An Introduction to the Finite Element Method

Department of Engineering Science & Mechanics
Virginia Polytechnic Institute & State University
Blacksburg, VA 24061-0219

An Introduction to the Finite Element Method
CONTENTS

Chapter 1: Introduction

1.1 What is the Finite Element Method?

The Finite Element Method (FEM) is a technique to numerically find an approximate solution of a given initial-boundary-value (IBV) problem. The approximate solution usually gives very good values of the unknown function or functions and of its/their derivatives at discrete points in the domain of study which is the region occupied by the body at a certain time in the time duration of interest.

An IBV problem usually represents a *mathematical model* of a physical phenomenon such as transient heat conduction in a body, fluid flow in a pipe, flow of blood in an artery pumped by the heart, wind flow in a hurricane, motion of ground in an earthquake, collision between two cars, impact between two rigid or deformable bodies, torsion of a shaft, bending of a beam, free and forced vibrations of a structure, flow of air around an airplane, and the propagation of electromagnetic or sound waves in a medium.

1.2 Mathematical Model

A *mathematical model* is comprised of either ordinary or partial differential or integral or algebraic equations together with a set of initial and boundary conditions. The number of equations generally equals the number of degrees of freedom in a problem. For example, a heat conduction problem has only temperature as the unknown and the mathematical model involves only one differential equation. However, in a hurricane an air particle can move in any direction in space, therefore, it has three translational degrees of freedom and the mathematical representation of this phenomenon will involve either three scalar partial differential equations, one for movement of a particle in each direction of motion, or a vector partial differential equation with the vector having three linearly independent components. The mathematical model of a thermomechanical problem will have, in general, four coupled partial differential equations, three for the translational degrees of freedom of a particle and one for temperature. Initial conditions describe in mathematical terms the initial state of a body. In a mechanical problem, boundary conditions specify loads acting either on a part or on the entire boundary of a body. In stead of prescribing loads, one could also give the final positions of material particles of the boundary. The material of a body is described through constitutive

relations that may be algebraic, differential or integral equations. Hooke's law describing a linear relation between stresses and strains is a constitutive relation for a linear elastic material. Similarly, Fourier's law of heat conduction is a constitutive relation for a heat conducting body, and Ohm's law is a constitutive relation for an electric conductor.

We now construct a mathematical model of the problem of the bending of a cantilever beam of uniform rectangular cross-section loaded by a distributed load per unit length, f, on its top surface. A schematic sketch of the problem is shown in Fig. 1.2.1 and it is generally studied in a first course

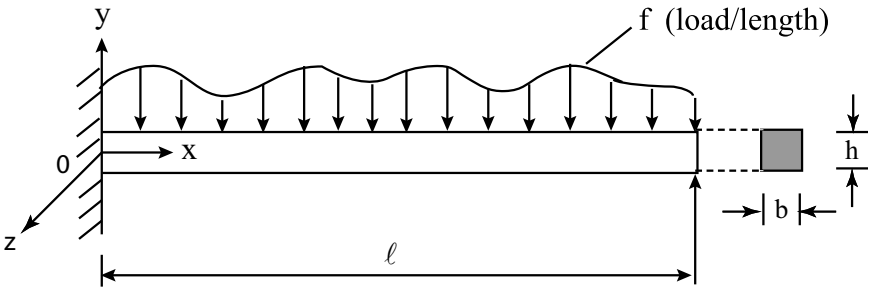


Figure 1.2.1: A beam loaded by a uniformly distributed load f on the top surface, clamped at the left end, and simply supported at the right end.

on Mechanics of Deformable Bodies under the following assumptions:

(i) Plane sections initially perpendicular to the midsurface of the beam remain plane and perpendicular to the deformed shape of the midsurface.

(ii) The state of stress in the beam can be approximated by a uniaxial stress along the beam axis.

(iii) The material of the beam is homogeneous and isotropic, strains induced are infinitesimal, and the material obeys Hooke's law.

The deflection, w, of the beam is governed by the following 4th-order ordinary differential equation:

$$EI\frac{d^4w}{dx^4} + f(x) = 0,\ 0 < x < \ell, \tag{1.2.1}$$

and boundary conditions:

$$\begin{aligned} &\text{at } x = 0,\ w = 0,\ \frac{dw}{dx} = 0, \\ &\text{at } x = \ell,\ EI\frac{d^2w}{dx^2} = 0,\ w = 0. \end{aligned} \tag{1.2.2}$$

Since it is an equilibrium problem, no initial conditions are needed. In equations (1.2.1) and (1.2.2) E equals Young's modulus of the material of the beam, and $I = b\int_{-h/2}^{h/2} y^2 dy$ the second moment of area about the neutral axis. For a beam made of a homogeneous material, for the problem shown in Fig. 1.2.1, the neutral axis is the z-axis passing through the centroid of the cross-section of the beam.

Exercise 1.2.1 Draw a free body diagram of a section of the beam, and derive equation (1.2.1). [If you are not an engineering major, ignore this exercise].

Equations (1.2.1) and (1.2.2) constitute a mathematical model of the problem that is based on the above-listed assumptions (i) through (iii). It involves a 4th-order linear ordinary differential equation. For simple functions f describing the distributed load, these equations can be solved analytically and a closed-form expression for w can be found. Should the deflection w so computed not come close to that observed experimentally, then one or more of the assumptions are incorrect and one needs to improve upon the mathematical model. Incidentally, for a beam with length/height about 5 the computed deflection does not match well with that observed experimentally and one needs to replace assumption (i) by the following: plane sections remain plane during bending but need not stay orthogonal to the deformed shape of the midsurface of the beam. This refinement will be considered later.

This course is concerned with finding an approximate solution of a given IBV, and whenever possible, determining the error in the approximate solution. Thus refinements of mathematical models will not be discussed.

A generalization of the above problem is that of a railroad track that rests on the ground. As the track deflects, the force exerted by the ground on the track depends upon the deflection of a point. One way to simulate the interaction between the track and the ground is to replace the ground by a series of springs distributed along the length of the track as depicted in Fig. 1.2.2. Depending upon the soil and its degree of saturation, the force exerted by it on a point of the track may be a nonlinear function of the deflection of that point. Thus the deflection is now governed by the following differential equation obtained from equation (1.2.1) by setting $f = \tilde{f} - Kw$ where K equals the spring constant of the soil, and $\tilde{f}$ is the force/length exerted by the weight of the rail car.

One may assume that the weight of the car is transferred to the railway track as point loads acting at the contact points between the wheels and the track.

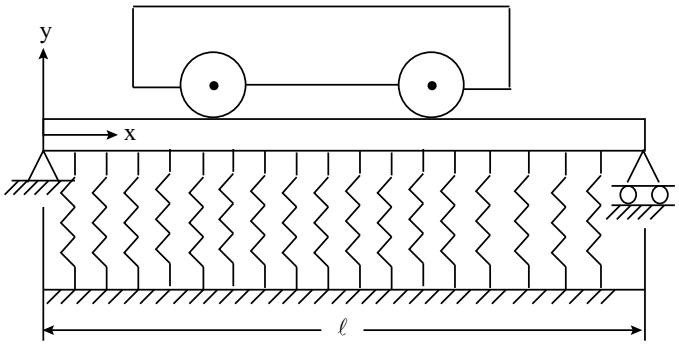


Figure 1.2.2: Schematic of a railway track supported on soil.

Boundary conditions depend upon how edges of the track are supported. For *simply supported edges*, we have

$$\begin{aligned} &\text{at } \; x = 0, \; w = 0, \; EI \, w'' = 0, \\ &\text{at } \; x = \ell, \; w = 0, \; EI \, w'' = 0, \end{aligned} \tag{1.2.3}$$

where $w' = dw/dx$. If the spring constant K of the soil depends upon w, then the mathematical model described by equations (1.2.1) and (1.2.3) involves a 4th-order nonlinear ordinary differential equation.

We now analyze transient heat conduction in a bar of uniform cross-section and assume that the diameter of the smallest circle enclosing a cross-section is much smaller than the length of the bar. It is thus reasonable to assume that the temperature varies only along the length of the bar. As depicted in Figure 1.2.3, we consider a free body diagram of an element of the bar of length Δx and denote the heat flux into the bar at the left end by q (q equals the rate of energy per unit area input into the bar) and that flowing out of it at the right end by $q + \Delta q$. If $\tilde{r}$ is the rate of thermal energy per unit volume of the bar received from the surroundings, ρ the mass density, c the specific heat, and θ the present temperature of the bar, then the balance of energy gives

$$\rho c \frac{\partial \theta}{\partial t} \Delta x = -(q + \Delta q) + q + \tilde{r} \Delta x, \tag{1.2.4}$$

or

$$\rho c \frac{\partial \theta}{\partial t} = -\frac{\partial q}{\partial x} + \tilde{r} \tag{1.2.5}$$

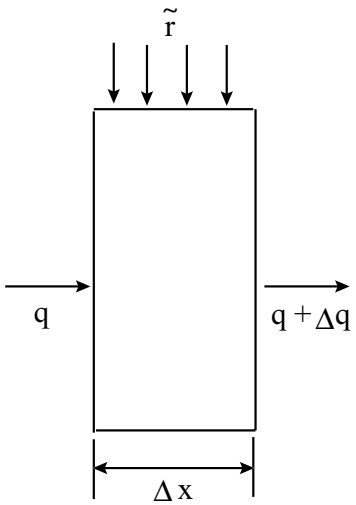


Figure 1.2.3: Free body diagram of an element of the bar showing heat flux at the two end faces and thermal energy received through radiation.

where t is time. Equation (1.2.5) has two unknowns in it, namely q and θ, and is supplemented by the following Fourier's law of heat conduction

$$q = -k\frac{\partial \theta}{\partial x}, \tag{1.2.6}$$

in which $k > 0$ is the thermal conductivity of the material of the bar. For a homogeneous bar k is independent of the location of a point on the bar, and for moderate changes in its temperature k can be assumed to be a constant. For an inhomogeneous bar k depends upon the location x of a point on the bar, and for large temperature changes k also varies with the temperature. Substitution for q from equation (1.2.6) into equation (1.2.5) gives

$$\rho c\frac{\partial \theta}{\partial t} = k\frac{\partial^2 \theta}{\partial x^2} + \tilde{r} \tag{1.2.7}$$

which is a 2nd-order inhomogeneous linear partial differential equation. Pertinent initial and boundary conditions are:

$$\begin{aligned} &\theta(x,0) = \theta_0(x), \\ &\theta(0,t) = \alpha(t),\ k\frac{d\theta}{dx}\bigg|_{x=\ell} = h(t). \end{aligned} \tag{1.2.8}$$

Equations (1.2.7) and (1.2.8) are a mathematical model of transient heat conduction in a homogeneous bar that obeys the Fourier law of heat conduction.

Even though there is only one space dimension, equation (1.2.7) governing the evolution of temperature is a partial differential equation because the temperature depends upon time t and the

space variable x. The order of the differential equation is determined by the highest order derivative present in it, the number of initial conditions equals the order of the highest time derivative, and the number of boundary conditions generally equals one-half of the highest spatial derivative present in the differential equation. *Mathematical models of most physical problems involve differential equations of even order.*

1.3 Classification of Boundary Conditions

Boundary conditions are classified as essential and natural. For a differential equation of order $2m$, boundary conditions involving derivatives of order at most $(m-1)$ are called *essential* and others are called *natural*. Essential boundary conditions are also sometimes called *kinematic* or *displacement* or *temperature* type, and natural boundary conditions *kinetic* or *traction* or *heat flux* type. If all boundary conditions are essential, then the boundary-value problem (BVP) is called *Dirichlet* and if all of them are natural, then it is called *Neumann*. Usually, a Neumann BVP does not have a unique solution, and two solutions of a BVP differ by a rigid body motion for a mechanical problem and a constant temperature for a thermal problem.

A BVP problem may have essential boundary conditions on a part of the boundary and natural boundary conditions on the rest of the boundary; such a problem is called mixed BVP.

1.4 Classification of Problems into Linear and Nonlinear

The BVP is called *linear* if the governing differential equation and the boundary conditions are linear in the unknown function or functions, and it is called *nonlinear* if at least one of these is nonlinear in the unknown function or functions. Thus for a nonlinear BVP either the differential equation or at least one of the boundary conditions or both are nonlinear in the unknown function. A linear IBV problem generally has a unique solution but a nonlinear IBV problem may have multiple solutions. For a nonhomogeneous differential equation

$$k\frac{d^2u}{dx^2} - \alpha u = 5,\ 0 < x < \ell, \tag{1.4.1}$$

where k and α are constants, we look at its homogeneous part to decide if it is linear or nonlinear. Examples of nonlinear differential equations are

$$\begin{aligned} k\frac{d^2u}{dx^2} - \alpha u^2 = 3,\ 0 < x < \ell, \\ k\frac{du}{dx}\frac{d^2u}{dx^2} - \alpha u = 5,\ 0 < x < \ell. \end{aligned} \tag{1.4.2}$$

1.5 Approximate Solution of a Problem

For both linear and nonlinear IBV problems, the FEM determines an approximate solution that very likely will not exactly satisfy either the given differential equation or side conditions such as the initial and the boundary conditions or both. The following questions arise:

(i) How good is the approximate solution and how do we determine the error?

(ii) How can we improve upon the accuracy of the numerical solution?

Following chapters address these questions.

Chapter 2: Mathematical Preliminaries

2.1 Summation Convention, Dummy Indices

Consider the sum

$$s = a_1x_1 + a_2x_2 + \ldots + a_nx_n \ . \tag{2.1.1}$$

We can write it in a compact form as

$$s = \sum_{i=1}^{n} a_ix_i = \sum_{j=1}^{n} a_jx_j = \sum_{m=1}^{n} a_mx_m \ . \tag{2.1.2}$$

It is obvious that the index $i,\ j$ or m in eqn. (2.1.2) is dummy in the sense that the sum is independent of the letter used. This is analogous to the dummy variable in an integral of a function over a finite interval:

$$I = \int_a^b f(x)dx = \int_a^b f(y)dy = \int_a^b f(t)dt \ . \tag{2.1.3}$$

Throughout this book we use "eqn." as an abbreviation for "equation". The three dots in the term on the right-hand side of eqn. (2.1.1) stand for the $(n-3)$ missing terms. *The common convention is to denote such missing terms by three, and not any other number of dots.* The first digit in an equation number stands for the Chapter, the second for the section and the third for the equation number in a particular section.

We can simplify the writing of eqn. (2.1.2) by adopting the following convention, sometimes called Einstein's summation convention. Whenever an index is repeated once in the same term, (i.e., it appears twice in the same term), it implies summation over the specified range of the index. Using the summation convention, eqn. (2.1.2) can be written as

$$s = a_ix_i = a_jx_j = a_mx_m, \tag{2.1.4}$$

where indices $i,\ j$ and m take values 1 through n. Note that expressions such as $a_ib_ix_i$ are *not* defined according to this convention. That is, an index should *never be repeated more than once* in the same term for the summation convention to be implied. Therefore, an expression of the form $\sum_{i=1}^{n} a_ib_ix_i$ must retain the summation sign.

In this Chapter, unless otherwise specified, we shall take n to be 3. Thus

$$\begin{aligned} a_ix_i = a_mx_m = a_1x_1 + a_2x_2 + a_3x_3 \ , \\ a_{ii} = a_{mm} = a_{11} + a_{22} + a_{33} \ . \end{aligned} \tag{2.1.5}$$

The summation convention can obviously be used to express a double sum, a triple sum, *etc*. For example, we can write $\sum_{i=1}^{3}\sum_{j=1}^{3} a_{ij}x_ix_j$ simply as $a_{ij}x_ix_j$. This expression equals the sum of nine terms:

$$\begin{aligned} a_{ij}x_ix_j &= a_{i1}x_ix_1 + a_{i2}x_ix_2 + a_{i3}x_ix_3 \ , \\ &= a_{11}x_1x_1 + a_{21}x_2x_1 + a_{31}x_3x_1 \\ &\quad + a_{12}x_1x_2 + a_{22}x_2x_2 + a_{32}x_3x_2 \\ &\quad + a_{13}x_1x_3 + a_{23}x_2x_3 + a_{33}x_3x_3 \ . \end{aligned} \tag{2.1.6}$$

Similarly, the triple sum $\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3} a_ib_jc_kx_ix_jx_k$ will simply be written as $a_ib_jc_kx_ix_jx_k$, and it represents the sum of 27 terms. It is again emphasized that expressions such as $a_{ii}x_ix_jx_j$ or $a_{ij}x_ix_jx_ix_j$ are *not* defined in the summation convention since the index i appears three times in $a_{ii}x_ix_jx_j$ and both i and j appear three times in $a_{ij}x_ix_jx_ix_j$.

2.2 Free Indices

Consider the following system of three equations:

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = a_{1i}x_i, \\ y_2 &= a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = a_{2i}x_i, \\ y_3 &= a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = a_{3i}x_i. \end{aligned} \tag{2.2.1}$$

These can be shortened to

$$y_i = a_{ij}x_j, \ i = 1, 2, 3. \tag{2.2.2}$$

An index which appears only once *in each term* of an equation such as the index i in eqn. (2.2.2) is called a "free index". A free index takes on the value 1, 2, or 3 *one at a time*. Thus eqn. (2.2.2) is a short way of writing three equations each having the sum of three terms on its right-hand side.

Note that the free index appearing in every term of an equation *must* be the same. Thus

$$a_i = b_j \tag{2.2.3}$$

is a *meaningless* equation. However, the following equations are meaningful.

$$\begin{aligned} a_i + k_i &= c_i, \\ a_i + b_ic_jd_j &= 0. \end{aligned} \tag{2.2.4}$$

If there are two free indices appearing in an equation such as

$$T_{ij} = A_{im}A_{jm},\ i = 1, 2, 3;\ j = 1, 2, 3; \tag{2.2.5}$$

then it is a short way of writing 9 equations. For example, eqn. (2.2.5) represents 9 equations; each one has the sum of 3 terms on the right-hand side. In fact

$$\begin{array}{l} T_{11} = A_{1m}A_{1m} = A_{11}A_{11} + A_{12}A_{12} + A_{13}A_{13}, \\ T_{12} = A_{1m}A_{2m} = A_{11}A_{21} + A_{12}A_{22} + A_{13}A_{23}, \\ T_{13} = A_{1m}A_{3m} = A_{11}A_{31} + A_{12}A_{32} + A_{13}A_{33}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ T_{33} = A_{3m}A_{3m} = A_{31}A_{31} + A_{32}A_{32} + A_{33}A_{33}. \end{array} \tag{2.2.6}$$

Again, equations such as

$$T_{ij} = T_{jk},\ T_{i\ell} = A_{im}A_{\ell\ell}, \tag{2.2.7}$$

are meaningless since in eqn. $(2.2.7)_1$, the left-hand side has free indices i and j and the right-hand side has free indices j and k. In eqn. $(2.2.7)_2$, the left-hand side has free index ℓ but the right-hand side has free index m. The subscript "1" in $(2.2.7)_1$ implies the first of the two equations in eqn. (2.2.7).

The nine quantities T_{ij} in eqn. (2.2.5) can be written as a 3×3 matrix

$$\begin{bmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{bmatrix}.$$

The transpose of a matrix is obtained by interchanging its rows and columns. The matrix form of eqn. (2.2.5) is

$$[T] = [A][A]^T \tag{2.2.8}$$

where $[T]$ is the 3×3 matrix T_{ij} and $[A]^T$ equals the transpose of the matrix $[A]$. Recall that the product $[A][B]$ of matrices $[A]$ and $[B]$ is defined as

$$([A][B])_{ij} = A_{ik}B_{kj}. \tag{2.2.9}$$

The summed index k in the expression on the right-hand side of eqn. (2.2.9) appears next to each other on matrices $[A]$ and $[B]$.

A matrix $[A]$ is also usually written as $\mathbf{A}$.

2.3 Kronecker Delta

The Kronecker delta, denoted by δ_{ij}, is defined as

$$\delta_{ij} = \begin{cases} 1 & \text{if} \quad i = j, \\ 0 & \text{if} \quad i \neq j. \end{cases} \tag{2.3.1}$$

That is,

$$\begin{aligned} &\delta_{11} = \delta_{22} = \delta_{33} = 1, \\ &\delta_{12} = \delta_{13} = \delta_{21} = \delta_{23} = \delta_{31} = \delta_{32} = 0. \end{aligned} \tag{2.3.2}$$

In other words, the matrix

$$\begin{bmatrix} \delta_{11} & \delta_{12} & \delta_{13} \\ \delta_{21} & \delta_{22} & \delta_{23} \\ \delta_{31} & \delta_{32} & \delta_{33} \end{bmatrix}$$

is the identity matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

often denoted by $[1]$. We note the following relations

$$\text{(a)} \quad \delta_{ii} = \delta_{11} + \delta_{22} + \delta_{33} = 1 + 1 + 1 = 3, \tag{2.3.3}$$

$$\begin{aligned} \text{(b)} \quad &\delta_{1m} a_m = \delta_{11} a_1 + \delta_{12} a_2 + \delta_{13} a_3 = a_1, \\ &\delta_{2m} a_m = \delta_{21} a_1 + \delta_{22} a_2 + \delta_{23} a_3 = a_2, \\ &\delta_{3m} a_m = \delta_{31} a_1 + \delta_{32} a_2 + \delta_{33} a_3 = a_3. \end{aligned} \tag{2.3.4}$$

Or, in general

$$\delta_{im} a_m = a_i. \tag{2.3.5}$$

Similarly, one can show that

$$\delta_{im} T_{mj} = T_{ij}. \tag{2.3.6}$$

In particular,

$$\delta_{im}\delta_{mj} = \delta_{ij}; \ \delta_{im}\delta_{mj}\delta_{jn} = \delta_{in}. \tag{2.3.7}$$

Eqn. (2.3.6) states that the product of the identity matrix with the matrix $[T]$ equals the matrix $[T]$, and eqn. (2.3.7) implies that the product of the identity matrix with itself equals the identity matrix. The identity matrix is also called the unit matrix.

2.4 Index Notation

Usually, rectangular Cartesian coordinates of a point are denoted by (x, y, z) and unit vectors along the x-, the y- and the z-axes by $\mathbf{i},\ \mathbf{j}$ and $\mathbf{k}$ respectively. In this coordinate system, components of a vector $\mathbf{u}$ along the x-, the y- and the z-axes are denoted by $u_x,\ u_y$ and u_z respectively. The vector $\mathbf{u}$ has the representation

$$\mathbf{u} = u_x\mathbf{i} + u_y\mathbf{j} + u_z\mathbf{k}. \tag{2.4.1}$$

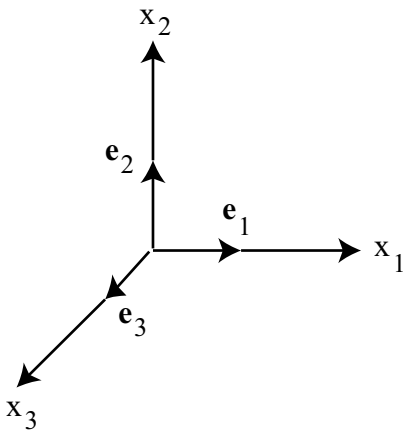


Fig. 2.4.1: Rectangular Cartesian coordinate axes

Throughout this Book we will usually denote a vector by a bold face lower case letter. The notation used in eqn. (2.4.1) does not lend itself to any abbreviation. Therefore, instead of denoting the coordinate axes by x, y, z we will denote them by x_1, x_2, x_3. Also we will denote unit vectors along x_1, x_2 and x_3 axes by $\mathbf{e}_1,\ \mathbf{e}_2$ and $\mathbf{e}_3$ respectively. Components of a vector $\mathbf{u}$ along $x_1,\ x_2$ and x_3 axes will be indicated by u_1, u_2 and u_3 respectively. Hence we can write

$$\begin{aligned} \mathbf{u} &= u_1\mathbf{e}_1 + u_2\mathbf{e}_2 + u_3\mathbf{e}_3, \\ &= u_j\mathbf{e}_j. \end{aligned} \tag{2.4.2}$$

Similarly,

$$\begin{aligned} \mathbf{v} &= v_1\mathbf{e}_1 + v_2\mathbf{e}_2 + v_3\mathbf{e}_3, \\ &= v_j\mathbf{e}_j. \end{aligned} \tag{2.4.3}$$

The dot or the inner or the scalar product, $\mathbf{u}\cdot\mathbf{v}$, between vectors $\mathbf{u}$ and $\mathbf{v}$ can simply be written as

$$\mathbf{u}\cdot\mathbf{v} = u_1v_1 + u_2v_2 + u_3v_3 = u_iv_i. \tag{2.4.4}$$

Since $\mathbf{e}_1,\ \mathbf{e}_2$ and $\mathbf{e}_3$ are mutually orthogonal unit vectors, therefore,

$$\begin{aligned}\mathbf{e}_1\cdot\mathbf{e}_1=\mathbf{e}_2\cdot\mathbf{e}_2=\mathbf{e}_3\cdot\mathbf{e}_3=1,\\ \mathbf{e}_1\cdot\mathbf{e}_2=\mathbf{e}_2\cdot\mathbf{e}_3=\mathbf{e}_3\cdot\mathbf{e}_1=0.\end{aligned} \tag{2.4.5}$$

These equations can be summarized as

$$\mathbf{e}_i\cdot\mathbf{e}_j=\delta_{ij}. \tag{2.4.6}$$

As another illustration of the use of the index notation, consider a line element with components $dx_1,\ dx_2,\ dx_3$. The square of the length, ds, of the line element is given by

$$\begin{aligned}ds^2&=dx_1^2+dx_2^2+dx_3^2,\\ &=dx_i dx_i,\\ &=\delta_{ij}dx_i dx_j.\end{aligned} \tag{2.4.7}$$

We have set $dx_i=\delta_{ij}dx_j$ (recall eqn. (2.3.5)) in going from eqn. $(2.4.7)_2$ to eqn. $(2.4.7)_3$.

Finally, we note that the differential, df, of a function $f(x_1,x_2,x_3)$ can be written as

$$\begin{aligned}df&=\frac{\partial f}{\partial x_1}dx_1+\frac{\partial f}{\partial x_2}dx_2+\frac{\partial f}{\partial x_3}dx_3,\\ &=\frac{\partial f}{\partial x_i}dx_i=f_{,i}dx_i,\end{aligned} \tag{2.4.8}$$

where $f_{,i}=\dfrac{\partial f}{\partial x_i}$; this notation for partial derivative of function f with respect to x_i is very often used in the Mechanics community.

2.5 Permutation Symbol

The permutation symbol, denoted by ϵ_{ijk}, is defined by

$$\epsilon_{ijk}=\left\{\begin{array}{r}1\\-1\\0\end{array}\right\}\text{ if }i,j,k\text{ form }\left\{\begin{array}{l}\text{an even}\\\text{an odd}\\\text{not a}\end{array}\right\}\text{permutation of 1, 2, 3.} \tag{2.5.1}$$

That is,

$$\begin{aligned}&\epsilon_{123}=\epsilon_{231}=\epsilon_{312}=1,\\ &\epsilon_{132}=\epsilon_{213}=\epsilon_{321}=-1,\\ &\epsilon_{111}=\epsilon_{211}=\epsilon_{133}=\ldots=0.\end{aligned} \tag{2.5.2}$$

We note that ϵ_{ijk} equals zero when any two of the three indices i, j, k have the same value, and

$$\epsilon_{ijk} = \epsilon_{jki} = \epsilon_{kij} = -\epsilon_{jik} = -\epsilon_{kji} = -\epsilon_{ikj}. \tag{2.5.3}$$

If $\mathbf{e}_1, \mathbf{e}_2$ and $\mathbf{e}_3$ form a right-handed triad of orthonormal vectors (i.e., mutually orthogonal unit vectors) then

$$\mathbf{e}_1 \times \mathbf{e}_2 = \mathbf{e}_3,\ \mathbf{e}_2 \times \mathbf{e}_3 = \mathbf{e}_1,\ etc., \tag{2.5.4}$$

which can be written as

$$\mathbf{e}_i \times \mathbf{e}_j = \epsilon_{ijk}\mathbf{e}_k. \tag{2.5.5}$$

Here $\mathbf{u} \times \mathbf{v}$ denotes the cross-product between vectors $\mathbf{u}$ and $\mathbf{v}$. Now, if $\mathbf{u} = u_i\mathbf{e}_i$ and $\mathbf{v} = v_i\mathbf{e}_i$, then

$$\begin{aligned} \mathbf{u} \times \mathbf{v} &= u_i\mathbf{e}_i \times v_j\mathbf{e}_j = u_iv_j\mathbf{e}_i \times \mathbf{e}_j, \\ &= u_iv_j\epsilon_{ijk}\mathbf{e}_k = \epsilon_{ijk}u_iv_j\mathbf{e}_k. \end{aligned} \tag{2.5.6}$$

The following useful identity, which can be verified by a long-hand calculation, should be memorized.

$$\epsilon_{ijm}\epsilon_{klm} = \delta_{ik}\delta_{jl} - \delta_{il}\delta_{jk}. \tag{2.5.7}$$

By using this identity, we prove the vector identity

$$\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) = (\mathbf{u} \cdot \mathbf{w})\mathbf{v} - (\mathbf{u} \cdot \mathbf{v})\mathbf{w}. \tag{2.5.8}$$

Proof: Let $\mathbf{v} \times \mathbf{w} = \mathbf{a}$. Then $\mathbf{a} = \epsilon_{ijk}v_iw_j\mathbf{e}_k$, and

$$\begin{aligned} &\mathbf{u} \times \mathbf{a} = \epsilon_{ijk}u_ia_j\mathbf{e}_k, \\ &a_j = \mathbf{a} \cdot \mathbf{e}_j = \epsilon_{ilm}v_iw_l\mathbf{e}_m \cdot \mathbf{e}_j, \\ &\qquad = \epsilon_{ilm}v_iw_l\delta_{mj} = \epsilon_{ilj}v_iw_l, \end{aligned} \tag{2.5.9}$$

where we have used $\mathbf{e}_m \cdot \mathbf{e}_j = \delta_{mj}$ and $\epsilon_{ilm}\delta_{mj} = \epsilon_{ilj}$. Thus

$$\begin{aligned} \mathbf{u} \times (\mathbf{v} \times \mathbf{w}) &= \epsilon_{ijk}u_i(\epsilon_{plj}v_pw_l)\mathbf{e}_k, \\ &= \epsilon_{ijk}\epsilon_{plj}u_iv_pw_l\mathbf{e}_k, \\ &= \epsilon_{kij}\epsilon_{plj}u_iv_pw_l\mathbf{e}_k, \end{aligned} \tag{2.5.10}$$

$$
\begin{aligned}
(\text{use eqn. } (2.5.7)) &= (\delta_{kp}\delta_{il} - \delta_{kl}\delta_{ip})u_i v_p w_l \mathbf{e}_k, \\
&= \delta_{kp}\delta_{il}u_i v_p w_l \mathbf{e}_k - \delta_{kl}\delta_{ip}u_i v_p w_l \mathbf{e}_k, \\
&= u_l w_l v_k \mathbf{e}_k - u_p v_p w_k \mathbf{e}_k, \\
(\text{use eqn. } (2.4.4)) &= (\mathbf{u}\cdot\mathbf{w})\mathbf{v} - (\mathbf{u}\cdot\mathbf{v})\mathbf{w}.
\end{aligned} \tag{2.5.11}
$$

We now write the determinant of a square (i.e., number of rows = number of columns) matrix $[A]$, denoted by $\det[A]$, in the index notation.

$$
\begin{aligned}
\det[A] &= \det \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}, \\
&= A_{11}(A_{22}A_{33} - A_{32}A_{23}) - A_{21}(A_{12}A_{33} - A_{32}A_{13}) + A_{31}(A_{12}A_{23} - A_{22}A_{13}), \\
&= A_{11}(\epsilon_{1jk}A_{j2}A_{k3}) - A_{21}(-\epsilon_{2jk}A_{j2}A_{k3}) + A_{31}(\epsilon_{3jk}A_{j2}A_{k3}), \\
&= A_{i1}\epsilon_{ijk}A_{j2}A_{k3}, \\
&= \epsilon_{ijk}A_{i1}A_{j2}A_{k3}.
\end{aligned} \tag{2.5.12}
$$

By following the same procedure, one can show that

$$
\det[A] = \epsilon_{ijk}A_{1i}A_{2j}A_{3k}. \tag{2.5.13}
$$

Recalling eqn. (2.5.2), one can conclude from eqns. (2.5.12) and (2.5.13) that $\det[A] = 0$ when either two rows or two columns of the matrix [A] are identical.

The permutation tensor or the permutation symbol is also called an alternating tensor or the Levi-Civita tensor.

A square matrix $[A]$ is called singular if $\det[A] = 0$; otherwise it is called non-singular. A non-singular matrix $[A]$ has an inverse $[A^{-1}]$ defined by

$$
[A][A^{-1}] = [A^{-1}][A] = [1]. \tag{2.5.14}
$$

2.6 Manipulations with the Indicial Notations

(a) Substitution

If

$$
a_i = v_{im}b_m, \tag{2.6.1}
$$

and

$$b_i = v_{im}c_m, \tag{2.6.2}$$

then, in order to substitute for b_i's from eqn. (2.6.2) into eqn. (2.6.1) we first change the dummy index in eqn. (2.6.2) from m to another letter, n, and then the free index in eqn. (2.6.2) from i to m, so that

$$b_m = v_{mn}c_n. \tag{2.6.3}$$

Now eqns. (2.6.1) and (2.6.3) give

$$a_i = v_{im}v_{mn}c_n. \tag{2.6.4}$$

Note that eqn. (2.6.4) represents three equations each having the sum of nine terms on its right-hand side since indices m and n are repeated.

(b) Multiplication

If

$$p = a_m b_m, \tag{2.6.5}$$

and

$$q = c_m d_m, \tag{2.6.6}$$

then

$$pq = a_m b_m c_n d_n. \tag{2.6.7}$$

It is important to note that $pq \neq a_m b_m c_m d_m$. In fact the right-hand side of this expression is not even defined in the summation convention since the index m appears four times, and further it is obvious that

$$pq \neq \sum_{m=1}^{3} a_m b_m c_m d_m.$$

(c) Factoring

If

$$T_{ij}n_j - \lambda n_i = 0, \tag{2.6.8}$$

then, using the Kronecker delta, we can write

$$n_i = \delta_{ij}n_j, \tag{2.6.9}$$

so that eqn. (2.6.8) becomes

$$T_{ij}n_j - \lambda\delta_{ij}n_j = 0. \tag{2.6.10}$$

Thus

$$(T_{ij} - \lambda\delta_{ij})n_j = 0. \tag{2.6.11}$$

(d) Contraction

The operation of setting two indices the same and so summing on them is known as contraction. For example, T_{ii} is the contraction of T_{ij},

$$T_{ii} = T_{11} + T_{22} + T_{33}, \tag{2.6.12}$$

and T_{ijj} is a contraction of T_{ijk},

$$T_{ijj} = T_{i11} + T_{i22} + T_{i33}. \tag{2.6.13}$$

Note that T_{iii} is *not* a contraction of T_{ijk}; T_{iii} is undefined since the index i appears three times. If

$$T_{ij} = \lambda E_{kk}\delta_{ij} + 2\mu E_{ij}, \tag{2.6.14}$$

then

$$\begin{aligned} T_{ii} &= \lambda E_{kk}\delta_{ii} + 2\mu E_{ii}, \\ &= (3\lambda + 2\mu)E_{kk}. \end{aligned} \tag{2.6.15}$$

The quantity T_{ii} is called the trace of T_{ij} and is denoted by $tr[T]$ or $tr\ \mathbf{T}$.

2.7 Translation and Rotation of Coordinate Axes

Consider two sets of rectangular Cartesian coordinate axes $O - x_1x_2$ and $O' - x_1'x_2'$ in a plane. If the set of coordinate axes $O' - x_1'x_2'$ is obtained from the set $O - x_1x_2$ by a shift of the origin and without a rotation of the axes, then, the transformation is a *translation*; for example see Fig. 2.7.1.

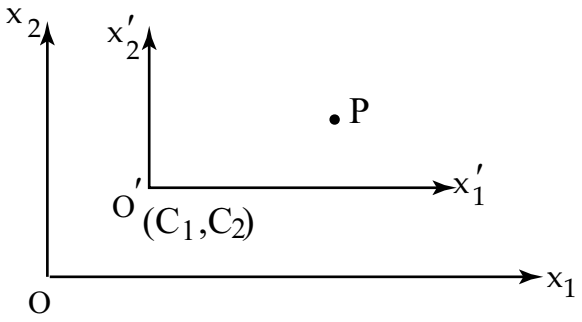


Fig. 2.7.1: Two rectangular Cartesian coordinate axes related by a translation.

If a point P has coordinates (x_1, x_2) and (x_1', x_2') with respect to $O - x_1x_2$ and $O' - x_1'x_2'$ respectively and (C_1, C_2) are the coordinates of O' with respect to $O - x_1x_2$, then

$$\begin{aligned} x_1 &= x_1' + C_1, \\ x_2 &= x_2' + C_2, \end{aligned} \tag{2.7.1}$$

or briefly

$$x_i = x_i' + C_i, \ i = 1, 2. \tag{2.7.2}$$

If the origin remains fixed, and the new axes $Ox_1', \ Ox_2'$ are obtained by rotating Ox_1 and Ox_2 through an angle θ in the counter-clockwise direction about a line perpendicular to the plane $O - x_1x_2$, then the transformation of axes, as illustrated in Fig. 2.7.2, is a *rotation*. Let the

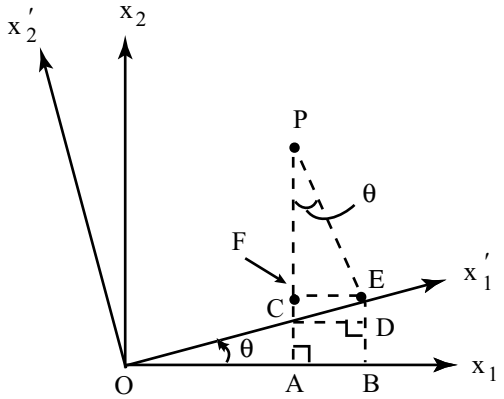


Fig. 2.7.2: Two rectangular Cartesian coordinate axes related by a rotation.

point P have coordinates (x_1, x_2) and (x_1', x_2') relative to $O - x_1x_2$ and $O - x_1'x_2'$ respectively. Then,

$$\begin{aligned} x_1 &= OA = OB - AB = OB - CD, \\ &= OE\cos\theta - PE\sin\theta, \\ &= x_1'\cos\theta - x_2'\sin\theta; \end{aligned} \tag{2.7.3}$$

$$\begin{aligned} x_2 &= AP = AF + FP = BE + FP, \\ &= x_1'\sin\theta + x_2'\cos\theta. \end{aligned} \tag{2.7.4}$$

We can write (x_1', x_2') in terms of (x_1, x_2) as

$$\begin{aligned} x_1' &= x_1\cos\theta + x_2\sin\theta, \\ x_2' &= -x_1\sin\theta + x_2\cos\theta. \end{aligned} \tag{2.7.5}$$

Using the index notation, the set of eqns. (2.7.5) can be written as

$$x_i' = a_{ij}x_j, \qquad i = 1, 2;\ j = 1, 2, \tag{2.7.6}$$

where a_{ij} are elements of the matrix $[a]$;

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}. \tag{2.7.7}$$

Before eqns. (2.7.1) and (2.7.5) are generalized to three dimensions, we give below an alternate method of arriving at eqn. (2.7.5). Let $\mathbf{e}_1'$ and $\mathbf{e}_2'$ denote unit vectors along x_1'- and x_2'-axes respectively, and $\mathbf{e}_1$- and $\mathbf{e}_2$-unit vectors along x_1- and x_2-axes respectively. Then the vector $\mathbf{OP}$ going from point O to point P can be written as

$$\begin{aligned} \mathbf{OP} &= x_1\mathbf{e}_1 + x_2\mathbf{e}_2, \\ &= x_1'\mathbf{e}_1' + x_2'\mathbf{e}_2'. \end{aligned} \tag{2.7.8}$$

Also

$$\begin{aligned} \mathbf{e}_1' &= \cos\theta\ \mathbf{e}_1 + \sin\theta\ \mathbf{e}_2, \\ \mathbf{e}_2' &= -\sin\theta\ \mathbf{e}_1 + \cos\theta\ \mathbf{e}_2. \end{aligned} \tag{2.7.9}$$

Therefore,

$$\begin{aligned} x_1' &= \mathbf{OP}\cdot\mathbf{e}_1', \\ &= (x_1\mathbf{e}_1 + x_2\mathbf{e}_2)\cdot(\cos\theta\mathbf{e}_1 + \sin\theta\mathbf{e}_2), \\ &= x_1\cos\theta + x_2\sin\theta; \end{aligned} \tag{2.7.10}$$

$$\begin{aligned} x_2' &= \mathbf{OP}\cdot\mathbf{e}_2', \\ &= -x_1\sin\theta + x_2\cos\theta. \end{aligned} \tag{2.7.11}$$

This latter approach can more easily be adopted to the 3-dimensional case. If the primed axes $O - x_1'x_2'x_3'$ are obtained from the unprimed axes $O - x_1x_2x_3$ just by a translation, then the coordinates of a point with respect to the two sets of axes are related by eqn. (2.7.2) with the index i ranging from 1 to 3. Now let the primed axes be obtained from the unprimed axes by a rotation only. We denote unit vectors along the x_1-, the x_2- and the x_3-axes by $\mathbf{e}_1, \mathbf{e}_2$ and $\mathbf{e}_3$ respectively and those along the x_1'-, the x_2'- and the x_3'-axes by $\mathbf{e}_1',\ \mathbf{e}_2'$ and $\mathbf{e}_3'$ respectively. If

$$a_{1j} = \text{cosine of the angle between } \mathbf{e}_1' \text{ and } \mathbf{e}_j, \tag{2.7.12}$$

then $a_{11},\ a_{12}$ and a_{13} are the direction cosines (or components) of $\mathbf{e}_1'$ with respect to the unprimed axes. We can write

$$\begin{aligned}\mathbf{e}_1' &= a_{11}\mathbf{e}_1 + a_{12}\mathbf{e}_2 + a_{13}\mathbf{e}_3,\\ &= a_{1i}\mathbf{e}_i.\end{aligned} \tag{2.7.13}$$

Similarly,

$$\mathbf{e}_2' = a_{2i}\mathbf{e}_i,\ \mathbf{e}_3' = a_{3i}\mathbf{e}_i. \tag{2.7.14}$$

Eqns. (2.7.13) and (2.7.14) can be written as

$$\mathbf{e}_i' = a_{ij}\mathbf{e}_j. \tag{2.7.15}$$

Note that the matrix $[a]$ in eqn. (2.7.15) is 3×3. Since

$$\mathbf{e}_i' \cdot \mathbf{e}_j' = \delta_{ij}, \tag{2.7.16}$$

therefore,

$$\begin{aligned}\delta_{ij} &= a_{ik}\mathbf{e}_k \cdot a_{jp}\mathbf{e}_p = a_{ik}a_{jp}\mathbf{e}_k \cdot \mathbf{e}_p,\\ &= a_{ik}a_{jp}\delta_{kp} = a_{ik}a_{jk}.\end{aligned} \tag{2.7.17}$$

Eqns. (2.7.17) are equivalent to the following six equations.

$$\begin{aligned}&a_{11}^2 + a_{12}^2 + a_{13}^2 = 1,\\ &a_{21}^2 + a_{22}^2 + a_{23}^2 = 1,\\ &a_{31}^2 + a_{32}^2 + a_{33}^2 = 1,\\ &a_{11}a_{21} + a_{12}a_{22} + a_{13}a_{23} = 0,\\ &a_{21}a_{31} + a_{22}a_{32} + a_{23}a_{33} = 0,\\ &a_{31}a_{11} + a_{32}a_{12} + a_{33}a_{13} = 0.\end{aligned} \tag{2.7.18}$$

Equations $(2.7.18)_1$, $(2.7.18)_2$ and $(2.7.18)_3$ imply that $\mathbf{e}_1',\ \mathbf{e}_2'$ and $\mathbf{e}_3'$ are unit vectors; equations $(2.7.18)_4$, $(2.7.18)_5$ and $(2.7.18)_6$ state that $\mathbf{e}_1'$, $\mathbf{e}_2'$ and $\mathbf{e}_3'$ are mutually orthogonal. Of course, we can write $\mathbf{e}_i$'s in terms of $\mathbf{e}_i'$'s. Since

$$a_{j1} = \text{cosine of the angle between } \mathbf{e}_1 \text{ and } \mathbf{e}_j', \tag{2.7.19}$$

therefore,

$$\mathbf{e}_1 = a_{j1}\mathbf{e}_j' \text{ or } \mathbf{e}_i = a_{ji}\mathbf{e}_j'. \tag{2.7.20}$$

From the point of view of the solution of a set of simultaneous linear equations, the matrix a_{ji} in eqn. (2.7.20) must be identified as the *inverse* of the matrix a_{ij}:

$$[a_{ij}]^{-1} = [a_{ji}] = [a_{ij}]^T. \tag{2.7.21}$$

Here $[a_{ij}]^T$ is the transpose of the matrix $[a_{ij}]$ and is obtained from it by interchanging rows and columns. A matrix $[a_{ij}]$ that satisfies eqn. (2.7.21) is called an *orthogonal* matrix. That is, the transpose of an orthogonal matrix equals its inverse. A transformation is said to be orthogonal if the matrix associated with it is orthogonal. The matrix $[a]$ in eqn. (2.7.15) defining a rotation of coordinate axes is orthogonal. For an orthogonal matrix

$$[a][a]^T = 1. \tag{2.7.22}$$

Therefore

$$\begin{aligned} & \det([a][a]^T) = 1, \\ \text{or} \quad & \det[a]\det[a]^T = 1, \\ \text{or} \quad & \det[a]\det[a] = 1, \end{aligned} \tag{2.7.23}$$

and thus

$$\det[a] = \pm 1. \tag{2.7.24}$$

An orthogonal matrix whose determinant equals $+1$ is called a *proper* orthogonal matrix and one whose determinant equals -1 is called an *improper* orthogonal matrix. A proper orthogonal matrix transforms a right-handed set of axes into a right-handed set of axes whereas an improper orthogonal matrix transforms a right-handed set of axes into a left-handed set of axes or vice-a-versa.

Consider a vector $\mathbf{OP}$ emanating from the origin O and ending at a point P. With respect to the primed and the unprimed axes,

$$\begin{aligned} \mathbf{OP} &= x_1'\mathbf{e}_1' + x_2'\mathbf{e}_2' + x_3'\mathbf{e}_3', \\ &= x_j'\mathbf{e}_j', \\ &= x_j\mathbf{e}_j. \end{aligned} \tag{2.7.25}$$

Recall that

$$\begin{aligned} x_i' &= \mathbf{OP}\cdot\mathbf{e}_i', \\ &= (x_j\mathbf{e}_j)\cdot(a_{ik}\mathbf{e}_k), \\ &= x_j a_{ik}\delta_{jk}, \\ &= a_{ij}x_j. \end{aligned} \tag{2.7.26}$$

Example: The components of a vector $\mathbf{A}$ with respect to unprimed axes are $A_i = \delta_{i2} + \delta_{i3}$. Consider a set of primed coordinate axes obtained by rotating the unprimed axes through an angle of 30° about the x_3-axis (see Fig. 2.7.3). What are components, A_i', of this vector with respect to the primed axes?

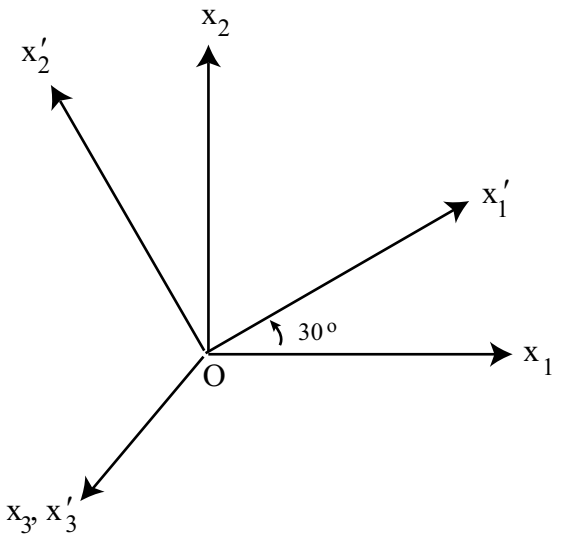


Fig. 2.7.3: Two rectangular Cartesian coordinate axes related by a rotation about the x_3-axis.

Solution:

$$\mathbf{e}_3' = \mathbf{e}_3;\ \mathbf{e}_1' = \cos 30^\circ\ \mathbf{e}_1 + \sin 30^\circ\ \mathbf{e}_2,$$
$$\mathbf{e}_2' = -\sin 30^\circ\ \mathbf{e}_1 + \cos 30^\circ\ \mathbf{e}_2.$$

Therefore

$$[a_{ij}] = [\mathbf{e}_i'\cdot\mathbf{e}_j] = \begin{bmatrix} \cos 30^\circ & \sin 30^\circ & 0 \\ -\sin 30^\circ & \cos 30^\circ & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The components A_i' given by

$$A_i' = a_{ij}A_j,$$

can be written as

$$\left\{ \begin{array}{c} A_1' \\ A_2' \\ A_3' \end{array} \right\} = \left[\begin{array}{ccc} \cos 30^\circ & \sin 30^\circ & 0 \\ -\sin 30^\circ & \cos 30^\circ & 0 \\ 0 & 0 & 1 \end{array} \right] \left\{ \begin{array}{c} 0 \\ 1 \\ 1 \end{array} \right\} = \left\{ \begin{array}{c} \sin 30^\circ \\ \cos 30^\circ \\ 1 \end{array} \right\}.$$

Hence

$$A_i' = 0.5\delta_{i1} + 0.866\delta_{i2} + \delta_{i3}.$$

Summarizing our discussion of the transformation of coordinate axes, we note that a general transformation from unprimed to primed axes combines a translation and a rotation of the axes. This can be written as

$$x_i' = a_{ij}x_j + c_i, \tag{2.7.27}$$

where $[a]$ is an orthogonal matrix and c_i is a constant. Under this transformation, the components of a vector $\mathbf{A}$ in the two sets of axes are related as

$$A_i' = a_{ij}A_j, \text{ or } \{A'\} = [a]\{A\}. \tag{2.7.28}$$

Here we have written the three components, A_1, A_2 and A_3, of the vector $\mathbf{A}$ as a column matrix. To save space, we will sometimes also write it as

$$\mathbf{A} = (A_1,\ A_2,\ A_3). \tag{2.7.29}$$

Under an orthogonal transformation the length of a line element and the angle between any two lines do not change.

2.8 Tensors

2.8.1 A linear transformation

Let $\mathbf{T}$ be a transformation from a vector space into the same vector space. That is, for any vector $\mathbf{u}$, $\mathbf{Tu}$ is also a vector of the same dimensions as $\mathbf{u}$. Then $\mathbf{T}$ is linear if and only if

$$\mathbf{T}(\alpha\mathbf{u} + \beta\mathbf{w}) = \alpha\mathbf{Tu} + \beta\mathbf{Tw} \tag{2.8.1}$$

for all real numbers α and β. [**NOTE:** f is a linear function of x if and only if $f(x) = \alpha x$ where α is a real number. For example, $f(x) = 2x + 3$ is *not* a linear function of x even though it is often referred to as such; $f(x) = 2x + 3$ is an *affine* function of x, or a polynomial of degree one in x.]

A linear transformation from a vector space into another vector space is also called a **second-order** tensor.

2.8.2 Tensor product between two vectors

The tensor product (or the dyadic product) $\otimes$ between vectors $\mathbf{a}$ and $\mathbf{b}$ is defined as

$$(\mathbf{a} \otimes \mathbf{b})\mathbf{c} = (\mathbf{b} \cdot \mathbf{c})\mathbf{a} \tag{2.8.2}$$

for every vector $\mathbf{c}$. Thus $(\mathbf{a}\otimes\mathbf{b})$ transforms a vector $\mathbf{c}$ into a vector parallel to $\mathbf{a}$. Since it transforms a vector into a vector and satisfies eqn. (2.8.1), it is a linear transformation. Note that

$$\mathbf{a} \otimes \mathbf{b} \neq \mathbf{b} \otimes \mathbf{a}. \tag{2.8.3}$$

The inner product $\mathbf{a} \cdot \mathbf{b}$ between vectors $\mathbf{a}$ and $\mathbf{b}$ is a scalar, the cross product $\mathbf{a} \times \mathbf{b}$ between vectors $\mathbf{a}$ and $\mathbf{b}$ is a vector perpendicular to the plane of $\mathbf{a}$ and $\mathbf{b}$, and the tensor product $\mathbf{a} \otimes \mathbf{b}$ is a second-order tensor. Note that for $\mathbf{a} = a_i\mathbf{e}_i$,

$$\mathbf{a} \otimes \mathbf{b} = a_i b_j \mathbf{e}_i \otimes \mathbf{e}_j. \tag{2.8.4}$$

2.8.3 Components of a second-order tensor

Let $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ be a set of orthonormal base vectors (i.e., $\mathbf{e}_1, \mathbf{e}_2$ and $\mathbf{e}_3$ are mutually orthogonal unit vectors). For any vector $\mathbf{a}$,

$$\mathbf{a} = a_i\mathbf{e}_i. \tag{2.8.5}$$

Let $\mathbf{b} = \mathbf{T}\mathbf{a}$. Then

$$\mathbf{b} = \mathbf{T}(a_j\mathbf{e}_j) = a_j(\mathbf{T}\mathbf{e}_j), \tag{2.8.6}$$

or

$$b_i\mathbf{e}_i = a_j(\mathbf{T}\mathbf{e}_j). \tag{2.8.7}$$

The inner product of both sides of eqn. (2.8.6) or eqn. (2.8.7) with $\mathbf{e}_k$ gives

$$b_k = \mathbf{e}_k \cdot a_j(\mathbf{T}\mathbf{e}_j) = a_j\mathbf{e}_k \cdot (\mathbf{T}\mathbf{e}_j) = a_j T_{kj}, \tag{2.8.8}$$

where

$$T_{kj} = \mathbf{e}_k \cdot (\mathbf{T}\mathbf{e}_j), \tag{2.8.9}$$

is the component of $\mathbf{T}$ with respect to the bases $\mathbf{e}_i$. For computational purposes, the three equations in eqn. (2.8.8) are written as

$$\left\{ \begin{array}{c} b_1 \\ b_2 \\ b_3 \end{array} \right\} = \left[\begin{array}{ccc} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{array} \right] \left\{ \begin{array}{c} a_1 \\ a_2 \\ a_3 \end{array} \right\}. \tag{2.8.10}$$

Analogous to the representation (2.8.5) for vector $\mathbf{a}$, we have the following representation for the second-order tensor $\mathbf{T}$:

$$\mathbf{T} = T_{ij}\mathbf{e}_i \otimes \mathbf{e}_j. \tag{2.8.11}$$

Because of inequality (2.8.3), T_{ij} need not equal T_{ji}. In order to see that eqn. (2.8.11) is equivalent to eqn. (2.8.8), we evaluate $\mathbf{Ta}$.

$$\begin{aligned}\mathbf{b} = \mathbf{Ta} &= (T_{ij}\mathbf{e}_i \otimes \mathbf{e}_j)(a_k\mathbf{e}_k),\\ &= T_{ij}a_k(\mathbf{e}_i \otimes \mathbf{e}_j)\mathbf{e}_k,\\ &= T_{ij}a_k\mathbf{e}_i(\mathbf{e}_j \cdot \mathbf{e}_k),\\ &= T_{ij}a_k\mathbf{e}_i\delta_{jk} = T_{ij}a_j\mathbf{e}_i,\end{aligned} \tag{2.8.12}$$

which is equivalent to $b_i = T_{ij}a_j$ or eqn. (2.8.8).

It is clear from eqn. (2.8.11) that the components T_{ij} of $\mathbf{T}$ depend upon the choice of bases $\mathbf{e}_i$. Let

$$\mathbf{e}'_i = Q_{ij}\mathbf{e}_j, \tag{2.8.13}$$

where $\mathbf{Q}$ is an orthogonal matrix (i.e. $\mathbf{QQ}^T = \mathbf{1}$, where $\mathbf{1}$ is the identity matrix or the identity tensor and $\mathbf{QQ}^T$ equals the product of matrices $[Q]$ and $[Q]^T$). Then

$$\begin{aligned}\mathbf{T} = T'_{ij}\mathbf{e}'_i \otimes \mathbf{e}'_j &= T'_{ij}(Q_{ik}\mathbf{e}_k) \otimes (Q_{jl}\mathbf{e}_l),\\ &= T'_{ij}Q_{ik}Q_{jl}(\mathbf{e}_k \otimes \mathbf{e}_l).\end{aligned} \tag{2.8.14}$$

Equating the two expressions for $\mathbf{T}$ given in eqns. (2.8.11) and (2.8.14), we obtain

$$T_{kl} = T'_{ij}Q_{ik}Q_{jl}, \tag{2.8.15}$$

and in matrix notation,

$$[T] = [Q]^T[T'][Q], \tag{2.8.16}$$

and since $\mathbf{Q}$ is orthogonal,

$$[T'] = [Q][T][Q]^T. \tag{2.8.17}$$

The transpose $\mathbf{T}^T$ of a second-order tensor $\mathbf{T}$ is defined by

$$\mathbf{a} \cdot (\mathbf{T}^T\mathbf{b}) = \mathbf{b} \cdot (\mathbf{Ta}) \tag{2.8.18}$$

for every vector $\mathbf{a}$ and $\mathbf{b}$. The components of $\mathbf{T}$ and $\mathbf{T}^T$ are related by

$$(T^T)_{ij} = T_{ji}. \tag{2.8.19}$$

A second-order tensor $\mathbf{T}$ is said to be symmetric if

$$\mathbf{T} = \mathbf{T}^T \text{ or } T_{ij} = T_{ji}, \tag{2.8.20}$$

and it is skew-symmetric or antisymmetric if

$$\mathbf{T} = -\mathbf{T}^T \text{ or } T_{ij} = -T_{ji}. \tag{2.8.21}$$

Thus a symmetric second-order tensor has six independent components, and a skew-symmetric second-order tensor has three independent components. All diagonal components of a skew-symmetric second-order tensor identically vanish, i.e., $T_{11} = T_{22} = T_{33} = 0$. Bases for a symmetric second-order tensor and a skew-symmetric second-order tensor are $(\mathbf{e}_i \otimes \mathbf{e}_j + \mathbf{e}_j \otimes \mathbf{e}_i)/2$ and $(\mathbf{e}_i \otimes \mathbf{e}_j - \mathbf{e}_j \otimes \mathbf{e}_i)/2$ respectively.

Since a skew-symmetric second-order tensor $\mathbf{T}^a$ has only three independent components, it can also be written as a vector $\mathbf{w}$ given by

$$w_i = \frac{1}{2}\epsilon_{ijk}T^a_{jk}. \tag{2.8.22}$$

Equivalently

$$T^a_{ij} = \epsilon_{ijk}w_k. \tag{2.8.23}$$

Furthermore

$$\mathbf{T}^a\mathbf{b} = \mathbf{b} \times \mathbf{w}. \tag{2.8.24}$$

A second-order tensor $\mathbf{T}$ has a unique additive decomposition into its symmetric part $\mathbf{T}^s$ and skew-symmetric part $\mathbf{T}^a$:

$$\mathbf{T} = \mathbf{T}^s + \mathbf{T}^a, \tag{2.8.25}$$

where

$$\mathbf{T}^s = (\mathbf{T} + \mathbf{T}^T)/2, \ \mathbf{T}^a = (\mathbf{T} - \mathbf{T}^T)/2. \tag{2.8.26}$$

2.8.4 Tensors of higher-order

A third-order tensor is a linear transformation from the space of second-order tensors into vectors or vectors into second-order tensors, and can be represented as

$$\mathbf{A} = A_{ijk}\mathbf{e}_i \otimes \mathbf{e}_j \otimes \mathbf{e}_k. \tag{2.8.27}$$

Under a change of bases given by eqn. (2.8.13), the transformation rule for its components can be derived as follows:

$$\begin{aligned} \mathbf{A} &= A'_{ijk}\mathbf{e}'_i \otimes \mathbf{e}'_j \otimes \mathbf{e}'_k, \\ &= A'_{ijk}Q_{il}\mathbf{e}_l \otimes Q_{jm}\mathbf{e}_m \otimes Q_{kn}\mathbf{e}_n, \\ &= A'_{ijk}Q_{il}Q_{jm}Q_{kn}\mathbf{e}_l \otimes \mathbf{e}_m \otimes \mathbf{e}_n. \end{aligned} \tag{2.8.28}$$

From eqns. (2.8.27) and (2.8.28) we conclude that

$$A_{lmn} = A'_{ijk}Q_{il}Q_{jm}Q_{kn}, \tag{2.8.29}$$

or

$$A'_{ijk} = Q_{il}Q_{jm}Q_{kn}A_{lmn}. \tag{2.8.30}$$

A fourth-order tensor is a linear transformation from the space of second-order tensors to second-order tensors, and has the representation

$$\mathbf{C} = C_{ijkl}\mathbf{e}_i \otimes \mathbf{e}_j \otimes \mathbf{e}_k \otimes \mathbf{e}_l. \tag{2.8.31}$$

It has eighty-one components. Under the transformation (2.8.13) of base vectors, its components will transform as

$$C'_{ijkl} = Q_{ip}Q_{jq}Q_{kr}Q_{ls}C_{pqrs}. \tag{2.8.32}$$

A fourth-order tensor is symmetric if it maps symmetric second-order tensors into symmetric second-order tensors. Thus for a symmetric fourth-order tensor

$$C_{ijkl} = C_{jikl} = C_{ijlk}. \tag{2.8.33}$$

It has 36 independent components, and can be written as a 6×6 matrix. The bases for a fourth-order symmetric tensor are

$$\frac{1}{2}(\mathbf{e}_i \otimes \mathbf{e}_j + \mathbf{e}_j \otimes \mathbf{e}_i) \otimes \frac{1}{2}(\mathbf{e}_k \otimes \mathbf{e}_l + \mathbf{e}_l \otimes \mathbf{e}_k). \tag{2.8.34}$$

The correspondence between the components of a fourth-order symmetric tensor $\mathbf{C}$ and the equivalent 6×6 matrix $[E]$ is given below.

$$\begin{aligned}
&E_{11} = C_{1111},\ E_{12} = C_{1122},\ E_{13} = C_{1133},\ E_{14} = C_{1123},\ E_{15} = C_{1131},\ E_{16} = C_{1112},\\
&E_{21} = C_{2211},\ E_{22} = C_{2222},\ E_{23} = C_{2233},\ E_{24} = C_{2223},\ E_{25} = C_{2231},\ E_{26} = C_{2212},\\
&E_{31} = C_{3311},\ E_{32} = C_{3322},\ E_{33} = C_{3333},\ E_{34} = C_{3323},\ E_{35} = C_{3331},\ E_{36} = C_{3312},\\
&E_{41} = C_{2311},\ E_{42} = C_{2322},\ E_{43} = C_{2333},\ E_{44} = C_{2323},\ E_{45} = C_{2331},\ E_{46} = C_{2312},\\
&E_{51} = C_{3111},\ E_{52} = C_{3122},\ E_{53} = C_{3133},\ E_{54} = C_{3123},\ E_{55} = C_{3131},\ E_{56} = C_{3112},\\
&E_{61} = C_{1211},\ E_{62} = C_{1222},\ E_{63} = C_{1233},\ E_{64} = C_{1223},\ E_{65} = C_{1231},\ C_{66} = C_{1212}.
\end{aligned} \tag{2.8.35}$$

The correspondence between the two indices on the 6×6 matrix $[E]$ and the four indices on $\mathbf{C}$ is as follows:

$$1 \to 11,\ 2 \to 22,\ 3 \to 33,\ 4 \to 23,\ 5 \to 31,\ 6 \to 12. \tag{2.8.36}$$

Note that under a change of bases from $\mathbf{e}$ to $\mathbf{e}'$, components of the 6×6 matrix $[E]$ *do not* transform as eqn. (2.8.17). They can be found by first determining $\mathbf{C}'$ from eqn. (2.8.32) and then using the identification (2.8.36).

If in addition to the symmetries (2.8.33) $\mathbf{C}$ also satisfies

$$C_{ijkl} = C_{klij}, \tag{2.8.37}$$

then the 6×6 matrix $[E]$ is symmetric. Both $\mathbf{C}$ and the matrix $[E]$ have twenty-one independent components.

For a fourth-order symmetric tensor $\mathbf{C}$ and a second-order skew symmetric tensor $\mathbf{T}^a$,

$$C_{ijkl} T^a_{kl} = 0, \tag{2.8.38}$$

and

$$C_{ijkl} T_{kl} = C_{ijkl} T^s_{kl}, \tag{2.8.39}$$

where $T^s_{kl} = (T_{kl} + T_{lk})/2$ equals the symmetric part of T_{kl} as given by eqn. $(2.8.26)_1$.

In direct notation, $C_{ijkl}T_{kl}$ is often written as $\mathbf{C} : \mathbf{T}$ or simply as $\mathbf{CT}$.

2.8.5 Isotropic tensors

A tensor is called isotropic if its components with respect to every set of orthonormal base vectors are the same. Thus, for an isotropic vector $\mathbf{v}$, a second-order tensor $\mathbf{T}$, a third-order tensor $\mathbf{A}$, and a fourth-order tensor $\mathbf{C}$,

$$v_i = Q_{ij} v_j, \tag{2.8.40}$$

$$T_{ij} = Q_{ik} Q_{jl} T_{kl}, \tag{2.8.41}$$

$$A_{ijk} = Q_{il} Q_{jm} Q_{kn} A_{lmn}, \tag{2.8.42}$$

$$C_{ijkl} = Q_{ip} Q_{jq} Q_{kr} Q_{ls} C_{pqrs}, \tag{2.8.43}$$

for every orthogonal matrix $[Q]$. Eqns. (2.8.40), (2.8.41) and (2.8.43) are satisfied if and only if

$$\mathbf{v} = \mathbf{0},\ T_{ij} = \alpha \delta_{ij},\ C_{ijkl} = \lambda \delta_{ij} \delta_{kl} + \mu \delta_{ik} \delta_{jl} + \gamma \delta_{il} \delta_{jk}, \tag{2.8.44}$$

where $\alpha,\ \lambda,\ \mu$ and γ are constants. If eqn. (2.8.42) is to hold for both proper and improper orthogonal matrices $[Q]$ then its only solution is $\mathbf{A} = \mathbf{0}$. However, if matrix $[Q]$ is restricted to be a proper orthogonal matrix, then eqn. (2.8.42) has the solution

$$A_{ijk} = \alpha \epsilon_{ijk}, \tag{2.8.45}$$

where α is a constant. Thus the only isotropic vector is the null vector, an isotropic second-order tensor is a scalar multiple of the Kronecker delta, an isotropic third-order tensor under proper orthogonal transformations is the permutation tensor, an isotropic third-order tensor under both proper and improper orthogonal transformations is a null tensor, and bases for a fourth-order isotropic tensor are the tensor products of Kronecker deltas.

2.8.6 Eigenvalues and eigenvectors

Eigenvalues and eigenvectors are defined for an even-order tensor. A unit vector $\mathbf{a}$ is an eigenvector of a second-order tensor $\mathbf{T}$ if and only if there exists a scalar α such that

$$\mathbf{T}\mathbf{a} = \alpha \mathbf{a} \text{ or } T_{ij} a_j = \alpha a_i. \tag{2.8.46}$$

The scalar α is called the eigenvalue corresponding to the eigenvector $\mathbf{a}$. Regarding eqn. (2.8.46) as three linear simultaneous equations for $a_1,\ a_2$ and a_3, the necessary and sufficient condition for

α to be an eigenvalue of $\mathbf{T}$ is that

$$\det[\mathbf{T} - \alpha\mathbf{1}] = 0. \tag{2.8.47}$$

The left-hand side of eqn. (2.8.47) is a cubic polynomial in α. Thus eqn. (2.8.47) has either three real roots or one real and two complex conjugate roots. For a symmetric $\mathbf{T}$, all three roots of eqn. (2.8.47) are real. When $\mathbf{T}$ is symmetric and positive definite (i.e., $T_{ij}b_ib_j > 0$ for every $\mathbf{b} \neq \mathbf{0}$), then all three roots of eqn. (2.8.47) are positive. For each root $\alpha^{(1)},\ \alpha^{(2)}$ and $\alpha^{(3)}$ of eqn. (2.8.47), one can solve any two of the following three eqns.

$$T_{ij}a_j^{(1)} = \alpha^{(1)}a_i^{(1)}, \tag{2.8.48}$$

and $a_i^{(1)}a_i^{(1)} = 1$ for $a_1^{(1)},\ a_2^{(1)}$ and $a_3^{(1)}$.

A second-order tensor $\mathbf{B}$ is an eigenvector of a fourth-order tensor $\mathbf{C}$ if there exists a scalar λ such that

$$C_{ijkl}B_{kl} = \lambda B_{ij}. \tag{2.8.49}$$

λ is called an eigenvalue of $\mathbf{C}$ corresponding to the eigenvector $\mathbf{B}$. Similar to the identification (2.8.37), we set

$$1 \to 11,\ 2 \to 22,\ 3 \to 33,\ 4 \to 23,\ 5 \to 31,\ 6 \to 12,\ 7 \to 32,\ 8 \to 13,\ 9 \to 21. \tag{2.8.50}$$

Thus eqn. (2.8.49) can be written as

$$E_{\alpha\beta}b_\beta = \lambda b_\alpha,\ \alpha, \beta = 1, 2, \ldots, 9, \tag{2.8.51}$$

where the 9×9 matrix $[E]$ is equivalent to the fourth-order tensor $\mathbf{C}$. The nine linear simultaneous eqns. (2.8.51) have a nontrivial solution if and only if

$$\det[E_{\alpha\beta} - \lambda\delta_{\alpha\beta}] = 0. \tag{2.8.52}$$

The left-hand side of eqn. (2.8.52) is a polynomial of degree nine. Therefore, eqn. (2.8.52) has nine roots of which one must be real and the remaining eight may occur as four pairs of complex conjugate roots. For a symmetric $\mathbf{C}$, i.e., $C_{ijkl} = C_{klij}$, $E_{\alpha\beta} = E_{\beta\alpha}$ and all nine roots of eqn. (2.8.52) and hence all nine eigenvalues of $\mathbf{C}$ are real. In addition to being symmetric, if $\mathbf{C}$ is also positive-definite (i.e., $C_{ijkl}A_{ij}A_{kl} > 0$ for every non-zero $\mathbf{A}$), then all nine eigenvalues of $\mathbf{C}$ are positive.

If a symmetric $\mathbf{C}$ also satisfies eqn. (2.8.33) then an eigenvector $\mathbf{B}$ of the fourth-order tensor $\mathbf{C}$ is a symmetric second-order tensor. With the correspondence of indices given in eqn. (2.8.35), we get eqn. (2.8.52) with $\alpha, \beta = 1, 2, \ldots, 6$. Thus $\mathbf{C}$ has only six real eigenvalues and eigenvectors.

2.8.7 Magnitude of a tensor

The magnitude (or the length) of a tensor $\mathbf{v}$ of order one, $\mathbf{T}$ of order two, $\mathbf{A}$ of order three, and $\mathbf{C}$ of order four is defined as follows:

$$\begin{aligned} |\mathbf{v}| &= |\mathbf{v}\cdot\mathbf{v}|^{1/2} = (v_i v_i)^{1/2}, \\ |\mathbf{T}| &= |\mathbf{T}\cdot\mathbf{T}|^{1/2} = (tr(\mathbf{T}\mathbf{T}^T))^{1/2} = (T_{ij}T_{ij})^{1/2}, \\ |\mathbf{A}| &= |\mathbf{A}\cdot\mathbf{A}|^{1/2} = (A_{ijk}A_{ijk})^{1/2}, \\ |\mathbf{C}| &= |\mathbf{C}\cdot\mathbf{C}|^{1/2} = (C_{ijkl}C_{ijkl})^{1/2}. \end{aligned} \tag{2.8.53}$$

With the correspondence (2.8.50) between indices, one can see the equivalence between the definitions of lengths of second-order and fourth-order tensors.

One can introduce other definitions of the length of a tensor. In a finite-dimensional space, however, all of these are equivalent to one another.

2.8.8 Invariants of a second-order tensor

Rewriting eqn. (2.8.17) in the index notation as

$$T'_{ij} = Q_{ik}Q_{jl}T_{kl}, \tag{2.8.54}$$

and contracting indices i and j, we arrive at

$$T'_{ii} = Q_{ik}Q_{il}T_{kl} = \delta_{kl}T_{kl} = T_{kk}, \tag{2.8.55}$$

where we have used the orthogonality of matrix $[Q]$. Thus

$$\mathrm{tr}(\mathbf{T}) = \mathrm{tr}(\mathbf{T}'). \tag{2.8.56}$$

Taking the determinant of both sides of eqn. (2.8.17) and recalling that the determinant of the product of two matrices equals the product of their determinants, we obtain

$$\det[T'] = \det[Q]\det[T]\det[Q]^T = \det[T] \tag{2.8.57}$$

since $\det[Q] = \det[Q]^T = \pm 1$. One can easily show that

$$\mathbf{T}'^2 = \mathbf{Q}\mathbf{T}^2\mathbf{Q}^T, \tag{2.8.58}$$

which implies that $\mathrm{tr}((\mathbf{T}')^2) = \mathrm{tr}(\mathbf{T}^2)$. Hence $\mathrm{tr}(\mathbf{T})$, $\mathrm{tr}(\mathbf{T}^2)$ and $\det[T]$ have the same value in any two orthonormal coordinate axes. Because of this property, they are called principal invariants of $\mathbf{T}$.

Eqn. (2.8.47) can be written as

$$\alpha^3 - I_T\alpha^2 + II_T\alpha - III_T = 0, \tag{2.8.59}$$

where

$$\begin{aligned} I_T &= \mathrm{tr}(\mathbf{T}), \\ II_T &= \frac{1}{2}[(\mathrm{tr}(\mathbf{T}))^2 - \mathrm{tr}(\mathbf{T}^2)], \\ III_T &= \det[T]. \end{aligned} \tag{2.8.60}$$

From the preceding discussion it is clear that I_T, II_T and III_T also have the same value in every orthonormal coordinate axes, and are also principal invariants of $\mathbf{T}$. However, in a three-dimensional space, there are only three linearly independent invariants of $\mathbf{T}$. The invariants

$$i_T = \mathrm{tr}(\mathbf{T}),\ ii_T = \mathrm{tr}(\mathbf{T}^2) \text{ and } iii_T = \mathrm{tr}(\mathbf{T}^3), \tag{2.8.61}$$

are related to those given in eqns. (2.8.60) as follows.

$$i_T = I_T,\ ii_T = I_T^2 - 2II_T,\ iii_T = 3III_T + I_T^3 - 3I_TII_T. \tag{2.8.62}$$

Equation (2.8.59) is usually called the characteristic equation of the second-order tensor $\mathbf{T}$.

2.8.9 Hamilton-Cayley theorem

The Hamilton-Cayley theorem states that a matrix $\mathbf{T}$ satisfies its own characteristic equation. That is,

$$\mathbf{T}^3 - I_T\mathbf{T}^2 + II_T\mathbf{T} - III_T\mathbf{1} = \mathbf{0}. \tag{2.8.62}$$

For a non-singular matrix $\mathbf{T}$, i.e., $III_T \neq 0$, eqn. (2.8.62) can be solved for $\mathbf{T}^{-1}$ to obtain

$$\mathbf{T}^{-1} = (\mathbf{T}^2 - I_T\mathbf{T} + II_T\mathbf{1})/III_T. \tag{2.8.63}$$

Taking the trace of both sides of eqn. (2.8.63) and using eqn. $(2.8.60)_2$ we obtain

$$II_T = III_T \, tr(\mathbf{T}^{-1}). \tag{2.8.64}$$

2.9 The Divergence Theorem

The Divergence Theorem is also known as Gauss's Theorem or Green's Theorem. In a three-dimensional space, let a body occupy the region Ω bounded by the smooth or piecewise smooth surface $\partial\Omega$. Let $\mathbf{F}$ be a smooth (i.e., differentiable) vector field defined on Ω, and $\mathbf{n}$ denote the outward unit normal to $\partial\Omega$. Then

$$\int_{\partial\Omega} \mathbf{F}\cdot\mathbf{n} dA = \int_{\Omega} (\text{div } \mathbf{F}) dV, \tag{2.9.1}$$

where $dA\mathbf{n}$ is the vector area element on $\partial\Omega$, dV the volume element in Ω, and $\text{div } \mathbf{F}$ is the divergence of $\mathbf{F}$. In rectangular Cartesian coordinates, $\text{div } \mathbf{F} = F_{i,i} = \partial F_i/\partial x_i$. Thus, the divergence theorem transforms a surface integral to a volume integral. Its counterpart in the two-dimensional space transforms a line integral to an area integral.

For $\mathbf{F}$ equal to a vector field $\mathbf{v}$ or a second-order tensor field $\mathbf{T}$, eqn. (2.9.1) in components form becomes

$$\int_{\partial\Omega} v_i n_i dA = \int_{\Omega} v_{i,i} dV, \tag{2.9.2}$$

$$\int_{\partial\Omega} T_{ij} n_j dA = \int_{\Omega} T_{ij,j} dV. \tag{2.9.3}$$

2.10 Differentiation of Tensor Fields

A tensor $\mathbf{T}$ defined on a region Ω of either the one- or the two- or the three-dimensional space is called a tensor field. Differentiation of $\mathbf{T}$ with respect to the coordinate x_i involves differentiating each component of $\mathbf{T}$ with respect to x_i and results in a tensor field of order one greater than the order of $\mathbf{T}$. For a second-order tensor $\mathbf{T}$ the operations grad, div and curl of $\mathbf{T}$ are defined as:

$$(\text{grad } \mathbf{T})_{ijk} = \frac{\partial T_{ij}}{\partial x_k} = T_{ij,k}, \tag{2.10.1}$$

$$(\text{div } \mathbf{T})_i = \frac{\partial T_{ij}}{\partial x_j} = T_{ij,j}, \tag{2.10.2}$$

$$(\text{curl } \mathbf{T})_{ij} = \epsilon_{ilm} T_{jl,m}. \tag{2.10.3}$$

The gradient of a second-order tensor is a third-order tensor, and the divergence of a second-order tensor is a vector. The curl of a second-order tensor is a second-order tensor with respect to proper orthogonal transformations of the coordinate axes and is defined only in a three-dimensional space.

2.11 Cylindrical Coordinates

We take the z-axis of the cylindrical coordinate system (r, θ, z) coincident with the x_3-axis of the rectangluar Cartesian coordinate axes x_1, x_2, x_3, and denote unit vectors along the r and θ directions by $\mathbf{e}_r$ and $\mathbf{e}_\theta$ respectively. Referring to Fig. 2.11.1, we have

$$\begin{aligned} \mathbf{e}_r &= \cos\theta\, \mathbf{e}_1 + \sin\theta\, \mathbf{e}_2, \\ \mathbf{e}_\theta &= -\sin\theta\, \mathbf{e}_1 + \cos\theta\, \mathbf{e}_2, \end{aligned} \tag{2.11.1}$$

or equivalently

$$\begin{aligned} \mathbf{e}_1 &= \cos\theta\, \mathbf{e}_r - \sin\theta\, \mathbf{e}_\theta, \\ \mathbf{e}_2 &= \sin\theta\, \mathbf{e}_r + \cos\theta\, \mathbf{e}_\theta. \end{aligned} \tag{2.11.2}$$

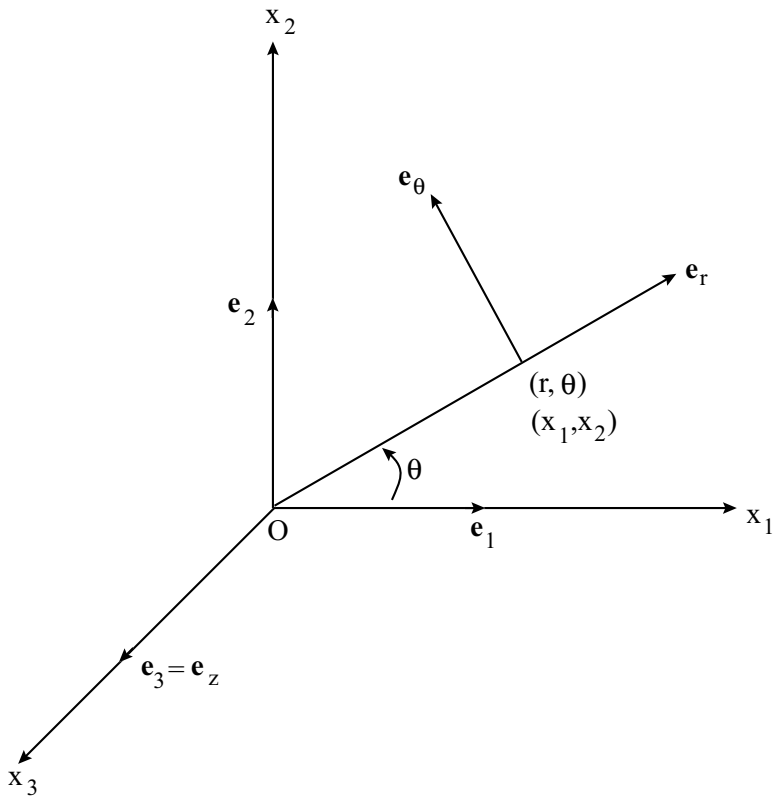


Fig. 2.11.1: Cartesian (x_1, x_2) and cylindrical (r, θ) coordinates of a point in the plane $x_3 = z = \text{constant}$.

Whereas directions of unit vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ do not change with the coordinates (x_1, x_2, x_3) of a point, those of unit vectors $\mathbf{e}_r$ and $\mathbf{e}_\theta$ depend upon the angular position θ of a point. Differentiation of equations (2.11.1) with respect to θ yields

$$\frac{\partial \mathbf{e}_r}{\partial \theta} = \mathbf{e}_\theta, \ \frac{\partial \mathbf{e}_\theta}{\partial \theta} = -\mathbf{e}_r. \tag{2.11.3}$$

We now evaluate the gradient of a scalar function and the divergence of a vector function in cylindrical coordinates. It follows from Fig. 2.11.1 that

$$\begin{aligned} x_1 &= r\cos\theta, \ x_2 = r\sin\theta \\ r^2 &= x_1^2 + x_2^2, \ \theta = \tan^{-1}\left(\frac{x_2}{x_1}\right). \end{aligned} \tag{2.11.4}$$

Thus

$$\text{grad}\, f(x_1, x_2, x_3) = \frac{\partial f}{\partial x_1}\mathbf{e}_1 + \frac{\partial f}{\partial x_2}\mathbf{e}_2 + \frac{\partial f}{\partial x_3}\mathbf{e}_3. \tag{2.11.5}$$

Using the chain rule of calculus, we get

$$\begin{aligned} \frac{\partial f}{\partial x_1} &= \frac{\partial f}{\partial r}\,\frac{\partial r}{\partial x_1} + \frac{\partial f}{\partial \theta}\,\frac{\partial \theta}{\partial x_1}, \\ &= \frac{\partial f}{\partial r}\,\frac{x_1}{r} + \frac{\partial f}{\partial \theta}\left(-\frac{x_2}{r^2}\right), \\ &= \frac{\partial f}{\partial r}\,\cos\theta - \frac{\partial f}{\partial \theta}\,\frac{\sin\theta}{r}; \end{aligned} \tag{2.11.6}$$

$$\begin{aligned} \frac{\partial f}{\partial x_2} &= \frac{\partial f}{\partial r}\,\frac{\partial r}{\partial x_2} + \frac{\partial f}{\partial \theta}\,\frac{\partial \theta}{\partial x_2}, \\ &= \frac{\partial f}{\partial r}\,\frac{x_2}{r} + \frac{\partial f}{\partial \theta}\,\frac{x_1}{r}. \end{aligned} \tag{2.11.7}$$

Substitution for $\dfrac{\partial f}{\partial x_1}$ and $\dfrac{\partial f}{\partial x_2}$ from eqns. (2.11.6) and (2.11.7), and for $\mathbf{e}_1$ and $\mathbf{e}_2$ from eqn. (2.11.2) into eqn. (2.11.5) gives

$$\text{grad}\, f = \frac{\partial f}{\partial r}\mathbf{e}_r + \frac{1}{r}\,\frac{\partial f}{\partial \theta}\mathbf{e}_\theta + \frac{\partial f}{\partial z}\mathbf{e}_z. \tag{2.11.8}$$

Thus

$$\text{grad} = \mathbf{e}_r\frac{\partial}{\partial_r} + \frac{\mathbf{e}_\theta}{r}\,\frac{\partial}{\partial \theta} + \mathbf{e}_z\frac{\partial}{\partial z}. \tag{2.11.9}$$

Recalling that

$$\begin{aligned}
\operatorname{div}\mathbf{v} &= \operatorname{grad}\cdot\mathbf{v},\\
&= \left(\mathbf{e}_r\frac{\partial}{\partial r}+\frac{\mathbf{e}_\theta}{r}\frac{\partial}{\partial\theta}+\mathbf{e}_z\frac{\partial}{\partial z}\right)\cdot\left(v_r\mathbf{e}_r+v_\theta\mathbf{e}_\theta+v_z\mathbf{e}_z\right),\\
&= \mathbf{e}_r\cdot\left(\frac{\partial v_r}{\partial r}\mathbf{e}_r+v_r\frac{\partial\mathbf{e}_r}{\partial r}+\frac{\partial v_\theta}{\partial r}\mathbf{e}_\theta+v_\theta\frac{\partial\mathbf{e}_\theta}{\partial r}+\frac{\partial v_z}{\partial r}\mathbf{e}_z+v_z\frac{\partial\mathbf{e}_z}{\partial r}\right)\\
&\quad+\frac{\mathbf{e}_\theta}{r}\cdot\left(\frac{\partial v_r}{\partial\theta}\mathbf{e}_r+v_r\frac{\partial\mathbf{e}_r}{\partial\theta}+\frac{\partial v_\theta}{\partial\theta}\mathbf{e}_\theta+v_\theta\frac{\partial\mathbf{e}_\theta}{\partial\theta}+\frac{\partial v_z}{\partial\theta}\mathbf{e}_z+v_z\frac{\partial\mathbf{e}_z}{\partial\theta}\right)\\
&\quad+\mathbf{e}_z\cdot\left(\frac{\partial v_r}{\partial z}\mathbf{e}_r+v_r\frac{\partial\mathbf{e}_r}{\partial z}+\frac{\partial v_\theta}{\partial z}\mathbf{e}_\theta+v_\theta\frac{\partial\mathbf{e}_\theta}{\partial z}+\frac{\partial v_z}{\partial z}\mathbf{e}_z+v_z\frac{\partial\mathbf{e}_z}{\partial z}\right),\\
&= \frac{\partial v_r}{\partial r}+\frac{v_r}{r}+\frac{1}{r}\frac{\partial v_\theta}{\partial\theta}+\frac{\partial v_z}{\partial z},
\end{aligned}\tag{2.11.10}$$

where we have used eqns. (2.11.3) and

$$\frac{\partial\mathbf{e}_r}{\partial r}=\mathbf{0},\ \frac{\partial\mathbf{e}_z}{\partial r}=\mathbf{0},\ \frac{\partial\mathbf{e}_\theta}{\partial r}=\mathbf{0},\ \frac{\partial\mathbf{e}_z}{\partial\theta}=\mathbf{0},\ \frac{\partial\mathbf{e}_r}{\partial z}=\frac{\partial\mathbf{e}_\theta}{\partial z}=\frac{\partial\mathbf{e}_z}{\partial z}=\mathbf{0}.\tag{2.11.11}$$

2.12 Linearly Independent Functions

A set S of functions $\phi_1(x),\ \phi_2(x),\ldots,\phi_n(x)$ defined on $[0,\ell]$ is said to be linearly independent if and only if the equation

$$C_1\phi_1(x)+C_2\phi_2(x)+\ldots+C_n\phi_n(x)=0\ \textbf{for every}\ x\in[0,\ell]\tag{2.12.1}$$

has a trivial solution $C_1=C_2=\ldots=C_n=0$. In order to find whether or not the given set of functions is linearly independent, we can select n different points in the interval $[0,\ell]$ and substitute them in eqn. (2.12.1). The result is n equations in n unknowns, namely, $C_1,C_2,\ldots,C_n$. These equations will have a trivial solution if and only if

$$\det\begin{bmatrix}\phi_1(x_1) & \phi_2(x_1) & \ldots & \phi_n(x_1)\\ \phi_1(x_2) & \phi_2(x_2) & \ldots & \phi_n(x_2)\\ \vdots & & & \\ \phi_1(x_n) & \phi_2(x_n) & \ldots & \phi_n(x_n)\end{bmatrix}\neq 0.\tag{2.12.2}$$

We should add that inequality (2.12.2) needs to be satisfied for *every* choice of n-points $x_1,x_2,\ldots,x_n$ in $[0,\ell]$. For linearly independent functions, the inequality (2.12.2) may be violated at a countable number of points in $[0,\ell]$. An alternative is the following and assumes that functions $\phi_1,\phi_2,\ldots,\phi_n$ are differentiable. Differentiating both sides of eqn. (2.12.1) with respect to x repeatedly upto

$(n-1)$ times, we obtain

$$
\begin{aligned}
&C_1\phi_1'(x) + C_2\phi_2'(x) + \ldots + C_n\phi_n'(x) = 0,\\
&C_1\phi_1''(x) + C_2\phi_2''(x) + \ldots + C_n\phi_n''(x) = 0,\\
&\vdots\\
&C_1\phi_1^{n-1}(x) + C_2\phi_2^{n-1}(x) + \ldots + C_n\phi_n^{n-1}(x) = 0.
\end{aligned}
\tag{2.12.3}
$$

Equation (2.12.1) and $(n-1)$ eqns. (2.12.3) constitute a set of n homogeneous equations for the n-unknowns $C_1, C_2, \ldots, C_n$. They will have a trivial solution, i.e., $C_1 = C_2 = \ldots = C_n = 0$ if and only if

$$
\det \begin{bmatrix} \phi_1 & \phi_2 & \ldots & \phi_n \\ \phi_1' & \phi_2' & \ldots & \phi_n' \\ \vdots & & & \\ \phi_1^{n-1} & \phi_2^{n-1} & \ldots & \phi_n^{n-1} \end{bmatrix} \neq 0 \tag{2.12.4}
$$

at *every* point in $[0, \ell]$. The determinant in (2.12.4) is usually referred to as the *Wronskian*.

The importance of the word "every" in the definition (2.12.1) of linearly independent functions, and in the necessary and sufficient conditions (2.12.2) and (2.12.4) can not be under-estimated. For example, functions $\sin n\pi x,\ n = 1, 2, \ldots$, defined on $[0, 1]$ are linearly independent even though the condition (2.12.1) is satisfied at $x = 0$ for non-trivial choices of $C_1, C_2, \ldots, C_n$. Similarly functions $x, x^2, x^3, \ldots, x^{10}$ defined on $[-1, 1]$ are linearly independent even though conditions (2.12.2) and (2.12.4) are violated at $x = 0$. Other examples of linearly independent functions defined on [0,1] are $e^x, e^{2x}, e^{3x} \ldots$; Legendre polynomials $L_0(x), L_1(x), L_2(x), \ldots$; and Lagrange polynomials $P_0(x), P_1(x), P_2(x), \ldots, P_n(x)$.

If a set of n functions defined on an interval is not linearly independent, then it is said to be linearly dependent. In a set of *linearly dependent* functions, at least one function from the set can be expressed as a *linear* combination of the remaining functions in the set.

The maximum number of linearly independent functions in the set S equals the *dimensionality* of S. Note that all functions in the set S are defined on the same domain. Even though functions 1, $1/x$ and $1/x^2$ defined on $[1, 2]$ are linearly independent, they are not linearly independent on the domain $[-1, 1]$ since $1/x^2$ and $1/x$ are not defined at $x = 0$. Note that there are infinitely many linearly independent functions defined on a given domain; e.g. $\sin \pi x,\ \sin 2\pi x, \ldots$; $1,\ \cos \pi x,\ \cos 2\pi x,\ \cos 3\pi x, \ldots$; and $1,\ e^x,\ e^{2x}, \ldots$; defined on $[0, 1]$ are linearly independent.

Reference

R. M. Bowen and C.-C. Wang, Introduction to Vectors and Tensors, Linear and Multilinear Algebra, Vol. 1, Plenum Press, New York, 1976.

Exercises:

2.1 Given $[a_{ij}] = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 2 \\ 3 & 0 & 3 \end{bmatrix}$.
Evaluate (a) a_{ii}, (b) $a_{ij}a_{ij}$, and (c) $a_{jk}a_{kj}$.

2.2 Given $[b_{ij}] = \begin{bmatrix} 3 & 2 & 4 \\ 1 & 5 & 3 \\ 5 & 7 & 9 \end{bmatrix}$.
Evaluate (a) $c_{ij} = b_{ik}b_{kj}$, (b) $d_{ij} = b_{ik}b_{jk}$, (c) $e_{ij} = b_{ki}b_{kj}$, (d) c_{ii}, (e) d_{ii}, (f) e_{ii}.

2.3 Using the index notation, write expressions for

(1) the magnitude of a vector $\mathbf{u}$,

(2) $\cos\theta;\ \theta$ being the angle between vectors $\mathbf{u}$ and $\mathbf{v}$.

2.4 Using the index notation write an expression for $|\sin\theta|$ where θ is the angle between vectors $\mathbf{u}$ and $\mathbf{v}$.

2.5 Show that

(a) $\epsilon_{ijk}\epsilon_{jki} = 6$,

(b) $\epsilon_{ijk}A_jA_k = 0$.

2.6 Show that

(a) If $\epsilon_{ijk}T_{jk} = 0$, then $T_{ij} = T_{ji}$,

(b) $\epsilon_{ilm}\epsilon_{jlm} = 2\delta_{ij}$, and

(c) if $T_{ij} = -T_{ji}$, then $T_{ij}a_ia_j = 0$.

2.7 Show that $\epsilon_{ijk}A_{il}A_{jm}A_{kn} = (\det[A])\epsilon_{lmn}$.

2.8 Given that $T_{ij} = \lambda E_{kk}\delta_{ij} + 2\mu E_{ij}$, $W = \frac{1}{2}T_{ij}E_{ij},\ P = T_{ij}T_{ij}$, show that

$$W = \mu E_{ij}E_{ij} + \frac{\lambda}{2}(E_{kk})^2,$$
$$P = 4\mu^2 E_{ij}E_{ij} + \lambda(E_{kk})^2(4\mu + 3\lambda).$$

2.9 Consider a cube formed by the orthonormal vectors $\mathbf{e}_1',\ \mathbf{e}_2'$ and $\mathbf{e}_3'$, where $\mathbf{e}_i' = a_{ij}\mathbf{e}_j$. By setting the volume of this cube equal to 1, show that $\det[a] = 1$.

2.10 Show that the tensor $B_{ij} = \epsilon_{ijk} v_k$ is skew symmetric, i.e., $B_{ij} = -B_{ji}$.

2.11 Let $\sigma_{ij} = \lambda e_{kk}\delta_{ij} + 2\mu e_{ij}$ where λ and μ are positive constants. Solve for e_{ij} in terms of σ_{ij}. Note that no term involving e_{ij} should appear on the right-hand side of the equation

$$e_{ij} = \ldots .$$

2.12 Under a change of orthonormal bases

$$\mathbf{e}'_i = a_{ij}\mathbf{e}_j,$$

the components of a second-order tensor $\mathbf{T}$ transform as

$$T'_{ij} = a_{ik}a_{jl}T_{kl} \ .$$

Show that

(i) $T'_{ii} = T_{kk}$,

(ii) $T'_{ij}T'_{ji} = T_{kl}T_{lk}$,

(iii) $\det[T'] = \det[T]$.

2.13 Under a change of orthonormal bases

$$\mathbf{e}'_i = a_{ij}\mathbf{e}_j,$$

show that the components of a fourth-order tensor $\mathbf{D}$ transform as

$$D'_{ijkl} = a_{ip}a_{jq}a_{kr}a_{ls}D_{pqrs}.$$

2.14 Show that for a second-order tensor $\mathbf{A}$,

$$\det[A] = \frac{1}{6}[(tr\ \mathbf{A})^3 + 2tr(\mathbf{A}^3) - 3(tr(\mathbf{A}))(tr(\mathbf{A}^2))].$$

2.15 Let the orthogonal matrix $[Q(t)]$ be a differential function of t. Show that the matrix $\left[\frac{dQ}{dt}\right][Q]^T$ is skew-symmetric.

Chapter 3: Weak Formulation of a Model Problem

3.1 Problem Statement

We study steady state heat conduction in a bar and assume that its left end is kept at a fixed temperature, the heat flux is prescribed at the right end, and $\tilde{r} = -4k\theta - 2k$ in eqn. (1.2.7). The BV problem to be solved is

$$\theta'' - 4\theta = 2,\ 0 < x < 1, \tag{3.1.1}$$

$$\theta(0) = 0,\ \theta'\big|_{x=1} = 3. \tag{3.1.2}$$

where $\theta' = d\theta/dx$. Equation (3.1.1) is a second-order $(2m = 2)$ linear ordinary differential equation. Boundary condition $(3.1.2)_1$ involving zeroth order $(\leq (m-1))$ derivative is essential, and boundary condition $(3.1.2)_2$ containing first-order derivative is natural.

3.2 Approximate Solution

Let $\tilde{\theta}$ be an approximate solution of the problem satisfying the essential boundary condition $(3.1.2)_1$. Even though it is not absolutely necessary we usually require that the approximate solution satisfy exactly the essential boundary conditions. We call $\tilde{\theta}$ the trial solution, and set

$$\tilde{\theta}'' - 4\tilde{\theta} - 2 = r(x),\ 0 < x < 1, \tag{3.2.1}$$

$$\tilde{\theta}'\big|_{x=1} - 3 = \bar{r}. \tag{3.2.2}$$

We wish to find $\tilde{\theta}$ so that the function $r(x)$ is essentially zero on $(0,1)$ and $\bar{r}$ is as close to zero as possible.

In order for the term $\tilde{\theta}''$ to make a contribution we should choose our trial solution $\tilde{\theta}$ so that $\tilde{\theta}''$ is non-zero on $(0,1)$. The simplest choice for $\tilde{\theta}$ that satisfies the essential boundary condition $(3.1.2)_1$ is

$$\tilde{\theta}(x) = bx^2,\ 0 < x < 1, \tag{3.2.3}$$

where b is a constant to be determined. However, unless required by the essential boundary conditions, we include all powers of x less than or equal to the minimum needed. Thus we take the trial solution to be

$$\tilde{\theta}(x) = ax + bx^2,\ 0 < x < 1, \tag{3.2.4}$$

where a and b are constants to be determined. These can be found by any one of the following methods.

3.2.1 Collocation Method

In order to find the two constants a and b in eqn. (3.2.4) we need two linearly independent equations. There are several possibilities. One of them is to require that $\bar{r} = 0$ and $r\left(\frac{1}{2}\right) = 0$. We thus obtain

$$\begin{aligned} a + 2b - 3 &= 0, \\ 2b - 4\left(\frac{a}{2} + \frac{b}{4}\right) - 2 &= 0, \end{aligned} \tag{3.2.5}$$

and the approximate solution is

$$\tilde{\theta} = \frac{x}{5}(-1 + 8x). \tag{3.2.6}$$

The exact or the analytical solution of the BV problem defined by eqns. (3.1.1) and (3.1.2) is

$$\theta_{\text{anal}} = \frac{1 + 3e^2}{2(e^4 + 1)} e^{2x} + \frac{e^2(e^2 - 3)}{2(e^4 + 1)} e^{-2x} - \frac{1}{2}. \tag{3.2.7}$$

There are several ways to define the error in the approximate solution; of these three are listed below.

$$\begin{aligned} \|e\|_{\text{sup}} &= \sup_{x \in (0,1)} \left[|\theta_{\text{anal}}(x) - \tilde{\theta}(x)| / |\tilde{\theta}_{\text{anal}}(x)| \right], \\ \|e\|_0 &= \left[\int_0^1 (\theta_{\text{anal}}(x) - \tilde{\theta}(x))^2 dx \Big/ \int_0^1 \theta_{\text{anal}}^2(x) dx \right]^{1/2}, \\ \|e\|_1 &= \left[\frac{\int_0^1 (\theta_{\text{anal}}(x) - \tilde{\theta}(x))^2 dx}{\int_0^1 (\theta_{\text{anal}}(x))^2 dx} + \frac{\int_0^1 (\theta'_{\text{anal}}(x) - \tilde{\theta}'(x))^2 dx}{\int_0^1 (\theta'_{\text{anal}}(x))^2 dx} \right]^{1/2}. \end{aligned} \tag{3.2.8}$$

Whereas the error $\|e\|_{\text{sup}}$ gives the maximum relative difference between the analytical and the approximate solution over the interval $(0, 1)$, the error $\|e\|_0$ determines the average difference between the analytical and the approximate solutions. If $\|e\|_{\text{sup}} \ll 1$, then $\|e\|_0 \ll 1$ but the converse need not be true. The error measures $\|e\|_0$ and $\|e\|_{\text{sup}}$ compare the analytical and the approximate solutions. The error $\|e\|_1$ also compares the first derivatives of the two solutions. Thus $\|e\|_1 \ll 1$ implies $\|e\|_0 \ll 1$ but the converse may not hold. Similarly $\|e\|_{\text{sup}} \ll 1$ does not imply that $\|e\|_1 \ll 1$. Thus whether or not the approximate solution is close to the analytical

solution depends upon the quantities of interest. If the temperature is the primary quantity of interest then $\|e\|_{\sup}$ or $\|e\|_0$ will suffice to measure the error in the approximate solution. However, if the heat flux is also desired to be accurate, then one should measure the error in the approximate solution with $\|e\|_1$. Since $\theta(0) = 0$, one cannot compute $\|e\|_0$ on $[0,\ 1]$. One may also need to find $\|e_0$ over the region $[\epsilon,\ 1]$ where $0 < \epsilon \ll 1$, and find the limiting value of $\|e\|_0$ as $\epsilon \to 0$.

Having found the error in the approximate solution, how do we reduce it or said differently how to improve the quality of the approximate solution? We do so by increasing the number of terms in the expression (3.2.4) for the approximate solution. That is, we take the approximate solution to be

$$\tilde{\theta}(x) = a_1 x + a_2 x^2 + a_3 x^3 + \dots a_n x^n. \tag{3.2.9}$$

This is akin to improving the representation of a function by using Fourier series. If we require that the natural boundary condition (note that the essential boundary condition $(3.1.2)_1$ is satisfied by the assumed solution (3.2.9)) $(3.1.2)_2$ be satisfied, then we need to choose $(n-1)$ points in $(0, 1)$ and set $r(x)$ equal to zero there. We thus will have n linear algebraic equations for the n unknowns $a_1, a_2, \dots, a_n$. One can continue increasing the number of terms in eqn. (3.2.9) till the error in the approximate solution is less than the prescribed value.

Note that in expression (3.2.9) for $\tilde{\theta}(x)$ we included all terms upto x^n except for x^0 which is ruled out by the requirement that the trial solution $\tilde{\theta}(x)$ satisfy the prescribed homogeneous essential boundary condition $(3.1.2)_1$. Intuitively one expects that increasing the number of terms in eqn. (3.2.9) will decrease the error; however, the computational effort increases significantly and there is no guarantee that the error will indeed decrease with an increase in the value of n. A necessary (but not sufficient) condition to ensure the decrease in the error with an increase in n is to not skip any term in the series solution (3.2.9). The rate of decrease of the error with an increase in the number of terms in eqn. (3.2.9) depends upon the degree of the complete polynomials included in (3.2.9); i.e., the value of n provided that no term involving x^{m-1}, $1 \le m \le n$, is missing except of course those needed to satisfy the essential boundary conditions. The error may also depend upon the choice of the $(n-1)$ points; i.e., their x-coodrinates. In the absence of knowledge of the optimum locations of these points, one places them uniformly over the domain $(0, 1)$.

3.2.2 Method of Least Squares

In order to find n unknowns in eqn. (3.2.9) one does not need to satisfy eqn. (3.1.1) at n different points. For example, we can compute $r(x)$ at m points $x_1, x_2, \ldots, x_m$ in $(0,1)$ and set

$$\begin{aligned} E_i &= r(x_i), \quad i = 1, 2, \ldots, m, \\ E &= E_i E_i, \end{aligned} \tag{3.2.10}$$

and minimize E by setting

$$\frac{\partial E}{\partial a_i} = 0, \quad i = 1, 2, \ldots, n. \tag{3.2.11}$$

The n simultaneous linear eqns. (3.2.11) are solved for $a_1, a_2, \ldots, a_n$.

A modification of the definition $(3.2.10)_2$ of the error E is not to assume that $\tilde{\theta}(x)$ satisfies essential boundary conditions, and replace eqn. $(3.2.10)_1$ by

$$\begin{aligned} r_1 &= \tilde{\theta}(0), \\ r_2 &= \left.\frac{d\tilde{\theta}}{dx}\right|_{x=1} - 3, \\ E_i &= r(x_i), \quad i = 3, 4, \ldots, m, \\ E &= \sum_{i=1}^{m} \beta_i E_i E_i, \end{aligned} \tag{3.2.12}$$

where $x_3, x_4, \ldots, x_m$ are distinct points in $(0,1)$, and $\beta_1, \beta_2, \ldots, \beta_m$ are positive known constants. Depending upon the relative values assigned to $\beta_1, \beta_2, \ldots, \beta_m$, the error at different points is weighted. If β_1 is very large as compared to $\beta_2, \ldots, \beta_m$, then r_1 will be very small as compared to $r_2, \ldots, r_n$. This technique may be called the *method of weighted least squares*.

3.2.3 Method of Subdomains

In stead of setting $r(x) = 0$ at discrete points, we equate to zero the value of r integrated over n, not necessarily disconnected, subdomains Ω_i of $(0,1)$. That is

$$\int_{\Omega_i} r(x) dx = 0, \quad i = 1, 2, \ldots, n. \tag{3.2.13}$$

If $\tilde{\theta}(x)$ is not required not to satisfy *a priori* the essential boundary conditions, then one can take $i = 1, 2, \ldots, n-1$ in eqn. (3.2.13) and obtain the nth equation by satisfying the essential boundary condition. Similarly, one can also approximately satisfy the natural boundary condition, and take $i = 1, 2, \ldots, n-2$ in eqn. (3.2.13).

An alternative to eqn. (3.2.13) is

$$\int_{\Omega_i} r(x)W(x)dx = 0, \;\; i = 1, 2, \ldots, n, \tag{3.2.14}$$

where W is a positive-valued function on Ω_i. The weighting function W need not be same on every subdomain Ω_i.

3.2.4 Method of Weighted Residuals

An alternative to eqn. (3.2.14) is

$$\int_0^1 r(x)W_i(x)dx = 0, \;\; i = 1, 2, \ldots, n, \tag{3.2.15}$$

where $W_1(x), W_2(x), \ldots, W_n(x)$ are linearly independent functions defined on $(0, 1)$. For the approximate solution given by

$$\tilde{\theta}(x) = a_i\psi_i(x), \tag{3.2.16}$$

where $\psi_1(x), \psi_2(x), \ldots, \psi_n(x)$ are linearly independent functions, one could take

$$W_i(x) = \psi_i(x) \tag{3.2.17}$$

Since $r(x)$ may be positive at some points of the domain and negative at other points, one can improve upon the quality of the approximate solution by modifying eqn. (3.2.15) to the following:

$$\int_0^1 r^2\tilde{W}_i(x)dx = 0, \;\; i = 1, 2, \ldots, n, \tag{3.2.18}$$

where $\tilde{W}_i(x) \geq 0$ for every $x \in (0, 1)$. The n algebraic equations arising from eqn. (3.2.18) will be nonlinear in the unknowns, and may not be easy to solve.

Remarks.

1. The quality of the approximate solution depends upon the choice of basis functions $\phi_1(x), \phi_2(x), \ldots, \phi_n(x)$ (e.g., $1, x, x^2, \ldots, x^n$ in eqn. (3.2.4), and $\psi_1(x), \psi_2(x), \ldots, \psi_n(x)$ in eqn. (3.2.16)). Note that the basis functions are defined on $[0,\ 1]$ even though the differential equation holds on $(0, 1)$; i.e., basis functions are also defined at the boundary points.

2. The basis functions should be such that 1 and x can be represented exactly by their linear combination. This follows from the requirement that a rigid body motion (translation in a one-dimensional problem) and the state of uniform strain can be exactly produced by the

approximate solution. For the heat conduction problem it means that the state of uniform temperature and constant temperature gradient can be exactly reproduced.

3. The approximate solution is generally required to satisfy exactly the essential boundary conditions.

4. The natural boundary conditions may not be exactly satisfied.

5. There are several ways to measure errors in the approximate solution.

6. For the BV problem studied above with the four methods, the basis functions need to be twice differentiable. Therefore, their first-order derivatives must be continuous, since otherwise the second-order derivative will be infinite at a point where the first-order derivative is discontinuous.

3.3 Reduction of the Order of the Given Differential Equation

For the BVP defined by eqns. (3.1.1) and (3.1.2), we set $\eta = \theta'$ and reduce the BVP to

$$\begin{aligned} &\theta' = \eta, \;\; 0 < x < 1, \\ &\eta' - 4\theta - 2 = 0, \;\; 0 < x < 1, \\ &\theta(0) = 0, \;\; \eta(1) = 3. \end{aligned} \tag{3.3.1}$$

Since only first order derivatives appear in eqns. (3.3.1) we require the basis functions $\psi_1(x), \psi_2(x), \ldots, \psi_n(x)$ defined on $[0, 1]$ to be continuous, and express the approximate solution as

$$\begin{aligned} \tilde{\theta}(x) &= a_i \psi_i(x), \;\; i = 1, 2, \ldots, n, \\ \tilde{\eta}(x) &= b_i(x) \psi_i(x), \; i = 1, 2, \ldots, n. \end{aligned} \tag{3.3.2}$$

Thus $2n$ constants, $a_1, a_2, \ldots, a_n$, $b_1, b_2, \ldots, b_n$ need to be determined by using one of the four methods outlined above. Recalling that a differentiable function must be continuous, basis functions used in Section 3.2 can be employed in eqns. (3.3.2). However, the converse is not necessarily true. Thus the class of trial solutions for the BVP defined by eqns. (3.3.1) is much larger than that for the same problem defined by eqns. (3.1.1) and (3.1.2). Note that there are $2n$ unknowns to be found. One could also use different basis functions for $\tilde{\theta}(x)$ and $\tilde{\eta}(x)$ in eqn. (3.3.2).

Because of the appearance of odd order derivatives in eqn. (3.3.1), the classification of boundary conditions into essesntial and natural is not valid. We rely on eqns. (3.1.1) and (3.1.2) to decide which boundary conditions are essential and must be exactly satisfied.

It should be evident that a BVP involving a differential equation of order $2m$ can be written as one involving $2m$ differential equations of first order. The main advantage of doing so is to be able to choose basis functions from a large set at the cost of possibly increasing the number of unknowns.

Thus far the points chosen in the domain $(0,1)$ where the residual $r(x)$ is set equal to zero have been chosen arbitrarily, and no connection amongst them has been presumed. Even though the BVP defined by eqns. (3.1.1) and (3.1.2) is defined on a one-dimensional domain $(0,1)$, the proposed techniques of finding its approximate solution are valid even if the problem were defined on a two- or a three-dimensional domain.

3.4 Weak Formulation of the Problem

We now assume that the weight function W in eqn. (3.2.14) is twice differentiable. If eqn. (3.2.14) holds for every choice of W, then $r(x) = 0$ on Ω_i since one can set $W(x) = r(x)$ on Ω_i. Henceforth, we require that eqn. (3.2.14) holds for every choice of W, call W the test function, and denote it by ϕ.

Substituting in eqn. (3.2.14) for $r(x)$ from eqn. (3.2.1), we get

$$\begin{aligned} 0 &= \int_{\Omega_e} (\tilde{\theta}'' - 4\tilde{\theta} - 2)\phi dx, \\ &= -\int_{\Omega_e} (\tilde{\theta}'\phi' + 4\tilde{\theta}\phi + 2\phi)dx + (\phi\tilde{\theta}')\Big|_{\partial\Omega_e}, \end{aligned} \tag{3.4.1}$$

where we have used the divergence theorem (see Section 2.9) on the first term, and $\partial\Omega_e$ is the boundary of the domain Ω_e. For the 1-D problem, $\Omega_e = (x_e, x_{e+1})$, and $\partial\Omega_e$ consists of two points $x = x_e$ and $x = x_{e+1}$. We note that steps in eqn. (3.4.1) are reversible. The requirement that it hold for arbitrary test functions implies that we can recover eqns. (3.1.1) and $(3.1.2)_1$.

An advantage of using eqn. (3.4.1) is that it involves first-order derivatives of the trial solution $\tilde{\theta}$ and the test function ϕ. Writing eqn. (3.4.1) as

$$\int_{\Omega_e} (\tilde{\theta}'\phi' + 4\tilde{\theta}\phi)dx = -2\int_{\Omega_e} \phi\, dx + (\phi\tilde{\theta}')\Big|_{\partial\Omega_e}, \tag{3.4.2}$$

we see that the left-hand side of eqn. (3.4.2) is symmetric in $\tilde{\theta}$ and ϕ. Let

$$\tilde{\theta} = d_j\psi_j(x),\ \phi = c_i\psi_i(x);\ \ i,j = 1,2,\ldots,n, \tag{3.4.3}$$

where $\psi_1, \psi_2, \ldots, \psi_n$ are basis functions defined on Ω_e and are such that their first-order derivative is square integrable. That is

$$\int_{\Omega_e} (\psi_i')^2 dx < \text{constant}. \tag{3.4.4}$$

Substitution for $\tilde{\theta}$ and ϕ from eqns. (3.4.3) into eqn. (3.4.2) gives

$$c_i \left[\int_{\Omega_e} (\psi_i'\psi_j' + 4\psi_i\psi_j)dx\right] d_j = -2c_i \int_{\Omega_e} \psi_i dx + (\psi_i\psi_j')\Big|_{\partial\Omega_e} c_i d_j. \tag{3.4.5}$$

With the definitions

$$\begin{aligned} K_{ij}^e &\equiv \int_{\Omega_e} (\psi_i'\psi_j' + 4\psi_i\psi_j)dx, \\ f_i^e &\equiv -2\int_{\Omega_e} \psi_i dx, \\ g_{ij}^e &\equiv (\psi_i\psi_j')\Big|_{\partial\Omega_e}, \end{aligned} \tag{3.4.6}$$

eqn. (3.4.5) becomes

$$c_i[K_{ij}^e d_j - f_i^e - g_{ij}^e]d_j = 0. \tag{3.4.7}$$

Recall that different choices of constants $c_1, c_2, \ldots, c_n$ in eqn. $(3.4.3)_2$ will generate different functions ϕ. Since eqn. (3.4.2) holds for artibrary choices of ϕ, therefore eqn. (3.4.7) holds for all choices of $c_1, c_2, \ldots, c_n$. The choices $c_i = \delta_{i1}$, $c_i = \delta_{i2}, \ldots, c_i = \delta_{in}$ give n simultaneous linear equations

$$K_{ij}^e d_j = f_i^e + g_{ij}^e d_j,\ i,j = 1,2,\ldots,n. \tag{3.4.8}$$

The second term on the right-hand side of eqn. (3.4.8) links Ω_e to subdomains $\Omega_{(e-1)}$ and $\Omega_{(e+1)}$ in $(0,1)$.

The choice $\Omega_e = \Omega$ gives

$$\begin{aligned} g_{ij}d_j &= \psi_i(1)\psi_j'(1)d_j - \psi_i(0)\psi_j'(0)d_j, \\ &= \psi_i(1)\tilde{\theta}'(1) - \psi_i(0)\tilde{\theta}'(0), \\ &= 3\psi_i(1) - \psi_i(0)\tilde{\theta}'(0), \end{aligned} \tag{3.4.9}$$

where we have used the natural boundary condition $(3.1.2)_2$. Since θ is prescribed at $x = 0$, therefore $\tilde{\theta}'(0)$ is unknown. We thus have $(n+1)$ equations

$$\begin{aligned} K_{ij}d_j &= f_i + 3\psi_i(1) - \psi_i(0)\tilde{\theta}'(0), \\ \psi_j(0)d_j &= 0, \end{aligned} \tag{3.4.10}$$

for the $(n+1)$ unknowns $\tilde{\theta}'(0),\ d_1, d_2, \ldots, d_n$. Once eqns. (3.4.10) have been solved for $\theta'(0), d_1, d_2, \ldots, d_n$, eqn. $(3.4.3)_1$ gives an approximate solution of the problem.

As mentioned earlier there are several choices for the basis functions $\psi_1, \psi_2, \ldots, \psi_n$. In one-dimensional problems, examples of basis functions include $1, x, x^2, \ldots$; $L_0(x), L_1(x), \ldots$; $e^0, e^x, e^{2x}, \ldots$; $\sin x,\ \sin 2\pi x, \ldots$; $1, \cos x, \cos \pi x, \cos 2\pi x, \ldots$. The finite element method (FEM) provides a systematic way of generating polynomial basis functions $\psi_i(x)$.

Chapter 4: One-dimensional Problems

4.1 Two Model Problems

The finite element method (FEM) is a technique to find an approximate solution of a given BV or of an IBV problem. Steps involved in finding an approximate solution of a BV problem are

(i) derive a weak/variational formulation of the problem,

(ii) derive a matrix formulation of the problem,

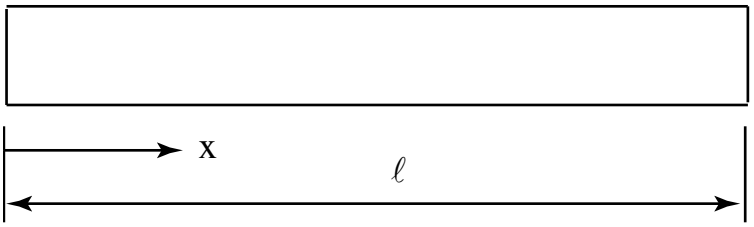
(iii) write and debug the code,

(iv) verify the code,

(v) compute and interpret the results, and

(vi) find error in the numerical solution.

Why do we need a weak formulation of the problem? In order to answer this question, we consider steady state heat conduction in a bar. Assume that heat flows only in the longitudinal direction, and that the bar can absorb or emit heat through its bounding surfaces. The governing equation is



$$k\frac{d^2u}{dx^2} - \alpha u = 0,\ 0 < x < \ell, \tag{4.1.1}$$

$$k\frac{du}{dx}\bigg|_{x=0} = -f_1,\ u = 20^\circ\text{C at } x = \ell. \tag{4.1.2}$$

Here k is the constant thermal conductivity, u the temperature, α a constant, f_1 the prescribed heat flux at the end $x = 0$, and the temperature at $x = \ell$ is assigned to be 20°C. We make the following observations. The given ordinary differential equation (ODE) is linear, homogeneous and is of second order; the order of the ODE is determined by the highest order derivative appearing in it. The ODE is homogeneous because when written with all terms involving the unknown u on the left-hand side, the right-hand side is zero.

The real-valued function $f(x)$ defined on the interval $[0,1]$ is a linear function of x if and only if $f(\alpha x+\beta y)=\alpha f(x)+\beta f(y)$ for every real α and β. Of course, $\alpha,\ \beta,\ x$ and y should be such that x, y and $\alpha x+\beta y\in[0,1]$. Thus $f(x)=3x$ is a linear function of x but $f(x)=3x+2$ is not. Writing eqn. (4.1.1) as $Lu=0$ where $L=k\frac{d^2}{dx^2}-\alpha$, we see that L is a linear differential operator since

$$\begin{aligned} L(au+bv) &= \left(k\frac{d^2}{dx^2}-\alpha\right)(au+bv),\\ &= k\frac{d^2}{dx^2}(au+bv)-\alpha(au+bv),\\ &= a\left(k\frac{d^2u}{dx^2}-\alpha u\right)+b\left(k\frac{d^2v}{dx^2}-\alpha v\right),\\ &= aLu+bLv. \end{aligned} \tag{4.1.3}$$

The order of the highest-order derivative in the boundary conditions (BCs) is less than the order of the highest-order derivative in the ODE. This is true for most BV problems in classical mechanics. Note that the BC at $x=0$ involves first-order derivative of u and that at $x=\ell$ involves no derivative. The BC at $x=0$ is called a natural BC and that at $x=\ell$ an essential one. How do we decide which BC is essential and which is natural?

As stated in Section 1.3, for a DE of order $2m$ (nearly all physical problems studied in mechanics involve DEs of even order; $2m$ equals the highest order derivative in the given DE) BCs involving derivatives of order *atmost* $(m-1)$ (i.e., $\leq(m-1)$) are called essential; other BCs are called natural. Let us consider a fourth-order DE associated with the bending of a beam.

Governing equation:

$$EI\frac{d^4\omega}{dx^4}+f(x)=0,\ 0<x<\ell. \tag{4.1.4}$$

BCs:

$$\text{at } x=0,\ \omega=0,\ \frac{d\omega}{dx}=0, \tag{4.1.5}$$

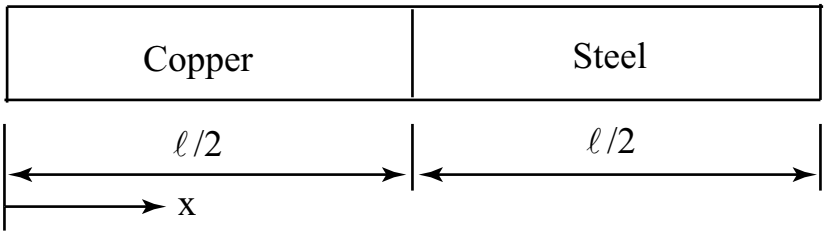
$$\text{at } x=\ell,\ EI\frac{d^2\omega}{dx^2}=0,\ \omega=0. \tag{4.1.6}$$

Here $2m=4$, $m-1=1$. Therefore, $\omega(0)=0,\ \left.\frac{d\omega}{dx}\right|_{x=0}=0,\ \omega(\ell)=0$ are essential BCs and $\left.EI\frac{d^2\omega}{dx^2}\right|_{x=\ell}=0$ is a natural BC.

Note that at $x = 0$ both BCs are essential but at $x = \ell$ one BC is natural and the other is essential. Thus at a point on the boundary, both essential and natural BCs may be prescribed simultaneously. Can it always be done? The answer is yes, so long as they are *linearly independent*.

4.2 Weak Formulation of a Problem

Now let us return to the problem defined by eqns. (4.1.1) and (4.1.2). By a classical or strong solution of the problem, we mean a function u defined on $[0, \ell]$ that is twice differentiable at every point in $(0, \ell)$ and satisfies eqn. (4.1.1) in $(0, \ell)$ and BCs (4.1.2) at the end points. For heat conduction in a bar with two parts made of different materials say copper (Cu) and steel (St),



the BCs at $x = \ell/2$ are

$$\begin{aligned} u\big|_{(\ell/2)^-} &= u\big|_{(\ell/2)^+}, \\ ku'\Big|_{(\ell/2)^-} &= ku'\Big|_{(\ell/2)^+}. \end{aligned} \tag{4.2.1}$$

Note that $k\Big|_{(\ell/2)^-} = k$ for Cu, $k\Big|_{(\ell/2)^+} = k$ for St and since $k_{\rm Cu} \neq k_{\rm St}$, $u'\Big|_{(\ell/2)^-} \neq u'\Big|_{(\ell/2)^+}$ and u'' is undefined at $x = \ell/2$. Thus we do not have a classical solution of the problem of heat conduction in a bar with two parts made of different materials. A similar situation will arise for the problem of the bending of a beam when at a point either the cross-section changes suddenly or the material changes abruptly. Since such problems are encountered in daily life and are physically meaningful, we need to modify or relax the definition of a solution of eqns. (4.1.1) and (4.1.2); one such modification is the concept of a weak solution.[1] By a **weak solution** of eqns. (4.1.1) and (4.1.2), we mean a function u defined on $[0, \ell]$ that satisfies exactly the essential BC $(4.1.2)_2$, but the natural BCs and eqn. (4.1.1) in a weak sense over the respective domains rather than at every

[1] An alternative approach is to solve two heat conduction problems; one on the domain $(0, \ell/2)$ and the other on $(\ell/2, 1)$ and use the conditions (4.2.1) to match the two solutions.

point of the domain. A weak solution generally has derivatives of order m when the given DE is of order $2m$. We elaborate upon this below.

Let $\phi : [0, \ell] \to \mathbb{R}$ be a "smooth" function, i.e., it has as many derivatives as we need. Multiply both sides of eqn. (4.1.1) by ϕ and integrate the result over the domain $(0, \ell)$ to obtain

$$\int_0^\ell \left(k\frac{d^2u}{dx^2} - \alpha u \right) \phi dx = 0. \tag{4.2.2}$$

Now integrate the first term by parts:[2]

$$\begin{aligned} \int_0^\ell k\frac{d^2u}{dx^2}\phi dx &= \left[\phi k \frac{du}{dx} \right] \Big|_0^\ell - \int_0^\ell k\frac{du}{dx}\frac{d\phi}{dx}dx, \\ &= \phi(\ell)ku'(\ell) + \phi(0)f_1 - \int_0^\ell ku'\phi' dx, \end{aligned} \tag{4.2.3}$$

where $u' = du/dx$. Note that we have used the natural BC at $x = 0$. Substitution from eqn. (4.2.3) into eqn. (4.2.2) gives

$$\int_0^\ell \left(ku'\phi' + \alpha u\phi \right) dx = \phi(0)f_1 + \phi(\ell)g(\ell), \tag{4.2.4}$$

where $g(\ell) = ku'(\ell)$ in the unknown heat flux at $x = \ell$. If functions u and ϕ are twice differentiable, then one can obtain eqns. (4.1.1) and (4.1.2) from eqn. (4.2.4) by following steps in the reverse order and making a suitable choice of ϕ. Thus for a classical solution of the problem, eqn. (4.2.4) embodies all of the information contained in eqns. (4.1.1) and (4.1.2). For a problem corresponding to heat conduction in a bar with different portions made of different materials, eqn. (4.2.4) is meaningful but eqn. (4.1.1) is not. Let

$$H^1 = \left\{ \psi \big| \psi : [0, \ell] \to \mathbb{R},\ \int_0^\ell \left(\psi^2 + \psi'^2 \right) dx < \infty \right\}. \tag{4.2.5}$$

By a *weak* solution of the BVP defined by eqns. (4.1.1) and (4.1.2) we mean a function $u : [0, \ell] \to \mathbb{R},\ u \in H^1,\ u(\ell) = 20$, such that eqn. (4.2.4) holds for *every* $\phi \in H^1$. The superscript 1 in H^1 signifies that every function in H^1 is differentiable and the square of the first derivative is intergrable on $[0, \ell]$. Note that a weak solution generally has derivatives of order $\mathbf{m}$ when the strong solution has derivatives of order $\mathbf{2m}$ and does not satisfy exactly the natural BC. We can now state formally a weak formulation of the problem defined by eqns. (4.1.1) and (4.1.2):

Find a function $u : [0, \ell] \to \mathbb{R}$ such that (i) $u(\ell) = 20$, (ii) $u \in H^1$, and (iii) eqn. (4.2.4) holds for *every* $\phi \in H^1$.

[2] Recall the integration by parts formula: $\int_a^b udv = uv\big|_a^b - \int_a^b vdu$. This is basically the divergence theorem in 1-D.

Thus to solve the problem defined by eqns. (4.1.1) and (4.1.2), we can choose a function $u \in H^1$ that satisfies $u(\ell) = 20$, substitute it into eqn. (4.2.4) and see if it is satisfied for *all* choices of $\phi \in H^1$. If eqn. (4.2.4) is satisfied, then the chosen function u is a weak solution of the given problem, otherwise it is not. Remember that eqn. (4.2.4) must be satisfied for every choice of the test function $\phi \in H^1$. The function u is called a trial solution, and the function ϕ the test function. Functions u and ϕ are from the set H^1 of functions.

When the trial solution and the test function are from the same space of functions, then the formulation is called the Galerkin formulation. The main advantage of the **Galerkin formulation** is that we need to construct only one space of functions. Thus the Galerkin formulation of the problem defined by eqns. (4.1.1) and (4.1.2) can be stated as follows:

Find $u \in H^1$ such that u satisfies the essential BC (4.1.2)$_1$ and eqn. (4.2.4) holds for every $\phi \in H^1$.

Note that the space of functions H^1 is infinite dimensional meaning that there are infinitely many linearly independent functions in it. A set of linearly independent functions constitutes basis functions in H^1, i.e., any function in H^1 can be expressed as a linear combination of the basis functions. Since there may be several sets of linearly independent functions in H^1, the choice of basis functions is not unique. For example, functions $\{\sin \frac{n\pi x}{\ell},\ n = 1, 2, \ldots\}$, $\{1, x, x^2, \ldots\}$, $\{(x-\ell),\ (x-\ell)^2, \ldots\}$ in H^1 are linearly independent. For an exact representation of u in terms of basis functions in H^1 one needs to determine infinitely many unknowns. In numerical work, the evaluation of infinitely many unknows is impossible, thus we work with a finite dimensional subset of H^1. Let H^{1n} be a n-dimensional subset of H^1, and $u^n \in H^{1n}$, $\phi^n \in H^{1n}$. Then the *Galerkin approximation* of the given BV problem defined by eqns. (4.1.1) and (4.1.2) is:

Find $u^n \in H^{1n}$ such that u^n satisfies the given essential BC and

$$\int_0^\ell (k u^{n\prime} \phi^{n\prime} + \alpha u^n \phi^n)\, dx = \phi^n(0) f_1 + \phi^n(\ell) g(\ell) \tag{4.2.6}$$

for *every* $\phi^n \in H^{1n}$.

Let $\psi_1, \psi_2, \ldots, \psi_n$ be basis functions in H^{1n}. Then

$$u^n(x) = d_1\psi_1(x) + d_2\psi_2(x) + \ldots + d_n\psi_n(x), \tag{4.2.7}$$

$$\phi^n(x) = c_1\psi_1(x) + c_2\psi_2(x) + \ldots + c_n\psi_n(x), \tag{4.2.8}$$

where $d_1, d_2, \ldots, d_n$ and $c_1, c_2, \ldots, c_n$ are constants. One can think of d_i as a component of $u^n(x)$ along the basis function (vector) $\psi_i(x)$. The basis functions $\psi_1, \psi_2, \ldots, \psi_n$ need not be orthonormal.[3] The problem of finding an approximate solution of the given problem reduces to that of finding n constants $d_1, d_2, \ldots, d_n$. Since the right-hand sides of equations $(4.2.7)$ and $(4.2.8)$ are sums of a finite number of terms, they can be differentiated term by term. Substituting for u^n and ϕ^n from eqs. (4.2.7)and (4.2.8) into eqn. (4.2.6) we arrive at

$$\sum_{i,j=1}^{n} c_i K_{ij} d_j = \sum_{i=1}^{n} c_i (F_i), \tag{4.2.9}$$

where

$$K_{ij} = \int_0^\ell \left(k\psi_i' \psi_j' + \alpha \psi_i \psi_j \right) dx, \tag{4.2.10a}$$

and

$$F_i = \psi_i(0) f_1 + \psi_i(\ell) g(\ell). \tag{4.2.11}$$

The square matrix K_{ij} is usually referred to as the "stiffness" matrix and the vector F_i the "load" vector. Note that $K_{ij} = K_{ji}$, i.e., $\mathbf{K}$ is symmetric. The matrix $\mathbf{K}$ will be symmetric if the given DE (or the mathematical model of the physical problem) involves derivatives of even order only. Since eqn. (4.2.6) holds for every $\phi^n \in H^{1n}$, therefore, eqn. (4.2.9) must hold for every choice of n constants $c_1, c_2, \ldots, c_n$. Note that a different choice of constants $c_1, c_2, \ldots, c_n$ will give a different function $\phi^n \in H^{1n}$. We choose $c_1 = 1,\ c_2 = c_3 = \ldots = c_n = 0$. Then eqn. (4.2.9) gives

$$\sum_{j=1}^{n} K_{1j} d_j = F_1. \tag{4.2.12}$$

Now choose $c_1 = 0,\ c_2 = 1,\ c_3 = c_4 = \ldots = c_n = 0$. The result is

$$\sum_{j=1}^{n} K_{2j} d_j = F_2. \tag{4.2.13}$$

Similarly, one can select $c_i = \delta_{i3}$, δ_{i4} etc. and obtain

$$\sum_{j=1}^{n} K_{ij} d_j = F_i,\ i = 1, 2, \ldots, n; \text{ or simply } \mathbf{Kd} = \mathbf{F}. \tag{4.2.14}$$

[3] Functions $\psi_1,\ \psi_2, \ldots \in H^1$ are orthonormal if

$$\int_0^\ell \psi_i \psi_j dx = \delta_{ij}, \text{ where } \delta_{ij} \text{ is the Kronecker delta.}$$

Another way to conclude eqn. (4.2.14) from eqn. (4.2.9) is the following. Let $e_i = \sum_{j=1}^{n} K_{ij} d_j - F_i$. Then eqn. (4.2.9) requires that $\sum_{i=1}^{n} e_i c_i = 0$. That is the n-dimensional vector $\mathbf{e}$ is perpendicular to every n-dimensional vector $\mathbf{c}$, since only such vector is a null vector, therefore, $\mathbf{e} = \mathbf{0}$ which is eqn. (4.2.14). Thus once the basis functions $\psi_1, \psi_2, \ldots, \psi_n$ have been selected, the "stiffness matrix" $\mathbf{K}$ and the "load vector" $\mathbf{F}$ can be evaluated.

For the trial solution $u^n(x)$ to satisfy the essential BC $(4.1.2)_2$, we must have

$$\sum_{i=1}^{n} d_i \psi_i(\ell) = 20. \tag{4.2.15}$$

Assuming that the matrix $\mathbf{K}$ is invertible, then $(n+1)$ eqns. (4.2.14) and (4.2.15) can be solved for $(n+1)$ unknowns $d_1, d_2, \ldots, d_n$ and $g(\ell)$, and the approximate solution to the given BV problem can be determined. If the given BV problem, i.e., the DE (4.1.1) and BCs $(4.1.2)_1$ with $(4.1.2)_2$ replaced by $u = 0$ at $x = \ell$, has a unique solution then the stiffness matrix is invertible. If $k > 0$ and $\alpha > 0$, then the matrix $\mathbf{K}$ is positive-definite. To show this consider

$$c_i K_{ij} c_j = \int_0^\ell \left[k \left(c_i \psi_i' \right) \left(c_j \psi_j' \right) + \alpha \left(c_i \psi_i \right) \left(c_j \psi_j \right) \right] dx, \tag{4.2.16}$$

where we have used the summation convention. Since $c_i \psi_i = \phi^n$, therefore

$$c_i K_{ij} c_j = \int_0^\ell \left[k \left(\phi^{n\prime} \right)^2 + \alpha \left(\phi^n \right)^2 \right] dx \geq 0, \tag{4.2.17}$$

and the equality holds if and only if $\phi^n = 0$ or equivalently $c_1 = c_2 = \ldots = c_n = 0$. Thus $\mathbf{K}$ is a positive definite matrix and is invertible.

Steps involved to find an approximate solution of a given BV problem can be summarized as follows:

1. Obtain a weak formulation of the problem. It is also called a variational formulation of the problem, or the weighted residual form of the problem. For the BV problem defined by eqns. (4.1.1) and (4.1.2), the weak formulation is stated immediately after eqn. (4.2.5).

2. Obtain the Galerkin approximation of the problem. For the given BV problem, it is stated as eqn. (4.2.6).

3. Choose a finite number of basis functions and obtain the matrix formulation of the problem. For the BV problem under discussion, it is given by eqn. (4.2.14).

4. Solve the system of linear eqs. (4.2.14) and (4.2.15), and examine the quality of the approximate solution u^n.

5. Find error in the approximate solution.

As is evident from the definitions $(4.2.10)$ and $(4.2.11)$ of K_{ij} and F_i, the quality of the approximate solution depends upon the dimensionality of the space H^{1n} and the choice of basis functions in H^{1n}.

An Example

Consider the BV problem

$$u'' + 1 = 0,\ 0 < x < 1,\ u' = \frac{du}{dx}, \tag{4.2.18}$$

$$u(0) = 0,\ u(1) = 0. \tag{4.2.19}$$

a) Obtain a weak formulation of the problem.

b) Derive a Galerkin approximation of the problem.

c) Deduce a matrix formulation of the problem.

d) Select $\psi_i = \sin i\pi x,\ i = 1, 2, 3$. Calculate K_{ij} and F_i and solve for unknowns d_1, d_2, d_3.

e) Choose $\psi_1 = x(1-x), \psi_2 = x^2(1-x), \psi_3 = x^3(1-x)$. Calculate K_{ij} and F_i and solve for d_1, d_2, d_3.

Solution

(a) Let $\phi : [0, 1] \to \mathbb{R}$ be a smooth function. Multiplying both sides of eqn. $(4.2.18)_1$ by ϕ, integrating the resulting equation over (0,1), and using the integration by parts formula, we obtain

$$\begin{aligned}
\int_0^1 0\phi dx = \int_0^1 (u'' + 1)\phi dx &= \int_0^1 u''\phi dx + \int_0^1 \phi dx, \\
&= \phi u'\Big|_0^1 - \int_0^1 u'\phi' dx + \int_0^1 \phi dx, \\
&= \phi(1)g - \phi(0)h - \int_0^1 u'\phi' dx + \int_0^1 \phi dx, \\
\int_0^1 u'\phi' dx &= \phi(1)g - \phi(0)h + \int_0^1 \phi dx
\end{aligned} \tag{4.2.20}$$

where $g = u'(1)$ and $h = u'(0)$ are unknowns. Let

$$H^1 = \left\{ \psi \big| \psi : [0,1] \to \mathbb{R},\ \int_0^1 \left(\psi'^2 + \psi^2\right) dx < \infty \right\}.$$

Then a weak formulation of the problem is: Find $u \in H^1$ such that

$$u(0) = 0,\ u(1) = 0 \text{ and eqn. (4.2.20) holds for every } \phi \in H^1.$$

(b) Let H^{1n} be a n-dimensional subspace of H^1, and ϕ^n and u^n be in H^{1n}. Then the Galerkin formulation of the given problem is: Find $u^n \in H^{1n}$ such that

$u^n(0) = 0,\ u^n(1) = 0$ and the eqn. $\displaystyle\int_0^1 u^{n\prime}\phi^{n\prime} dx = \phi^n(1)g - \phi^n(0)h + \int_0^1 \phi^n dx$ holds for every $\phi^n \in H^{1n}$.

(c) Let $u^n = \displaystyle\sum_{i=1}^{n} d_i\psi_i$, and $\phi^n = \sum_{j=1}^{n} c_j\psi_j$. Then $g,\ h$ and d_i's are solutions of

$K_{ij}d_j = F_i,\ d_i\psi_i(0) = 0,\ d_i\psi_i(\ell) = 0,$ where $K_{ij} = \displaystyle\int_0^1 \psi_i'\psi_j' dx,\ F_i = \int_0^1 \psi_i dx + \psi_i(1)g - \psi_i(0)h.$

(d) For $\psi_i = \sin i\pi x,\ \psi_i' = (i\pi)\cos i\pi x,\ i = 1, 2, 3,$

$$K_{ii} = \int_0^1 (i\pi)^2 \cos^2 i\pi x dx = (i\pi)^2 \left[\frac{x}{2} + \frac{\sin i2\pi x}{4i\pi}\right]\Bigg|_0^1 = \frac{i^2\pi^2}{2},\ \text{no sum on } i.$$

For $i \neq j$,

$$K_{ij} = \int_0^1 ij \cos i\pi x \cos j\pi x dx = \frac{ij}{2}\left[\frac{\sin(i+j)\pi x}{(i+j)\pi} + \frac{\sin(i-j)\pi x}{(i-j)\pi}\right]_0^1 = 0,$$
$$F_i = \int_0^1 \sin i\pi x dx + g\sin i\pi - (0)h = \frac{-\cos i\pi x}{i\pi}\Bigg|_0^1 = \frac{1 - \cos i\pi x}{i\pi}.$$

Thus

$$K_{ij} = \begin{bmatrix} \frac{\pi^2}{2} & 0 & 0 \\ 0 & \frac{4\pi^2}{2} & 0 \\ 0 & 0 & \frac{9\pi^2}{2} \end{bmatrix},\ F_i = \begin{bmatrix} 2/\pi \\ 0 \\ 2/(3\pi) \end{bmatrix},$$
$$K_{ij}d_j = F_i \text{ gives } d_1 = 4/\pi^3,\ d_2 = 0,\ d_3 = 4/(27\pi^3),$$
$$u^3 = \frac{4}{\pi^3}\sin\pi x + \frac{4}{27\pi^3}\sin 3\pi x.$$

Since unknowns g and h disappear from eqs. (4.2.20), they cannot be determined for this choice of basis functions. Note that u^n exactly satisfies the essential BCs which generally is not the case. The exact solution is $u_e = \frac{x(1-x)}{2}$.

(e)

$$\psi_i = x^i(1-x),\ i=1,2,3.\ \psi_i' = ix^{i-1} - (i+1)x^i,$$
$$K_{ij} = \int_0^1 \psi_i'\psi_j' dx = \frac{ij}{i+j-1} - \frac{i(j+1)+(i+1)j}{(i+j)} + \frac{(i+1)(j+1)}{(i+j+1)},$$
$$F_i = \int_0^1 \psi_i dx = (0)g - (0)h\frac{1}{i+1} - \frac{1}{i+2}.$$

Therefore,

$$K_{ij} = \begin{bmatrix} 1/3 & 1/6 & 1/10 \\ 1/6 & 2/15 & 1/10 \\ 1/10 & 1/10 & 3/335 \end{bmatrix},\ F_i = \begin{bmatrix} 1/6 \\ 1/12 \\ 1/20 \end{bmatrix},$$
$$K_{ij}d_j = F_i \text{ gives } d_1 = 1/2,\ d_2 = 0,\ d_3 = 0,$$
$$u^3 = \frac{x(1-x)}{2} = u_{\text{exact.}}$$

As for part (d) unknowns g and h do not contribute to the load vector. However, in general, this is not the case.

Remarks:

Note that for the second choice of basis functions, the approximate solution agrees with the exact solution. This is so because the analytical solution can be represented as a linear combination of the basis functions. This is always true, i.e., whenever the analytical solution of the given BV problem can be represented as a linear combination of the chosen basis functions, the Galerkin approximation will give the analytical solution of the problem. *This simple exercise illustrates the importance of the selection of basis functions.* In general, constants g and h will appear in the load vector, and one will need to use the essential BCs to find them.

Another question we need to answer is "how close is the approximate solution to the analytical solution of the problem?"

We first discuss the selection of the finite element basis functions and will then focus on the discussion of the error in the approximate solution.

4.3 Finite Element Basis Functions

As we saw in the example problem of Section 4.2, the quality of the approximate solution depends upon the choice of basis functions. The finite element method (FEM) provides an easy way to select these basis functions and their number can be increased systematically. The first step is to

divide the given domain into a finite number of *disjoint* subdomains, Ω_e, called finite elements such that their union equals the given domain Ω. We select special points in an element and call them nodes. The collection of finite elements and nodes is called the FE mesh. A simple choice is to select end points of an element as nodes; the discussion in the remainder of this section is limited to this choice of nodes. Let the nodes and elements be numbered consecutively starting from the left end. Thus nodes 3 and 4 belong to the third element, and for a FE mesh with $(m-1)$ elements, there will be m nodes.

The FE mesh in which all elements are of the same size is called a uniform mesh; otherwise it is a non-uniform mesh. Note that for the one-dimensional problem the union of subdomains matches exactly with the given domain but this *may* not be true for two- and three-dimensional problems.

The guidelines for constructing the FE basis functions are:

1. The basis functions are generated by simple polynomials defined piecewise-element by element- over the FE mesh.

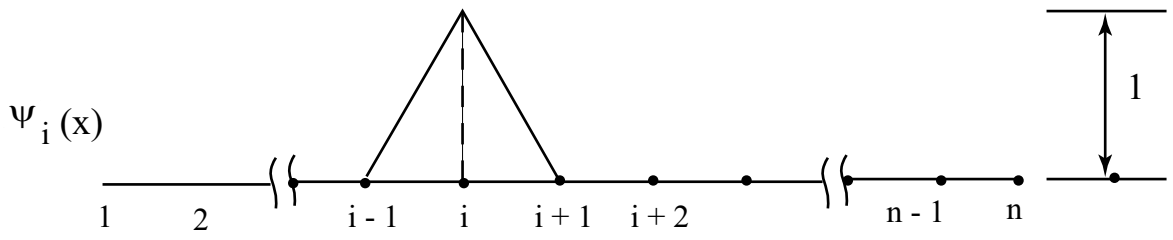
2. The basis functions are smooth enough to be in the appropriate space of functions, e.g., H^{1n} for the problem defined by eqns. (4.1.1) and (4.1.2).

3. The basis function ψ_i equals 1 at node i and 0 at the remaining nodes. That is $\psi_i(x_j) = \delta_{ij}$ where x_j is the x-coordinate of node j.

We note that the number of FE basis functions equals the number of nodes in the FE mesh. For a second-order DE the basis functions are required to be in H^1 and a simple choice is the following. For an interior node i, let

$$\psi_i(x) = \begin{cases} 0 & , \quad 0 \le x \le x_{i-1}, \\ \dfrac{x - x_{i-1}}{x_i - x_{i-1}} & , \quad x_{i-1} \le x \le x_i, \\ \dfrac{x_{i+1} - x}{x_{i+1} - x_i} & , \quad x_i \le x \le x_{i+1}, \\ 0 & , \quad x_{i+1} \le x \le \ell, \end{cases} \tag{4.3.1}$$

where x_i is the x-coordinate of the i^{it} node and the positive x-axis points to the right. A graph of $\psi_i(x)$ is shown below.



Functions $\psi_1(x)$ and $\psi_n(x)$ are defined as

$$\psi_1(x) = \begin{cases} \dfrac{x_2 - x}{x_2 - x_1} & , \quad x_1 \le x \le x_2, \\ 0 & , \quad x \ge x_2; \end{cases} \tag{4.3.2a}$$

$$\psi_n(x) = \begin{cases} 0 & , \quad 0 \le x \le x_{n-1}, \\ \dfrac{x - x_{n-1}}{x_n - x_{n-1}} & , \quad x_{n-1} \le x \le x_n. \end{cases} \tag{4.3.2b}$$

From the definition of $\psi_i(x)$, it is clear that $\psi_i(x)$ is continuous on the entire domain $[0, \ell]$ and is non-zero only on elements that meet at node i. Also $\psi_i'(x)$ is piece-wise constant; it is non-zero only on the interior of elements that meet at node i and is discontinuous at nodes belonging to these elements. A similar statement is valid for $(\psi_i'(x))^2$. Since $(\psi_i'(x))^2$ is discontinuous at a finite number of points and is finite everywhere else, therefore, it is integrable over the domain and the FE basis functions generated above are indeed in H^1. To see that they are linearly independent, consider the linear equation

$$\sum_{i=1}^{m} c_i \psi_i(x) = 0. \tag{4.3.3}$$

Evaluating (4.3.3) at $x = x_j$ and recalling that $\psi_i(x_j) = \delta_{ij}$, we conclude from eqn. (4.3.3) that $c_j = 0$, $j = 1, 2, \ldots, n$. Thus functions $\psi_1(x), \psi_2(x), \ldots, \psi_n(x)$ are linearly independent.

We now discuss details of obtaining an approximate solution of the BV problem defined by eqns. (4.1.1) and (4.1.2). Let

$$u^n(x) = \sum_{i=1}^{n} d_i \psi_i(x), \ \phi^n(x) = \sum_{i=1}^{n} c_i \psi_i(x). \tag{4.3.4}$$

Since $\psi_i(x_j) = \delta_{ij}$,

$$u^n(x_j) = \sum_{i=1}^{n} d_i \psi_i(x) = d_j. \tag{4.3.5}$$

That is, d_j equals the value of the trial solution at node j. This holds only if $\psi_i(x_j) = \delta_{ij}$.

Functions $u^n(x)$ and $\phi^n(x)$ are continuous and piecewise linear on $[0, \ell]$ and have constant derivative on the interior of each element but $u^{n\prime}(x)$ and $\phi^{n\prime}(x)$ are, in general, discontinuous

across inter-element boundaries. Regarding the stiffness matrix, we note that

$$K_{ij} = \int_0^\ell \left(k\psi_i'\psi_j' + \alpha\psi_i\psi_j\right) dx = \sum_{e=1}^{n-1} \int_{x_e}^{x_{e+1}} \left(k\psi_i'\psi_j' + \alpha\psi_i\psi_j\right) dx, \tag{4.3.6}$$

$$= \sum_{e=1}^{n-1} K_{ij}^e,\ i,j = 1,2,\ldots,n-1. \tag{4.3.7}$$

Here K_{ij}^e is the *element stiffness matrix* and K_{ij} is termed the *global stiffness matrix*. Thus the global stiffness matrix is obtained by summing together the element stiffness matrices. For the load vector we get

$$F_i = f_1\psi_i(0) + ku'(\ell)\psi_i(\ell). \tag{4.3.8}$$

Whereas

$$K_{ij} = \sum_e K_{ij}^e \tag{4.3.9}$$

is always valid, for the FE basis functions, K_{ij}^e has a simple form. For every element, K_{ij}^e is a $n \times n$ matrix since indices i and j take values $1, 2, \ldots, n$. However, for the e^{th} element,

$$\psi_i' = 0,\ \psi_i = 0,\ \text{if } i < e \text{ or } i > e+1; \tag{4.3.10}$$

thus

$$K_{ij}^e = 0 \text{ if } i,j < e \text{ or } i,j > (e+1), \tag{4.3.11}$$

implying thereby that there are only 4 non-zero entries in the $n \times n$ matrix K_{ij}^e. That is, in the $n \times n$ matrix K_{ij}^e, there is only one 2×2 non-zero matrix.

The 2×2 non-zero matrix $\overline{\mathbf{K}}^e$ is referred to as the "element stiffness matrix". The size of $\overline{\mathbf{K}}^e$ equals the number of nodes on an element multiplied by the number of degrees of freedom per node.

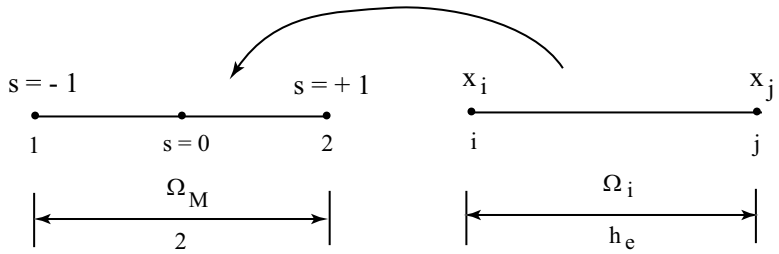
We now evaluate the 2×2 non-zero matrix $\overline{\mathbf{K}}^e$ in the $n \times n$ matrix $\mathbf{K}^e$. The nonzero basis functions on Ω_e are ψ_e and ψ_{e+1}; *the restrictions of basis functions to an element are called shape functions*. Thus for the piecewise linear basis functions being studied here, there are only two shape functions for an element. In general, there is one shape function associated with each node and since in this case an element has two nodes, there are only two shape functions for the element. Note that a shape function associated with a node equals one at that node and vanishes at other nodes on the element. The basis function $\psi_i(x)$ for an interior node i is obtained by patching

together the two shape functions, one when it is viewed as a node belonging to the element Ω_{i-1} and the other when it is regarded as a node on the element Ω_i.

Let N_i be the restriction of ψ_i on element Ω_i. That is, $N_i(x) = \psi_i(x),\ x \in \Omega_i$. Then N_i, N_j $(j = i+1)$ are shape functions for the i^{th} and the j^{th} nodes belonging to Ω_i, and

$$K_{ij}^e = \int_{\Omega_i} \left(k\psi_i'\psi_j' + \alpha\psi_i\psi_j\right) dx = \int_{\Omega_i} \left(kN_i'N_j' + \alpha N_iN_j\right) dx. \tag{4.3.12}$$

The limits of integration are the x-coordinates of the left and the right nodes of the element Ω_i. In order to facilitate the evaluation of this definite integral, we introduce the transformation



$$s = \frac{x - (x_i + x_j)/2}{(x_j - x_i)/2} = \frac{2x - (x_i + x_j)}{h_e}, \tag{4.3.13}$$

$$N_i(x) = \psi_i(x) = \frac{x_j - x}{x_j - x_i} = (1-s)/2 \equiv N_1(s), \tag{4.3.14}$$

$$N_j(x) = \psi_j(x) = \frac{x - x_i}{x_j - x_i} = (1+s)/2 \equiv N_2(s), \tag{4.3.15}$$

$$N_i' = \frac{dN_i}{dx} = \frac{dN_i}{ds}\frac{ds}{dx}, \tag{4.3.16}$$

$$N_1' = \left(-\frac{1}{2}\right)\frac{2}{h_e} = -1/h_e, \tag{4.3.17}$$

$$N_2' = \frac{1}{2}\frac{2}{h_e} = 1/h_e. \tag{4.3.18}$$

Here $h_e = (x_j - x_i)$ is the length of element Ω_i. We note that the domain $[x_i, x_j]$ is transformed into the domain [-1,1] irrespective of the values of x_i and x_j. The element with domain $[-1, 1]$ is called a **master element**, henceforth denoted by Ω_M. $N_1(s)$ and $N_2(s)$ are termed shape functions for Ω_M. Note that nodes on the master element are numbered by starting from 1, and the coordinate of a point on Ω_M is generally referred to as its local coordinate. Thus, for $a, b = 1, 2$,

$$K_{ab}^e = \int_{-1}^{1} \left(kN_a'N_b' + \alpha N_aN_b\right)\frac{dx}{ds}ds = \frac{h_e}{2}\int_{-1}^{1} \left(kN_a'N_b' + \alpha N_aN_b\right) ds, \tag{4.3.19}$$

$$[K^e] = \frac{h_e}{2}\int_{-1}^{1} \begin{bmatrix} k\left(\frac{1}{h_e}\right)^2 + \alpha\left(\frac{1-s}{2}\right)^2 & -\frac{k}{h_e^2} + \frac{\alpha}{4}\left(1-s^2\right) \\ -\frac{k}{h_e^2} + \frac{\alpha}{4}\left(1-s^2\right) & \frac{k}{h_e^2} + \alpha\left(\frac{1+s}{2}\right)^2 \end{bmatrix} ds, \tag{4.3.20}$$

$$= \begin{bmatrix} \frac{k}{h_e} + \alpha\frac{h_e}{3} & -\frac{k}{h_e} + \alpha\frac{h_e}{6} \\ -\frac{k}{h_e} + \alpha\frac{h_e}{6} & \frac{k}{h_e} + \alpha\frac{h_e}{3} \end{bmatrix}. \tag{4.3.21}$$

The expression (4.3.21) is valid for every element in the FE mesh. In order to generate the $n \times n$ matrix $\mathbf{K}^e$ from the 2×2 matrix $\overline{\mathbf{K}}^e$, we need to know the correspondence between the nodes on Ω_e and those on Ω_M; the array or the matrix giving this correspondence is called the connectivity array or the **connectivity matrix**. The number of rows in the **connectivity matrix** equals the number of elements in the FE mesh and the number of columns equals the number of nodes on an element. Thus for the FE mesh consisting of n elements with each element having two nodes, the size of the connectivity matrix, IC, will be $n \times 2$, and for the FE mesh under study it will have the following entries.

$$IC(1,1) = 1,\ IC(1,2) = 2;\ IC(2,1) = 2,\ IC(2,2) = 3, \ldots, IC(n,1) = n,\ IC(n,2) = n+1 \tag{4.3.22}$$

For the e^{th} element, $IC(e,1) = e$ and $IC(e,2) = e+1$ imply that nodes 1 and 2 on the master element correspond to nodes e and $e+1$ in the global mesh. Thus $\overline{K}^e_{11}$, $\overline{K}^e_{12}$, $\overline{K}^e_{21}$ and $\overline{K}^e_{22}$ correspond to K^e_{ee}, $K^e_{e(e+1)}$, $K^e_{(e+1)e}$ and $K^e_{(e+1)(e+1)}$, respectively, and the remaining entries in K^e_{ij} are zeroes. The next step is to add the $n \times n$ matrices $\mathbf{K}^e$ to obtain the global stiffness matrix; a process referred to as **assembling the global stiffness matrix**.

In order to illustrate the computation of the global stiffness matrix, we consider a uniform mesh of 3 elements and 4 nodes. Thus $h_e = \ell/3$ and

$$[\overline{K}^e] = \begin{bmatrix} \frac{3k}{\ell} + \alpha\frac{\ell}{9} & -\frac{3k}{\ell} + \alpha\frac{\ell}{18} \\ -\frac{3k}{\ell} + \frac{\alpha\ell}{18} & \frac{3k}{\ell} + \frac{\alpha\ell}{9} \end{bmatrix} \equiv \begin{bmatrix} A & B \\ B & A \end{bmatrix}, \tag{4.3.24}$$

$$[K^1] = \begin{bmatrix} A & B & 0 & 0 \\ B & A & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},\ [K^2] = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & A & B & 0 \\ 0 & B & A & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},\ [K^3] = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & A & B \\ 0 & 0 & B & A \end{bmatrix},$$

$$[K] = [K^1] + [K^2] + [K^3] = \begin{bmatrix} A & B & 0 & 0 \\ B & 2A & B & 0 \\ 0 & B & 2A & B \\ 0 & 0 & B & A \end{bmatrix}. \tag{4.3.24}$$

We note that the non-zero entries in $\mathbf{K}^e$ are obtained from those in $\mathbf{K}^{(e-1)}$ by translating them to the right and downwards by one.

We now compute the element load vector defined by eqn. (4.3.8). For this problem the load vector does not involve any domain integration. It follows from $\psi_i(x_j) = \delta_{ij}$ that

$$F_1 = f_1,\ F_2 = 0,\ F_3 = 0, \ldots,\ F_{n-1} = 0,\ F_n = ku'(\ell). \tag{4.3.25}$$

Thus for $n = 4$, we need to solve the following system of simultaneous equations.

$$\begin{bmatrix} A & B & 0 & 0 \\ B & 2A & B & 0 \\ 0 & B & 2A & B \\ 0 & 0 & B & A \end{bmatrix} \begin{Bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \end{Bmatrix} = \begin{Bmatrix} f_1 \\ 0 \\ 0 \\ ku'(\ell) \end{Bmatrix}. \tag{4.3.26}$$

Recall that we also need to satisfy the essential BC $u(\ell) = 20$. For $n = 4$,

$$u^4(x) = \sum_{i=1}^{4} d_i \psi_i(x) \text{ and } u^4(\ell) = 20 \tag{4.3.27}$$

give

$$d_4 = 20. \tag{4.3.28}$$

Thus the 4^{th} equation in (4.3.26) that has the unknown $ku'(\ell)$ on the right-hand side can be replaced by eqn. (4.3.28). Doing so results in the following system of equations.

$$\begin{bmatrix} A & B & 0 & 0 \\ B & 2A & B & 0 \\ 0 & B & 2A & B \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{Bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \end{Bmatrix} = \begin{Bmatrix} f_1 \\ 0 \\ 0 \\ 20 \end{Bmatrix} \tag{4.3.29}$$

After having solved eqn. (4.3.29) for $d_1,\ d_2,\ d_3$ and d_4, we can find the unknown $u'(\ell)$ from

$$Bd_3 + Ad_4 = ku'(\ell)$$

which is the 4^{th} eqn. in (4.3.26).

4.4 Interpretation of the Finite Element Solution

In our model problem defined by eqns. (4.1.1) and (4.1.2), let $k = 2,\ \alpha = 3,\ f_1 = 5$ and $\ell = 3$. Then for a uniform FE mesh of 3 elements, $h_e = 3/3 = 1$, $A = \frac{k}{h_e} + \alpha\frac{h_e}{3} = 3,\ B = -3/2$ and the set of eqns. (4.3.29) becomes

$$\begin{bmatrix} 3 & -\frac{3}{2} & 0 & 0 \\ -\frac{3}{2} & 6 & -\frac{3}{2} & 0 \\ 0 & -\frac{3}{2} & 6 & -\frac{3}{2} \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{Bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \end{Bmatrix} = \begin{Bmatrix} 5 \\ 0 \\ 0 \\ 20 \end{Bmatrix}, \tag{4.5.1}$$

whose solution is

$$d_1 = 35/13,\ d_2 = 80/39,\ d_3 = 215/39,\ d_4 = 20.$$

Thus the approximate solution of the problem is

$$u^4(x) = \frac{35}{13}\psi_1(x) + \frac{80}{39}\psi_2(x) + \frac{215}{39}\psi_3(x) + 20\psi_4(x)\ . \tag{4.5.2}$$

In order to plot this solution we note that

$$u^4(0) = \frac{35}{13},\ u^4(1) = \frac{80}{39},\ u^4(2) = \frac{215}{39},\ u^4(3) = 20, \tag{4.5.3}$$

and u^4 varies linearly between any two points. The graph of $u^4(x)$ is depicted in Fig. 4.3.

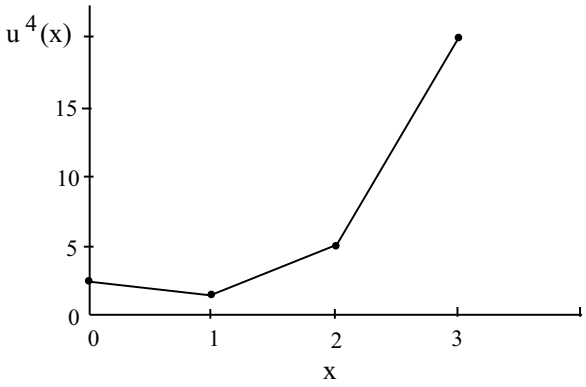


Fig. 4.3: Plot of an approximate solution of eqns. (4.1.1)and (4.1.2)

To find $u^n(x)$ for a value $\overline{x}$ of x not coinciding with a node, we first locate the element to which $\overline{x}$ belongs and then evaluate $u^n(\overline{x})$ by interpolation. For example, let $\overline{x}$ belong to the element e

with left node i and right node j. Then

$$u^n(\overline{x}) = u^n(x_i) + \frac{u^n(x_j) - u^n(x_i)}{x_j - x_i}(\overline{x} - x_i), \tag{4.5.4a}$$

$$= u_i^n + \frac{u_j^n - u_i^n}{x_j - x_i}(\overline{x} - \overline{x}_i), \tag{4.5.4b}$$

where u_i^n and u_j^n are values of the approximate solution u^n at nodes i and j, respectively. Equivalently, let $\overline{s}$ be the co-ordinate of the point in the master element that corresponds to $\overline{x}$ in Ω_e through the mapping (4.3.13). Then

$$u^n(\overline{x}) = u_i^n N_1(\overline{s}) + u_j^n N_2(\overline{s}) \tag{4.5.5}$$

which can be shown to be the same as the right-hand side of eqn. (4.5.4b).

We note that the essential BC at $x = 3$ is identically satisfied. This should not be surprising since we forced it to be so. However, the natural BC at $x = 0$ that was embedded in the weak formulation of the problem need not be well satisfied. Since $u^{4\prime}(x)$ is constant over each element, and $u^{4\prime}$ is not uniquely defined at the nodes, we need a different strategy. The common practice is to assign the constant value of $u^{4\prime}$ on an element to the centroid of the element and then use an interpolation or curve fitting technique to evaluate $u^{4\prime}$ at other points. This is referred to as **postprocessing** the solution. For our problem,

$$u^{4\prime}\left(\frac{1}{2}\right) = -0.6410,\ u^{4\prime}\left(\frac{3}{2}\right) = 3.4615,\ u^{4\prime}\left(\frac{5}{2}\right) = 14.4872, \tag{4.5.6}$$

and a quadratic curve passing through $\left(\frac{1}{2},\ -0.6410\right)$, $\left(\frac{3}{2},\ 3.4615\right)$ and $\left(\frac{5}{2},\ 14.4872\right)$ is

$$u^{4\prime}(x) = -0.0961 - 2.8207x + 3.4616x^2. \tag{4.5.7}$$

Hence $u^{4\prime}(0) = -0.0961$ which is far from the prescribed value of -2.5, and is a worse approximation than $-0.641 \times 2 = -1.282$. In order to improve upon this approximation we either need to use a finer mesh (i.e., more elements) or higher-order polynomials in the basis functions. The procedure for the first alternative is the same as that discussed in Section 4.3, higher-order basis functions are discussed below. However, in general, the FE solution does not satisfy exactly the natural boundary conditions.

4.5 Lagrange Shape Functions

As discussed in Section 4.3, the piecewise linear FE basis function for node i is obtained by patching together the shape functions for node i corresponding to the two elements that meet at node i.

Also the number of shape functions for an element equals the number of nodes on the element and the shape function for node A equals 1 at that node and zero at the remaining nodes on the element. The former property results in continuous basis functions. Thus we may discuss how to construct the shape functions rather than the basis functions. Also, when discussing the shape functions it is convenient to use the master element and work in terms of the local coordinate s. By using the transformation (4.3.13) one can obtain expressions for the basis functions in terms of x.

Let us consider a master element Ω_M with n-nodes and their local coordinates be $s_1, s_2, \ldots, s_n$. In order for the basis functions to be continuous, we require that $s_1 = -1$ and $s_n = +1$. The requirements that the shape functions be simple polynomials and that N_i vanish at every other node except node i suggests the following expression for $N_i(s)$:

$$N_i(s) = C(s - s_1)(s - s_2)\ldots(s - s_{i-1})(s - s_{i+1})\ldots(s - s_n), \tag{4.6.1}$$

where C is a constant. Note that $N_i(s)$ will be a polynomial of degree $(n-1)$ since we have $(n-1)$ factors linear in s in the product on the right-hand side of eqn. (4.6.1). We evaluate C by requiring that $N_i(s) = 1$ for $s = s_i$. Thus

$$\begin{aligned} N_i(s) =& \frac{(s - s_1)(s - s_2)\ldots(s - s_{i-1})(s - s_{i+1})\ldots(s - s_n)}{(s_i - s_1)(s_i - s_2)\ldots(s_i - s_{i-1})(s_i - s_{i+1})\ldots(s_i - s_n)}, \\ =& \prod_{\substack{j=1 \\ j\neq i}}^{n}(s - s_j) / \prod_{\substack{j=1 \\ j\neq i}}^{n}(s_i - s_j). \end{aligned} \tag{4.6.2}$$

The polynomial shape functions defined by eqn. (4.6.2) are called *Lagrange shape functions*. Examples:

(1) $n = 2. \quad s_1 = -1, \quad s_2 = +1,$

$$N_1(s) = \frac{(s-1)}{(-1-1)} = (1-s)/2, \tag{4.6.3}$$
$$N_2(s) = \frac{(s-(-1))}{(1+1)} = (1+s)/2.$$

These are the same as those given by eqns. (4.3.14) and(4.3.15).

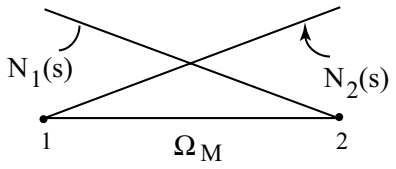


Fig. 4.4: Lagrange shape functions for 2-node element

(ii) $n = 3$. $s_1 = -1$, $s_2 = 0$, $s_3 = +1$.

$$N_1(s) = \frac{(s-0)(s-1)}{(-1-0)(-1-1)} = \frac{1}{2}s(s-1),$$
$$N_2(s) = \frac{(s-(-1))(s-1)}{(0+1)(0-1)} = (1-s^2), \qquad (4.6.4)$$
$$N_3(s) = \frac{(s-(-1))(s-0)}{(1+1)(1-0)} = \frac{1}{2}s(1+s).$$

The three shape functions are plotted in the Fig. 4.5.

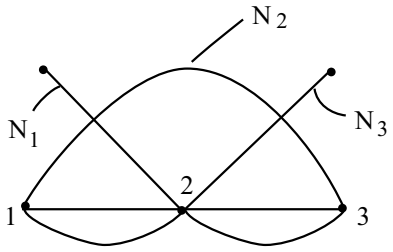


Fig. 4.5: Lagrange shape functions for 3-node element

The corresponding basis functions for a uniform FE mesh of two elements with each element having three nodes are depicted in the Fig. below.

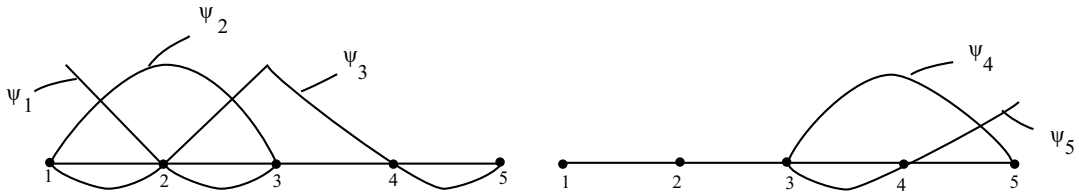


Fig. 4.6: Piece-wise quadratic basis functions in H^1.

(iii) $n = 4$. $s_1 = -1$, $s_2 = -\frac{1}{3}$, $s_3 = \frac{1}{3}$, $s_4 = 1$.

$$N_1(s) = \frac{(s-(-\frac{1}{3}))(s-\frac{1}{3})(s-1)}{(-1+\frac{1}{3})(-1-\frac{1}{3})(-1-1)} = \frac{(s^2-\frac{1}{9})(s-1)}{(-\frac{2}{3})(-\frac{4}{3})(-2)} = \frac{-1}{16}(9s^3-9s^2-s+1),$$
$$N_2(s) = \frac{(s+1)(s-\frac{1}{3})(s-1)}{(\frac{2}{3})(-\frac{2}{3})(-\frac{4}{3})} = \frac{9}{16}(s^2-1)(3s-1) = \frac{9}{16}(3s^3-s^2-3s+1), \qquad (4.6.5)$$
$$N_3(s) = \frac{(s+1)(s+\frac{1}{3})(s-1)}{(\frac{4}{3})(\frac{2}{3})(-\frac{2}{3})} = -\frac{9}{16}(s^2-1)(3s+1) = -\frac{9}{16}(3s^3+s^2-3s-1),$$
$$N_4(s) = \frac{(s+1)(s+\frac{1}{3})(s-\frac{1}{3})}{(2)(\frac{4}{3})(\frac{2}{3})} = \frac{1}{16}(s+1)(9s^2-1) = \frac{1}{16}(9s^3+9s^2-s-1).$$

Note that the basis functions are continuous at the nodes but their first-order derivative is discontinuous or suffers a jump there. Since ψ_i' is discontinuous only at a finite number of points, and is finite everywhere else, $(\psi_i')^2$ is integrable and the piecewise quadratic basis functions are in H^1. All basis functions generated from the Lagrange shape functions are in H^1.

One could also derive expressions for the higher-order shape functions by proceeding as follows. We give details of obtaining an expression for $N_1(s)$ for the quadratic shape function. Let

$$N_1(s) = a + bs + cs^2,$$

where a, b, c are constants to be determined from

$$N_1(-1) = 1,\ N_1(0) = 0,\ N_1(+1) = 0.$$

Thus

$$1 = a - b + c,\ 0 = a,\ 0 = a + b + c,$$

which give $a = 0,\ b = -1/2,\ c = +1/2$, and hence

$$N_1(s) = -\frac{1}{2}s + \frac{1}{2}s^2 = \frac{1}{2}s(s-1).$$

For the cubic shape function, we start with

$$N_1(s) = a + bs + cs^2 + ds^3,$$

place nodes 1, 2, 3 and 4 at $s = -1,\ -1/3,\ 1/3$ and 0 respectively. We use $N_1(-1) = 1,\ N_1(-\frac{1}{3}) = 0,\ N_1(\frac{1}{3}) = 0,\ N_1(1) = 0$ to obtain

$$1 = a - b + c - d,\ 0 = a - \frac{b}{3} + \frac{c}{9} - \frac{d}{27},$$
$$0 = a + \frac{b}{3} + \frac{c}{9} + \frac{d}{27},\ 0 = a + b + c + d,$$

whose solution is

$$a = -1/16,\ b = 1/16,\ c = 9/16,\ d = -9/16.$$

Hence

$$N_1(s) = -(9s^3 - 9s^2 - s + 1)/16.$$

4.6 Completeness of Shape Functions

Shape functions $N_1(s), N_2(s), \ldots, N_n(s)$ are said to be complete polynomials of order m if $1, s$, $s^2, \ldots, s^m$ can be expressed as a *linear* combination of $N_1, N_2, \ldots, N_n$. The completeness of shape functions plays a crucial role in the convergence of the solution as the FE mesh is refined. We note that Lagrange shape functions $N_1(s), N_2(s), \ldots, N_n(s)$ are complete polynomials of order $(n-1)$. Thus

$$1 = C_1N_1(s) + C_2N_2(s) + \ldots + C_nN_n(s). \tag{4.7.1}$$

Let $s = s_j$, the s-coordinate of the jth node. Since $N_i(s_j) = \delta_{ij}$, the Kronecker delta, therefore, $C_i = 1$ for every i, and we have

$$1 = N_1(s) + N_2(s) + \ldots + N_n(s). \tag{4.7.2}$$

Also

$$0 = \frac{dN_1}{ds} + \frac{dN_2}{ds} + \ldots + \frac{dN_n}{ds} = \sum_{i=1}^{n} \frac{dN_i}{ds}. \tag{4.7.3}$$

Basis functions derived from complete shape functions are also complete and they satisfy

$$1 = \sum_i \psi_i(x), \tag{4.7.4}$$

$$0 = \sum_i \frac{d\psi_i}{dx}(x). \tag{4.7.5}$$

For complete basis functions, the approximate solution of a linear BV problem converges to the analytical solution of the problem as the FE mesh is refined.

4.7 Imposition of Essential Boundary Conditions

It should become clear from eqn. (4.3.26) that the value of the load vector at the node where essential BC is prescribed is not fully determined. Whereas it was simple to eliminate the corresponding equation from consideration in the example problem studied, in a practical problem essential BCs are prescribed at several nodes that may be distributed throughout the structure. Eliminating corresponding equations is rather tedious. In practice we follow one of the following two procedures. In each case the global stiffness matrix and the global load vector are assembled for the entire mesh **by ignoring contributions to the load vector from reaction forces at nodes where essential boundary conditions are prescribed**.

Let us assume that u is prescribed to be β at node i, i.e., $u_i = \beta$. One can use either method A or method B to satisfy $u_i = \beta$.

Method A.

(i) We first multiply the ith column of the global stiffness matrix $\mathbf{K}$ by β and subtract the resulting vector from the load vector; this is equivalent to adding the contributions from the term $-\beta K_{ij}$ to the load vector.

(ii) Entries in the ith row and the ith column of $\mathbf{K}$ are set equal to zeroes,

(iii) K_{ii} (no sum on i) is replaced by 1 and F_i by β.

(iv) Thus the ith equation is replaced by $d_i = \beta$, and $d_i = \beta$ has been substituted in the remaining equaitons. The procedure is repeated for every node where an essential BC is assigned.

Method B.

(i) We replace K_{ii} (no sum on i) by $K_{ii}+\gamma,\ F_i$ by $\beta\gamma$ where γ is very large as compared to a typical entry in $\mathbf{K}$,

(ii) solve the resulting system of equations.

The value of γ depends upon the word length in the computer being used and generally equals 10^6 (absolute value of a typical entry in $\mathbf{K}$). Thus the ith equation becomes

$$K_{i1}d_1 + K_{i2}d_2 + \ldots + (K_{ii}+\gamma)d_i + \ldots + K_{in}d_n = \gamma\beta, \text{ no sum on } i \text{ and } n.$$

Since $(K_{ii}+\gamma) >> |K_{ij}|$, the i^{th} equation is effectively reduced to

$$d_i = \frac{\gamma}{(K_{ii}+\gamma)}\beta = \beta/(1+K_{ii}/\gamma), \text{ no sum on } i,$$

and as $\gamma \to \infty,\ d_i \to \beta$. Thus the larger the value of γ, the better is the essential BC satisfied at node i. However, one cannot make γ too large since the set of equations may become ill conditioned. This method can be viewed as the *penalty method* in which we are penalizing the diagonal term of the stiffness matrix corresponding to the i^{th} equation.

Chapter 5: Fourth-Order Differential Equations

5.1 A Model Problem

Consider the following BV problem

$$(b(x)w''(x))'' = f(x),\ 0 < x < \ell, \tag{5.1.1}$$

$$w(0) = 0,\ w'(0) = 0,\ [b(x)w''(x)]_{x=\ell} = M_0,\ [b(x)w''(x)]'\Big|_{x=\ell} = F_0. \tag{5.1.2}$$

Here $w' = \frac{dw}{dx}$. The problem corresponds to analyzing the deflection w (displacement along the z-axis) of a cantilever beam of length ℓ loaded by a distributed force $f(x)$ acting upwards, rigidly clamped at the left end $x = 0$, and subjected to a point shearing force F_0 and a bending moment M_0 at the right end (cf. Fig. 5.1).

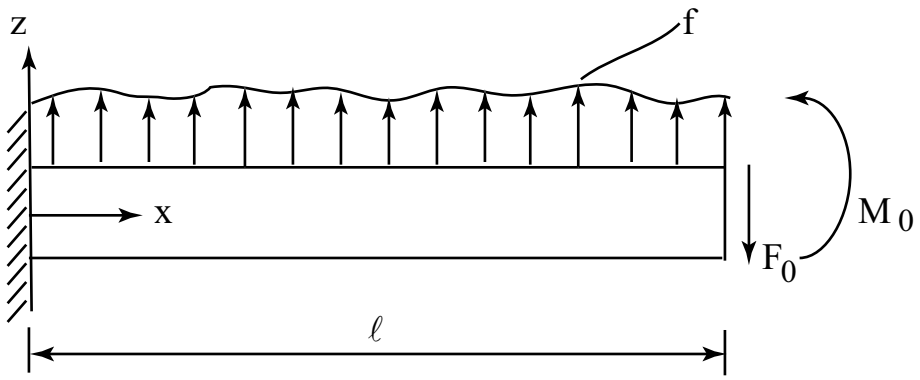


Fig. 5.1 A cantilever beam with applied loads and moments.

The function $b(x) = E(x)I(x)$ characterizes the bending stiffness of a cross-section of the beam; here E is Young's modulus of the material of the beam and I the second moment of area about the neutral axis. The DE $(5.1.1)_1$ is 4th order, therefore, $2m = 4$, $m - 1 = 1$, and BCs involving derivatives of order 1 and lower are essential and other BCs are natural. Thus BCs prescribed at the end $x = 0$ of the beam are essential and those assigned at $x = \ell$ are natural.

5.2 Galerkin Formulation of the Problem

Let $\phi : [0, \ell] \to \mathrm{R}$ be a smooth function. By multiplying both sides of eqn. (5.1.1) by ϕ, and integrating the resulting equation over the domain $(0, \ell)$, we obtain

$$\int_0^\ell [\phi(bw'')'' - \phi f]\, dx = 0. \tag{5.2.1}$$

Recall that

$$\int_0^\ell \phi(bw'')''dx = [\phi(bw'')' - \phi'(bw'')]_0^\ell + \int_0^\ell \phi''bw''dx,$$
$$= \phi(\ell)F_0 - \phi'(\ell)M_0 - [\phi(0)F_s - \phi'(0)M_s] + \int_0^\ell \phi''bw''dx. \tag{5.2.2}$$

Here we have integrated twice by parts to obtain the same order derivatives on ϕ and w, have substituted for the natural BCs prescribed at the end $x = \ell$, and set $(bw'')'\big|_{x=0} = F_s$, $(bw'')\big|_{x=0} = M_s$. Thus F_s and M_s denote, respectively, the shear force and the bending moment at the clamped edge $x = 0$. Equations (5.2.1) and (5.2.2) yield

$$\int_0^\ell \phi''bw''dx = -\phi(\ell)F_0 + \phi'(\ell)M_0 + [\phi(0)F_s - \phi'(0)M_s] + \int_0^\ell \phi f dx. \tag{5.2.3}$$

Since ϕ'' and w'' appear in the integrand on the left-hand side of eqn. (5.2.3), the test function ϕ and the trial solution w should be such that their second-order derivatives are square-integrable; hence their first-order derivatives must be continuous. Let

$$H^2 \equiv \left\{ \psi | \psi : [0,\ell] \to \mathrm{R},\ \int_0^\ell \left(\psi''^2 + \psi'^2 + \psi^2\right) dx < \infty \right\}. \tag{5.2.4}$$

The weak formulation of the problem defined by eqns. (5.1.1) and (5.1.2) can be stated as follows:

$$\text{Find } w \in H^2 \text{ such that } w(0) = 0,\ w'(0) = 0 \text{ and eqn. (5.2.3) holds for every } \phi \in H^2. \tag{5.2.5}$$

Note that the natural BCs are embedded in the weak formulation (5.2.3) of the problem and the essential BCs are to be satisfied.

Let H^{2n} be a finite dimensional subset of H^2 and let $\psi_1^0, \psi_1^1, \psi_2^0, \psi_2^1, \ldots, \psi_n^0, \psi_n^1$ be basis functions in H^n. Then for any function $w^n \in H^{2n}$

$$w^n(x) = \sum_{i=1}^{n} [w_i \psi_i^0(x) + w_i' \psi_i^1(x)]. \tag{5.2.6}$$

Henceforth, we will adopt the convention that a repeated index implies summation over the range of the index. The basis functions $\psi_1^0, \psi_2^0, \ldots, \psi_n^0$ ensure the continuity of the function $w^n(x)$ over the domain $(0, \ell)$, and basis functions $\psi_1^1, \psi_2^1, \ldots, \psi_n^1$ ensure the continuity of the first derivative of $w^n(x)$ over the domain $(0, \ell)$. Note that $\psi_1^0, \psi_1^1, \psi_2^0, \psi_2^1, \ldots$ are linearly independent and the dimensionality of H^{2n} is $2n$ rather than n. Furthermore, whereas ψ_i^0 are nondimensional, the dimension of ψ_i^1 is that of length in order for every term in eqn. (5.2.6) to have the same units.

Analogous to eqn. (5.2.6) we have

$$\phi^n(x) = \left(C_i^0 \psi_i^0(x) + C_i^1 \psi_i^1(x)\right). \tag{5.2.7}$$

Replacing w and ϕ by w^n and ϕ^n in eqn. (5.2.3), substituting from eqns. (5.2.6) and (5.2.7), and rearranging terms we get

$$\begin{aligned} &C_i^0 \left[\left(\int_0^\ell b\psi_i^{0\prime\prime}\psi_j^{0\prime\prime} dx \right) w_j + \left(\int_0^\ell b\psi_i^{0\prime\prime}\psi_j^{1\prime\prime} dx \right) w_j' \right] + \\ &C_i^1 \left[\left(\int_0^\ell b\psi_i^{1\prime\prime}\psi_j^{0\prime\prime} dx \right) w_j + \left(\int_0^\ell b\psi_i^{1\prime\prime}\psi_j^{1\prime\prime} dx \right) w_j' \right] = \\ &C_i^0 F_i^0 + C_i^1 F_i^1, \end{aligned} \tag{5.2.8}$$

where

$$F_i^0 = -\psi_i^0(\ell) F_0 + \psi_i^{0\prime}(\ell) M_0 + \left[\psi_i^0(0) F_s - \psi_i^{0'}(0) M_s \right] + \int_0^\ell \psi_i^0 f dx, \tag{5.2.9}$$

$$F_i^1 = -\psi_i^1(\ell) F_0 + \psi_i^{1\prime}(\ell) M_0 + \left[\psi_i^1(0) f_s - \psi_i^{1'}(0) M_s \right] + \int_0^\ell \psi_i^1 f dx. \tag{5.2.10}$$

Since different choices of C_i^0 and C_i^1 will yield different functions ϕ^n in eqn. (5.2.7), eqn. (5.2.8) must hold for all choices of constants $C_1^0, C_1^1, C_2^0, C_2^1, \ldots$ etc. For example, we can take $C_1^0 = 1,\ C_2^0 = 0,\ C_3^0 = 0,\ C_4^0 = 0 \ldots,\ C_1^1 = C_2^1 = \ldots = 0$. Next we can take $C_1^0 = 0,\ C_2^0 = 1,\ C_3^0 = C_4^0 = \ldots = 0,\ C_1^1 = C_2^1 = \ldots = 0$. Similarly we can take one C_i^1 to be 1 and the remaining C_i^1's and all C_i^0's to be zero. It yields the following $2n$ equations for the determination of w_i and w_i':

$$\left(\int_0^\ell b\psi_i^{0\prime\prime}\psi_j^{0\prime\prime} dx \right) w_j + \left(\int_0^\ell b\psi_i^{0\prime\prime}\psi_j^{1\prime\prime} dx \right) w_j' = F_i^0, \tag{5.2.11}$$

$$\left(\int_0^\ell b\psi_i^{1\prime\prime}\psi_j^{0\prime\prime} dx \right) w_j + \left(\int_0^\ell b\psi_i^{1\prime\prime}\psi_j^{1\prime\prime} dx \right) w_j' = F_i^1. \tag{5.2.12}$$

With

$$\begin{aligned} K_{ij}^{00} &= \int_0^\ell b\psi_i^{0\prime\prime}\psi_j^{0\prime\prime} dx,\ K_{ij}^{01} = \int_0^\ell b\psi_i^{0\prime\prime}\psi_j^{1\prime\prime} dx, \\ K_{ij}^{10} &= \int_0^\ell b\psi_i^{1\prime\prime}\psi_j^{0\prime\prime} dx,\ K_{ij}^{11} = \int_0^\ell b\psi_i^{1\prime\prime}\psi_j^{1\prime\prime} dx, \end{aligned} \tag{5.2.13}$$

$$K_{ij} = \begin{bmatrix} K_{ij}^{00} & K_{ij}^{01} \\ K_{ij}^{10} & K_{ij}^{11} \end{bmatrix},\ F_i = \begin{bmatrix} F_i^0 \\ F_i^1 \end{bmatrix},\ d_i = \begin{bmatrix} w_i \\ w_i' \end{bmatrix}, \tag{5.2.14}$$

we can write eqns. (5.2.11) and (5.2.12) as

$$K_{ij} d_j = F_i. \tag{5.2.15}$$

We note that each of the matrices $K_{ij}^{00},\ K_{ij}^{01},\ K_{ij}^{10}$ and K_{ij}^{11} is a $n \times n$ matrix. Also at each node we have two unknowns, the value of the function and the value of the slope of the function.

5.3 Basis Functions

Substitution of $x = x_j$, the x-coordinate of node j, in eqn. (5.2.6) yields

$$w^n(x_j) = w_i\psi_i^0(x_j) + w_i'\psi_i^1(x_j). \tag{5.3.1}$$

In FE work, we usually require that $w^n(x_j) = w_j$. One way to satisfy this is to have

$$\psi_i^0(x_j) = \delta_{ij},\ \psi_i^1(x_j) = 0. \tag{5.3.2}$$

Similarly,

$$w^{n\prime}(x_j) = w_i\psi_i^{0\prime}(x_j) + w_i'\psi_i^{1\prime}(x_j), \tag{5.3.3}$$

and again the requirement $w^{n\prime}(x_j) = w_j'$ can be satisfied by having

$$\psi_i^{0\prime}(x_j) = 0,\ \psi_i^{1\prime}(x_j) = \delta_{ij}. \tag{5.3.4}$$

Equations (5.3.2) and (5.3.4) can be used to generate the basis functions.

At each node the value of the basis function and of its slope are given. Thus if an element has two nodes, we have four conditions to satisfy and, therefore, assume that

$$\psi_i^0(x) = a + bx + cx^2 + dx^3,\ x \in \Omega_e, \tag{5.3.5}$$

over an element with left node at x_i and right node at x_j. Recalling that the restriction of a basis function to an element is called a shape function, we have four shape functions $N_i^0(x),\ N_i^1(x),\ N_j^0(x)$ and $N_j^1(x)$ for an element, and on Ω_e

$$N_i^0(x_i) = 1,\ N_i^0(x_j) = 0,\ N_i^{0\prime}(x_i) = 0,\ N_i^{0\prime}(x_j) = 0,\ \text{no sum on } i, \tag{5.3.6}$$

and a similar set of conditions holds for $N_j^0(x),\ N_i^1(x)$ and $N_j^1(x)$. With $N_i(x)$ given by the right-hand side of eqn. (5.3.5), we have

$$\begin{aligned} 1 &= a + bx_i + \ cx_i^2 + dx_i^3, \\ 0 &= a + bx_j + \ cx_j^2 + dx_j^3, \\ 0 &= \quad b \quad + 2cx_i + 3dx_i^2, \\ 0 &= \quad b \quad + 2cx_j + 3dx_j^2. \end{aligned} \tag{5.3.7}$$

Solving eqn. (5.3.7) for $a,\ b,\ c$ and d and substituting the result into eqn. (5.3.5) we obtain

$$N_i^0(x) = 1 - 3\left(\frac{x-x_i}{h_e}\right)^2 + 2\left(\frac{x-x_i}{h_e}\right)^3, h_e = x_j - x_i. \tag{5.3.8}$$

Similarly, we can derive

$$\begin{aligned} N_j^0(x) &= 3\left(\frac{x-x_i}{h_e}\right)^2 - 2\left(\frac{x-x_i}{h_e}\right)^3, \\ N_i^1(x) &= (x-x_i) - 2\frac{(x-x_i)^2}{h_e} + \frac{(x-x_i)^3}{h_e^2}, \\ N_j^1(x) &= -\frac{(x-x_i)^2}{h_e} + \frac{(x-x_i)^3}{h_e^2}. \end{aligned} \tag{5.3.9}$$

We note that the algebra would have been less involved had we assumed

$$N_i(x) = a + b(x-x_i) + c(x-x_i)^2 + d(x-x_i)^3,$$

and used eqn. (5.3.6) to solve for a, b, c and d. We still would have gotten eqns. (5.3.8) and (5.3.9). The shape functions (5.3.8) and (5.3.9) are called *Hermitian shape functions* and are plotted below in Fig. 5.3.1; the corresponding basis functions are called *Hermitian basis functions*.

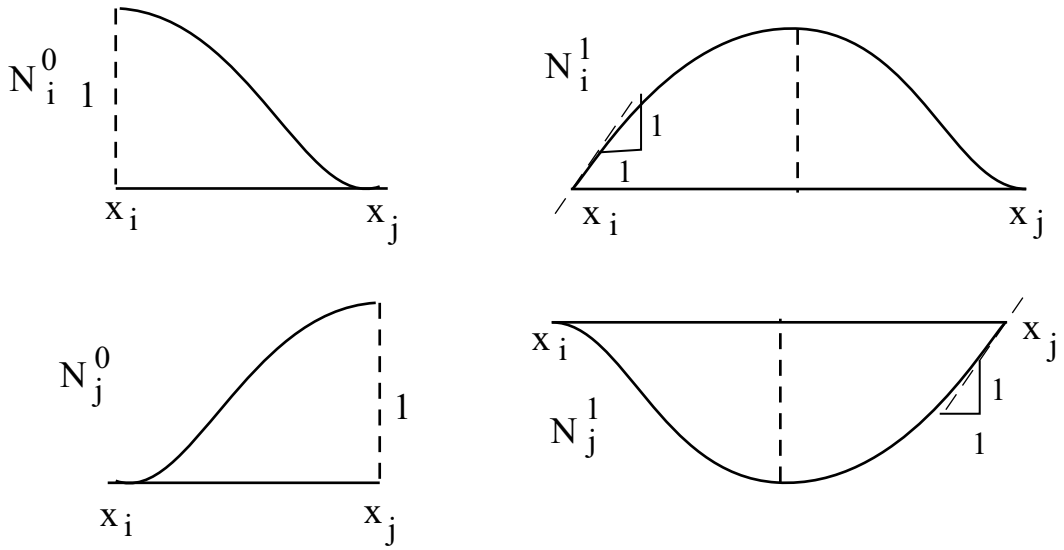


Fig. 5.3.1: Hermitian shape functions for nodes i and j.

The basis functions $\psi_i^0(x)$ and $\psi_i^1(x)$ are plotted below in Fig. 5.3.2.

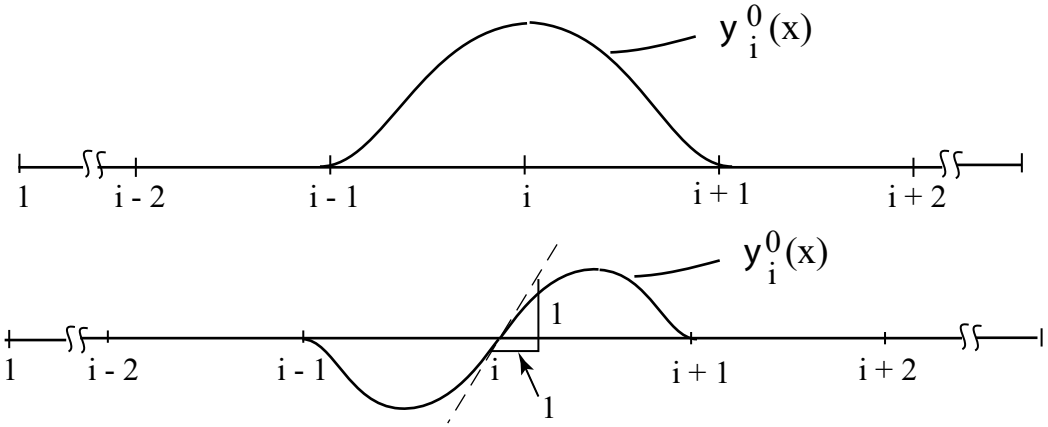


Fig. 5.3.2: Hermitian basis functions for node i.

As should be clear from the way $N_i^0,\ N_j^0,\ N_i^1$ and N_j^1 have been derived (cf. eqn. (5.3.7)) these four shape functions are complete polynomials of degree 3. In terms of the local coordinate

$$s = \frac{x - (x_i + x_j)/2}{h_e/2} = \frac{2x - (x_i + x_j)}{h_e}, \tag{5.3.10}$$

or

$$\begin{aligned} x &= \frac{(x_i + x_j)}{2} + \frac{h_e s}{2}, \\ x - x_i &= \frac{(x_j - x_i)}{2} + \frac{h_e s}{2} = \frac{h_e}{2}(1 + s), \end{aligned} \tag{5.3.11}$$

$$\begin{aligned} N_1^0(s) &= \frac{1}{4}(2 - 3s + s^3), \\ N_2^0(s) &= \frac{1}{4}(2 + 3s - s^3), \\ N_1^1(s) &= \frac{h_e}{8}(1 - s - s^2 + s^3), \\ N_2^1(s) &= \frac{h_e}{8}(-1 - s + s^2 + s^3). \end{aligned} \tag{5.3.12}$$

We note that

$$\begin{aligned} &N_1^0(-1) = 1,\ N_1^0(1) = 0,\ \left.\frac{dN_1^0}{dx}\right|_{s=\pm 1} = 0, \\ &N_2^0(-1) = 0,\ N_2^0(1) = 1,\ \left.\frac{dN_2^0}{dx}\right|_{s=\pm 1} = 0, \\ &N_1^1(\pm 1) = 0,\ \left.\frac{dN_1^1}{dx}\right|_{s=-1} = 1,\ \left.\frac{dN_1^1}{dx}\right|_{s=+1} = 0, \\ &N_2^1(\pm 1) = 0,\ \left.\frac{dN_2^1}{dx}\right|_{s=-1} = 0,\ \left.\frac{dN_2^1}{dx}\right|_{s+1} = 1. \end{aligned} \tag{5.3.13}$$

The derivative of shape functions N_1^1 and N_2^1 with respect to x rather than s equals 1 or 0 at local nodes 1 and 2. Whereas x has dimensional units, s is dimensionless. The presence of h_e in eqns. $(5.3.12)_3$ and $(5.3.12)_4$ makes $\frac{dN_1^1}{dx}$ and $\frac{dN_2^1}{dx}$ dimensionless.

5.4 Evaluation of Element Matrices

It is clear that at each node we have two unknowns - the value of the function and its first derivative. It is easier to number them and the basis functions as follows:

$$d_1 = w_1,\ d_2 = w_1',\ d_3 = w_2,\ d_4 = w_2', \ldots, d_{2i-1} = w_i,\ d_{2i} = w_i', \ldots, \tag{5.4.1}$$

$$\eta_1(x) = \psi_1^0(x),\ \eta_2(x) = \psi_1^1(x), \ldots, \eta_{2i-1}(x) = \psi_i^0(x),\ \eta_{2i}(x) = \psi_i^1(x), \ldots. \tag{5.4.2}$$

Thus we can write

$$w^n(x) = d_i \eta_i(x),\ i = 1, 2, \ldots, 2n, \tag{5.4.3}$$

$$\phi^n(x) = c_j \eta_j(x),\ j = 1, 2, \ldots, 2n. \tag{5.4.4}$$

Replacing w and ϕ in eqns. (5.2.11) and (5.2.12) by w^n and ϕ^n and rearranging terms we obtain

$$c_i K_{ij} d_j = c_i F_i, \tag{5.4.5}$$

where

$$K_{ij} = \int_0^\ell \eta_i'' b \eta_j'' dx, \tag{5.4.6a}$$

$$F_i = -\eta_i(\ell) F_0 + \eta_i'(\ell) M_0 + [\eta_i(0) F_s - \eta_i'(0) M_s] + \int_0^\ell \eta_i f dx. \tag{5.4.6b}$$

Denoting the element stiffness matrix corresponding to the global ones given in eqn. (5.4.6a) by a superscript e, we have

$$K_{ij}^e = \int_{x_\ell}^{x_r} b \eta_i'' \eta_j'' dx, \tag{5.4.7}$$

where a prime denotes differentiation with respect to x, and x_ℓ and x_r are the x-coordinates of the left and the right nodes of the element. The map (5.3.11) from a typical element to the master element can be written as

$$\begin{aligned} x &= x_\ell \frac{(1-s)}{2} + x_r \frac{(1+s)}{2}, \\ &= x_\ell N_1(s) + x_r N_2(s). \end{aligned} \tag{5.4.8}$$

Here N_1 and N_2 are the Lagrange shape functions for nodes 1 and 2, respectively. Thus shape functions used to map the master element onto the actual element are different from those used to approximate the trial solution w on the actual element. Transforming the variable of integration in the integral of eqn. (5.4.7) from x to s, we obtain

$$K_{ab}^e = \int_{-1}^{1} b\eta_a''\eta_b'' \frac{dx}{ds} ds, \ a,b = 1,2,3,4,$$

where

$$\eta_a'' = \frac{d^2\eta_a}{dx^2} = \frac{d^2\eta_a}{ds^2}\left(\frac{ds}{dx}\right)^2 = \frac{d^2\eta_a}{ds^2}\left(\frac{2}{h_e}\right)^2,$$

$$\begin{aligned} K_{ab}^e &= \left(\frac{4}{h_e^2}\right)^2 \int_{-1}^{1} b(x(s)) \frac{d^2\eta_a}{ds^2}\frac{d^2\eta_b}{ds^2}\left(\frac{h_e}{2}ds\right), \\ &= \frac{8}{h_e^3}\int_{-1}^{1} b(x(s))\frac{d^2\eta_a}{ds^2}\frac{d^2\eta_b}{ds^2}ds. \end{aligned}$$

Henceforth we assume that $b(x) = \text{constant} = \alpha$. Then

$$\begin{aligned} K_{11}^e &= \frac{8\alpha}{h_e^3}\int_{-1}^{1}\left(\frac{3}{2}s\right)^2 ds = \left(\frac{8\alpha}{h_e^3}\frac{9}{4}\right)\left[\frac{s^3}{3}\Big|_{-1}^{1}\right], \\ &= \frac{18\alpha}{h_e^3}\frac{2}{3} = \frac{12\alpha}{h_e}. \end{aligned} \tag{5.4.9}$$

Similarly,

$$\begin{aligned} K_{12}^e &= \frac{8\alpha}{h_e^3}\int_{-1}^{1}\frac{d^2\eta_1}{ds^2}\frac{d^2\eta_2}{ds^2}ds, \\ &= \frac{8\alpha}{h_e^3}\int_{-1}^{1}\left(\frac{3s}{2}\right)\left(\frac{h_e(-1+3s)}{4}\right)ds, \\ &= \frac{3\alpha}{h_e^2}\left[-\frac{s^2}{2}+3\frac{s^3}{3}\right]_{-1}^{1} = \frac{6\alpha}{h_e^2}, \end{aligned} \tag{5.4.10}$$

and one can evaluate other entries in the 4×4 element stiffness matrix K_{ab}^e.

We now evaluate the third term on the right-hand side of the load vector $(5.4.6b)$ for the case of constant f. Denoting the corresponding element term by $\overline{F}_i^e$, we have

$$\begin{aligned} \overline{F}_i^e &= \int_{x_\ell}^{x_r}\eta_i f dx = f\int_{x_\ell}^{x_r}\eta_i dx, \\ &= f\int_{-1}^{1}\eta_i(x(s))\frac{dx}{ds}ds, \\ &= f\frac{h_e}{2}\int_{-1}^{1}\eta_i(s)ds\ . \end{aligned} \tag{5.4.11}$$

Thus

$$\begin{aligned}
\overline{F}_1^e &= f\frac{h_e}{2}\int_{-1}^{1}\frac{1}{4}\left(2-3s+s^3\right)ds = \frac{fh_e}{2}\,,\\
\overline{F}_2^e &= f\frac{h_e}{2}\int_{-1}^{1}\frac{h_e}{8}\left(1-s-s^2+s^3\right)ds = \frac{fh_e^2}{12}\,,\\
\overline{F}_3^e &= \frac{fh_e}{2}\int_{-1}^{1}\frac{1}{4}\left(2+3s-s^2\right)ds = \frac{fh_e}{2}\,,\\
\overline{F}_4^e &= \frac{fh_e}{2}\int_{-1}^{1}\frac{h_e}{8}\left(-1-s+s^2+s^3\right)ds = -\frac{fh_e^2}{12}\,.
\end{aligned} \tag{5.4.12}$$

Recalling that f has units of force/length, units of $\bar{F}_1^e$ and $\bar{F}^e$ are those of force, and units of $\bar{F}_2^e$ and $\bar{F}_3^e$ are those of moment. For a uniformly distributed force, the resultant force at the two nodes equals one-half of the total force on the element. However, moments at the two nodes are equal and opposite.

5.5 Assembly of Element Matrices

In order to assemble element matrices we need the connectivity array that gives destinations of different entries in the global stiffness matrix and the global load vector. We assume that in the FE mesh, nodes are numbered consecutively starting from 1 on the left end. Then the connectivity array for the e^{th} element with the left node numbered e and the right node $(e+1)$ is

$$IC(e,1)=(2e-1),\ IC(e,2)=2e,\ IC(e,3)=2e+1,\ IC(e,4)=(2e+2)\,. \tag{5.5.1}$$

Thus if $i,j(=1,2,\ldots,2n)$ denote the row and the column in the global stiffness matrix corresponding to indices $a,b(=1,2,3,4)$ of the element stiffness matrix K_{ab}^e, we have

$$i=IC(e,a),\ j=IC(e,b)\,. \tag{5.5.2}$$

To assemble the global stiffness matrix we first initialize K_{ij} to be zero, and then for each element e, add K_{ab}^e to K_{ij} through the correspondence established by eqn. (5.5.2). That is

$$\begin{aligned}
&\text{DIMENSION } SK(\cdot,\cdot), SKE(4,4),\ IC(\cdot,\cdot)\\
&\qquad DO\ 100\ I=1,2n\\
&\qquad DO\ 100\ J=1,2n
\end{aligned}$$

$100\ SK(I,J) = 0.0E + 0$

$DO\ 200\ N = 1,\ NELM$

$DO\ 190\ NA = 1, 4$

$I = IC(N, NA)$

$DO\ 190 NB = 1, 4$

$J = IC(N, NB)$

$190\ SK(I,J) = SK(I,J) + SKE(NA, NB)$

200 CONTINUE

Here SK and SKE denote the global and the element stiffness matrices and their dimensions should have been declared in the beginning of the program; NELM equals the number of elements in the mesh.

5.6 Solution of Equations (5.1.1) by using Lagrange Basis Functions

Recalling that Lagrange basis functions developed in Section 4.6 can only be used to solve a second-order DE, we set

$$M = bw'', \tag{5.6.1}$$

and write equations (5.1.1) as

$$\begin{aligned} bw'' - M = 0, \quad 0 < x < \ell, \\ M'' - f = 0, \quad 0 < x < \ell, \end{aligned} \tag{5.6.2}$$

$$w(0) = 0,\ w'(0) = 0,\ M(\ell) = M_0,\ M'(\ell) = F_0. \tag{5.6.3}$$

Thus we need to solve two second-order coupled ordinary DEs for w and M. Note that both BCs on w are now defined at $x = 0$ and those on M at $x = \ell$. It implies that we no longer have two classical BV problems, and we abandon the classification of BCs at the end points into essential and natural BCs. Henceforth we assume that b in eqn. (5.6.1) is a constant.

In contrast to the unknowns as the function w and its derivative w' (i.e., the deflection and the slope), we now have w and M (i.e., the deflection and the bending moment, respectively) as unknowns. The problem with w and M as unknowns is usually referred to as the mixed formulation.

Let ϕ_1 and ϕ_2 be two smooth real valued functions defined on $[0, \ell]$. Multiplication of eqn. $(5.6.2)_1$ with ϕ_1, of eqn. $(5.6.2)_2$ with ϕ_2, the integration of the resulting equations over the domain $[0, \ell]$

and the use of the integration by parts formula yield the following two equations.

$$\begin{aligned}
&\int_0^\ell (bw'\phi_1' + M\phi_1)dx - \phi_1(\ell)bw'(\ell) + \phi_1(0)bw'(0) = 0,\\
&\int_0^\ell \phi_2' M' dx - \phi_2(\ell)M'(\ell) + \phi_2(0)M'(0) = -\int_0^\ell f\phi_2 dx.
\end{aligned} \tag{5.6.4}$$

We now substitute $w'(0) = 0$ and $M'(\ell) = F_0$ from eqn. (5.6.3) into eqn. (5.6.4) to obtain

$$\begin{aligned}
&\int_0^\ell (bw'\phi_1' + M\phi_1)dx - b\phi_1(\ell)w'(\ell) = 0,\\
&\int_0^\ell \phi_2' M' dx + \phi_2(0)M'(0) = F_0\phi_2(\ell) - \int_0^\ell f\phi_2 dx.
\end{aligned} \tag{5.6.5}$$

It is clear that $w,\ \phi_1,\ M$ and ϕ_2 must be in H^1 for the integrals in eqn. (5.6.5) to be well defined. Let H^{1n} be a finite dimensional subset of H^1, and $w^n,\ \phi_1^n,\ M^n$ and ϕ_2^n be in H^{1n}. Then a weak approximation of the given problem can be stated as follows: Find $w^n \in H^{1n},\ M^n \in H^{1n}$ such that $w^n(0) = 0$, $M(\ell) = M_0$, and

$$\begin{aligned}
&\int_0^\ell (bw^{n'}\phi_1^{n'} + M^n\phi_1^n)dx - b\phi_2^n(\ell)w^{n'}(\ell) = 0,\\
&\int_0^\ell M^{n'}\phi_2^{n'} dx + M^{n'}(0)\phi_2^n(0) = F_0\phi_2^n(\ell) - \int_0^\ell f\phi_2^n dx,
\end{aligned} \tag{5.6.6}$$

hold for every $\phi_1 \in H^{1n}$ and $\phi_2 \in H^{2n}$.

In terms of the basis functions $\psi_1(x),\ \psi_2(x), \ldots, \psi_n(x)$ in H^{1n}, we write

$$\begin{aligned}
w^n(x) &= \sum_{j=1}^n w_j\psi_j(x),\ M^n(x) = \sum_{j=1}^n M_j\psi_j(x),\\
\phi_1^n(x) &= \sum_{i=1}^n C_i^1\psi_i(x),\ \phi_2^n(x) = \sum_{i=1}^n C_i^2\psi_i(x),
\end{aligned} \tag{5.6.7}$$

and exploiting the condition that equations (5.6.6) hold for all choices of ϕ_1^n and ϕ_2^n, we obtain

$$\begin{aligned}
&\left(\int_0^\ell b\psi_i'\psi_j' dx\right) w_j + \left(\int_0^\ell \psi_i\psi_j dx\right) M_j - b\psi_i(\ell)w^{n'}(\ell) = 0,\\
&\left(\int_0^\ell \psi_i'\psi_j' dx\right) M_j + \psi_i(0)M^{n'}(0) = F_0\psi_i(\ell) - \int_0^\ell f\psi_i dx.
\end{aligned} \tag{5.6.8}$$

Let

$$K_{ij} = \int_0^\ell \psi_i'\psi_j' dx,\ \alpha_{ij} = \int_0^\ell \psi_i\psi_j dx,\ F_i = -\int_0^\ell f\psi_i dx, \tag{5.6.9}$$

then for a uniform FE mesh of 2 elements and 3 nodes with nodes numbered consecutively from left to right, equations (5.6.8) become

$$\begin{bmatrix} bK_{11} & \alpha_{11} & bK_{12} & \alpha_{12} & bK_{13} & \alpha_{13} \\ 0 & K_{11} & 0 & K_{12} & 0 & K_{13} \\ bK_{21} & \alpha_{21} & bK_{22} & \alpha_{22} & bK_{23} & \alpha_{23} \\ 0 & K_{21} & 0 & K_{22} & 0 & K_{23} \\ bK_{31} & \alpha_{31} & bK_{32} & \alpha_{32} & bK_{33} & \alpha_{33} \\ 0 & K_{31} & 0 & K_{32} & 0 & K_{33} \end{bmatrix} \begin{Bmatrix} w_1 \\ M_1 \\ w_2 \\ M_2 \\ w_3 \\ M_3 \end{Bmatrix} = \begin{Bmatrix} 0 \\ F_1 - M^{n'}(0) \\ 0 \\ F_2 \\ bw^{n'}(\ell) \\ F_0 + F_3 \end{Bmatrix}. \tag{5.6.10}$$

We now impose the boundary condition $w(0) = w_1 = 0$ in the 5$^{\text{th}}$ equation, and $M(\ell) = M_3 = M_0$ in the 2$^{\text{nd}}$ equation. The reason for these choices is that $w^{n'}(\ell)$ in the 5$^{\text{th}}$ equation and $M^{n'}(0)$ in the 2$^{\text{nd}}$ equation are unknown. Note that equations (5.6.10) are 6 algebraic equations which need to be solved simultaneously for the six unknowns $w_1,\ M_1,\ w_2,\ M_2,\ w_3$ and M_3. Since there is no w_1 appearing in the 2$^{\text{nd}}$ equation, it can not be used for satisfying $w_1 = 0$. Even though w_1 appears in the 1$^{\text{st}}$ and 3$^{\text{rd}}$ equations, using either one of them to satisfy $w_1 = 0$ will not eliminate the unknown $M^{n'}(0)$ or $w^{n'}(\ell)$ from the right-hand side of equations (5.6.10).

Using method A outlined in Section 4.4 to enforce the boundary conditions $w_1 = 0,\ M_3 = M_0$, we arrive at the following set of algebraic equations.

$$\begin{bmatrix} 0 & \alpha_{11} & bK_{12} & \alpha_{12} & bK_{13} & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & \alpha_{21} & bK_{22} & \alpha_{22} & bK_{23} & 0 \\ 0 & K_{21} & 0 & K_{22} & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & K_{31} & 0 & K_{32} & 0 & 0 \end{bmatrix} \begin{Bmatrix} w_1 \\ M_1 \\ w_2 \\ M_2 \\ w_3 \\ M_3 \end{Bmatrix} = \begin{Bmatrix} 0 \\ M_0 \\ 0 \\ F_2 \\ 0 \\ F_0 + F_3 \end{Bmatrix}. \tag{5.6.11}$$

It should be emphasized that the essential BC $w'(0) = 0$ of the original problem may not be exactly satisfied by using the mixed formulation of this section. However, the natural BC $(bw'')'\Big|_{x=\ell} = M_0$ of the original problem is now exactly satisfied.

When b is not a constant, then the weak formulation of the original fourth order DE involves terms symmetric in ϕ and w. If we proceed as usual, the weak formulation of the two second order DEs will involve derivatives of b and will be unsymmetric in ϕ and w thereby resulting in a nonsymmetric stiffness matrix. However, it can be rendered symmetric by taking the test or the weight function for equation (5.6.2)$_1$ as ϕ_1/b.

Chapter 6: Numerical Integration

As should be evident from the evaluation of the element stiffness matrix and the element load vector discussed in Sections 4.3 and 5.4, we need to evaluate integrals of the type $\int_{-1}^{1} f(x)dx$. If the coefficients of $u,\ u',\ u''$ etc. and the non-homogeneous term in the given DE are complicated functions of x, then the integrand $f(x)$ will not be a polynomial function of x. However, in the FE work we normally approximate coefficients of $u,\ u'$ etc. by their interpolants obtained by using the FE basis functions. Also when the variable of integration is changed to the local coordinate on the master element, the integrals are evaluated over $[-1, 1]$. We first review two methods learned earlier in a Calculus course to evaluate these integrals.

6.1 Trapezoidal Rule

The interval $[-1, 1]$ is divided into n subintervals, usually of equal length, and on each subinterval the function f is approximated by a straight line. Let $s_1 = -1, s_2, \ldots, s_{n+1} = 1$ be points on $[-1, 1]$. Then over the interval (s_i, s_{i+1}) the function $f(s)$ is approximated by a straight line joining $(s_i, f(s_i))$ and $(s_{i+1}, f(s_{i+1}))$ and the integral $\displaystyle\int_{s_i}^{s_{i+1}} f(s)ds$ is approximated by the area of the trapezoid with height $(s_{i+1} - s_i)$ and the lengths of the two sides as $f(s_i)$ and $f(s_{i+1})$. Thus

$$\begin{aligned} I =& \int_{-1}^{1} f(s)ds \simeq \sum_{i=1}^{n} \frac{f(s_{i+1}) + f(s_i)}{2}(s_{i+1} - s_i), \\ =& \frac{h}{2}f(s_1) + hf(s_2) + hf(s_3) + \ldots + hf(s_n) + \frac{h}{2}f(s_{n+1}); \qquad (6.1.1) \\ h =& (s_{i+1} - s_i) = (s_{n+1} - s_1)/n. \end{aligned}$$

We will call $s_1, s_2, \ldots, s_{n+1}$ the sampling points, i.e., points where the value of the function is determined and these values are used to evaluate the integral; $\frac{h}{2}, h, h, \ldots, \frac{h}{2}$ are called weights. With

$$W_1 = \frac{h}{2},\ W_2 = W_3 = \ldots = W_n = h,\ W_{n+1} = \frac{h}{2}, \qquad (6.1.2)$$

$$I \simeq \sum_{i=1}^{n+1} W_i f(s_i). \qquad (6.1.3)$$

Of course, one way to improve upon the accuracy in evaluating I is to increase the number of sampling points; this will result in higher computational cost. We note that the trapezoidal rule integrates exactly a polynomial of degree one by using two sampling points.

6.2 Simpson's Rule

The interval $[-1,1]$ is divided into $(2n+1)$ sampling points and over the interval (s_{i-1}, s_{i+1}), $f(s)$ is approximated by a parabola passing through $(s_{i-1}, f(s_{i-1}))$, $(s_i, f(s_i))$ and $(s_{i+1}, f(s_{i+1}))$. The integral $I = \int_{s_{i-1}}^{s_{i+1}} f(s)ds$ is approximated by the area under the parabola. Thus

$$I \simeq \frac{f(s_{i-1}) + 4f(s_i) + f(s_{i+1})}{3}\left(\frac{s_{i+1} - s_{i-1}}{2}\right). \tag{6.2.1}$$

If the interval $[-1,1]$ is divided into $2n$ subintervals of equal length, then

$$I \simeq \frac{h}{3}f(s_1) + \frac{4h}{3}f(s_2) + \frac{2h}{3}f(s_3) + \frac{4h}{3}f(s_4) + \ldots + \frac{h}{3}f(s_{2n+1}), \tag{6.2.2}$$

$$= \sum_{i=1}^{2n+1} W_i f(s_i), \tag{6.2.3}$$

where $W_1 = \frac{h}{3},\ W_2 = \frac{4h}{3},\ W_3 = \frac{2h}{3},\ W_4 = \frac{4h}{3}, \ldots, W_{2n+1} = \frac{h}{3}$. The Simpson rule integrates exactly a polynomial of degree 2 by using three sampling points. However, both the trapezoidal rule and the Simpson rule require more sampling points than are needed to integrate exactly polynomials of degree one and two respectively. The sampling points are also called quadrature points.

6.3 Gauss-Quadrature Rule

We now investigate the minimum number of sampling points required to integrate exactly a polynomial function and find the corresponding weights. The first attempt is essentially ad-hoc; subsequently we give an elegant solution to the problem.

When $f(s)$ is a polynomial of degree 1, i.e., $f(s) = a + bs$, we have

$$I = \int_{-1}^{1} (a + bs)ds = 2a = 2f(0). \tag{6.3.1}$$

Thus $s_1 = 0$ and $W_1 = 2$ will evaluate exactly the integral from -1 to $+1$ of a polynomial function of degree 1. Recall that the trapezoidal rule integrates exactly a polynomial of degree one by using two quadrature points. Now consider $f(s) = a + bs + cs^2 + ds^3$. Let us assume that two sampling points are needed; then we need to find W_1, s_1, W_2, s_2 such that

$$\int_{-1}^{1} f(s)ds = W_1 f(s_1) + W_2 f(s_2), \tag{6.3.2}$$

i.e.,

$$2a + \frac{2}{3}c = W_1\left(a + bs_1 + cs_1^2 + ds_1^3\right) + W_2\left(a + bs_2 + cs_2^2 + ds_2^3\right). \tag{6.3.3}$$

Since (6.3.3) has to hold for every choice of a, b, c and d, therefore

$$\begin{aligned} 2 &= W_1 + W_2, \\ 0 &= W_1 s_1 + W_2 s_2, \\ \frac{2}{3} &= W_1 s_1^2 + W_2 s_2^2, \\ 0 &= W_1 s_1^3 + W_2 s_2^3. \end{aligned} \tag{6.3.4}$$

These algebraic equations are nonlinear and difficult to solve in general. We assume that the sampling points are symmetrically located around the origin, i.e., $s_1 = -s_2$. Then $(6.3.4)_2$ and $(6.3.4)_4$ imply that $W_1 = W_2$ which together with $(6.3.4)_1$ gives $W_1 = W_2 = 1$, and $(6.3.4)_3$ gives $-s_1 = s_2 = \sqrt{\frac{1}{3}}$. Thus

$$\int_{-1}^{1} f(s)ds = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right) \tag{6.3.5}$$

for a polynomial of degree 3. We note that Simpson's rule requires three sampling points to integrate exactly a polynomial of degree 3.

For a fifth order polynomial, it is reasonable to assume that we need three sampling points. That is, for $f(s) = a + bs + cs^2 + ds^3 + es^4 + gs^5$,

$$\int_{-1}^{1} f(s)ds = W_1 f(s_1) + W_2 f(s_2) + W_3 f(s_3). \tag{6.3.6}$$

If it has to hold for every choice of a, b, c, d, e and g, then we must have

$$\begin{aligned} 2 &= W_1 + W_2 + W_3, \\ 0 &= W_1 s_1 + W_2 s_2 + W_3 s_3, \\ \frac{2}{3} &= W_1 s_1^2 + W_2 s_2^2 + W_3 s_3^2, \\ 0 &= W_1 s_1^3 + W_2 s_2^3 + W_3 s_3^3, \\ \frac{2}{5} &= W_1 s_1^4 + W_2 s_2^4 + W_3 s_3^4, \\ 0 &= W_1 s_1^5 + W_2 s_2^5 + W_3 s_3^5. \end{aligned} \tag{6.3.7}$$

In order to solve (6.3.7) we assume that $s_2 = 0$ and $s_1 = -s_3$, i.e., $s = 0$ is a sampling point and the other two points are symmetrically located around it. Equation $(6.3.7)_2$ or $(6.3.7)_4$ or $(6.3.7)_6$ gives $W_1 = W_3$, and $(6.3.7)_3$ and $(6.3.7)_5$ yield $-s_1 = s_3 = \sqrt{3}/\sqrt{5}$. Substitution for s_1, s_2 and s_3 in $(6.3.7)_3$ and using $W_1 = W_3$, we obtain $W_1 = W_3 = 5/9$ and then $(6.3.7)_1$ gives $W_2 = 8/9$. Thus, for a polynomial of degree 5,

$$\int_{-1}^{1} f(s)ds = \frac{5}{9} f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9} f(0) + \frac{5}{9} f\left(\sqrt{\frac{3}{5}}\right). \tag{6.3.8}$$

In summary, to integrate polynomials of degree 1, 3 and 5, we need 1, 2 and 3 sampling points. Thus for a polynomial of degree $(2n+1)$ we need $(n+1)$ sampling points. We note that the sum of the weights always equals 2; this is so because all of these rules must integrate exactly a polynomial of degree zero or a constant function. As should be clear from the above stated procedure, the algebra involved gets more tedious as the degree of the polynomial to be integrated increases. We now discuss the more general case.

Let $f(s)$ be a polynomial of degree $(2n+1)$ that is to be integrated on the interval $[-1,1]$. We note that the Legendre polynomial of degree n is defined by

$$P_n(s) = \frac{1}{2^n n!}\frac{d^n(s^2-1)^n}{ds^n}, \tag{6.3.9}$$

and $P_0(s), P_1(s), \ldots, P_n(s)$ are linearly independent and are mutually orthogonal in the sense that

$$\int_{-1}^{1} P_i(s)P_j(s)ds = 0, \text{ if } i \neq j. \tag{6.3.10}$$

Note that $P_0(s) = 1$, $P_1(s) = s$, $P_2(s) = (3s^2-1)/2$, $P_3(s) = (5s^3-3s)/2$, $P_4(s) = (35s^4-30s^2+3)/8$ and $P_5(s) = (63s^5-70s^3+15s)/8$. We can divide $f(s)$ by $P_{n+1}(s)$ to obtain the quotient $b(s)$ and the remainder $r(s)$. Both $b(s)$ and $r(s)$ will be polynomials of degree n. Thus

$$f(s) = b(s)P_{n+1}(s) + r(s). \tag{6.3.11}$$

We use $P_0(s), P_1(s), P_2(s), \ldots, P_n(s)$ as basis functions to write

$$b(s) = \sum_{i=0}^{n} c_i P_i(s), \tag{6.3.12}$$

and use Lagrange basis functions $N_1(s), N_2(s), \ldots, N_{n+1}(s)$ to write

$$r(s) = \sum_{i=1}^{n+1} r(s_i)N_i(s), \tag{6.3.13}$$

where $s_1, s_2, s_3, \ldots, s_{n+1}$ are in $[-1,1]$ and are yet to be determined. Substitution from (6.3.12) and (6.3.13) into (6.3.11), and integration of the resulting equation from -1 to 1 yield

$$\int_{-1}^{1} f(s)ds = \int_{-1}^{1} b(s)\left(\sum_{i=0}^{n} c_i P_i(s)\right)P_{n+1}(s)ds + \int_{-1}^{1}\sum_{i=1}^{n+1} r(s_i)N_i(s)ds = \sum_{i=1}^{n+1} r(s_i)W_i, \tag{6.3.14}$$

where

$$W_i = \int_{-1}^{1} N_i(s)ds, \tag{6.3.15}$$

and we have used (6.3.10).

We select $s_1, s_2, \ldots, s_{n+1}$ to be roots of

$$P_{n+1}(s) = 0. \tag{6.3.16}$$

Then, from (6.3.11), $f(s_i) = r(s_i)$ and (6.3.14) becomes

$$\int_{-1}^{1} f(s)ds = \sum_{i=1}^{n+1} f(s_i)W_i, \tag{6.3.17}$$

where W_i is defined by (6.3.15). Recall that

$$N_i(s) = \prod_{\substack{j=1 \\ j\neq i}}^{n+1}(s - s_j)/\prod_{\substack{j=1 \\ j\neq i}}^{n+1}(s_i - s_j). \tag{6.3.18}$$

Thus to integrate exactly a polynomial of degree $(2n+1)$ on $[-1,1]$ we need $(n+1)$ sampling points which are roots of $P_{n+1}(s) = 0$ where $P_{n+1}(s)$ is a Legendre polynomial of degree $(n+1)$ in s. The corresponding weights are values of $\int_{-1}^{1} N_i(s)ds$, $N_i(s)$ being the ith Lagrange shape function. To see that we get the same results as those obtained by the ad-hoc procedure, we take $n = 2$. Thus $f(s)$ is a polynomial of degree 5. Equation (6.3.9) gives

$$P_3(s) = \left(5s^3 - 3s\right)/2, \tag{6.3.19}$$

and $P_3(s) = 0$ has roots $-s_1 = s_3 = \sqrt{3}/\sqrt{5}$, $s_2 = 0$. The corresponding Lagrange shape functions are

$$N_1(s) = \frac{5}{6}\left(s\left(s - \sqrt{\frac{3}{5}}\right)\right),$$
$$N_2(s) = \frac{5}{3}\left(-s^2 + \frac{3}{5}\right),$$
$$N_3(s) = \frac{5}{6}\left(s\left(s + \frac{\sqrt{3}}{\sqrt{5}}\right)\right), \tag{6.3.20}$$

and

$$W_1 = \int_{-1}^{1} \frac{5}{6}\left(s^2 - \sqrt{\frac{3}{5}}s\right)ds = \frac{5}{9},$$

$W_2 = 8/9$ and $W_3 = 5/9$, which are the same as those obtained before. The number of sampling points obtained herein is the minimum needed to integrate exactly a polynomial function over the

interval $[-1, 1]$. These sampling points are called Gauss points or Gauss quadrature points and the corresponding quadrature rule is termed as the Gaussian rule. The weights and sampling points for integrating an even order polynomial on $[-1, 1]$ are the same as that for the next odd order polynomial.

In general, the ratio of two polynomial functions will not be a polynomial function. Should the integrand be not a polynomial function, then one should experiment with an increasing number of Gauss points till the difference between two successive values obtained is less than the acceptable tolerance.

The sampling points and the corresponding weights for integrating polynomials of different order are listed in Table 6.1 below.

Table 6.1: ABSCISSAE AND WEIGHT COEFFICIENTS OF THE GAUSSIAN QUADRATURE FORMULA $\int_{-1}^{1} f(x)dx = \sum_{j=1}^{n} H_i f(a_j)$

$\pm a$	H
$n = 1$	
0	2.000 000 000 000 000
$n = 2$	
0.577 350 269 189 626	1.000 000 000 000 000
$n = 3$	
0.774 596 669 241 483	0.555 555 555 555 556
0.000 000 000 000 000	0.888 888 888 888 889
$n = 4$	
0.861 136 311 594 953	0.347 854 845 137 454
0.339 981 043 584 856	0.652 145 154 862 546
$n = 5$	
0.906 179 845 938 664	0.236 926 885 056 189
0.538 469 310 105 683	0.478 628 670 499 366
0.000 000 000 000 000	0.568 888 888 888 889
$n = 6$	
0.932 469 514 203 152	0.171 324 492 379 170
0.661 209 386 466 265	0.360 761 573 048 139
0.238 619 186 083, 197	0.467 913 934 572 691
$n = 7$	
0.949 107 912 342 759	0.129 484 966 168 870
0.741 531 185 599 394	0.279 705 391 489 277
0.405 845 151 377 397	0.381 830 050 505 119
0.000 000 000 000 000	0.417 959 183 673 469
$n = 8$	
0.960 289 856 497 536	0.101 228 536 290 376
0.796 666 477 413 627	0.222 381 034 453 374
0.525 532 409 916 329	0.313 706 645 877 887
0.183 434 642 495 650	0.362 683 783 378 362
$n = 9$	
0.968 160 239 507 626	0.081 274 388 361 574
0.836 031 107 326 636	0.180 648 160 694 857
0.613 371 432 700 590	0.260 610 696 402 935
0.324 253 423 403 809	0.312 347 077 040 003
0.000 000 000 000 000	0.330 239 355 001 260
$n = 10$	
0.973 906 528 517 172	0.066 671 344 308 688
0.865 063 366 688 985	0.149 451 349 150 581
0.679 409 568 299 024	0.219 086 362 515 982
0.433 395 394 129 247	0.269 266 719 309 996
0.148 874 338 981 631	0.295 524 224 714 753

Chapter 7: Two-Dimensional Problems

7.1 A Model Problem

As a typical two-dimensional problem we analyze the following BV problem:

$$-\nabla\cdot(k\nabla u)+\alpha u=f \text{ in } \Omega\,, \tag{7.1.1}$$

$$-k\frac{\partial u}{\partial n}=\gamma\,(u-u_0) \text{ on } \Gamma_1\,, \tag{7.1.2}$$

$$u=\hat{u} \text{ on } \Gamma_2\,. \tag{7.1.3}$$

Here ∇ is the gradient operator ($\nabla = \hat{i}\frac{\partial}{\partial x}+\hat{j}\frac{\partial}{\partial y}$ in rectangular Cartesian coordinates where $\hat{i}$ and $\hat{j}$ are unit vectors along x- and y-axes, respectively), k is the thermal conductivity, γ is the heat transfer coefficient between the body and the surroundings, Ω is a regular bounded domain, and Γ_1 and Γ_2 are complementary parts of the boundary of Ω where heat flux and temperature are prescribed, respectively. u_0 may be thought of as the ambient temperature and $\hat{u}$ is a known function of (x,y) for points on Γ_2. $\mathbf{A}\cdot\mathbf{B}$ denotes the inner product between two vectors $\mathbf{A}$ and $\mathbf{B}$. $k,\ \alpha,\ f,\ \gamma$ and u_0 are known functions of x and y, and $\frac{\partial u}{\partial n}=\nabla u\cdot\mathbf{n}$ where $\mathbf{n}$ is a unit outward normal to the boundary, $\partial\Omega$, of Ω. We note that the given BV problem involves a partial DE rather than an ordinary DE and the highest order derivative appearing in it is two. Therefore, BCs involving derivatives of order zero $(m-1=0)$ and lower are essential and others are natural; the BC (7.1.3) is essential and (7.1.2) is natural. We follow the procedure similar to that used to analyze the one-dimensional problem.

7.2 Weak Formulation

Let $\phi:\overline{\Omega}\rightarrow\mathbb{R}$ be a smooth function. Multiplying both sides of eqn. (7.1.1) by ϕ and integrating the result over the domain Ω we obtain

$$-\int_\Omega \nabla\cdot(k\nabla u)\,\phi d\Omega+\int_\Omega \alpha u\phi d\Omega=\int_\Omega f\phi d\Omega\,. \tag{7.2.1}$$

By using the chain rule of differentiation, we get

$$\int_\Omega \nabla\cdot(k\nabla u)\,\phi d\Omega=\int_\Omega \nabla\cdot(\phi k\nabla u)\,d\Omega-\int_\Omega \nabla\phi\cdot k\nabla u d\Omega\,. \tag{7.2.2}$$

The use of the divergence theorem yields

$$\int_\Omega \nabla\cdot(\phi k\nabla u)\,d\Omega = \int_{\partial\Omega}\phi k\nabla u\cdot\mathbf{n}d\Gamma = \int_{\partial\Gamma}\phi k\frac{\partial u}{\partial n}d\Gamma,$$
$$= -\int_{\Gamma_1}\phi\gamma\,(u-u_0)\,d\Gamma - \int_{\Gamma_2}\phi q_s d\Gamma, \tag{7.2.3}$$

where we have set $-k\frac{\partial u}{\partial n} = q_s$ on Γ_2 and used the natural BC (7.1.2). We note that q_s is not known on Γ_2 but $u = \hat{u}$ is given there. Substitution from eqn. (7.2.3) into eqn. (7.2.2) and the result into eqn. (7.2.1) gives

$$\int_\Omega(\nabla\phi\cdot k\nabla u + \alpha u\phi)\,d\Omega + \int_{\Gamma_1}\phi\gamma u d\Gamma = \int_\Omega f\phi d\Omega + \int_{\Gamma_1}\gamma\phi u_0 d\Gamma - \int_{\Gamma_2}\phi q_s d\Gamma\,. \tag{7.2.4}$$

With the definitions

$$B\,(\phi,u) \equiv \int_\Omega(k\nabla\phi\cdot\nabla u + \alpha\phi u)\,d\Omega + \int_{\Gamma_1}\gamma\phi u d\Gamma\,, \tag{7.2.5}$$

$$\ell\,(\phi) \equiv \int_\Omega f\phi d\Omega + \int_{\Gamma_1}\gamma\phi u_0 d\Gamma - \int_{\Gamma_2}\phi q_s d\Gamma, \tag{7.2.6}$$

we can write eqn. (7.2.4) as

$$B\,(\phi,u) = \ell(\phi). \tag{7.2.7}$$

$B(\cdot,\cdot)$ is called a bilinear form since it is linear in each of its arguments and $\ell(\phi)$ is a linear functional of ϕ. In order for the integrals appearing in the definitions of $B(\cdot,\cdot)$ and $\ell(\cdot)$ to be finite, $\int_\Omega|\nabla\phi|^2 d\Omega$, $\int_\Omega\phi^2 d\Omega$ and $\int_{\partial\Gamma}\phi^2 d\Gamma$ should be finite. Because of Poincaré's inequality (4.2.6) and the trace theorem, it is sufficient to require that $\int_\Omega|\nabla\phi|^2 d\Omega$ be finite provided that Γ_2 is a nonempty set. Let

$$H^1 = \left\{\psi|\psi:\overline{\Omega}\to\mathbb{R},\ \int_\Omega\nabla\psi\cdot\nabla\psi d\Omega<\infty\right\}, \tag{7.2.8}$$

We can now state the weak form, W, of the given problem as follows.

W: Find $u\in H^1$ such that $u=\hat{u}$ on Γ_1 and eqn. (7.2.7) holds for every $\phi\in H^1$.

Let H^{1n} be a finite dimensional subset of H^1 and $v^n\in H^{1n}$, $\phi^n\in H^{1n}$. Then the Galerkin approximation of the given problem is: find $u^n\in H^{1n}$ such that $u^n=\hat{u}$ on Γ_2 and

$$B\,(\phi^n,u^n) = \ell\,(\phi^n)\,, \tag{7.2.9}$$

for every $\phi^n\in H^{1n}$.

Let $\psi_1,\psi_2,\ldots,\psi_n$ denote a set of basis functions in H^{1n}. Then we can write

$$\phi^n = c_i\psi_i,\ u^n = d_j\psi_j, \tag{7.2.10}$$

where summation on a repeated index is implied and indices i and j range over 1 through n. The linearity of $B(\cdot,\cdot)$ in each of its arguments implies that

$$B\left(\phi^n, u^n\right) = B\left(c_i\psi_i, d_j\psi_j\right) = c_i B\left(\psi_i, \psi_j\right) d_j = c_i K_{ij} d_j, \tag{7.2.11}$$

where

$$K_{ij} = B\left(\psi_i, \psi_j\right). \tag{7.2.12}$$

Similarly, with

$$F_i = \ell(\psi_i),\ \ell(\phi^n) = \ell(c_i\psi_i) = c_i\ell(\psi_i) = c_i F_i\ , \tag{7.2.13}$$

eqn. (7.2.9) becomes

$$c_i K_{ij} d_j = c_i F_i. \tag{7.2.14}$$

Since eqn. (7.2.9) holds for every ϕ^n, therefore eqn. (7.2.14) should hold for every choice of $c_1, c_2, \ldots, c_n$ which is possible if and only if

$$K_{ij} d_j = F_i,\ i = 1, 2, \ldots, n. \tag{7.2.17}$$

Matrices $\mathbf{K}$ and $\mathbf{F}$ are called the stiffness matrix and the load vector, respectively. For our problem,

$$B\left(\phi, u\right) = B\left(u, \phi\right), \tag{7.2.16}$$

i.e., B is a symmetric bilinear form, therefore,

$$K_{ij} = K_{ji}, \tag{7.2.17}$$

i.e., the stiffness matrix is symmetric. Also if $k\left(x, y\right) > 0$ and $\alpha\left(x, y\right) > 0$ at every point $(x, y) \in \Omega$, then

$$B\left(\phi, \phi\right) > 0 \tag{7.2.18}$$

for every $\phi \not\equiv 0$ and the bilinear form is positive definite. Thus if $k(x, y) > 0$ and $\alpha(x, y) > 0$ everywhere in Ω, then the stiffness matrix will be positive definite.

Recall that the right-hand side of eqn. (7.2.15) involves unknowns q_s on Γ_2. Equations having $\int_{\Gamma_2} q_s\psi_i d\Gamma$ as part of F_i are modified by one of the two methods discussed in Section to satisfy the essential boundary condition $u = \hat{u}$ on Γ_2.

Remarks:

1. The quality of the approximate solution depends upon the choice of basis functions $\psi_1, \psi_2, \ldots, \psi_n$.
2. The natural BC (7.1.2) is embedded in the weak formulation of the problem, the essential BC (7.1.3) needs to be explicitly satisfied.
3. As for a one-dimensional problem, the FE basis function for node i is obtained by patching together the shape functions for node i corresponding to different elements that meet at node i. Thus we need to first find shape functions for a node.

7.3 Finite Element Shape Functions and Basis Functions

In one-dimensional problems the domain can be divided into subdomains or finite elements such that the union of subdomains equals exactly the given domain. However, in two-dimensional problems this need not hold. Also, the choice of types of subdomains or finite elements (e.g.. triangular, quadrilateral etc.) is wider in two-dimensional problems as compared to that in one-dimensional problems. Recalling that basis functions can be obtained by patching together appropriate shape functions, we construct here shape functions. The shape functions that correspond to continuous global basis functions must satisfy the following conditions: (i) the shape function N_i for node i must equal one at node i, (ii) N_i vanishes at all other nodes of the element, and (iii) N_i vanishes on all sides of the element that do not pass through node i. We generally require that the shape functions are polynomials of the lowest degree.

The simplest two-dimensional element is a triangular element with three nodes; one at each vertex. We now develop expressions for shape functions for this element.

Let three nodes of a triangular element be identified as i, j, k; then the shape function N_i for node i should equal 1 at node i and zero at nodes j and k. Since N_i has to satisfy three conditions, we assume that

$$N_i(x, y) = a + bx + cy, \tag{7.3.1}$$

where constants $a,\ b,\ c$ are to be determined from

$$N_i\left(x_i, y_i\right) = 1,\ \text{(no sum on } i\text{)}, \tag{7.3.2}$$

$$N_i\left(x_j, y_j\right) = 0, \tag{7.3.3}$$

$$N_i\left(x_k, y_k\right) = 0. \tag{7.3.4}$$

The evaluation of constants a, b, and c from eqns. (7.3.2)-(7.3.4) and the substitution of the result into eqn. (7.3.1) yields

$$N_i(x, y) = \frac{1}{2\Delta}\left[\left(x_j y_k - x_k y_j\right) + x\left(y_j - y_k\right) + y\left(x_k - x_j\right)\right], \tag{7.3.5}$$

where

$$2\Delta = \left(x_i - x_k\right)\left(y_j - y_k\right) - \left(x_j - x_k\right)\left(y_i - y_k\right), \tag{7.3.6}$$

and Δ is the area of the triangular element. Similarly, one can derive expressions for N_j and N_k. Alternatively, one can obtain these from eqn. (7.3.5) by a cyclic permutation of indices i, j, k; the result is

$$N_j = \frac{1}{2\Delta}\left[\left(x_k y_i - x_i y_k\right) + x\left(y_k - y_i\right) + y\left(x_i - x_k\right)\right], \tag{7.3.7}$$

$$N_k = \frac{1}{2\Delta}\left[\left(x_i y_j - x_j y_i\right) + x\left(y_i - y_j\right) + y\left(x_j - x_i\right)\right]. \tag{7.3.8}$$

The shape function N_i is exhibited in Fig. 7.3.1. Since N_i equals zero at nodes j and k and is an

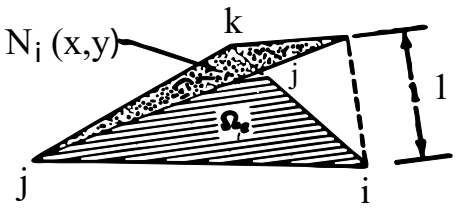


Fig. 7.3.1: Shape function for node i of a triangular element

affine function of x and y, it must be zero everywhere on the line joining nodes j and k. Thus N_i satisfies all conditions laid out above for a shape function. The basis function ψ_i corresponding to node i is obtained by patching together shape functions N_i's; one for every element that meets at node i. For example, in Fig. 7.3.2, five triangular elements meet at node 6 and the corresponding basis function looks like a tent with unit height and five triangular surfaces.

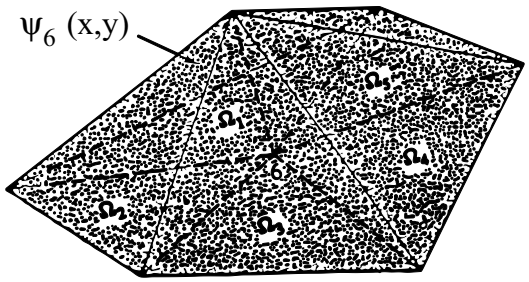


Fig. 7.3.2: Basis function for node 6 common to triangular elements $\Omega_1, \Omega_2, \Omega_3, \Omega_4$ and Ω_5.

In terms of the FE basis functions, the approximate solution can be written as

$$u^m = \sum_{j=1}^{m} d_j \psi_j + \sum_{j=m+1}^{n} d_j \psi_j, \tag{7.3.9}$$

where nodes $m+1$ through n are assumed to be on the boundary Γ_2. Since

$$\psi_i \left(x_j, y_j\right) = \delta_{ij}, \tag{7.3.10}$$

therefore,

$$u^m \left(x_j, y_j\right) = d_j \equiv u_j, \ j = 1, 2, \ldots, m, \tag{7.3.11}$$

for points in Ω, and

$$u^m \left(x_j, y_j\right) = d_j = \hat{u} \left(x_j, y_j\right) \equiv u_j, \ j = m+1, m+2, \ldots, n, \tag{7.3.12}$$

for points on Γ_2. Thus the value of d_j equals that of the approximate solution at node j, and we will henceforth identify d_j with u_j - the value of the approximate solution at node j. Thus

$$u^n = \sum_{j=1}^{n} u_j \psi_j. \tag{7.3.13}$$

When we restrict ourselves to an element, eqn. (7.3.13) becomes

$$u^e(x, y) = u_i N_i(x, y) + u_j N_j(x, y) + u_k N_k(x, y), \ \text{(no sum on } i, j, k),$$

where nodes have been identified as i, j and k, and, therefore, no summation is implied on repeated indices. Since N_i, N_j, N_k are affine functions of x and y, $\frac{\partial u^e}{\partial x}$ and $\frac{\partial u^e}{\partial y}$ will be constants. If u^e denotes the displacement of a point, then $\frac{\partial u^e}{\partial x}$ and $\frac{\partial u^e}{\partial y}$ define components of the strain tensor at a point. Thus components of the strain tensor will be constants throughout the element. For this reason this element is usually referred to as a constant strain triangular (CST) element.

The shape functions for the CST element are complete polynomials of degree one in x and y. The next higher order triangular element should have shape functions which are complete polynomials of degree two. Such a polynomial has six terms in it and therefore we should have six conditions for their determination. This in turn necessitates that we have six nodes on the triangular element. Even though one can follow the procedure outlined above for the determination of shape functions for a six-noded triangular element, it is rather cumbersome. This provides a motivation for introducing a master triangular element and area coordinates.

A master triangular element is a right-angle triangle of unit base and unit height. The area coordinates (r, s, t) of a point P are defined as

$$r = \text{area } P23/\text{area } 123;\ s = \text{area } P31/\text{area } 123,\ t = \text{area } P12/\text{area } 123. \tag{7.3.14}$$

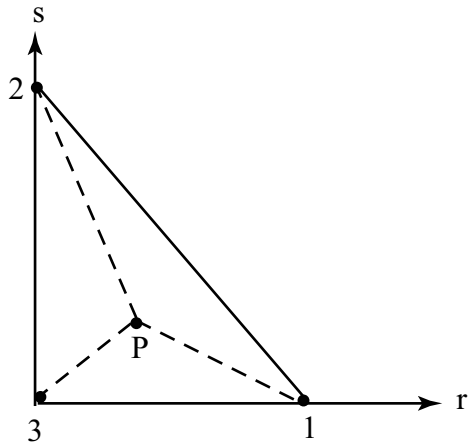


Fig. 7.3.3: Master triangular element

Since

$$r + s + t = 1, \tag{7.3.15}$$

r, s, t are not independent of each other. Lines $r = \text{const.}$ are parallel to the side 23; $r = 0$ coincides with side 23 and $r = 1$ passes through node 1. Similarly, lines $s = \text{const.}$ and $t = \text{const.}$ are parallel to sides 13 and 12, respectively. In terms of area co-ordinates the shape functions for nodes 1, 2 and 3 are given by

$$N_1 = r,\ N_2 = s,\ \text{and } N_3 = t. \tag{7.3.16}$$

We now need to find a transformation T_e that will map the master triangular element onto an actual element. This transformation T_e should be simple, 1-1, onto, differentiable and should have

a differentiable inverse. One such transformation is

$$\begin{aligned} x =& x_i N_1 + x_j N_2 + x_k N_3 = x_i r + x_j s + x_k t, \\ T_e: \quad =& x_k + (x_i - x_k)\, r + (x_j - x_k)\, s, \\ y =& y_i N_1 + y_j N_2 + y_k N_3 = y_k + (y_i - y_k)\, r + (y_j - y_k)\, s, \end{aligned} \tag{7.3.17}$$

where $(x_i, y_i), (x_j, y_j)$ and (x_k, y_k) are, respectively, coordinates of nodes i, j and k of the element onto which the master triangle is mapped. It is evident that T_e is uniquely defined for each element, and is 1-1. A necessary and sufficient condition for T_e to have a differentiable inverse is that

$$J \equiv \det \begin{bmatrix} \dfrac{\partial x}{\partial r} & \dfrac{\partial x}{\partial s} \\ \dfrac{\partial y}{\partial r} & \dfrac{\partial y}{\partial s} \end{bmatrix} \neq 0, \tag{7.3.18}$$

everywhere in the master element; J is called the Jacobian of the transformation T_e. For the transformation defined by eqn. (7.3.17),

$$J = 2\Delta, \tag{7.3.19}$$

and is positive. Thus T_e^{-1} exists and is differentiable. For the transformation (7.3.17), T_e^{-1} is given below.

$$\begin{aligned} r &= x(y_j - y_k) - y(x_j - x_k) + [x_k(y_k - y_j) + y_k(x_j - x_k)]\,/A, \\ s &= [x(y_k - y_i) - y(x_k - x_i) + [x_k(y_i - y_k) + y_k(x_k - x_i)]\,/A\ , \\ A &= (x_i - x_k)(y_j - y_k) - (y_i - y_k)(x_j - x_k). \end{aligned} \tag{7.3.20}$$

Note that T_e^{-1} may exist even when $J = 0$. For example, consider

$$y = x^{3/2},\ x \in [0, 1]. \tag{7.3.21}$$

Then

$$J = \frac{dy}{dx} = \frac{3}{2} x^{1/2}, \tag{7.3.22}$$

and

$$J(0) = 0. \tag{7.3.23}$$

Since

$$x = y^{2/3} \tag{7.3.24}$$

is defined on $[0,1]$. thus T_e^{-1} exists even though $J(0)=0$. However,

$$\frac{dx}{dy}=\frac{2}{3}y^{-1/3} \tag{7.3.25}$$

is not defined at $y=0$. Thus the inverse map (7.3.24) is not differentiable for all $y\in[0,1]$.

7.4 Numerical Integration

In order to evaluate the element stiffness matrix, one needs to determine the derivative of a function, say ψ, with respect to x and y. Also, the area element $dxdy$ needs to be expressed in terms of dr and ds. A theorem in calculus gives

$$dxdy = Jdrds. \tag{7.4.1}$$

Also,

$$\frac{\partial\psi}{\partial x}=\frac{\partial\psi}{\partial r}\frac{\partial r}{\partial x}+\frac{\partial\psi}{\partial s}\frac{\partial s}{\partial x}, \tag{7.4.2a}$$

$$\frac{\partial\psi}{\partial y}=\frac{\partial\psi}{\partial r}\frac{\partial r}{\partial y}+\frac{\partial\psi}{\partial s}\frac{\partial s}{\partial y}. \tag{7.4.2b}$$

To evaluate $\frac{\partial r}{\partial x}$, $\frac{\partial r}{\partial y}$ etc., we note that

$$dr=\frac{\partial r}{\partial x}dx+\frac{\partial r}{\partial y}dy, \tag{7.4.3a}$$

$$ds=\frac{\partial s}{\partial x}dx+\frac{\partial s}{\partial y}dy, \tag{7.4.3b}$$

and

$$\left\{\begin{array}{c} dx \\ dy \end{array}\right\}=\left[\begin{array}{cc} \dfrac{\partial x}{\partial r} & \dfrac{\partial x}{\partial s} \\ \dfrac{\partial y}{\partial r} & \dfrac{\partial y}{\partial s} \end{array}\right]\left\{\begin{array}{c} dr \\ ds \end{array}\right\}, \tag{7.4.4}$$

or

$$\left\{\begin{array}{c} dr \\ ds \end{array}\right\}=\left[\begin{array}{cc} \dfrac{\partial x}{\partial r} & \dfrac{\partial x}{\partial s} \\ \dfrac{\partial y}{\partial r} & \dfrac{\partial y}{\partial s} \end{array}\right]^{-1}\left\{\begin{array}{c} dx \\ dy \end{array}\right\}=\frac{1}{J}\left[\begin{array}{cc} \dfrac{\partial y}{\partial s} & -\dfrac{\partial x}{\partial s} \\ -\dfrac{\partial y}{\partial r} & \dfrac{\partial x}{\partial r} \end{array}\right]\left\{\begin{array}{c} dx \\ dy \end{array}\right\}. \tag{7.4.5}$$

A comparison of eqns. (7.4.5) and (7.4.3) yields

$$\frac{\partial r}{\partial x}=\frac{1}{J}\frac{\partial y}{\partial s},\ \frac{\partial r}{\partial y}=-\frac{1}{J}\frac{\partial x}{\partial s},\ \frac{\partial s}{\partial x}=-\frac{1}{J}\frac{\partial y}{\partial r},\ \frac{\partial s}{\partial y}=\frac{1}{J}\frac{\partial x}{\partial r}. \tag{7.4.6}$$

Substitution from (7.4.6) into (7.4.2) results in

$$\frac{\partial \psi}{\partial x} = \frac{1}{J}\left[\frac{\partial \psi}{\partial r}\frac{\partial y}{\partial s} - \frac{\partial \psi}{\partial s}\frac{\partial y}{\partial r}\right], \tag{7.4.7a}$$

$$\frac{\partial \psi}{\partial y} = \frac{1}{J}\left[-\frac{\partial \psi}{\partial r}\frac{\partial x}{\partial s} + \frac{\partial \psi}{\partial s}\frac{\partial x}{\partial r}\right]. \tag{7.4.7b}$$

Since $\frac{\partial x}{\partial r}$, $\frac{\partial x}{\partial s}$ etc. can be evaluated directly from (7.3.17) there is no need to find T_e^{-1}.

For the problem discussed in Section 7.1,

$$\overline{K}_{ab}^{e} = \int_{\Omega_e} f_{ab}(x,y)\, dxdy, \tag{7.4.8}$$

where

$$f_{ij} = k\left(\frac{\partial N_a}{\partial x}\frac{\partial N_b}{\partial x} + \frac{\partial N_a}{\partial y}\frac{\partial N_b}{\partial y}\right) + \alpha N_a N_b. \tag{7.4.9}$$

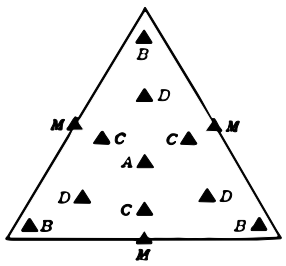
Thus

$$\begin{aligned}\overline{K}_{ab}^{e} &= \int_{\Omega_m} f_{ab}\left(x(r,s), y(r,s)\right) J\, drds\ , \\ &= \int_0^1 dr \int_0^{1-r} g_{ab}(r,s)\, ds, \\ &= \sum_{k=1}^{n_{int}} W_k g_{ab}(r_k, s_k)\ ,\end{aligned} \tag{7.4.10}$$

where $g_{ab} \equiv J f_{ab}$, (r_k, s_k) are coordinates of the kth quadrature point, and W_k is the corresponding weight. The coordinates of quadrature points and the corresponding weights are listed below in Table 7.1.

Numerical Integration for Triangles. Let ϕ be a function of area coordinates r, s and t. The quadrature rule is

$$\int_A \phi\, dA = \frac{1}{2}\sum_{i=1}^{n} W_i J_i \phi_i \tag{7.4.11}$$



in which ϕ_i is the value of ϕ at an integration point in the triangle, W_i is the weight appropriate to this point, n is the number of sampling points used, $dA = J\, d\xi\, d\eta$. The factor of $\frac{1}{2}$ appears in eqn. (7.4.11) because the area of the reference triangle in area coordinates is $\frac{1}{2}$. For an undistorted triangle of unit area in Cartesian coordinates, $J = 2$ throughout. Hence, since $\Sigma W_i = 1$ (see data in Table 7.1), we obtain $\int dA = 1$ when $\phi = 1$ in eqn. (7.4.11).

TABLE 7.1: GAUSS QUADRATURE FORMULAS FOR INTEGRATION OVER A TRIANGLE ACCORDING TO EQUATION (7.4.5). APPROXIMATE LOCATIONS OF ENTRIES IN THE "POINTS" COLUMN ARE SHOWN IN FIG. 14.4, ns = POINTS OF MULTIPLICITY 6, NOT SHOWN IN FIG. 14.4.

Points	Multiplicity	Area Coordinates $r,\ s,\ t$			Weights W_i
		1-point formula	degree of precision 1		
A	1	0.33333 33333 33333	0.33333 33333 33333	0.33333 33333 33333	1, 00000 00000 00000
		3-point formula	degree of precision 2		
B	3	0.66666 66666 66667	0.16666 66666 66667	0.16666 66666 66667	0.33333 33333 33333
		3-point formula	degree of precision 2		
M	3	0.50000 00000 00000	0.50000 00000 00000	0.00000 00000 00000	0.33333 33333 33333
		4-point formula	degree of precision 3		
A	1	0.33333 33333 33333	0.33333 33333 33333	0.33333 33333 33333	−0.56250 00000 00000
B	3	0.60000 00000 00000	0.20000 00000 00000	0.20000 00000 00000	0.52083 33333 33333
		6-point formula	degree of precision 4		
B	3	0.81684 75729 80459	0.09157 62135 09771	0.09157 62135 09771	0.10995 17436 55322
C	3	0.10810 30181 68070	0.44594 84909 15965	0.44594 84909 15965	0.22338 15896 78011
		7-point formula	degree of precision 5		
A	1	0.33333 33333 33333	0.33333 33333 33333	0.22500 00000 00000	
B	3	0.79742 69853 53087	0.10128 65073 23456	0.10128 65073 23456	0.12593 91805 44827
C	3	0.47014 20641 05115	0.47014 20641 05115	0.05971 58717 89770	0.13239 41527 88506
		12-point formula	degree of precision 6		
B	3	0.87382 19710 16996	0.06308 90144 91502	0.06308 90144 91502	0.5084 49063 70207
D	3	0.50142 65096 58179	0.24928 67451 70910	0.24928 67451 70910	0.11678 62757 26379
ns	6	0.63650 24991 21399	0.31035 24510 33784	0.05314 50498 44817	0.08285 10756 18374
		13-point formula	degree of precision 7		
A	1	0.33333 33333 33333	0.33333 33333 33333	0.33333 33333 33333	−0.14957 00444 67682
D	3	0.47930 80678 41920	0.26034 59660 79040	0.26034 59660 79040	0.17561 52574 33208
B	3	0.86973 97941 95568	0.06513 01029 02216	0.06513 01029 02216	0.05334 72356 08838
ns	6	0.63844 41885 69810	0.31286 54960 04874	0.04869 03154 25316	0.07711 37608 90257

7.5 Higher Order Triangular Element

The shape functions for a three-noded triangular element discussed in the previous section are complete polynomials of degree 1. These result in basis functions that are complete polynomials of degree one. In order to generate basis functions which are complete polynomials of higher order, one should look at the Pascal triangle given below.

$$
\begin{array}{ccccc}
 & & 1 & & \\
 & & x \quad y & & \\
 & x^2 & xy & y^2 & \\
x^3 & x^2y & xy^2 & y^3 & \\
 & & \ldots & &
\end{array}
$$

Fig. 7.5.1. Pascal's triangle

Thus to generate basis functions which are complete polynomials of degree 2, we should have shape functions which contain terms $1,\ x,\ y,\ x^2,\ xy$ and y^2. This necessitates 6 nodes on a triangular element. The Pascal triangle not only gives terms to be included in a complete polynomial of a given degree, but it also gives the location for the nodes. For shape functions which are complete

polynomials of degree one, the three nodes should be located at the three vertices of a triangle; for shape functions of degree two, the nodes should be located at the vertices and the midpoints of each side. For shape functions which are complete polynomials of degree 3, the nodes ought to be at the vertices, two on each side and one at the centroid of the triangular element. The two nodes on a side usually divide it into three equal parts. For brevity, the above three elements are usually referred to as linear, quadratic and cubic triangular elements, respectively.

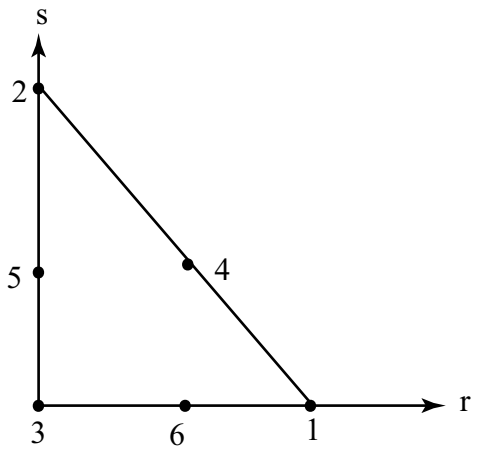


Fig. 7.5.2. A six-noded triangular element

To develop shape functions for a quadratic element, we note that the shape function for node 1 should vanish at nodes 2 through 6 as well as on side 23. Also it should be a polynomial of the lowest order. Equations of lines 253 and 46 are $r = 0$ and $r - \frac{1}{2} = 0$, respectively. Thus the product $(r)(r - 1/2)$ vanishes at nodes 2 through 6, on side 23, and is a polynomial of degree 2. However, it equals $(1)(1 - \frac{1}{2}) = 1/2$ at node 1. Thus

$$N_1 = \frac{r(r - 1/2)}{1/2} = r(2r - 1). \tag{7.5.1}$$

Similarly,

$$N_2 = s(2s - 1), \tag{7.5.2}$$

$$N_3 = t(2t - 1). \tag{7.5.3}$$

For the midside node 4, the shape function should vanish at nodes 1, 2, 3, 5 and 6 and on sides 23 and 31. Equations of lines 23 and 31 are $r = 0$ and $s = 0$; thus the product rs vanishes at nodes 1, 2, 3, 5 and 6 and on sides 23 and 31. Since rs equals $\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{4}$ at node 4, therefore

$$N_4 = \frac{rs}{1/4} = 4rs\ . \tag{7.5.4}$$

Similarly,

$$N_5 = 4st \ , \tag{7.5.5}$$

$$N_6 = 4tr \ . \tag{7.5.6}$$

Because of the constraint (7.3.15), r, s and t are not linearly independent. Note that shape functions N_1 through N_6 contain at most quadratic terms in r, s and t; collectively they are complete polynomials of degree 2 in r, s and t.

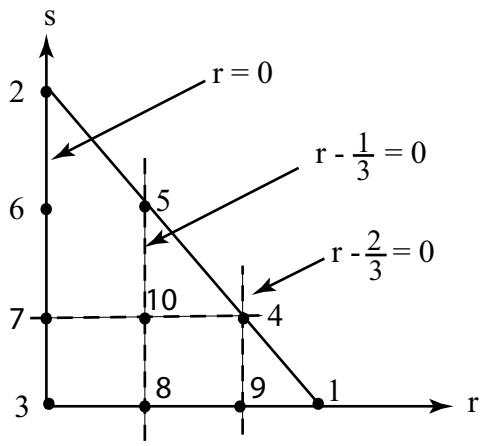


Fig. 7.5.3 A ten-noded triangular element

By following the procedure outlined above for generating shape functions for the six-noded triangular element, we can develop shape functions for the ten-noded triangular element shown in Figure 7.5.3. For example, the shape function for node 1 should vanish on nodes 2 through 10 and on side 23 of the triangle, and should be a polynomial of the lowest order possible. Equations of lines 23, 58 and 49 are $r = 0$, $r = 1/3$ and $r = 2/3$, respectively. Thus the product $r(r - 1/3)(r - 2/3)$ vanishes on nodes 2 through 9 and on side 23, and is a good candidate for the shape function for node 1. We normalize it so that it equals one at node 1 thereby arriving at the following.

$$N_1 = \frac{r\left(r - \frac{1}{3}\right)\left(r - \frac{2}{3}\right)}{1\left(1 - \frac{1}{3}\right)\left(1 - \frac{2}{3}\right)} = r\,(3r - 1)\,(3r - 2)\,/2 \ . \tag{7.5.7}$$

Similarly

$$N_2 = s(3s - 1)(3s - 2)/2, \tag{7.5.8}$$

$$N_3 = t(3t - 1)(3t - 2)/2 \ . \tag{7.5.9}$$

The shape function for node 4 should vanish on sides 23 and 31 and also at the other nine nodes. Equations of lines 23, 31 and 58 are $r = 0$, $s = 0$, $r - \frac{1}{3} = 0$, respectively. Thus

$$N_4 = \frac{rs\left(r-\frac{1}{3}\right)}{\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)\left(\frac{2}{3}-\frac{1}{3}\right)} = \frac{9}{2}rs(3r-1)\ . \tag{7.5.10}$$

Similarly

$$N_6 = \frac{9}{2}st(3s-1)\ , \tag{7.5.11}$$

$$N_8 = \frac{9}{2}tr(3t-1)\ , \tag{7.5.12}$$

$$N_5 = \frac{rs\left(s-\frac{1}{3}\right)}{\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\left(\frac{2}{3}-\frac{1}{3}\right)} = \frac{9}{2}rs(3s-1)\ , \tag{7.5.13}$$

$$N_7 = \frac{9}{2}st(3t-1)\ , \tag{7.5.14}$$

$$N_9 = \frac{9}{2}tr(3r-1)\ . \tag{7.5.15}$$

The shape function for node 10 should vanish on nodes 1 through 9 and also on all sides of the triangular element. The vanishing of N_{10} on all sides of the triangular element automatically satisfies the requirement that it vanish on the remaining nodes. Thus

$$N_{10} = \frac{rst}{\frac{1}{3}\frac{1}{3}\frac{1}{3}} = 27rst\ . \tag{7.5.16}$$

For a cubic triangular element, the fifteen nodes are located as shown in Fig. 7.5.4, and the shape functions are given below.

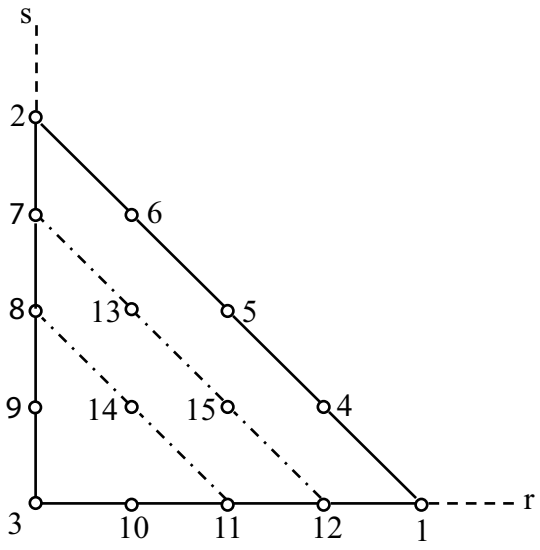


Fig. 7.5.4 A cubic triangular element

$$
\begin{aligned}
N_1 &= \frac{r\left(r-\frac{1}{4}\right)\left(r-\frac{2}{4}\right)\left(r-\frac{3}{4}\right)}{1\left(1-\frac{1}{4}\right)\left(1-\frac{1}{2}\right)\left(1-\frac{3}{4}\right)} = \frac{1}{3}r(4r-1)(2r-1)(4r-3),\\
N_2 &= \frac{1}{3}s(4s-1)(2s-1)(4s-3),\\
N_3 &= \frac{1}{3}t(4t-1)(2t-1)(4t-3),\\
N_4 &= \frac{rs\left(r-\frac{1}{4}\right)\left(r-\frac{1}{2}\right)}{\frac{3}{4}\cdot\frac{1}{4}\left(\frac{3}{4}-\frac{1}{4}\right)\left(\frac{3}{4}-\frac{1}{2}\right)} = \frac{16}{3}rs(4r-1)(2r-1),\\
N_5 &= \frac{rs\left(r-\frac{1}{4}\right)\left(s-\frac{1}{4}\right)}{\frac{1}{2}\cdot\frac{1}{2}\left(\frac{1}{2}-\frac{1}{4}\right)\left(\frac{1}{2}-\frac{1}{4}\right)} = 4rs(4r-1)(4s-1),\\
N_6 &= \frac{rs\left(s-\frac{1}{2}\right)\left(s-\frac{1}{4}\right)}{\frac{1}{4}\cdot\frac{3}{4}\left(\frac{3}{4}-\frac{1}{2}\right)\left(\frac{3}{4}-\frac{1}{4}\right)} = \frac{16}{3}rs(2s-1)(4s-1),\\
N_7 &= \frac{16}{3}st(4s-1)(2s-1),\\
N_8 &= 4st(4s-1)(4t-1),\\
N_9 &= \frac{16}{3}st(4t-1)(2t-1),\\
N_{10} &= \frac{16}{3}tr(4t-1)(2t-1),\\
N_{11} &= 4tr(4t-1)(4r-1),\\
N_{12} &= \frac{16}{3}tr(2r-1)(4r-1),\\
N_{13} &= \frac{rst\left(s-\frac{1}{4}\right)}{\frac{1}{4}\cdot\frac{2}{4}\cdot\frac{1}{4}\left(\frac{2}{4}-\frac{1}{4}\right)} = 32rst(4s-1),\\
N_{14} &= \frac{rst\left(t-\frac{1}{4}\right)}{\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{2}{4}\left(\frac{2}{4}-\frac{1}{4}\right)} = 32rst(4t-1),\\
N_{15} &= 32rst(4r-1)\ .
\end{aligned}
\tag{7.5.17}
$$

7.6 Isoparametric, Subparametric and Superparametric Maps

The shape functions we have generated in Section 7.5 are defined on a master triangular element Ω_M. However, the weak formulation of a given BV problem involves derivatives of the trial solution with respect to x and y and integrations on the physical domain or actual elements Ω_e. The map T_e defined by eqn. (7.3.17) links points in Ω_e to those in Ω_M and vice-versa. For higher order (quadratic, cubic etc.) elements, one may still use 3-noded master triangular element to generate the map T_e from Ω_M to Ω_e but use a quadratic or cubic approximation of the trial solution u^e on

the element Ω_e. That is,

$$u^e\left(x\left(r,s\right),y\left(r,s\right)\right)=\sum_{i=1}^{m}u_iN_i(r,s), \tag{7.6.1}$$

$$T_e: \quad \begin{aligned} x &= \sum_{i=1}^{n} x_i\hat{N}_i(r,s), \\ y &= \sum_{i=1}^{n} y_i\hat{N}_i(r,s) \end{aligned} \quad . \tag{7.6.2}$$

The map T_e is called isoparametric, subparametric and superparametric according as $n = m$, $n < m$ and $n > m$ respectively. We recall that for the beam element, $m = 4$, and the corresponding shape functions are Hermitian, but the map T_e is defined by

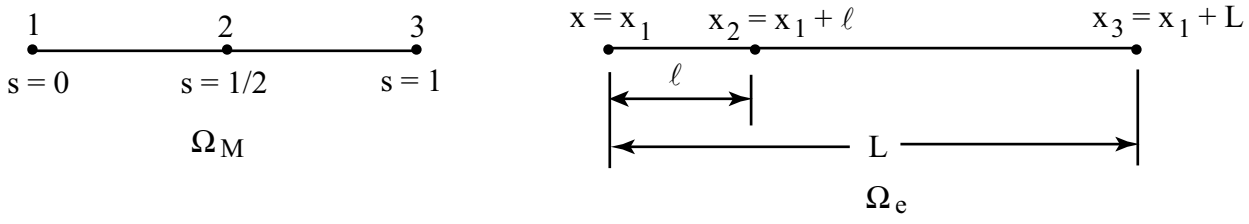
$$T_e = x_1N_1(s) + x_2N_2(s)\ .$$

Thus the map T_e is subparametric. In two-dimensional problems one will use a superparametric map if Ω_e has curved sides but the trial solution u^e is expected to vary slowly over the element Ω_e. Note that if $m \neq n$, different shape functions are used in eqns. (7.6.1) and (7.6.2). In order to avoid using two different sets of shape functions, one for u^e and the other for T_e, the isoparametric map is most often used.

Whereas the location of nodes in Ω_M not coinciding with the vertices of a triangle is generally taken to be that given by the Pascal triangle, their location in the actual element Ω_e should be such as to satisfy $J > 0$ where J is the Jacobian of the map T_e (cf. eqn. (7.3.18)). This restricts the location of nodes on the sides of Ω_e.

7.7 Restrictions on the Location of Nodes

Let us consider the map T_e from a quadratic Lagrangian element onto a 3-noded element with the interior node located at a distance ℓ from the left node. Here we have intentionally taken



the origin at node 1 in Ω_M so that $s \geq 0$ everywhere in Ω_M. Lagrange shape functions for Ω_M are

$$N_1 = 2\left(s - \frac{1}{2}\right)(s-1), \tag{7.7.1}$$

$$N_2 = -4s(s-1), \tag{7.7.2}$$

$$N_3 = 2s\left(s - \frac{1}{2}\right), \tag{7.7.3}$$

and assuming that T_e is isoparametric,

$$\begin{aligned} T_e: \quad x =& x_1 N_1 + x_2 N_2 + x_3 N_3, \\ =& x_1(2s-1)(s-1) + 4x_2 s(1-s) + x_3 s(2s-1). \end{aligned} \tag{7.7.4}$$

Therefore,

$$J = \frac{dx}{ds} = 4s(L - 2\ell) + (4\ell - L). \tag{7.7.5}$$

Without loss in generality we may assume that $\ell \leq L/2$. Since $s \geq 0$, the minimum value of J occurs at $s = 0$ and in order for $J > 0$, $4\ell - L > 0$ or $\ell > L/4$. Thus the distance of the interior node in Ω_e from the leftmost node must be greater than one-fourth of the length of the element. If $\ell = L/4$ then $J = 0$ at $s = 0$ and the map T_e will be singular at the leftmost node.

Similarly, one can derive restrictions on nodes located on the sides of a triangular element.

7.8 Quadrilateral Elements

The simplest quadrilateral element has four nodes located at its vertices. One way to derive the corresponding shape functions is to assume that

$$N_i(x, y) = \alpha_i + \beta_i x + \gamma_i y + \delta_i xy, \ i = 1, 2, 3, 4, \tag{7.8.1}$$

and then use

$$N_i(x_j, y_j) = \delta_{ij}, \ i, j = 1, 2, 3, 4, \tag{7.8.2}$$

where δ_{ij} is the Kronecker delta. Thus for each i, we have four equations for the determination of $\alpha_i, \beta_i, \gamma_i$ and δ_i.

An alternative is to work on the master element which is taken to be a square of side 2. Figure 7.8.1 depicts a choice of the coordinate system and the equations of bounding lines. The shape

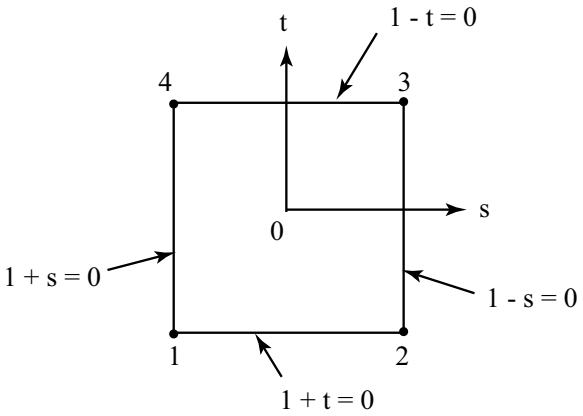


Fig. 7.8.1 Master quadrilateral element

function N_1 for node 1 should vanish on nodes 2, 3 and 4 and on sides 23 and 34 not passing through node 1. The product $(1-s)(1-t)$ satisfies these conditions; however, we need to normalize it so that it equals one at node 1. Therefore,

$$N_1(s,t) = \frac{(1-s)(1-t)}{(1-(-1))(1-(-1))} = \frac{1}{4}(1-s)(1-t)\ . \tag{7.8.3}$$

Similarly

$$N_2(s,t) = \frac{1}{4}(1+s)(1-t), \tag{7.8.4}$$

$$N_3(s,t) = \frac{1}{4}(1+s)(1+t), \tag{7.8.5}$$

$$N_4(s,t) = \frac{1}{4}(1+t)(1-s)\ . \tag{7.8.6}$$

Expressions (7.8.3) - (7.8.6) can be written as

$$N_i(s,t) = \frac{1}{4}(1+s_i s)(1+t_i t),\ i=1,2,3,4, \tag{7.8.7}$$

where (s_i, t_i) are coordinates of the ith node. The function $N_3(s,t)$ is plotted in Fig. 7.8.2.

Fig. 7.8.2 Shape function $N_3(s,t)$

The four-noded square master element can also be obtained by taking the tensor product of two linear elements, one in the s-direction and one in the t-direction as shown below.

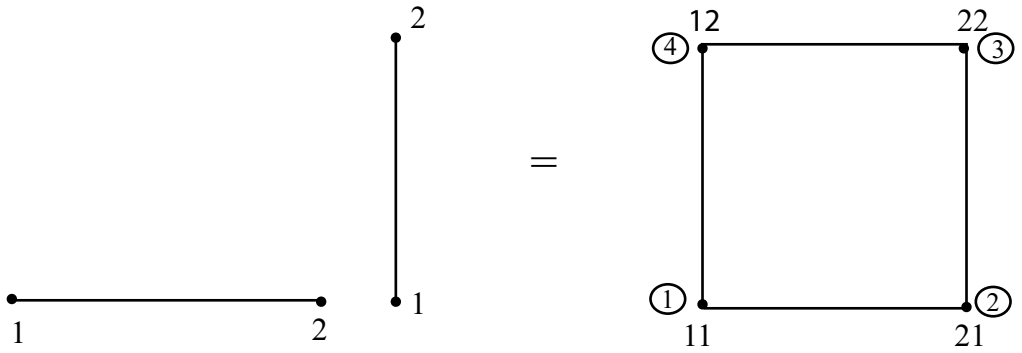


Fig. 7.8.3 A 4-noded quadrilateral element obtained as the tensor product of two linear elements

The first index on the square element in Fig. 7.8.3 corresponds to the node on the horizontal one-dimensional element and the second index to the node on the vertical one-dimensional element. Comparing it with the master element depicted in Fig. 7.8.1 we get

$$N_1 = N_1^v N_1^h,\ N_2 = N_1^v N_2^h,\ N_3 = N_2^v N_2^h, N_4 = N_2^v N_1^h\ . \tag{7.8.8}$$

Since

$$N_1^v = \frac{1}{2}(1-t),\ N_2^v = \frac{1}{2}(1+t),\ N_1^h = \frac{1}{2}(1-s),\ N_2^h = \frac{1}{2}(1+s), \tag{7.8.9}$$

therefore, we get relations (7.8.3) - (7.8.6) for the shape functions $N_i(s,t),\ i = 1,2,3,4$. This procedure can be easily extended to derive shape functions for higher-order elements. For example, the tensor product of two one-dimensional quadratic elements will result in a nine-noded quadrilateral element as shown below.

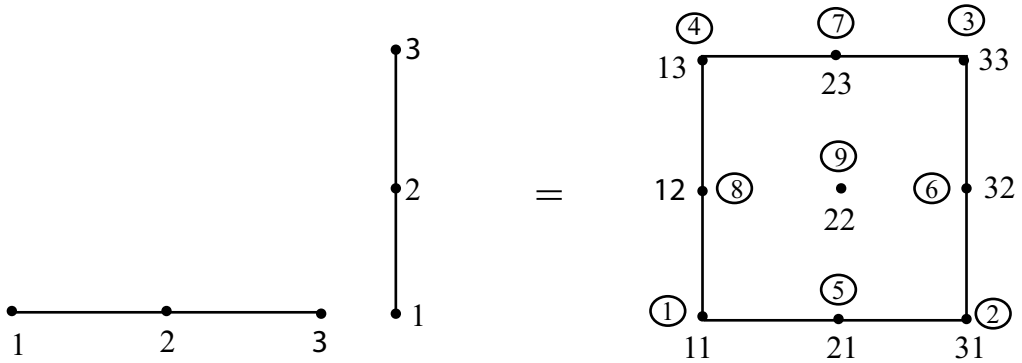


Fig. 7.8.4 A 9-noded quadrilateral element obtained by taking the tensor product of two one-dimensional quadratic elements

In order to derive the shape functions we note that

$$N_1 = N_1^v N_1^h,\ N_5 = N_1^v N_2^h,\ N_2 = N_1^v N_3^h \text{ etc.} \tag{7.8.10}$$

Since $N_1^v = \frac{1}{2}t(t-1),\ N_2^v = (1-t^2),\ N_3^v = \frac{1}{2}t(1+t)$, and $N_1^h = \frac{1}{2}s(s-1),\ N_2^h = (1-s^2),\ N_3^h = \frac{1}{2}s(1+s)$, we arrive at the following expressions for the shape functions.

$$\begin{aligned}
N_1 &= \frac{1}{4}st(t-1)(s-1),\\
N_2 &= \frac{1}{4}st(t-1)(1+s),\\
N_3 &= \frac{1}{4}st(t+1)(s+1),\\
N_4 &= \frac{1}{4}st(t+1)(s-1),\\
N_5 &= \frac{1}{2}t(t-1)(1-s^2)\\
N_6 &= \frac{1}{2}s(1-t^2)(1+s),\\
N_7 &= \frac{1}{2}t(1+t)(1-s^2),\\
N_8 &= \frac{1}{2}s(1-t^2)(s-1),\\
N_9 &= (1-s^2)(1-t^2).
\end{aligned} \tag{7.8.11}$$

Even though N_1 through N_9 contain cubic and quartic terms in s and t, they are complete polynomials of degree 2 only. It becomes transparent by looking at the Pascal triangle shown below

$$\begin{array}{ccccc}
 & & 1 & & \\
 & s & & t & \\
s^2 & & st & & t^2\\
 s^3 & & s^2t & & st^2 \quad t^3\\
s^4 & s^3t & & s^2t^2 & st^3 \quad t^4
\end{array}$$

Fig. 7.8.5 Terms of Pascal triangle included in shape functions for 4- and 9-noded quadrilateral elements.

Six-noded quadrilateral elements can be obtained by taking the tensor product of a quadratic one-dimensional element and a linear one-dimensional element; the corresponding shape functions can be derived by following the aforestated procedure.

We note that for the 9-noded quadrilateral element, node 9 is at the centroid of the element and, in a FE mesh, will not be connected to any other element. Since the shape functions have

three terms in addition to those needed for a complete polynomial of degree two, node 9 can easily be deleted thus obtaining 8-noded serendipity element. Shape functions for the serendipity element can not be obtained from those for the one-dimensional element since it is not derivable by taking the tensor product of one-dimensional elements. Therefore, we follow the procedure used

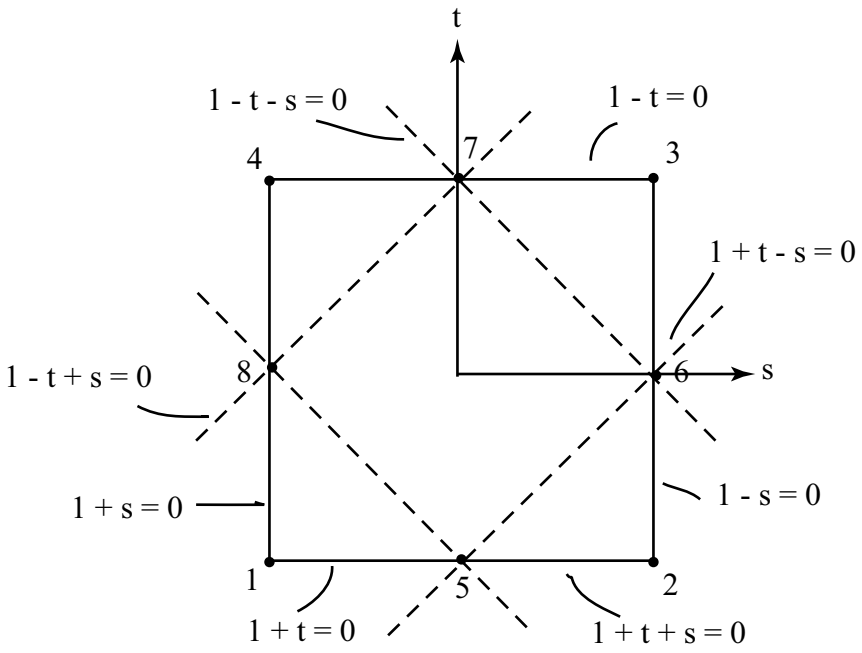


Fig. 7.8.6 8-noded serendipity element and equations of different lines

to obtain shape functions for a higher order triangular element. Since the shape function for node 1 should vanish on the remaining nodes as well as on sides 23 and 34 not passing through node 1, we take it to be the product of polynomials of degree 1 that correspond to equations of lines 23, 34, and 58 and then normalize it so that it equals one at node 1. Thus

$$
\begin{aligned}
N_1(s,t) &= \frac{(1-s)(1-t)(1+s+t)}{(1-(-1))(1-(-1))(1-1-1)} = -\frac{1}{4}(1-s)(1-t)(1+s+t),\\
N_2(s,t) &= -\frac{1}{4}(1+s)(1-t)(1+t-s),\\
N_3(s,t) &= -\frac{1}{4}(1+s)(1+t)(1-t-s),\\
N_4(s,t) &= -\frac{1}{4}(1-s)(1+t)(1-t+s),\\
N_5(s,t) &= \frac{1}{2}(1-s^2)(1-t),
\end{aligned}
\tag{7.8.12}
$$

$$N_6(s,t) = \frac{1}{2}(1+s)(1-t^2),$$
$$N_7(s,t) = \frac{1}{2}(1-s^2)(1+t),$$
$$N_8(s,t) = \frac{1}{2}(1-s)(1-t^2).$$

Even though shape functions given in (7.8.11) for nodes 1 through 8 meet all of the requirements for the shape functions for a 8-noded serendipity element, they should not be used since all of them vanish at the centroid of the master element. Thus a trial solution obtained by using them will vanish at an interior point of an element into which the centroid is mapped.

In many practical problems one needs to use an element with 5, 6 or 7 nodes. We now develop shape functions for such an element. The first step is to write shape functions for the corner nodes since these are always present.

$$N_i(s,t) = \frac{1}{4}(1+s_i s)(1+t_i t),\ i = 1,2,3,4\ . \tag{7.8.13}$$

Here (s_i, t_i) are coordinates of node i. If node 5 is present at the mid-point of side 12, then

$$N_5(s,t) = \frac{1}{2}(1-s^2)(1-t), \tag{7.8.14}$$

and N_1 and N_2 need to be modified since they now should vanish at node 5; N_3 and N_4 satisfy this requirement. We note that N_1 and N_2 equal $1/2$ at node 5 and since N_5 equals 1 there, therefore, $N_1 - \frac{1}{2}N_5$ and $N_2 - \frac{1}{2}N_5$ will vanish at node 5. Thus, if node 5 is present, then

$$\begin{aligned} N_1 &\leftarrow N_1 - \frac{1}{2}N_5 \\ N_2 &\leftarrow N_2 - \frac{1}{2}N_5 \end{aligned} \tag{7.8.15}$$

where $\leftarrow$ signifies that the quantity on the left is replaced by the one on the right. Similarly, if node 6 is present, then

$$\begin{aligned} N_6 &= \frac{1}{2}(1+s)(1-t^2), \\ N_2 &\leftarrow N_2 - \frac{1}{2}N_6, \\ N_3 &\leftarrow N_3 - \frac{1}{2}N_6\ . \end{aligned} \tag{7.8.16}$$

Note that if both nodes 5 and 6 are present, then N_2 will need to be modified twice, once because of the presence of node 5 and second time because of the presence of node 6. Similarly, the

presence of nodes 7 and 8 will necessitate the modifications of $N_3,\ N_4$ and $N_4,\ N_1$ respectively. The presence of node 9 will require the modification of N_1 through N_8. In order to minimize the necessary modifications, it is better to check for the presence of node 9 first and modify N_1 through N_4 if necessary. Of course, expressions (7.8.14) and $(7.8.16)_1$ for N_5, and N_6 will need to be changed if node 9 is present.

To see whether or not this method yields expressions (7.8.12) for the shape functions, we assume that nodes 5 and 6 are present. Thus

$$\begin{aligned} N_2 =& \frac{1}{4}(1+s)(1-t) - \frac{1}{4}(1-s^2)(1-t) - \frac{1}{4}(1+s)(1-t^2), \\ =& \frac{1}{4}(1+s)(1+t)(-1+s-t), \end{aligned} \tag{7.8.17}$$

which agrees with $(7.8.12)_2$. However, in general, the shape functions for higher-order (quadratic and higher) elements are not uniquely determined.

7.9 Numerical Integration on Quadrilateral Elements

In Section 6.3 we discussed Gauss Quadrature Rule to evaluate

$$I = \int_{-1}^{1} f(s)ds \tag{7.9.1}$$

where $f(s)$ is a polynomial of degree n in s. For two dimensional problems involving quadrilateral elements, we will need to evaluate

$$I = \int_{-1}^{1} dt \int_{-1}^{1} f(s,t)ds\ . \tag{7.9.2}$$

If $f(s,t)$ is a polynomial function of degree $(2m+1)$ in s and $(2n+1)$ in t, then we can use $(m+1)$ Gauss quadrature points in the s-direction and $(n+1)$ Gauss quadrature points in the t-direction to evaluate exactly I in (7.9.2). That is

$$I = \sum_{j=1}^{n+1} W_j^t \sum_{i=1}^{m+1} f(s_i,t_j) W_i^s \tag{7.9..3}$$

where W_i^s and W_j^t are the corresponding weight functions in the s- and t-direction. Thus we need $(m+1)(n+1)$ quadrature points.

The double summations in (7.9.3) and the triple summations in (8.3.4) can be reduced to a single summation by renumbering the quadrature points. That is, for $(m+1)=2$ and $(n+1)=2$,

the four quadrature points can be consecutively numbered from 1 to 4 with the result

$$I = \sum_{k=1}^{4} W_k f(s_k, t_k) \tag{7.9.4}$$

where the weights W_k are the products of the corresponding weights in the s and t directions. The locations and corresponding weights for the 2×2 and 3×3 integration rule over a master quadrilateral element are depicted in Fig. 7.9.1.

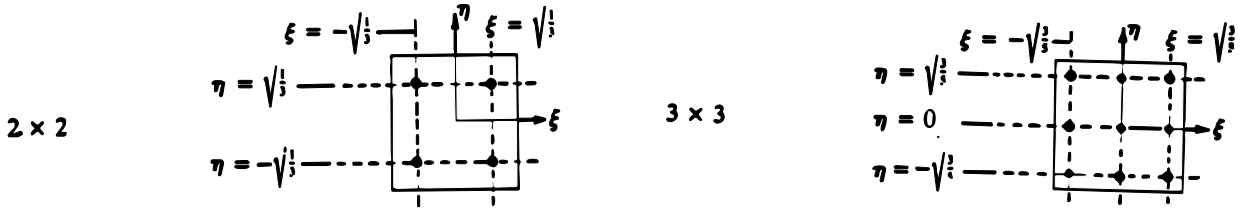


Fig. 7.9.1 Locations of quadrature points for the 2×2 and 3×3 integration rules.

When evaluating the element stiffness matrix, we note that the derivatives of shape functions with respect to x and y can be evaluated by using (7.4.7) and the area element can be expressed in terms of local coordinates by (7.4.1). After having substituted these in (7.3.14) we find that the integrand K_{ij}^e for the element stiffness matrix contains J in the denominator. For the simplest 4-noded bilinear quadrilateral element, the integrand for K_{ij}^e is a ratio of two polynomials which need not result in a polynomial. The quadrature rule that will integrate exactly a non-polynomial integrand has not been established. In practice, one experiments with quadrature rules of different orders until sufficiently accurate evaluation of the desired integrand has been attained.

Chapter 8: Three-Dimensional Problems

8.1 Two/Three Dimensional Problems in Linear Elasticity

We study infinitesimal deformations of a linear elastic homogeneous body that is stress free in the reference configuration. With respect to a set of rectangular Cartesian coordinates, let u_i denote the displacement of a material point x_i. Thus

$$x_i = u_i + X_i$$

where x_i gives the current position of the material point X_i. The displacements u_i are given by

$$\sigma_{ij,j} + f_i = 0 \text{ in } \Omega,\ i = 1,2,3 \tag{8.1.1}$$

$$\sigma_{ij} n_j = h_i \text{ on } \Gamma_1,\ i = 1,2,3 \tag{8.1.2}$$

$$u_i = \hat{u}_i \text{ on } \Gamma_2,\ i = 1,2,3 \tag{8.1.3}$$

$$\sigma_{ij} = D_{ijk\ell} e_{k\ell} \text{ in } \Omega, \tag{8.1.4}$$

$$e_{ij} = (u_{i,j} + u_{j,i})/2 = u_{(i,j)} \text{ in } \Omega, \tag{8.1.5}$$

where

$$\overline{\Gamma}_1 \cup \overline{\Gamma}_2 = \partial\Omega,\ \Gamma_1 \cap \Gamma_2 = \phi, \tag{8.1.6}$$

$D_{ijk\ell}$ is the elasticity tensor, e_{ij} is the infinitesimal strain tensor, n_j is a unit outward normal to the boundary $\partial\Omega$ of Ω, h_i is the prescribed surface traction on Γ_1, and $\hat{u}_i$ the prescribed surface displacements on Γ_2; Γ_1 and Γ_2 are complementary parts of the boundary $\partial\Omega$. The elasticity tensor $D_{ijk\ell}$ exhibits the following symmetries

$$D_{ijk\ell} = D_{jik\ell} = D_{ij\ell k} = D_{k\ell ij} \tag{8.1.7}$$

and is often assumed to satisfy

$$D_{ijk\ell} e_{ij} e_{k\ell} \geq 0 \tag{8.1.8}$$

for every symmetric tensor e_{ij}, and the equality holds only when $e_{ij} = 0$.

With (8.1.7) there are 21 independent components of $D_{ijk\ell}$. In the FE work, it is convenient to use the following compressed notations for σ_{ij}, e_{ij} and $D_{ijk\ell}$: $\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{12}, \sigma_{23}$ and σ_{31} are denoted by $\sigma_1, \sigma_2, \ldots, \sigma_6$; $e_{11}, e_{22}, e_{33}, 2e_{12}, 2e_{23}$ and $2e_{31}$ by $e_1, e_2, \ldots, e_6$ and the corresponding

elasticity matrix is a symmetric 6×6 matrix. The relation (8.1.4) is written in the matrix notation as

$$\{\sigma\} = [D]\{e\} \ . \tag{8.1.9}$$

Substitution from (8.1.4), (8.1.5) into (8.1.1) and recalling (8.1.7) give

$$(D_{ijk\ell}u_{k,\ell})_{,j} + f_i = 0 \text{ in } \Omega \ . \tag{8.1.10}$$

Thus the given boundary-value problem involves solving a set of three coupled second-order partial differential equations for u_1, u_2 and u_3 subjected to natural boundary conditions (8.1.2) and essential boundary conditions (8.1.3). We now derive a weak formulation of the problem.

Let, for $i = 1, 2, 3, \phi_i : \overline{\Omega} \to \mathbb{R}$ be smooth functions such that $\phi_i = 0$ on Γ_2. Taking the inner product of (8.1.1) with ϕ_i, integrating the result over Ω we arrive at

$$\int_\Omega \sigma_{ij,j}\phi_i d\Omega + \int_\Omega f_i\phi_i d\Omega = 0 \ . \tag{8.1.11}$$

Because of the divergence theorem,

$$\int_\Omega \sigma_{ij,j}\phi_i d\Omega = \int_{\partial\Omega} \sigma_{ij}\phi_i n_j d\Gamma - \int_\Omega \sigma_{ij}\phi_{i,j} d\Omega, \tag{8.1.12}$$

$$= \int_{\Gamma_1} h_i\phi_i d\Gamma - \int_\Omega \sigma_{ij}(\phi_{(i,j)} + \phi_{[i,j]}) d\Omega, \tag{8.1.13}$$

$$= \int_{\Gamma_1} h_i\phi_i d\Gamma - \int_\Omega \sigma_{ij}\phi_{(i,j)} d\Omega, \tag{8.1.14}$$

where

$$\phi_{[i,j]} = \frac{1}{2}(\phi_{i,j} - \phi_{j,i}), \ \phi_{(i,j)} = \frac{1}{2}(\phi_{i,j} + \phi_{j,i}) \ , \tag{8.1.16}$$

and we have used $\phi_i = 0$ on Γ_2, the natural boundary condition (8.1.2) and

$$\sigma_{ij}\phi_{[i,j]} \equiv 0 \ . \tag{8.1.16}$$

Substitution from (8.1.14), (8.1.4) and (8.1.5) into (8.1.11) yields

$$B(\mathbf{u}, \boldsymbol{\phi}) = \ell(\boldsymbol{\phi}), \tag{8.1.17}$$

where

$$B(\mathbf{u}, \boldsymbol{\phi}) = \int_\Omega u_{(i,j)} D_{ijk\ell}\phi_{(k,\ell)} d\Omega \ , \tag{8.1.18}$$

$$\ell(\boldsymbol{\phi}) = \int_\Omega f_i\phi_i d\Omega + \int_{\Gamma_1} h_i\phi_i d\Gamma, \tag{8.1.19}$$

are the bilinear form and the linear functional respectively. Because of (8.1.8), the bilinear form $B(\cdot,\cdot)$ is positive unless $\mathbf{u}$ or $\boldsymbol{\phi}$ correspond to a rigid body motion.

We note that in order for the bilinear form (8.1.18) to be well defined, first order derivatives of functions $\mathbf{u}$ and $\boldsymbol{\phi}$ should be square integrable over Ω. We define

$$H^1 = \left\{ \boldsymbol{\psi} | \boldsymbol{\psi} : \overline{\Omega} \to \mathbb{R}^3, \int_\Omega (\psi_i \psi_i + \psi_{i,j}\psi_{i,j})\, d\Omega < \infty \right\} \tag{8.1.20}$$

$$H_0^1 = \left\{ \boldsymbol{\psi} | \boldsymbol{\psi} \in H^1,\ \boldsymbol{\psi} = \mathbf{0} \text{ on } \Gamma_2 \right\} . \tag{8.1.21}$$

Thus a **weak formulation** of the problem defined by (8.1.1) - (8.1.6) is: find $\mathbf{u} \in H^1$ such that $u_i = \hat{u}_i$ on Γ_2 and (8.1.17) holds for every $\boldsymbol{\phi} \in H_0^1$.

Let $g_i : \overline{\Omega} \to \mathbb{R}$ be such that $g_i \in H^1$ and $g_i = \hat{u}_i$ on Γ_2. Then for any function $\mathbf{u} \in H^1$, we can find a function $\mathbf{v} \in H_0^1$ such that $\mathbf{u} = \mathbf{v} + \mathbf{g}$. This when substituted into (8.1.17) gives

$$B(\mathbf{v}, \boldsymbol{\phi}) = \ell(\boldsymbol{\phi}) - B(\mathbf{g},\ \boldsymbol{\phi}), \tag{8.1.22}$$

and the **Galerkin formulation** of the problem can be stated as follows: find $\mathbf{v} \in H_0^1$ such that (8.1.22) holds for every $\boldsymbol{\phi} \in H_0^1$. Let $H_0^{1n} \subset H_0^1$ be a finite dimensional subset of H_0^1. Then the **Galerkin approximation** of the problem (8.1.1)-(8.1.6) is: find $\mathbf{v}^n \in H_0^{1n}$ such that

$$B(\mathbf{v}^n, \boldsymbol{\phi}^n) = \ell(\boldsymbol{\phi}^n) - B(\mathbf{g}, \boldsymbol{\phi}^n) \tag{8.1.23}$$

holds for every $\boldsymbol{\phi}^n \in H_0^{1n}$. Let $\psi_1, \psi_2, \ldots \psi_n$ be basis functions in H_0^{1n}, then we can write

$$v_i^n(x_1, x_2, x_3) = \sum_{\alpha=1}^{n} \psi_\alpha(x_1, x_2, x_3) d_{\alpha i},\ i = 1, 2, 3 \tag{8.1.24a}$$

$$\phi_i^n(x_1, x_2, x_3) = \sum_{\alpha=1}^{n} \psi_\alpha(x_1, x_2, x_3) c_{\alpha i},\ i = 1, 2, 3\ . \tag{8.1.24b}$$

The first index in $d_{\alpha i}$ and $c_{\alpha i}$ refers to the basis function and the second index to the x_i direction. In the FEM, the basis function $\psi_\alpha(x_1, x_2, x_3)$ is generally associated with the node α. Thus there are three unknowns $d_{\alpha 1}, d_{\alpha 2}$ and $d_{\alpha 3}$ at each node α, and the total number of d's and c's equal $3n$. Dropping the summation sign, we note that

$$v_{i,j}^n = \psi_{\alpha,j} d_{\alpha i}\ , \tag{8.1.25}$$

$$e_{ij}^v \equiv v_{(i,j)}^n = (\psi_{\alpha,j} d_{\alpha i} + \psi_{\alpha,i} d_{\alpha j})/2\ , \tag{8.1.26a}$$

or

$$\{e^v\} = [B]\{d\} \ . \tag{8.1.26b}$$

Similarly

$$\{e^\phi\} = [B]\{c\} \ . \tag{8.1.26c}$$

We have denoted the strain-displacement matrix by B to stay consistent with the common notation. The context should make clear whether the matrix B or the bilinear form B is being used. The elements of the strain-displacement matrix B are partial derivatives of the basis functions. Recalling (8.1.9) we can write the bilinear form $B(\mathbf{v}^n, \boldsymbol{\phi}^n)$ as

$$B(\mathbf{v}^n, \boldsymbol{\phi}^n) = \int_\Omega \{e^\phi\}^T [D]\{e^v\} d\Omega, \tag{8.1.27a}$$

$$= \{c\}^T \left(\int_\Omega [B]^T [D][B] d\Omega \right) \{d\} \ , \tag{8.1.27b}$$

$$= \{c\}^T [K]\{d\}, \tag{8.1.27c}$$

where

$$[K] = \int_\Omega [B]^T [D][B] d\Omega \tag{8.1.28}$$

is called the stiffness matrix. Similarly,

$$B(\mathbf{g}, \boldsymbol{\phi}^n) = \{c\}^T [K]\{g\} \tag{8.1.29}$$

where $\{g\}$ is the matrix of the values of g_i at node points. Since the function g is known, these values can be determined. Recalling (8.1.19),

$$\ell(\boldsymbol{\phi}^n) = \int_\Omega \{\phi^n\}^T \{f\} d\Omega + \int_{\Gamma_1} \{\phi^n\}^T \{h\} d\Gamma, \tag{8.1.30a}$$

$$= \{c\}^T \left(\int_\Omega [\psi]^T \{f\} d\Omega + \int_{\Gamma_1} [\psi]^T \{h\} d\Gamma \right), \tag{8.1.30b}$$

$$= \{c\}^T \{\overline{F}\}, \tag{8.1.30c}$$

where

$$\{\overline{F}\} = \int_\Omega [\psi]^T \{f\} \, d\Omega + \int_{\Gamma_1} [\psi]^T \{h\} d\Gamma, \tag{8.1.31}$$

is the matrix of nodal forces that are equivalent to the given body force $\mathbf{f}$ in Ω and surface tractions $\mathbf{h}$ on Γ_1. Substitution from (8.1.27), (8.1.29) and (8.1.30) into (8.1.23) and recalling that (8.1.23) holds for every $\boldsymbol{\phi}^n$ and hence every choice of $\{c\}$, we obtain

$$Kd = F \tag{8.1.32}$$

where $F = \overline{F} - Kg$. Equations (8.1.32) do not include degrees of freedom at nodes where essential boundary conditions are prescribed. In the FE work it is common to include these and work with $Kd = \overline{F}$ and apply essential boundary conditions by one of the methods discussed in Section 4.4.

After having solved (8.1.32) for d's, one can evaluate strains and stresses at quadrature (Gauss) points and then at other desired points by extrapolating or interpolating these values. We now state explicitly the element strain-displacement matrix. On an element, equation $(8.1.24a)$ becomes

$$u_i^e(x_1, x_2, x_3) = \sum_{a=1}^{\#\text{ nodes}} N_a(x_1, x_2, x_3) d_{ai}, \ i = 1, 2, 3, \tag{8.1.33a}$$

or in matrix notation

$$\left\{ \begin{array}{c} u_1 \\ u_2 \\ u_3 \end{array} \right\} = \left[\begin{array}{ccccccccc} N_1 & 0 & 0 & N_2 & 0 & 0 & N_3 & 0 & \dots \\ 0 & N_1 & 0 & 0 & N_2 & 0 & 0 & N_3 & \dots \\ 0 & 0 & N_1 & 0 & 0 & N_2 & 0 & N_3 & \dots \end{array} \right] \left\{ \begin{array}{c} d_{11} \\ d_{12} \\ d_{13} \\ d_{21} \\ d_{22} \\ d_{23} \\ \vdots \end{array} \right\}. \tag{8.1.33b}$$

Usually, the shape functions are defined on the master element in terms of local coordinates. Their derivatives with respect to global coordinates x_1, x_2, x_3 can be computed by using relations like (7.4.7). Recalling $(8.1.26a)$, we get

$$\left\{ \begin{array}{c} e_{11}^u \\ e_{22}^u \\ e_{33}^u \\ 2e_{12}^u \\ 2e_{23}^u \\ 2e_{31}^u \end{array} \right\} = \left[\begin{array}{ccccccccccc} N_{1,1} & 0 & 0 & N_{2,1} & 0 & 0 & N_{3,1} & 0 & 0 & \dots \\ 0 & N_{1,2} & 0 & 0 & N_{2,2} & 0 & 0 & N_{3,2} & 0 & \dots \\ 0 & 0 & N_{1,3} & 0 & 0 & N_{2,3} & 0 & 0 & N_{3,3} & \dots \\ N_{1,2} & N_{1,1} & 0 & N_{2,2} & N_{2,1} & 0 & N_{3,2} & N_{3,1} & 0 & \dots \\ 0 & N_{1,3} & N_{1,2} & 0 & N_{2,3} & N_{2,2} & 0 & N_{3,3} & N_{3,2} & \dots \\ N_{1,3} & 0 & N_{1,1} & N_{2,3} & 0 & N_{2,1} & N_{3,3} & 0 & N_{3,1} & \dots \end{array} \right] \left\{ \begin{array}{c} d_{11} \\ d_{12} \\ d_{13} \\ d_{21} \\ d_{22} \\ d_{23} \\ \vdots \end{array} \right\} \tag{8.1.34}$$

We note that the formulation given in this section is valid for anisotropic materials; the material anisotropy is exhibited in the stress-strain relation (8.1.9). For isotropic materials the elements of the stress-strain matrix D will be made up of only two independent material parameters say Young's modulus and Poisson's ratio.

By changing the range of indices i and j from 1, 2, 3 to 1, 2, the foregoing equations reduce to that for two-dimensional (plane stress, plane strain and axisymmetric) problems. For axisymmetric problems displacements, strains and stresses are independent of the angular coordinate θ, the

tangential displacement (in the θ-direction) vanishes and one usually considers a sector of the body that subtends an angle of one radian at the center. Therefore, the volume element $d\Omega = rdrdz(1)$.

For plane strain problems with the cross-section of the body in the x_1x_2-plane, u_3-the displacement of a body point in the x_3-direction is taken to vanish identically, and u_1 and u_2 are assumed to be functions of x_1 and x_2 only. Thus for plane strain problems, $e_{33} = e_{31} = e_{32} = 0$. However, for plane stress problems, with the cross-section of the body in the x_1x_2-plane, the dimension of the body in the x_3-direction is taken to be very small as compared to that in the x_1- or the x_2-direction and the two end-faces are not subjected to surface tractions simultaneously. One generally assumes that $\sigma_{33} = 0 = \sigma_{31} = \sigma_{32}$, and $u_i = u_i(x_1, x_2)$. One generally solves the eqn. $\sigma_{33} = 0$ for e_{33}, and substitutes for e_{33} in the expressions for σ_{11} and σ_{22} to obtain the modified (or the reduced) constitutive equations. The displacement $u_3(x_1, x_2)$ can be related to u_1 and u_2, and one thus solves for $u_1(x_1, x_2)$ and $u_2(x_1, x_2)$.

The assumption of plane stress is generally reasonable for thin structural elements such as plates and shells, and that of plane strain for long prismatic bodies such as a long tunnel.

8.2 Shape Functions

A simple three-dimensional master element is a cube of side 2 shown in Fig. 8.2.1; shape functions for its nodes can be generated by taking the tensor product of three linear one-dimensional

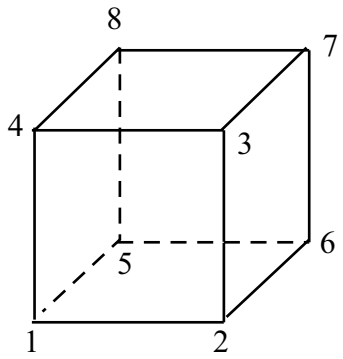


Fig.8.2.1 A cubic element

elements. Thus

$$N_i(r, s, t) = \frac{1}{8}(1 + r_i r)(1 + s_i s)(1 + t_i t),\ i = 1, 2, \ldots, 8, \tag{8.2.1}$$

where (r_i, s_i, t_i) are coordinates of the ith node. Even though N_i given by (8.2.1) contains monomials of degree 3 (e.g. the rst term), the shape functions are complete polynomials of degree one only. By taking the tensor product of three quadratic one-dimensional elements we obtain a triquadratic cubic element whose shape functions are complete polynomials of degree two.

Another simple three-dimensional master element is a tetrahedron, depicted in Fig. 8.2.2, whose edges are of unit length. For any point P, we can define volume coordinates (r, s, t, u) by

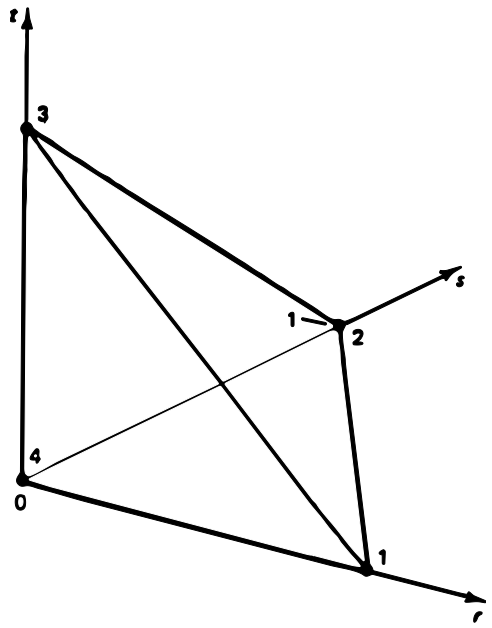


Fig. 8.2.2 A 4-noded tetrahedron element

$$\begin{aligned} r &= \text{vol. of tetrahedron } P234/\text{vol. of tetrahedron } 1234, \\ s &= \text{vol. of tetrahedron } P314/\text{vol. of tetrahedron } 1234, \\ t &= \text{vol. of tetrahedron } P124/\text{vol. of tetrahedron } 1234, \\ u &= \text{vol. of tetrahedron } P123/\text{vol. of tetrahedron } 1234. \end{aligned} \tag{8.2.2}$$

Thus

$$r + s + t + u = 1 \,,$$

and shape functions for the four nodes are given by

$$N_1 = r,\ N_2 = s,\ N_3 = t,\ N_4 = u = 1 - r - s - t \,. \tag{8.2.3}$$

Shape functions (8.2.3) are complete polynomials of degree one and do not contain monomials of higher degree. By adding a node at the midside of each edge, we get a tetrahedron with ten nodes;

the corresponding shape functions can be developed by following a procedure analogous to that used for the 6-noded triangular element.

8.3 Numerical Integration on Quadrilateral, Cubic and Tetrahedral Elements

Folloowing the reasoning given in Section 7.9, for integrating a polynomial function $f(r,s,t)$ of degree $(2m+1)$ in r, $(2n+1)$ in s and $(2p+1)$ in t over the master cube, we use $(m+1)$ Gauss points in the r-direction, $(n+1)$ in the s-direction and $(p+1)$ in the t-direction. That is

$$\begin{aligned} I &= \int_{-1}^{1} dr \int_{-1}^{1} ds \int_{-1}^{1} f(r,s,t)dt \\ &= \sum_{i=1}^{m+1} W_i^r \sum_{j=1}^{n+1} W_j^s \sum_{k=1}^{p+1} f(r_i,s_j,t_k) W_k^t \,. \end{aligned} \tag{8.3.1}$$

We note that the quadrature rule indicated in (8.3.1) is not necessarily optimum as it is for the one-dimensional case. Nevertheless, it is the most-often used.

For integration over a tetrahedron, we note that

$$\int_{\Omega} r^{\alpha} s^{\beta} t^{\gamma} u^{\delta} d\Omega = \frac{\alpha!\beta!\gamma!\delta!}{(\alpha+\beta+\gamma+\delta+3)!} 6V \tag{8.3.2}$$

where V, the volume of the tetrahedron is given by

$$6V = \det \begin{bmatrix} 1 & x_1 & y_1 & z_1 \\ 1 & x_2 & y_2 & z_2 \\ 1 & x_3 & y_3 & z_3 \\ 1 & x_4 & y_4 & z_4 \end{bmatrix} \tag{8.3.3}$$

x_a, y_a, z_a are the coordinates of vertex a. If the integrand is not a monomial, numerical integration needs to be employed. The following Table 8.3.1 lists a few integration rules; others may be found in P.C. Hammer, O.P. Marlowe and A.H. Strand, Numerical Integration over Simplexes and Cones, *Mathematical Tables and Aids to Computation*, 10 (1956) 130-137, and Y. Jinyun, Symmetric Gaussian Quadrature Formulae for Tetrahedral Regions, *Computer Methods in Applied Mechanics and Engineering*, 43 (1984), 349-353.

Table 8.3.1: Integration Rules for a Tetrahedron (Hammer et al.)

W_a	r_a	s_a	t_a	u_a	Multiplicity
	1-point formula				
$1/6$	$1/4$	$1/4$	$1/4$	$1/4$	1
	4-point formula				
$1/24$	0.58541020	0.13819660	0.13819660	0.13819660	4
	5-point formula				
$-2/15$	$1/4$	$1/4$	$1/4$	$1/4$	1
$3/40$	$1/3$	$1/6$	$1/6$	$1/6$	4

8.4 Shape Functions for Singular Problems

For problems involving cracks and defects, the deformation field near a crack tip and in the neighborhood of defects exhibits singularities in the sense that gradients of the deformation field go to infinity as the crack tip is approached. The analysis of such problems by the finite element method is facilitated by the use of special elements for which the trial solution exhibits the expected singular behavior. One possibility is to locate nodes on the sides of the element Ω_e at certain specific locations; an example being to locate the interior node in a quadratic one-dimensional element at $L/4$ from the left end where L is the length of Ω_e. Another way is to use in the expressions for shape functions terms that mimic the expected singular behavior. We now discuss the second alternative for one-dimensional elements. We assume that the origin is located at the left end of the element and denote the distance of a point on the element from the left end by r.

For a two-node element with node 1 at $r = 0$ and node 2 at $r = 1$, shape functions

$$N_1(r) = 1 - r^\alpha,\ N_2(r) = r^\alpha, \qquad r \in [0, 1], \tag{8.4.1}$$

enable one to express the trial solution in terms of r^α. For $\alpha < 1,\ \frac{du}{dr}$ will be singular at $r = 0$. If we wish to express u as a linear combination of $1,\ r$ and r^α, then we can take a three-noded element with node 1 at $r = 0$, node 2 at $r = 1/2$ and node 3 at $r = 1$. We temporarily take

$$N_1 = 1 - 2r,\ N_2 = 2r \text{ and } N_3 = r^\alpha\ . \tag{8.4.2}$$

Shape functions N_1 and N_2 vanish at nodes 2 and 1, respectively, and equal 1 at nodes 1 and 2. However, they do not vanish at node 3 and N_3 does not vanish at nodes 1 and 2. We first modify

N_3 so that it equals 0, 0 and 1 at nodes 1, 2 and 3 respectively by

$$N_3 \leftarrow \frac{N_3 - N_3(0)N_1 - N_3\left(\frac{1}{2}\right)N_2}{N_3(1) - N_3(0)N_1(1) - N_3\left(\frac{1}{2}\right)N_2(1)}, \tag{8.4.3}$$

where $\leftarrow$ implies that the left-hand side is replaced by the expression on the right-hand side. Thus

$$N_3 = \frac{r^\alpha - \left(\frac{1}{2}\right)^\alpha (2r)}{1 - \left(\frac{1}{2}\right)^\alpha (2)} = \left(r^\alpha - 2^{1-\alpha} r\right) / \left(1 - 2^{1-\alpha}\right). \tag{8.4.4}$$

We now modify N_1 and N_2 as follows:

$$N_1 \leftarrow (N_1 - N_1(1)N_3), \tag{8.4.5}$$

$$N_2 \leftarrow (N_2 - N_2(1)N_3). \tag{8.4.6}$$

Hence

$$N_1 = 1 - 2r + \left(r^\alpha - 2^{1-\alpha} r\right) / \left(1 - 2^{1-\alpha}\right), \tag{8.4.7}$$

$$N_2 = 2r - 2\left(r^\alpha - 2^{1-\alpha} r\right) / \left(1 - 2^{1-\alpha}\right). \tag{8.4.8}$$

As a third example, we consider the case when we like to develop shape functions capable of representing exactly $1,\ r^\alpha,\ r^{2\alpha},\ r^\beta$. We start with a four-node element with nodes 1, 2, 3 and 4 located at $r = 0,\ 1/3,\ 2/3$ and 1 respectively. We take for $N_1,\ N_2,\ N_3$ Lagrange polynomials generated by setting $s = r^\alpha$. That is

$$N_1 = \frac{\left(r^\alpha - \left(\frac{1}{3}\right)^\alpha\right)\left(r^\alpha - \left(\frac{2}{3}\right)^\alpha\right)}{\left(0 - \left(\frac{1}{3}\right)^\alpha\right)\left(0 - \left(\frac{2}{3}\right)^\alpha\right)} = \left(3^\alpha r^\alpha - 1\right)\left(3^\alpha r^\alpha - 2^\alpha\right)/2^\alpha, \tag{8.4.9}$$

$$N_2 = \frac{\left(r^\alpha - 0^\alpha\right)\left(r^\alpha - \left(\frac{2}{3}\right)^\alpha\right)}{\left(\left(\frac{1}{3}\right)^\alpha - 0^\alpha\right)\left(\left(\frac{1}{3}\right)^\alpha - \left(\frac{2}{3}\right)^\alpha\right)} = 3^\alpha r^\alpha \left(3^\alpha r^\alpha - 2^\alpha\right) / \left(1 - 2^\alpha\right), \tag{8.4.10}$$

$$N_3 = \frac{\left(r^\alpha - 0^\alpha\right)\left(r^\alpha - \left(\frac{1}{3}\right)^\alpha\right)}{\left(\left(\frac{2}{3}\right)^\alpha - 0^\alpha\right)\left(\left(\frac{2}{3}\right)^\alpha - \left(\frac{1}{3}\right)^\alpha\right)} = 3^\alpha r^\alpha \left(3^\alpha r^\alpha - 1\right)/2^\alpha \left(2^\alpha - 1\right). \tag{8.4.11}$$

We temporarily take

$$N_4 = r^\beta \tag{8.4.12}$$

and modify $N_1,\ N_2,\ N_3$ and N_4 to satisfy the requirements for shape functions:

$$N_4 \leftarrow \frac{N_4 - N_4(0)N_1 - N_4\left(\frac{1}{3}\right)N_2 - N_4\left(\frac{2}{3}\right)N_3}{N_4(1) - N_4(0)N_1(1) - N_4\left(\frac{1}{3}\right)N_2(1) - N_4\left(\frac{2}{3}\right)N_3(1)}. \tag{8.4.13}$$

For $a = 1, 2, 3$

$$N_a \leftarrow \left(N_a - \sum_{\substack{b=1 \\ b \neq a}}^{4}, N_a\left(r_b\right) N_b\right) \tag{8.4.14}$$

or

$$N_1 \leftarrow (N_1 - N_1(1)N_4)\,, \tag{8.4.15}$$

$$N_2 \leftarrow (N_2 - N_2(1)N_4)\,, \tag{8.4.16}$$

$$N_3 \leftarrow (N_3 - N_3(1)N_4)\,. \tag{8.4.17}$$

Thus one can see the advantage of having as many shape functions as possible satisfy the requirement that they equal one at a given node and vanish at the remaining nodes.

It should be noted that whereas the trial solution on an element is expressed in terms of the above given shape functions, the transformation T_e that maps Ω_M onto Ω_e is expressed in terms of regular polynomial shape functions. Thus for shape functions (8.4.1)

$$u^e = u_1\,(1 - r^\alpha) + u_2 r^\alpha \tag{8.4.18}$$

and

$$T_e: \quad x = x_1(1 - r) + x_2 r \tag{8.4.19}$$

Substitution for r from (8.4.19) into (8.4.18) yields

$$u^e = u_1 + (u_2 - u_1)\left(\frac{x - x_1}{x_2 - x_1}\right)^\alpha \tag{8.4.20}$$

which implies that for $\alpha < 1$, $\frac{du}{dx}^e$ will be singular at $x = x_1$. Had we used

$$T_e': \quad x = x_1\,(1 - r^\alpha) + x_2 r^\alpha\,, \tag{8.4.21}$$

then we would have obtained

$$u^e = u_1 + (u_2 - u_1)\left(\frac{x - x_1}{x_2 - x_1}\right), \tag{8.4.22}$$

and u^e will not exhibit the desired singular behavior.

Two-dimensional elements that exhibit the desired singular behavior can be developed by taking the tensor product of a one-dimensional singular element with a one-dimensional regular element. Thus

$$N_1 = (1 - r^\alpha)\,(1 - s),\ N_2 = r^\alpha(1 - s)\,, \tag{8.4.23}$$

$$N_3 = r^\alpha s,\ N_4 = (1 - r^\alpha)\,s \tag{8.4.24}$$

are shape functions for the 4-node quadrilateral element shown in Fig. 8.4.1 that exhibits the singular behavior for $\alpha < 1$ in the s-direction.

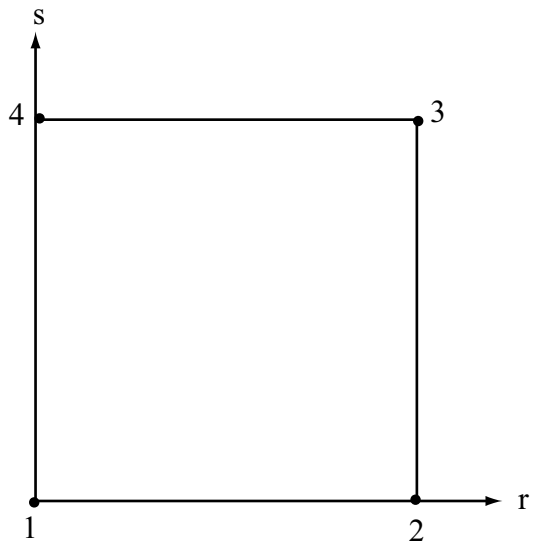


Fig. 8.4.1 A 4-node quadrilateral element

8.5 Characteristics of the Galerkin Approximate Solution

We now discuss the optimal properties of the Galerkin approximate solution. We note that boundary-value problems involving second-order differential equations can be written in the following weak form. Find $u \in H^1$ such that

$$B(\phi, u) = \ell(\phi) \tag{8.5.1}$$

for every $\phi \in H_0^1$ where functions in H^1 satisfy the given essential boundary conditions and functions in H_0^1 satisfy the corresponding homogeneous essential boundary conditions. In (8.5.1) $B(\cdot, \cdot)$ is the bilinear form and $\ell(\cdot)$ the linear functional associated with the given boundary-value problem. The approximate finite element solution $u^n \in H^{1n}$ satisfies

$$B(\phi^n, u^n) = \ell(\phi^n) \tag{8.5.2}$$

for every $\phi^n \in H_0^{1n}$ where H^{1n} and H_0^{1n} are finite-dimensional subsets of H^1 and H_0^1 respectively. We assume that

(i) $B(\cdot, \cdot)$ is symmetric, bilinear and positive definite, and

(ii) $B(\cdot, \cdot)$ and $\| \cdot \|_m$ define equivalent norms on H^1, that is,

$$\|\phi\|_m \leq C_1 B(\phi, \phi) \leq C_2 \|\phi\|_m \tag{8.5.3}$$

where

$$\|\phi\|_m^2 = \int_\Omega \left(\phi^2 + \phi_{,i}\phi_{,i} + \phi_{,ij}\phi_{,ij} + \ldots + \phi_{,i_1 i_2 \ldots i_m}\phi_{,i_1 i_2 \ldots i_m}\right) d\Omega, \tag{8.5.4}$$

and C_1 and C_2 are constants that are independent of ϕ. For solid mechanics problems, $B(\phi, \phi)$ is called the strain energy associated with the deformation field ϕ. We now state and prove the main result. Let

$$e = u^n - u \tag{8.5.5}$$

denote the error in the finite element approximation. Then

(i) $B(\psi^n, e) = 0 \quad \forall \psi^n \in H_0^{1n}$,

(ii) $B(e, e) \leq B(U^n - u,\ U^n - u) \quad \forall\, U^n \in H^{1n}$.

Proof. To prove (i) we note that (8.5.1) holds for every $\phi \in H_0^1$. Therefore, (8.5.1) also holds for $\phi^n \in H_0^{1n}$. Thus

$$B(\phi^n, u) = \ell(\phi^n) \quad \forall \phi^n \in H_0^{1n} \tag{8.5.6}$$

Subtraction of (8.5.6) from (8.5.2) yields

$$B(\phi^n, u^n - u) = 0 \text{ or } B(\phi^n, e) = 0\ \forall \phi^n \in H_0^{1n}. \tag{8.5.7}$$

We now prove (ii). Let $\phi^n \in H_0^{1n}$. Then

$$B(e + \phi^n, e + \phi^n) = B(e, e) + 2B(\phi^n, e) + B(\phi^n, \phi^n) \tag{8.5.8}$$

which follows from the bilinearity of B. Since $B(\phi^n, e) = 0\ \forall \phi^n \in H_0^{1n}$ and $B(\phi^n, \phi^n) \geq 0$, therefore

$$B(e, e) \leq B(e + \phi^n,\ e + \phi^n) \quad \forall \phi^n \in H_0^{1n}\ . \tag{8.5.9}$$

Any $U^n \in H^{1n}$ can be written as

$$U^n = u^n + \phi^n\ . \tag{8.5.10}$$

Therefore,

$$e + \phi^n = u^n - u + \phi^n = U^n - u, \tag{8.5.11}$$

which when substituted in the right-hand side of (8.5.9) gives (ii).

Remarks: The result (i) implies that the error function e is orthogonal to every $\phi^n \in H_0^{1n}$. Thus the error lies in the orthogonal complement of H_0^{1n}. By making H_0^{1n} large, one can make its orthogonal

complement quite small. The result (ii) means that the strain energy associated with the error in the Galerkin finite element approximation is least. In this sense, the Galerkin approximate solution u^n is the "best" approximate solution of the problem.

When the given essential boundary conditions are homogeneous, then $H^{1n} = H_0^{1n}$, and we can show that

$$B(u,u) = B(u^n,u^n) + B(e,e) \tag{8.5.12}$$

To prove it, we note that $B(u^n,e) = 0$, thus

$$\begin{aligned} B(u,u) = B(u^n-e,u^n-e) =& B(u^n,u^n) - 2B(u^n,e) + B(e,e). \\ =& B(u^n,u^n) + B(e,e) \end{aligned}$$

Thus

$$B(u^n,u^n) \leq B(u,u), \tag{8.5.13}$$

and the strain-energy of the approximate solution is atmost equal to that of the analytical solution. Equation (8.5.12) when written as $B(e,e) = -\left(B(u^n,u^n) - B(u,u)\right)$ implies that the energy of the error equals the negative of the error of the energy.

The quantity $I(U)$, defined as

$$I(U) = \frac{1}{2}B(U,U) - (U,f)_\Omega - (U,h)_\Gamma \tag{8.5.14}$$

where

$$(U,f)_\Omega = \int_\Omega U f d\Omega \ , \tag{8.5.15}$$

$$(U,h)_\Gamma = \int_\Gamma U h d\Gamma \ , \tag{8.5.16}$$

represent the work done by the body force f and prescribed surface tractions h on Γ, is called the potential energy of the system. We now show that the potential energy is minimized by the analytical solution u. For any $\phi \in H_0^1$ and $\varepsilon \in \mathbb{R}$,

$$U_\varepsilon = u + \varepsilon\phi \in H^1 \ . \tag{8.5.17}$$

We now regard $I(U_\varepsilon)$ as a real-valued function of scalar ε. $I(U_\varepsilon)$ assumes a minimum value at u if and only if I as a function of ε takes on a minimum value at $\varepsilon = 0$.

$$\begin{aligned}\hat{I}(\varepsilon) =& I(U_\varepsilon) = \frac{1}{2}B(u,u) + \varepsilon B(u,\phi) + \frac{1}{2}\varepsilon^2 B(\phi,\phi) - (u,f)_\Omega \\ & - \varepsilon(\phi,f)_\Omega - (u,h)_\Gamma - \varepsilon(\phi,h)_\Gamma \ . \end{aligned} \tag{8.5.18}$$

We first assume that u is an analytical solution of the problem and show that $\hat{I}$ takes on a minimum value at $\varepsilon = 0$. Since an analytical solution u satisfies

$$B(u,\phi) = (u,f)_\Omega + (u,h)_\Gamma \ , \tag{8.5.19}$$

therefore

$$\begin{aligned}\hat{I}(\varepsilon) =& \frac{1}{2}B(u,u) + \frac{1}{2}\varepsilon^2 B(\phi,\phi) - (u,f)_\Omega - (u,h)_\Gamma \\ =& I(u) + \frac{1}{2}\varepsilon^2 B(\phi,\phi) \ . \end{aligned} \tag{8.5.20}$$

which takes on its minimum value at $\varepsilon = 0$.

To prove the converse, we assume that $\hat{I}$ is minimum at $\varepsilon = 0$. Therefore,

$$\left.\frac{d\hat{I}}{d\varepsilon}\right|_{\varepsilon=0} = B(u,\phi) - (\phi,f)_\Omega - (\phi,h)_\Gamma = 0 \tag{8.5.21}$$

which implies that u satisfies (8.5.1) with $\ell(\phi) = (\phi,f)_\Omega + (\phi,h)_\Gamma$ for every $\phi \in H_0^1$ and is therefore a solution of the problem. Since

$$\left.\frac{d^2\hat{I}}{d\varepsilon^2}\right|_{\varepsilon=0} = B(\phi,\phi) > 0 \quad \forall \phi \in H_0^1, \ \phi \neq 0, \tag{8.5.22}$$

therefore $\hat{I}$ is minimum at $\varepsilon = 0$.

The analytical solution u minimizes the potential energy implies that

$$I(u) \leq I(u^n) \ . \tag{8.5.23}$$

Thus if a sequence of approximate solutions obtained by refining the finite element mesh gives successively lower values of the potential energy, then the approximate solutions are converging to the analytical solution. This does not imply point-wise convergence, that is, the displacement at every node converges to the value of the analytical solution there.

Chapter 9: Vibrations

9.1 A Model Problem

We study free vibrations of a bar with its cross-section in the x_2x_3-plane and its length along the x_1-axis. For a bar, a typical dimension of its cross-section is much smaller than its length and it is usual to assume that $u_2 = u_3 = 0$ and $u_1 = u(x)$ where we have set $x_1 = x$. The only non-zero component of the infinitesimal strain tensor is $e_{11} = u_{1,1} = \partial u/\partial x$ and the only nonzero component of the stress tensor is $\sigma_{11} = \sigma = E\frac{\partial u}{\partial x}$ where E is the Young's modulus for the material of the bar. When the bar is vibrating freely, external forces do not do any work. Thus we consider the case when there is no body force applied, the end $x = 0$ of the bar is traction free and the

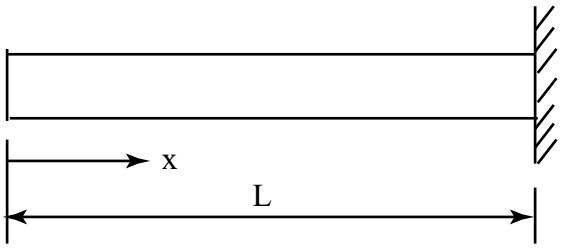


Fig. 9.1.1 A freely vibrating bar

end $x = L$ is clamped. The pertinent governing equation and boundary conditions are:

$$\frac{\partial}{\partial x}\left(E\frac{\partial u}{\partial x}\right) = \rho\frac{\partial^2 u}{\partial t^2},\ 0 < x < L\ , \tag{9.1.1}$$

$$\left(E\frac{\partial u}{\partial x}\right)\Big|_{x=0} = 0,\ u(L,t) = 0\ . \tag{9.1.2}$$

Here ρ is the mass density per unit volume. Because of infinitesimal deformations involved, one can either use the referential or the spatial description of motion. We seek solutions of equations (9.1.1) and (9.1.2) of the form

$$u(x,t) = \tilde{u}(x)e^{i\omega t} \tag{9.1.3}$$

Substitution of (9.1.3) into (9.1.1) and (9.1.2) gives

$$\frac{d}{dx}\left(E\frac{du}{dx}\right) + \lambda\rho u = 0,\ 0 < x < L \tag{9.1.4}$$

$$\left(E\frac{du}{dx}\right)\Big|_{x=0} = 0,\ u(L) = 0 \tag{9.1.5}$$

where we have dropped the superimposed tildas on u and set $\lambda = \omega^2$. Equations (9.1.4) and (9.1.5) are homogeneous linear equations in u and, of course, have a trivial solution $u = 0$. Also, if u is a solution of (9.1.4) and (9.1.5) so is αu where α is an arbitrary constant. We remove this nonuniqueness in u by requiring that

$$\int_0^L \rho u^2 dx = 1 \, . \tag{9.1.6}$$

Other normalization techniques such as

$$\sup_{0<x<L} |u(x)| = 1 \tag{9.1.7}$$

will also be equally good. For certain discrete values of λ, equations (9.1.4) and (9.1.5) have a nontrivial solution; these values of λ are called eigenvalues and the corresponding nontrivial u's are called eigenfunctions. So our task is to find the eigenpair (λ, u). Here we will seek an approximate solution of the problem.

9.2 Weak Formulation

Let $\phi : [0, L] \to \mathbb{R}$ be a smooth function such that $\phi(L) = 0$. Multiplication of both sides of (9.1.4) by ϕ, integration of the result over $(0, L)$, integration by parts of the first term, and the use of boundary condition $(9.1.5)_1$ and $\phi(L) = 0$ gives

$$-\int_0^L E\frac{du}{dx}\frac{d\phi}{dx}dx + \lambda \int_0^L \rho u \phi dx = 0 \, . \tag{9.2.1}$$

With

$$B(u, \phi) = \int_0^L E\frac{du}{dx}\frac{d\phi}{dx} \, , \tag{9.2.2a}$$

$$(\rho u, \phi) = \int_0^L \rho u \phi dx \tag{9.2.2b}$$

we write (9.2.1) as

$$B(u, \phi) = \lambda(\rho u, \phi) \tag{9.2.3}$$

Let

$$H^1 = \left\{ \psi | \psi : [0, L] \to \mathbb{R}, \ \int_0^L \left(\frac{d\psi}{dx}\right)^2 dx < \infty \right\} \tag{9.2.4a}$$

$$H_0^1 = \left\{ \psi | \psi \in H^1, \ \psi(L) = 0 \right\} , \tag{9.2.4b}$$

then a weak statement of the problem defined by equations (9.1.4) and (9.1.5) can be stated as follows: Find an eigenpair (λ, u), $\lambda \in \mathbb{R},\ u \in H_0^1$ such that equation (24.10) holds for every $\phi \in H_0^1$. We note that this is also the Galerkin formulation of the problem. Let $H_0^{1n} \subset H_0^1$ be a finite dimensional subset of H_0^1; then the Galerkin approximation of the problem can be stated as: find (λ^n, u^n), $\lambda^n \in \mathbb{R}$, $u^n \in H_0^{1n}$ such that

$$B(u^n, \phi^n) = \lambda^n (\rho u^n, \phi^n) \tag{9.2.5}$$

for every $\phi^n \in H_0^{1n}$. With $\psi_1, \psi_2, \ldots, \psi_n$ as the basis functions in H_0^{1n}, we can write

$$\phi^n = \sum_{i=1}^{n} c_i \psi_i,\ u^n = \sum_{i=1}^{n} d_i \psi_i\ , \tag{9.2.6}$$

$$B(u^n, \phi^n) = c_i \left(\int_0^L E \frac{d\psi_i}{dx} \frac{d\psi_j}{dx} dx \right) d_j = c_i K_{ij} d_j\ , \tag{9.2.7}$$

$$(\rho u^n, \phi^n) = c_i \left(\int_0^L \rho \psi_i \psi_j dx \right) d_j = c_i M_{ij} d_j\ , \tag{9.2.8}$$

where we have set

$$K_{ij} = \int_0^L E \frac{d\psi_i}{dx} \frac{d\psi_j}{dx} dx = K_{ji}\ , \tag{9.2.9a}$$

$$M_{ij} = \int_0^L \rho \psi_i \psi_j dx = M_{ji}\ , \tag{9.2.9b}$$

M_{ij} is the mass matrix and K_{ij} the usual stiffness matrix. Thus equation (9.2.5) becomes

$$c_i \left(K_{ij} - \lambda^n M_{ij} \right) d_j = 0 \tag{9.2.10}$$

which should hold for arbitrary values of $c_1, c_2, \ldots, c_n$. Thus

$$\left(K_{ij} - \lambda M_{ij} \right) d_j = 0 \tag{9.2.11}$$

where we have dropped the superscript n from λ^n since it is understood that (9.2.11) refers to the evaluation of the approximate solution. Since $u^n \in H_0^{1n}$, the essential boundary condition $u^n(L) = 0$ is built into (9.2.11); otherwise we need to modify the set of equations (9.2.11) to satisfy the essential boundary conditions. The normalization condition (9.1.6) takes the form

$$d_i M_{ij} d_j = 1\ . \tag{9.2.12}$$

Equations (9.2.11) will have a non-trivial solution if and only if

$$\det\left(K - \lambda M \right) = 0 \tag{9.2.13}$$

whose n roots $\lambda_{(1)}, \lambda_{(2)}, \ldots, \lambda_{(n)}$ are the approximate n-eigenvalues of the given problem. For any one of these values of $\lambda,\ (n-1)$ equations out of (9.2.11) and (9.2.12) are used to find the corresponding d's; then the approximate eigenfunction is given by $(9.2.6)_2$. Henceforth we assume that both the mass matrix and the stiffness matrix are positive-definite; this will be the case if E and ρ are positive everywhere which is a reasonable assumption. We can now establish the following results.

(a) All of the eigenvalues are positive. Taking the inner product of (9.2.11) with d_i, we obtain

$$d_i K_{ij} d_j = \lambda d_i M_{ij} d_j = \lambda \,, \tag{9.2.14}$$

and since K_{ij} is a positive definite matrix, therefore $\lambda > 0$.

(b) Eigenfunctions corresponding to distinct eigenvalues are mutually orthogonal, i.e., if $\lambda_{(m)} \neq \lambda_{(n)}$, then $d_i^{(m)} M_{ij} d_j^{(n)} = 0$.
Eigenpairs $(\lambda_{(m)},\ u_{(m)}),\ (\lambda_{(n)},\ u_{(n)})$ satisfy

$$K_{ij} d_j^{(m)} = \lambda_{(m)} M_{ij} d_j^{(m)} \tag{9.2.15a}$$

$$K_{ij} d_j^{(n)} = \lambda_{(n)} M_{ij} d_j^{(n)} \tag{9.2.15b}$$

Taking the inner product of $(9.2.15a)$ with $d_i^{(n)}$, of $(9.2.15b)$ with $d_i^{(m)}$, subtracting one from the other, and recalling that both K and M are symmetric we obtain

$$\left(\lambda_{(m)} - \lambda_{(n)}\right) d_i^{(m)} M_{ij} d_j^{(n)} = 0 \tag{9.2.16}$$

Since $\lambda_{(m)} - \lambda_{(n)} \neq 0$, therefore $d_i^{(m)} M_{ij} d_j^{(n)} = 0$ implying thereby that $\mathbf{d}^{(m)}$ and $\mathbf{d}^{(n)}$ are mutually orthogonal with respect to the mass matrix.

(c) The error in the frequency is of the same order of magnitude as that in the eigenvalue. Let

$$\lambda_\ell / \lambda_{(\ell)} = 1 + \epsilon \,, \tag{9.2.17}$$

where λ_ℓ is the analytical value corresponding to $\lambda_{(\ell)}$. Since

$$\frac{\omega_\ell}{\omega_{(\ell)}} = \left(\frac{\lambda_\ell}{\lambda_{(\ell)}}\right)^{1/2} = (1+\epsilon)^{1/2} = 1 + \frac{1}{2}\epsilon + O(\epsilon^2) \,, \tag{9.2.18}$$

therefore,

$$\omega_\ell / \omega_{(\ell)} - 1 = O(\epsilon) \,. \tag{9.2.19}$$

(d) Let the eigenvalues be ordered as

$$0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \lambda_n \leq \lambda_{n+1} \ldots \tag{9.2.20}$$

then one can show that

$$\lambda_\ell \leq \lambda_{(\ell)} \leq \lambda_\ell + ch^{2(k+1-m)}(\lambda_\ell)^{(k+1)/m} \tag{9.2.21}$$

where h is the mesh parameter, k is the degree of the complete polynomial in the basis functions, $2m$ is the order of the given differential equation, and c is a constant independent of h, k and m. Typically, h equals the diameter of the smallest circle or sphere enclosing the largest element in the mesh. Thus the approximate eigenvalue is always greater than or equal to the corresponding analytical one, and the error in the higher eigenvalue increases because of the presence of λ_ℓ in the upper bound for $\lambda_{(\ell)}$. The error in the eigenfunctions is given by

$$\|u_\ell - u_{(\ell)}\|_m \leq ch^{(k+1-m)}(\lambda_\ell)^{(k+1)/2m} \,. \tag{9.2.22}$$

Thus the rate of convergence of an eigenfunction with the refinement of the mesh is one-half that of the corresponding eigenvalue. The error in the L_2-norm (or H^0-norm) in the eigenfunction is given by

$$\|u_\ell - u_{(\ell)}\|_0 \leq ch^{\sigma}(\lambda_\ell)^{(k+1)/2m} \tag{9.2.23}$$

where $\sigma = \min(k+1,\ 2(k+1-m))$.

(e) The accuracy of the eigenvalues and eigenfunctions depends upon the quality of both the mass matrix M and the stiffness matrix K.

(f) If numerical quadrature is employed to evaluate M and K, then a rule accurate enough to exactly integrate all monomials through degree $\overline{k} + k - 2m$ is sufficient to maintain full rate of convergence. A sufficient condition for convergence is that the integration rule exactly integrate monomials through degree $\overline{k} - m$. Here $\overline{k}$ equals the order of the highest-order monomial appearing in the element shape functions. For a bilinear quadrilateral element $\overline{k} = 2$ because of the st term, for a biquadratic quadrilateral $\overline{k} = 4$ due to the s^2t^2 term, and for a trilinear brick element, $\overline{k} = 3$ because of the rst term.

9.3 Diagonal Mass Matrices

9.3.1 Techniques to diagonalize the mass matrix

The consistent mass matrix

$$M_{ij}^e = \int_{\Omega_e} \rho N_i N_j d\Omega \tag{9.3.1}$$

is generally fully populated. In developing algorithms for numerically integrating the parabolic and/or hyperbolic set of equations, it is more convenient to use a diagonal or lumped mass matrix. These diagonal mass matrices can be obtained in the following ways.

(a) Use nodes as quadrature points.

$$M_{ij}^e = \int_{\Omega_e} \rho N_i N_j d\Omega = \int_{\Omega_M} \rho N_i N_j J ds dt \tag{9.3.2a}$$

$$= \sum_{a=1}^{\text{nodes}} \rho(s_a, t_a) N_i(s_a, t_a) N_j(s_a, t_a) J(s_a, t_a) W_a \tag{9.3.2b}$$

$$= \sum_{a=1}^{\text{nodes}} \delta_{ia}\delta_{ja} \rho(s_a, t_a) W_a J(s_a, t_a) \tag{9.3.2c}$$

We note that this technique can result in negative masses which are undesirable since they result in a blow-up of the solution in time. Also, if nodal quadrature is used for axisymmetric problems, zero masses result for nodes on the axis of symmetry. Zero masses are equally undesirable.

(b) Row-Sum Technique

In this case the element ii of the diagonal mass matrix is obtained by summing all of the elements of the ith row in the consistent mass matrix. That is,

$$M_{ii} = \sum_{j=1}^{\text{nodes}} \int_{\Omega_e} \rho N_i N_j d\Omega = \int_{\Omega_e} \rho N_i d\Omega, \text{ no sum on } i\ . \tag{9.3.3}$$

It eliminates the problem of zero masses at the nodes but can produce negative masses. This is the case for the corner nodes of the eight-node serendipity element. In (9.3.3) $\sum_{j=1}^{\text{nodes}} N_j(s,t) = 1$ has been used.

(c) Special Lumping Technique

The elements of the diagonal mass matrix are taken to be proportional to the corresponding elements of the consistent mass matrix; the factor of proportionality is such that the total mass in

the two cases is the same. That is

$$M^e_{ii} = \alpha \int_{\Omega_e} \rho N_i^2 d\Omega,\ M^e_{ij} = 0,\ i \neq j \tag{9.3.4a}$$

$$\alpha = \int_{\Omega_e} \rho d\Omega / \left(\sum_i \int_{\Omega_e} \rho N_i^2 d\Omega \right) . \tag{9.3.4b}$$

Remark: Although the lumped mass matrix has been used successfully in solid mechanics, structural mechanics and heat transfer problems, some disappointing results have been obtained in fluid mechanics.

In one-dimensional problems, $\bar{k} = k$ and full rate of convergence is maintained if the integration rule is capable of exactly integrating polynomials of degree $2(k-m)$; the minimum condition for convergence is a rule that can exactly integrate polynomials of degree $(k-m)$.

9.3.2 Expressions for diagonal mass matrices:

(a) One-dimensional linear element.

The integration rule should be capable of exactly integrating polynomials of degree $2(k-m) = 2(1-1) = 0$ or constants. The trapezoidal rule that uses end-points as integration points should suffice. Thus for $a = 1, 2,\ s_a = -1, 1,\ W_a = 1, 1$.

$$\begin{aligned} M^e_{11} =& \int_0^h \rho N_1^2 dx = \int_{-1}^1 \rho N_1^2 \frac{dx}{ds} ds = \rho \frac{h}{2} \int_{-1}^1 N_1^2 ds \\ =& \frac{\rho h}{2} \left[(1)^2 \cdot 1 + 0^2 \cdot 1 \right] = \frac{\rho h}{2} . \\ M_{22} =& \rho \frac{h}{2} \int_{-1}^1 N_2^2 ds = \frac{\rho h}{2} \left[1 \cdot 0^2 + (1)^2 \cdot 1 \right] = \frac{\rho h}{2} . \\ M^e =& \frac{\rho h}{2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} . \end{aligned} \tag{9.3.5}$$

Here we have assumed that the mass density is constant, and the length of the element equals h.

(b) One dimensional quadratic element

For full rate of convergence, the integration rule should integrate exactly polynomials of degree $2(k-m) = 2(2-1) = 2$. Simpson's rule that uses node points as integration points will suffice. Thus

$$\begin{aligned} & s_1 = -1,\ s_2 = 0,\ s_3 = 1, \\ & W_1 = \frac{1}{3},\ W_2 = \frac{4}{3},\ W_3 = \frac{1}{3}, \end{aligned}$$

$$
\begin{aligned}
M^e_{11} &= \int_0^h \rho N_1^2 dx = \rho \int_{-1}^1 N_1^2 \frac{h}{2} ds = \frac{\rho h}{2} \int_{-1}^1 \left(\frac{1}{2}s(s-1)\right)^2 ds \\
&= \frac{\rho h}{2}\left[\frac{1}{3}(1)^2 + \frac{4}{3}(0) + \frac{1}{3}(0)\right] = \frac{\rho h}{6}\,, \\
M^e_{22} &= \int_0^h \rho N_2^2 dx = \rho \int_{-1}^1 N_2^2 \frac{h}{2} ds = \frac{\rho h}{2} \int_{-1}^1 \left((1-s^2)\right)^2 ds \\
&= \frac{\rho h}{2}\left[\frac{1}{3}(0) + \frac{4}{3}(1) + \frac{1}{3}(0)\right] = \frac{4\rho h}{6}\,,
\end{aligned}
$$

and

$$
M^e = \frac{\rho h}{6}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{9.3.6}
$$

(c) One dimensional cubic element.

In order to maintain full rate of convergence the integration rule should integrate exactly polynomials of degree $2(3-1) = 4$. Thus the problem is to find the weights W_1, W_2, W_3 and W_4 so that when quadrature points are located at $s_1 = -1,\ s_2 = -\frac{1}{3},\ s_3 = \frac{1}{3},\ s_4 = 1$, the integration rule integrates exactly terms $1, s, s^2, s^3$ and s^4 over $(-1, 1)$. That is

$$
\int_{-1}^1 g(s) ds = \sum_{a=1}^4 g(s_a) W_a \tag{9.3.7}
$$

For $g(s) = 1, s, s^2, s^3$ and s^4, equation (9.3.7) gives

$$
2 = W_1 + W_2 + W_3 + W_4 \tag{9.3.8a}
$$

$$
0 = -W_1 - \frac{1}{3}W_2 + \frac{1}{3}W_3 + W_4 \tag{9.3.8b}
$$

$$
\frac{2}{3} = W_1 + \frac{1}{9}W_2 + \frac{1}{9}W_3 + W_4 \tag{9.3.8c}
$$

$$
0 = W_1 - \frac{1}{27}W_2 + \frac{1}{27}W_3 + W_4 \tag{9.3.8d}
$$

$$
\frac{2}{5} = W_1 + \frac{1}{81}W_2 + \frac{1}{81}W_3 + W_4\,. \tag{9.3.8e}
$$

The first four of these give $W_1 = 1/4 = W_4$, $W_2 = 3/4 = W_3$, which do not satisfy $(9.3.8e)$. To remedy this, we assume that the interior nodes are not located at a distance of 1/3 from the origin but are at a distance of C which is to be determined. Without loss in generality we can assume that

$W_1 = W_4$ and $W_2 = W_3$. Equations analogous to $(9.3.8a)$ through $(9.3.8e)$ are

$$2 = W_1 + W_2 + W_3 + W_4 = 2(W_1 + W_2)\ , \tag{9.3.9a}$$

$$0 = -W_1 - CW_2 + CW_3 + W_4 \tag{9.3.9b}$$

$$\frac{2}{3} = 2W_1 + 2W_2C^2 \tag{9.3.9c}$$

$$0 = -W_1 - C^3W_2 + C^3W_3 + W_4 \tag{9.3.9d}$$

$$\frac{2}{5} = 2(W_1 + C^4W_2) \tag{9.3.9e}$$

A solution of equations $(9.3.9a)$ - $(9.3.9e)$ is

$$W_1 = 1/6,\ W_2 = 5/6,\ C = 1/\sqrt{5}\ .$$

Thus if the nodes of a cubic element are located at $-1,\ -1/\sqrt{5},\ 1/\sqrt{5},\ 1$ and node points are used as the quadrature points then the resulting mass matrix will be the following diagonal matrix.

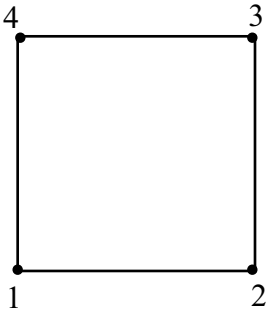
$$M^e = \frac{\rho h}{12}\begin{bmatrix} 1 & & & \\ & 5 & & 0 \\ & & 5 & \\ & 0 & & 1 \end{bmatrix} \tag{9.3.10}$$

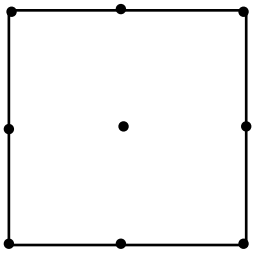
and full rate of convergence will be preserved.

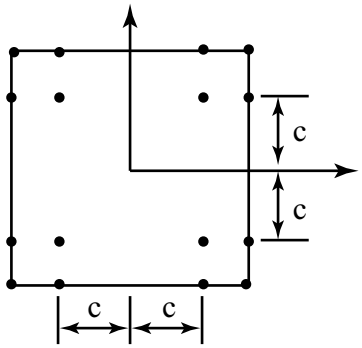
Quadrature (Integration) rules that use endpoints as integration points are called Lobatto's rule. The trapezoidal rule and Simpson's rule are special cases of Lobatto's rule. The preceding example establishes integration points and the corresponding weights for a fourth-order Lobatto's rule.

(d) Two-dimensional elements.

When a two-dimensional element can be regarded as the tensor product of two one dimensional elements, then the corresponding Lobatto's rule can be easily established. Since node-points are taken as integration points, we only need to specify the weights; these are given below for three quadrilateral elements.







$W_a = 1,\ a = 1, 2, 3, 4$

$W_a = \frac{1}{9}$ for corner nodes
$= \frac{4}{9}$ for midside nodes
$= \frac{16}{9}$ for the central node.

$W_a = \frac{1}{36}$ for corner nodes
$= \frac{5}{36}$ for nodes on the sides
$= \frac{25}{36}$ for interior nodes.

An eight-node serendipity element cannot be regarded as the tensor product of two one-dimensional elements. In order to maintain the full rate of convergence, the integration rule should integrate all monomials of degree $\overline{k} + k - 2m = 3 + 2 - 2 = 3$. Because of symmetry, we assume that $W_1 = W_2 = W_3 = W_4$, and $W_5 = W_6 = W_7 = W_8$. The equation

$$\int_{-1}^{1} dt \int_{-1}^{1} g(s,t)ds = \sum_{a=1}^{8} g(s_a, t_a)W_a \tag{9.3.11}$$

for $g(s,t) = 1, s, t, s^2, st, t^2, s^3, s^2t, st^2$ and t^3 gives

$$4 = 4(W_1 + W_5) \tag{9.3.12a}$$

$$\frac{4}{3} = 4W_1 + 2W_5 \tag{9.3.12b}$$

$$\frac{4}{3} = 4W_1 + 2W_5 \tag{9.3.12c}$$

and the remaining equations are $0 = 0$. A solution of equations (9.3.12) is $W_1 = -1/3,\ W_5 = 4/3$; thus the corresponding diagonal mass matrix will have negative masses at locations 11, 22, 33 and 44.

Chapter 10: Transient Parabolic Problems

10.1 Classification of Partial Differential Equations

The partial differential equation

$$a(x,y)u_{,xx} + 2b(x,y)u_{,xy} + c(x,y)u_{,yy} = f(x,y,u_{,x},u_{,y},u) \tag{10.1.1}$$

is elliptic, parabolic or hyperbolic at a point according as $b^2 - ac < 0, = 0$, or > 0, respectively, at that point. In (10.1.1) $u_{,x} = \partial u/\partial x,\ u_{,xx} = \partial^2 u/\partial x^2$ etc. This terminology is chosen by analogy with the general equation of a conic section. The equation $ax^2 + 2bxy + cy^2 + dx + ey + f = 0$ where $a, b, c, \ldots, f$ are constants, represents an ellipse, a parabola or an hyperbola, according as $b^2 - ac < 0, = 0$, or > 0 respectively. Generally speaking, elliptic equations correspond to equilibrium problems and have smooth solutions usually free of discontinuities and/or waves. For parabolic equations, such as the transient heat conduction equation and the Navier-Stokes equation, the wave speed is infinite; thus a discontinuity present at a point spreads out instantaneously. For hyperbolic equations, such as those governing the dynamic deformations of an elastic body, the wave speed is finite; these problems also admit shock waves.

The steady-state heat equation, the Laplace equation and the equations of elastostatics studied in the previous sections are elliptic. The diffusion equation, $\alpha^2 u_{,xx} = u_{,t}$ where the diffusivity α^2 is a constant and t is the time, is parabolic. The wave equation $c^2 u_{,xx} = u_{,tt}$ where c is a constant and t is the time is hyperbolic. We will first study parabolic equations and then hyperbolic equations in the next section.

10.2 A Model Problem

The transient heat conduction equation is a typical parabolic equation. For an anisotropic and nonhomogeneous body, this equation is

$$\rho c \frac{\partial u}{\partial t} = -q_{i,i} + f \quad \text{in} \quad \Omega \times (0,T), \tag{10.2.1}$$

$$q_i = -\kappa_{ij} u_{,j} \quad \text{in} \quad \Omega \times (0,T), \tag{10.2.2}$$

$$-q_i n_i = h(\mathbf{x},t) \quad \text{on} \quad \Gamma_1 \times (0,T), \tag{10.2.3}$$

$$u(\mathbf{x},t) = \hat{u}(\mathbf{x},t) \quad \text{on} \quad \Gamma_2 \times (0,T), \tag{10.2.4}$$

$$u(\mathbf{x},0) = u_0(\mathbf{x}) \quad \text{in} \quad \Omega. \tag{10.2.5}$$

Here ρ is the mass density, c the specific heat, q_i the heat flux, and f the source of energy per unit volume. In general, ρ and c are functions of $\mathbf{x}$, and f is a function of $\mathbf{x}$ and time t. Equation (10.2.2) is the Fourier law of heat conduction, and is called the constitutive equation for the body, $\boldsymbol{\kappa}$ is the thermal conductivity tensor and for a nonhomogenous body is a function of $\mathbf{x}$. $\boldsymbol{\kappa}$ is assumed to be positive definite. For an isotropic body

$$\kappa_{ij} = \kappa\delta_{ij} \tag{10.2.6}$$

where δ_{ij} is the Kronecker delta. Γ_1 and Γ_2 are complementary parts of the boundary $\partial\Omega$ of Ω, on Γ_1 heat flux is prescribed and at points of Γ_2 the temperature is assigned as a function of $\mathbf{x}$ and t. Since we are studying a transient problem, initial conditions need to be specified, and equation (10.2.5) gives these. Substitution from (10.2.2) into (10.2.1) and (10.2.3) yields

$$\rho c\frac{\partial u}{\partial t} = (\kappa_{ij}u_{,j})_{,i} + f \text{ in } \Omega \times (0,T) \tag{10.2.7}$$

$$(\kappa_{ij}u_{,j})n_i = h \quad \text{on} \quad \Gamma_1 \times (0,T). \tag{10.2.8}$$

It is clear from (10.2.7) that the order of the given differential equation is 2; hence, from section 1, it follows that the boundary condition (10.2.8) is natural and boundary condition (10.2.4) is essential.

In order to classify the given differential equation (10.2.7) into elliptic, parabolic or hyperbolic, we assume that the body is a bar and study heat conduction along the length of the bar. Thus u is a function of x and t only and equation (10.2.7) becomes

$$\rho c\frac{\partial u}{\partial t} = \kappa_{11}\frac{\partial^2 u}{\partial x^2} + \frac{\partial \kappa_{11}}{\partial x}\frac{\partial u}{\partial x} + f. \tag{10.2.9}$$

Comparing it with equation (10.1.1) with y set equal to t we see that

$$a = \kappa_{11},\ b = 0,\ c = 0. \tag{10.2.10}$$

Thus $b^2 - ac = 0$ and equation (10.2.9) is parabolic.

By a classical solution of the initial-boundary-value problem defined by equations (10.2.1) - (10.2.5), we mean the following: Given $f, h, \hat{u}$, and u_0, find $u : \overline{\Omega} \times [0,T] \to \mathbb{R}$ such that equations (10.2.1) through (10.2.5) are satisfied.

Here we assume that we cannot find a classical solution of the given initial-boundary-value problem and proceed to find its approximate solution. The first step is to obtain a semidiscrete formulation of the problem which consists of a set of coupled ordinary differential equations in time and the corresponding initial conditions.

10.3 Semi-discrete Formulation of the Model Problem

Let $\phi : \overline{\Omega} \to \mathbb{R}$ be a smooth function such that $\phi = 0$ on Γ_2. Note that ϕ is a function of $\mathbf{x}$ only. Throughout the remainder of this section, we consider time t to be fixed. Multiplication of both sides of eqn. (10.2.7) by ϕ, and integration of the resulting equation over the domain Ω yields

$$\int_\Omega \rho c \frac{\partial u}{\partial t} \phi d\Omega = \int_\Omega (\kappa_{ij} u_{,j})_{,i} \phi d\Omega + \int_\Omega f \phi d\Omega. \tag{10.3.1}$$

Noting that

$$\begin{aligned} \int_\Omega (\kappa_{ij} u_{,j})_{,i} \phi d\Omega &= \int_\Omega (\kappa_{ij} u_{,j} \phi)_{,i} d\Omega - \int_\Omega \kappa_{ij} u_{,j} \phi_{,i} d\Omega, \\ &= \int_{\partial\Omega} \kappa_{ij} u_{,j} \phi n_i d\Gamma - \int_\Omega \kappa_{ij} u_{,j} \phi_{,i} d\Omega, \\ &= \int_{\Gamma_1} h\phi d\Gamma - \int_\Omega \kappa_{ij} u_{,j} \phi_{,i} d\Omega, \end{aligned} \tag{10.3.2}$$

wherein we have used the divergence theorem, the boundary condition (10.2.3) and $\phi = 0$ on Γ_2, we obtain from (10.3.1) the following.

$$\int_\Omega \rho c \dot{u} \phi d\Omega + \int_\Omega \kappa_{ij} u_{,j} \phi_{,i} d\Omega = \int_{\Gamma_1} h\phi d\Gamma + \int_\Omega f \phi d\Omega. \tag{10.3.3}$$

Here $\dot{u} = \partial u / \partial t$. With the notations

$$(\phi, \rho c \dot{u})_\Omega = \int_\Omega \rho c \dot{u} \phi d\Omega \,, \tag{10.3.4}$$

$$B(\phi, u) = \int_\Omega \kappa_{ij} u_{,j} \phi_{,i} d\Omega \,, \tag{10.3.5}$$

$$(\phi, h)_{\Gamma_1} = \int_{\Gamma_1} \phi h d\Gamma \,, \text{and} \tag{10.3.6}$$

$$(\phi, f)_\Omega = \int_\Omega f \phi d\Omega \,, \tag{10.3.7}$$

we write eqn. (10.3.3) as

$$(\phi, \rho c \dot{u})_\Omega + B(\phi, u) = (\phi, h)_{\Gamma_1} + (\phi, f)_\Omega \,. \tag{10.3.8}$$

We note that $(\cdot,\cdot)_\Omega$, $B(\cdot,\cdot)$ and $(\cdot,\cdot)_{\Gamma_1}$ are bilinear forms, and $B(\cdot,\cdot)$ is positive definite if κ_{ij} is positive definite everywhere in Ω. In order for expressions in (10.3.4) through (10.3.7) to be well defined, the functions ϕ and u should be such that their first-order derivative is square integrable. Let

$$H^1 = \left\{ \psi | \psi : \overline{\Omega} \to \mathbb{R},\ \int_\Omega (\psi_{,i}\psi_{,i} + \psi^2) d\Omega < \infty \right\}, \tag{10.3.9}$$

$$H_0^1 = \left\{ \psi | \psi \in H^1,\ \psi = 0 \text{ on } \Gamma_2 \right\}. \tag{10.3.10}$$

Note that (10.3.10) is not a standard notation.

We now convert the initial conditions (10.2.5) into their weak form. Multiplication of both sides of (26.6) by $\rho c \phi$ and integration over the domain Ω gives

$$(\phi, \rho c u(\mathbf{x}, 0))_\Omega = (\phi, \rho c u_0)_\Omega \,. \tag{10.3.11}$$

Thus a weak formulation of the given initial-boundary-value problem is: find $u : \overline{\Omega} \times (0,T) \to \mathbb{R}$ such that $u \in H^1$, $u = \hat{u}$ on $\Gamma_1 \times (0,T)$ for every value of $t \in (0,T)$, and equations (10.3.8) and (10.3.11) hold for every $\phi \in H_0^1$. Here it is tacitly implied that (26.18) holds for $t > 0$ and (10.3.11) for $t = 0$.

In order to derive the Galerkin formulation of the problem, we select a function $g \in H^1$ which also satisfies the essential boundary conditions. Then, for every $v \in H_0^1$, $u = (v + g) \in H^1$ and $u = \hat{u}$ on $\Gamma_2 \times (0,T)$. Thus the problem of finding a u reduces to that of finding a $v \in H_0^1$; the function $g(x,t)$ is kept fixed. Substitution of $u = (v + g)$ in (10.3.8) and (10.3.11) gives

$$(\phi, \rho c \dot{v})_\Omega + B(\phi, v) = (\phi, h)_{\Gamma_1} + (\phi, f)_\Omega - (\phi, \rho c \dot{g})_\Omega - B(\phi, g) \tag{10.3.12}$$

$$(\phi, \rho c v(\mathbf{x}, 0))_\Omega = (\phi, \rho c u_0)_\Omega - (\phi, \rho c g(\mathbf{x}, 0))_\Omega \,. \tag{10.3.13}$$

The Galerkin formulation of the given problem is: find $v(\cdot, t) \in H_0^1$ such that equations (10.3.12) and (10.3.13) hold for every $\phi \in H_0^1$.

Let H_0^{1n} be a finite-dimensional subset of H_0^1, and $v^n \in H_0^{1n}$, $\phi^n \in H_0^{1n}$. Since v^n and ϕ^n are also in H_0^1, we can replace v and ϕ in (10.3.12) and (10.3.13) by v^n and ϕ^n; this yields the Galerkin approximation of the given initial-boundary-value problem which can be stated as follows. Find

$v^n(\cdot,t) \in H_0^{1n}$ such that

$$(\phi^n, \rho c\dot{v}^n)_\Omega + B(\phi^n, v^n) = (\phi^n, h)_{\Gamma_1} + (\phi^n, f)_\Omega - (\phi^n, \rho c\dot{g})_\Omega - B(\phi^n, g), \tag{10.3.14}$$

$$(\phi^n, \rho c v^n(\mathbf{x}, 0))_\Omega = (\phi^n, \rho c u_0)_\Omega - (\phi^n, \rho c g(\mathbf{x}, 0))_\Omega \text{ for every } \phi^n \in H_0^{1n} . \tag{10.3.15}$$

Let $\psi_1, \psi_2, \ldots \psi_n$ be basis functions in H_0^{1n}; ψ_1, ψ_2 etc. are functions of $\mathbf{x}$ only. Then

$$\phi^n(\mathbf{x}) = \sum_{A=1}^{n} C_A \psi_A(\mathbf{x}), \tag{10.3.16}$$

$$v^n(\mathbf{x}, t) = \sum_{B=1}^{n} d_B(t) \psi_B(\mathbf{x}) . \tag{10.3.17}$$

The expression on the right-hand side of (10.3.17) is analogous to that assumed in the method of separation of variables for solving a partial differential equation. However, $v^n(x,t) \neq f(x)g(t)$. With (10.3.16) and (10.3.17),

$$\begin{aligned}(\phi^n, \rho c\dot{v}^n)_\Omega =& \int_\Omega C_A \psi_A \dot{d}_B \psi_B d\Omega = C_A \left(\int_\Omega \psi_A \psi_B d\Omega \right) \dot{d}_B , \\ =& C_A M_{AB} \dot{d}_B , \end{aligned} \tag{10.3.18}$$

$$\begin{aligned} a(\phi^n, v^n) =& \int_\Omega C_A \psi_{A,i} \kappa_{ij} d_B \psi_{B,j} d\Omega, \\ =& C_A \left(\int_\Omega \psi_{A,i} \kappa_{ij} \psi_{B,j} d\Omega \right) d_B = C_A K_{AB} d_B , \end{aligned} \tag{10.3.19}$$

$$(\phi^n, h)_{\Gamma_1} + (\phi^n, f)_\Omega = C_A \left[\int_{\Gamma_1} \psi_A h d\Gamma + \int_\Omega \psi_A f d\Omega \right] = C_A F_A^1, \tag{10.3.20}$$

$$(\phi^n, \rho c\dot{g})_\Omega - a(\phi^n, g) = C_A \left[\int_\Omega \psi_A \rho c \dot{g} d\Omega + \int_\Omega \psi_{A,i} \kappa_{ij} g_{,j} d\Omega \right] = C_A F_A^2, \tag{10.3.21}$$

$$(\phi^n, \rho c u_0)_\Omega - (\phi^n, \rho c g(\mathbf{x}, 0))_\Omega = C_A \left[\int_\Omega \rho c \left(u_0(\mathbf{x}) - g(\mathbf{x}, 0) \right) \psi_A d\Omega \right] = C_A F_A^0, \tag{10.3.22}$$

where the repeated index A or B implies summation over the range of the index. Thus equations (10.3.14) and (10.3.15) can be written as

$$C_A \left(M_{AB} \dot{d}_B + K_{AB} d_B - F_A^1 - F_A^2 \right) = 0, \tag{10.3.23}$$

$$C_A \left(M_{AB} d_B(0) - F_A^0 \right) = 0. \tag{10.3.24}$$

In order for equations (10.3.14) and (10.3.15) to hold for every $\phi^n \in H_0^{1n}$, equations (10.3.23) and (10.3.24) must hold for every choice of $C_1, C_2, \ldots$, which is the case if and only if the terms in the

parentheses vanish. Thus

$$M_{AB}\dot{d}_B + K_{AB}d_B = F_A^1 + F_A^2, \tag{10.3.25}$$

$$M_{AB}d_B(0) = F_A^0\,. \tag{10.3.26}$$

Equations (10.3.25) and (10.3.26) are a set of coupled ordinary differential equations for $d_1, d_2, \ldots, d_n$ and the corresponding initial conditions. Their solution will determine d's as a function of time which when substituted into (10.3.17) will give an approximate solution of the problem. Since (10.3.25) are differential rather than algebraic equations, the formulation is called semidiscrete.

In the finite element work, the function g introduced in equations (10.3.12) and (10.3.13) is found as follows. The basis functions in H_0^{1n} are taken to be the finite element basis functions. Note that n equals the number of nodes in the finite element discretization of Ω where the essential boundary conditions are not prescribed and hence the solution is to be found. Let nodes $(n+1), (n+2), \ldots m$ lie on Γ_2 and the corresponding finite element basis functions be $\psi_{(n+1)}(\mathbf{x}), \psi_{(n+2)}(\mathbf{x}) \ldots, \psi_m(\mathbf{x})$. The function g is defined as

$$g(\mathbf{x}, t) = \sum_{A=n+1}^{m} \hat{u}(x_A, t)\psi_A(\mathbf{x}) \tag{10.3.27}$$

where x_A stands for the coordinates of node A.

When solving a practical problem in which the number of nodes $(m-n)$ on Γ_2 is very small as compared to n, we ignore the function g and assemble matrices $\mathbf{M}$ and $\mathbf{K}$ for all of the nodes (including those on Γ_2). We consider equations

$$\mathbf{M}\dot{\mathbf{d}} + \mathbf{K}\mathbf{d} = \mathbf{F}^1 \equiv \mathbf{F}, \tag{10.3.28}$$

$$d_A(0) = d_{0_A} \equiv u_0(x_A),\ A = 1, 2, \ldots, m, \tag{10.3.29}$$

and apply the essential boundary conditions by one of the the two methods outlined in section 5 (it will become clear in the next section).

10.4 A Generalized Trapezoidal Algorithm

In order to solve equations (10.3.28) and (10.3.29), we divide the time interval $(0, T)$ into subintervals of not necessarily equal length. Let $t_1 = 0,\ t_2 = t_1 + \Delta t,\ t_3 = t_2 + \Delta t$, etc. be points on $(0, T)$ and

$$\mathbf{v}_{n+1} \simeq \dot{\mathbf{d}}(t_{n+1}),\ \mathbf{d}_{n+1} \simeq \mathbf{d}(t_{n+1}),\ \mathbf{F}_{n+1} = \mathbf{F}(t_{n+1}). \tag{10.4.1}$$

Here $\mathbf{v}_{n+1}$ and $\mathbf{d}_{n+1}$ are approximations to the values of $\dot{\mathbf{d}}$ and $\mathbf{d}$ at time t_{n+1}. The generalized trapezoidal algorithm that encompasses the forward-difference method, the central-difference method and the backward-difference method can be stated as follows.

$$\mathbf{M}\mathbf{v}_{n+1} + \mathbf{K}\mathbf{d}_{n+1} = \mathbf{F}_{n+1}, \tag{10.4.2}$$

$$\mathbf{d}_{n+1} = \mathbf{d}_n + \Delta t \mathbf{v}_{n+\alpha}, \tag{10.4.3}$$

$$\mathbf{v}_{n+\alpha} = (1-\alpha)\mathbf{v}_n + \alpha \mathbf{v}_{n+1}, \tag{10.4.4}$$

where $\alpha \in [0,1]$. The objective is to find $\mathbf{v}_{n+1}$ and $\mathbf{d}_{n+1}$ from a knowledge of $\mathbf{v}_n$ and $\mathbf{d}_n$ or said differently, from equations (10.4.2) - (10.4.4) a recursive relation among $\mathbf{v}_{n+1}, \mathbf{d}_{n+1}, \mathbf{v}_n$ and $\mathbf{d}_n$ can be derived. The trapezoidal algorithm reduces to the forward-difference method if $\alpha = 0$, to the central-difference method or the Crank-Nicolson method if $\alpha = 1/2$, and to the backward-difference method if $\alpha = 1$. The algorithm can be implemented in one of the following two ways.

(a) $\mathbf{v}_{n+1}$ - implementation.

Substitution from (10.4.4) into (10.4.3) yields

$$\mathbf{d}_{n+1} = \tilde{\mathbf{d}}_{n+1} + \alpha \Delta t \mathbf{v}_{n+1}, \tag{10.4.5}$$

$$\tilde{\mathbf{d}}_{n+1} = \mathbf{d}_n + \Delta t(1-\alpha)\mathbf{v}_n. \tag{10.4.6}$$

Note that $\tilde{\mathbf{d}}_{n+1}$ can be computed from $\mathbf{d}_n$ and $\mathbf{v}_n$ which are presumed to be known for the time being. From (10.4.5) and (10.4.2) we obtain

$$\mathbf{K}^{eff}\mathbf{v}_{n+1} = \mathbf{F}^{eff}_{n+1}, \tag{10.4.7}$$

where

$$\mathbf{K}^{eff} = \mathbf{M} + \alpha \Delta t \mathbf{K}, \tag{10.4.8}$$

$$\mathbf{F}^{eff}_{n+1} = \mathbf{F}_{n+1} - \mathbf{K}\tilde{\mathbf{d}}_{n+1}\ . \tag{10.4.9}$$

We modify equations (10.4.7) to satisfy the essential boundary conditions by using one of the two methods outlined in section 5, and solve (10.4.7) for $\mathbf{v}_{n+1}$. Recalling that $\mathbf{d}_0$ can be evaluated from the initial data $\mathbf{u}_0(\mathbf{x}_A)$, we solve

$$\mathbf{M}\mathbf{v}_0 + \mathbf{K}\mathbf{d}_0 = \mathbf{F}_0 \tag{10.4.10}$$

for $\mathbf{v}_0$. Knowing $\mathbf{v}_0$, $\mathbf{d}_0$, we can compute $\tilde{\mathbf{d}}_1$ from (10.4.6), $\mathbf{F}_1^{eff}$ and $\mathbf{K}^{eff}$ from (10.4.9) and (10.4.8) respectively, and $\mathbf{v}_1$ from (10.4.7) and then $\mathbf{d}_1$ from (10.4.5). Thus the approximate solution $\mathbf{v}_1, \mathbf{d}_1$ at time t_1 has been determined and one can march forward in time.

Since the determination of $\mathbf{v}_{n+1}, \mathbf{d}_{n+1}$ requires a knowledge of the solution at one previous time step, the generalized trapezoidal algorithm is a one-step method. In a multistep method, values at more than one previous time step are needed to march forward the solution in time; a multistep method may allow a larger time step size than that allowed by a single-step method.

(b) $\mathbf{d}_{n+1}$ - implementation

Substitution for $\mathbf{v}_{n+1}$ from (10.4.5) into (10.4.2) yields

$$\left(\frac{1}{\alpha\Delta t}\mathbf{M} + \mathbf{K}\right)\mathbf{d}_{n+1} = \mathbf{F}_{n+1} + \frac{1}{\alpha\Delta t}\mathbf{M}\tilde{\mathbf{d}}_{n+1}\ , \tag{10.4.11}$$

which can be solved for $\mathbf{d}_{n+1}$. Knowing $\mathbf{d}_{n+1},\ \mathbf{v}_{n+1}$ is computed from (10.4.5). In (10.4.11) it is tacitly assumed that $\alpha > 0$. For a diagonal mass matrix, the evaluation of the second term on the right-hand side of (10.4.11) requires less computational effort as compared to that needed to evaluate the corresponding term in (10.4.9). It is clear from (10.4.11) that as $\alpha\Delta t \rightarrow \infty$, the solution of the problem approaches that of the stationary or the equilibrium problem. Thus one way to solve an equilibrium problem is to solve the corresponding transient problem by choosing large values of $\alpha\Delta t$; of course the algorithm employed should allow this for the computation of a stable solution.

We notice from equations (1.4.7) and (10.4.8) that for $\alpha = 0,\ \mathbf{K}^{\text{eff}} = \mathbf{M}$. For a diagonal or a lumped mass matrix, equations (10.4.7) are uncoupled and there is no need to solve a system of simultaneous equations. An alogrithm in which the solution at time t_{n+1} can be found from a knowledge of variables at time t_n without solving a system of simultaneous equations is called *explicit*; otherwise it is called *impicit*. Note that for $\alpha > 0$, the generalized trapezoidal alogorithm is implicit since $\mathbf{K}$ is not a diagonal matrix and therefore $\mathbf{K}^{\text{eff}}$ is not a diagonal matrix even if $\mathbf{M}$ is a diagonal matrix.

10.5 Stability of the Generalized Trapezoidal Algorithm

An algorithm is said to be stable if $\mathbf{d}_n \simeq \mathbf{d}(t_n)$ stays bounded for all times. In order to discuss the stability of the generalized trapezoidal algorithm, we first uncouple equations (10.3.28); an easy way to accomplish this is to use eigenvectors as base vectors. Recall that we are now studying the problem in the time-domain so these basis functions are quite different from the finite element basis functions employed earlier in the space domain.

Let $(\lambda_{(i)}, \boldsymbol{\psi}_{(i)})$ be an eigen-solution of the eigenvalue problem

$$(\mathbf{K} - \lambda \mathbf{M})\boldsymbol{\psi} = \mathbf{0} \; . \tag{10.5.1}$$

Then, as discussed in Section 9.1,

$$\boldsymbol{\psi}_{(m)}^T \mathbf{M} \boldsymbol{\psi}_{(l)} = \delta_{ml} \; , \tag{10.5.2}$$

$$\mathbf{K}\boldsymbol{\psi}_{(l)} = \lambda_{(l)} \mathbf{M} \boldsymbol{\psi}_{(l)}, \tag{10.5.3}$$

and if $\mathbf{K}$ and $\mathbf{M}$ are positive definite, then

$$0 < \lambda_{(1)} \leq \lambda_{(2)} \ldots \leq \lambda_{(m)} \; . \tag{10.5.4}$$

With $\boldsymbol{\psi}_{(1)}, \boldsymbol{\psi}_{(2)}, \ldots, \boldsymbol{\psi}_{(m)}$ taken as the base vectors, we can write

$$\mathbf{d}(t) = \sum_{i=1}^{m} d_{(i)}(t) \boldsymbol{\psi}_{(i)} \; , \tag{10.5.5}$$

$$\mathbf{v}(t) = \sum_{i=1}^{m} v_{(i)}(t) \boldsymbol{\psi}_{(i)} \; , \tag{10.5.6}$$

$$\mathbf{F}(t) = \sum_{i=1}^{m} f_{(i)}(t) \mathbf{M} \boldsymbol{\psi}_i(i) \; . \tag{10.5.7}$$

Thus $d_{(i)}(t)$ can be regarded as the component of $\mathbf{d}$ along the base vector $\boldsymbol{\psi}_{(i)}$; some times it is called the Fourier component. Substitution from (10.5.5)-(10.5.7) into (10.3.28), premultiplication of the resulting equation by $\boldsymbol{\psi}_{(j)}^T$ and the use of (10.5.2) and (10.5.3) yields

$$v_{(i)} + \lambda_{(i)} d_{(i)} = f_{(i)}, \tag{10.5.8}$$

which is an equation for $d_{(i)}$ only. Similarly, one can derive the initial conditions for $d_{(i)}$. Henceforth, we will consider equation (10.5.8) and therefore drop the subscript i. Also, in order to study

the stability of the algorithm, we will study the free problem and thus set the right-hand side of (10.5.8) equal to zero.

The algorithm (10.4.2) - (10.4.4) when applied to equation (10.5.8) with $f = 0$ gives

$$v_{n+1} + \lambda d_{n+1} = 0, \tag{10.5.9}$$

$$d_{n+1} = d_n + \Delta t[(1-\alpha)v_n + \alpha v_{n+1}]. \tag{10.5.10}$$

Substituting $v_{n+1} = -\lambda d_{n+1}$, and $v_n = -\lambda d_n$ into (10.5.10), we obtain

$$d_{n+1} = A d_n, \tag{10.5.11}$$

where

$$A = \frac{1 - \Delta t\lambda(1-\alpha)}{1 + \alpha\lambda\Delta t}, \tag{10.5.12}$$

is called the amplification factor. The necessary and sufficient condition for the solution to stay bounded for all times is that $|A| \leq 1$. Thus the generalized trapezoidal algorithm will be stable if and only if

$$-1 \overset{(i)}{\leq} \frac{1 - \Delta t(1-\alpha)\lambda}{1 + \alpha\Delta t\lambda} \overset{(ii)}{\leq} 1\ . \tag{10.5.13}$$

Inequality (i) is equivalent to

$$\Delta t\lambda(1 - 2\alpha) < 2 \tag{10.5.14}$$

and inequality (ii) is always satisfied. We note that for $\alpha \geq 1/2$, (10.5.14) holds irrespective of the values of Δt and λ. Thus the mid-point or the central difference or the Crank-Nicolson method for which $\alpha = 1/2$, and the backward-difference method for which $\alpha = 1$ are stable for all values of Δt; the accuracy of the algorithm does depend upon Δt. For $\alpha < 1/2$, inequality (10.5.14) holds only if

$$\Delta t < \frac{2}{\lambda(1-2\alpha)}, \tag{10.5.15}$$

and in particular for the forward-difference method (i.e. $\alpha = 0$)

$$\Delta t < \frac{2}{\lambda}\ . \tag{10.5.16}$$

Note that (10.5.16) should be satisfied for all values of λ. If (10.5.16) does not hold for any one of the m-values $\lambda_1, \lambda_2, \ldots, \lambda_m$ then the solution corresponding to modes for which (10.5.16) is violated will blow up in time and the solution will become unbounded. Thus

$$\Delta t < \frac{2}{\lambda_{\max}} \tag{10.5.17}$$

for the forward-difference method to be stable. Generally speaking, the value of $\lambda_{\max}$ increases with the refinement of the mesh in the spatial domain, and for the heat conduction problem, is inversely proportional to the square of the size of the smallest element. Thus if the size of the smallest element is halved, then Δt will need to be decreased by a factor of 4 in order to compute a stable solution with the forward difference method.

An algorithm which is stable for all values of Δt is called *unconditionally stable*, and the one which is stable for restricted values of Δt is called *conditionally stable*. Thus the forward difference method is conditionally stable, and the central difference and the backward difference methods are unconditionally stable. Generalized trapezoidal algorithms with $\alpha \geq \frac{1}{2}$ are unconditionally stable and those with $\alpha < \frac{1}{2}$ are conditionally stable.

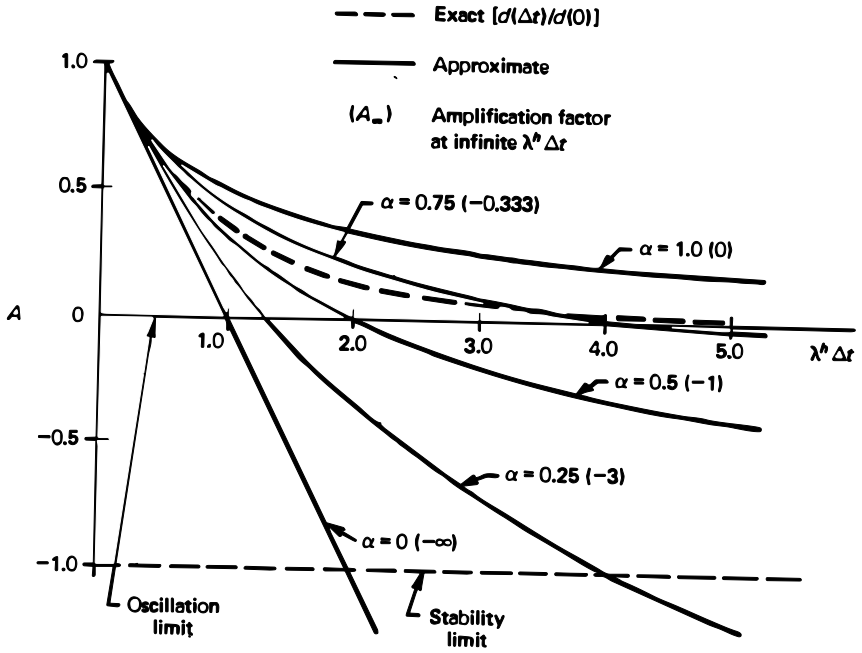


Fig. 10.5.1 Amplification factor for typical one-step methods.

Figure 10.5.1 depicts the plot of the amplification factor, A, vs $\lambda\Delta t$ for several values of α. The value of $\lambda\Delta t$ for which $A = 0$ is called the oscillation limit since for greater values of $\lambda\Delta t$, $A < 0$ and the sign of A^n changes from one time step to the next. For the unconditionally stable algorithms, i.e., those with $\alpha \geq \frac{1}{2}$, the asymptotic value of the amplification factor satisfies $|A_\infty| \leq 1$. It follows from eqn. (10.5.12) that $\lim_{\lambda\Delta t \to \infty} A = \frac{\alpha - 1}{\alpha}$. Thus for $\alpha \geq \frac{1}{2}$, $|A| < 1$ and all spurious high modal components decay. However, for $\alpha = \frac{1}{2}$ and $\lambda\Delta t \gg 1,\ A = -1$. Thus high modal components behave like $(-1)^n$ and the solution oscillates in a sawtooth pattern. These spurious

high frequency modes may be filtered out by averaging the solution over two consecutive time steps.

10.6 Convergence

The algorithm (10.4.2)-(10.4.4) when applied to equation (10.5.8) yields

$$d_{n+1} - Ad_n - L_n = 0 \tag{10.6.1}$$

where $L_n = \Delta t f_{n+\alpha}/(1+\alpha\Delta t\lambda)$. When d_n and d_{n+1} in (10.6.1) are replaced by their exact expressions, then we obtain

$$d(t_{n+1}) - Ad(t_n) - L_n = \Delta t\tau(t_n), \tag{10.6.2}$$

where $\tau(t_n)$ is called the truncation error. If $|\tau(t)| \leq c\Delta t^k$ for all $t \in [0,T]$ where c is a constant independent of Δt, and $k > 0$, the algorithm defined by (10.4.2)-(10.4.4) is called *consistent*, and k is called the *rate of convergence* or *the order of accuracy*.

The following proposition shows that the algorithm (10.4.2)-(10.4.4) is consistent.

Proposition: The generalized trapezoidal methods are consistent, and furthermore $k = 1$ for all $\alpha \in [0,1]$, except for $\alpha = 1/2$, in which case $k = 2$.

Proof:

Expand $d(t_{n+1})$ and $d(t_n)$ about $t_{n+\alpha}$ in finite Taylor series.

$$\begin{aligned} d(t_{n+1}) &= d(t_{n+\alpha}) + (1-\alpha)\Delta t\dot{d}(t_{n+\alpha}) \\ &\quad + \frac{(1-\alpha)^2\Delta t^2}{2!}\ddot{d}(t_{n+\alpha}) + \frac{(1-\alpha)^3\Delta t^3}{3!}\dddot{d}(t_{n+\alpha}) + \bigcirc(\Delta t^4), \end{aligned} \tag{10.6.3}$$

$$\begin{aligned} d(t_n) &= d(t_{n+\alpha}) + (-\alpha\Delta t)\dot{d}(t_{n+\alpha}) \\ &\quad + \frac{(-\alpha\Delta t)^2}{2!}\ddot{d}(t_{n+\alpha}) + \frac{(-\alpha\Delta t)^3}{3!}\dddot{d}(t_{n+\alpha}) + \bigcirc(\Delta t^4). \end{aligned} \tag{10.6.4}$$

Substituting for A from (10.5.12) into (10.6.1), and for $d(t_{n+1})$ and $d(t_n)$ from (10.6.3) and (10.6.4) into the resulting equation, we obtain

$$\begin{aligned} \Delta t(1+\alpha\Delta t\lambda)\tau(t_n) &= (1+\alpha\Delta t\lambda)d(t_{n+1}) - [1-(1-\alpha)\lambda\Delta t]d(t_n) - \Delta t f_{n+\alpha}, \\ &= [(1+\alpha\Delta t\lambda) - (1-(1-\alpha)\lambda\Delta t]d(t_{n+\alpha}) \\ &\quad + [(1+\alpha\Delta t\lambda)(1-\alpha)\Delta t - (1-(1-\alpha)\lambda\Delta t)(-\alpha\Delta t)]\dot{d}(t_{n+\alpha}) \\ &\quad + \left[\frac{(1-\alpha)^2\Delta t^2}{2!}(1+\alpha\Delta t\lambda) - (1-(1-\alpha)\lambda\Delta t)\frac{(-\alpha\Delta t)^2}{2!}\right]\ddot{d}(t_{n+\alpha}) \end{aligned}$$

$$
\begin{aligned}
&+\left[\frac{(1-\alpha)^3\Delta t^3}{3!}(1+\alpha\Delta t\lambda)-(1-(1-\alpha)\lambda\Delta t)\frac{(-\alpha\Delta t)^3}{3!}\right]\dddot{d}(t_{n+\alpha})\\
&-\Delta t\left[(1-\alpha)\left(f(t_{n+\alpha})+(-\alpha\Delta t)\dot{f}(t_{n+\alpha})+\frac{(-\alpha\Delta t)^2}{2!}\ddot{f}(t_{n+\alpha})\right.\right.\\
&\left.+\frac{(-\alpha\Delta t)^3}{3!}\dddot{f}(t_{n+\alpha})\right)+\alpha\left(f(t_{n+\alpha})+(1-\alpha)\Delta t\dot{f}(t_{n+\alpha})\right.\\
&\left.\left.+\frac{(1-\alpha)^2}{2!}\Delta t^2\ddot{f}(t_{n+\alpha})+\frac{(1-\alpha)^3\Delta t^3}{3!}\dddot{f}(t_{n+\alpha})\right]+\bigcirc(\Delta t^4),\right.\\
=\;&[1+\alpha\Delta t\lambda-1+\lambda\Delta t-\alpha\Delta t\lambda]d(t_{n+\alpha})\\
&+[\Delta t+\alpha\Delta t^2\lambda-\alpha^2\Delta t^2\lambda-\alpha\lambda\Delta t^2+\alpha^2\lambda\Delta t^2]\dot{d}(t_{n+\alpha})\\
&+\frac{\Delta t^2}{2}[(1+\alpha^2-2\alpha)(1+\alpha\Delta t\lambda)-(1-\lambda\Delta t+\alpha\lambda\Delta t)\alpha^2]\ddot{d}(t_{n+\alpha})\\
&+\frac{\Delta t^3}{3!}[(1-\alpha^3-3\alpha+3\alpha^2)(1+\alpha\Delta t\lambda)+\alpha^3(1-\lambda\Delta t+\alpha\lambda\Delta t)]\dddot{d}(t_{n+\alpha})\\
&-\Delta t\left[(1-\alpha+\alpha)f(t_{n+\alpha})+(-\alpha\Delta t+\alpha\Delta t-\alpha^2\Delta t)\dot{f}(t_{n+\alpha})\right.\\
&\left.+\frac{\Delta t^2}{2}(\alpha^2-\alpha^3+\alpha+\alpha^3-2\alpha^2)\ddot{f}(t_{n+\alpha})+\frac{\Delta t^3}{3!}(-\alpha^3+\alpha^4+\alpha-\alpha^4-3\alpha^2+3\alpha^3)\dddot{f}(t_{n+\alpha})\right]\\
&+\bigcirc(\Delta t^4),\\
=\;&\lambda\Delta t d(t_{n+\alpha})+\Delta t\dot{d}(t_{n+\alpha})\\
&+\frac{\Delta t^2}{2!}[1-2\alpha+\alpha\Delta t\lambda+\alpha^3\Delta t\lambda-2\alpha^2\Delta t\lambda+\alpha^2\lambda\Delta t-\alpha^3\lambda\Delta t]\ddot{d}(t_{n+\alpha})\\
&+\frac{\Delta t^3}{3!}[1-\alpha^3-3\alpha+3\alpha^2+\alpha\Delta t\lambda-\alpha^4\Delta t\lambda-3\alpha^2\Delta t\lambda+3\alpha^3\Delta t\lambda\\
&+\alpha^3-\alpha^3\lambda\Delta t+\alpha^4\lambda\Delta t]\dddot{d}(t_{n+\alpha})]\\
&-\Delta t[f(t_{n+\alpha})+\frac{\Delta t^2}{2!}\alpha(1-\alpha)\ddot{f}(t_{n+a})\\
&+\frac{\Delta t^3}{3!}(2\alpha^3-3\alpha^2+\alpha)\dddot{f}(t_{n+\alpha})]+\bigcirc(\Delta t^4),\\
=\;&\Delta t[\dot{d}(t_{n+\alpha})+\lambda d(t_{n+\alpha})-f(t_{n+\alpha})]\\
&+\frac{\Delta t^2}{2!}[1-2\alpha+\alpha\lambda\Delta t-\alpha^2\lambda\Delta t]\ddot{d}(t_{n+\alpha})\\
&+\frac{\Delta t^3}{3!}[1-3\alpha+2\alpha^3\Delta t\lambda+\alpha\Delta t\lambda+3\alpha^2-3\alpha^2\Delta t\lambda]\dddot{d}(t_{n+\alpha})\\
&-\frac{\Delta t^3}{2!}\alpha(1-\alpha)\ddot{f}(t_{n+\alpha})\\
&-\frac{\Delta t^4}{6}(2\alpha^3-3\alpha^2+\alpha)\dddot{f}(t_{n+\alpha})+\bigcirc(\Delta t^4).
\end{aligned}
$$

Noting that $\dot{d}(t_{n+\alpha}) + \lambda d(t_{n+\alpha}) - f(t_{n+\alpha}) = 0$, and dividing throughout by Δt, we obtain

$$\begin{aligned}(1+\alpha\Delta t\lambda)\tau(t_n) &= \frac{\Delta t}{2}[1-2\alpha+\alpha\lambda\Delta t-\alpha^2\lambda\Delta t]\ddot{d}(t_{n+\alpha})\\ &+\frac{\Delta t^2}{6}[1-3\alpha+\alpha\lambda\Delta t+2\alpha^3\Delta t\lambda+3\alpha^2-3\alpha^2\Delta t\lambda]\dddot{d}(t_{n+\alpha})\\ &-\frac{\Delta t^2}{2}\alpha(1-\alpha)\ddot{f}(t_{n+\alpha})\\ &-\frac{\Delta t^3}{6}(2\alpha^3-3\alpha^2+\alpha)\dddot{f}(t_{n+\alpha})+\bigcirc(\Delta t^3),\\ \tau(t_n) &= (1-2\alpha)\bigcirc(\Delta t)+\bigcirc(\Delta t^2). \end{aligned} \tag{10.6.5}$$

Thus the central-difference method $(\alpha = 1/2)$ *or the midpoint rule is the only member of the family of generalized trapezoidal methods that is 2nd-order accurate.*

The following Theorem proves that a stable and a consistent algorithm is convergent.

Theorem: Let the trapezoidal algorithm be stable and consistent. Then, for a fixed value of t_{n+1}, the error

$$e_{n+1} = d(t_{n+1}) - d_{n+1} \tag{10.6.6}$$

goes to zero as $\Delta t \to o$.

Proof: Substracting (10.6.1) from (10.6.2) we get

$$e_{n+1} = Ae_n + \Delta t\tau(t_n). \tag{10.6.7}$$

Replacing the subscript n by $(n-1)$ in (10.6.7) we obtain e_n. Substitution for e_n in (10.6.7) yields

$$e_{n+1} = A^2e_{n-1} + \Delta t[A\tau(t_{n-1}) + \tau(t_n)].$$

Repeating the same procedure $(n-1)$ times, we arrive at

$$e_{n+1} = A^{n+1}e_0 + \Delta t\sum_{i=0}^{n} A^i\tau(t_{n-i}). \tag{10.6.8}$$

Note that $e_0 = 0$ since the approximate and the analytical solutions satisfy the same initial condition. Therefore, equation (10.6.8) simplifies to

$$e_{n+1} = \Delta t\sum_{i=0}^{n} A^i\tau(t_{n-i}). \tag{10.6.9}$$

Taking the absolute values of both sides and using the triangle and the arithmetic-geometric mean inequalities, we obtain

$$|e_{n+1}| \le \Delta t\sum_{i=0}^{n} |A|^i|\tau(t_{n-i})|. \tag{10.6.10}$$

The stability of the algorithm implies that $|A| \leq 1$ and its consistency gives $|\tau(t_{n-i})| = c\Delta t^k$. Therefore, we conclude from (10.6.10) that

$$|e_{n+1}| \leq \Delta t(n+1)c\Delta t^k = t_{n+1}c\Delta t^k. \tag{10.6.11}$$

Thus $e_{n+1} \to 0$ as $\Delta t \to 0$ which proves the result.

10.7 Method of Weighted Residuals

Following the work of Zienkiewicz,[4] we study a systematic way of deriving the generalized trapezoidal algorithm. Consider the equation

$$\begin{aligned} &\dot{a} + \lambda a = f,\ 0 < t < T, \\ &a(0) = a_0. \end{aligned} \tag{10.7.1}$$

Let $W : [t_n,\ t_{n+1}] \to \mathrm{R}$ be a smooth function. For the initial-value problem, the weight function is not required to satisfy homogeneous initial conditions. Multiply both sides of eqn. (10.7.1) by W and integrate the resulting eqn. over $[t_n,\ t_{n+1}]$ to obtain

$$\int_{t_n}^{t_{n+1}} W(\dot{a} + \lambda a)dt = \int_{t_n}^{t_{n+1}} W f\, dt. \tag{10.7.2}$$

Let

$$\xi = \frac{t - (t_{n+1} + t_n)/2}{(t_{n+1} - t_n)/2}. \tag{10.7.3}$$

Then $\xi \in [-1,\ 1]$ for $t \in [t_n,\ t_{n+1}]$. Define shape functions

$$N_1(\xi) = \frac{1-\xi}{2},\ N_2(\xi) = \frac{1+\xi}{2}, \tag{10.7.4}$$

and on the interval $[t_n,\ t_{n+1}]$, write

$$a(t(\xi)) = a_n N_1(\xi) + a_{n+1} N_2(\xi) \tag{10.7.5}$$

where

$$a_n \simeq a(t_n),\ a_{n+1} \simeq a(t_{n+1}).$$

[4] O.C. Zienkiewicz, A new look at Newmark, Houbolt and other time stepping formulae. A weighted residual approach. Univ. of Wales, Swansea Civil Eng'g Report C/R/273/76.
A more readily available reference is W. L. Wood, A further look at Newmark, Houbolt, etc. time-stepping formulae, *Int. J. Num. Meth. Engng.*, 20, 1009-1017, 1984.

Then

$$\begin{aligned}\dot{a} &= \frac{da}{d\xi}\cdot\frac{d\xi}{dt} = \left[a_n\left(-\frac{1}{2}\right)+a_{n+1}\left(\frac{1}{2}\right)\right]\frac{2}{\Delta t};\ \Delta t = t_{n+1}-t_n.\\ &= \frac{a_{n+1}-a_n}{\Delta t}.\end{aligned} \tag{10.7.6}$$

Substitution from (10.7.5) and (10.7.6) into (10.7.2) yields

$$\int_{-1}^{1}\left\{W\left(\frac{a_{n+1}-a_n}{\Delta t}\right)+\lambda\left[a_n\frac{(1-\xi)}{2}+a_{n+1}\frac{(1+\xi)}{2}\right]\right\}\frac{\Delta t}{2}d\xi = \int_{-1}^{1} Wf\frac{\Delta t}{2}d\xi. \tag{10.7.7}$$

Assume that $\displaystyle\int_{-1}^{1} W d\xi \neq 0$, and let

$$\begin{aligned}\alpha &\equiv \int_{-1}^{1} W\xi d\xi \Big/ \int_{-1}^{1} W d\xi,\\ \tilde{f} &= \Delta t\int_{-1}^{1} Wf d\xi \Big/ \int_{-1}^{1} W d\xi.\end{aligned} \tag{10.7.8}$$

Dividing (10.7.7) by $\displaystyle\int_{-1}^{1} W d\xi$ and using (10.7.8), we obtain

$$\begin{aligned}&(a_{n+1}-a_n)+\frac{\lambda\Delta t}{2}[a_n(1-\alpha)+a_{n+1}(1+\alpha)] = \tilde{f},\\ &a_{n+1}\left(1+\frac{\lambda\Delta t}{2}(1+\alpha)\right)+a_n\left(-1+(1-\alpha)\frac{\lambda\Delta t}{2}\right) = \tilde{f}.\end{aligned} \tag{10.7.9}$$

Thus the amplification factor, A, is given by

$$A = -\left[\frac{-1+(1-\alpha)\dfrac{\lambda\Delta t}{2}}{1+(1+\alpha)\dfrac{\lambda\Delta t}{2}}\right]. \tag{10.7.10}$$

For stability,

$$-1\overset{(i)}{\leq}A\overset{(ii)}{\leq}1.$$

$$\begin{aligned}\text{(i)}\quad &-1+(1-\alpha)\frac{\lambda\Delta t}{2}\leq 1+(1+\alpha)\frac{\lambda\Delta t}{2},\\ &-\alpha\frac{\lambda\Delta t}{2}-\alpha\frac{\lambda\Delta t}{2}\leq 2\\ \text{or}\quad &-\alpha\lambda\Delta t\leq 2.\\ \text{(ii)}\quad &1-(1-\alpha)\frac{\lambda\Delta t}{2}\leq 1+(1+\alpha)\frac{\lambda\Delta t}{2},\\ &-\frac{\lambda\Delta t}{2}(2)\leq 0.\end{aligned}$$

Since $\lambda \geq 0$, (ii) is always satisfied. For $\alpha > 0$, (i) is also always satisfied. However, if $\alpha < 0$, then (i) requires that $\Delta t \leq 2/(-\alpha\lambda)$. Thus the algorithm is unconditionally stable for $\alpha \geq 0$, and conditionally stable for $\alpha < 0$. Note that α is well defined only when $\int_{-1}^{1} W d\xi \neq 0$. Hence W cannot be an odd function of ξ. We consider three choices for W.

$$
\begin{aligned}
&\text{(a)} \quad W = N_1(\xi) = (1-\xi)/2, \\
&\qquad \alpha = \int_{-1}^{1} \xi(1-\xi)/2 d\xi \Big/ \int_{-1}^{1} (1-\xi)/2 d\xi, \\
&\qquad = \frac{\left(\frac{\xi^2}{2} - \frac{\xi^3}{3}\right)\Big|_{-1}^{1}}{\left(\xi - \frac{\xi^2}{2}\right)\Big|_{-1}^{1}} = \frac{-\frac{2}{3}}{+2} = -\frac{1}{3}; \\
&\text{(b)} \quad W = N_2(\xi) = (1+\xi)/2, \\
&\qquad \alpha = \frac{\left(\frac{\xi^2}{2} + \frac{\xi^3}{3}\right)\Big|_{-1}^{1}}{\left(\xi + \frac{\xi^2}{2}\right)\Big|_{-1}^{1}} = \frac{\frac{2}{3}}{2} = \frac{1}{3}; \\
&\text{(c)} \quad W = 1, \\
&\qquad \alpha = \frac{\left(\frac{\xi^2}{2}\right)\Big|_{-1}^{1}}{(\xi)\Big|_{-1}^{1}} = 0.
\end{aligned}
$$

The algorithm (10.7.9) is a one-step algorithm since the value of a at only one previous time step is needed to compute its present value. We could similarly consider quadratic and cubic shape functions over the interval $[t_n,\ t_{n+1}]$ and derive multistep algorithms.

In order to derive a two-step alogrithm, we select a test function defined on $[t_n,\ t_{n+2}]$ and instead of (10.7.2) we have

$$
\int_{t_n}^{t_{n+2}} W(\dot{a} + \lambda a) dt = \int_{t_n}^{t_{n+2}} W f dt. \tag{10.7.11}
$$

Let

$$
\xi = \frac{t - (t_{n+2} + t_n)/2}{(t_{n+2} - t_n)/2}, \tag{10.7.12}
$$

and

$$
N_1 = \xi(\xi - 1)/2,\ N_2 = 1 - \xi^2,\ N_3 = \xi(\xi + 1)/2 \tag{10.7.13}
$$

be quadratic shape functions defined on $[-1,\ 1]$. On the interval $[t_n,\ t_{n+2}]$ we write

$$a(t(\xi)) = a_n N_1(\xi) + a_{n+1} N_2(\xi) + a_{n+2} N_3(\xi). \tag{10.7.14}$$

Thus

$$\dot{a} = [a_n(2\xi - 1) + a_{n+1}(-4\xi) + a_{n+2}(2\xi + 1)]/(t_{n+2} - t_n). \tag{10.7.15}$$

Substitution from (10.7.14) and (10.7.15) into (10.7.11) yields

$$\int_{-1}^{1} W[(-a_n + a_{n+2} + 2\xi(a_n - 2a_{n+1} + a_{n+2}))/2\Delta t + \lambda(\xi^2(a_n - 2a_{n+1} + a_{n+2})/2 - \xi(a_n - a_{n+2})/2 + a_{n+1})]d\xi = \int_{-1}^{1} W f d\xi. \tag{10.7.16}$$

With the assumption that $\int_{-1}^{1} W d\xi \neq 0$, and the definitions

$$\alpha = \int_{-1}^{1} W\xi d\xi / \int_{-1}^{1} W d\xi,\ \beta \equiv \int_{-1}^{1} W\xi^2 d\xi / \int_{-1}^{1} W d\xi, \tag{10.7.17}$$

we rewrite (10.7.16) as

$$-a_n + a_{n+2} + 2\alpha(a_n - 2a_{n+1} + a_{n+2}) + 2\lambda\Delta t[\beta(a_n - 2a_{n+1} + a_{n+2})/2 - \alpha(a_n - a_{n+2})/2 + a_{n+1}] = f. \tag{10.7.18}$$

Note that equation (10.7.18) relating a_n, a_{n+1} and a_{n+2} is a two-step method. For a free body, $f = 0$. With

$$a_{n+1} = Aa_n,\ a_{n+2} = A^2 a_n, \tag{10.7.19}$$

the amplification factor A is given by

$$A^2((1 + 2\alpha) + \lambda\Delta t(\alpha + \beta)) + 2A(-2\alpha + \lambda\Delta t(1 - \beta)) - (1 - 2\alpha + \lambda\Delta t(\alpha - \beta)) = 0, \tag{10.7.20}$$

and depends upon $\lambda\Delta t$, α and β. We first examine values of A for two limiting values of $\lambda\Delta t$. For $\lambda\Delta t \ll 1$, i.e., $\lambda\Delta t \to 0$, equation (10.7.20) reduces to

$$A^2(1 + 2\alpha) - 4A\alpha - (1 - 2\alpha) = 0, \tag{10.7.21}$$

and

$$A = \frac{2\alpha \pm 1}{2\alpha_1}. \tag{10.7.22}$$

the condition for the algorithm stability is

$$\alpha \geq 0. \tag{10.7.23}$$

For $\lambda\Delta t \gg 1$, equation (10.7.20) can be approximated as

$$A^2(\alpha_\beta) + 2A(1-\beta) - (\alpha - \beta) = 0 \tag{10.7.24}$$

and the stability condition can be written as

$$1 - 2\alpha - 3\beta \leq \pm\sqrt{1 - 2\beta + \alpha^2} \leq 1 + 2\alpha + \beta. \tag{10.7.25}$$

For $W = N_1, N_2, N_3$, values of α and β found from equation (10.7.17) are listed below.

W	α	β
N_1	-1	$3/5$
N_2	0	$1/5$
N_3	1	$3/5$

For $W = N_1$, inequality (10.7.23) is violated, and for $W = N_2$, inequality $(10.7.25)_1$ is not satisfied. the choice $W = N_3$ satisfies inequalities (10.7.23) and (10.7.25) and gives an unconditionally stable alagorithm. For this case equation (10.7.20) becomes

$$A^2(15 + 8\lambda\Delta t) + 4A(-5 + \lambda\Delta t) + (5 - 2\lambda\Delta t) = 0, \tag{10.7.26}$$

and has roots

$$A = \frac{2(5 - \lambda\Delta 5) \pm \sqrt{25 - 5\lambda\Delta t + 40(\lambda\Delta t)^2}}{15 + 8\lambda\Delta t}. \tag{10.7.27}$$

10.8 Modal Analysis

An alternative to the time integration schemes described above is to use eigenvectors as base vectors, and express the time dependent solution as

$$d_A(t) = \sum_\ell a_{(\ell)}(t)\psi_{(\ell)A}. \tag{10.8.1}$$

Here $\boldsymbol{\psi}_{(1)}$, $\boldsymbol{\psi}_{(2)}, \ldots$ are eigenvectors, and $a_{(\ell)}(t)$ may be regarded as the time-dependent component of $\mathbf{d}$ along $\boldsymbol{\psi}_{(\ell)}$. Note that $\psi_{(\ell)A}$ is the value of the ℓth eigenvector at node A. The summation index ℓ in eqn. (10.8.1) ranges from 1 to the desired number, n_{modes}, of modes deemed necessary to get a good solution of the problem. Of course, $n_{\text{modes}} \leq n_{\text{eqn}}$, where n_{eqn} equals the number of degrees of freedom in the finite element model of the problem. For each mode selected, the

eigenvalue $\lambda_{(\ell)}$ and the eigenvector $\psi_{(\ell)}$ should be good approximations to the exact eigenvalue and eigenvector. This requires that $n_{\rm modes} \ll n_{\rm eqn}$. However, the number of modes considered should be large enough to accurately represent $\mathbf{d}$ and the load vector $\mathbf{F}$ in terms of the series expansion in first $n_{\rm modes}$.

The next step is to integrate with respect to time t the $n_{\rm modes}$ scalar ordinary differential equations

$$\dot{a}_{(\ell)} + \lambda_{(\ell)} a_{(\ell)} = f_{(\ell)}, \tag{10.8.2}$$

$$a_{(\ell)}(0) = a_{0(\ell)}, \tag{10.8.3}$$

by using the generalized trapezoidal algorithm. Substitution for $a_{(\ell)}(t)$ into (31.1) gives the desired solution.

Because of the computational effort involved in finding the eigenvalues and eigenfunctions, the modal analysis is more efficient if many analyses of the same configuration are needed, and if only a small number of modes participate in the solution.

Chapter 11: Linear Elastodynamics

11.1 Problem Statement

Using the notations of Section 8.1, equations governing the transient deformations of a linear elastic homogeneous body are

$$\sigma_{ij,j} + f_i = \rho \ddot{u}_i, \text{ in } \Omega \times (0,T), \ i = 1,2,3, \tag{11.1.1}$$

$$\sigma_{ij} n_j = h_i(x,t) \text{ on } \Gamma_1 \times (0,T), \ i = 1,2,3, \tag{11.1.2}$$

$$u_i(x,t) = \hat{u}_i(x,t) \text{ on } \Gamma_2 \times (0,T), \ i = 1,2,3, \tag{11.1.3}$$

$$\sigma_{ij}(x,t) = D_{ijk\ell} e_{k\ell} \text{ in } \Omega \times (0,T), \ i = 1,2,3, \tag{11.1.4}$$

$$e_{ij}(x,t) = (u_{i,j} + u_{j,i})/2 = u_{(i,j)} \text{ in } \Omega \times (0,T), \ i = 1,2,3, \tag{11.1.5}$$

$$u_i(x,0) = u_i^0(x), \text{ in } \Omega, \tag{11.1.6}$$

$$\dot{u}_i(x,0) = \dot{u}_i^0(x), \text{ in } \Omega, \tag{11.1.7}$$

where

$$\bar{\Gamma}_1 \cup \bar{\Gamma}_2 = \partial\Omega, \ \Gamma_1 \cap \Gamma_2 = \phi. \tag{11.1.8}$$

Here ρ is the mass density, and $(0,T)$ is the time duration of interest. Because of the presence of two time derivatives in (11.1.1), the initial displacements and the initial velocities need to be prescribed.

In order to delineate whether the problem described by equations (11.1.1) through (11.1.8) is parabolic or hyperbolic, we consider the one-dimensional problem, namely, that of wave propagation in a bar. With $\mathbf{f} = \mathbf{0}$, equation (11.1.1) reduces to

$$E\frac{\partial^2 u}{\partial x^2} = \rho \frac{\partial^2 u}{\partial t^2}, \tag{11.1.9}$$

where E is Young's modulus. A comparison of (11.1.9) with (10.1.1) yields $a = E$, $b = 0$ and $c = -\rho$ where we have set $y = t$. Thus $b^2 - ac = E\rho > 0$ and equation (11.1.9) is hyperbolic. Thus the system of equations (11.1.1) is hyperbolic. In a hyperbolic problem waves propagate with a finite speed. It will be interesting to see if the numerical solution correctly predicts the shape and size of a propagating pulse.

11.2 Semi-discrete Formulation

Following the procedures of Section 8.1 and 10, equations (11.1.1)-(11.1.8) result in the following system of coupled ordinary differential equations.

$$\mathbf{M}\ddot{\mathbf{d}} + \mathbf{K}\mathbf{d} = \mathbf{F}, \tag{11.2.1}$$

$$\mathbf{d}(0) = \mathbf{d}_0,\ \dot{\mathbf{d}}(0) = \dot{\mathbf{d}}_0. \tag{11.2.2}$$

In eqn. (11.2.2), $\mathbf{d}_0$ and $\dot{\mathbf{d}}_0$ are known. For a linear viscoelastic body, the left-hand side of eqn. (11.2.1) has an additional term $\mathbf{C}\dot{\mathbf{d}}$ where $\mathbf{C}$ is derived from the viscoelastic constants of the material. In structures made of linear elastic members, the damping matrix $\mathbf{C}$ accounts for the frictional effects at the joints, and is ususally taken to equal $a\mathbf{M} + b\mathbf{K}$ where a and b are constants; the damping in this case is called the Rayleigh damping. Below we study the numerical solution of the set of equations

$$\mathbf{M}\ddot{\mathbf{d}} + \mathbf{C}\dot{\mathbf{d}} + \mathbf{K}\mathbf{d} = \mathbf{F}, \tag{11.2.3}$$

$$\mathbf{d}(0) = \mathbf{d}_0,\ \dot{\mathbf{d}}(0) = \dot{\mathbf{d}}_0. \tag{11.2.4}$$

11.3 The Newmark Method

The Newmark method involves finding $\mathbf{a}_{n+1},\ \mathbf{v}_{n+1}$ and $\mathbf{d}_{n+1}$ from $\mathbf{a}_n,\ \mathbf{v}_n$ and $\mathbf{d}_n$ by using the following three sets of equations.

$$\mathbf{M}\mathbf{a}_{n+1} + \mathbf{C}\mathbf{v}_{n+1} + \mathbf{K}\mathbf{d}_{n+1} = \mathbf{F}_{n+1}, \tag{11.3.1}$$

$$\mathbf{d}_{n+1} = \mathbf{d}_n + \Delta t\mathbf{v}_n + \frac{\Delta t^2}{2}[(1-2\beta)\mathbf{a}_n + 2\beta\mathbf{a}_{n+1}] \tag{11.3.2}$$

$$\mathbf{v}_{n+1} = \mathbf{v}_n + \Delta t[(1-\gamma)\mathbf{a}_n + \gamma\mathbf{a}_{n+1}]. \tag{11.3.3}$$

Here $\mathbf{a}_n \simeq \ddot{\mathbf{d}}(t_n)$, $\mathbf{v}_n \simeq \dot{\mathbf{d}}(t_n)$ and $\mathbf{d}_n \simeq \mathbf{d}(t_n)$; β and γ are constants. Values of β and γ determine the stability, the order of accuracy, and other properties of the algorithm. Solving (11.3.2) for $\mathbf{a}_{n+1}$, substituting the result into (11.3.3) and then substituting for $\mathbf{v}_{n+1}$ and $\mathbf{a}_{n+1}$ into (11.3.1) we obtain

$$\left(\frac{1}{\beta\Delta t^2}\mathbf{M} + \frac{\gamma}{\beta\Delta t}\mathbf{C} + \mathbf{K}\right)\mathbf{d}_{n+1} = \mathbf{F}_{n+1} - \mathbf{F}_n^{\text{int}} \tag{11.3.4}$$

$$\begin{aligned}\mathbf{F}_n^{\text{int}} = &\left(\frac{1}{\beta\Delta t^2}\mathbf{M} - \frac{\gamma}{\beta\Delta t}\mathbf{C}\right)\mathbf{d}_n + \left(\frac{1}{\beta\Delta t}\mathbf{M} + \left(1-\frac{\gamma}{\beta}\right)\mathbf{C}\right)\mathbf{v}_n \\ &+ \left(\frac{1-2\beta}{2\beta}\mathbf{M} - \Delta t\left(1-2\gamma+\frac{\gamma}{\beta}\right)\mathbf{C}\right)\mathbf{a}_n.\end{aligned} \tag{11.3.5}$$

Having found $\mathbf{d}_{n+1}$ from (11.3.4), one can determine $\mathbf{a}_{n+1}$ and $\mathbf{v}_{n+1}$ from (11.3.4) and (11.3.5) respectively. This implementation of the Newmark method suggests that it is a one step method since $\mathbf{d}_{n+1}$ is determined from a knowledge of the values of $\mathbf{a}_n,\ \mathbf{v}_n$ and $\mathbf{d}_n$, i.e., values at the immediately preceding time.

11.4 Analysis of the Stability of the Newmark Method

In order to study the stability of the method, we first use the modal analysis to uncouple the system (11.2.3) of equations. Let $(\lambda_{(1)},\ \boldsymbol{\psi}_{(1)}),\ (\lambda_{(2)},\ \boldsymbol{\psi}_{(2)})\ldots$ be eigensolutions of

$$(\mathbf{K}-\lambda\mathbf{M})\boldsymbol{\Psi}=\mathbf{0}, \tag{11.4.1}$$

$$\boldsymbol{\psi}_{(\ell)}^T\mathbf{M}\boldsymbol{\psi}_{(m)}=\delta_{\ell m}, \tag{11.4.2}$$

where $\lambda=\omega^2$ and ω is a natural frequency of the undamped system. For the Rayleigh damping

$$\begin{aligned}\boldsymbol{\psi}_{(\ell)}^T\mathbf{C}\boldsymbol{\psi}_{(m)} &= a\boldsymbol{\psi}_{(\ell)}^T\mathbf{M}\boldsymbol{\psi}_{(m)}+b\boldsymbol{\psi}_{(\ell)}^T\mathbf{K}\boldsymbol{\psi}_{(m)},\\ &= a\delta_{\ell m}+b\omega_{(\ell)}^2\delta_{\ell m},\\ &= \omega_{(\ell)}[a/\omega_{(\ell)}+b\omega_{(\ell)}]\delta_{\ell m},\\ &= 2\xi_{(\ell)}\omega_{(\ell)}\delta_{\ell m},\end{aligned} \tag{11.4.3}$$

where $\xi_{(\ell)}=(a/\omega_{(\ell)}+b\omega_{(\ell)})/2$ is called the damping ratio. Writing

$$d_A(t)=\sum_{\ell=1}^{\text{nodes}} d_{(\ell)}(t)\psi_{(\ell)A}, \tag{11.4.4}$$

and following the steps employed in going from eqn. (10.3.28) to eqn. (10.5.8) we get

$$\begin{aligned}&\ddot{d}_{(\ell)}+2\xi_{(\ell)}\omega_{(\ell)}\dot{d}_{(\ell)}+\omega_{(\ell)}^2 d_{(\ell)}=f_{(\ell)},\ \ell=1,2,\ldots,\\ &d_{(\ell)}(0)=d_{0(\ell)},\ \dot{d}_{(\ell)}(0)=\dot{d}_{0(\ell)}.\end{aligned} \tag{11.4.5}$$

In the remainder of this section we drop the subscript ℓ enclosed in parentheses. In order to study the stability of the algorithm we set $f=0$. The application of the algorithm (11.3.4) and (11/3/5) to $\ddot{d}+2\xi\omega\dot{d}+\omega^2 d=0$ gives

$$\mathbf{A}_{(1)}\mathbf{y}_{n+1}=\mathbf{A}_{(2)}\mathbf{y}_n, \tag{11.4.6}$$

where

$$\mathbf{y}_{n+1} = \left\{ \begin{array}{c} d_{n+1} \\ \Delta t v_{n+1} \end{array} \right\},\ \mathbf{A}_{(1)} = \left[\begin{array}{cc} 1+\beta\Omega^2 & 2\beta\xi\Omega\Delta t \\ \Omega^2\gamma & (1+2\xi\Omega\gamma) \end{array} \right],$$
$$\mathbf{A}_{(2)} = \left[\begin{array}{cc} 1-(1-2\beta)\dfrac{\Omega^2}{2} & (1-\Omega(1-2\beta)) \\ -\Omega^2(1-\gamma) & (1-2\xi\Omega(1-\gamma)) \end{array} \right]. \tag{11.4.7}$$

Note that we are now simultaneously finding d_{n+1} and v_{n+1} from d_n and v_n. Rewriting eqn. (11.4.6) as

$$\mathbf{y}_{n+1} = \mathbf{A}\mathbf{y}_n \tag{11.4.8}$$

where $\mathbf{A} = \mathbf{A}_{(1)}^{-1}\mathbf{A}_{(2)}$, we see that the ampification matrix $\mathbf{A}$ is a 2×2 matrix.

Let λ_1 and λ_2 be (possibly complex) eigenvalues of $\mathbf{A}$, and $|\lambda_1| = (\lambda_1\bar{\lambda}_1)^{1/2}$ where $\bar{\lambda}_1$ equals the complex conjugate of λ_1. The spectral radius $\hat{\rho}$ of the matrix $\mathbf{A}$ is defined as

$$\hat{\rho}(\mathbf{A}) = \max(|\lambda_1|, |\lambda_2|). \tag{11.4.9}$$

The algorithm defined by eqns. (11.3.4) and (11.3.5) is spectrally stable if

$$\begin{aligned} \hat{\rho}(\mathbf{A}) &\leq 1 \quad \text{when} \quad \lambda_1 \neq \lambda_2, \\ &< 1 \quad \text{when} \quad \lambda_1 = \lambda_2. \end{aligned} \tag{11.4.10}$$

The reason for requiring $\hat{\rho}(\mathbf{A}) < 1$ for $\lambda_1 = \lambda_2$ is the following. When $\lambda_1 \neq \lambda_2$, $\mathbf{A}$ has the decomposition

$$\mathbf{A} = \mathbf{P}\mathbf{\Lambda}(\mathbf{A})\mathbf{P}^{-1},\ \mathbf{\Lambda}(\mathbf{A}) = \left[\begin{array}{cc} \lambda_1(\mathbf{A}) & 0 \\ 0 & \lambda_2(\mathbf{A}) \end{array} \right], \tag{11.4.11}$$

where $\mathbf{P}$ is the matrix of linearly independent eigenvectors of $\mathbf{A}$. Therefore,

$$\mathbf{A}^n = \mathbf{P}\mathbf{\Lambda}(\mathbf{A})^n\mathbf{P}^{-1},\ \mathbf{\Lambda}(\mathbf{A})^n = \left[\begin{array}{cc} \lambda_1(\mathbf{A})^n & 0 \\ 0 & \lambda_2(\mathbf{A})^n \end{array} \right]. \tag{11.4.12}$$

However, when $\lambda_1 = \lambda_2 = \lambda$, the decomposition of $\mathbf{A}$ is

$$\mathbf{A} = \mathbf{P}\mathbf{J}(\mathbf{A})\mathbf{P}^{-1},\ \mathbf{J}(\mathbf{A}) = \left[\begin{array}{cc} \lambda(\mathbf{A}) & 1 \\ 0 & \lambda(\mathbf{A}) \end{array} \right] \tag{11.4.13}$$

and

$$\mathbf{J}(\mathbf{A})^n = \left[\begin{array}{cc} \lambda(\mathbf{A})^n & n\lambda(\mathbf{A})^{n-1} \\ 0 & \lambda(\mathbf{A})^n \end{array} \right]. \tag{11.4.14}$$

Note that $|\lambda(\mathbf{A})| < 1$ implies that $|n\lambda(\mathbf{A})^{n-1}| \to 0$ as $n \to \infty$. However, if $|\lambda(\mathbf{A})| = 1$, then $|n\lambda(\mathbf{A})^n| = n$ and the solution will grow with n.

The eigenvalues λ_1 and λ_2 of $\mathbf{A}$ are given by

$$\det[\mathbf{A} - \lambda \mathbf{1}] = 0 \text{ or } \lambda^2 - \lambda(A_{11} + A_{22}) + A_{11}A_{22} - A_{12}A_{21} = 0, \tag{11.4.15}$$

where $\mathbf{1}$ is the 2×2 identity matrix, and $A_{ij}(i, j = 1, 2)$ are elements of the matrix $\mathbf{A}$. With the notations

$$A_1 = (A_{11} + A_{22})/2,\ A_2 = A_{11}A_{22} - A_{12}A_{21}, \tag{11.4.16}$$

eqn. $(11.4.15)_2$ becomes

$$\lambda^2 - 2\lambda A_1 + A_2 = 0. \tag{11.4.17}$$

From the definition (11.4.10) of stability, it follows that the boundary of the stability region consists of all points at which $\hat{\rho}(\mathbf{A}) = 1$. It may be determined by setting $\lambda = \lambda(\alpha) = e^{-i\alpha}$, where $i = \sqrt{-1}$ and $\alpha \in [0,\ 2\pi]$. Substituting $\lambda = e^{-i\alpha}$ into (11.4.17) and recalling the trigonometric identities

$$e^{i\alpha} = \cos\alpha + i\sin\alpha\ ,\ \cos 2\alpha = 2\cos^2\alpha - 1\ ,\ \sin 2\alpha = 2\sin\alpha\cos\alpha,$$

we obtain

$$\begin{aligned} 0 &= (\cos 2\alpha - i\sin 2\alpha) - 2A_1(\cos\alpha - i\sin\alpha) + A_2\ , \\ &= (2\cos^2\alpha - 1 - 2A_1\cos\alpha + A_2) + i[-2\sin\alpha\cos\alpha + 2A_1\sin\alpha]\ , \\ &= [2\cos\alpha(\cos\alpha - A_1) + A_2 - 1] + i[2\sin\alpha(A_1 - \cos\alpha)]\ . \end{aligned} \tag{11.4.18}$$

Equating the real and imaginary parts equal to zero gives

$$0 = 2\cos\alpha(\cos\alpha - A_1) + A_2 - 1\ ,\ 0 = 2\sin\alpha(A_1 - \cos\alpha) \tag{11.4.19}$$

We consider the following three cases:

(i) $\alpha = 0, \qquad \lambda = +1$. For this case $(11.4.19)_2$ is satisfied and $(11.4.19)_1$ gives

$$0 = 1 - 2A_1 + A_2\ . \tag{11.4.20}$$

(ii) $\alpha = \pi, \qquad \lambda = -1$. Again $(11.4.19)_2$ is identically satisfied, and $(11.4.19)_1$ gives

$$0 = 1 + 2A_1 + A_2\ . \tag{11.4.21}$$

(iii) $0 < \alpha < \pi$ or $\pi < \alpha < 2\pi$. For these values of α, $\sin\alpha \neq 0$. Thus $(11.4.19)_2$ gives

$$A_1 = \cos\alpha \tag{$11.4.22_1$}$$

and $(11.4.19)_1$ yields

$$A_2 = 1. \tag{11.4.22$_2$}$$

Equations (11.4.20), (11.4.21) and (11.4.22) define 3 straight lines in the $A_1 - A_2$ plane; these are plotted in Fig. 11.1. Thus $\hat{\rho}(\mathbf{A}) = 1$ on the boundary of the triangle, but at no point interior to it. At the origin, $A_1 = A_2 = 0$, and from eqn. (11.4.17) $\lambda = 0$, and thus $\hat{\rho}(\mathbf{A}) = 0$. Since $\hat{\rho}(\mathbf{A})$ is a continuous function of A_1 and A_2, $\hat{\rho}(\mathbf{A}) < 1$ inside the triangle. We must now exclude from the triangle all points which give rise to double roots of modulus unity. From (11.4.17) we conclude that double roots can only occur along the parabola $A_1^2 = A_2$ which intersects the boundaries of the triangle at two points,

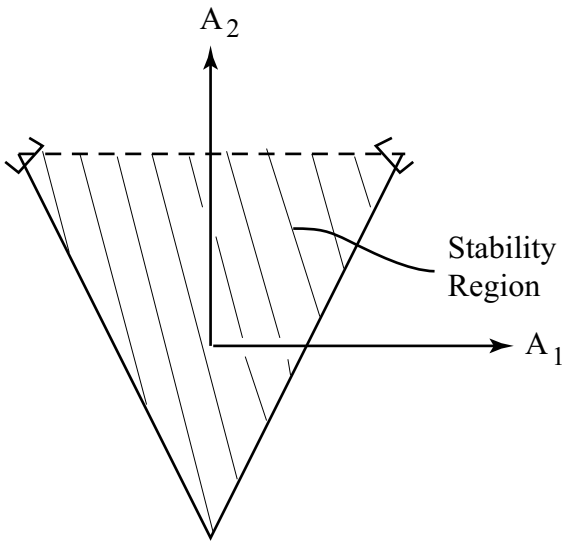


Fig. 11.1: Boundaries of the stability region are indicated by straight lines

$(A_1, A_2) = (\pm 1, 1)$, and double roots of modulus unity occur at each of these points. Hence the stability region is the closed triangle minus the two points $(A_1, A_2) = (\pm 1, 1)$.

A useful analytical description of the stability region is given in terms of the following two sets of conditions

$$-(A_2 + 1)/2 \leq A_1 \leq (A_2 + 1)/2, \ -1 \leq A_2 < 1, \tag{11.4.23}$$

and

$$-1 < A_1 < 1, \ A_2 = 1. \tag{11.4.24}$$

Condition (11.4.23) pertains to the interior and side boundaries of the triangle, and (11.4.24) to the upper boundary.

Equations (11.4.16), and (11.4.7) and the definition of the matrix $\mathbf{A}$ yield

$$\begin{aligned} A_1 &= 1 - \left[\xi\Omega + \Omega^2\left(\gamma + \frac{1}{2}\right)/2\right] \Big/ D, \\ A_2 &= 1 - \left[2\xi\Omega + \Omega^2\left(\gamma - \frac{1}{2}\right)\right] \Big/ D \end{aligned} \tag{11.4.25}$$

where

$$D = 1 + 2\xi\gamma\Omega + \beta\Omega^2 \, . \tag{11.4.26}$$

Expressions (11.4.25) and (11.4.26) when substituted into (11.4.23) and (11.4.24) yield the following stability conditions for the Newmark method. [For details, see

(i) H. M. Hilber, Analysis and Design of Numerical Integration Methods in Structural Dynamics, EERC Report No. 76-29, Earthquake Eng'g Research Center, Univ. of California, Berkeley, CA, November 1976.

(ii) H. M. Hilber and T. J. R. Hughes, Collocation, Dissipation and 'Overshoot' for Time Integration Schemes in Structural dynamics, Earthquake Eng'g & Structural dynamics, Vol. 6, 99-118, 1978.

(iii) H. M. Hilber, T. J. R. Hughes and R. L. Taylor, Improved Numerical Dissipation for Time Integration Algorithms in Structural dynamics, Earthquake Eng'g & Structural dynamics, Vol. 5, 281-292, 1977.]

Unconditional Stability

$$2\beta \geq \gamma \geq \frac{1}{2} \, . \tag{11.4.27}$$

Conditional Stability

$$\gamma \geq \frac{1}{2}, \; \beta < \frac{\gamma}{2}, \; \omega_{\max}\Delta t \leq \Omega_{\text{crit}} \, , \tag{11.4.28}$$

where

$$\Omega_{\text{crit}} = \frac{\xi\left(\gamma - \frac{1}{2}\right) + \left[\frac{\gamma}{2} - \beta + \xi^2\left(\gamma - \frac{1}{2}\right)^2\right]^{1/2}}{\left(\frac{\gamma}{2} - \beta\right)} \, . \tag{11.4.29}$$

Remarks

1. Note that $\omega_{\max}$ may be bounded by the maximum frequency of the individual element.

2. If $\gamma = 1/2$, viscous damping has no effect on stability. Furthermore, if $\gamma > \dfrac{1}{2}$, the damping increases Ω_{crit}. Thus, the undamped critical frequency,

$$\Omega_{\text{crit}} = \left(\frac{\gamma}{2} - \beta\right)^{-1/2} \tag{11.4.30}$$

serves as a conservative value when an estimate of the modal damping coefficient is not available. We can now discuss the stability of several well-known methods.

Method	β	γ	Stability Condition ($\xi = 0$)
Average accel.	$\frac{1}{4}$	$\frac{1}{2}$	Unconditional
Linear accel.	$\frac{1}{6}$	$\frac{1}{2}$	$\Omega_{\text{crit}} = 2\sqrt{3} = 3.464$
Fox-Goodwin	$\frac{1}{12}$	$\frac{1}{2}$	$\Omega_{\text{crit}} = \sqrt{6} = 2.449$
Central-diff.	0	$\frac{1}{2}$	$\Omega_{\text{crit}} = 2.$

These methods are at most 2nd-order accurate; 2nd order accuracy is achieved if and only if $\gamma = 1/2$.

Oscillatory Response

In practice, it is generally advisable to satisfy conditions slightly more stringent than (11.4.27) - (11.4.29) required for stability. These conditions follow from the requirement that the eigenvalues of the amplification matrix be complex conjugates. This is the case if

$$A_1^2 < A_2. \tag{11.4.31}$$

The Newmark methods satisfy (11.4.31) and the spectral stability requirements when the following conditions hold.

Unconditional Stability:

$$0 \le \xi < 1,\ \gamma \ge \frac{1}{2},\ \beta \ge \left(\gamma + \frac{1}{2}\right)^2 /4, \tag{11.4.32}$$

Conditional Stability:

$$0 \leq \xi < 1,\ \gamma \geq \frac{1}{2},\ \Omega < \Omega_{bif}\ , \tag{11.4.33}$$

$$\Omega_{bif} = \frac{\frac{1}{2}\xi\left(\gamma - \frac{1}{2}\right) + \left[\frac{1}{4}\left(\gamma + \frac{1}{2}\right)^2 - \beta + \xi^2\left(\beta - \frac{1}{2}\gamma\right)\right]^{1/2}}{\frac{1}{4}\left(\gamma + \frac{1}{2}\right)^2 - \beta}\ ; \tag{11.4.34}$$

in the undamped case

$$\Omega_{bif} = \left[\frac{1}{4}\left(\gamma + \frac{1}{2}\right)^2 - \beta\right]^{-1/2}\ . \tag{11.4.35}$$

Ω_{bif} is the value of Ω at which complex conjugate eigenvalues bifurcate into real, distinct eigenvalues.

11.5 Viscous Damping and High Frequency Behavior

11.5.1 Viscous Damping

We now show that the addition of viscous damping may not filter out high frequency modes. Figure 33.2 depicts $\hat{\rho}$ vs. $\Delta t/T$ for the average acceleration method with $\gamma = 1/2$. It is clear that only a middle band of frequencies is substantially affected. Since $\hat{\rho}_\infty = 1$, higher frequencies are virtually unaffected.

We note that for unconditionally stable algorithms such as the average acceleration method, one selects Δt small enough to accurately represent the low modes. Thus Δt is very large as compared with the time period of higher modes, so that $\Delta t/T \gg 1$ for these modes.

Stiffness proportional damping gives ξ proportional to ω and $\rho_\infty = 1$. Thus it is ineffective as a mechanism for removing high-frequency response.

11.5.2 High-Frequency Behavior

The high frequencies and mode shapes of the spatially discretized equations generally do not represent accurately the behavior of the original problem. It is therefore desirable to filter out the response of these modes in transient analysis. To study this phenomenon, one should consider the spectral radius.

If $\hat{\rho}(\mathbf{A}) < 1$, then $\mathbf{A}^n$ decays like $\hat{\rho}(\mathbf{A})^n$. The closer $\hat{\rho}(\mathbf{A})$ is to one, the slower is the decay. Thus for high frequency modes to be damped out, $\hat{\rho}(\mathbf{A}) < 1$ for these modes.

For $\gamma = 0.9$ and $\xi = 0$, Fig. 11.3 below

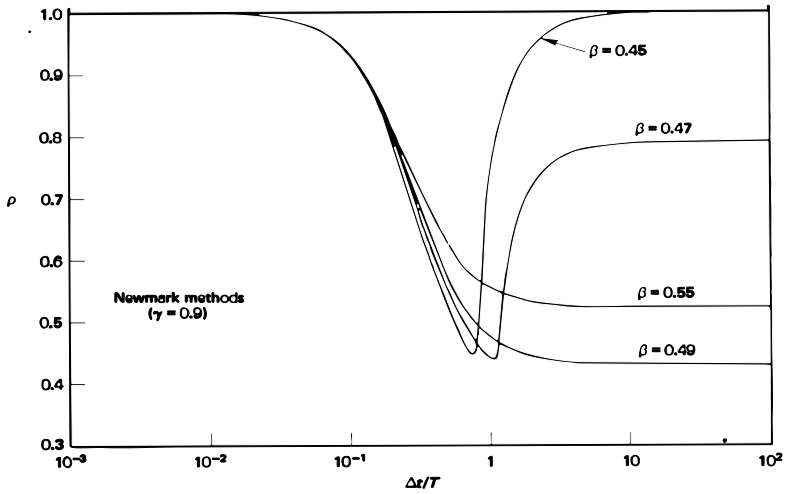


Fig. 11.3: Spectral radii for Newmark methods for varying β

gives a plot of the spectral radius, $\hat{\rho}(\mathbf{A})$, versus $\Delta t/T$. Let

$$\hat{\rho}_\infty = \lim_{\Delta t/T \to \infty} \hat{\rho}(\mathbf{A}), \tag{11.5.1}$$

where $T = 2\pi/\omega$ is the time period. If β is selected according to (11.4.32) with equality holding, then $\hat{\rho}_\infty = 0.43$ and there is strong damping in the high modes. Increasing, or decreasing, β reduces $\hat{\rho}_\infty$. The minimum value of β which insures unconditional stability is $\beta = \frac{\gamma}{2} = 0.45$. However, for this value of β, $\hat{\rho}_\infty = 1$ and the high frequency modes are undamped. The minimum points of the $\beta = 0.45$ and $\beta = 0.47$ curves represent the values of $\Delta t/T$ at which bifurcation of the complex conjugate roots occurs. Since $\beta = 0.49$ and $\beta = 0.55$ satisfy (11.4.32), no bifurcation occurs. The behavior represented in Fig. 11.3 is representative of the general case and for this reason it is usually recommended that β be selected according to (11.4.32) rather than (11.4.27). If (11.4.32) is not satisfied, then Δt should be selected according to (11.4.33) rather than (11.4.28).
Note: $\beta = \frac{1}{4}\left(\gamma + \frac{1}{2}\right)^2 \Rightarrow \hat{\rho}_\infty = \left|1 - 2/\left(\gamma + \frac{1}{2}\right)\right|$. Thus $\beta = 1,\ \gamma = 3/2$ entails $\hat{\rho}_\infty = 0$, or maximal damping of high frequencies. Unfortunately, this algorithm has very poor accuracy in the low frequency modes, and thus is not useful for transient analysis.

11.5.3 Numerical Dissipation and Dispersion

The single degree of freedom problem

$$\ddot{d} + 2\xi\omega d + \omega^2 d = 0\ , \tag{11.5.2}$$

subject to initial conditions

$$d(0) = D_0,\ \dot{d}(0) = V_0\ , \tag{11.5.3}$$

has the analytical solution given below.

Case 1: $0 \leq \xi < 1$ (underdamped)

$$d(t) = e^{-\xi\omega t}\left(D_0 \cos\omega_d t + \frac{V_0 + \xi\omega D_0}{\omega_d}\sin\omega_d t\right)\ , \tag{11.5.4}$$

where

$$\omega_d = (1-\xi^2)^{1/2}\omega\ , \tag{11.5.5}$$

is the damped natural frequency.

Case 2: $\xi = 1$ (overdamped)

$$d(t) = e^{-\omega t}(D_0 + (V_0 + \omega D_0)t)\ . \tag{11.5.6}$$

Case 3: $\xi > 1$ (overdamped)

$$d(t) = e^{-\xi\omega t}\left(D_0 \cosh\hat{\omega} t + \frac{V_0 + \xi\omega D_0}{\hat{\omega}}\sinh\hat{\omega} t\right)\ , \tag{11.5.7}$$

where

$$\hat{\omega} = (\xi^2 - 1)^{1/2}\omega\ . \tag{11.5.8}$$

The discrete analogs of these cases are given below.

continuous	discrete
underdamped	complex conjugate roots λ_1, λ_2
critically damped	real and identical roots
overdamped	real and distinct roots

To see, write

$$\left\{\begin{array}{c} d_{n+1} \\ v_{n+1} \end{array}\right\} = [A]\left\{\begin{array}{c} d_n \\ v_n \end{array}\right\}\ , \tag{11.5.9}$$

in the expanded form

$$\begin{aligned} d_{n+1} &= A_{11}d_n + A_{12}v_n\ , \\ v_{n+1} &= A_{21}d_n + A_{22}v_n\ . \end{aligned} \tag{11.5.10}$$

Replace n by $(n-1)$ in (11.5.10) and eliminate v_n. The result is

$$d_{n+1} - 2A_1 d_n + A_2 d_{n-1} = 0\ , \tag{11.5.11}$$

where A_1 and A_2 are given by (11.4.16) If the eigenvalues of $\mathbf{A}$ are distinct, then the general solution of (11.5.11) is

$$d_n = c_1\lambda_1^n + c_2\lambda_2^n \ , \tag{11.5.12}$$

where c_1 and c_2 are constants. Note that c_1, c_2 used in different equations may have different values. If $\mathbf{A}$ has two equal eigenvalues, then the general solution of (11.5.11) is

$$d_n = (c_1 + c_2 n)\lambda^n \ . \tag{11.5.13}$$

If $|\lambda| = 1$, the possibility of a "weak instability" arises whenever d_n grows as a polynomial in n (e.g. $n, n^2, n^3 \ldots$), the instability is referred to as a weak instability. This type of growth is considerably weaker than that of instabilities characterized by $|\lambda|^n$, where $|\lambda| > 1$.

The most typical case in structural dynamics is that of underdamping. We thus develop the discrete analog when ω_d is given by (11.5.5). Let

$$\mathbf{y}(t_{n+1}) = \mathbf{E}\mathbf{y}(t_n) \ , \tag{11.5.14}$$

where

$$\mathbf{y} = \left\{ \begin{array}{c} d \\ v \end{array} \right\} \ , \tag{11.5.15}$$

and $\mathbf{E}$ is the exact amplification matrix. The eigenvalues of $\mathbf{E}$ can be written as

$$\lambda_{1,2}(\mathbf{E}) = e^{-\xi\Omega \pm i\Omega_d} \ , \tag{11.5.16}$$

where $\Omega = \omega\Delta t$ and $\Omega_d = \omega_d\Delta t$. The invariants of $\mathbf{E}$ are

$$E_1 = \frac{1}{2}\mathrm{tr}\,\mathbf{E} = e^{-\xi\Omega}\cos\Omega_d \ , \tag{11.5.17}$$

$$E_2 = \det\mathbf{E} = e^{-2\xi\Omega} \ , \tag{11.5.18}$$

and the spectral radius is

$$\hat{\rho}(\mathbf{E}) = \sqrt{E_2} = e^{-\xi\Omega} \ . \tag{11.5.19}$$

One can verify that the exact analog of (11.5.12), namely

$$d(t_n) = c_1\lambda_1(\mathbf{E})^n + c_2\lambda_2(\mathbf{E})^n \tag{11.5.20}$$

leads back to (11.5.2) upon substitution of (11.5.7) and the use of the initial conditions.

For the discrete problem, eigenvalues of $\mathbf{A}$ can be written as

$$\lambda_{1,2}(\mathbf{A}) = e^{-\bar{\xi}\bar{\Omega}\pm i\bar{\Omega}_d} \tag{11.5.21}$$

where $\bar{\Omega} = \bar{\omega}\Delta t$, $\bar{\omega}_d = (1-\xi^2)^{1/2}\bar{\Omega}$, and

$$\bar{\Omega}_d = \tan^{-1}\left(\frac{A_2}{A_1^2} - 1\right)^{1/2},\ \bar{\xi} = -\frac{1}{2\bar{\Omega}}\ln A_2\ . \tag{11.5.22}$$

Thus

$$A_1 = e^{-\bar{\xi}\bar{\Omega}}\cos\bar{\Omega}_d\ ,\ A_2 = e^{-2\bar{\xi}\bar{\Omega}}\ ,\ \hat{\rho}(\mathbf{A}) = e^{-\bar{\xi}\bar{\Omega}}\ . \tag{11.5.23}$$

Employing (11.5.21) in (11.5.12), and evaluating the result at $n=0$ and $n=1$, gives

$$d_n = e^{-\bar{\xi}\bar{\omega}t_n}(D_0\cos\bar{\omega}_d t_n + \bar{c}\sin\bar{\omega}_d t_n)\ , \tag{11.5.24}$$

where

$$\bar{c} = \frac{d_1 - A_1 D_0}{(A_2 - A_1^2)^{1/2}} = \frac{\frac{1}{2}(A_{11} - A_{22})D_0 + A_{12}V_0}{(A_2 - A_1^2)^{1/2}}\ . \tag{11.5.25}$$

Here $\bar{\xi}$ and $\bar{\omega}$ are the algorithmic counterparts of ξ and ω, respectively. Their values may be found by substituting for A_1 and A_2 into (11.5.22). $\bar{\xi}$ may be regarded as a measure of the numerical dissipation, and $(\bar{T}-T)/T$ of relative period error where $\bar{T} = 2\pi/\bar{\omega}$ and $T = 2\pi/\omega$. In general, it is difficult to obtain analytical expressions for $\bar{\xi}$ and $(\bar{T}-T)/T$. Consequently, one must resort to numerical evaluation.

The following partial analytical results may be obtained for the Newmark method:

$$\bar{\xi} = \xi + \Omega\left(\gamma - \frac{1}{2}\right)/2 + \bigcirc(\Omega^2)\ , \tag{11.5.26}$$

$$\frac{\bar{T}-T}{T} = \bigcirc(\Omega^2)\ . \tag{11.5.27}$$

Remarks:

1. First-order errors resulting from $\gamma \neq 1/2$ manifest themselves only in the form of excess numerical dissipation, and not in period discrepancies.

2. If $\xi = 0$ and $2\beta \geq \gamma = \frac{1}{2}$, then $\bar{\xi} = 0$. That is, there is no numerical damping.

3. If $\xi = 0$ and $2\beta \geq \gamma > \frac{1}{2}$, then $\bar{\xi} > 0$. That is, values of $\gamma > \frac{1}{2}$ may be associated with numerical damping.

11.6 Matched Methods

Assume $\xi = 0$ and $\gamma = \frac{1}{2}$, thus $\bar{\xi} = 0$, and there is no amplitlude decay. The periods of individual modes of the finite element model will, however, be distorted by the particular integrator. For example, if $\beta = 1/4$, periods will be elongated. and for the central-difference method, $\beta = 0$, periods will be shortened.

Note that the periods of the finite element model are already in error when compared with those of the original "exact" (analytical) problem. For the consistent mass matrix, one can show that the finite element frequencies are bounded from above by the corresponding analytical ones. Numerical experiments show that lumped mass matrices tend to behave in the opposite fashion.

Thus the transient integrators and mass matrices should be matched so that the induced errors may cancel with each other. For example, trapezoidal rule $(\beta = 1/4)$ and consistent mass matrix, and central difference and lumped mass matrix, would be appropriate matches.

The following example taken from R. D. Krieg and S. W. Key, Transient Shell Response by Numerical Time Integration, *Int. J. Num. Meth. Engng.*, vol. 7, 273-286, 1973, illustrates these ideas.

For the wave equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} \,, \tag{11.6.1}$$

consider a finite element mesh of uniform linear elements of length $h = 1/n_{el}$ where n_{el} equals the number of elements. The element stiffness matrix K^e is given by

$$K^e = \frac{1}{h} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} . \tag{11.6.2}$$

A one-parameter (r) family of element mass matrices may be written as

$$M^e_{(r)} = h \begin{bmatrix} \frac{1}{2} - r & r \\ r & \frac{1}{2} - r \end{bmatrix} , \tag{11.6.3}$$

and $r = 1/6,\ 0$ and $1/12$ correspond to the consistent, lumped and higher-order mass matrices.

Let

$$\begin{aligned}\omega &= \text{exact frequency},\\ \omega^h &= \text{frequency produced by the finite element discretization},\\ \bar{\omega}^h &= \text{frequency produced by the time integration algorithm}\\ &\qquad \text{in conjunction with the finite element spatial discretization.}\end{aligned}$$

Thus we see $\bar{\omega}^h$ but will like to see ω. The semi-discrete formulation of the problem corresponding to (11.6.1) is

$$\mathbf{M}_{(r)}\ddot{\mathbf{d}} + \mathbf{K}\mathbf{d} = 0\,. \tag{11.6.4}$$

Let $d_A(t)$ denote the displacement at node A, and

$$\mathbb{C}d_A = d_{A+1} - 2d_A + d_{A-1}, \tag{11.6.5}$$

be the undivided second central difference operator. For an interior node A, eqn. (11.6.6) gives

$$h(1 + r\mathbb{C})\,\ddot{d}_A - \frac{1}{h}\mathbb{C})\,d_A = 0\,. \tag{11.6.6}$$

We wish to solve (11.6.6) exactly. Let the solution be of the form

$$d_A(t) = S_A T(t)\,. \tag{11.6.7}$$

Substituting (11.6.7) into (11.6.6) and adding and substracting $(\omega^h)^2 T(1 + r\mathbb{C})S_A$ yields

$$[\ddot{T} + (\omega^h)^2 T](1 + r\mathbb{C})S_A - \left[(\omega^h)^2(1 + r\mathbb{C})S_A + \frac{1}{h^2}\mathbb{C}S_A\right] T = 0\,. \tag{11.6.8}$$

This eqn. is satisfied if

$$\ddot{T} + (\omega^h)^2 T = 0\,, \tag{11.6.9}$$

$$(\omega^h)^2(1 + r\mathbb{C})S_A + \frac{1}{h^2}\mathbb{C}S_A = 0. \tag{11.6.10}$$

We assume that the general solution of (11.6.10) is of the form

$$S_A = c_1 \sin\frac{A\lambda}{n_{el}} + c_2 \cos\frac{A\lambda}{n_{el}} \tag{11.6.11}$$

where values of λ are determined by imposing homogeneous boundary conditions. (Note that λ here has a different meaning from that in the previous sections). For

$$\text{fixed-fixed, and free-free ends } \lambda = \ell\pi, \tag{11.6.12}$$

$$\text{free-fixed, and fixed-free ends } \lambda = (2\ell - 1)\pi/2, \tag{11.6.13}$$

where ℓ is an integer. These values of λ coincide with the exact frequencies, ω, obtained by solving the eigenvalue problem associated with (11.6.1).

For the fixed-fixed case, $c_2 = 0$ in (11.6.11), and (11.6.10) becomes

$$\left(\omega^h h\right)^2 \sin\frac{A\omega}{n_{el}} + [1 + r(\omega^h h)^2]\left(\sin\frac{(A-1)\omega}{n_{el}} - 2\sin\frac{A\omega}{n_{el}} + \sin\frac{(A+1)\omega}{n_{el}}\right) = 0. \quad (11.6.14)$$

Recalling that $\sin(a \pm b) = \sin a \cos b \pm \cos a \sin b$, we get

$$\left(\frac{\omega^h h}{2}\right)^2 \left[1 + 2r\left(\cos\frac{\omega}{n_{el}} - 1\right)\right] + \frac{1}{2}\left(\cos\frac{\omega}{n_{el}} - 1\right) = 0\,. \quad (11,6.15)$$

Using $\sin^2 a/2 = (1 - \cos a)/2$, eqn. (11.6.15) becomes

$$\left(\frac{\omega^h h}{2}\right)^2 \left[1 - 4r\sin^2\frac{\omega}{2n_{el}}\right] - \sin^2\frac{\omega}{2n_{el}} = 0\,. \quad (11.6.16)$$

Since $n_{el} = 1/h$, therefore

$$\frac{\omega^h}{\omega} = \frac{\sin\omega h/2}{\frac{\omega h}{2}\left[1 - 4r\sin^2\left(\frac{\omega h}{2}\right)\right]^{1/2}}\,. \quad (11.6.17)$$

For the Newmark method, we assume $\xi = 0$ and $\gamma = 1/2$. Then eqn. (11.5.22) and expressions (11.4.16) for A_1 and A_2 become

$$\bar{\omega}^h \Delta t = \tan^{-1}\left[\frac{(A_2 - A_1^2)^{1/2}}{A_1}\right]\,, \quad (11.6.18)$$

where

$$A_1 = 1 - \frac{(\omega^h \Delta t)^2}{2(1 + \beta(\omega^h \Delta t)^2)}\,, \quad (11,6.19)$$

$$A_2 = 1. \quad (11.6.20)$$

We now use the trigonometric identities

$$\begin{aligned} \cot a &= \cot 2a + \operatorname{cosec} 2a, \text{ and} \\ \operatorname{cosec} a &= (\cot^2 a + 1)^{1/2}\ . \end{aligned} \quad (11.6.21)$$

These enable us to conclude from (11.6.18) that

$$\cot^2\frac{\bar{\omega}^h \Delta t}{2} = \frac{1 + A_1}{1 - A_1} = -1 + 4\beta + \left(\frac{2}{\omega^h \Delta t}\right)^2\,. \quad (11.6.22)$$

Using $\sin^2 a = 1/(1 + \cot^2 a)$ we get

$$\sin^2\frac{\bar{\omega}^h \Delta t}{2} = \frac{\sin^2(\omega h/2)}{4[\beta - r(h/\Delta t)^2]\sin^2\left(\frac{\omega h}{2}\right) + \left(\frac{h}{\Delta t}\right)^2}\,. \quad (11.6.23)$$

Note that if $h = \Delta t$ and $\beta = r$, then (11.6.23) implies that $\bar{\omega}^h = \omega$. That is, the errors introduced by the finite element spatial discretization, the particular mass matrix and the temporal algorithm will cancel to yield exact results.

Remarks

1. If eqn. (11.6.1) is replaced by
$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \tag{11.6.24}$$
where $c^2 = E/\rho$, then h should be replaced by h/c in (11.6.23). Thus $\Delta t = h/c$, and $\beta = r$ will give numerical results which are exact at the nodes no mattter how few elements are employed. The simplest choice is $\beta = r = 0$, i.e., the central difference method and the lumped mass matrix.

2. Note that reducing Δt while holding the element length fixed can only worsen the results. In this case we converge to the exact solution of the spatially discrete, temporally continuous system (i.e. mass points and springs) rather than the exact solution of (11.6.24).

3. The value $\Delta t = h/c$ represents the stability limit of the lumped mass, central-difference method. Therefore, computations can not be performed with $\Delta t > h/c$. It is generally considered advisable to use Δt as close to h/c as possible.

4. In more general settings, e.g., unequal element lengths, variable material properties, multi-dimensional problems etc., results obtained by matched methods, such as central differences and lumped mass, will not be exact. However, results obtained with matched methods will generally be superior to those obtained with inappropriate combinations, such as consistent mass and central differences.

11.7 An Alternative Method to Study the Stability of the Algorithm

In order to study the stability of the Newmark family of methods, it is sufficient to consider a single degree of freedom problem, i.e., assume that the governing equations have been uncoupled. Consider the equation
$$\ddot{d} + \omega^{h^2} d = 0\ . \tag{11.7.1}$$

According to the Newmark method

$$d_{n+1} = d_n + \Delta t \dot{d}_n + \frac{\Delta t^2}{2}[(1-2\beta)\ddot{d}_n + 2\beta \ddot{d}_{n+1}] \ . \tag{11.7.2}$$

Substitution for $\ddot{d}_n$ and $\ddot{d}_{n+1}$ from (11.7.1) into (11.7.2) yields

$$\begin{aligned} d_{n+1} &= d_n + \Delta t \dot{d}_n + \frac{\Delta t^2}{2}[(1-2\beta)(-\omega^{h^2})d_n + 2\beta(-\omega^{h^2})d_{n+1}] \ , \\ &= d_n \left(1 - \frac{\Omega^2}{2} + \beta\Omega^2\right) + \Delta t \dot{d}_n - \beta\Omega^2 d_{n+1} \ , \end{aligned} \tag{11.7.3}$$

where $\Omega = \omega^h \Delta t$. Rewriting eqn. (11.7.3) as

$$(d_{n+1} - d_n)(1+\beta\Omega^2) = -\frac{\Omega^2}{2} d_n + \Delta t \dot{d}_n \ , \tag{11.7.4}$$

and recalling that in the Newmark family of methods

$$\begin{aligned} \dot{d}_n &= \dot{d}_{n-1} + \Delta t[(1-\gamma)\ddot{d}_{n-1} + \gamma \ddot{d}_n] \ , \\ &= \dot{d}_{n-1} + \Delta t(1-\gamma)(-\omega^{h^2} d_{n-1}) + \Delta t \gamma(-\omega^{h^2} d_n) \ , \end{aligned}$$

where we have substituted for $\ddot{d}_n$ from (11.7.1). Multiplying both sides of this equation by Δt and using (11.7.4), we obtain

$$\Delta t \dot{d}_n = \Delta t \dot{d}_{n-1} - \Omega^2(1-\gamma) d_{n-1} - \gamma \Omega^2 d_n \tag{11.7.5}$$

Substitute for $\Delta t \dot{d}_n$ and $\Delta t \dot{d}_{n-1}$ from (11.7.4) into (11.7.5) to get

$$(1+\beta\Omega^2)(d_{n+1} - d_n) + \frac{\Omega^2}{2} d_n = (1+\beta\Omega^2)(d_n - d_{n-1}) + \frac{\Omega^2}{2} d_{n-1} - \Omega^2(1-\gamma) d_{n-1} - \gamma\Omega^2 d_n \ ,$$

which can be rewritten as

$$(1+\beta\Omega^2)(d_{n+1} - 2d_n + d_{n-1}) + \Omega^2 \left(\gamma - \frac{1}{2}\right)(d_n - d_{n-1}) + \Omega^2 d_n = 0. \tag{11.7.6}$$

With

$$\delta = \gamma - \frac{1}{2} \text{ and } \alpha^2 = \Omega^2/(1+\beta\Omega^2) \ , \tag{11.7.7}$$

eqn. (11.7.6) can be written as

$$(d_{n+1} - 2d_n + d_{n-1}) + \delta\alpha^2(d_n - d_{n-1}) + \alpha^2 d_n = 0 \ \ .$$

Let $d_n = \mu^n$, then we get

$$\mu^2 + \mu(\alpha^2 + \alpha^2\delta - 2) + (1 - \alpha^2\delta) = 0 \ ,$$

which is of the form

$$\mu^2 + \mu b + c = 0 \tag{11.7.8}$$

where

$$b = \alpha^2(1+\delta) - 2\ , \quad c = 1 - \alpha^2\delta\ . \tag{11.7.9}$$

Set

$$\mu = Re^{i\theta}\ . \tag{11.7.10}$$

For the algorithm to be stable, θ must be real and $|R| \leq 1$, otherwise the solution will grow unboundedly. From (11.7.8) and (11.7.10) we obtain

$$R(\cos\theta + i\sin\theta) = \mu = \frac{-b \pm \sqrt{b^2-4c}}{2} = \frac{-b \pm i\sqrt{4c-b^2}}{2}\ .$$

Equating real and imaginary parts on both sides of the equation gives

$$R\cos\theta = -\frac{b}{2}\ , \quad R\sin\theta = \frac{\sqrt{4c-b^2}}{2}\ \ .$$

Therefore

$$\begin{aligned} R^2 &= \frac{b^2}{4} + \frac{4c-b^2}{4} = c = 1 - \alpha^2\delta = 1 - \alpha^2\left(\gamma - \frac{1}{2}\right)\ , \\ \tan\theta &= -\frac{\sqrt{4c-b^2}}{b}, \text{ and for real } \theta,\ 4c \geq b^2\ . \end{aligned} \tag{11.7.11}$$

Thus, for stability

$$1 - \frac{\Omega^2\left(\gamma - \frac{1}{2}\right)}{(1+\beta\Omega^2)} \leq 1, \text{ and } 4(1-\alpha^2\delta) \geq (\alpha^2(1+\delta)-2)^2\ . \tag{11.7.12}$$

Equations $(11.7.12)_1$ and $(11.7.12)_2$ are equivalent to

$$\Omega^2\left(\gamma - \frac{1}{2}\right) \geq 0, \text{ and } 4 + \Omega^2\left(4\beta - \left(\gamma + \frac{1}{2}\right)^2\right) \geq\ 0. \tag{11.7.13}$$

Thus if $\gamma \geq \dfrac{1}{2}$ and $\beta \geq \dfrac{\left(\gamma+\frac{1}{2}\right)^2}{4}$, then there is no restriction on the time step size and the algorithm is unconditionally stable. Otherwise, we satisfy (11.7.12) and (11.7.13) by requiring that

$$\gamma \geq \frac{1}{2}\ , \quad \text{and} \quad \Omega \leq \left[\frac{1}{\frac{1}{4}\left(\gamma+\frac{1}{2}\right)^2 - \beta}\right]^{1/2}\ . \tag{11.7.14}$$

11.8 Time Periods of the Newmark Algorithm

A general solution of eqn. (11.7.1) according to the Newmark algorithm is

$$d_n = A\cos n\theta + B\sin n\theta \tag{11.8.1}$$

where

$$\tan\theta = \frac{\sqrt{4c-b^2}}{-b},\ b=\alpha^2(1+\delta)-2,\ c=1-\alpha^2\delta\ .$$

We now consider the case of $\gamma = \dfrac{1}{2}$ only. Thus $\delta = \gamma - \dfrac{1}{2} = 0$, $c=1$, $b=\alpha^2-2$, and

$$\tan\theta = \frac{\sqrt{4-(\alpha^2-2)^2}}{-(\alpha^2-2)} = \frac{\alpha(1-\alpha^2/4)^{1/2}}{(1-\alpha^2/2)}\ .$$

Recalling that $\sec^2\theta = 1+\tan^2\theta$, we obtain

$$\cos^2\theta = \frac{1}{1+\dfrac{\alpha^2\left(1-\dfrac{\alpha^2}{4}\right)}{1+\dfrac{\alpha^4}{4}-\alpha^2}} = \left(1-\frac{\alpha^2}{2}\right)^2\ ,$$

and thus

$$\cos\theta = 1-\frac{\alpha^2}{2} = 1-2\sin^2\theta/2\ ,$$

which implies that

$$\frac{\alpha^2}{4} = \sin^2\frac{\theta}{2} \text{ or } \theta = 2\sin^{-1}\frac{\alpha}{2}\ . \tag{11.8.2}$$

An analytical solution of (11.7.1) is

$$\begin{aligned} d_n &= A_1\cos\omega^h t + B_1\sin\omega^h t\ \ , \\ &= A_1\cos\Omega n + B_1\sin\Omega n\ \ , \end{aligned} \tag{11.8.3}$$

where we have set $t = n\Delta t$, and $\Omega = \omega^h\Delta t$. A comparison of (11.8.1) and (11.8.3) yields

$$\frac{\omega^{\text{algor.}}}{\omega^{\text{analyt.}}} = \frac{\theta}{\Omega} = \frac{2\sin^{-1}\dfrac{\alpha}{2}}{\Omega^h\Delta t} = \frac{\sin^{-1}\left(\dfrac{\omega^h\Delta t/2}{\sqrt{1+\beta(\omega^h\Delta t)^2}}\right)}{\omega^h\Delta t/2}\ .$$

If $\omega^h\Delta t \ll 1$, then

$$\begin{aligned} \frac{T^{\text{algor.}}}{T^{\text{analyt.}}} &= \frac{\omega^h\Delta t/2}{\dfrac{\omega^h\Delta t/2}{\sqrt{1+\beta(\omega^h\Delta t)^2}}} = \sqrt{1+\beta(\omega^h\Delta t)^2}\ , \\ &\simeq 1+\frac{1}{2}\beta(\omega^h\Delta t)^2. \end{aligned}$$

Thus

$$\frac{T^{\text{algor.}} - T^{\text{analyt.}}}{T^{\text{analyt.}}} \simeq \frac{1}{2}\beta(\omega^h \Delta t)^2 = \bigcirc(\Delta t^2) \, .$$

The difference in the time periods of the numerical and analytical solutions is proportional to $\beta\Omega^2$. The error in the computed time periods will increase as the square of the frequency and the time step size implying that computed higher frequencies will have larger errors.

11.9 Time-Step Estimates for Some Simple Finite Elements

Recall that

$$\omega_{\max} \leq \max_e \{\omega^e_{\max}\} \, . \tag{11.9.1}$$

Thus we need to evaluate the maximum frequency of free vibration of an element; we do so below for some simple finite elements.

11.9.1 Two-node linear rod element

The stiffness matrix and the lumped mass matrix for this element are

$$M^e = \frac{\rho h}{2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \; K^e = \frac{E}{h}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

Therefore

$$\det[K^e - \lambda M^e] = 0$$

gives

$$\left(\frac{E}{h} - \lambda\frac{\rho h}{2}\right)^2 - \left(\frac{E}{h}\right)^2 = 0 \, ,$$

or

$$\lambda_{\max} = \frac{4c^2}{h^2} \text{ where } c^2 = E/\rho \, .$$

Thus

$$\omega_{\max} = 2c/h, \text{ and } \Delta t_{\text{crit}} = \frac{2}{\omega_{\max}} = \frac{h}{c}. \tag{11.9.2}$$

The cirtical time step for the Newmark method with $\beta = 0$, $\gamma = 1/2$, i.e., the central difference method equals the time required for a wave to traverse one element in the rod.

For a consistent mass matrix,

$$M^e = \frac{\rho h}{6}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \, ,$$

and $\det[M^e - \lambda K^e] = 0$ gives

$$\left(\frac{E}{h} - \lambda\frac{\rho h}{3}\right)^2 - \left(\frac{E}{h} + \lambda\frac{\rho h}{6}\right)^2 = 0 \,,$$

or

$$\lambda_{\max} = \frac{12c^2}{h^2} = \omega_{\max}^2.$$

The critical time step for the central difference method is given by

$$\Delta t_{\rm crit} = \frac{2}{\omega_{\max}} = \frac{h}{\sqrt{3}c} \,. \tag{11.9.3}$$

This result is typical: *Consistent mass matrices yield smaller critical time steps than lumped-mass matrices*.

When studying the heat conduction problem with one-dimensional linear elements, we have

$$M^e = \frac{\rho c h}{2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},\ K^e = \frac{k}{h}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \,.$$

Therefore

$$\lambda_{\max} = 4k/(\rho c h^2) \,,$$

and

$$\Delta t_{\rm crit} = 2/\lambda_{\max} = h^2/(k/\rho c) \,.$$

Thus the critical time step varies as the square of the length of the element for the heat conduction problem.

11.9.2 Three-node quadratic element

If we assume a lumped mass matrix based on Simpson's rule weighting (i.e. the ratio of the middle node mass to the end node masses is 4), we get

$$\omega_{\max}^h = \frac{2\sqrt{6}c}{h},\ \Delta t_{\rm crit} = \frac{h}{\sqrt{6}c}. \tag{11.9.4}$$

Based on equal nodal spacing, the allowable time-step is 0.8164 of that for linear elements with lumped mass.

11.9.3 Quadrilateral and hexahedral elements

Flanagan and Belytschko (D. P. Flanagan and T. Belytschko, A Uniform Strain Hexahedron and Quadrilateral with Orthogonal Hourglass Control, *Int. J. Num. Meth. Engng.*, 17, 679-706, 1981)

have estimated the maximum frequency for the one-point quadrature, 4-node quadrilateral and 8-node hexahedron elements. Their result is

$$\omega_{\max} \leq C_d g^{1/2}, \tag{11.9.5}$$

where $C_d^2 = (\lambda + 2\mu)/\rho$ is the dilatational wave speed, λ and μ are the Lamé constants, and g is a geometric parameter. For a quadrilateral

$$g = \frac{4}{A^2} \sum_{i=1}^{2} \sum_{a=1}^{4} B_{ia} B_{ia} \tag{11.9.6}$$

$$B_{ia} = \frac{1}{2} \begin{bmatrix} y_2 - y_4 & y_3 - y_1 & y_4 - y_2 & y_1 - y_3 \\ x_4 - x_2 & x_1 - x_3 & x_2 - x_4 & x_3 - x_1 \end{bmatrix} \tag{11.9.7}$$

where $(x_a,\ y_a)$ are the coordinates of node a and A is the area of the element. The estimate (11.4.2) leads to a sufficient condition for stability. For the central-difference method,

$$\Delta t \leq \frac{2}{C_d g^{1/2}}. \tag{11.9.8}$$

For a rectangular element with side lengths h_1 and h_2,

$$\Delta t \leq \frac{1}{C_d (1/h_1^2 + 1/h_2^2)^{1/2}}. \tag{11.9.9}$$

For higher-order elements, very little of precise nature has been done.

11.10 Another Look at the Newmark Method

This is a one-step method in which $d(t)$ is approximated by a cubic polynomial

$$d(t) = C_0 + C_1 t + C_2 t^2 + C_3 t^3 \,, \tag{11.10.1}$$

over the time interval $[t_{n-1}, t_n]$. The coefficients C_0 and C_1 are determined by interpolating $d(t)$ and $\dot{d}(t)$ at time t_{n-1}. Therefore (11.10.1) becomes

$$d(t) = d_{n-1} + v_{n-1}(t - t_{n-1}) + C_2 (t - t_{n-1})^2 + C_3 (t - t_{n-1})^3 \,. \tag{11.10.2}$$

where we have set $v = \dot{d}$. The other two coefficients are determined by the following two conditions on $d(t)$ and $\dot{d}(t)$ at time $t = t_n$.

$$\begin{aligned} d(t_n) &\simeq d_n = d_{n-1} + v_{n-1}\Delta t + \frac{1}{2}[(1 - 2\beta)a_{n-1} + 2\beta a_n]\Delta t^2 \,, \\ \dot{d}(t_n) &\simeq v_n = v_{n-1} + \Delta t[(1 - \gamma)a_{n-1} + \gamma a_n] \,, \end{aligned} \tag{11.10.3}$$

where β and γ are constants and $a = \ddot{d}$. Equations (11.10.2) and (11.10.3) yield the following cubic polynomial:

$$d(t) = d_{n-1} + v_{n-1}(t-t_n) + \frac{1}{2}\left[\frac{a_{n-1}+a_n}{2} + \left(6\beta - 2\gamma - \frac{1}{2}\right)(a_n - a_{n-1})\right] \times (t-t_{n-1})^2 + (\gamma - 2\beta)\left(\frac{a_n - a_{n-1}}{\Delta t}\right)(t-t_{n-1})^3 \ . \tag{11.10.4}$$

It degenerates to a quadratic polynomial if $\gamma - 2\beta = 0$; then the acceleration will be constant over the interval $[t_{n-1}, t_n]$.

To derive the recurrence relation, we write the equation of motion at $t = t_n$,

$$Ma_n + Cv_n + Kd_n = F_n \ . \tag{11.10.5}$$

Eliminating a_n and v_n from (11.10.3)$_1$, (11.10.3)$_2$ and (11.10.5), we obtain the following

$$\left[\frac{1}{\beta\Delta t^2}M + \frac{\gamma}{\beta\Delta t}C + K\right] d_n = F_n + \left[\frac{1}{\beta\Delta t^2}M + \frac{\gamma}{\beta\Delta t}C\right] d_{n-1} + \left[\frac{1}{\beta\Delta t}M + \left(\frac{\gamma}{\beta} - 1\right)C\right] v_{n-1} + \left[\left(\frac{1}{2\beta} - 1\right)M + \left(\frac{\gamma}{2\beta} - 1\right)\Delta t C\right] a_{n-1}. \tag{11.10.6}$$

Knowing d_n, a_n is computed from (11.10.3)$_1$ and then v_n from (11.10.3)$_2$.

There are two cases that are especially popular.

1. $\gamma = \frac{1}{2},\ \beta = \frac{1}{4}$ (the average acceleration method).

   Equation (34.4) gives

   $$\begin{aligned} d(t) &= d_{n-1} + v_{n-1}(t-t_{n-1}) + \frac{1}{2}\left(\frac{a_{n-1}+a_n}{2}\right)(t-t_{n-1})^2, \\ v(t) &= v_{n-1} + \left(\frac{a_{n-1}+a_n}{2}\right)(t-t_{n-1}), \\ a(t) &= \frac{a_{n-1}+a_n}{2}. \end{aligned} \tag{11.10.7}$$

   Thus the acceleration over the interval $[t_{n-1}, t_n]$ stays constant and equals the average of its values at the end points.

2. $\gamma = \dfrac{1}{2},\ \beta = \dfrac{1}{6}$ (the linear acceleration method).

   Equation (11.10.4) yields

   $$\begin{aligned} d(t) &= d_{n-1} + v_{n-1}(t-t_{n-1}) + \frac{1}{2}a_{n-1}(t-t_{n-1})^2 + \frac{1}{6}\frac{a_n - a_{n-1}}{\Delta t}(t-t_{n-1})^3, \\ v(t) &= v_{n-1} + a_{n-1}(t-t_{n-1}) + \frac{1}{2}\frac{a_n - a_{n-1}}{\Delta t}(t-t_{n-1})^2, \\ a(t) &= a_{n-1} + \frac{a_n - a_{n-1}}{\Delta t}(t-t_{n-1}). \end{aligned} \tag{11.10.8}$$

Here $d(t)$ is a Taylor series expansion about t_{n-1}, and the acceleration linearly interpolates a_{n-1} and a_n.

11.11 The Houbolt Method

This is a three-step method in which $d(t)$ is approximated by a cubic polynomial that interpolates $d(t)$ at times t_{n-3}, t_{n-2}, t_{n-1} and t_n; that is,

$$d(t) = d_{n-3}N_1(t) + d_{n-2}N_2(t) + d_{n-1}N_3(t) + d_nN_4(t) \tag{11.11.1}$$

where $N_1,\ N_2,\ N_3$ and N_4 are the Lagrange cubic shape functions. Differentiating (11.11.1) twice with respect to t, evaluating $v(t)$ and $a(t)$ at $t = t_n$, and substituting into (11.10.5), we arrive at the following.

$$\begin{aligned}\left[\frac{2}{\Delta t^2}M + \frac{11}{6\Delta t}C + K\right] d_n = F_n + \left[\frac{5}{\Delta t^2}M + \frac{3}{\Delta t}C\right] d_{n-1} \\ - \left[\frac{4}{\Delta t^2}M + \frac{3}{2\Delta t}C\right] d_{n-2} + \left[\frac{1}{\Delta t^2}M + \frac{1}{3\Delta t}C\right] d_{n-3}.\end{aligned} \tag{11.11.2}$$

Differentiation of (11.11.1) with respect to time t also yields

$$a(t) = a_{n-2}\frac{t_{n-1} - t}{\Delta t} + a_{n-1}\frac{t - t_{n-2}}{\Delta t}\ , \tag{11.11.3}$$

where

$$\begin{aligned}a_{n-1} &= \frac{d_n - 2d_{n-1} + d_{n-2}}{\Delta t^2}\ , \\ a_{n-2} &= \frac{d_{n-1} - 2d_{n-2} + d_{n-3}}{\Delta t^2}\ .\end{aligned} \tag{11.11.4}$$

Equation (11.11.3) linearly interpolates the acceleration at times t_{n-1} and t_{n-2}, and a_{n-1} and a_{n-2} are evaluated by using the central-difference method. The recurrence relation (11.11.2) assumes uniform time steps. However, uniform steps are clearly not necessary. Such a "generalized Houbolt" relation is used in a large commercial FE program. Since the recursive relation (11.11.2) expresses d_n in terms of $d_{n-1},\ d_{n-2}$, and d_{n-3}, it is sometimes described as a backward-difference method.

It is implicit, unconditionally stable for linear problems, quite "robust" even for nonlinear problems. It tends to act somewhat "overdamped".

It needs a special starting procedure since values of d_{-1} and d_{-2} are needed to compute d_1. The relation (11.11.1) can easily be generalized to express $d(t)$ as a quartic or higher-order polynomial in t.

11.12 The Wilson-θ Method

It is a one-step method in which $d(t)$ is approximated as a cubic polynomial:

$$d(t) = c_0 + c_1 t + c_2 t^2 + c_3 t^3 \tag{11.12.1}$$

defined over $[t_{n-1},\ t_n + \theta\Delta t]$. The parameter θ determines how far beyond t_n the range of definition extends. The coefficients $c_0,\ c_1,\ c_2$, and c_3 are determined by interpolating $d(t),\ \dot{d}(t)$, and $\ddot{d}(t)$ at t_{n-1} and $\ddot{d}(t)$ at $t_\theta = t_n + \theta\Delta t$. The result is

$$d(t) = d_{n-1} + v_{n-1}(t - t_{n-1}) + \frac{1}{2}a_{n-1}(t - t_{n-1})^2 + \frac{1}{6}\frac{a_\theta - a_{n-1}}{\theta\Delta t}(t - t_{n-1})^3. \tag{11.12.2}$$

Thus

$$\begin{aligned} a(t) &= a_{n-1} + \frac{a_\theta - a_{n-1}}{\theta\Delta t}(t - t_{n-1}), \\ v(t) &= v_{n-1} + a_{n-1}(t - t_{n-1}) + \frac{1}{2}\frac{a_\theta - a_{n-1}}{\theta\Delta t}(t - t_{n-1})^2. \end{aligned} \tag{11.12.3}$$

The equation of motion is now evaluated at time t_θ:

$$Ma_\theta + Cv_\theta + Kd_\theta = F_\theta \tag{11.12.4}$$

where $F_\theta = (1-\theta)F_{n-1} + \theta F_n$, and expressions for a_θ and v_θ are obtained from (11.12.3) by setting $t = t_\theta$. These result in the following

$$\begin{aligned} \left[\frac{6}{\theta^2\Delta t^2}M + \frac{3}{\theta\Delta t}C + M\right] d_\theta &= F_{n-1} + \theta(F_n - F_{n-1}) \\ &+ \left[\frac{6}{\theta^2\Delta t^2}M + \frac{3}{\theta\Delta t}C\right] d_{n-1} + \left(\frac{6}{\theta\Delta t}M + 2C\right) v_{n-1} + \left[2M + \frac{\theta\Delta t}{2}C\right] a_{n-1}\,. \end{aligned} \tag{11.12.5}$$

Equation (11.12.5) is the Wilson recurrence relation. After solving it for d_θ, a_θ is obtained from (11.12.2) and then $d_n,\ v_n$ and a_n are computed from (11.12.2) and (11.12.3).

The Wilson-θ method is one-step, implicit, and unconditionally stable for linear problems only if $\theta \geq 1.37;\ \theta = 1.40$ is normally used in practice.

11.13 Park's Method

For the first-order, linear , ordinary differential equation

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, t) \tag{11.13.1}$$

Park suggested the following algorithm.

$$\left[\sum_{i=0}^{3} \alpha_i y_{n+1-i} + \Delta t \beta_i f(y_{n+1-i},\ t_{n+1-i})\right] = 0, \tag{11.13.2}$$

with

$$\alpha_0 = -1,\ \alpha_1 = \frac{3}{2},\ \alpha_2 = -\frac{3}{5},\ \alpha_3 = \frac{1}{10},\ \beta_0 = \frac{3}{5},\ \beta_1 = \beta_2 = \beta_3 = 0. \tag{11.13.3}$$

Thus, the recursive relation is

$$-y_{n+1} + \frac{3}{2}y_n - \frac{3}{5}y_{n-1} + \frac{1}{10}y_{n-2} + \Delta t \frac{3}{5} f(y_{n+1},\ t_{n+1}) = 0\ . \tag{11.13.4}$$

The second-order, linear, ordinary differential equation

$$M\ddot{d} + C\dot{d} + kd = F \tag{11.13.5}$$

can be written as

$$\{y\} = \left\{ \begin{array}{c} d \\ \dot{d} \end{array} \right\},\ G = \left[\begin{array}{cc} 0 & I \\ -M^{-1}K & -M^{-1}C \end{array} \right],\ H(t) = \left\{ \begin{array}{c} 0 \\ M^{-1}F(t) \end{array} \right\} \tag{11.13.6}$$

and

$$\{\dot{y}\} = Gy + H(t)\ . \tag{11.13.7}$$

Thus

$$\begin{aligned} &\left\{ \begin{array}{c} d_{n+1} \\ v_{n+1} \end{array} \right\} + \frac{3\Delta t}{5} \left[\begin{array}{cc} 0 & -I \\ M^{-1}K & M^{-1}C \end{array} \right] \left\{ \begin{array}{c} d_{n+1} \\ v_{n+1} \end{array} \right\} \\ &\quad = \frac{3}{2} \left\{ \begin{array}{c} d_n \\ v_n \end{array} \right\} - \frac{3}{5} \left\{ \begin{array}{c} d_{n-1} \\ v_{n-1} \end{array} \right\} + \frac{1}{10} \left[\begin{array}{c} d_{n-2} \\ v_{n-2} \end{array} \right\} + \frac{3\Delta t}{5} \left\{ \begin{array}{c} 0 \\ M^{-1}F_{n+1} \end{array} \right\}\ . \end{aligned} \tag{11.13.8}$$

The method is 2nd-order accurate, unconditionally stable, retains good accuracy in the low frequencies and strong dissipative characteristics in the high frequencies. Its disadvantages are that it entails significantly more historical data pool and requires a starting procedure. Park's method is more accurate than the Houbolt method.

11.14 Collocation Schemes

Collocation methods generalize and combine aspects of the Newmark method and the Wilson-θ method; they are defined by the following equations.

$$\left.\begin{aligned}
&Ma_{n+\theta} + Cv_{n+\theta} + Kd_{n+\theta} = F_{n+\theta} \\
&a_{n+\theta} = (1-\theta)a_n + \theta a_{n+1} \\
&F_{n+\theta} = (1-\theta)F_n + \theta F_{n+1} \\
&d_{n+\theta} = d_n + \theta\Delta t v_n + \frac{(\theta\Delta t)^2}{2}\left\{(1-2\beta)a_n + 2\beta a_{n+\theta}\right\}, \\
&v_{n+\theta} = v_n + \theta\Delta t\left\{(1-\gamma)a_n + \gamma a_{n+\theta}\right\} \\
&d_{n+1} = d_n + \Delta t v_n + \frac{\Delta t^2}{2}\left[(1-2\beta)a_n + 2\beta a_{n+1}\right] \\
&v_{n+1} = v_n + \Delta t\left[(1-\gamma)a + \gamma a_{n+1}\right].
\end{aligned}\right\} \tag{11.14.1}$$

θ is called the collocation parameter. For $\theta = 1$, the method becomes the Newmark method. If $\beta = \frac{1}{6}$ and $\gamma = \frac{1}{2}$, the Wilson-θ methods are obtained. A necessary and sufficient condition for 2nd order accuracy is that $\gamma = 1/2$. Unconditionally stable, 2nd-order accurate schemes are defined by

$$\gamma = \frac{1}{2},\ \theta \geq 1,\ \frac{\theta}{2(1+\theta)} \geq \beta \geq \frac{2\theta^2 - 1}{4(2\theta^3 - 1)}. \tag{11.14.2}$$

The one-parameter subfamily of methods with $\gamma = \dfrac{1}{2}$, $\theta = \theta^*(\beta)$ along with their algorithmic damping ratio and relative period error for $\Delta t/T = 0.1$ are listed below.

β	θ^*	$\overline{\xi}$	$(\overline{T} - T)/T$
$1/4$	1	0	0.032
0.24	1.021712	0.6×10^{-4}	0.032
0.23	1.047364	0.27×10^{-3}	0.033
0.20	1.159772	0.27×10^{-2}	0.039
0.17	1.381914	0.13×10^{-1}	0.060
0.16	1.514951	0.21×10^{-1}	0.075

11.15 α-Method (Hilber-Hughes-Taylor Method)

In this method, the finite difference formulas of the Newmark method are retained and the equation of motion is modified as follows:

$$Ma_{n+1} + (1+\alpha)Cv_{n+1} - \alpha Cv_n + (1+\alpha)Kd_{n+1} - \alpha Kd_n = F(t_{n+\alpha}) \tag{11.15.1}$$

where $t_{n+\alpha} = (1+\alpha)t_{n+1} - \alpha t_n = t_{n+1} + \alpha\Delta t$. If $\alpha = 0$, the method reduces to the Newmark method. If $\alpha \in [-\frac{1}{3}, 0]$, $\gamma = (1-2\alpha)/2$, and $\beta = (1-\alpha)^2/4$, then an unconditionally stable,

second-order accurate scheme results. For $\alpha = 0$, we have the trapezoidal rule. Decreasing α increases the amount of numerical dissipation.

11.16 Discussion of Time-Stepping Algorithms

For a method to be competitive, it should possess the following attributes.

1. Unconditional stability when applied to linear problems.
2. No more than one set of implicit equations to be solved at each time step.
3. Second-order accuracy.
4. Controllable algorithmic damping in higher modes.
5. Self-starting.

In complicated structural models containing slender members exhibiting bending effects, conditionally stable algorithms require time steps that are much smaller than those needed for accuracy, especially when only low-mode response is of interest. For this reason, unconditionally stable algorithms are generally preferred.

Numerical experiments have indicated that in structural dynamics, second-order accurate methods are vastly superior to first-order accurate methods. The second-order method with the smallest error constant is the trapezoidal rule or the average acceleration method $\left(\beta = \frac{1}{4},\ \gamma = \frac{1}{2}\right)$. However, the trapezoidal rule does not have numerical dissipation needed to damp out any spurious participation of the high frequency modes.

11.17 Overshoot

Goudreau and Taylor (G. L. Goudreau and R. L. Taylor, Evaluation of Numerical Methods in Elastodynamics, *J. of Computer Methods in Applied Mechs.& Engng.*, Vol. 2, 69-97, 1973) discovered a peculiar property of the Wilson-θ method. Even though the method is unconditionally stable, numerical experiments indicated a tendency to significantly "overshoot" exact solutions in the first few time steps. To see this, consider the following hypothetical amplification matrix.

$$A = \begin{bmatrix} \epsilon & k \\ 0 & \epsilon \end{bmatrix}$$

where $0 < \epsilon < 1$ and $k \gg 1$. The spectral radius of A is ϵ.

$$A^n = \begin{bmatrix} \epsilon^n & n\epsilon^{n-1}k \\ 0 & \epsilon^n \end{bmatrix}$$

and every term of A^n goes to zero as $n \to \infty$. However, due to the presence of k, the term $n\epsilon^{n-1}k$, is very large for small enough n. From the example we see that the long-term, or asymptotic, behavior is governed by the spectral properties of A. However, the short term potential for overshoot requires an examination of all elements of A.

11.18 Runge-Kutta Method

For the first-order ordinary differential equations

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, t) \tag{11.18.1}$$

the second-order Runge-Kutta method can be stated as follows

$$\begin{aligned} \mathbf{k}_1 &= \Delta t \mathbf{f}(\mathbf{y}_n, t_n) \\ \mathbf{k}_2 &= \Delta t \mathbf{f}\left(\mathbf{y}_n + \frac{\mathbf{k}_1}{2},\ t_n + \frac{\Delta t}{2}\right) \\ \mathbf{y}_{n+1} &= \mathbf{y}_n + \mathbf{k}_2 + \bigcirc(\Delta t^3) \end{aligned} \tag{11.18.2}$$

The name "second-order" follows from the following definition. A method is called nth order if its error term is $\bigcirc(\Delta t^{n+1})$.

There are many ways to evaluate the right-hand side $\mathbf{f}(\mathbf{y}, t)$ which all agree to first-order, but which have different coefficients of higher-order error terms. Adding up the right combination of these, we can eliminate the error terms order by order. The fourth-order Runge-Kutta formula is:

$$\begin{aligned} \mathbf{k}_1 &= \Delta t \mathbf{f}(\mathbf{y}_n, t_n), \\ \mathbf{k}_2 &= \Delta t \mathbf{f}\left(\mathbf{y}_n + \frac{\mathbf{k}_1}{2},\ t_n + \frac{\Delta t}{2}\right) \\ \mathbf{k}_3 &= \Delta t \mathbf{f}\left(\mathbf{y}_n + \frac{\mathbf{k}_2}{2},\ t_n + \frac{\Delta t}{2}\right), \\ \mathbf{k}_4 &= \Delta t \mathbf{f}(\mathbf{y}_n + \mathbf{k}_3,\ t_n + \Delta t) \\ \mathbf{y}_{n+1} &= \mathbf{y}_n + \frac{\mathbf{k}_1}{6} + \frac{\mathbf{k}_2}{3} + \frac{\mathbf{k}_3}{3} + \frac{\mathbf{k}_4}{6} + \bigcirc(\Delta t^5) \end{aligned} \tag{11.18.3}$$

We note that higher-order does not always mean higher accuracy. However, fourth-order Runge-Kutta is *generally* superior to second-order, and also to higher-order Runge-Kutta schemes. For orders M higher than four, more than M function evaluations but not more than $M+2$ are required. Thus fourth-order is a natural breakpoint.

Adaptive Stepsize Control for Runge-Kutta Method

The purpose of the adaptive stepsize control is to achieve predetermined accuracy in the solution with minimum computational effort. Implementation of adaptive stepsize control requires that the stepping algorithm return information about its performance, most important, an estimate of its truncation error. With fourth-order Runge-Kutta, the most straightforward technique by far is step doubling. We take each step twice, once as a full step, then, independently, as two half steps since the basic method is fourth order, the true solution and the two numerical approximations are related by

$$\begin{aligned} y(t+2\Delta t) &= y_1(t) + (2\Delta t)^5\phi + \bigcirc(\Delta t^6) + \dots \\ y(t+2\Delta t) &= y_2(t) + 2(\Delta t)^5\phi + \bigcirc(\Delta t^6) + \dots \end{aligned} \tag{11.18.4}$$

where, to order Δt^5, the value ϕ remains constant over the step, and its magnitude is of order $y^{(5)}(t)/5$. The first expression involves $(2\Delta t)^5$ since the stepsize is $2\Delta t$, while the second expression involves $2(\Delta t)^5$ since the error on each step is $\Delta t^5\phi$. The difference between the two numerical estimates is a convenient indicator of truncation error.

$$\Delta \equiv y_2 - y_1 \tag{11.18.5}$$

Since Δ scales as Δt^5, therefore,

$$\Delta t_0 = \Delta t_1 \left| \frac{\Delta_0}{\Delta_1} \right|^{0.2} \tag{11.18.6}$$

Thus if Δ_0 denotes the desired accuracy, we can estimate Δt_0. Thus if Δ_1 is larger than Δ_0 in magnitude, equation (11.18.6) gives how much to decrease the stepsize when we retry the present (failed) step.

How to choose Δ_0 in a set of equations whose dependent variables differ enormously in magnitude? One possibility is to use fractional errors, $\Delta_0 = \epsilon y$, where ϵ is a number like 10^{-6}. However, if the solution oscillates and passes through zero but is bounded by some maximum values, then $\Delta_0 = \epsilon$ (maximum value). Another possibility is

$$\Delta_{0i} = \epsilon y\mathrm{scal}_i, \ i = 1, 2, \dots \tag{11.18.7}$$

where ϵ is the overall tolerance, $y\mathrm{scal}_i$ is the value of y_i at the beginning of the step, and Δ_{0i} is the desired accuracy for the ith equation.

In some applications, it may be desirable to control "global" accumulation of errors, from the beginning to the end of the integration. In the worst possible case, the errors are presumed to add with the same sign. Then, the smaller the stepsize Δt, the smaller the value Δ_0 that will need to be imposed since there will be more total number of steps. In such cases

$$\Delta_0 = \epsilon \delta t \dot{y}_i \tag{11.18.8}$$

This enforces fractional accuracy not on y but on the increments of y during each step. If Δ_0 has an implicit scaling with Δt, then the exponent 0.2 is no longer correct. When the stepsize is reduced from a too-large value, then the new predicted Δt will fail to meet the desired accuracy when $y$scale's are altered to this new Δt value. Instead of $0.20 = 1/5$, we must scale by the exponent $0.25 = 1/4$ for things to work out.

Since exponents 0.25 and 0.20 are not really very different, we adopt the following pragmatic approach. Whenever we decrease a stepsize, we use the larger value of the exponent whether we need it or not, and whenever we increase a stepsize, we use the smaller exponent. Furthermore, because our estimates of error are not exact, we use a safety factor S which is a few percent smaller than unity. Thus

$$\begin{aligned} \Delta t_0 &= S \Delta t_1 \left| \frac{\Delta_0}{\Delta_1} \right|^{0.2}, \ \Delta_0 \geq \Delta_1, \\ \Delta t_0 &= S \Delta t_1 \left| \frac{\Delta_0}{\Delta_1} \right|^{0.25}, \ \Delta_0 < \Delta_1. \end{aligned} \tag{11.18.9}$$

11.19 Stiff Sets of Equations

As soon as one deals with more than one first-order differential equation, the possibility of a stiff set of equations arises. Stiffness occurs in a problem when there are two or more very different scales of the independent variables on which the dependent variables are changing. For example, consider the second-order differential equation

$$\ddot{y} = f(t) y = 100 y \tag{11.19.1}$$

with initial conditions

$$y(0) = 1, \ \dot{y}(0) = -10. \tag{11.19.2}$$

Then the true solution is

$$y = e^{-10t}. \tag{11.19.3}$$

However, the numerical integration methods would give a solution that will start off decaying as e^{-10t}, but would then "explode" as e^{10t} as t becomes large; the reason is any roundoff or truncation error as we start the integration. Thus

$$y_{\text{num}} \approx e^{-10t} + \epsilon e^{10t} \, . \tag{11.19.4}$$

No matter how small ϵ is made by taking a very small stepsize, sooner or later the second term in (11.19.4) dominates.

All stiff systems need not have divergent solutions. For example, consider equations

$$\begin{aligned} \dot{u} &= 998u + 1998v \\ \dot{v} &= -999u - 1999v \end{aligned} \tag{11.19.5}$$

with initial conditions

$$u(0) = 1, \; v(0) = 0. \tag{11.19.6}$$

By means of the transformation

$$u = 2y - z, \; v = -y + z \tag{11.19.7}$$

we find the solution

$$\begin{aligned} u &= 2e^{-t} - e^{-1000t} \\ v &= -e^{-t} + e^{-1000t}. \end{aligned} \tag{11.19.8}$$

Thus an integration method would require a stepsize $\Delta t \ll 1/1000$ for the method to be stable even though the e^{-1000t} term is completely negligible in determining the values of u and v as soon as one is away from the origin. We are required to follow the variation in the solution on the shortest length scale to maintain stability of the integration, even though accuracy requirements allow a much larger stepsize.

The simplest cure is to resort to an implicit method. For example, consider the single equation

$$\dot{y} = -cy, \; c > 0. \tag{11.19.9}$$

The backward Euler scheme gives

$$\begin{aligned} y_{n+1} &= y_n + \Delta t \dot{y}_{n+1} = y_n + \Delta t(-cy_{n+1}) \\ \text{or } y_{n+1} &= y_n/(1 + c\Delta t) \, . \end{aligned} \tag{11.19.10}$$

As $\Delta t \rightarrow \infty,\ y_{n+1} \rightarrow 0$, which is the correct solution of the differential equation. This nice feature of implicit methods holds only for linear systems. In the general case, implicit methods give better stability.

For the system of equations

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}) \tag{11.19.11}$$

implicit differencing gives

$$\begin{aligned} \mathbf{y}_{n+1} &= \mathbf{y}_n + \Delta t \mathbf{f}(t_{n+1},\ \mathbf{y}_{n+1})\ , \\ &\simeq \mathbf{y}_n + \Delta t \left[\mathbf{f}(t_{n+1}, \mathbf{y}_n) + \left.\frac{\partial \mathbf{f}}{\partial \mathbf{y}}\right|_{\mathbf{y}_n} \cdot (\mathbf{y}_{n+1} - \mathbf{y}_n) \right] \end{aligned} \tag{11.19.12}$$

so at each step we have to invert the matrix $\left(I - \Delta t \frac{\partial \mathbf{f}}{\partial \mathbf{y}} \right)$ to find $\mathbf{y}_{n+1}$. It is not guaranted to be stable, but it usually is, because the behavior is locally similar to the case when $\left(I - \Delta t \frac{\partial \mathbf{f}}{\partial \mathbf{y}} \right)$ is a constant matrix.

We can construct a second-order method by taking the average of the explicit and the implicit first-order methods.

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{\Delta t}{2} \left[\mathbf{f}(t_{n+1}, \mathbf{y}_n) + \left.\frac{\partial \mathbf{f}}{\partial \mathbf{y}}\right|_{\mathbf{y}_n} \cdot (\mathbf{y}_{n+1} - \mathbf{y}_n) + \mathbf{f}(t_n, \mathbf{y}_n) \right] . \tag{11.19.13}$$

For higher-order methods, see Gear or Stoer and Bulirsch.

Gear, C. William, 1971, Numerical Initial Value Problems in Ordinary Differential Equations, Prentice-Hall, Englewood Cliffs, N.J.

Stoer, J. and Bulirsch, R., 1980, Introduction to Numerical Analysis, Springer-Verlag, New York.

11.20 Element-by-Element Implicit Methods

Unconditionally stable, second-order accurate, implicit methods, such as the central-difference method, perform very well in heat conduction analysis. the main drawback of these methods is the large storage or memory requirement and the need to solve a system of simultaneous equations. One way to overcome these shortcomings is to approximate the global matrices by a product of element matrices. The inversion of the coefficient matrix is then replaced by the sequential inversion of element matrices.

Drawbacks: Storage
Equation solving burden

Solution: A product approximation of the element assembly is made so that the inversion of the coefficient matrix is replaced by sequential inveresions of element matrices.

The generalized trapezoidal algorithm can be written as

$$(M + \alpha\Delta t K)d_{n+1} = (M - (1-\alpha)\Delta t K)d_n + \Delta t F_{n+\alpha} \tag{11.20.1}$$

$$\begin{aligned} M + \alpha\Delta t K &= M^{1/2}(M^{1/1} + \alpha\Delta t m^{-1/2} K) \\ &= M^{1/2}(I + \alpha\Delta t M^{-1/2} K M^{-1/2})M^{1/2} \\ &= M^{1/2}(I + \alpha\Delta t C)M^{1/2} \end{aligned} \tag{11.20.2}$$

$$C = M^{-1/2} K M^{-1/2} \tag{11.20.3}$$

$$(M - (1-\alpha)\Delta t K) = M^{1/2}(I - (1-\alpha)\Delta t C)M^{+1/2} \tag{11.20.4}$$

Here $M^{1/2}$ is the square-root of M. For a diagonal mass matrix, $M^{1/2}$ is easily computed.