$
$$=\begin{Bmatrix}\partial w/\partial x\\ \partial w/\partial y\end{Bmatrix}^{\mathrm{T}}\begin{bmatrix}k_{xx} & k_{xy}\\ k_{yx} & k_{yy}\end{bmatrix}\begin{Bmatrix}\partial T/\partial x\\ \partial T/\partial y\end{Bmatrix} \tag{24}$$

Comparison of (23) with (24) thus shows again that the least squares formulation produces diffusion now with an anisotropic diffusion matrix.

## D.5 GRADIENT LEAST SQUARES FUNCTIONAL, D-C-R EQUATION

### D.5.1 One dimension

The governing differential equation is

$$R(\phi)\equiv L(\phi)-f\equiv\frac{\mathrm{d}}{\mathrm{d}x}\left(-D\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)+\frac{\mathrm{d}}{\mathrm{d}x}(u\phi)+c\phi-f=0\quad\text{in }\Omega=]a,b[ \tag{1}$$

and we need to pay no attention to the boundary conditions. To stabilize oscillations due to large reaction, the least squares formulation of Section D.4 is of no help. The way to proceed is to generate a new differential equation by differentiating both sides of (1) with respect to $x$:

$$\frac{\mathrm{d}R}{\mathrm{d}x}\equiv\frac{\mathrm{d}}{\mathrm{d}x}L(\phi)-\frac{\mathrm{d}f}{\mathrm{d}x}\equiv\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(-D\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)+\frac{\mathrm{d}^2}{\mathrm{d}x^2}(u\phi)+\frac{\mathrm{d}}{\mathrm{d}x}(c\phi)-\frac{\mathrm{d}f}{\mathrm{d}x}=0 \tag{2}$$

We now form a least squares functional from this:

$$\boxed{\Pi(\phi)=\frac{1}{2}\int_\Omega\left(\frac{\mathrm{d}R}{\mathrm{d}x}\right)^2\mathrm{d}\Omega} \tag{3}$$

The variation is



$$\delta\Pi=\frac{1}{2}\int_\Omega 2\frac{\mathrm{d}R}{\mathrm{d}x}\delta\frac{\mathrm{d}R}{\mathrm{d}x}\mathrm{d}\Omega=\int_\Omega\frac{\mathrm{d}R}{\mathrm{d}x}\frac{\mathrm{d}\delta R}{\mathrm{d}x}\mathrm{d}\Omega=\int_\Omega\frac{\mathrm{d}R}{\mathrm{d}x}\frac{\mathrm{d}L(\delta\phi)}{\mathrm{d}x}\mathrm{d}\Omega \tag{4}$$

Again the interpretation $\delta\phi=w$ is introduced to give

$$\boxed{\int_\Omega\frac{\mathrm{d}L(w)}{\mathrm{d}x}\frac{\mathrm{d}R}{\mathrm{d}x}\mathrm{d}\Omega=0} \tag{5}$$

which we will call the *gradient least squares weak form* (gradientti pienimmän neliön heikko muoto). This is in detail

$$\int_\Omega\left[\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(-D\frac{\mathrm{d}w}{\mathrm{d}x}\right)+\frac{\mathrm{d}^2}{\mathrm{d}x^2}(uw)+\underline{\frac{\mathrm{d}}{\mathrm{d}x}(cw)}\right]\cdot$$
$$\cdot\left[\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(-D\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)+\frac{\mathrm{d}^2}{\mathrm{d}x^2}(u\phi)+\underline{\frac{\mathrm{d}}{\mathrm{d}x}(c\phi)}-\frac{\mathrm{d}f}{\mathrm{d}x}\right]\mathrm{d}\Omega=0 \tag{6}$$

The underlined terms are essential in sensitizing reaction dominated cases. The explanation is analogous to the one given in Section D.4.1 in connection with convection. Quantity $c^2$ can be interpreted as a positive diffusivity (or thermal conductivity).

### D.5.2 Two dimensions

In two dimensions the governing differential equation (D.4.19) can be differentiated in two independent directions to produce two new equations:

$$\frac{\partial R}{\partial x}=0,\qquad\frac{\partial R}{\partial y}=0\quad\text{in }\Omega \tag{7}$$

The most general least squares functional corresponding to this is

$$\Pi(\phi)=\frac{1}{2}\int_\Omega\begin{Bmatrix}\partial R/\partial x\\ \partial R/\partial y\end{Bmatrix}^{\mathrm{T}}\begin{bmatrix}\tau_{xx} & \tau_{xy}\\ \tau_{yx} & \tau_{yy}\end{bmatrix}\begin{Bmatrix}\partial R/\partial x\\ \partial R/\partial y\end{Bmatrix}\mathrm{d}\Omega \tag{8}$$

The *weight factor matrix* (painotekijämatriisi) or here the sensitizing parameter matrix

$$[\tau]=\begin{bmatrix}\tau_{xx} & \tau_{xy}\\ \tau_{yx} & \tau_{yy}\end{bmatrix} \tag{9}$$

is taken to be a symmetric. Symmetry does not prevent generality (see Remark D.4). The election of the elements of the sensitizing parameter matrix is discussed in Chapter 7.

Taking the variation and going through the details similarly as in the previous section gives the gradient least squares weak form (see Remark D.5)

$$\int_\Omega \begin{Bmatrix} \partial L(w)/\partial x \\ \partial L(w)/\partial y \end{Bmatrix}^{\mathrm{T}} \begin{bmatrix} \tau_{xx} & \tau_{xy} \\ \tau_{yx} & \tau_{yy} \end{bmatrix} \begin{Bmatrix} \partial R/\partial x \\ \partial R/\partial y \end{Bmatrix} \mathrm{d}\Omega = 0 \tag{10}$$

in which

$$\begin{aligned}
\frac{\partial L(w)}{\partial x} &= \frac{\partial^2}{\partial x^2}\left(-D\frac{\partial w}{\partial x}\right) + \frac{\partial^2}{\partial x \partial y}\left(-D\frac{\partial w}{\partial y}\right) \\
&\quad + \frac{\partial^2}{\partial x^2}(uw) + \frac{\partial}{\partial x \partial y}(vw) + \underline{\frac{\partial}{\partial x}(cw)} \\
\frac{\partial L(w)}{\partial y} &= \frac{\partial^2}{\partial y \partial x}\left(-D\frac{\partial w}{\partial x}\right) + \frac{\partial^2}{\partial y^2}\left(-D\frac{\partial w}{\partial y}\right) \\
&\quad + \frac{\partial^2}{\partial y \partial x}(uw) + \frac{\partial^2}{\partial y^2}(vw) + \underline{\frac{\partial}{\partial y}(cw)} \\
\frac{\partial R(\phi)}{\partial x} &= \frac{\partial^2}{\partial x^2}\left(-D\frac{\partial \phi}{\partial x}\right) + \frac{\partial^2}{\partial x \partial y}\left(-D\frac{\partial \phi}{\partial y}\right) \\
&\quad + \frac{\partial^2}{\partial x^2}(u\phi) + \frac{\partial}{\partial x \partial y}(v\phi) + \underline{\frac{\partial}{\partial x}(c\phi)} - \frac{\partial f}{\partial x} \\
\frac{\partial R(\phi)}{\partial y} &= \frac{\partial^2}{\partial y \partial x}\left(-D\frac{\partial \phi}{\partial x}\right) + \frac{\partial^2}{\partial y^2}\left(-D\frac{\partial \phi}{\partial y}\right) \\
&\quad + \frac{\partial^2}{\partial y \partial x}(u\phi) + \frac{\partial^2}{\partial y^2}(v\phi) + \underline{\frac{\partial}{\partial y}(c\phi)} - \frac{\partial f}{\partial y}
\end{aligned} \tag{11}$$

The underlined terms are essential in reaction dominated cases. Multiplying them together gives the integrand contribution (constant $c$ has been assumed for simplicity)

$$\begin{Bmatrix} \partial w/\partial x \\ \partial w/\partial y \end{Bmatrix}^{\mathrm{T}} \begin{bmatrix} c^2\tau_{xx} & c^2\tau_{xy} \\ c^2\tau_{yx} & c^2\tau_{yy} \end{bmatrix} \begin{Bmatrix} \partial \phi/\partial x \\ \partial \phi/\partial y \end{Bmatrix} \tag{12}$$



Again a diffusion type term has emerged.

**Remark D.5.** The derivation of equation (10) from expression (8) with matrix notation can be performed as follows. Let us denote the integrand in (8) as

$$I = \{\}^{\mathrm{T}}[\,]\{\} \tag{13}$$

The square matrix does not depend on $\phi$. The variation is

$$\delta I = \delta\{\}^{\mathrm{T}}[\,]\{\} + \{\}^{\mathrm{T}}[\,]\delta\{\} = 2\delta\{\}^{\mathrm{T}}[\,]\{\} \tag{14}$$

The last equality follows from the fact that the two terms in the sum are scalars (1×1-matrices) and the value of a scalar does not change in transposition. Thus

$$\left(\{\}^{\mathrm{T}}[\,]\delta\{\}\right)^{\mathrm{T}} = \delta\{\}^{\mathrm{T}}[\,]^{\mathrm{T}}\left(\{\}^{\mathrm{T}}\right)^{\mathrm{T}} = \delta\{\}^{\mathrm{T}}[\,]\{\} \tag{15}$$

Here use is made of the symmetry of the square matrix. □

## REFERENCES

Crandall, S. (1956). *Engineering Analysis*, McGraw-Hill, New York.

Lanczos, C. (1970). *The Variational Principles of Mechanics*, 4th ed., University of Toronto Press, Toronto.

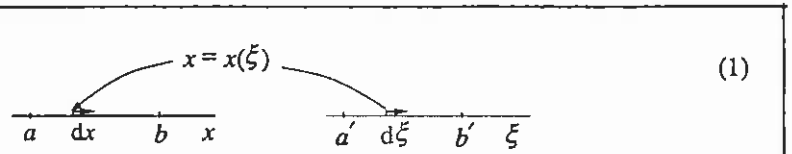
# APPENDIX E FORMULAS FOR MAPPED ELEMENTS

## E.1 TRANSFORMATION OF INTEGRALS

Finite element calculations consist in practice of evaluation of definite integrals in the reference element space. This means that formulas must be available for transforming integrals from the global space to the reference space. The necessary formulas are given in the following mostly without proofs. The theme is a part of classical mathematics. An excellent Finnish reference is Väisälä (1960).

### E.1.1 One dimension

**Table E.1** One dimension




$$\mathrm{d}x = M\,\mathrm{d}\xi \tag{2}$$

$$\int_a^b f\,\mathrm{d}x = \int_{a'}^{b'} f M\,\mathrm{d}\xi \tag{3}$$

$$M = \mathrm{d}x/\mathrm{d}\xi \tag{4}$$

Mapping (1) in Table E.1 transforms a differential line element $\mathrm{d}\xi$ of the $\xi$-axis to a differential line element $\mathrm{d}x$ of the $x$-axis:

$$\mathrm{d}x = \frac{\mathrm{d}x}{\mathrm{d}\xi}\mathrm{d}\xi = M(\xi)\mathrm{d}\xi \tag{1}$$

The coefficient $M$ might be called scaling or *magnification factor* (suurennustekijä) giving the ratio between the lengths of the line elements. Let us consider the integral

$$\int_a^b f(x)\mathrm{d}x \tag{2}$$

with $x$ as the integration variable. Remembering the meaning of the definite integral as the limit of the Riemann sum $\sum f(x)\Delta x$, we realize that this sum can be formed also in the $\xi$-space just by multiplying the line elements $\Delta\xi$ by the factor $M = \mathrm{d}x/\mathrm{d}\xi$. The transformation formula for integrals is thus





$$\int_a^b f(x)\mathrm{d}x = \int_{a'}^{b'} f\big(x(\xi)\big)M(\xi)\mathrm{d}\xi \tag{3}$$

or shortly formula (3) of Table E.1.

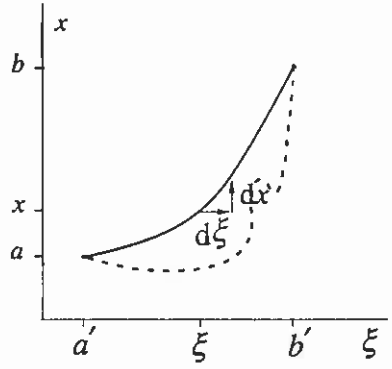



**Figure E.1** Mapping $x = x(\xi)$.

In the isoparametric mapping where

$$x = x(\xi) = \sum N_i(\xi)x_i \tag{4}$$

the normal requirement is that the mapping is one-to-one or bijection (kääntäen yksikäsitteinen, bijektio). If we consider the mapping as a curve in the $\xi x$-plane (Figure E.1), the graph must be either ascending or descending in the interval $[a',b']$. In other words, the derivative $\mathrm{d}x/\mathrm{d}\xi$ must not change its sign; the solid line in the figure. Otherwise the reference element can partly fold on the $x$-axis or even map outside the interval $[a,b]$: the dashed line in the figure. For instance, let the nodes of a four-noded reference element follow each other in the normal order 1′, 2′, 3′, 4′ in the positive direction of the $\xi$-axis. Let us now put the nodes in the global element, say in the order 1, 3, 2, 4 in the positive direction of the $x$-axis. The corresponding mapping will obviously not be bijective.

The shape function derivative in global space can be solved from the equation

$$\frac{\mathrm{d}N_i(\xi)}{\mathrm{d}\xi} = \frac{\mathrm{d}N_i}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}\xi} \tag{5}$$

produced by chain differentiation (cf. formula (3.2.13)). If $\mathrm{d}x/\mathrm{d}\xi = 0$ at a certain point — compare the dashed line in Figure E.1 — formula (5) gives an infinite value for the derivative $\mathrm{d}N_i/\mathrm{d}x$. This is of course not admissible in the

approximation of smoothly behaving functions. However, the possibility of producing infinite derivatives has been employed with profit in two and three dimensions in fracture mechanics where certain derivatives at the crack tip are known to have infinite values.

The use of isoparametric elements in one dimension is of course not in principle necessary if the global elements are not curved. Lagrange shape functions can be easily generated for any however nonuniform node geometry in the global space. The presentation above should be considered as an introduction in a case where some features of the mapping can be easily visualized.

**Example E.1** The length coordinates $L_1$ and $L_2$ were described in Section 2.2.1. We will derive the closed form integration formula

$$\int_{x_1}^{x_2} L_1^{\alpha} L_2^{\beta}\,\mathrm{d}x = \frac{\alpha!\beta!}{(\alpha+\beta+1)!}(x_2 - x_1) \tag{a}$$

where $\alpha$ and $\beta$ are integers and $0! \equiv 1$ This formula is convenient in some analytical finite element calculations.


**Figure (a)**

We take $\xi = L_2$ as the independent natural coordinate for the reference element (Figure (a)). The length coordinates $L_1$ and $L_2$ are also the shape functions $N_1$ and $N_2$ for a two-noded line element. The mapping

$$\begin{aligned} x &= \sum_{i=1}^{2} N_i x_i = L_1 x_1 + L_2 x_2 = (1 - L_2)x_1 + L_2 x_2 \\ &= x_1 + L_2(x_2 - x_1) = x_1 + \xi(x_2 - x_1) \end{aligned} \tag{b}$$

which maps the $\xi$-axis interval [0,1] onto the $x$-axis interval $[x_1, x_2]$ can thus also be considered as an isoparametric mapping of a two-noded line element. Here the magnification factor

$$M = \frac{\mathrm{d}x}{\mathrm{d}\xi} = x_2 - x_1 \tag{c}$$

which is the element length, does not depend on position. According to formula (3) of Table E.1

$$\int_{x_1}^{x_2} L_1^{\alpha} L_2^{\beta}\,\mathrm{d}x = \int_0^1 L_1^{\alpha} L_2^{\beta}(x_2 - x_1)\,\mathrm{d}\xi = (x_2 - x_1)\int_0^1 \xi^{\beta}(1-\xi)^{\alpha}\,\mathrm{d}\xi \tag{d}$$

We make use of the general formula



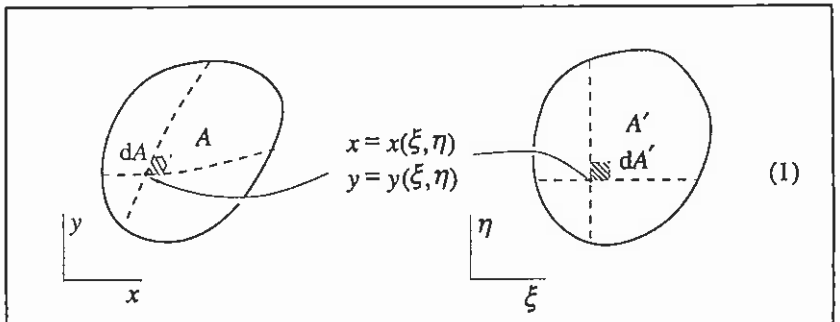
$$\int_0^a t^m (a-t)^n\,\mathrm{d}t = \frac{m!n!}{(m+n+1)!}a^{m+n+1} \tag{e}$$

which can be proved by integration by parts. When this is applied in equation (d) we get the result (a).

### E.1.2 Two dimensions

Table E.2 contains the essential formulas needed in applications. The meaning of the notations should be rather obvious.

**Table E.2** Two dimensions




(1)

$$\mathrm{d}A = M\,\mathrm{d}A' \tag{2}$$

$$\int_A f\,\mathrm{d}A = \int_{A'} f M\,\mathrm{d}A' \tag{3}$$

$$M = \det[J] \tag{4}$$

$$J = \left[\frac{\partial(x,y)}{\partial(\xi,\eta)}\right] = \begin{bmatrix} \partial x/\partial\xi & \partial x/\partial\eta \\ \partial y/\partial\xi & \partial y/\partial\eta \end{bmatrix} \tag{5}$$

Formula (3) of the table expressed in more detail is

$$\int_A f(x,y)\mathrm{d}A = \int_{A'} f\big(x(\xi,\eta), y(\xi,\eta)\big) M(\xi,\eta)\,\mathrm{d}A' \tag{6}$$

or

$$\int_A f(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_{A'} f\big(x(\xi,\eta), y(\xi,\eta)\big) M(\xi,\eta)\,\mathrm{d}\xi\,\mathrm{d}\eta \tag{7}$$

The figure in the table shows how a rectangular differential surface element $\mathrm{d}A'$ is mapped onto a trapezoidal differential surface element $\mathrm{d}A$. Thus, if we write according to formula (7) formally $\mathrm{d}x\,\mathrm{d}y = M\,\mathrm{d}\xi\,\mathrm{d}\eta$, it must be realized that this

does not mean that two rectangular differential elements are mapped on each other.

In the isoparametric technique where

$$x = x(\xi,\eta) = \sum N_i(\xi,\eta)\,x_i \qquad y = y(\xi,\eta) = \sum N_i(\xi,\eta)\,y_i \tag{8}$$

the mapping must again be a bijection for the same reasons as described in the previous section. The right-hand side of formula (4) in Table E.2 must be put in the general case inside the absolute value signs. Here we have assumed that the mapping distorts the domain so mildly that the angle between the coordinate lines $\eta = \text{constant}$, $\xi = \text{constant}$ in the $xy$-plane measured from the former line in the anticlockwise direction remains in the interval $(0,180°)$. In this case the determinant $\det[J]$ of the Jacobian matrix $[J]$ (see Section 3.2.4), called the jacobian determinant or shortly the *jacobian* (Jacobin funktionaali-determinantti) remains positive, provided the images of the line elements $d\xi$ and $d\eta$ do not shrink to zero lengths.

In numerical integration the value of the Jacobian $\det[J]$ is evaluated at each integration point. The program should contain a test, which stops the run if $\det[J] \le 0$ at some integration point and give the element in which this happens. In this case the element is too distorted. This test can also detect geometric errors in the mesh due to erroneous input data. If the form $|\det[J]|$ is employed, more freedom in element nodal numbering order in the global space is achieved but the test for possible errors is lost.

Literature contains some rules to avoid too strong distortion. For instance, the angles between the sides of a four-noded quadrilateral should be smaller than 180°. The side midnodes of a nine-noded quadrilateral should situate roughly in the midpoint area etc.

The position vector in the $xy$-plane is

$$\mathbf{r} = \mathbf{r}(\xi,\eta) = x(\xi,\eta)\mathbf{i} + y(\xi,\eta)\mathbf{j} \tag{9}$$

and its differential is

$$d\mathbf{r} = \left(\frac{\partial x}{\partial \xi}d\xi + \frac{\partial x}{\partial \eta}d\eta\right)\mathbf{i} + \left(\frac{\partial y}{\partial \xi}d\xi + \frac{\partial y}{\partial \eta}d\eta\right)\mathbf{j} \tag{10}$$



The derivation of result (4) of Table E.2 can be based on this formula by studying how the differential vectors $d\mathbf{r}_1' = d\xi\mathbf{i}$ and $d\mathbf{r}_2' = d\eta\mathbf{j}$ in the $\xi\eta$-plane are mapped to differential vectors $d\mathbf{r}_1$ and $d\mathbf{r}_2$ on the $xy$-plane.

**Example E.2.** The surface coordinates $L_1$, $L_2$, $L_3$ for triangles were described in Section 3.2.1. We will derive the closed form integration formula

$$\int_A L_1^\alpha L_2^\beta L_3^\gamma \, dA = \frac{\alpha!\beta!\gamma!}{(\alpha+\beta+\gamma+2)!}2A \tag{a}$$

This is a direct extension of the corresponding formula (a) in Example E.1 concerning length coordinates. $A$ is the area of the triangle.

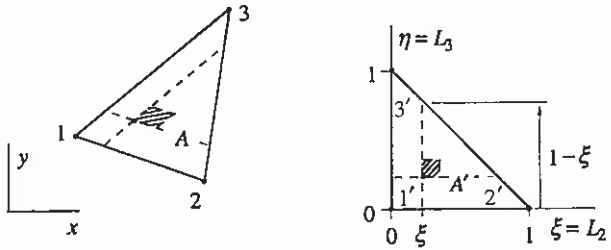



**Figure (a)**

We select as the independent natural coordinates $\xi = L_2$ and $\eta = L_3$ (Figure (a)). The mapping from the $\xi\eta$-plane on the $xy$-plane can be thought as an isoparametric mapping of a three-noded element:

$$x = \sum N_i x_i = L_1x_1 + L_2x_2 + L_3x_3 = (1-\xi-\eta)x_1 + \xi x_2 + \eta x_3$$
$$y = \sum N_i y_i = L_1y_1 + L_2y_2 + L_3y_3 = (1-\xi-\eta)y_1 + \xi y_2 + \eta y_3 \tag{b}$$

The Jacobian matrix

$$[J] = \begin{bmatrix} \partial x/\partial\xi & \partial x/\partial\eta \\ \partial y/\partial\xi & \partial y/\partial\eta \end{bmatrix} = \begin{bmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{bmatrix} \tag{c}$$

and similarly the magnification factor

$$\det[J] = \begin{vmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 \\ x_1 & x_2 - x_1 & x_3 - x_1 \\ y_1 & y_2 - y_1 & y_3 - y_1 \end{vmatrix} = \begin{vmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{vmatrix} = 2A \tag{d}$$

are here constants with respect to position. The $2\times2$ determinant has been expanded to a $3\times3$ determinant to get a more symmetric form. Some determinant manipulation rules have been made use of. (For instance, the value of a determinant is not changed if a row multiplied by a constant is added to another row.) According to formula (2) of Table E.2, $dA/dA' = M$ and because $M$ is a constant, the ratio of the finite areas $A$ and $A'$ is also the same: $A/A' = M = \det[J]$ and as $A' = 1/2$ we have $A = \det[J]/2$. This explains the last equation (d).

Now (see Figure (a))

$$\int_A L_1^\alpha L_2^\beta L_3^\gamma \, dA = \int_{A'} L_1^\alpha L_2^\beta L_3^\gamma \, 2A \, dA' = 2A \int_0^1 \xi^\beta \left[ \int_0^{1-\xi} \eta^\gamma (1-\xi-\eta)^\alpha \, d\eta \right] d\xi \tag{e}$$

Formula (e) of Example E.1 is applied twice:

$$\int_0^{1-\xi} \eta^\gamma (1-\xi-\eta)^\alpha \, d\eta = \frac{\gamma!\alpha!}{(\gamma+\alpha+1)!}(1-\xi)^{\gamma+\alpha+1} \tag{f}$$

$$\int_A L_1^\alpha L_2^\beta L_3^\gamma \, dA = 2A \frac{\gamma!\alpha!}{(\gamma+\alpha+1)!} \int_0^1 \xi^\beta (1-\xi)^{\gamma+\alpha+1} \, d\xi$$

$$= 2A \frac{\gamma!\alpha!}{(\gamma+\alpha+1)!} \frac{\beta!(\gamma+\alpha+1)!}{(\beta+\gamma+\alpha+2)!} = \frac{\alpha!\beta!\gamma!}{(\alpha+\beta+\gamma+2)!} 2A \tag{g}$$

For the triangle, it is possible to express the area coordinates in closed form in the global coordinates. We can write the equations

$$\begin{aligned} 1 &= L_1 + L_2 + L_3 \\ x &= x_1 L_1 + x_2 L_2 + x_3 L_3 \\ y &= y_1 L_1 + y_2 L_2 + y_3 L_3 \end{aligned} \tag{h}$$

the first on being the constraint on the area coordinates and the second and third the isoparametric mappings. In matrix notation:

$$\begin{Bmatrix} 1 \\ x \\ y \end{Bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{bmatrix} \begin{Bmatrix} L_1 \\ L_2 \\ L_3 \end{Bmatrix} \tag{i}$$

and the solution is, say using Cramer's rule,

$$L_1 = \frac{1}{2A}\left[ x_2 y_3 - x_3 y_2 + (y_2 - y_3)x + (x_3 - x_2)y \right] \tag{j}$$

...

The dots mean that the formulas for $L_2$ and $L_3$ are obtained by cyclic permutation in the order 1, 2, 3 of the indices. These formulas are given also in Section F.2.1.

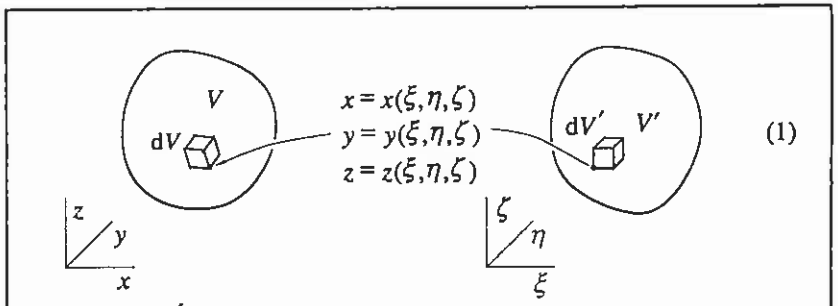
### E.1.3 Three dimensions

The contents of Table E.3 correspond to that of Table E.2 in three dimensions. Again the mapping should distort the domain so mildly that the images of the differential line elements $d\xi$, $d\eta$, $d\zeta$ still form a right-handed system. The Jacobian $\det[J]$ remains then positive. The isoparametric mapping is of the form



$$\begin{aligned} x &= x(\xi,\eta,\zeta) = \Sigma N_i(\xi,\eta,\zeta) x_i \\ y &= y(\xi,\eta,\zeta) = \Sigma N_i(\xi,\eta,\zeta) y_i \\ z &= z(\xi,\eta,\zeta) = \Sigma N_i(\xi,\eta,\zeta) z_i \end{aligned} \tag{11}$$

**Table E.3** Three dimensions




(1)

$$dV = M \, dV' \tag{2}$$

$$\int_V f \, dV = \int_{V'} f M \, dV' \tag{3}$$

$$M = \det[J] \tag{4}$$

$$J = \left[ \frac{\partial(x,y,z)}{\partial(\xi,\eta,\zeta)} \right] = \begin{bmatrix} \partial x/\partial\xi & \partial x/\partial\eta & \partial x/\partial\zeta \\ \partial y/\partial\xi & \partial y/\partial\eta & \partial y/\partial\zeta \\ \partial z/\partial\xi & \partial z/\partial\eta & \partial z/\partial\zeta \end{bmatrix} \tag{5}$$

## E.2 TRANSFORMATION OF INTEGRALS WITH DIFFERING NUMBER OF SPACE DIMENSIONS

### E.2.1 One independent variable

Table E.4 is concerned with the mapping of a line to a space curve. In mathematical terminology equations (1) in the table give a parametric representation of the curve and $\xi$ is called the curve parameter.

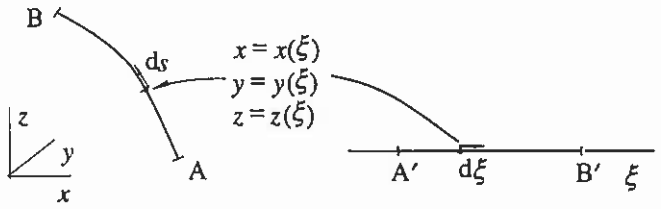
Let us measure the curve arclength $s$ in the direction of increasing $\xi$, that is, from point A towards point B. In principle we need to evaluate integrals $\int f(x(s), y(s), z(s)) ds$ or shortly $\int f(s) ds$. When using the finite element method the integrand is however approximated so that the integrals to be determined are finally of the type

$$\int_A^B f(\xi) ds \tag{1}$$

Thus no explicit dependence $s = s(\xi)$ is needed; it is enough to have the relationship $ds = M(\xi)d\xi$. It has been given in the table. The transformation formula (3) of the table is thus in more detail

$$\int_A^B f(\xi)\,ds = \int_{A'}^{B'} f(\xi)M(\xi)\,d\xi \tag{2}$$

**Table E.4** Line to three dimensions




(1)

$$ds = M\,d\xi \tag{2}$$

$$\int_A^B f\,ds = \int_{A'}^{B'} f M\,d\xi \tag{3}$$

$$M = (M_x^2 + M_y^2 + M_z^2)^{1/2} \tag{4}$$

$$M_x = dx/d\xi, \quad M_y = dy/d\xi, \quad M_z = dz/d\xi \tag{5}$$

Table E.5 describes the mapping of a line to a plane curve. In applications the expression for the unit normal vector $\mathbf{n}(\xi)$ to the curve is often needed. When using the formulas of the table, it should be noted that $\mathbf{n}$ has been directed here $90^\circ$ to the anticlockwise direction from the increasing direction of the variable $\xi$.

In the isoparametric mapping

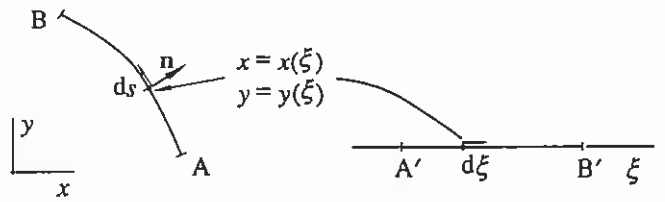
$$\begin{aligned} x &= x(\xi) = \Sigma N_i(\xi)x_i \\ y &= y(\xi) = \Sigma N_i(\xi)y_i \end{aligned} \tag{3}$$

and for instance

$$\begin{aligned} M_x &= \Sigma dN_i/d\xi \cdot x_i \\ M_y &= \Sigma dN_i/d\eta \cdot y_i \end{aligned} \tag{4}$$



**Table E.5** Line to two dimensions




(1)

$$ds = M\,d\xi \tag{2}$$

$$\int_A^B f\,ds = \int_{A'}^{B'} f M\,d\xi \tag{3}$$

$$\mathbf{n} = n_x\mathbf{i} + n_y\mathbf{j} = \frac{M_y}{M}\mathbf{i} - \frac{M_x}{M}\mathbf{j} \tag{4}$$

$$M = (M_x^2 + M_y^2)^{1/2} \tag{5}$$

$$M_x = dx/d\xi, \quad M_y = dy/d\xi \tag{6}$$

### E.2.2 Two independent variables

Table E.6 is concerned with the mapping of a plane domain to a space surface. The variables $\xi$ and $\eta$ can be called in analogy with the discussion in the previous section as surface parameters. A differential plane element $dA'$ is mapped onto a differential surface element $dS$. In the formulas of the table the unit normal vector $\mathbf{n}$ has been directed so that the images of $d\xi$, $d\eta$ and the vector $\mathbf{n}$ form in this order a right-handed system. In practice integrals of the form $\int_S f(\xi,\eta)dS$ are needed and formula (3) of the table means in detail

$$\int_S f(\xi,\eta)dS = \int_{A'} f(\xi,\eta)M(\xi,\eta)dA' \tag{5}$$

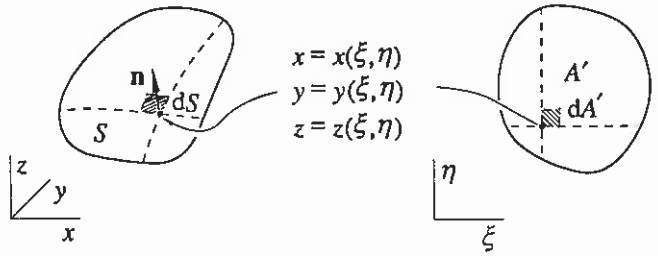
Sometimes it is necessary to generate a local rectangular coordinate system at certain points of the surface so that one axis — for instance $z'$ — coincides with the normal direction and the two remaining — $x'$ and $y'$ — coincide with the tangent plane. The treatment is based on the fact that $\partial\mathbf{r}/\partial\xi$ and $\partial\mathbf{r}/\partial\eta$ are seen to be tangent vectors to the surface and they thus span the tangent plane. Here $\mathbf{r}$ means the position vector of the surface:

$$\mathbf{r} = \mathbf{r}(\xi,\eta) = x(\xi,\eta)\mathbf{i} + y(\xi,\eta)\mathbf{j} + z(\xi,\eta)\mathbf{k} \tag{6}$$

In the isoparametric mapping

$$x = x(\xi,\eta) = \sum N_i(\xi,\eta)\, x_i$$
$$y = y(\xi,\eta) = \sum N_i(\xi,\eta)\, y_i \qquad (7)$$
$$z = z(\xi,\eta) = \sum N_i(\xi,\eta)\, z_i$$

**Table E.6** Plane to three dimensions




(1)

$$dS = M\, dA' \qquad (2)$$

$$\int_S f\, dS = \int_{A'} f M\, dA' \qquad (3)$$

$$\mathbf{n} = n_x\mathbf{i} + n_y\mathbf{j} + n_z\mathbf{k} = \frac{M_x}{M}\mathbf{i} + \frac{M_y}{M}\mathbf{j} + \frac{M_z}{M}\mathbf{k} \qquad (4)$$

$$M = (M_x^2 + M_y^2 + M_z^2)^{1/2} \qquad (5)$$

$$M_x = \det\left[\frac{\partial(y,z)}{\partial(\xi,\eta)}\right] = \det\begin{bmatrix} \partial y/\partial\xi & \partial y/\partial\eta \\ \partial z/\partial\xi & \partial z/\partial\eta \end{bmatrix}$$

$$M_y = \det\left[\frac{\partial(z,x)}{\partial(\xi,\eta)}\right] = \det\begin{bmatrix} \partial z/\partial\xi & \partial z/\partial\eta \\ \partial x/\partial\xi & \partial x/\partial\eta \end{bmatrix} \qquad (6)$$

$$M_z = \det\left[\frac{\partial(x,y)}{\partial(\xi,\eta)}\right] = \det\begin{bmatrix} \partial x/\partial\xi & \partial x/\partial\eta \\ \partial y/\partial\xi & \partial y/\partial\eta \end{bmatrix}$$

## REFERENCE

Väisälä, K. (1960). *Matematiikka III*, Teknillinen korkeakoulu, moniste n:o 11, Helsinki.

# APPENDIX F SHAPE FUNCTION INTEGRALS

Some ready evaluated shape function integrals for simple elements and simple geometries are given below to ease hand calculations.

## F.1 ONE-DIMENSIONAL ELEMENTS

### F.1.1 Linear line element

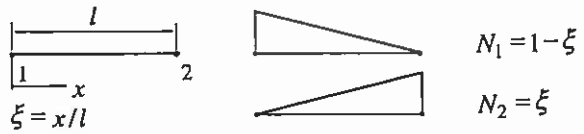



**Figure F.1** Linear element and shape functions.

Some integrals:

$$\int_0^l N_i N_j \,dx \triangleq \frac{l}{6}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

$$\int_0^l N_i N_j' \,dx \triangleq \frac{1}{2}\begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}$$

$$\int_0^l N_i' N_j' \,dx \triangleq \frac{1}{l}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \qquad (1)$$

$$\int_0^l N_i \,dx \triangleq \frac{l}{2}\begin{Bmatrix} 1 \\ 1 \end{Bmatrix}$$

$$\int_0^l N_i' \,dx \triangleq \begin{Bmatrix} -1 \\ 1 \end{Bmatrix}$$

The dash refers to differentiation with respect to $x$. The left-hand side indicates to an element of the right hand side matrix. Index $i$ refers to row and index $j$ to column.





### F.1.2 Quadratic line element

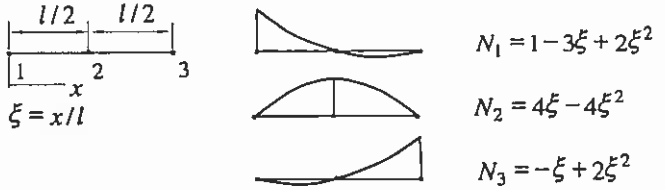



**Figure F.2** Quadratic element and shape functions.

Some integrals:

$$\int_0^l N_i N_j \,dx \triangleq \frac{l}{30}\begin{bmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{bmatrix} \qquad \int_0^l N_i N_j' \,dx \triangleq \frac{1}{6}\begin{bmatrix} -3 & 4 & -1 \\ -4 & 0 & 4 \\ 1 & -4 & 3 \end{bmatrix}$$

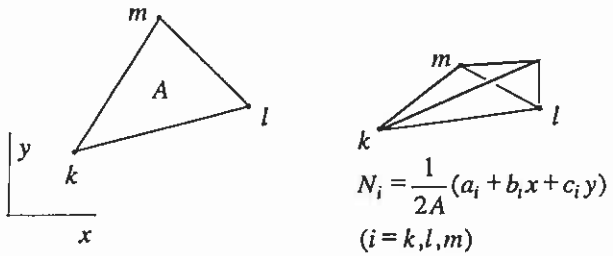
$$\int_0^l N_i N_j'' \,dx \triangleq \frac{2}{3l}\begin{bmatrix} 1 & -2 & 1 \\ 4 & -8 & 4 \\ 1 & -2 & 1 \end{bmatrix} \qquad \int_0^l N_i' N_j' \,dx \triangleq \frac{1}{3l}\begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{bmatrix}$$

$$\int_0^l N_i' N_j'' \,dx \triangleq \frac{4}{l^2}\begin{bmatrix} -1 & 2 & -1 \\ 0 & 0 & 0 \\ 1 & -2 & 1 \end{bmatrix} \qquad \int_0^l N_i'' N_j'' \,dx \triangleq \frac{16}{l^3}\begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{bmatrix} \qquad (2)$$

$$\int_0^l N_i \,dx \triangleq \frac{l}{6}\begin{Bmatrix} 1 \\ 4 \\ 1 \end{Bmatrix} \qquad \int_0^l N_i' \,dx \triangleq \begin{Bmatrix} -1 \\ 0 \\ 1 \end{Bmatrix} \qquad \int_0^l N_i'' \,dx \triangleq \frac{4}{l}\begin{Bmatrix} 1 \\ -2 \\ 1 \end{Bmatrix}$$

## F.2 TWO-DIMENSIONAL ELEMENTS

### F.2.1 Linear triangular element




$$N_i = \frac{1}{2A}(a_i + b_i x + c_i y)$$
$$(i = k, l, m)$$

**Figure F.3** Linear triangular element and shape functions.

The quantities $a$, $b$, $c$ are obtained by applying cyclic permutation in the order $k$, $l$, $m$ to the following formulas:

$$\begin{aligned} a_k &= x_l y_m - x_m y_l \\ b_k &= y_l - y_m \\ c_k &= x_m - x_l \end{aligned} \tag{1}$$

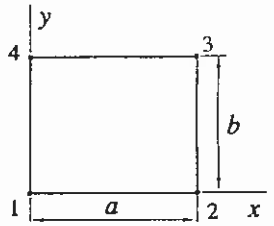
Some integrals ($i, j = k, l, m$):

$$\begin{aligned} &\int_A N_i N_j \,dA = \frac{1}{6}A \quad (j = i), \quad = \frac{1}{12}A \quad (j \neq i) \\ &\int_A N_i N_{j,x} \,dA = \frac{1}{6} b_j \qquad \int_A N_i N_{j,y} \,dA = \frac{1}{6} c_j \\ &\int_A N_{i,x} N_{j,x} \,dA = \frac{1}{4A} b_i b_j \\ &\int_A N_{i,x} N_{j,y} \,dA = \frac{1}{4A} b_i c_j \\ &\int_A N_{i,y} N_{j,y} \,dA = \frac{1}{4A} c_i c_j \\ &\int_A N_i \,dA = \frac{1}{3}A \qquad \int_A N_{i,x} \,dA = \frac{1}{2} b_i \qquad \int_A N_{i,y} \,dA = \frac{1}{2} c_i \end{aligned} \tag{2}$$

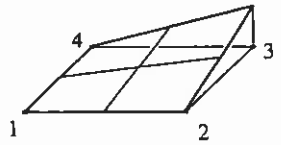
$A$ is the area of the triangle. The notations $,x$ and $,y$ refer to partial differentiation with respect to $x$ and $y$.



### F.2.2 Bilinear rectangular element




$$\xi = x/a$$
$$\eta = x/b$$




$$N_1 = (1-\xi)(1-\eta)$$
$$N_2 = \xi(1-\eta)$$
$$N_3 = \xi\eta$$
$$N_4 = (1-\xi)\eta$$

**Figure F.4** Bilinear rectangular element and shape functions.

Some integrals:

$$\int_A N_i N_j \,dA \triangleq \frac{ab}{36}\begin{bmatrix} 4 & 2 & 1 & 2 \\ 2 & 4 & 2 & 1 \\ 1 & 2 & 4 & 2 \\ 2 & 1 & 2 & 4 \end{bmatrix}$$

$$\int_A N_i N_{j,x} \,dA \triangleq \frac{b}{12}\begin{bmatrix} -2 & 2 & 1 & -1 \\ -2 & 2 & 1 & -1 \\ -1 & 1 & 2 & -2 \\ -1 & 1 & 2 & -2 \end{bmatrix}$$

$$\int_A N_i N_{j,y} \,dA \triangleq \frac{a}{12}\begin{bmatrix} -2 & -1 & 1 & 2 \\ -1 & -2 & 2 & 1 \\ -1 & -2 & 2 & 1 \\ -2 & -1 & 1 & 2 \end{bmatrix}$$

$$\int_A N_{i,x} N_{j,y} \,dA \triangleq \frac{1}{4}\begin{bmatrix} 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 \end{bmatrix}$$

(3)

$$\int_A N_i N_{j,xy}\,dA \triangleq \frac{1}{4}\begin{bmatrix} 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \end{bmatrix}$$

$$\int_A N_{i,x} N_{j,x}\,dA \triangleq \frac{1}{6}\frac{b}{a}\begin{bmatrix} 2 & -2 & -1 & 1 \\ -2 & 2 & 1 & -1 \\ -1 & 1 & 2 & -2 \\ 1 & -1 & -2 & 2 \end{bmatrix}$$

$$\int_A N_{i,y} N_{j,y}\,dA \triangleq \frac{1}{6}\frac{a}{b}\begin{bmatrix} 2 & 1 & -1 & -2 \\ 1 & 2 & -2 & -1 \\ -1 & -2 & 2 & 1 \\ -2 & -1 & 1 & 2 \end{bmatrix}$$

$$\int_A N_i\,dA \triangleq \frac{ab}{4}\begin{Bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{Bmatrix} \qquad \int_A N_{i,x}\,dA \triangleq \frac{b}{2}\begin{Bmatrix} -1 \\ 1 \\ 1 \\ -1 \end{Bmatrix} \qquad \int_A N_{i,y}\,dA \triangleq \frac{a}{2}\begin{Bmatrix} -1 \\ -1 \\ 1 \\ 1 \end{Bmatrix}$$