1 Introduction and Background to the FEM

The **Finite Element Method** (FEM) is a way of obtaining approximate (numerical) solutions to differential equations. Broadly speaking, the FEM is used to reduce differential equation(s) to systems of equations which can be more easily solved. There are two usual ways to derive this system of equations: using

(1) **Galerkin's weighted residual method**

or

(2) a **variational method** together with the **Rayleigh-Ritz scheme**

Both of these approaches are discussed below but the former, being the more general of the two, is the one which will be followed in most of this text. The Galerkin method is described in §1.1 and §1.2 and the Variational approach is briefly discussed in §1.3.

1.1 Weighted Residual Methods

The FEM using the Galerkin method is more specifically called the **Galerkin Finite Element Method** (GFEM). Before discussing the GFEM, which is done in the next Chapter, it is worthwhile discussing **Galerkin's Method**, from which it derives. (In fact, many of the important concepts of the FEM are touched upon in this chapter.) Galerkin's Method is what one might use to obtain a solution to a differential equation if one did not have a computer. It was only with the development of the computer in the 1950s that the Galerkin Method was generalised to the Galerkin FEM.

Galerkin's method[1] is one of a number of numerical techniques known as **Weighted Residual Methods**. These various weighted residual methods are often as effective as each other, but it is the Galerkin method which leads naturally into the Finite Element

[1] Boris Grigoryevich Galerkin was a Russian engineer who taught in the St. Petesburg Polytechnic. His method, which he originally devised to solve some structural mechanics problems, and which he published in 1915, now forms the basis of the Galerkin Finite Element method. I.G. Bubnov independently devised a similar method around the same time, and Galerkin's method is known also as the Bubnov-Galerkin method

Method[2]. The two other most commonly encountered weighted residual methods are the **Collocation Method** and the **Method of Least Squares**; these are special cases of the most general **Petrov-Galerkin Method**, which is described in §1.1.4.

The Collocation, Least Squares and Galerkin methods will be illustrated here through the following simple one dimensional example problem: solve the linear ordinary differential equation (ODE)

$$\frac{d^2u}{dx^2} - u = -x, \qquad u(0) = 0, \quad u(1) = 0 \tag{1.1}$$

[the exact solution is $u(x) = x - \frac{\sinh x}{\sinh(1)}$]

Begin by assuming some form to u, usually a polynomial[3]. For example, take as a **trial function**

$$\tilde{u}(x) = a + bx + cx^2 \tag{1.2}$$

This trial function has three unknowns, two of which can immediately be obtained from the boundary conditions (BC's), leading to a trial function which automatically satisfies these BC's:

$$\tilde{u}(x) = b\left(x - x^2\right) \tag{1.3}$$

It now remains to determine b.

1.1.1 The Collocation Method

The most direct method is to satisfy the differential equation at some point in the interval, $x \in [0,1]$ - this is the Collocation Method. Which point one chooses is arbitrary, but it makes sense to choose the midpoint, which usually yields best results, in which case, substituting (1.3) into (1.1) and setting $x = 1/2$, one finds that $b = 2/9$ and the approximate solution is

[2] rather, the most commonly encountered FEM is that based on the Galerkin method, but it is possible to derive FEM equations using other weighted residual methods, most importantly the Petrov-Galerkin Method (see later)

[3] it is not necessary to use polynomials, e.g. one could use sinusoids, $\sum_i \sin(ix)$

$$\tilde{u}(x) = \tfrac{2}{9}\left(x - x^2\right) \tag{1.4}$$

Slightly different results will be obtained by choosing to enforce the differential equation at different points.

More accuracy can be achieved by choosing higher order polynomials. For example, one could begin with a cubic and so have the trial function which satisfies the BC's

$$\tilde{u}(x) = bx + cx^2 - (b+c)x^3 \tag{1.5}$$

With two unknowns, one needs two equations. For example, enforcing (1.1) at the equi-spaced points $x = 1/3$ and $x = 2/3$ leads to the system of equations {▲Problem 2}

$$\begin{aligned} 62b + 2c &= 9 \\ 59b + 29c &= 9 \end{aligned} \tag{1.6}$$

and solving these leads to the approximate solution

$$\tilde{u}(x) = \frac{1}{560}\left[81x + 9x^2 - 90x^3\right] \tag{1.7}$$

The "1-term" (Eqn. 1.4) and "2-term" (Eqn. 1.7) approximate solutions are graphed in Fig. 1.1, the latter being virtually indistinguishable from the exact solution at this scale. Using the methods described here, one would expect the 1-term solution to be within, perhaps, 10-20% of the exact solution. Ever more accurate solutions can be obtained by increasing the order of the polynomial and solving systems with ever greater numbers of equations.

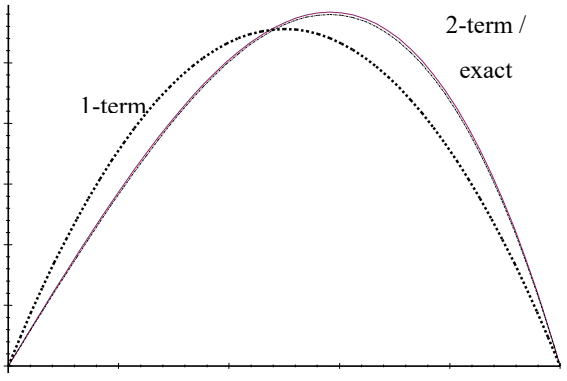


Figure 1.1: Collocation Method solution to Eqn. 1.1

1.1.2 The Method of Least Squares

Consider now an alternative solution procedure, wherein the differential equation is multiplied across by some weight function $\omega(x)$ and the complete equation is integrated over the domain:

$$\int_0^1 \left[\frac{d^2u}{dx^2} - u + x \right] \omega(x) dx = 0 \tag{1.8}$$

Again, choose a trial function which satisfies the boundary conditions, for example the quadratic (1.3), leading to

$$\int_0^1 \left[\frac{d^2\tilde{u}}{dx^2} - \tilde{u} + x \right] \omega(x) dx = \int_0^1 \left[bx^2 + (1-b)x - 2b \right] \omega(x) dx = 0 \tag{1.9}$$

The term inside the square brackets is the **residual** *R*, and it is this which one wants to drive to zero. The idea here is that if this integral is zero for any arbitrary weight function ω, then the residual should be zero also.

There is one unknown in (1.9) and the question now is: what function $\omega(x)$ does one choose? Again, the choice here is somewhat arbitrary (but see below). In the Least Squares method, one chooses $\omega(x) = \partial R / \partial b$, leading to a cubic integrand in Eqn. 1.9; integration then leads to the equation {▲Problem 3}

$$\tfrac{47}{10} b - \tfrac{13}{12} = 0 \quad \rightarrow \quad b = \tfrac{65}{282} \tag{1.10}$$

which is close to the Collocation Method solution $b = 2/9$.

Note the primary difference between the Collocation Method and the Least Squares Method. In the former, the differential equation is satisfied at one or more *particular points*. In the latter, (1.9), the differential equation is forced to zero in some *average* way, determined by the weight function, over the complete domain.

As before, choose now a higher order polynomial to improve the solution, say the cubic trial function of (1.5), rewritten as

$$\tilde{u}(x) = a_1 x + a_2 x^2 - (a_1 + a_2) x^3 \tag{1.11}$$

Again, this is substituted into (1.8). This time, with two unknown coefficients, one requires two equations, which are obtained by choosing two different weight functions, namely

$$\begin{aligned} \omega_1(x) &= \frac{\partial R}{\partial a_1} = -7x + x^3 \\ \omega_2(x) &= \frac{\partial R}{\partial a_2} = 2 - 6x - x^2 + x^3 \end{aligned} \tag{1.12}$$

leading to two integral equations which can be evaluated to obtain

$$\begin{bmatrix} \frac{1436}{105} & \frac{2783}{420} \\ \frac{2783}{420} & \frac{449}{105} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \frac{32}{15} \\ \frac{21}{10} \end{bmatrix} \quad \rightarrow \quad a_1 = 0.1485,\ a_2 = 0.0154 \tag{1.13}$$

Substituting (1.13) back into (1.11) gives the approximate solution, which is close to the exact solution to the problem.

The reason why the Least Squares Method works is as follows: if the integral (1.8) is zero for the complete set of functions[4]

$$\omega_1 = 1, \quad \omega_2 = x, \quad \omega_3 = x^2, \quad \cdots \quad \omega_n = x^n, \quad \cdots \tag{1.14}$$

then the residual will be identically zero. As one takes higher order polynomial trial functions, the weight functions $\omega_i = \partial R / \partial a_i$ contain higher order terms in x and more and more terms from the complete set of functions (1.14) are included, and the solution is obtained with ever increasing accuracy.

The Collocation Method as a Weighted Residual Method

Re-visiting now the Collocation Method, it can be seen that this is also a weighted residual method, with the weights chosen to be

$$\omega_i(x) = \delta(x - x^i) \tag{1.15}$$

[4] by which is meant that any function can be represented as a linear combination of these functions

where $\delta(x-x^i)=1$ if $x=x^i$ and zero otherwise (the Dirac delta function). For example, the solution (1.4) is derived by considering the integral/equation

$$I=\int_0^1\left(\frac{d^2u}{dx^2}-u+x\right)\delta\left(x-\tfrac{1}{2}\right)dx=\left(\frac{d^2u}{dx^2}-u+x\right)_{x=\frac{1}{2}}=0 \tag{1.16}$$

Symmetry of the Least Squares Method

The coefficient matrix of the Least Squares systems of equations (1.13) is symmetric. This is always desirable, particularly for large systems of equations, since special rapid equation-solver algorithms are available for symmetric coefficient matrices. The Least Squares coefficient matrix is always symmetric provided the differential equation is *linear*. This can be shown as follows: write the differential equation in operator form

$$L[u]=f(x) \tag{1.17}$$

Substituting in the approximation $\tilde{u}=\sum a_i x^i$ leads to, provided L is a linear operator,

$$L[\sum a_i x^i]=\sum a_i L[x^i]=f(x), \tag{1.18}$$

with the weight functions

$$\omega_j(x)=\frac{\partial R}{\partial a_j}=\frac{\partial}{\partial a_j}\left\{\sum a_i L\left[x^i\right]-f(x)\right\}=L\left[x^j\right] \tag{1.19}$$

leading to the system of integral equations, one for each weight,

$$\sum a_i\int L\left[x^i\right]L\left[x^j\right]dx-\int f(x)L\left[x^j\right]dx,\quad j=1,2,\cdots \tag{1.20}$$

For symmetry one requires that the coefficient of a_i in equation j be equal to the coefficient of a_j in equation i, and (1.20) clearly results in a symmetric coefficient matrix.

1.1.3 Galerkin's Method

In Galerkin's Method, the weight functions are chosen through

$$\omega_i = \frac{\partial \tilde{u}}{\partial a_i} \tag{1.21}$$

As with the Method of Least Squares, the higher the order of the approximating polynomial $\tilde{u}$, the higher the order of the terms x^i included in the weight functions, so that the more weight functions are chosen, the more of the complete set of functions (1.14) will be chosen, and the closer the residual will be to zero.

Again considering problem (1.1) and using the trial function which satisfies the boundary conditions, Eqn. 1.3, one has the single weight function

$$\omega = \frac{\partial \tilde{u}}{\partial b} = x - x^2 \tag{1.22}$$

which, when substituted into (1.9) and the integral is evaluated, leads to

$$-\tfrac{11}{30} b + \tfrac{1}{12} = 0 \quad \rightarrow \quad b = \tfrac{5}{22} \tag{1.23}$$

The cubic polynomial satisfying the boundary conditions, Eqn. 1.5, together with the weight functions (with $a_1 = b$, $a_2 = c$)

$$\omega_1 = \frac{\partial \tilde{u}}{\partial a_1} = x - x^3, \quad \omega_2 = \frac{\partial \tilde{u}}{\partial a_2} = x^2 - x^3 \tag{1.24}$$

leads to the system of integral equations

$$\begin{aligned} I_1 &= \int_0^1 \left(2a_2 + x(1 - 7a_1 - 6a_2) - a_2 x^2 + (a_1 + a_2)x^3\right)(x - x^3)\,dx = 0 \\ I_2 &= \int_0^1 \left(2a_2 + x(1 - 7a_1 - 6a_2) - a_2 x^2 + (a_1 + a_2)x^3\right)(x^2 - x^3)\,dx = 0 \end{aligned} \tag{1.25}$$

Evaluating the integrals leads to the system of equations

$$\begin{bmatrix} -\frac{92}{105} & -\frac{137}{420} \\ -\frac{137}{420} & -\frac{1}{7} \end{bmatrix}\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}+\begin{bmatrix} \frac{2}{15} \\ \frac{1}{20} \end{bmatrix}=\begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \rightarrow \quad a_1=\tfrac{69}{473}, \quad a_2=\tfrac{8}{473} \tag{1.26}$$

and hence to the approximate solution

$$\tilde{u}(x)=\frac{1}{473}\left(69x+8x^2-77x^3\right) \tag{1.27}$$

From the definition of the Galerkin weight function (1.21), the trial function can be written in the alternative form

$$\tilde{u}(x)=\omega_1(x)a_1+\omega_2(x)a_2+\cdots, \qquad \omega_i(x)=\frac{\partial\tilde{u}}{\partial a_i} \tag{1.28}$$

This form of the trial function will be used in most of what follows.

Integration by Parts & the Weak Form

The Galerkin method as presented gives reasonably accurate numerical solutions to differential equations. A modified version of the method involves integrating by parts the term inside the weighted integral which contains the highest, second order, term from the differential equation. For the example considered above, this means re-writing (1.8) as

$$\int_0^1\left[\frac{d^2u}{dx^2}\omega-u\omega+x\omega\right]dx=0 \quad \rightarrow \tag{1.29a}$$

$$\int_0^1\left[\frac{du}{dx}\frac{d\omega}{dx}+u\omega-x\omega\right]dx=\left[\frac{du}{dx}\omega\right]_0^1 \tag{1.29b}$$

There are two advantages to integrating by parts:

i) a linear trial function can be used
ii) the Galerkin coefficient matrix is symmetric for certain equations

Before discussing these points, introduce the following terminology. The original differential equation and BC's, Eqn. 1.1, is referred to as the **strong statement** of the problem. The weighted residual equation of (1.29b) is referred to as the **weak statement** of the problem. The terminology weak statement (or **weak problem** or **weak formulation** or **weak form**) is used to mean two different things: it most often used to mean that the

problem is stated in integral form, contrasting with the strong form of the differential equation, which must be satisfied at all points on the interval of interest; in that sense Eqns. 1.29a,b are weak forms. However, the weak form more correctly means that the required differentiability of the solution is of an order less than that in the original differential equation; in that sense Eqn. 1.29a is a strong form whereas Eqn. 1.29b is a weak form. To avoid ambiguity, we will here maintain the terminology "weak form" to mean the form of Eqn. 1.29b.

Regarding (i), it is clear that the second derivative of a linear trial function $\tilde{u}(x) = a + bx$ is zero and that the first term in the equation on the left of 1.29. This is not the case for the weak form, which retains this information. Of course the two coefficients in a linear trial function can be immediately found from the boundary conditions and so the weighted residual (1.29) is not necessary to obtain what is a trivial solution; however, linear trial functions can be used in the Galerkin FEM, as outlined in the next Chapter, and the integration by parts is then essential.

Regarding (ii), the Galerkin coefficient matrix in (1.26) is symmetric, but this is fortuitous – in general, it is not. However, the weak form of (1.29) *is* symmetric {▲Problem 4}. To generalise this, consider the arbitrary linear ODE

$$p(x)\frac{d^2u}{dx^2} + q(x)\frac{du}{dx} + r(x)u = f(x) \tag{1.30}$$

Multiplying by ω and integrating over the complete domain leads to

$$\int \left(pu''\omega + qu'\omega + ru\omega\right)dx = \int f\omega dx \tag{1.31}$$

To integrate the first term by parts, first note that an integration by parts without the ω gives $\int pu''dx = pu' - \int p'u'dx$. This suggests that one adds and subtracts a term to/from Eqn. 1.31:

$$\int \left[\left(pu'' + p'u'\right)\omega - p'u'\omega + qu'\omega + ru\omega\right]dx = \int f\omega dx \tag{1.32}$$

so that an integration by parts of (1.31) gives the weak form

$$\int \left(-pu'\omega' - p'u'\omega + qu'\omega + ru\omega\right)dx = \int f\omega dx - \left[pu'\omega\right] \tag{1.33}$$

The terms on the left, involving u, contribute to the coefficient matrix. Writing $\tilde{u} = \sum a_i \omega_i$ leads to a system of equations, and the relevant terms are:

$$\int \left\{ -p\sum a_i \omega_i' \omega_j' - p'\sum a_i \omega_i' \omega_j + q\sum a_i \omega_i' \omega_j + r\sum a_i \omega_i \omega_j \right\} dx \tag{1.34}$$

For symmetry, this integral should be unchanged if i and j are interchanged. This will be so if $p' = q$, so a second-order equation leading to symmetry is the equation

$$\frac{d}{dx}\left(p(x)\frac{du}{dx} \right) + r(x)u = f(x) \tag{1.35}$$

This is known as the **self-adjoint ODE**. When $p' = q = 0$, one has the equation

$$p_0 \frac{d^2 u}{dx^2} + r(x)u = f(x) \tag{1.36}$$

where p_0 is a constant, and an equation of the form $b(x)u'' + c(x)u = d(x)$ can always be put in the form (1.36) by dividing through by $b(x)$. It can be seen that Eqn. 1.1 is of this form, hence the symmetric matrix of Eqn. 1.2.6.

Consider again now the example problem (1.1), only this time using the weak form of (1.29):

$$\int_0^1 \left[\frac{d\tilde{u}}{dx}\frac{d\omega}{dx} + \tilde{u}\omega - x\omega \right] dx = \left[\frac{d\tilde{u}}{dx}\omega \right]_0^1 \tag{1.37}$$

The quadratic trial function satisfying the BC's, Eqn. 1.3, leads to

$$\left\{ \int_0^1 \left[(1-2x)^2 + \left(x - x^2\right)^2 \right] dx - \left[(1-2x)\left(x - x^2\right) \right]_0^1 \right\} b = \int_0^1 x\left(x - x^2\right) dx \tag{1.38}$$

which gives $b = 10/44$. Moving to the cubic polynomial $\tilde{u} = a_1\omega_1(x) + a_2\omega_2(x)$ with the weights as in (1.24) leads to

$$I_1 = \int_0^1 \left(\frac{du}{dx}\frac{d\omega_1}{dx} + u\omega_1 - x\omega_1 \right) dx - \left[\frac{d\tilde{u}}{dx}\omega_1 \right]_0^1 = 0$$
$$I_2 = \int_0^1 \left(\frac{du}{dx}\frac{d\omega_2}{dx} + u\omega_2 - x\omega_2 \right) dx - \left[\frac{d\tilde{u}}{dx}\omega_2 \right]_0^1 = 0 \qquad (1.39)$$

and the (symmetric) system of equations and solution

$$\begin{bmatrix} \frac{92}{105} & \frac{137}{420} \\ \frac{137}{420} & \frac{1}{7} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \frac{2}{15} \\ \frac{1}{20} \end{bmatrix} \quad \rightarrow \begin{matrix} a_1 = \frac{69}{473} \\ a_2 = \frac{8}{473} \end{matrix} \quad \rightarrow \tilde{u} = \frac{1}{473}\left(69x + 8x^2 - 77x^3\right) \qquad (1.40)$$

This is actually the same system as was obtained without the integration by parts (see Eqns. 1.26, 1.27); in general though, this will not be the case.

A further approximation would be the quartic polynomial

$$\tilde{u} = a + bx + cx^2 + dx^3 + ex^4 \qquad (1.41)$$

giving the trial function satisfying the BC's

$$\tilde{u} = a_1\omega_1(x) + a_2\omega_2(x) + a_3\omega_3(x) \qquad (1.42)$$

with

$$\omega_1 = x - x^4, \quad \omega_2 = x^2 - x^4, \quad \omega_3 = x^3 - x^4 \qquad (1.43)$$

leading to the integrals, symmetric system and solution

$$I_i = \int_0^1 \left(\frac{d\tilde{u}}{dx}\frac{d\omega_i}{dx} + \tilde{u}\omega_i - x\omega_i \right) dx - \left[\frac{d\tilde{u}}{dx}\omega_i \right]_0^1 = 0$$

$$\begin{bmatrix} \frac{88}{63} & \frac{929}{1260} & \frac{769}{2520} \\ \frac{929}{1260} & \frac{4}{9} & \frac{493}{2520} \\ \frac{769}{2520} & \frac{493}{2520} & \frac{113}{1260} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{6} \\ \frac{1}{12} \\ \frac{1}{30} \end{bmatrix}, \quad \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} \frac{14427}{96406} \\ -\frac{539}{96406} \\ -\frac{6041}{48203} \end{bmatrix} \qquad (1.44)$$

$$\rightarrow \tilde{u} = 0.1497x - 0.0056x^2 - 0.1253x^3 - 0.0187x^4$$

Differential Equations with Non-constant Coefficients

Differential equations with non-constant coefficients can be handled as above. As an illustration, consider the problem

$$x^2 \frac{d^2u}{dx^2} - 2u = 1, \qquad u(1) = 0, \quad u(2) = 0 \tag{1.45}$$

[the exact solution is $\frac{1}{14}\left(6/x - 7 + x^2\right)$]

The weighted residual is

$$I = \int_1^2 \left(x^2 \frac{d^2u}{dx^2} - 2u - 1 \right) \omega(x) dx = 0 \tag{1.46}$$

To integrate this type of function by parts, one needs to add and subtract terms (as was done above in Eqns. 1.31-32). First, integrate the second-order term by parts:

$$\int x^2 \frac{d^2u}{dx^2} dx = x^2 \frac{du}{dx} - \int 2x \frac{du}{dx} dx \quad \rightarrow \quad \int \left(x^2 \frac{d^2u}{dx^2} + 2x \frac{du}{dx} \right) dx = x^2 \frac{du}{dx} \tag{1.47}$$

Adding and subtracting this term $2x(du/dx)$ then leads to

$$\begin{aligned} I &= \int_1^2 \left[\left(x^2 \frac{d^2u}{dx^2} + 2x \frac{du}{dx} \right) w - 2x \frac{du}{dx} w - 2u\omega - \omega \right] dx = 0 \\ &\rightarrow \quad \int_1^2 \left[x^2 \frac{du}{dx} \frac{d\omega}{dx} + 2x \frac{du}{dx} w + 2u\omega + \omega \right] dx - \left[x^2 \frac{du}{dx} \omega \right]_1^2 = 0 \end{aligned} \tag{1.48}$$

Using a quadratic trial function which satisfies the BC's,

$$\tilde{u} = a(x-1)(x-2), \qquad \omega(x) = (x-1)(x-2) \tag{1.49}$$

one has

$$\begin{aligned} I &= \int_1^2 \left[x^2 a(2x-3)^2 + 2ax(2x-3)(x-1)(x-2) + 2a(x-1)^2(x-2)^2 + (x-1)(x-2) \right] dx \\ &= \tfrac{5}{6} a - \tfrac{1}{6} = 0 \end{aligned} \tag{1.50}$$

and the solution

$$\tilde{u} = 0.2(x-1)(x-2) = 0.4 - 0.6x + 0.2x^2 \tag{1.51}$$

A higher order trial solution would be $\tilde{u} = a + bx + cx^2 + dx^3$. Application of the BC's gives $\tilde{u} = a_1(x-1)(x-2) + a_2(x-1)(x-2)(x+3)$. Actually, one can just as well use the simpler trial function $\tilde{u} = a_1(x-1)(x-2) + a_2 x(x-1)(x-2)$, which satisfies the BC's and is cubic. This latter function results in the system of equations

$$\begin{bmatrix} \frac{5}{6} & \frac{7}{5} \\ \frac{13}{10} & \frac{241}{105} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{6} \\ \frac{1}{4} \end{bmatrix} \quad \rightarrow \quad \begin{aligned} a_1 &= 0.351027 \\ a_2 &= -0.0898973 \end{aligned} \tag{1.52}$$

and the solution

$$\begin{aligned} \tilde{u} &= 2a_1 + (2a_2 - 3a_1)x + (a_1 - 3a_2)x^2 + a_2 x^3 \\ &= 0.702054 - 1.232877x + 0.620719x^2 - 0.0898973x^3 \end{aligned} \tag{1.53}$$

Note that the coefficient matrix here is not symmetric; this is as expected since (1.45) is not of the form (1.35).

The 1-term and 2-term solutions are plotted below.

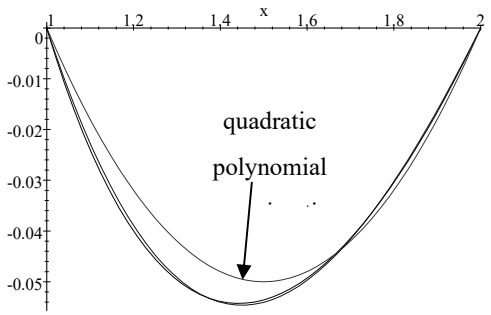


Figure 1.2: Galerkin Method solution to Eqn. 1.45

1.1.4 The Petrov-Galerkin Method

The Petrov-Galerkin method is the most general type of weighted residual method. Basically it is a catch-all term for all weighted residual methods, more specifically for those in which the weight functions are not as chosen in the Least Squares or Galerkin methods. As long as the weight functions embody the complete set of functions (1.14), the method will converge as higher order trial functions are chosen.

1.1.5 Limitations of Weighted Residual Method

The Weighted Residual Methods involve selecting an appropriate trial function to represent the solution. Greater accuracy is achieved by increasing the degree of the approximating trial function. Unfortunately, the resulting coefficient matrix might well become ill-conditioned for polynomials of high degree. For example, shown in Fig. 1.3 is a plot of x^n over [0,1] and it can be seen the curves become very close to each other for larger n, and computer rounding will inevitably mean that these curves will be indistinguishable.

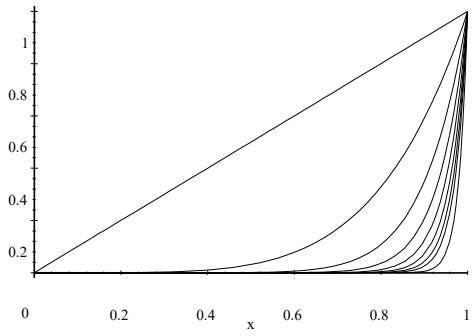


Figure 1.3: A plot of x^n over [0,1]

Further, it might be the case that the actual solution is a highly complex function, perhaps a two or three dimensional function, and that the boundary conditions themselves are complex. In these cases, instead of selecting a trial function to encompass the complete domain, it is better (necessary) to use some other numerical method, the natural extension to Galerkin's method being the Galerkin Finite Element Method, described in the next Chapter.

1.2 Galerkin's Method: Further Applications

So far, Galerkin's method has been used to solve the second order differential equation (1.1). Here, it is shown how the method can be used to solve a wide variety of problems: linear problems with various types of boundary condition, non-linear ODEs and partial differential equations.

Many of the problems considered here are trivial, in that an exact solution can easily be obtained; they are used to illustrate the method, which can be used to tackle more complex problems.

1.2.1 Essential and Natural Boundary Conditions

An **essential** (or **Dirichlet**) boundary condition for a second-order differential equation is one on the unknown u. A **natural** (or **Von Neumann**) boundary condition is one on the first derivative, u'. These boundary conditions can be homogeneous (their value at the boundary is zero) or non-homogeneous. For example, the problem of Eqn. 1.1 involves homogeneous essential BC's.

Homogeneous & Non-Homogeneous Essential BC's

Consider a general problem involving homogeneous essential BC's, $u(0)=0$ and $u(l)=0$. A trial function $\tilde{u}=\sum_{i=0}^{n} a_i x^i$ satisfying these BC's is {▲Problem 8}

$$\tilde{u}=\sum_{i=1}^{n-1} a_i \omega_i , \qquad \omega_i = x^i - \frac{x^n}{l^{n-i}} \tag{1.54}$$

Note that *the weight functions satisfy the essential boundary conditions*, i.e. $w_i(0)=w_i(l)=0$. This is an important property of the weight functions in the Galerkin method; it has to be the case since the coefficients a_i in $\tilde{u}=\sum_{i=1}^{n-1} a_i \omega_i$ are arbitrary and $\tilde{u}$ satisfies the essential BC's.

Consider now the case of non-homogeneous BC's, $u(0)=\bar{u}_0$ and $u(l)=\bar{u}_l$. In this case the trial function is written as $\tilde{u}=\sum_{i=0}^{n} a_i x^i + \beta(x)$. The first part, the sum, again satisfies the essential BC's and the extra term involving β ensures that the non-homogeneous BC's

are satisfied, for example one might let $\beta = (1-x/l)\bar{u}_0 + (x/l)\bar{u}_l$. The weight functions are the same as for the homogeneous BC case, since $\beta(x)$ is independent of the a_i.

An important consequence of the Galerkin formulation is that, when one integrates by parts the second-order term to obtain a boundary term of the form (see, for example, Eqn. 1.37)

$$\left[\frac{\partial u}{\partial x}\omega_j\right]_0^l \tag{1.55}$$

the weight functions are zero at each end and so the *boundary term is zero*.

To illustrate these points, consider the following example problem with non-homogeneous BC's:

$$\frac{d^2u}{dx^2} = 1, \qquad u(0) = 0, \quad u(1) = \tfrac{3}{2} \tag{1.56}$$

[the exact solution is $\frac{1}{2}x^2 + x$]

Forming the weighted residual and integrating by parts to obtain the weak form, one has

$$I = \int_0^1 \left[\frac{du}{dx}\frac{d\omega}{dx} + \omega\right]dx - \left[\frac{du}{dx}\omega\right]_0^1 = 0 \tag{1.57}$$

Choose a quadratic polynomial trial function of the form $\tilde{u}(x) = \sum_{i=0}^{2} a_i x^i + \beta(x)$. Applying the BC's gives

$$\tilde{u} = a(x - x^2) + \tfrac{3}{2}x \tag{1.58}$$

leading to

$$\begin{aligned} I &= \int_0^1 \left\{\left[a(1-2x) + \tfrac{3}{2}\right](1-2x) + \left(x - x^2\right)\right\}dx - \cancel{\left[\frac{du}{dx}\omega\right]_0^1} \\ &= \tfrac{1}{3}a + \tfrac{1}{6} = 0 \\ &\rightarrow a = -\tfrac{1}{2} \qquad \rightarrow \tilde{u} = x + 0.5x^2 \end{aligned} \tag{1.59}$$

Note that this is actually the exact solution; a quadratic trial function was used and the exact solution is quadratic.

Again, re-stating the two important points made: (i) the weight function $x - x^2$ satisfies the essential BC's and (ii) the boundary term in (1.59) is zero.

Natural Boundary Conditions

It is not necessary to have the trial function satisfy the natural boundary conditions – they only have to satisfy the essential BC's (hence the name *essential*). For example, consider the following problem:

$$\frac{d^2u}{dx^2} = 1, \qquad u(0) = 1, \quad \left.\frac{\partial u}{\partial x}\right|_{x=1} = 2 \tag{1.60}$$

[the exact solution is $\frac{1}{2}x^2 + x + 1$]

Choosing a quadratic trial function, $\tilde{u} = a_o + a_1 x + a_2 x^2$, which satisfies the essential BC, one has

$$\tilde{u} = a_1 x + a_2 x^2 + 1, \qquad \omega_1 = x, \quad \omega_2 = x^2 \tag{1.61}$$

There is an extra unknown coefficient and a second weight function, since the natural boundary condition has not yet been applied. As usual, the weight functions satisfy the homogeneous essential BC. The name *natural* is used since these BC's arise naturally in the weak statement of the problem:

$$\begin{aligned} I_j &= \int_0^1 \left[\frac{du}{dx}\frac{d\omega_j}{dx} + \omega_j \right] dx - \left.\frac{du}{dx}\right|_{x=1} \omega_j(1) = 0, \qquad j = 1,2 \\ &= \int_0^1 \left[\frac{du}{dx}\frac{d\omega_j}{dx} + \omega_j \right] dx - u'(1) = 0, \qquad j = 1,2 \end{aligned} \tag{1.62}$$

Substituting in the trial function and evaluating the integrals leads to

$$\begin{aligned} I_1 &= a_1 + a_2 + \tfrac{1}{2} - u'(1) \\ I_2 &= a_1 + \tfrac{4}{3}a_2 + \tfrac{1}{3} - u'(1) \\ &\rightarrow \begin{bmatrix} 1 & 1 \\ 1 & \frac{4}{3} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} \frac{1}{2} - u'(1) \\ \frac{1}{3} - u'(1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{aligned} \tag{1.63}$$

Applying the natural boundary condition finally leads to

$$\begin{bmatrix} 1 & 1 \\ 1 & \frac{4}{3} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} -\frac{3}{2} \\ -\frac{5}{3} \end{bmatrix} = 0, \quad a_1 = 1, a_2 = \frac{1}{2}$$
$$\rightarrow \tilde{u} = 1 + a_1 x + a_2 x^2 = \tfrac{1}{2} x^2 + x + 1 \tag{1.64}$$

which is the exact solution.

Unlike the case of two essential BC's, the boundary term here is non-zero.

A 4th – order ODE

The Galerkin method can be used to deal with equations of higher order. For example, here the above ideas are generalized to solve a fourth-order differential equation:

$$\frac{d^4 u}{dx^4} = 0, \qquad u(0) = 0, u(1) = 1, u'(0) = 0, u'(1) = 2 \tag{1.65}$$

[the exact solution is $u(x) = x^2$]

The weighted residual is

$$I = \int_0^1 \frac{d^4 u}{dx^4} \omega dx = 0 \tag{1.66}$$

and integrating twice by parts gives

$$I = \int_0^1 \frac{d^2 u}{dx^2} \frac{d^2 \omega}{dx^2} dx + \left[\frac{d^3 u}{dx^3} \omega \right]_0^1 - \left[\frac{d^2 u}{dx^2} \frac{dw}{dx} \right]_0^1 = 0 \tag{1.67}$$

In this problem, the essential boundary conditions are (see the ω in the two boundary terms)

$$u \text{ and } \frac{du}{dx} \text{ specified at the end-points} \tag{1.68}$$

and the natural boundary conditions are evidently (again, see the boundary terms)

$$\frac{d^2u}{dx^2} \text{ and } \frac{d^3u}{dx^3} \text{ specified at the end-points} \qquad (1.69)$$

Thus in this example there are four essential boundary conditions and no natural boundary conditions.

Using a quartic trial function then leads to the 1-term approximate solution $\widetilde{u} = 2x^3 - x^4$ {▲Problem 13}, which is plotted in Fig. 1.4.

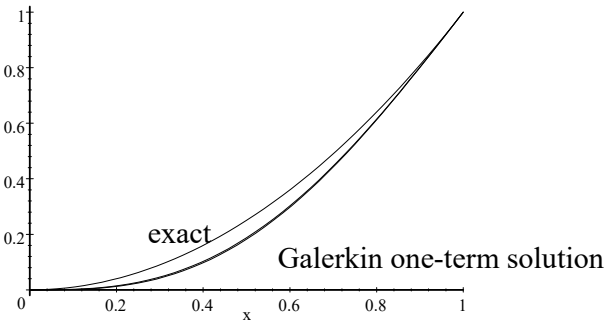


Figure 1.4: Exact and 1-term Galerkin solutions to the 4th-order ODE (1.65)

1.2.2 Non – Linear Ordinary Differential Equations

The solution method for non-linear equations is essentially the same as for linear equations. The new feature here is that one ends up having to solve a system of *non-linear* equations for the unknown coefficients. For example, consider the equation

$$2\frac{du}{dx}\frac{d^2u}{dx^2} + 1 = 0, \qquad u(0) = 0, \quad \left.\frac{\partial u}{\partial x}\right|_{x=1} = 1 \qquad (1.70)$$

[the exact solution is $u(x) = \frac{2}{3}\left(2^{3/2} - (2-x)^{3/2}\right)$]

The weak statement of the problem, after an integration by parts (see the method encompassed in Eqns. 1.46-48), is {▲Problem 15}

$$I = \int_0^1 \left[\left(\frac{du}{dx}\right)^2 \frac{d\omega}{dx} - \omega\right] dx - \left(u'(1)\right)^2 \omega(1) = 0 \qquad (1.71)$$

Using a quadratic trial function satisfying the essential BC, $\tilde{u} = a_1 x + a_2 x^2$, leads to the *non-linear* system of equations {▲Problem 15}

$$\begin{aligned} a_1^2 + 2a_1 a_2 + \tfrac{4}{3} a_2^2 - \tfrac{3}{2} &= 0 \\ a_1^2 + \tfrac{8}{3} a_1 a_2 + 2a_2^2 - \tfrac{4}{3} &= 0 \end{aligned} \tag{1.72}$$

These equations are fairly simple and can actually be solved using elementary elimination methods; there are two possible solutions

$$(a_1, a_2) = (2.2298, -2.1114),\ (1.4241, -0.2051) \tag{1.73}$$

Note, however, that, in general, a system of non-linear equations cannot be solved by elementary elimination methods; numerical methods such as the **Newton-Raphson technique** (see Chapter 5) needs to be employed.

The solution is not unique. Usually, the physics of the problem lets one know which solution to choose. The second solution, $(1.4241, -0.2051)$, is plotted in Fig. 1.5. It is very accurate since the exact solution can be well approximated by a quadratic polynomial.

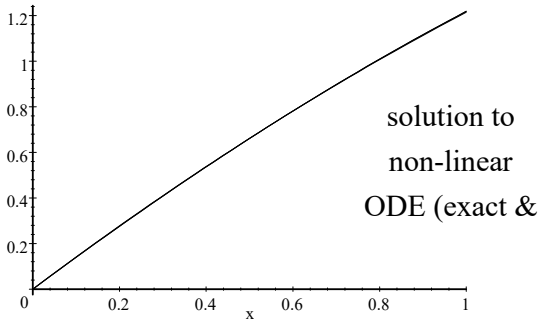


Figure 1.5: Exact and quadratic Galerkin solutions to the non-linear ODE (1.70)

1.2.3 Partial Differential Equations

Galerkin's method can also be used to solve partial differential equations, in particular those containing both time and spatial derivatives. The Galerkin approach is used to

reduce the spatial terms, leaving a system of ordinary differential equations in time. For example, consider the equation

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0 \tag{1.74}$$

subject to

initial conditions: $u(x,0) = \sin \pi x + x$

boundary conditions: $u(0,t) = 0, \quad u(1,t) = 1$

[the exact solution is $u(x,t) = \sin(\pi x)e^{-\pi^2 t} + x$]

Introduce the trial function

$$u(x,t) = a(t)\left(x - x^2\right) + \sin \pi x + x, \qquad a(0) = 0 \tag{1.75}$$

As usual, this trial function explicitly satisfies the essential boundary conditions and the weight function satisfies the two homogeneous essential boundary conditions.

Note that the coefficient *a* is now a function of time whereas the weight function is a function of *x* only. Following the Galerkin procedure,

$$\begin{aligned} \int_0^1 \left(\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} \right) \omega(x) dx &= \int_0^1 \frac{\partial u}{\partial t} \omega dx + \int_0^1 \frac{\partial u}{\partial x} \frac{\partial \omega}{\partial x} dx - \left[\frac{\partial u}{\partial x} \omega \right]_0^1 \\ &= \frac{\partial a}{\partial t} \int_0^1 \omega^2 dx + \int_0^1 \frac{\partial u}{\partial x} \frac{\partial \omega}{\partial x} dx \\ &= \frac{1}{30} \frac{\partial a}{\partial t} + \left(\frac{1}{3} a + \frac{4}{\pi} \right) = 0 \end{aligned} \tag{1.76}$$

This now yields an *ordinary* differential equation in *time*, which must be integrated to obtain the solution:

$$\frac{da}{dt} + 10a + \frac{120}{\pi} = 0, \qquad a(0) = 0 \tag{1.77}$$

giving

$$a(t) = \frac{12}{\pi}\left(e^{-10t} - 1\right) \tag{1.78}$$

The solution is thus

$$u(x,t) = \frac{12}{\pi}\left(e^{-10t} - 1\right)\left(x - x^2\right) + \sin \pi x + x \tag{1.79}$$

The solution is plotted in Fig. 1.6. Note how the perturbation dies away over time, leaving the linear distribution $u = x$, the solution to $u'' = 0$, as the steady state solution.

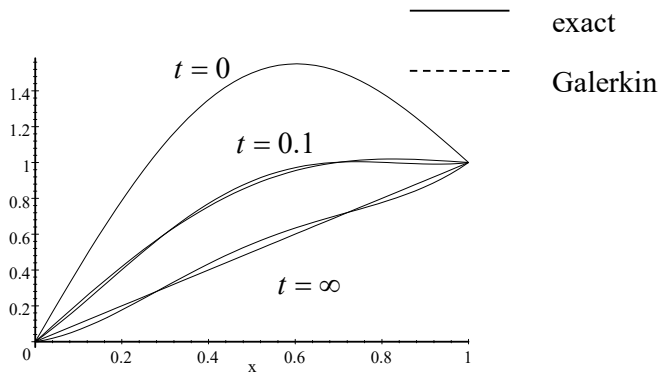


Figure 1.6: Exact and Galerkin solutions to the PDE (1.74)

1.3 The Variational Approach

It has been shown above how one can use the Weighted Residual Methods to reduce differential equations to systems of equations which may be solved to obtain approximate solutions. An alternative approach is to use a variational method. There are many similarities between this and the weighted residual methods, as will be seen.

What follows here is a brief first step into the branch of mathematics known as the **Calculus of Variations**, which is primarily concerned with the minimisation of the values of certain integrals.

First, introduce the concept of the **variation**: consider the function $u(x)$ and a second function

$$\hat{u}(x) = u(x) + \varepsilon\eta(x) \tag{1.80}$$

where $\eta(x)$ is some arbitrary function and ε is an infinitesimal scalar parameter. Thus $\hat{u}(x)$ is everywhere at most infinitesimally close to $u(x)$. The variation of u is the difference between these two, $\varepsilon\eta(x)$, also denoted by $\delta u(x)$, as illustrated in Fig. 1.7:

$$\delta u(x) = \varepsilon\eta(x) = \hat{u}(x) - u(x) \tag{1.81}$$

The variation varies over the domain but it is everywhere infinitesimal, and it is *zero where essential BC's are applied*.

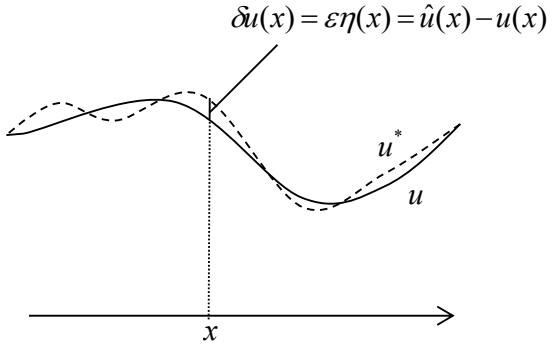


Figure 1.7: The variation of a function u

Note how the variation $\delta u(x)$, which is a small change to u at a *fixed* point x, differs from the increment Δu used in calculus, which is a small change in u due to a small change Δx in x.

Considering again the problem of Eqn. 1.60, multiplying the equation across by the variation of u, integrating over the domain, integrating by parts and using the fact that $\delta u(0) = 0$,

$$\int_0^1 \left[\frac{d^2u}{dx^2}\delta u - \delta u \right] dx \quad \rightarrow \quad \int_0^1 \left[\frac{du}{dx}\frac{d(\delta u)}{dx} + \delta u \right] dx - u'(1)\delta u(1) = 0 \tag{1.82}$$

This is none other than the weak statement (1.62), with the weight function here being the variation. Now, from Eqns. 1.81 and 1.80,

$$\delta\left(\frac{du}{dx}\right)=\frac{d\hat{u}}{dx}-\frac{du}{dx}=\varepsilon\eta'(x),\qquad \frac{d}{dx}(\delta u)=\frac{d}{dx}(\varepsilon\eta(x))=\varepsilon\eta'(x) \tag{1.83}$$

and one has the important identity

$$\delta\left(\frac{du}{dx}\right)=\frac{d}{dx}(\delta u) \tag{1.84}$$

To continue, consider a functional $F(x,u)$, that is a function of another function $u(x)$. When u undergoes a variation δu and changes to $\hat{u}$, F undergoes a consequent change

$$\begin{aligned}\delta F &= F\left(x,u+\delta u\right)-F\left(x,u\right)\\ &=\delta u\frac{\partial F}{\partial u}+O\left(\left(\delta u\right)^2\right)\\ &=\varepsilon\eta\frac{\partial F}{\partial u}+O\left(\varepsilon^2\right)\end{aligned} \tag{1.85}$$

and the **first variation** of F, that is the change in F for small ε, is

$$\delta F=\frac{\partial F}{\partial u}\delta u \tag{1.86}$$

Similarly, with

$$\begin{aligned}\delta\left(F^2\right)&=F^2\left(x,u+\varepsilon\eta\right)-F^2\left(x,u\right)\\ &=\left[F\left(x,u+\varepsilon\eta\right)+F\left(x,u\right)\right]\left[F\left(x,u+\varepsilon\eta\right)-F\left(x,u\right)\right]\end{aligned} \tag{1.87}$$

it follows from Eqns. 1.85-86, that

$$\delta\left(F^2\right)=2F\delta F \tag{1.88}$$

(as in the formula for the ordinary differentiation). Using (1.84) and (1.88), (1.82) becomes

$$\int_0^1\delta\left[\frac{1}{2}\left(\frac{du}{dx}\right)^2+u\right]dx-u'(1)\delta u(1)=0 \tag{1.89}$$

Finally, since

$$\begin{aligned}\int \delta F(x)dx &= \int \left[F(x,u+\varepsilon\eta)-F(x,u)\right]dx \\ &= \int F(x,u+\varepsilon\eta)dx - \int F(x,u)dx, \\ &= \delta\left(\int F dx\right)\end{aligned} \tag{1.90}$$

and $u'(1)$ is constant,

$$\delta W(u) = \delta\left\{\int_0^1 \left[\frac{1}{2}\left(\frac{du}{dx}\right)^2 + u\right]dx - u'(1)u(1)\right\} = 0 \tag{1.91}$$

This is an alternative weak statement to the problem: the variation of the functional $W(u)$, i.e. the function of the function $u(x)$, inside the curly brackets, is zero. The problem is therefore now: find the function $u(x)$ which causes W to be stationary.

Just as the Galerkin method was used to reduce the weighted residual weak statement to a system of equations to be solved for an approximate solution, the variational weak statement can be reduced to a system of equations using the Rayleigh-Ritz method, which is discussed next.

The Rayleigh-Ritz Method

In the Rayleigh-Ritz Method, a trial function satisfying the essential BC's is chosen, say $\tilde{u} = a_1 x + a_2 x^2 + 1$, as in (1.61). The functional in Eqn. 1.91 can then be written as a function of the unknown coefficients; $W\left(u(x)\right) \to W\left(a_1, a_2\right)$:

$$\begin{aligned}W(a_1,a_2) &= \int_0^1 \left[\frac{1}{2}\left(a_1+2a_2x\right)^2 + \left(a_1x+a_2x^2+1\right)\right]dx - 2\left(a_1+a_2+1\right) \\ &= \tfrac{1}{2}a_1^2 + a_1a_2 + \tfrac{2}{3}a_2^2 - \tfrac{3}{2}a_1 - \tfrac{5}{3}a_2 - 1\end{aligned} \tag{1.92}$$

We require that $\delta W\left(a_1, a_2\right) = 0$. With $\delta W = \left(\partial W / \partial a_1\right)\delta a_1 + \left(\partial W / \partial a_2\right)\delta a_2$, we require that $\partial W / \partial a_i = 0$. Evaluating these partial derivatives leads to the system of equations

$$\begin{bmatrix} 1 & 1 \\ 1 & \frac{4}{3} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} -\frac{3}{2} \\ -\frac{5}{3} \end{bmatrix} = 0 \tag{1.93}$$

which is the exact same system as obtained using the Galerkin method (this will always be the case for differential equations of the self-adjoint form, i.e. as in (1.35)).

Here, the variational weak statement (1.90) was derived from the strong differential equation statement of the problem (1.60). It should be noted that in many applications the variational statement appears quite naturally, for example by using the principle of virtual displacements in certain mechanics problems, and that the variational statement can be converted back into the strong statement using the Calculus of Variations.

1.4 Summary

In summary then, in this Chapter have been discussed two broadly different ways of tackling a boundary value problem, the Weighted Residual approach and the Variational approach. The former involves converting the strong differential equation statement of the problem into the weak weighted residual statement of the problem, and solving using the Galerkin or other similar numerical method. The Variational approach involves converting the strong statement of the problem into a stationary functional problem (or perhaps beginning with the stationary problem), and solving numerically using the Rayleigh-Ritz approach. This is summarized in the Fig. 1.8. In the figure, the bold arrows show the method which will be primarily used in this text.

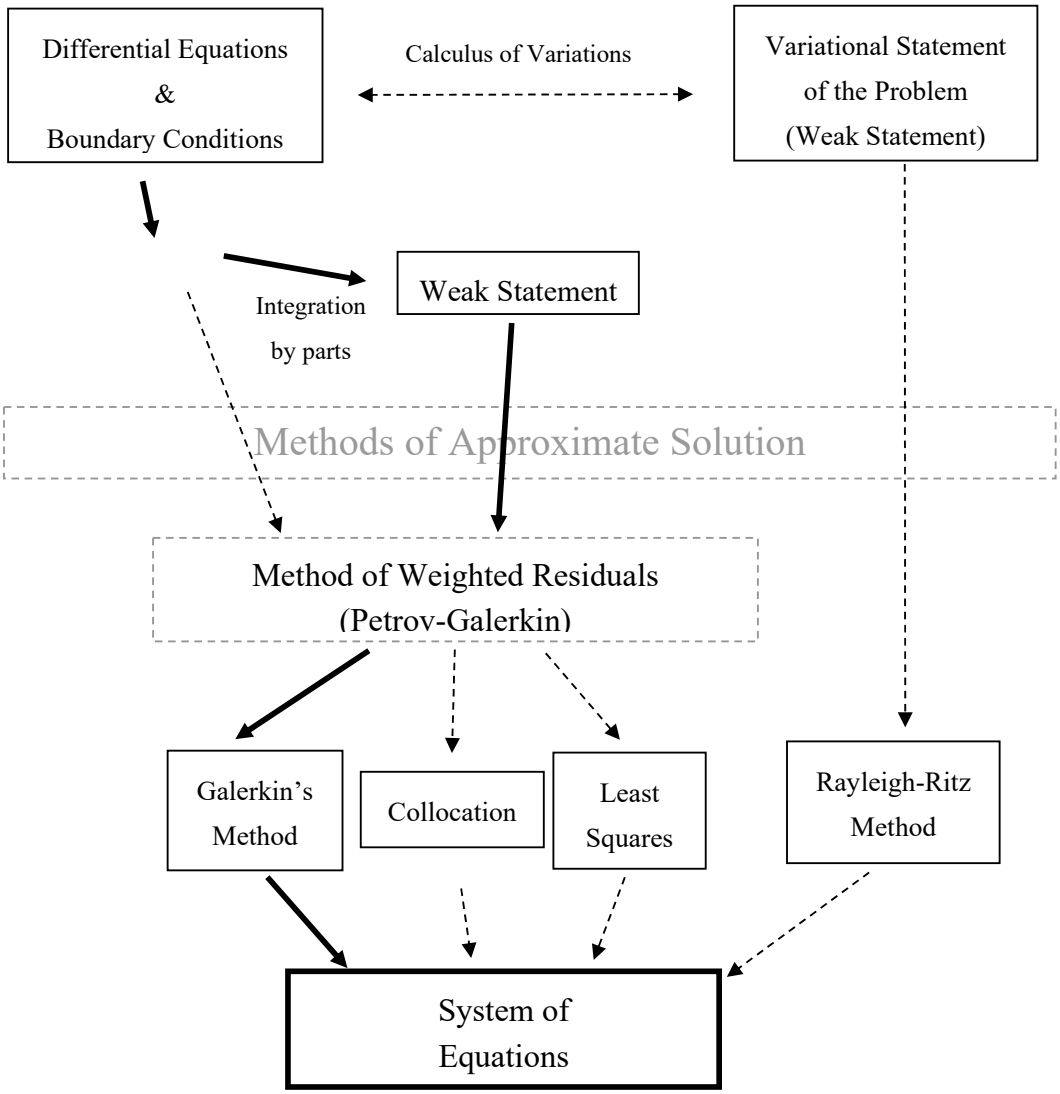


Fig. 1.8. Weighted Residual and Variational Solution of Differential Equations

1.5 Problems

1. The approximate solution (1.4), $\tilde{u}(x) = \frac{2}{9}\left(x - x^2\right)$, to the differential equation (1.1) was obtained by using the collocation method, with (1.1) enforced at $x = 1/2$. What is the solution when the point chosen is $x = 1/4$. Which is the better approximation?
2. Derive the system of equations (1.6), resulting from a cubic trial function for the differential equation (1.1) and the collocation method.
3. From the weighted residual integral (1.9) and the Least Squares relation $\omega(x) = \partial R / \partial b$, derive the equation (1.10) for the unknown coefficient *b*.

4. Show that the weak form of (1.29) leads to a symmetric coefficient matrix when the weight functions are chosen according to the Galerkin Method, Eqn. 1.21 (Do not consider the boundary term).
5. Use the weighted residual methods, (i) Collocation, (ii) Least Squares, (iii) Galerkin (strong form), (iv) Galerkin (weak form), with a quadratic trial function, to solve the following differential equations:
   a) $u'' = 1, \quad u(0) = u(1) = 0$ [exact sln. $u(x) = \frac{1}{2}x(x-1)$]
   b) $u'' + u = 2x, \quad u(0) = u(2) = 0$ [exact sln. $u(x) = 2x - 4\frac{\sin x}{\sin 2}$]

   Which is the most accurate at the mid-point?
6. Use the Galerkin method (weak form) to solve the following ODE with non-constant coefficients:
   $2xu'' + x = 1, \quad u(1) = u(2) = 0$ [exact sln. $u(x) = -\frac{1}{2} + (1-x)\ln 2 + \frac{3}{4}x - \frac{1}{4}x^2 + \frac{1}{2}x\ln x$]

   Use a quadratic trial function. Would the coefficient matrix resulting from a higher order trial function be symmetric?
7. Consider again the problem (1.45) with non-constant coefficients leading to the unsymmetric matrix (1.52). Is it possible to rewrite the equation (1.45) so as to obtain a symmetric coefficient matrix? Re-solve the problem using this equation, with a quadratic trial function.
8. Derive the weight functions in (1.54) for the general one dimensional second-order problem involving the trial function $\tilde{u} = \sum_{i=0}^{n} a_i x^i$ and the homogeneous essential BC's, $u(0) = 0$, $u(l) = 0$.
9. Use the Galerkin method (weak form) to solve the following ODEs with two non-homogeneous BC's:
   a) $u'' = x, \quad u(1) = 1, \quad u(2) = 4$ [exact sln. $\frac{1}{6}x^3 + \frac{11}{6}x - 1$]
   b) $u'' + u' = 0, \quad u(1) = 1, \quad u(3) = 2$ [exact sln. $u(x) = \frac{1}{e^2-1}(2e^2 - 1 - e^{3-x})$]

   Are the boundary terms zero?
10. In the problem (1.56), the exact solution (1.59) was obtained using a quadratic trial function. What happens if one uses a cubic trial function?
11. In the solution of the ODE (1.60), the quadratic trial function which satisfies only the essential BC, $\tilde{u} = a_1 x + a_2 x^2 + 1$, was used. Re-solve the problem using a trial function which explicitly satisfies both the essential *and* the natural BC's. Again, use a quadratic trial function (which will yield the exact solution).
12. Use the Galerkin method (weak form) to solve the following ODE:
    $u'' + x = 0, \quad u(1) = 1, \quad \left.\frac{du}{dx}\right|_{x=2} = 1$ [exact sln. $u(x) = -\frac{11}{6} + 3x - \frac{1}{6}x^3$]
13. Using the weak statement (1.67) and a quartic trial function, obtain the 1-term solution for the 4th-order ODE of (1.65). (You should get the exact solution.)

14. Derive the weak statement (1.71) from the non-linear differential equation (1.70). Hence derive the non-linear system of equations (1.72).
15. Solve the following non-linear equation using Galerkin's method (weak form), first with one unknown coefficients, then with two unknown coefficients:
    $2uu'' + x = 0, \qquad u(0) = 0, \quad u(1) = 0$ [no exact sln.]
16. Re-solve the non-linear equation (1.70) by using a trial function which explicitly includes the natural BC.
17. Use the Rayleigh-Ritz Method to find an approximation to the function $u(x)$, satisfying the essential boundary conditions $u(0) = u(1) = 0$, which renders stationary the functional
    $I(u) = \int_0^1 \left\{ \frac{1}{2}(du/dx)^2 + \frac{1}{2}u^2 - u \right\} dx$ [exact sln. $u(x) = 1 - \frac{1}{1-e}\left(e^x + e^{1-x}\right)$]
    Use a quadratic trial function.

2 The (Galerkin) Finite Element Method

2.1 Approximate Solution and Nodal Values

In order to obtain a numerical solution to a differential equation using the Galerkin Finite Element Method (GFEM), the domain is subdivided into **finite elements**. The function is approximated by piecewise **trial functions** over each of these elements. This is illustrated below for the one-dimensional case, with *linear* functions used over each element, p being the dependent variable.

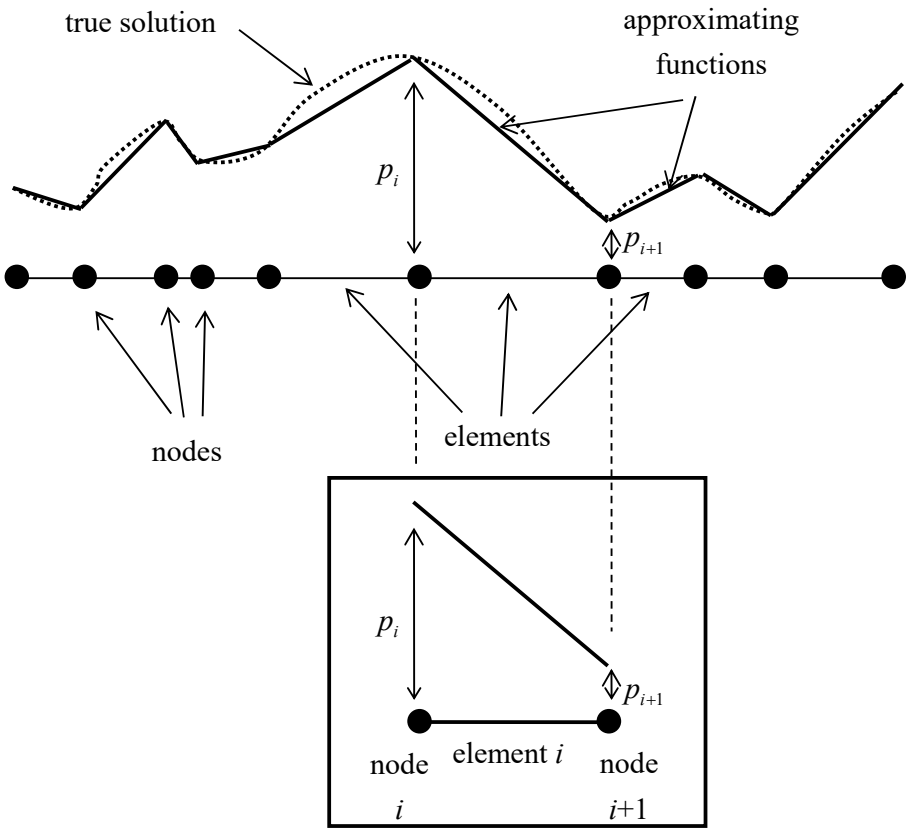


Figure 2.1: A mesh of *N* one dimensional Finite Elements

The unknowns of the problem are the **nodal values** of p, $p_i \;\; i = 1 \ldots N+1$, at the **element boundaries** (which in the 1D case are simply points). The (approximate) solution within each element can then be constructed once these nodal values are known.

2.2 Trial Functions

2.2.1 Lagrange and Hermite Elements

There are an endless number of different trial functions which one can use. In practice, these trial functions can be grouped into two broad types. The first consists of the **Lagrange** or C^0 trial functions (and corresponding Lagrange or C^0 element). These are trial functions which are continuous across element boundaries, but whose first derivatives are not continuous across boundaries. The elements in Fig. 2.1 are C^0 linear elements – there is a clear jump in the first derivative of the trial functions at the element boundaries (nodes) – the first derivative is piecewise continuous.

The second group consists of the **Hermite** or C^1 trial functions (elements). These are functions which are not only continuous across element boundaries, but whose first derivatives are also continuous across boundaries.

In general, a C^n element is one for which the trial functions are continuous up to the nth derivative, but elements with $n > 1$ are rarely used. In fact, three of the most commonly encountered types of element are those with a

(1) linear Lagrange trial function (C^0)
(2) quadratic Lagrange trial function (C^0)
(3) cubic Hermite trial function (C^1)

These three trial functions / elements will be discussed in what follows.

Obviously, the higher the order and the higher the continuity of the element, the better the accuracy one would expect, but the more computation which is required.

2.2.2 The C^0 Linear Element

The C^0 linear element is by far the most commonly used finite element. Consider one typical element of the domain, with end-points x_1, x_2, Fig. 2.2. Assuming a linear interpolation,

$$\tilde{p}(x) = a + bx\,. \tag{2.1}$$

Let the (unknown) end-point values be $\tilde{p}(x_1) = \tilde{p}_1$, $\tilde{p}(x_2) = \tilde{p}_2$. This gives two algebraic equations in two unknowns *a* and *b*,

$$\begin{aligned} \tilde{p}_1 &= a + bx_1 \\ \tilde{p}_2 &= a + bx_2 \end{aligned} \tag{2.2}$$

Solving the equations gives

$$a = \frac{\tilde{p}_1 x_2 - \tilde{p}_2 x_1}{x_2 - x_1}, \quad b = \frac{\tilde{p}_2 - \tilde{p}_1}{x_2 - x_1} \tag{2.3}$$

so that, after some algebra, one can write $\tilde{p}(x)$ in terms of the two unknowns $\tilde{p}_1$, $\tilde{p}_2$ (instead of in the form of Eqn. 2.1, which is in terms of the two values *a* and *b*)

Linear Trial Function:

$$\begin{aligned} \tilde{p}(x) &= N_1(x)\tilde{p}_1 + N_2(x)\tilde{p}_2 \\ N_1(x) &= \frac{x_2 - x}{L}, \quad N_2(x) = \frac{x - x_1}{L} \end{aligned} \tag{2.4}$$

where *L* is the length of the element, $L = x_2 - x_1$.

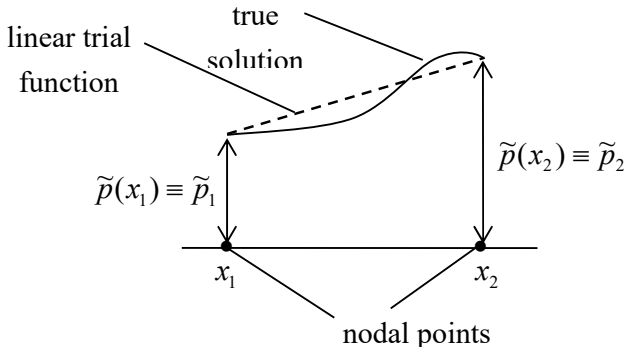


Figure 2.2: Linear trial function approximation over an element

The function $p(x)$ can now be approximated over each interval through

$$\tilde{p}(x) = \begin{array}{ll} N_1(x)\tilde{p}_1 + N_2(x)\tilde{p}_2, & x_1 < x < x_2 \\ N_1(x)\tilde{p}_2 + N_2(x)\tilde{p}_3, & x_2 < x < x_3 \\ \vdots & \\ N_1(x)\tilde{p}_N + N_2(x)\tilde{p}_{N+1}, & x_N < x < x_{N+1} \end{array} \tag{2.5}$$

Here, the **shape** (or **basis**) **functions** N_1, N_2 are the same over each interval (although they don't have to be – they could be interspersed with, for example, quadratic shape functions – see later).

Structure of the Linear Shape Functions

The shape functions, Eqns. 2.4, have a number of interesting properties. Most importantly, they have a value of either 0 or 1 at a node - the variation of the shape functions over an element is shown in Fig. 2.3. A second property of the shape functions is that they sum to 1, $\sum_{i=1}^{2} N_i = 1$.

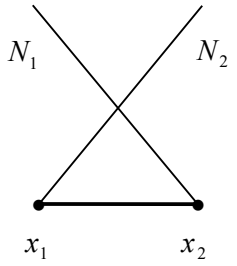


Figure 2.3: Shape functions for the linear trial function

2.3 The Standard Galerkin FEM

The Galerkin FEM for the solution of a differential equation consists of the following steps:

(1) multiply the differential equation by a **weight function** $\omega(x)$ and form the integral over the whole domain
(2) if necessary, integrate by parts to reduce the order of the highest order term

(3) choose the order of interpolation (e.g. linear, quadratic, etc.) and corresponding shape functions $N_i, i = 1 \ldots m$, with trial function $p = \tilde{p}(x) = \sum_{i=1}^{m} N_i(x) p_i$

(4) evaluate all integrals over each element, either exactly or numerically, to set up a system of equations in the unknown p_i's

(5) solve the system of equations for the p_i's.

The linear C^0 element will be used in what follows. Quadratic and cubic elements will be considered later.

2.3.1 A Single-Element Example

Consider the following problem: solve the following differential equation using *one* linear element:

$$\frac{d^2 p}{dx^2} = 0, \quad p'(0) = 1, \quad p(2) = 0 \tag{2.6}$$

[the exact solution is $p(x) = x - 2$]

First, multiply the equation across by $\omega(x)$ and integrating over $[0,2]$ to get the **weighted residual integral**

$$I = \int_0^2 \left(\frac{d^2 p}{dx^2} \right) \omega(x) dx = 0 \tag{2.7}$$

Integrating by parts (and multiplying across by -1) leads to the **weak form**

$$I = \int_0^2 \left(\frac{dp}{dx} \frac{d\omega}{dx} \right) dx - \left[\frac{dp}{dx} \omega \right]_0^2 = 0 . \tag{2.8}$$

This step is crucial to the FEM when linear trial functions are used, since if the trial function $\tilde{p}$ is linear, then $\tilde{p}'' = 0$, and one cannot work with (2.7). However, by first integrating by parts, there is no longer any second derivative and (2.8) is no longer the trivial $0 = 0$.

Choose the linear trial function[1] and, from Eqn. 2.4,

$$\tilde{p}(x) = N_1 p_1 + N_2 p_2 \quad N_1 = 1 - \frac{x}{2} \qquad N_2 = \frac{x}{2} \tag{2.9}$$

Now in the Galerkin FEM, one lets the weight functions simply be equal to the shape functions, i.e. $\omega_i = N_i$, so that[2] $\omega_i = \partial\tilde{p} / \partial p_i$. Thus one has two equations in two unknowns, *one equation for each weight function*. Note also that *the boundary term is not discretised* using (2.9), it is left as dp/dx, so that boundary conditions can be applied (see below),

$$\begin{aligned} \int_0^2 \left(\frac{d\left(N_1 p_1 + N_2 p_2\right)}{dx} \frac{dN_1}{dx} \right) dx - \left[\frac{dp}{dx} N_1 \right]_0^2 = 0 \\ \int_0^2 \left(\frac{d\left(N_1 p_1 + N_2 p_2\right)}{dx} \frac{dN_2}{dx} \right) dx - \left[\frac{dp}{dx} N_2 \right]_0^2 = 0 \end{aligned} \tag{2.10}$$

As mentioned, the shape functions have the following property:

$$\begin{aligned} N_1(0) = 1, \quad N_1(2) = 0 \\ N_2(0) = 0, \quad N_2(2) = 1 \end{aligned} \tag{2.11}$$

i.e. they are zero or one at one of the end-points, and so the boundary terms simplify to

$$\begin{aligned} p_1 \int_0^2 \frac{dN_1}{dx} \frac{dN_1}{dx} dx + p_2 \int_0^2 \frac{dN_2}{dx} \frac{dN_1}{dx} dx + \left[\frac{dp}{dx} \right]_{\text{node 1}} = 0 \\ p_1 \int_0^2 \frac{dN_1}{dx} \frac{dN_2}{dx} dx + p_2 \int_0^2 \frac{dN_2}{dx} \frac{dN_2}{dx} dx - \left[\frac{dp}{dx} \right]_{\text{node 2}} = 0 \end{aligned} \tag{2.12}$$

Substituting in the shape functions and evaluating the integrals leads to the equations

$$\frac{1}{2} \begin{bmatrix} +1 & -1 \\ -1 & +1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} -p'(0) \\ +p'(2) \end{bmatrix} \tag{2.13}$$

---

[1] this example is similar to the Galerkin examples of Chapter 1, the only difference being that here the unknowns in the trial function are the end-point values, rather than the *a*, *b* of (2.1)

[2] this is by far the most commonly used version of the FEM. However, there are some problems which are not well solved using the Galerkin FEM; in these cases other variations of the FEM can be used in which the weight functions chosen are not simply the shape functions

Applying the essential boundary condition $p(2)=0$, the first equation with the natural BC $p'(0)=1$ gives $\frac{1}{2}p_1=-p'(0)=-1\rightarrow p_1=-2$. From the second equation, one finds that $p'(2)=1$. The full solution is

$$p=(1-x/2)p_1+(x/2)p_2=x-2 \tag{2.14}$$

Note that the solution is exact – because the solution was assumed to be linear, and it is.

Consider now the general ordinary differential equation with constant coefficients:

$$a\frac{d^2u}{dx^2}+b\frac{du}{dx}+cu=d \tag{2.15}$$

The weighted residual integral is

$$I=\int_{x_1}^{x_2}\left(a\frac{du}{dx}\frac{d\omega}{dx}-b\frac{du}{dx}\omega-cu\omega\right)dx+\int_{x_1}^{x_2}d\,\omega dx-a\left[\frac{du}{dx}\omega\right]_{x_1}^{x_2}=0. \tag{2.16}$$

Using the linear trial function, and after some algebra, one arrives at the system of two equations

Equations for Linear Trial Function:

$$\left\{a\frac{1}{L}\begin{bmatrix}+1 & -1\\ -1 & +1\end{bmatrix}-b\frac{1}{2}\begin{bmatrix}-1 & +1\\ -1 & +1\end{bmatrix}-c\frac{L}{6}\begin{bmatrix}+2 & +1\\ +1 & +2\end{bmatrix}\right\}\begin{bmatrix}\tilde{u}_1\\ \tilde{u}_2\end{bmatrix}+d\frac{L}{2}\begin{bmatrix}1\\ 1\end{bmatrix}=a\begin{bmatrix}-u'(x_1)\\ +u'(x_2)\end{bmatrix} \tag{2.17}$$

As an example, let $a=1, b=-2, c=d=1$ and $x_1=0, x_2=1, L=1$. The exact solution for the boundary conditions $u(x_1)=\tilde{u}_1=1, u'(x_2)=2$ is $u(x)=1+xe^{x-1}$. Putting these values directly into the linear equations Eqn. 2.17 immediately yields $\tilde{u}_2=2.2$ and $u'(x_1)=0.2$ (compared with the exact solutions 2 and 0.368 respectively) so

$$\tilde{u}(x)=\tilde{u}_1(1-x)+\tilde{u}_2x=1+1.2x \tag{2.18}$$

This solution is compared with the exact solution in Fig. 2.4.

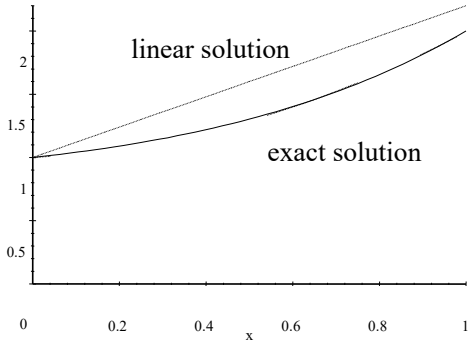


Figure 2.4: Single Linear C^0 Element Solution to Eqn. 2.15

2.3.2 Global and Local Formulations of the FEM

There are two ways in which the FEM can be formulated – the **global** and **local** formulations. In what follows, a simple example will be examined using both formulations; the emphasis here is on how the **global stiffness matrix** is formed, and how the boundary conditions are applied.

Consider the differential equation

$$\frac{d^2p}{dx^2} + 1 = 0, \qquad p(0) = 1, \quad p'(2) = 1 \tag{2.19}$$

[the exact solution is $p(x) = 1 + 3x - \frac{1}{2}x^2$]

The weighted residual, after an integration by parts, is

$$I = \int_0^2 \left(\frac{dp}{dx}\frac{d\omega}{dx} - \omega \right) dx - \left[\frac{dp}{dx}\omega \right]_0^2 = 0 \tag{2.20}$$

This equation will now be solved using two finite elements.

Solution: The Global Formulation

In the global formulation, one uses a single trial function which extends over the *complete domain*[3],

$$\tilde{p}(x) = \tilde{N}_1(x)p_1 + \tilde{N}_2(x)p_2 + \tilde{N}_3(x)p_3 \tag{2.21}$$

and these shape functions are shown below.

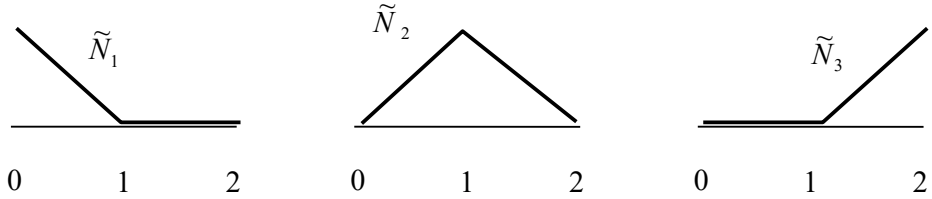


Figure 2.5: Linear shape functions

$$\tilde{N}_1 = \begin{cases} 1-x & 0 \le x < 1 \\ 0 & 1 \le x < 2 \end{cases} \qquad \tilde{N}_2 = \begin{cases} x & 0 \le x < 1 \\ 2-x & 1 \le x < 2 \end{cases} \qquad \tilde{N}_3 = \begin{cases} 0 & 0 \le x < 1 \\ x-1 & 1 \le x < 2 \end{cases} \tag{2.22}$$

Three weighted residuals are now formed, one for each shape function, leading to

$$\begin{aligned}
&p_1\left[+\int_0^1 dx\right] + p_2\left[-\int_0^1 dx\right] + \left[-\int_0^1 (1-x)dx\right] && + p'(0) = 0 \quad \left(\omega = \tilde{N}_1\right) \\
&p_1\left[-\int_0^1 dx\right] + p_2\left[+\int_0^1 dx\right] + \left[-\int_0^1 x dx\right] && \\
&\qquad + p_2\left[+\int_1^2 dx\right] + p_3\left[-\int_1^2 dx\right] + \left[-\int_1^2 (2-x)dx\right] && = 0 \quad \left(\omega = \tilde{N}_2\right) \\
&\qquad + p_2\left[-\int_1^2 dx\right] + p_3\left[+\int_1^2 dx\right] + \left[-\int_1^2 (x-1)dx\right] && - p'(2) = 0 \quad \left(\omega = \tilde{N}_3\right)
\end{aligned} \tag{2.23}$$

and the system of equations

[3] as in the standard Galerkin Method of Chapter 1

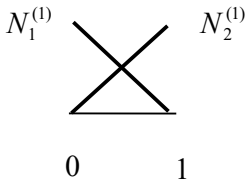
$$\begin{bmatrix} +1 & -1 & 0 \\ -1 & +2 & -1 \\ 0 & -1 & +1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} - p'(0) \\ 1 \\ \frac{1}{2} + p'(2) \end{bmatrix} \tag{2.24}$$

Global Stiffness Matrix

Solution: The Local Formulation

The local or *element* viewpoint is the traditional approach in engineering and is the more useful one when it comes to coding equations. Here, the calculations are done for each element separately:

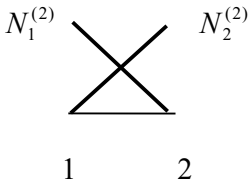
Element 1



$$N_1^{(1)} = 1 - x$$
$$N_2^{(1)} = x$$
$$\tilde{p}^{(1)} = (1-x)p_1 + xp_2$$

$$\begin{bmatrix} +1 & -1 \\ -1 & +1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} - p'(0) \\ \frac{1}{2} + p'(1) \end{bmatrix}$$

Element 2



$$N_1^{(2)} = 2 - x$$
$$N_2^{(2)} = x - 1$$
$$\tilde{p}^{(2)} = (2-x)p_2 + (x-1)p_3$$

$$\begin{bmatrix} +1 & -1 \\ -1 & +1 \end{bmatrix} \begin{bmatrix} p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} - p'(1) \\ \frac{1}{2} + p'(2) \end{bmatrix}$$

(2.25)

$$\begin{bmatrix} +1 & -1 & 0 \\ -1 & +1 & 0 \\ \hline 0 & +1 & -1 \\ 0 & -1 & +1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} - p'(0) \\ \frac{1}{2} + p'(1) \\ \hline \frac{1}{2} - p'(1) \\ \frac{1}{2} + p'(2) \end{bmatrix}$$

boundary terms cancel at internal boundaries

the full global matrix is constructed from the individual element matrices – from the global formulation, it can be seen that one must sum the contributions for common nodes

$$\begin{bmatrix} +1 & -1 & 0 \\ -1 & +2 & -1 \\ 0 & -1 & +1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} - p'(0) \\ 1 \\ \frac{1}{2} + p'(2) \end{bmatrix}$$

Boundary Conditions

Apply now the BC's $p(0) = 1,\ p'(2) = 1$ to get

$$\begin{bmatrix} 2 & -1 \\ -1 & +1 \end{bmatrix} \begin{bmatrix} p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} 2 \\ \frac{3}{2} \end{bmatrix}, \qquad p'(0) = p_2 - \tfrac{1}{2} \tag{2.26}$$

which can be solved for

$$p_2 = \tfrac{7}{2},\ p_3 = 5, \quad p'(0) = 3 \tag{2.27}$$

The solution here happens to be exact at the nodes – this will be the case for problems of the form $p'' = f(x)$.

Returning now to the trial functions (2.21) or (2.25), the full solution, graphed in Fig. 2.6, is

Element 1: $\widetilde{p}^{(1)} = 1 + \frac{5}{2}x$
Element 2: $\widetilde{p}^{(2)} = 2 + \frac{3}{2}x$

Note that, when programming the FE, it is best if rows and columns of the coefficient matrix are not eliminated as in going from (2.24) to (2.26), when applying boundary conditions. The essential BC corresponds to node 1, so we replace the first row with the essential BC. If one wants to preserve a symmetric coefficient matrix, one can also then replace the first column with zeros (and a 1). In this way, application of the boundary conditions to the above system of 3 equations would lead to (note how the "–1" in the first column, second row, of the original coefficient matrix, is brought over to the right hand side, changing the 1 to a 2)

$$\begin{bmatrix} +1 & 0 & 0 \\ 0 & +2 & -1 \\ 0 & -1 & +1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ \frac{3}{2} \end{bmatrix} \tag{2.28}$$

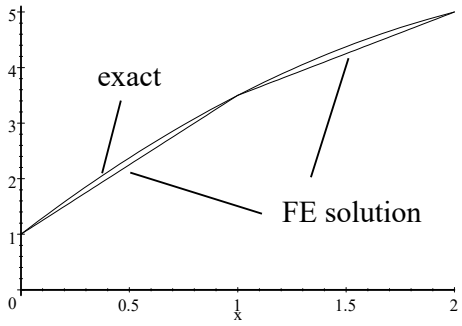


Figure 2.6: Two (Linear C^0) Element Solution to the ODE (2.19)

The First Derivative

The solution for the primary variable p is more accurate than the solution for the derivative p'; p was approximated by a linear C^0 function, so here p' can only be approximated as constant over each element, and will be discontinuous at element boundaries:

Element 1: $\widetilde{p}'^{(1)} = \frac{5}{2}$
Element 2: $\widetilde{p}'^{(2)} = \frac{3}{2}$

The amount of discontinuity can be used as a guide to the inaccuracy of the overall FE solution.

A **smoothed solution** for the first derivative, plotted in Fig. 2.7, can be obtained as follows: take the average of the two solutions at the boundary point, so set $p'(1) = \left(\frac{5}{2} + \frac{3}{2}\right)/2 = 2$, and also use the values at the end points, $p'(2) = 1$ (specified) and $p'(0) = 3$ (evaluated from the FE equations 2.24), then join them up linearly. In this simple example, the smoothed solution actually equals the exact solution. Note that although the smoothed solution looks good, it tends to hide the inaccuracy of the solution. FE software typically outputs smoothed results by default – one should remove this option if one is interested in examining the accuracy/reliability of FE software solutions.

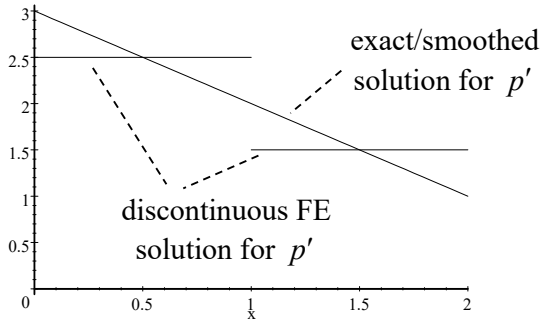


Figure 2.7: FEM solution for the Derivative using Linear Elements

Note the following:

(1) the solution can be made more accurate by either (or both)
   (i) dividing the domain into more elements (the ***h*-method**)
   (ii) selecting higher order elements (e.g. quadratic elements) (the ***p*-method**)

(2) This mesh has three **degrees of freedom**: there are three nodes, and each node has a single degree of freedom (there is only one variable, p, associated with each node). The final system of equations to be solved will be of size $n \times n$, where n is the number of degrees of freedom.

(3) It is not difficult to reformulate the problem with elements of unequal length

(4) The coefficient (Global Stiffness) matrix for this problem, in (2.24), is singular. Thus if one puts two natural boundary conditions into the system of equations one cannot obtain a solution.

2.4 Adaptive Meshing

The difference between the discontinuous FE solution for p' and the smoothed solution allows many FE softwares to automatically refine the mesh in regions where the accuracy is not good[4]. The FE and smoothed solutions are first obtained: $(p')_{\text{FE}}$, $(p')_{\text{smoothed}}$. The "error" is then $(p')_{\text{FE}} - (p')_{\text{smoothed}}$. Measures of these over each element can be determined by integrating over the element, say $\int_{L_i} \left[(p')_{\text{FE}} \right]^2 dx$, $\int_{L_i} \left[(p')_{\text{smoothed}} \right]^2 dx$ and the

[4] that is supposing that the first derivative *shouldn't* be discontinuous – it may well be discontinuous, in certain problems, for example where a stress field is discontinuous at the interface between different materials

element error (squared) $e_i = \int_{L_i} \left[(p')_{\text{FE}} - (p')_{\text{smoothed}} \right]^2 dx$. An example of a global relative error η would then be

$$\eta = \sqrt{\frac{\sum e_i}{\sum \left\{ \int_{L_i} \left[(p')_{\text{FE}} \right]^2 dx \right\} + \sum e_i}} \tag{2.29}$$

η is a global parameter – it doesn't measure error at a particular point/element (element mesh-adaptation parameters might be unreliable, since they would be large simply if $\int_{L_i} \left[(p')_{\text{FE}} \right]^2 dx$ were small). An algorithm might then be: evaluate η; if $\eta < 0.05$ terminate, otherwise refine the mesh in regions where e_i is large; stop after, say, 5 iterations. Note that neither η nor e can reveal the percentage error in the analysis, since the true solution is unknown.

2.5 Local Coordinate Systems

It is convenient to rewrite the FE expressions in terms of a **local** (or **natural**) **coordinate system** ξ. Here, all elements will be normalised over the interval $\xi = [-1,+1]$[5]. To this end, let

Change of Variable to Local Coordinates:

$$x = \frac{1}{2}(1-\xi)x_i + \frac{1}{2}(1+\xi)x_{i+1}, \qquad \xi = \frac{2(x-x_i)}{L} - 1, \qquad J \equiv \frac{dx}{d\xi} = \frac{L}{2} \tag{2.30}$$

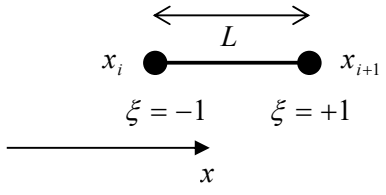


Figure 2.8: Local coordinates over an element

[5] here, the coordinate system used is $\xi = [-1,+1]$. Note that the interval $\xi = [0,+1]$ is sometimes used

From Eqn. 2.4,

Linear Shape Functions (Local Coordinates)

$$N_1(\xi)=\frac{1}{2}(1-\xi),\quad N_2(\xi)=\frac{1}{2}(1+\xi) \tag{2.31}$$

This change of coordinates results in integrals which appear again and again in different FE problems, and they only have to be evaluated on a *once-and-for-all basis*. A number of this type of important integral are evaluated in the Appendix to this chapter. In higher dimensional problems (2-D and 3-D), this normalisation allows one to obtain approximate solutions to the integrals using numerical integration rules.

As an example of using the local coordinate system, consider this problem: solve the following differential equation using two linear elements of equal length:

$$\frac{d^2p}{dx^2}+\frac{dp}{dx}=1,\quad p(1)=1,\quad \left.\frac{dp}{dx}\right|_{x=2}=2 \tag{2.32}$$

[the exact solution is $p(x)=e+x-e^{2-x}$]

One has

$$I=\int_1^2\left(\frac{dp}{dx}\frac{d\omega}{dx}-\frac{dp}{dx}\omega+w\right)dx-\left[\frac{dp}{dx}\omega\right]_1^2=0\,. \tag{2.33}$$

In the global coordinate system,

$$\begin{aligned}
&\underline{\text{Element 1:}}\quad \sum_{i=1}^{2}p_i\int_0^L\frac{dN_i^{(1)}}{dx}\frac{dN_j^{(1)}}{dx}dx-\sum_{i=1}^{2}p_i\int_0^L\frac{dN_i^{(1)}}{dx}N_j^{(1)}dx+\int_0^L N_j^{(1)}dx+\delta_{j1}p'(1)\\
&\underline{\text{Element 2:}}\quad \sum_{i=1}^{2}p_{i+1}\int_L^{2L}\frac{dN_1^{(2)}}{dx}\frac{dN_j^{(2)}}{dx}dx-\sum_{i=1}^{2}p_{i+1}\int_L^{2L}\frac{dN_1^{(2)}}{dx}N_j^{(2)}dx+\int_L^{2L}N_j^{(2)}dx-\delta_{j2}p'(2)
\end{aligned} \tag{2.34}$$

with $j=1,2$. Changing to local coordinates, note first the shape functions,

$$\underline{\text{Element 1}}:\ p^{(1)} = N_1^{(1)} p_1 + N_2^{(1)} p_2 = \left(\frac{1-\xi^{(1)}}{2}\right) p_1 + \left(\frac{1+\xi^{(1)}}{2}\right) p_2 \qquad \left(\xi^{(1)} = 2\,\tfrac{x-1}{L} - 1\right)$$

$$\underline{\text{Element 2}}:\ p^{(2)} = N_1^{(2)} p_2 + N_2^{(2)} p_3 = \left(\frac{1-\xi^{(2)}}{2}\right) p_2 + \left(\frac{1+\xi^{(2)}}{2}\right) p_3 \quad \left(\xi^{(2)} = 2\,\tfrac{x-3/2}{L} - 1\right) \tag{2.35}$$

with $L = \frac{1}{2}$. Using the chain rule, $dN/dx = \left(dN/d\xi\right)\left(d\xi/dx\right)$ and, from Eqn. 2.30, $d\xi/dx = 2/L$, this leads to

$$\begin{aligned} &\sum_{i=1}^{2} p_i \left\{ \left[\frac{2}{L} \int_{-1}^{+1} \frac{dN_i}{d\xi} \frac{dN_j}{d\xi} d\xi \right] - \left[\int_{-1}^{+1} \frac{dN_i}{d\xi} N_j d\xi \right] \right\} + \left[\frac{L}{2} \int_{-1}^{+1} N_j d\xi \right] + \delta_{j1} p'(1) \\ &\sum_{i=1}^{2} p_i \left\{ \left[\frac{2}{L} \int_{-1}^{+1} \frac{dN_i}{d\xi} \frac{dN_j}{d\xi} d\xi \right] - \left[\int_{-1}^{+1} \frac{dN_i}{d\xi} N_j d\xi \right] \right\} + \left[\frac{L}{2} \int_{-1}^{+1} N_j d\xi \right] - \delta_{j2} p'(2) \end{aligned} \tag{2.36}$$

The square bracketed terms/integrals are evaluated in the Appendix to this Chapter so that

Element 1:
$$\begin{aligned} \tfrac{5}{2} p_1 - \tfrac{5}{2} p_2 + \tfrac{1}{4} + p'(1) &= 0 \\ -\tfrac{3}{2} p_1 + \tfrac{3}{2} p_2 + \tfrac{1}{4} &= 0 \end{aligned}$$

Element 2:
$$\begin{aligned} \tfrac{5}{2} p_2 - \tfrac{5}{2} p_3 + \tfrac{1}{4} &= 0 \\ -\tfrac{3}{2} p_2 + \tfrac{3}{2} p_3 + \tfrac{1}{4} - p'(2) &= 0 \end{aligned}$$

and the following system of three equations is obtained:

$$\begin{bmatrix} +\frac{5}{2} & -\frac{5}{2} & 0 \\ -\frac{3}{2} & +4 & -\frac{5}{2} \\ 0 & -\frac{3}{2} & +\frac{3}{2} \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} -\frac{1}{4} - p'(1) \\ -\frac{1}{2} \\ -\frac{1}{4} + p'(2) \end{bmatrix} \tag{2.37}$$

The essential boundary condition is applied at 1 and the natural boundary condition at 2, which leads to

$$\begin{bmatrix} 4 & -\frac{5}{2} \\ -\frac{3}{2} & +\frac{3}{2} \end{bmatrix} \begin{bmatrix} p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{7}{4} \end{bmatrix} \ \rightarrow \ \begin{bmatrix} p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} \frac{47}{18} \\ \frac{34}{9} \end{bmatrix} = \begin{bmatrix} 2.6111 \\ 3.7777 \end{bmatrix} \tag{2.38}$$

This can be compared to the exact solution, which is $p_2 = 2.5696,\ p_3 = 3.7183$. The complete solution (plotted in Fig. 2.9) is then

Element 1: $p^{(1)} = (3-2x)p_1 + (2x-2)p_2 = -\frac{20}{9} + \frac{29}{9}x$

Element 2: $p^{(2)} = (4-2x)p_2 + (2x-3)p_3 = -\frac{8}{9} + \frac{21}{9}x$

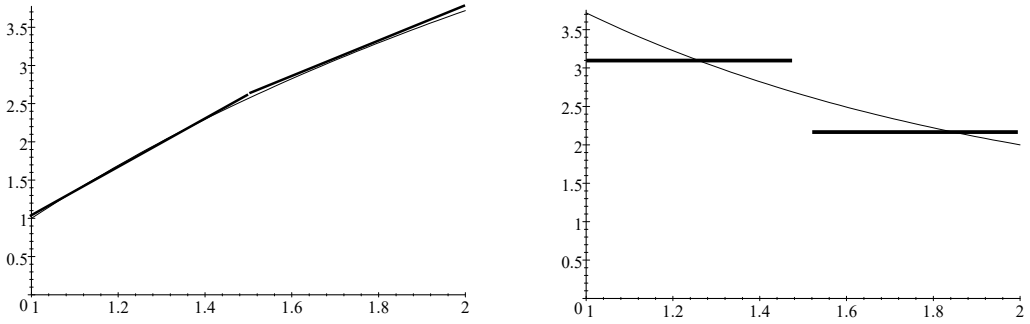


Figure 2.9: FEM solution for the ODE in (2.32) – left: *p*, right: d*p*/d*x*

Note that the error at $x = 2$ reduces as more elements are taken. The error is shown here:

L	*N*	*Error*
1	1	0.28172
0.5	2	0.05949
0.25	4	0.01433

When the element size is halved, the error is reduced by a factor of approximately 4. This general convergence behaviour occurs for linear differential equations and is explained in section 2.6.1 below.

Finally, consider the following general differential equation with constant coefficients using *n* linear elements of length L_i, $i = 1 \ldots n$, Fig. 2.10:

$$a\frac{d^2p}{dx^2} + b\frac{dp}{dx} + cp = d \tag{2.39}$$

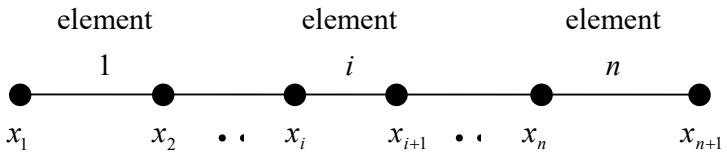


Figure 2.10: a 1-D FE mesh

The weighted residual integral for element *i* is

$$I = \int_{x_1}^{x_{n+1}} \left(a\frac{dp}{dx}\frac{d\omega}{dx} - b\frac{dp}{dx}\omega - cpw \right) dx + \int_{x_1}^{x_{n+1}} dw dx - a\left[\frac{dp}{dx}\omega \right]_{x_1}^{x_{n+1}} = 0 . \qquad (2.40)$$

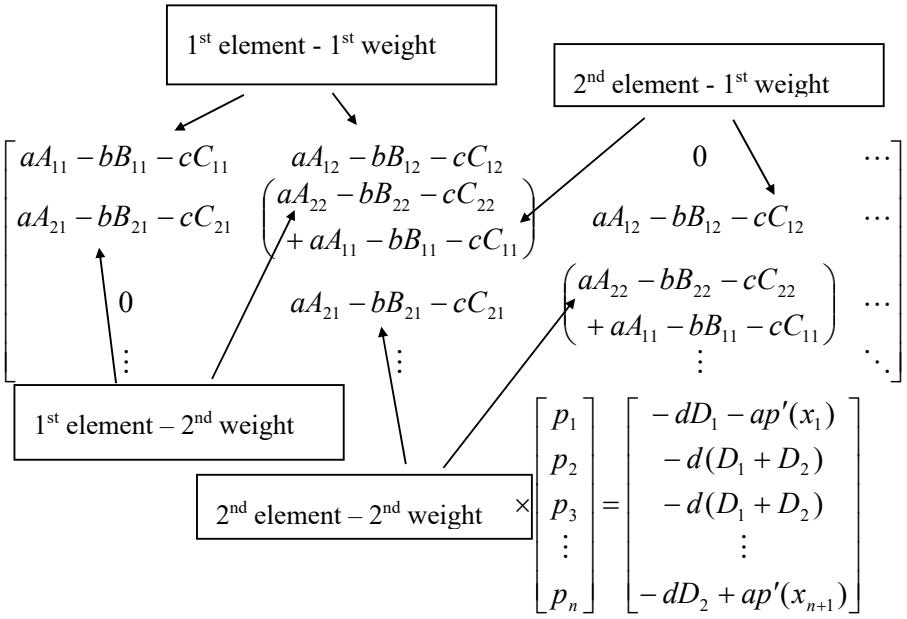
The integral is subdivided into two separate integrals: the first will give rise to matrices, the second to column vectors. Changing to local coordinates leads to

$$p_i \left\{ aA_{j1} - bB_{j1} - cC_{j1} \right\} + p_{i+1} \left\{ aA_{j2} - bB_{j2} - cC_{j2} \right\} + dD_j + \delta_{j1} p'(x_1) - \delta_{jn} p'(x_{n+1}) \qquad (2.41)$$

for element *i* and weight *j*, with (see the Appendix to this Chapter)

$$\begin{aligned} A_{jm} &= \frac{2}{L_i}\int_{-1}^{+1} \frac{dN_j}{d\xi}\frac{dN_m}{d\xi} d\xi = \frac{1}{L_i}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, & B_{jm} &= \int_{-1}^{+1} N_j \frac{dN_m}{d\xi} d\xi = \frac{1}{2}\begin{bmatrix} -1 & +1 \\ -1 & +1 \end{bmatrix} \\ C_{jm} &= \frac{L_i}{2}\int_{-1}^{+1} N_j N_m d\xi = \frac{L_i}{6}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, & D_j &= \frac{L_i}{2}\int_{-1}^{+1} N_j d\xi = \frac{L_i}{2}\begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{aligned} \qquad (2.42)$$

The final system of equations are therefore as shown here.



$$\begin{bmatrix} aA_{11}-bB_{11}-cC_{11} & aA_{12}-bB_{12}-cC_{12} & 0 & \cdots \\ aA_{21}-bB_{21}-cC_{21} & \begin{pmatrix} aA_{22}-bB_{22}-cC_{22} \\ +aA_{11}-bB_{11}-cC_{11} \end{pmatrix} & aA_{12}-bB_{12}-cC_{12} & \cdots \\ 0 & aA_{21}-bB_{21}-cC_{21} & \begin{pmatrix} aA_{22}-bB_{22}-cC_{22} \\ +aA_{11}-bB_{11}-cC_{11} \end{pmatrix} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ \vdots \\ p_n \end{bmatrix} = \begin{bmatrix} -dD_1 - ap'(x_1) \\ -d(D_1+D_2) \\ -d(D_1+D_2) \\ \vdots \\ -dD_2 + ap'(x_{n+1}) \end{bmatrix} \tag{2.43}$$

Note that the system matrix is **tridiagonal** meaning only the diagonal and adjacent elements are non-zero. Special efficient sparse-matrix algorithms can be used to solve such systems of equations.

The complete system 2.43 of equations is of the form

$$\mathbf{K}\,\mathbf{u} = \mathbf{F} \tag{2.44}$$

As mentioned earlier, the coefficient matrix **K** is called the **global stiffness matrix**; the right-hand side **F** is usually called the **force vector**. This terminology arises from the formulation of structural mechanics problems in which the unknown variables are the nodal displacements **u**.

2.6 Approximation Error in the Finite Element Method

The first step in solving real world problems with FEM is to construct a mathematical model to represent a physical system. The question as to how well the ideal mathematical model represents reality is the domain of physics and validation. Errors associated with the FEM are those which arise when one constructs a computational model/representation of

the mathematical model and solve that computational model. Apart from bugs in software and "mistakes" made by the user, the main sources of error in the FEM are **discretization error** and **solution error**.

The discretization error arises when the mathematical model is converted into a discrete computational model. It is the error introduced when representing a function of a continuous variable by its values at a discrete set of nodes; or, equivalently, when representing a differential equation by its FEM matrix equations. This error will depend on the element order (linear, quadratic, etc.) and also on the number of elements used to represent the domain of interest. One would expect that the discretization error will tend to zero as the mesh of elements is made smaller by reducing element size.

The discretization error is associated with the complete problem and so is a global measure. The discretization error itself depends on the more local element level on the **interpolation error**. This is the error which arises when a function is approximated using shape functions over a particular element, and is discussed further below[6].

Other errors which can arise are those associated with numerical integration (all integrals are evaluated exactly in this chapter, but only approximate values can be obtained in higher dimensions), and numerical/rounding errors, for example in inverting the stiffness **K** matrix.

2.6.1 Interpolation Error

The **interpolation error** is usually the largest source of error in the FEM. It can be calculated for any typical element; for example, with the linear element, one can proceed as follows:

The FE approximation is given by

$$\tilde{p}(\xi) = N_1(\xi)p_1 + N_2(\xi)p_2 \tag{2.45}$$

[6] One is most often interested in the interpolation error for functions; one can also examine the interpolation errors associated with the gradients of functions

With the linear element, the shape functions are $\frac{1}{2}(1 \mp \xi)$. Expand the true solution in a Taylor series,

$$p(\xi + h) = p(\xi) + h\frac{\partial p}{\partial \xi}\bigg|_{\xi} + \frac{h^2}{2}\frac{\partial^2 p}{\partial \xi^2}\bigg|_{\bar{\xi}\ \text{on}\ [\xi, \xi+h]} \tag{2.46}$$

Where the last term is the error associated with the Taylor series approximation, with $\bar{\xi}$ lying somewhere in the interval. Then, letting $h = \mp 1 - \xi$,

$$\begin{aligned} p(-1) = p_1 &= p(\xi) + (-1-\xi)\frac{\partial p}{\partial \xi}\bigg|_{\xi} + \frac{(-1-\xi)^2}{2}\frac{\partial^2 p}{\partial \xi^2}\bigg|_{\bar{\xi}\ \text{on}\ [-1, \xi]} \\ p(+1) = p_2 &= p(\xi) + (+1-\xi)\frac{\partial p}{\partial \xi}\bigg|_{\xi} + \frac{(+1-\xi)^2}{2}\frac{\partial^2 p}{\partial \xi^2}\bigg|_{\bar{\xi}\ \text{on}\ [+1, \xi]} \end{aligned} \tag{2.47}$$

When determining the interpolation error, we can assume that the nodal values p_1, p_2 are the exact values, since interpolation error is only associated with the approximation of the true function with a linear function, not with any error which might arise at the nodes. Thus, from (2.45), and assuming that p_1, p_2 are the exact nodal values,

$$\begin{aligned} \tilde{p}(\xi) = \{N_1(\xi) + N_2(\xi)\}p(\xi) + \{N_1(\xi)(-1-\xi) + N_2(\xi)(+1-\xi)\}\frac{\partial p}{\partial \xi}\bigg|_{\xi} \\ + \left\{ N_1(\xi)\frac{(-1-\xi)^2}{2}\frac{\partial^2 p}{\partial \xi^2}\bigg|_{\bar{\xi}\ \text{on}\ [-1, \xi]} + N_2(\xi)\frac{(+1-\xi)^2}{2}\frac{\partial^2 p}{\partial \xi^2}\bigg|_{\bar{\xi}\ \text{on}\ [+1, \xi]} \right\} \end{aligned} \tag{2.48}$$

The term inside the first curly bracket is 1, and that inside the second curly brackets is zero, so that the interpolation error is, using the change of coordinates (2.30), $\partial p / \partial \xi = (\partial p / \partial x)(\partial x / \partial \xi) = (L/2)(\partial p / \partial x)$,

$$\begin{aligned} e(\xi) &= \tilde{p}(\xi) - p(\xi) \\ &= N_1(\xi)\frac{(-1-\xi)^2}{2}\frac{\partial^2 p}{\partial \xi^2}\bigg|_{-1<\bar{\xi}<\xi} + N_2(\xi)\frac{(+1-\xi)^2}{2}\frac{\partial^2 p}{\partial \xi^2}\bigg|_{\xi<\bar{\xi}<1} \\ &= \frac{1}{4}\left(\frac{L}{2}\right)^2 (1-\xi^2)\left[(1+\xi)\frac{\partial^2 p}{\partial x^2}\bigg|_{-1<\bar{\xi}<\xi} + (1-\xi)\frac{\partial^2 p}{\partial x^2}\bigg|_{\xi<\bar{\xi}<1} \right] \end{aligned} \tag{2.49}$$

The maximum error in the element is then

$$\begin{aligned} |e|_0 &\le \frac{L^2}{16}\left|(1+\xi)\frac{\partial^2 p}{\partial x^2}\right|_{-1<\bar{\xi}<\xi} + (1-\xi)\frac{\partial^2 p}{\partial x^2}\Big|_{\xi<\bar{\xi}<1}\right| \\ &\le \frac{L^2}{8}\text{Max}\left|\frac{\partial^2 p}{\partial x^2}\right| \end{aligned} \tag{2.50}$$

Thus as the length of the element is decreased, the error decreases as the square of the element length. One says that the linear element is **second-order accurate**:

$$e \propto L^2 \tag{2.51}$$

In general, the error of an element is proportional to L^{k+1}, where L is the length, or a characteristic length, of the element, and k is the order of the interpolating polynomial. The error in the first derivative is proportional to L^k.

The same applies, to a certain extent, for higher dimensions. For example, the error for a linear 2-D triangular element is proportional to l^2, where l is a characteristic length of the element.

2.7 Quadratic C^0 Elements

Here the C^0 quadratic trial function is examined.

2.7.1 Quadratic Trial Function

A quadratic trial function is of the form $\tilde{u}(x) = a + bx + cx^2$, and one needs to choose three nodal points to evaluate the unknown coefficients. The obvious ones to take are the two end-points and the centre-point[7]. Assume then that the values at these nodal points are $\tilde{u}(x_1) = \tilde{u}_1$, $\tilde{u}(x_1 + L/2) = \tilde{u}_2$, $\tilde{u}(x_2) = \tilde{u}_3$, Fig. 2.11. There are then three equations to determine the three nodal values and one finds that {▲Problem 7}

[7] the third node does not have to be central; in some applications, e.g. fracture mechanics, it helps to place nodes one-quarter the way along elements

Quadratic Trial Function:

$$\tilde{u}(x) = N_1(x)\tilde{u}_1 + N_2(x)\tilde{u}_2 + N_3(x)\tilde{u}_3$$

$$N_1 = 1 - 3\frac{x - x_1}{L} + 2\left(\frac{x - x_1}{L}\right)^2 \qquad (2.52)$$

$$N_2 = 4\frac{x - x_1}{L} - 4\left(\frac{x - x_1}{L}\right)^2$$

$$N_3 = -\frac{x - x_1}{L} + 2\left(\frac{x - x_1}{L}\right)^2$$

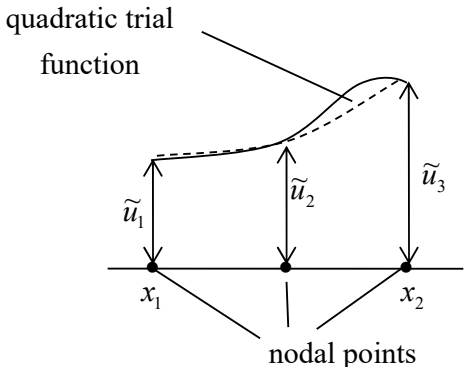


Figure 2.11: the quadratic trial function

Structure of the Weight Functions

As with the linear trial functions, the shape functions are either 0 or 1 at a node, Fig. 2.12, and they sum to 1.

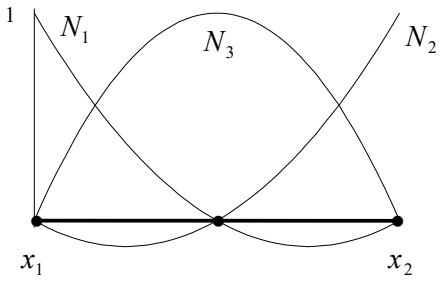


Figure 2.12: shape functions for the quadratic trial function

Looking again at the general ODE problem of Eqn. 2.15, with a single quadratic element one arrives at the system of three equations

Equations for Quadratic Trial Function:

$$\left\{ a\frac{1}{3L}\begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{bmatrix} - b\frac{1}{6}\begin{bmatrix} -3 & +4 & -1 \\ -4 & 0 & +4 \\ +1 & -4 & +3 \end{bmatrix} - c\frac{L}{30}\begin{bmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{bmatrix} \right\} \begin{bmatrix} \tilde{u}_1 \\ \tilde{u}_2 \\ \tilde{u}_3 \end{bmatrix} + d\frac{L}{6}\begin{bmatrix} 1 \\ 4 \\ 1 \end{bmatrix} = a\begin{bmatrix} -u'(x_1) \\ 0 \\ +u'(x_2) \end{bmatrix} \tag{2.53}$$

Again looking at the example $a = 1, b = -2, c = d = 1$ and $x_1 = 0, x_2 = 1, L = 1$, application of the boundary conditions $u(x_1) = \tilde{u}_1 = 1, u'(x_2) = 2$ leads to two equations in two unknowns, which yield $u_2 = 1.290, u_3 = 1.993$ (compared with the exact solutions 1.303 and 2 respectively). The complete quadratic solution is, Fig. 2.13,

$$\begin{aligned} \tilde{u}(x) &= \tilde{u}_1\left(1 - 3x + 2x^2\right) + \tilde{u}_2\left(4x - 4x^2\right) + \tilde{u}_3\left(-x + 2x^2\right) \\ &= 1 + 0.167x + 0.826x^2 \end{aligned} \tag{2.54}$$

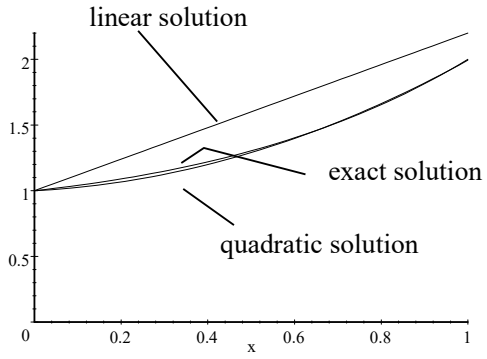


Figure 2.13: FEM solution for the ODE in (2.15) using Quadratic Elements

The quadratic shape functions can be expressed in terms of the local coordinates: from (2.30)

Quadratic Shape Functions (Local Coordinates)

$$N_1 = \frac{1}{2}\xi(\xi - 1), \quad N_2 = 1 - \xi^2, \quad N_3 = \frac{1}{2}\xi(\xi + 1) \tag{2.55}$$

Note that with a quadratic trial function, one does not need to integrate the higher order term by parts in order to obtain a solution. However, integration by parts will ensure that the resulting coefficient matrix is symmetric[8], and is done here.

Consider next the following differential equation, to be solved using two quadratic elements of equal length:

$$\frac{d^2p}{dx^2} + p = 1, \quad \left.\frac{dp}{dx}\right|_{x=0} = 1, \quad p(\pi) = 0 \tag{2.56}$$

[the exact solution is $p(x) = \cos x + \sin x + 1$]

The weighted residual integral is $I = \int_0^\pi (p'' + p - 1)\omega dx = 0$, leading to

$$I = \int_0^\pi \left(\frac{dp}{dx}\frac{d\omega}{dx} - p\omega + \omega \right) dx - \left[\frac{dp}{dx}\omega \right]_0^\pi = 0 \tag{2.57}$$

The finite element mesh consists of two elements and five nodes:

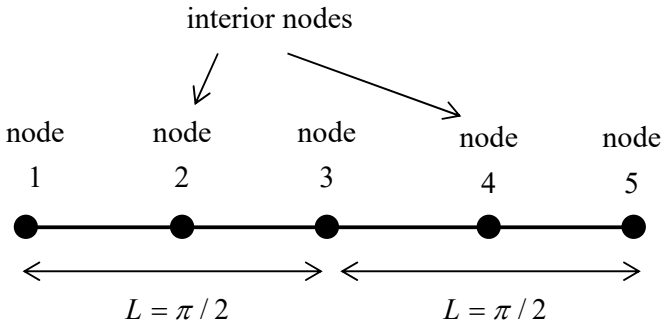


Figure 2.14: FEM solution for the ODE in (2.15) using Quadratic Elements

[8] if the DE to be solved is of the self-adjoint type, as discussed in Chapter 1

Element 1: $p^{(1)} = N_1^{(1)} p_1 + N_2^{(1)} p_2 + N_3^{(1)} p_3$

Element 2: $p^{(2)} = N_1^{(2)} p_3 + N_2^{(2)} p_4 + N_3^{(2)} p_5$

One now has the expressions:

Element 1:

$$\begin{aligned}
&\int_0^L \left(\frac{dp^{(1)}}{dx} \frac{dN_1^{(1)}}{dx} - p^{(1)} N_1^{(1)} + N_1^{(1)} \right) dx + p'(0) \\
&\int_0^L \left(\frac{dp^{(1)}}{dx} \frac{dN_2^{(1)}}{dx} - p^{(1)} N_2^{(1)} + N_2^{(1)} \right) dx = 0 \\
&\int_0^L \left(\frac{dp^{(1)}}{dx} \frac{dN_3^{(1)}}{dx} - p^{(1)} N_3^{(1)} + N_3^{(1)} \right) dx = 0
\end{aligned} \tag{2.58}$$

Element 2:

$$\begin{aligned}
&\int_L^{2L} \left(\frac{dp^{(2)}}{dx} \frac{dN_1^{(2)}}{dx} - p^{(2)} N_1^{(2)} + N_1^{(2)} \right) dx = 0 \\
&\int_L^{2L} \left(\frac{dp^{(2)}}{dx} \frac{dN_2^{(2)}}{dx} - p^{(2)} N_2^{(2)} + N_2^{(2)} \right) dx = 0 \\
&\int_L^{2L} \left(\frac{dp^{(2)}}{dx} \frac{dN_3^{(2)}}{dx} - p^{(2)} N_3^{(2)} + N_3^{(2)} \right) dx - p'(\pi) = 0
\end{aligned} \tag{2.59}$$

Substituting the shape functions into these six integrals, changing to the local variables, and making use of the integrals in the Appendix to this Chapter, leads to

Element 1:

$$\begin{aligned}
p_1\left[+\tfrac{7}{3L} - \tfrac{4L}{30}\right] + p_2\left[-\tfrac{8}{3L} - \tfrac{2L}{30}\right] + p_3\left[\tfrac{1}{3L} + \tfrac{L}{30}\right] + \tfrac{L}{6} + p'(0) &= 0 \\
p_1\left[-\tfrac{8}{3L} - \tfrac{2L}{30}\right] + p_2\left[+\tfrac{16}{3L} - \tfrac{16L}{30}\right] + p_3\left[-\tfrac{8}{3L} - \tfrac{2L}{30}\right] + \tfrac{4L}{6} &= 0 \\
p_1\left[+\tfrac{1}{3L} + \tfrac{L}{30}\right] + p_2\left[-\tfrac{8}{3L} - \tfrac{2L}{30}\right] + p_3\left[+\tfrac{7}{3L} - \tfrac{4L}{30}\right] + \tfrac{L}{6} &= 0
\end{aligned}$$

Element 2:

$$\begin{aligned}
p_3\left[+\tfrac{7}{3L} - \tfrac{4L}{30}\right] + p_4\left[-\tfrac{8}{3L} - \tfrac{2L}{30}\right] + p_5\left[+\tfrac{1}{3L} + \tfrac{L}{30}\right] + \tfrac{L}{6} &= 0 \\
p_3\left[-\tfrac{8}{3L} - \tfrac{2L}{30}\right] + p_4\left[+\tfrac{16}{3L} - \tfrac{16L}{30}\right] + p_5\left[-\tfrac{8}{3L} - \tfrac{2L}{30}\right] + \tfrac{4L}{6} &= 0 \\
p_3\left[+\tfrac{1}{3L} + \tfrac{L}{30}\right] + p_4\left[-\tfrac{8}{3L} - \tfrac{2L}{30}\right] + p_5\left[+\tfrac{7}{3L} - \tfrac{4L}{30}\right] + \tfrac{L}{6} - p'(\pi) &= 0
\end{aligned}$$

Summing the last equation for element 1 and the first equation for element 2, one arrives at the five equations (one for each degree of freedom)

$$\frac{1}{30L}\begin{bmatrix} +70-4L^2 & -80-2L^2 & +10+L^2 & 0 & 0 \\ -80-2L^2 & +160-16L^2 & -80-2L^2 & 0 & 0 \\ +10+L^2 & -80-2L^2 & +140-8L^2 & -80-2L^2 & +10+L^2 \\ 0 & 0 & -80-2L^2 & +160-16L^2 & -80-2L^2 \\ 0 & 0 & +10+L^2 & -80-2L^2 & +70-4L^2 \end{bmatrix}\begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{bmatrix} = \frac{1}{6}\begin{bmatrix} -L-6p'(0) \\ -4L \\ -2L \\ -4L \\ -L+6p'(\pi) \end{bmatrix} \tag{2.60}$$

Applying the natural boundary condition at $x=0$, $p'(0)=1$, and eliminating the last row by applying the essential boundary condition $p(\pi)=p_5=0$ leaves

$$\frac{1}{30L}\begin{bmatrix} +70-4L^2 & -80-2L^2 & +10+L^2 & 0 \\ -80-2L^2 & +160-16L^2 & -80-2L^2 & 0 \\ +10+L^2 & -80-2L^2 & +140-8L^2 & -80-2L^2 \\ 0 & 0 & -80-2L^2 & +160-16L^2 \end{bmatrix}\begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix} = \frac{1}{6}\begin{bmatrix} -L-6 \\ -4L \\ -2L \\ -4L \end{bmatrix} \tag{2.61}$$

Solving the equations, with $L=\pi/2$, gives

$$\begin{aligned} p_1 &= 2.011591 \\ p_2 &= 2.417818 \\ p_3 &= 2.000278 \\ p_4 &= 1.000196 \end{aligned} \quad . \qquad \text{The exact solution is} \qquad \begin{aligned} p_1 &= 2 \\ p_2 &= 2.41421 \\ p_3 &= 2 \\ p_4 &= 1 \end{aligned}$$

The full solution is

Element 1:

$$\begin{aligned} p^{(1)} &= \tfrac{1}{2}\xi^{(1)}\left(\xi^{(1)}-1\right)p_1 + \left(1-\xi^{(1)2}\right)p_2 + \tfrac{1}{2}\xi^{(1)}\left(\xi^{(1)}+1\right)p_3, \qquad \xi^{(1)} = \tfrac{2x}{L}-1 \\ &= \left[1-3\frac{x}{L}+2\left(\frac{x}{L}\right)^2\right]p_1 + \left[4\frac{x}{L}-4\left(\frac{x}{L}\right)^2\right]p_2 + \left[-\frac{x}{L}+2\left(\frac{x}{L}\right)^2\right]p_3 \\ &= 2.012 + 1.042x - 0.668x^2 \end{aligned}$$

Element 2:

$$\begin{aligned} p^{(2)} &= \tfrac{1}{2}\xi^{(2)}\left(\xi^{(2)}-1\right)p_3 + \left(1-\xi^{(2)2}\right)p_4 + \tfrac{1}{2}\xi^{(2)}\left(\xi^{(2)}+1\right)p_5, \qquad \xi^{(2)} = \tfrac{2x}{L}-3 \\ &= \left[6-7\frac{x}{L}+2\left(\frac{x}{L}\right)^2\right]p_3 + \left[-8+12\frac{x}{L}-4\left(\frac{x}{L}\right)^2\right]p_4 + \left[3-5\frac{x}{L}+2\left(\frac{x}{L}\right)^2\right]p_5 \\ &= 4.000 - 1.273x - 0.00009x^2 \end{aligned}$$

which is very accurate.

2.8 Cubic Hermite Finite Elements

Here the C^1 cubic Hermite trial function is examined.

2.8.1 Cubic Hermite Trial Function

The purpose of introducing the cubic Hermite element is to ensure that the first derivatives of the trial function are continuous at the nodes. To this end, consider four unknowns, the two values of p at the element ends and the two values of p' at the element ends. With four unknowns, one can use the cubic trial function

$$p(x) = a + bx + cx^2 + dx^3 \tag{2.62}$$

Using the four equations

$$p(x_1) = p_1, \quad p(x_2) = p_2, \quad p'(x_1) = p_1', \quad p'(x_2) = p_2' \tag{2.63}$$

to re-write the trial function in terms of the unknown nodal values, one arrives at (after some lengthy algebra)

Cubic Hermite Trial Function:

$$p(x) = N_1(x)p_1 + N_2(x)p_2 + N_3(x)p_1' + N_4(x)p_2'$$

where

$$N_i = \frac{1}{(x_1 - x_2)^3}\left[a_i + b_i x + c_i x^2 + d_i x^3\right]$$

with

$$a_i = \begin{Bmatrix} x_2^2\left((3x_1 - x_2\right) \\ x_1^2\left((x_1 - 3x_2\right) \\ -x_1x_2^2(x_1 - x_2) \\ -x_1^2x_2\left((x_1 - x_2\right) \end{Bmatrix}, \quad b_i = \begin{Bmatrix} -6x_1x_2 \\ +6x_1x_2 \\ x_2(2x_1 + x_2)(x_1 - x_2) \\ x_1(x_1 + 2x_2)(x_1 - x_2) \end{Bmatrix}$$

$$c_i = \begin{Bmatrix} +3(x_1 + x_2) \\ -3(x_1 + x_2) \\ -(x_1 + 2x_2)(x_1 - x_2) \\ -(2x_1 + x_2)(x_1 - x_2) \end{Bmatrix}, \quad d_i = \begin{Bmatrix} -2 \\ +2 \\ x_1 - x_2 \\ x_1 - x_2 \end{Bmatrix}$$

(2.64)

The cubic Hermite shape functions can be expressed in terms of the local coordinates: from (2.30),

Cubic Hermite Trial Function (Local Coordinates):

$$\begin{aligned} N_1 &= \frac{1}{4}\left(2 - 3\xi + \xi^3\right) = \frac{1}{4}(1 - \xi)^2(2 + \xi) \\ N_2 &= \frac{1}{4}\left(2 + 3\xi - \xi^3\right) = \frac{1}{4}\left(1 + \xi^2\right)(2 - \xi) \\ N_3 &= \frac{L}{8}\left(+1 - \xi - \xi^2 + \xi^3\right) = +\frac{L}{8}\left(1 - \xi^2\right)(1 - \xi) \\ N_4 &= \frac{L}{8}\left(-1 - \xi + \xi^2 + \xi^3\right) = -\frac{L}{8}\left(1 - \xi^2\right)(1 + \xi) \end{aligned} \tag{2.65}$$

Note that the third and fourth shape functions depend on the element length. Also, although these are zero at the nodes, their derivatives are one or zero there, as required.

Consider the following differential equation, to be solved using a single cubic Hermite element:

$$\frac{d^2p}{dx^2} = 1, \quad p(0) = 1, \quad \left.\frac{dp}{dx}\right|_{x=1} = 2 \tag{2.66}$$

[the exact solution is $p(x) = \frac{1}{2}x^2 + x + 1$]

The weighted residual integral is $I = \int_0^1 (p'' - 1)\omega dx = 0$, leading to

$$\int_0^1 \frac{dp}{dx}\frac{d\omega}{dx}dx = -\int_0^1 \omega dx + \left[\frac{dp}{dx}\omega\right]_0^\pi \tag{2.67}$$

Using the interpolation (2.64), re-labelling $p_1' = p_3$, $p_2' = p_4$, and using the integrals in the Appendix (Eqns. 2A.13, 2A.14), leads to the system of equations

$$\frac{1}{30}\begin{bmatrix} +36\frac{1}{L} & -36\frac{1}{L} & +3 & +3 \\ -36\frac{1}{L} & +36\frac{1}{L} & -3 & -3 \\ +3 & -3 & +4L & -L \\ +3 & -3 & -L & +4L \end{bmatrix}\begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix} = \frac{L}{12}\begin{bmatrix} -6 \\ -6 \\ -L \\ +L \end{bmatrix} + \begin{bmatrix} -p_1' \\ +p_2' \\ 0 \\ 0 \end{bmatrix} \tag{2.68}$$

The boundary conditions then lead to the exact solution, $p_2 = 5/2$, $p_1' = 1$.

The second derivatives of the shape functions are non-zero for the cubic Hermite trial function, and so the solution can be obtained using the weighted residual *without* the integration by parts, using the integrals in (2A.15), but the coefficient matrix will not be symmetric. In this case, when one is changing to local coordinates, one must use

$$\begin{aligned} \frac{dN_i}{dx} &= \frac{dN_i}{d\xi}\frac{d\xi}{dx} \\ \frac{d}{dx}\left(\frac{dN_i}{dx}\right) &= \frac{d}{dx}\left(\frac{dN_i}{d\xi}\frac{d\xi}{dx}\right) \\ &= \frac{d}{dx}\left(\frac{dN_i}{d\xi}\right)\frac{d\xi}{dx} + \frac{dN_i}{d\xi}\frac{d^2\xi}{dx^2} \\ &= \frac{d^2N_i}{d\xi^2}\left(\frac{d\xi}{dx}\right)^2 + \frac{dN_i}{d\xi}\frac{d^2\xi}{dx^2} \end{aligned} \tag{2.69}$$

2.9 Finite Differences

Finite Differences (FD) is an alternative method of obtaining a numerical solution to differential equations. Here, FD and its relation to FE is described. Focusing on a simple second order problem,

$$\frac{\partial^2 p}{\partial x^2} = 1, \qquad p(0) = 1, \quad p'(1) = 2 \tag{2.70}$$

[exact solution: $\frac{1}{2}x^2 + x + 1$]

the FD approach is to approximate derivatives of functions using truncated Taylor series. For example, consider the expansions

$$\begin{aligned} p(x_0 + \Delta x) &\approx p(x_0) + \Delta x \left.\frac{dp}{dx}\right|_{x_0} + \frac{1}{2}(\Delta x)^2 \left.\frac{d^2 p}{dx^2}\right|_{x_0} \\ p(x_0 - \Delta x) &\approx p(x_0) - \Delta x \left.\frac{dp}{dx}\right|_{x_0} + \frac{1}{2}(\Delta x)^2 \left.\frac{d^2 p}{dx^2}\right|_{x_0} \end{aligned} \tag{2.71}$$

Adding these leads to the approximation

$$\left.\frac{d^2 p}{dx^2}\right|_{x_0} = \frac{1}{(\Delta x)^2}\left[p(x_0 - \Delta x) - 2p(x_0) + p(x_0 + \Delta x)\right] + O(\Delta x)^2 \tag{2.72}$$

Using a grid [9] with $N+1$ nodes, so that $\Delta x = 1/N$, and for the natural boundary condition using the approximation

[9] the *mesh* of Finite Elements is usually called a *grid* in Finite Differences

$$
\begin{aligned}
p(x_0-2\Delta x) &\approx p(x_0)-2\Delta x\left.\frac{dp}{dx}\right|_{x_0}+2(\Delta x)^2\left.\frac{d^2p}{dx^2}\right|_{x_0}\\
-4p(x_0-\Delta x) &\approx -4p(x_0)+4\Delta x\left.\frac{dp}{dx}\right|_{x_0}-2(\Delta x)^2\left.\frac{d^2p}{dx^2}\right|_{x_0}\\
&\rightarrow\\
\left.\frac{dp}{dx}\right|_{x_0} &= \frac{1}{2(\Delta x)}\left[p(x_0-2\Delta x)-4p(x_0-\Delta x)+3p(x_0)\right]+O(\Delta x)^2
\end{aligned}
\tag{2.73}
$$

the ODE (2.70) is transformed into the system of equations

$$
N^2\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0\\
1 & -2 & 1 & & & \\
 & 1 & -2 & 1 & & \\
 & & 1 & -2 & & \\
 & & & & \ddots & \\
 & & & \frac{1}{2N} & -\frac{2}{N} & \frac{3}{2N}
\end{bmatrix}
\begin{bmatrix} p_1\\ p_2\\ p_3\\ \vdots\\ p_{n+1}\end{bmatrix}=
\begin{bmatrix} N^2\\ 1\\ 1\\ 1\\ 2\end{bmatrix}
\tag{2.74}
$$

This leads to the exact solution with 3 (or more) nodes.

Now consider a quadratic finite element. The interpolation can be written as

$$
p(x)=N_1p(x_0-\Delta x)+N_2p(x_0)+N_3p(x_0+\Delta x) \tag{2.75}
$$

where x_0 is the central node, the shape functions are given by (2.52), and Δx is the semi-element length. Forming the weighted residual of (2.70) and taking the weight function to be unity, $\omega(x)\equiv 1$ (or a constant), and integrating the shape functions over the element, but not integrating by parts, leads to

$$
\frac{1}{(\Delta x)^2}\left[p(x_0-\Delta x)-2p(x_0)+p(x_0+\Delta x)\right]=1 \tag{2.76}
$$

This scheme provides a single equation for this central node, exactly the same as the Finite Difference equation. Thus this FD scheme can be considered to be a FE scheme with a *constant* weight function.

2.10 Application: Static Elasticity (Structural Mechanics)

The FE methodology described in this Chapter is here used to analyse the problem of an elastic material subject to arbitrary loading conditions. The geometry of the problem is as shown below, a (one dimensional) rod of length *l*, possibly varying cross section $A(x)$ and Young's modulus $E(x)$, subjected to a given displacement or stress/force at its ends.

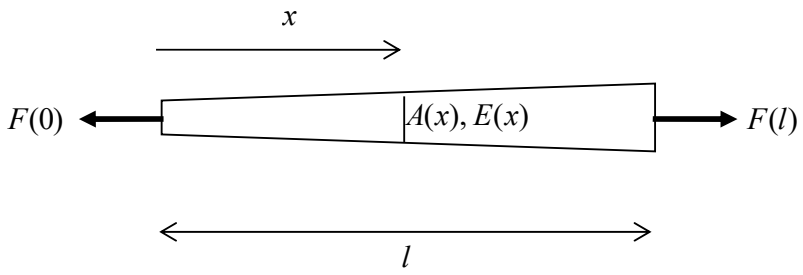


Figure 2.15: an elastic rod

This problem is discussed in detail in Solid Mechanics, Part II, section 2.1.

2.10.1 Governing Differential Equation

The equations governing the response of the rod are:

Governing Equations for Elastostatics:

Equation of Equilibrium:

$$\frac{d}{dx}(A\sigma) + Ab = 0 \tag{2.77}$$

Strain-Displacement Relation:

$$\varepsilon = \frac{du}{dx} \tag{2.78}$$

Constitutive Relation:

$$\sigma = E\varepsilon \tag{2.79}$$

The first two of these are derived in the Appendix to this Chapter, §2.12.2. The third is Hooke's experimental law, which is valid for elastic materials undergoing small strains.

In these equations, σ is the stress, ε is the small strain (change in length per original length) and u is the displacement. The constant of proportionality in the linear elastic constitutive law is E, the Young's modulus of the material; b is a body force (per unit volume), for example the force of gravity, and A is the cross-sectional area.

The strain-displacement relation and the constitutive equation can be substituted into the equation of equilibrium to obtain

1D Governing Equation for Static Elasticity:

$$\frac{d}{dx}\left(AE\frac{du}{dx}\right) + f = 0 \tag{2.80}$$

where $f(x)$ is the product of the body force (per unit volume) and the cross-sectional area, and so is a *force per unit length*.

Note the significance of the term inside the brackets in Eqn. 2.80: from Eqn. 2.78, the stresses acting on any cross-section of the rod are

$$\sigma(x) = E\frac{du}{dx} \tag{2.81}$$

and so the forces acting on any cross section are

$$F\left(x\right) = AE\frac{du}{dx} \tag{2.82}$$

An Exact Solution

When A, E and f are constant, the exact solution is obtained by integrating twice the governing differential equation to obtain

$$u = -\frac{f}{2AE}x^2 + \overline{A}x + \overline{B}, \qquad \varepsilon = -\frac{f}{AE}x + \overline{A}, \qquad \sigma = -\frac{f}{A}x + E\overline{A} \tag{2.83}$$

where $\overline{A}$ and $\overline{B}$ are constants to be determined from the boundary conditions.

2.10.2 FEM Formulation

Formally applying the Galerkin method to the one dimensional static elasticity problem and integrating by parts leads to

$$\int_0^l AE\frac{du}{dx}\frac{d\omega}{dx}dx = \int_0^l f\omega dx + \left[\left(AE\frac{du}{dx}\right)\omega\right]_0^l \tag{2.84}$$

Trial Function & Boundary Conditions

The trial function for the GFEM is of the form

$$\tilde{u}(x) = \sum_{i=1}^{n} \omega_i(x) u_i \tag{2.85}$$

where n is the number of nodes in the element, the u_i are the unknown nodal values and the ω_i are the weighting functions. With the shape functions N_i as the weights, substituting into Eqn. 2.84 gives

$$\sum_{i=1}^{n}\left[\int_0^l AE\frac{dN_i}{dx}\frac{dN_j}{dx}dx\right]u_i = \int_0^l f(x)N_j dx + \left[\left(AE\frac{du}{dx}\right)N_j\right]_0^l \tag{2.86}$$

The boundary conditions can involve the displacement u and its derivative du/dx. Boundary conditions on u are of the essential type and boundary conditions on du/dx are of the natural type. It can be seen that the natural boundary condition in effect involve a condition on the forces F acting at the ends of the rod:

Boundary conditions for Static Elasticity:

$$u(0) = u_0, \quad u(l) = u_l, \quad \left(AE \frac{du}{dx} \right)_{x=0} = A\sigma_0 = F_0, \quad \left(AE \frac{du}{dx} \right)_{x=l} = A\sigma_l = F_l \qquad (2.87)$$

essential boundary conditions

natural boundary conditions

The Stiffness Matrix and Force Vector

The FE equations corresponding to (2.84) take the form

$$\mathbf{Ku} = \mathbf{f} + \mathbf{F} \qquad (2.88)$$

where **f** is the distributed **body force vector** and **F** is the concentrated **force vector**. These two vectors represent the external loads acting on the rod, and together are called the **loads vector**. For example, taking A and E to be constant over an element, and neglecting the body force term, the element equations for a linear element of length L are

$$\frac{AE}{L} \begin{bmatrix} +1 & -1 \\ -1 & +1 \end{bmatrix} \begin{bmatrix} u_i \\ u_{i+1} \end{bmatrix} = \begin{bmatrix} -F(0) \\ +F(L) \end{bmatrix}, \qquad \mathbf{Ku} = \mathbf{F} \qquad (2.89)$$

The dimensions of each term here is one of force. The coefficient matrix **K**, including the AE/L term, is called the **stiffness matrix**. Note that the force $F(0)$ is positive when directed in the negative x direction (tension positive; see Fig. 2.15).

The formulation used here allows one to apply concentrated forces at internal nodes, i.e. within the rod. This can be achieved by having non-zero entries other than at the first and last nodes in the global force vector.

More generally, the form of the FE equations for an arbitrary linear element is, from the above,

$$\begin{bmatrix} \int_{x_i}^{x_{i+1}} AE\frac{dN_1}{dx}\frac{dN_1}{dx}dx & \int_{x_i}^{x_{i+1}} AE\frac{dN_1}{dx}\frac{dN_2}{dx}dx \\ \int_{x_i}^{x_{i+1}} AE\frac{dN_2}{dx}\frac{dN_1}{dx}dx & \int_{x_i}^{x_{i+1}} AE\frac{dN_2}{dx}\frac{dN_2}{dx}dx \end{bmatrix} \begin{bmatrix} u_i \\ u_{i+1} \end{bmatrix} = \begin{bmatrix} -F_i \\ +F_{i+1} \end{bmatrix} + \begin{bmatrix} \int_{x_i}^{x_{i+1}} f(x) N_1 dx \\ \int_{x_i}^{x_{i+1}} f(x) N_2 dx \end{bmatrix} \tag{2.90}$$

The stiffness matrix **K** is singular. This means that if no essential BC is applied (so that a row can be eliminated), then no solution can be obtained. Physically, an absence of an essential BC means that there is only an application of forces at the rod-ends (natural BCs), but no applied displacement. A singular matrix occurs when a structure is not adequately supported, in this case the rod can undergo an arbitrary rigid body translation – this must be prevented by at least one essential BC.

The diagonal terms of **K** must be positive. To see this, suppose that $u_{i+1} = 0$. Then, neglecting the body force term, $K_{11}u_i = -F_i$. F_i is defined positive in the negative x direction, and one must have the displacement u_i directed the same way as the force (so $u_i < 0$ if $u_{i+1} = 0$), and hence $K_{11} > 0$, and similarly for other diagonal terms.

It can also be proved that the elasticity stiffness matrix **K** is positive definite, i.e. $\mathbf{x}^{\mathrm{T}}\mathbf{K}\mathbf{x} > 0$, for any arbitrary vector **x**. For example, in the example given above, Eqn. 2.89,

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \frac{AE}{L} \begin{bmatrix} +1 & -1 \\ -1 & +1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{AE}{L}(x_1 - x_2)^2 > 0 \tag{2.91}$$

Example

Consider a rod with varying cross section $A(x) = A_0(1+x)$, constant E and f, and boundary conditions $u(0) = 0,\ F(l) = \overline{F}$. The exact solution to this problem is seen to be

$$u(x) = \frac{1}{EA_0}\left\{ f\left[(l+1)\ln(1+x) - x\right] + \overline{F}\ln(1+x)\right\} \tag{2.92}$$

For a quadratic element, the element equations are, assuming A to be constant,

$$\frac{AE}{3L}\begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = f\frac{L}{6}\begin{bmatrix} 1 \\ 4 \\ 1 \end{bmatrix} + \begin{bmatrix} -F(0) \\ 0 \\ +F(L) \end{bmatrix} \tag{2.93}$$

Here, A can be taken to be the average cross-sectional area in the element[10]. Taking the data $E = 10^9\,\text{Pa}$, $\overline{F} = 1\text{kN}$, $f = 500\,\text{N/m}$, $l = 1\text{m}$, $A_0 = 1\text{cm}$, the one and two-element solutions are as shown below.

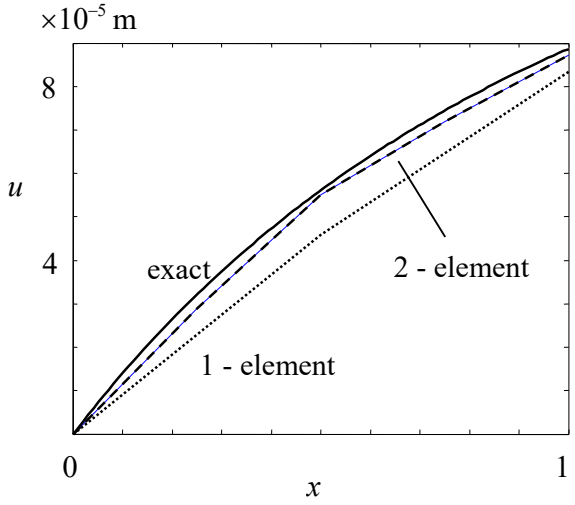


Figure 2.16: FEM solution for the Elastic Rod Example Problem

Work Equivalence of Distributed Loads

Consider now the term involving the body force load $f(x)$. In the real mathematical model (2.80), this force is distributed along the rod. In the FE model, forces cannot be represented at every material particle, only at the nodes; the Galerkin procedure distributes the load according to (see the integral on the right-hand side of Eqn. 2.84)

$$\int_0^L f(x) N_j dx \qquad \ldots \text{node } j \tag{2.94}$$

and this integral results in the distributed body force vector. For example, consider a constant force (per unit length) f distributed over an element of length L, so that the total force acting on the element is fL. For linear and quadratic elements, this force is distributed amongst the nodes as shown in Fig. 2.17.

[10] for greater accuracy, one could have used the explicit expression for $A(x)$ in (2.82) and carry out the integration to obtain a different stiffness matrix

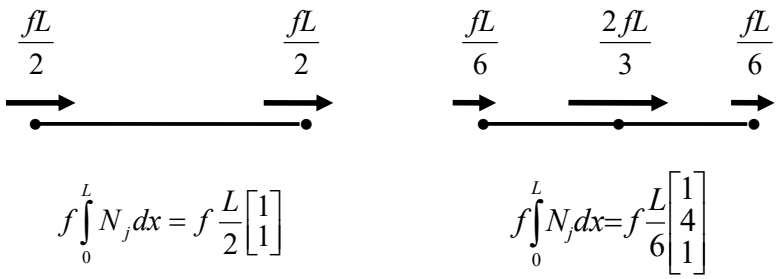


Figure 2.17: FEM distribution of loads at the nodes

In fact, the total force acting over an element in the real model is $F_{ext} = \int_0^L f(x)\,dx$. The total force acting over an element in the FE model is (see Eqn. 2.90)

$$\sum_j \int_0^L f(x) N_j(x)\,dx = \int_0^L f(x) \sum_j N_j(x) dx \tag{2.95}$$

which is seen to be the same as F_{ext}, since $\sum_j N_j(x) = 1$. Thus the forces in the FE model are the same as in the real model; they are **statically equivalent**. Not only that, the loads in both models are **work equivalent**. This is explained in what follows:

The small increment in work done by the body forces over an element in the real model as it undergoes a small displacement increment (over the length of the element) is $\delta W_{ext} = \int_0^L f(x)\,\delta u(x)\,dx$. Approximating the actual displacement by the FEM interpolation (which introduces some discretisation error),

$$\delta W_{ext} \approx \sum_j \delta u_j \int_0^L f(x) N_j(x)\,dx \tag{2.96}$$

The equivalent work in the FE model is

$$\sum_j \int_0^L f(x) N_j(x) dx \times \delta u_j \tag{2.97}$$

which is the same as that in the real model. Thus *the work done by the external loads is the same in the FE model as it is in the real model*. From a mechanics point of view, if the work done in both is the same, the FE model produces the "correct" results. Clearly, if the

forces were distributed in some other way, the work done in the FE model would not be the true work done, and the FE model would give spurious results.

Internal Forces and Internal Work

Internal forces arise within the elastic material to equilibrate the externally applied forces. Whereas the external forces are due to some external agency, for example gravity, or an applied load, the internal forces are a result of the stresses which arise within the deformed material. As the external forces perform work, so do the internal forces: the **internal work** is a result of the internal forces moving through some displacement.

As with the external forces, the internal forces are distributed amongst the nodes in the FE model. Since the FE equations are $\mathbf{Ku} = \mathbf{f} + \mathbf{F}$, and the right-hand side represents the externally applied forces, $\mathbf{Ku}$ must represent the equilibrating internal forces. For example, consider four linear elements, with $u(0) = 0$, a constant cross-section and concentrated forces F_2, F_3 applied at nodes 2 and 3. The FE equations are

$$\frac{AE}{L}\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}\begin{bmatrix} u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} F_2 \\ F_3 \\ 0 \end{bmatrix} \tag{2.98}$$

The internal forces at the nodes in this FE model are then as shown in Fig. 2.15.

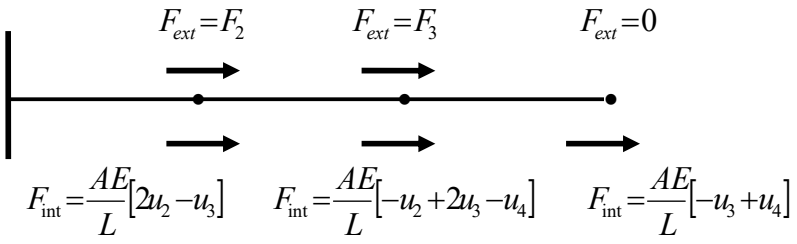


Figure 2.15: Internal forces in the FE model

The internal nodal forces in the FE model, also called the **element nodal point forces**, balance the external nodal forces at each node.

Since the internal forces equilibrate the external forces at each node, and the external forces are work equivalent to those in the real model, the internal forces are also work

equivalent to those in the real model. This point is better illustrated in the context of the potential energy of the deforming system, discussed in the following section.

2.10.3 Variational Formulation

The variational approach was introduced in Chapter 1, section 1.3. (See also the detailed discussion of the variational approach in sections 8.5 and 8.6 of Solid Mechanics, Part I.) Following that procedure, beginning again, the governing equation of static elasticity, Eqn. 2.80, is

$$\frac{d}{dx}\left(AE\frac{du}{dx}\right) + f = 0 \tag{2.99}$$

Multiplying the equation across by a small displacement $\delta u(x)$,

$$\frac{d}{dx}\left(AE\frac{du}{dx}\right)\delta u + f\,\delta u = 0 \tag{2.100}$$

Recall that each term in Eqn. 2.99 has units of force (per unit length), and so the terms in Eqn 2.100 have units of work (per unit length). Integrating over the element gives the total work due to the change in displacement:

$$\int_o^l \frac{d}{dx}\left(AE\frac{du}{dx}\right)\delta u\,dx + \int_0^l f\,\delta u\,dx = 0 \tag{2.101}$$

Integrating by parts and using the results from the Calculus of Variations derived and given in section 1.3, this can now be re-expressed as

$$\delta\left\{\int_o^l \frac{1}{2}AE\left(\frac{du}{dx}\right)^2 dx - \left[\left(AE\frac{du}{dx}\right)u\right]_0^l - \int_0^l f\,u\,dx\right\} = 0 \tag{2.102}$$

It will be recognised that the term $\int_0^l \frac{1}{2}AE\left(du/dx\right)^2 dx$ is the strain energy in the bar (see, for example, Eqn. 8.2.3 in Solid Mechanics, Part I, with the force given by Eqn. 2.82). Thus Eqn. 2.102 can be expressed as

$$\delta\Pi = 0 \tag{2.103}$$

where Π is the potential (strain) energy. Taking the potential energy to be a function of the displacement, Eqn. 2.103 can be expressed as

$$\delta\Pi = \frac{d\Pi(u)}{du}\delta u = 0 \tag{2.104}$$

which is an expression of the principle of minimum potential energy: the solution to the structural mechanics problem is that which causes the potential energy to be stationary, $d\Pi / du = 0$.

Using now the displacement interpolation $u = \sum N_i u_i$,

$$\Pi \approx \int_o^l \frac{1}{2} AE\left(\sum \frac{dN_i}{dx} u_i\right)^2 dx - \left[\left(AE\frac{du}{dx}\right)\sum N_i u_i\right]_0^l - \int_0^l f \sum N_i u_i \, dx \tag{2.105}$$

Introducing row vectors for the shape functions:

$$\left[\mathbf{N}\right] = \left[N_1 \quad N_2\right], \qquad \left[\mathbf{B}\right] = \left[dN_1/dx \quad dN_2/dx\right] \tag{2.106}$$

and the column vector of unknown nodal displacements

$$\left\{\mathbf{u}\right\} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{2.107}$$

Eqn. 2.105 can be expressed as (for linear elements, but this can be generalised in the same way for other elements):

$$\int_o^l \frac{1}{2} AE \left\{\mathbf{u}\right\}^T \left[\mathbf{B}\right]^T \left[\mathbf{B}\right]\left\{\mathbf{u}\right\} dx - \left[\left(AE\frac{du}{dx}\right)\left[\mathbf{N}\right]\left\{\mathbf{u}\right\}\right]_0^l - \int_0^l f\left[\mathbf{N}\right]\left\{\mathbf{u}\right\} dx \tag{2.108}$$

The stationary value of the energy can now be obtained by differentiating this expression with respect to the vector **u**. For the purposes of differentiation with respect to matrices/vectors, note the following rules:

Consider the scalar function of a vector $\mathbf{u}$: $\phi(\mathbf{u}) = [\mathbf{N}]\{\mathbf{u}\} = N_1 u_1 + N_2 u_2$. Differentiation of a scalar with respect to a vector produces a vector:

$$\frac{\partial \phi(\mathbf{u})}{\partial \mathbf{u}} = \begin{bmatrix} \partial\phi / \partial u_1 \\ \partial\phi / \partial u_2 \end{bmatrix} = \begin{bmatrix} N_1 \\ N_2 \end{bmatrix} = [\mathbf{N}]^T \tag{2.109}$$

Consider next the scalar function $\psi(\mathbf{u}) = \{\mathbf{u}\}^T [\mathbf{B}]^T [\mathbf{B}]\{\mathbf{u}\} = (B_1 u_1 + B_2 u_2)^2$. Differentiation gives:

$$\frac{\partial \psi(\mathbf{u})}{\partial \mathbf{u}} = \begin{bmatrix} \partial\psi / \partial u_1 \\ \partial\psi / \partial u_2 \end{bmatrix} = \begin{bmatrix} 2B_1 (B_1 u_1 + B_2 u_2) \\ 2B_2 (B_1 u_1 + B_2 u_2) \end{bmatrix} = 2[\mathbf{B}]^T [\mathbf{B}]\{\mathbf{u}\} \tag{2.110}$$

Thus the stationary point of Eqn. 2.108 is given by

$$\frac{d\Pi(\mathbf{u})}{d\mathbf{u}} = \int_o^l AE[\mathbf{B}]^T [\mathbf{B}]\{\mathbf{u}\}\, dx - \left[\left(AE \frac{du}{dx} \right) [\mathbf{N}]^{\mathbf{T}} \right]_0^l - \int_0^l f[\mathbf{N}]^{\mathbf{T}}\, dx \tag{2.111}$$

Setting this to zero then gives

$$\int_o^l AE[\mathbf{B}]^T [\mathbf{B}]\{\mathbf{u}\}\, dx = \left[\left(AE \frac{du}{dx} \right) [\mathbf{N}]^{\mathbf{T}} \right]_0^l + \int_0^l f[\mathbf{N}]^{\mathbf{T}}\, dx \tag{2.112}$$

which is exactly the same as Eqn. 2.90 derived earlier directly from the governing differential equation.

Thus the FE equations can be derived either directly from the governing differential equation or through the variational approach. When using the variational approach, the potential (strain) energy of the system enters directly, and in that sense one says that the finite element model is energy or work equivalent to the real model.

2.11 Problems

1. Solve the following problem using two linear elements of equal length:

$$\frac{d^2 u}{dx^2} = 6x, \quad u(0) = 1, \quad u(2) = 3 \qquad [\text{exact sln. } u(x) = x^3 - 3x + 1]$$

To deal with the non-homogeneous term, you might proceed in one of two ways:
(i) use the linear interpolation $6x \approx 6\left(x_i N_1 + x_{i+1} N_2\right)$ - which is in this case of course exact since the function $6x$ is linear
(ii) after converting to local coordinates, evaluate the resulting integrals $6\int x N_j d\xi$ exactly using the relation (2.A6)

$$\int_{-1}^{+1} N_i(\xi + A) d\xi = \frac{1}{3}\begin{bmatrix} 3A-1 \\ 3A+1 \end{bmatrix}$$

Evaluate also the FE solution for p'. Sketch the exact, FE and smoothed solutions for p'.

2. Re-solve the ODE of Problem 1, only now with a natural boundary condition at $x = 2$: $(dp/dx)_{x=2} = 6$ [the exact solution is $p(x) = x^3 - 6x + 1$]
3. Solve the problem $p'' + x = 0, \quad p(0) = 1, p'(3) = 0$ with two linear elements of lengths 1 and 2 respectively. Work through the problem in detail, using local coordinates. Derive an expression for p' in each element.
[answer: element 1 - $p' = \frac{13}{3}$, element 2 - $p' = \frac{7}{3}$]
4. What is a C^1 element?
5. What is adaptive meshing? How might an adaptive meshing procedure be implemented with C^0 elements (briefly explain)?
6. What are the sources of error in the FEM (explain any terminology you might use)?
7. Derive Eqns. 2.52, the shape functions for the quadratic element.
8. Consider the equation $p'' = f(x)$. Write down, as quick as you can, the global stiffness matrix for (consult the notes)
   a) a mesh of 5 linear 1-d elements, each of length 2.
   b) a mesh of 3 quadratic 1-d elements, each of length $\frac{1}{3}$.
9. For the (Galerkin) FEM, in a 2nd-order problem, why does one integrate by parts to reduce the order of the highest derivative to 1? Is this always necessary?
10. Note that, in the coefficient matrices encountered in this Chapter, e.g. Eqns. 2.24, 2.27, 2.53, 2.60, the entries of any row or column sum to zero. Using the properties of the shape functions, explain why this is so, and why it is not so for the matrix (2.68) of the cubic Hermite element.

2.12 Appendix to Chapter 2

2.12.1 Integrals involving the 1D Shape Functions

In the following, L is the length of the element, ξ is a local coordinate and A is a constant.

Linear Shape Functions

The shape functions for the standard linear element are

$$\begin{aligned} N_1 &= \frac{1}{2}(1-\xi), \qquad N_2 = \frac{1}{2}(1+\xi) \\ \frac{dN_1}{d\xi} &= -\frac{1}{2}, \qquad \frac{dN_2}{d\xi} = +\frac{1}{2} \end{aligned} \tag{2.A1}$$

The integrals are

1. $$\frac{2}{L}\int_{-1}^{+1}\frac{dN_1}{d\xi}\frac{dN_1}{d\xi}d\xi = +\frac{1}{L}, \qquad \frac{2}{L}\int_{-1}^{+1}\frac{dN_1}{d\xi}\frac{dN_2}{d\xi}d\xi = -\frac{1}{L}, \qquad \frac{2}{L}\int_{-1}^{+1}\frac{dN_2}{d\xi}\frac{dN_2}{d\xi}d\xi = +\frac{1}{L}$$

(2.A2)

2. $$\int_{-1}^{+1}\frac{dN_1}{d\xi}N_1 d\xi = -\frac{1}{2}, \quad \int_{-1}^{+1}\frac{dN_1}{d\xi}N_2 d\xi = -\frac{1}{2}, \quad \int_{-1}^{+1}\frac{dN_2}{d\xi}N_1 d\xi = +\frac{1}{2}, \quad \int_{-1}^{+1}\frac{dN_2}{d\xi}N_2 d\xi = +\frac{1}{2}$$

(2.A3)

3. $$\frac{L}{2}\int_{-1}^{+1}N_1 N_1 d\xi = \frac{L}{3}, \qquad \frac{L}{2}\int_{-1}^{+1}N_1 N_2 d\xi = \frac{L}{6}, \qquad \frac{L}{2}\int_{-1}^{+1}N_2 N_2 d\xi = \frac{L}{3}$$

(2.A4)

4. $$\frac{L}{2}\int_{-1}^{+1}N_1 d\xi = \frac{L}{2}, \qquad \frac{L}{2}\int_{-1}^{+1}N_2 d\xi = \frac{L}{2}$$

(2.A5)

5. $$\frac{L^2}{4}\int_{-1}^{+1}(\xi + A)N_1 d\xi = \frac{L^2}{12}(3A-1), \qquad \frac{L^2}{4}\int_{-1}^{+1}(\xi + A)N_2 d\xi = \frac{L^2}{12}(3A+1)$$

(2.A6)

Quadratic Shape Functions

The shape functions for the standard quadratic element are

$$N_1 = \frac{1}{2}\xi(\xi-1), \quad N_2 = 1-\xi^2, \quad N_3 = \frac{1}{2}\xi(\xi+1)$$
$$\frac{dN_1}{d\xi} = \xi - \frac{1}{2}, \quad \frac{dN_2}{d\xi} = -2\xi, \quad \frac{dN_3}{d\xi} = \xi + \frac{1}{2} \tag{2.A7}$$

The integrals are

1.
$$\frac{2}{L}\int_{-1}^{+1}\frac{dN_1}{d\xi}\frac{dN_1}{d\xi}d\xi = +\frac{7}{3L}, \quad \frac{2}{L}\int_{-1}^{+1}\frac{dN_1}{d\xi}\frac{dN_2}{d\xi}d\xi = -\frac{8}{3L}, \quad \frac{2}{L}\int_{-1}^{+1}\frac{dN_1}{d\xi}\frac{dN_3}{d\xi}d\xi = +\frac{1}{3L}$$
$$\frac{2}{L}\int_{-1}^{+1}\frac{dN_2}{d\xi}\frac{dN_2}{d\xi}d\xi = +\frac{16}{3L}, \quad \frac{2}{L}\int_{-1}^{+1}\frac{dN_2}{d\xi}\frac{dN_3}{d\xi}d\xi = -\frac{8}{3L}, \quad \frac{2}{L}\int_{-1}^{+1}\frac{dN_3}{d\xi}\frac{dN_3}{d\xi}d\xi = +\frac{7}{3L} \tag{2.A8}$$

2.
$$\int_{-1}^{+1}\frac{dN_1}{d\xi}N_1 d\xi = -\frac{1}{2}, \quad \int_{-1}^{+1}\frac{dN_1}{d\xi}N_2 d\xi = -\frac{2}{3}, \quad \int_{-1}^{+1}\frac{dN_1}{d\xi}N_3 d\xi = +\frac{1}{6}$$
$$\int_{-1}^{+1}\frac{dN_2}{d\xi}N_1 d\xi = +\frac{2}{3}, \quad \int_{-1}^{+1}\frac{dN_2}{d\xi}N_2 d\xi = 0, \quad \int_{-1}^{+1}\frac{dN_2}{d\xi}N_3 d\xi = -\frac{2}{3}$$
$$\int_{-1}^{+1}\frac{dN_3}{d\xi}N_1 d\xi = -\frac{1}{6}, \quad \int_{-1}^{+1}\frac{dN_3}{d\xi}N_2 d\xi = +\frac{2}{3}, \quad \int_{-1}^{+1}\frac{dN_3}{d\xi}N_3 d\xi = +\frac{1}{2} \tag{2.A9}$$

3.
$$\frac{L}{2}\int_{-1}^{+1}N_1 N_1 d\xi = +\frac{2L}{15}, \quad \frac{L}{2}\int_{-1}^{+1}N_1 N_2 d\xi = +\frac{L}{15}, \quad \frac{L}{2}\int_{-1}^{+1}N_1 N_3 d\xi = -\frac{L}{30}$$
$$\frac{L}{2}\int_{-1}^{+1}N_2 N_2 d\xi = +\frac{8L}{15}, \quad \frac{L}{2}\int_{-1}^{+1}N_2 N_3 d\xi = +\frac{L}{15}, \quad \frac{L}{2}\int_{-1}^{+1}N_3 N_3 d\xi = +\frac{2L}{15} \tag{2.A10}$$

4.
$$\frac{L}{2}\int_{-1}^{+1}N_1 d\xi = +\frac{L}{6}, \quad \frac{L}{2}\int_{-1}^{+1}N_2 d\xi = +\frac{2L}{3}, \quad \frac{L}{2}\int_{-1}^{+1}N_3 d\xi = +\frac{L}{6} \tag{2.A11}$$

5. $$\frac{L^2}{4}\int_{-1}^{+1}(\xi+A)N_1 d\xi = \frac{L^2}{12}(A-1), \quad \frac{L^2}{4}\int_{-1}^{+1}(\xi+A)N_2 d\xi = \frac{L^2 A}{3}$$
$$\frac{L^2}{4}\int_{-1}^{+1}(\xi+A)N_3 d\xi = \frac{L^2}{12}(A+1)$$

(2.A12)

Cubic Hermite Shape Functions

The shape functions for the standard quadratic element are given by (2.64). The integrals are

1. $$\frac{2}{L}\int_{-1}^{+1}\frac{dN_i}{d\xi}\frac{dN_j}{d\xi}d\xi = \frac{1}{30}\begin{bmatrix} +36\frac{1}{L} & -36\frac{1}{L} & +3 & +3 \\ -36\frac{1}{L} & +36\frac{1}{L} & -3 & -3 \\ +3 & -3 & +4L & -L \\ +3 & -3 & -L & +4L \end{bmatrix}$$

(2.A13)

2. $$\frac{L}{2}\int_{-1}^{+1}N_j d\xi = \frac{L}{12}\begin{bmatrix} +6 \\ +6 \\ +L \\ -L \end{bmatrix}$$

(2.A14)

3. $$\frac{2}{L}\int_{-1}^{+1}N_i\frac{d^2N_j}{d\xi^2}d\xi = \frac{1}{30}\begin{bmatrix} -36\frac{1}{L} & +36\frac{1}{L} & -33 & -3 \\ +36\frac{1}{L} & -36\frac{1}{L} & +3 & +33 \\ -3 & +3 & -4L & +L \\ -3 & +3 & +L & -4L \end{bmatrix}$$

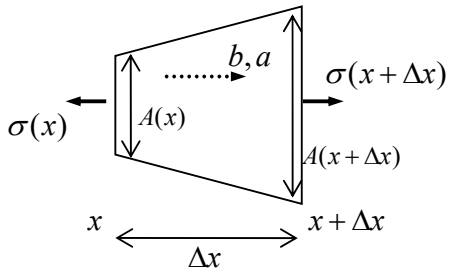
(2.A15)

2.12.2 Derivation of the Governing Equations for Elasticity

The following are discussed in more detail in Solid Mechanics, Part II, Sections 1.1, 1.2.

Force Balance

Consider a one-dimensional differential element of length Δx. The element has varying cross secton, with $A(x+\Delta x) = A(x) + O(\Delta x)$. Let a body force (per unit volume) b, e.g. the force of gravity, act on the element, again, with $b(x+\Delta x) = b(x) + O(\Delta x)$. Denote the acceleration and density of the element be a and ρ. Stresses σ act on the element.



The surface forces acting are $\sigma A\big|_{x+\Delta x} - \sigma A\big|_x$. The force due to b is $A(x)b(x)\Delta x + O(\Delta x)^2$. Applying Newton's second law,

$$\begin{aligned} \sigma A\big|_{x+\Delta x} - \sigma A\big|_x + A(x)b(x)\Delta x + O(\Delta x)^2 &= \rho a \Delta x \left(A\big|_{x+\Delta x} + A\big|_x \right)/2 \\ \rightarrow \frac{\sigma A\big|_{x+\Delta x} - \sigma A\big|_x}{\Delta x} + Ab &= A\rho a + O(\Delta x) \end{aligned} \tag{2A.16}$$

so that, by the definition of the derivative, in the limit as $\Delta x \rightarrow 0$, one has the *equation of motion*

$$\frac{d}{dx}(A\sigma) + Ab = A\rho a \tag{2A.17}$$

If the material is static, or if the accelerations are so low that they can be neglected, this equation reduces to the *equation of equilibrium*:

$$\frac{d}{dx}(A\sigma) + Ab = 0 \tag{2A.18}$$

Kinematics

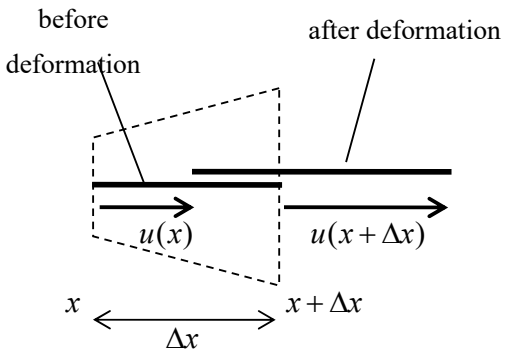
When a material is subjected to a stress/force, it deforms. Define the strain $\varepsilon(x)$ to be the change in length per unit length that a small line element positioned at x undergoes.

To derive a relationship between the stress and the displacement of material particles, consider one such line element of length Δx and emanating from position x. During the deformation, the end at x undergoes a displacement $u(x)$ and the other end undergoes a displacement $u(x+\Delta x)$. From the definition then, the strain is

$$\varepsilon(x) = \frac{\left[u(x+\Delta x) + \Delta x - u(x)\right] - \Delta x}{\Delta x} \tag{2A.19}$$

In the limit as $\Delta x \to 0$ then, this reduces to the relation

$$\varepsilon(x) = \frac{\partial u}{\partial x} \tag{2A.20}$$



3 NonStandard Galerkin FEM

The standard Galerkin FEM as described in chapter 2 is a powerful tool for the numerical solution of a wide variety of problems. However, there are certain problems which cannot be solved with adequate accuracy using the standard GFEM, for example the analysis of nearly-incompressible materials or convection/diffusion with large Reynolds numbers; new variants of the standard FEM have been proposed to deal with problems of this type. Other types of GFEM have been proposed as attractive alternatives to the standard strategies. As an introduction to these variants of the FEM, below are discussed briefly the **Penalty Method**, some **Mixed Methods** and a **Non-Mixed Conservative Method**. These methods are considered "advanced" and do not need to be studied on a first reading.

3.1 The Penalty Method

The Penalty Method involves a very simple idea – essential boundary conditions do not have to be *strongly* enforced, but can be imposed *weakly*.

It has been seen that to apply an essential boundary condition, to node 1 say, $p_1 = \overline{p}_1$, one can alter the FE system as follows:

$$\begin{bmatrix} K_{11} & K_{12} & K_{13} & \cdots & K_{1n} \\ K_{21} & K_{22} & K_{23} & \cdots & K_{2n} \\ K_{31} & K_{32} & K_{33} & \cdots & K_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ K_{n1} & K_{n2} & K_{n3} & \cdots & K_{nn} \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ \vdots \\ p_n \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_n \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & K_{22} & K_{23} & \cdots & K_{2n} \\ 0 & K_{32} & K_{33} & \cdots & K_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & K_{n2} & K_{n3} & \cdots & K_{nn} \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ \vdots \\ p_n \end{bmatrix} = \begin{bmatrix} \overline{p}_1 \\ f_2 - K_{21}\overline{p}_1 \\ f_3 - K_{31}\overline{p}_1 \\ \vdots \\ f_n - K_{n1}\overline{p}_1 \end{bmatrix} \tag{3.1}$$

This is the strong method of applying the boundary condition, building it into the system. An alternative method is to impose it weakly, as follows:

$$\begin{bmatrix} K_{11} & K_{12} & K_{13} & \cdots & K_{1n} \\ K_{21} & K_{22} & K_{23} & \cdots & K_{2n} \\ K_{31} & K_{32} & K_{33} & \cdots & K_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ K_{n1} & K_{n2} & K_{n3} & \cdots & K_{nn} \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ \vdots \\ p_n \end{bmatrix} + \eta \begin{bmatrix} p_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_n \end{bmatrix} + \eta \begin{bmatrix} \overline{p}_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{3.2}$$

Here, the term(s) involving η is called a **penalty** term; the penalty η must be large enough to drive p_1 to $\overline{p}_1$, but not so large as to cause numerical problems.

The method does not hold much of an advantage over the standard FEM, but it is worth studying because it forms the basis of more powerful Galerkin FEMs, the **Internal Penalty** (IP) method and the **Discontinuous** Galerkin FEM. In these latter methods, the trial polynomials *p* over each element may be discontinuous at common nodes – they are forced to be continuous by penalty.

More formally, consider the following problem with non-homogeneous essential boundary conditions:

$$\frac{\partial^2 p}{\partial x^2} + f(x) = 0, \qquad p(0) = p(1) = \overline{p} \tag{3.3}$$

The weak formulation including the penalty term is

$$\int \frac{\partial p}{\partial x} \frac{\partial \omega}{\partial x} dx - \left[\frac{\partial p}{\partial x} \omega \right] + \left[\eta (p - \overline{p}) \omega \right] = \int f \omega dx \tag{3.4}$$

For a mesh of $n-1$ linear elements of equal length *L*, the two integrals lead to the standard global system

$$\frac{1}{L} \begin{bmatrix} +1 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & \ddots & -1 \\ & & -1 & +1 \end{bmatrix} \begin{bmatrix} p_1 \\ \vdots \\ \vdots \\ p_n \end{bmatrix} = \begin{bmatrix} f_1 \\ \vdots \\ \vdots \\ f_n \end{bmatrix} \tag{3.5}$$

The two boundary terms are treated as follows (this formulation relies on the fact that the shape functions take the values 0 or 1 at a boundary node):

$$\left[\frac{\partial p}{\partial x}\omega\right]=\frac{1}{L}(p_{i+1}-p_i)[N_j] \;\rightarrow\; \frac{1}{L}\begin{bmatrix}+1 & -1 & & \\ & & & \\ & & -1 & +1\end{bmatrix}\begin{bmatrix}p_1\\ p_2\\ \vdots\\ p_n\end{bmatrix}$$
$$+\eta[(p-\overline{p})N_j]\rightarrow \;\eta\begin{bmatrix}-1 & & & \\ & & & \\ & & & +1\end{bmatrix}\begin{bmatrix}p_1\\ p_2\\ \vdots\\ p_n\end{bmatrix}=\eta\begin{bmatrix}-\overline{p}\\ \\ +\overline{p}\end{bmatrix} \tag{3.6}$$

leading to the final system

$$\left\{\frac{1}{L}\begin{bmatrix}+1 & 0 & & \\ -1 & 2 & -1 & \\ & -1 & \ddots & -1\\ & & 0 & +1\end{bmatrix}+\eta\begin{bmatrix}-1 & & & \\ & & & \\ & & & +1\end{bmatrix}\right\}\begin{bmatrix}p_1\\ p_2\\ \vdots\\ p_n\end{bmatrix}=\begin{bmatrix}f_1\\ f_2\\ \vdots\\ f_n\end{bmatrix}+\eta\begin{bmatrix}-\overline{p}\\ 0\\ \vdots\\ +\overline{p}\end{bmatrix} \tag{3.7}$$

Taking now $\eta = C/L$, where C is a constant, it can be seen that

$$p_1=-\frac{L}{C+1}f_1+\frac{C}{C+1}\overline{p},\qquad p_n=+\frac{L}{C+1}f_n+\frac{C}{C+1}\overline{p} \tag{3.8}$$

and sufficient accuracy is obtained by choosing C to be sufficiently large.

3.1.1 Symmetric Systems

The system can be made symmetric by including an extra boundary term (it can be included because $p\rightarrow 0$ at the boundary):

$$\int\frac{\partial p}{\partial x}\frac{\partial \omega}{\partial x}dx-\left[\frac{\partial p}{\partial x}\omega\right]-\left[\frac{\partial \omega}{\partial x}p\right]+[\eta\omega p]=\int f\omega dx \tag{3.9}$$

This results in the final system {▲Problem 1}

$$\left\{\frac{1}{L}\begin{bmatrix}+1 & 0 & & \\ 0 & 2 & -1 & \\ & -1 & \ddots & 0 \\ & & 0 & +1\end{bmatrix}+\eta\begin{bmatrix}-1 & & \\ & & \\ & & +1\end{bmatrix}\right\}\begin{bmatrix}p_1\\ p_2\\ \vdots\\ p_n\end{bmatrix}=\begin{bmatrix}f_1\\ f_2\\ \vdots\\ f_n\end{bmatrix}+\eta\begin{bmatrix}-\overline{p}\\ 0\\ \vdots\\ +\overline{p}\end{bmatrix}-\frac{\overline{p}}{L}\begin{bmatrix}+1\\ -1\\ \vdots\\ -1\\ +1\end{bmatrix} \tag{3.10}$$

For example, taking $f(x)=x$ (in which case the exact solution is $p=\overline{p}+x(1-x^2)/6$), the f vector is

$$\frac{L^2}{4}\begin{bmatrix}A_1-1/3\\ A_1+A_2\\ A_2+A_3\\ \vdots\\ A_{n-1}+1/3\end{bmatrix},\qquad A_i=\left(\frac{x_{i+1}+x_i}{x_{i+1}-x_i}\right) \tag{3.11}$$

The value of p at the right hand side, for $\overline{p}=1$ and 5 elements, is then as shown in the table below (convergence at the left hand end is much better for a given C).

C	$p(1)$
1	0.009333333
10	0.819878788
100	0.980382838
1000	0.998020646
10000	0.999801886

3.1.2 Natural Boundary Conditions

Natural boundary conditions can be treated as in the standard FEM. For example, consider next the following problem:

$$\frac{\partial^2 p}{\partial x^2}=A,\qquad p(0)=\overline{p}(0),\quad p'(1)=\overline{p}'(1) \tag{3.12}$$

[exact solution: $p(x)=\frac{1}{2}Ax^2+\left(\overline{p}'(1)-A\right)x+\overline{p}(0)$]

In this case one arrives at the system

$$\left\{\frac{1}{L}\begin{bmatrix}+1 & 0 & & \\ 0 & 2 & -1 & \\ & -1 & \ddots & -1 \\ & & -1 & +1\end{bmatrix}+\eta\begin{bmatrix}-1 & & & \\ & & & \\ & & & \\ & & & 0\end{bmatrix}\right\}\begin{bmatrix}p_1\\ p_2\\ \vdots\\ p_n\end{bmatrix}=\begin{bmatrix}f_1\\ f_2\\ \vdots\\ f_n\end{bmatrix}+\begin{bmatrix}0\\ 0\\ \vdots\\ +\bar{p}'(1)\end{bmatrix}+\eta\begin{bmatrix}-\bar{p}(0)\\ 0\\ \vdots\\ 0\end{bmatrix}-\frac{\bar{p}(0)}{L}\begin{bmatrix}+1\\ -1\\ \vdots\\ 0\\ 0\end{bmatrix} \tag{3.13}$$

3.2 The Mixed Method

It often happens that the most important variable to be solved for is not the independent variable p, but its derivative $q = \partial p / \partial x$. It has been seen that, with the standard Galerkin FEM, the solution for q is of an order less accurate than the solution for p. In mixed methods, equations are set up for the solution of p and q simultaneously (as with the cubic Hermite element). Different interpolation schemes (weight functions) can be used for p and q, depending on the accuracy required. In the most basic case, q is interpolated linearly between nodes whilst p is constant over an element (reversing the accuracy obtained with the standard FEM). Obtaining a more accurate solution for the derivative takes some more computational effort than obtaining a sufficiently accurate p.

3.2.1 The Standard Mixed Method

The standard Mixed Method will be illustrated by solving the problem

$$\frac{\partial^2 p}{\partial x^2} = -1 \text{ subject to } q(0) = 1, \quad p(2) = 0 \text{, where } q = \frac{\partial p}{\partial x}$$

[exact solution: $x - \frac{1}{2}x^2$] (3.14)

Solution I (standard Galerkin FEM)

First, recall the standard FEM: one writes

$$\int_{x_i}^{x_{i+1}} \frac{\partial^2 p}{\partial x^2} w dx = -\int_{x_i}^{x_{i+1}} w dx$$

$$\rightarrow \quad p_i \frac{1}{L}\begin{bmatrix}+1\\-1\end{bmatrix} + p_{i+1}\frac{1}{L}\begin{bmatrix}-1\\+1\end{bmatrix} = \begin{bmatrix}-p'(x_i)\\+p'(x_{i+1})\end{bmatrix} + \frac{L}{2}\begin{bmatrix}+1\\+1\end{bmatrix} \tag{3.15}$$

Two elements, with $L = 1$, give the exact pressures at the nodes as plotted below left, $\begin{bmatrix} p_1 & p_2 \end{bmatrix}^T = \begin{bmatrix} 0 & 1/2 \end{bmatrix}^T$. The system of equations to solve is $(N+1)\times(N+1)$ for N elements, but a little additional work needs to be done to evaluate the derivative q. For the two elements, the derivatives are $q = \partial p / \partial x = (p_{i+1} - p_i)/L$, so

element 1: $q = +\frac{1}{2}$

element 2: $q = -\frac{1}{2}$

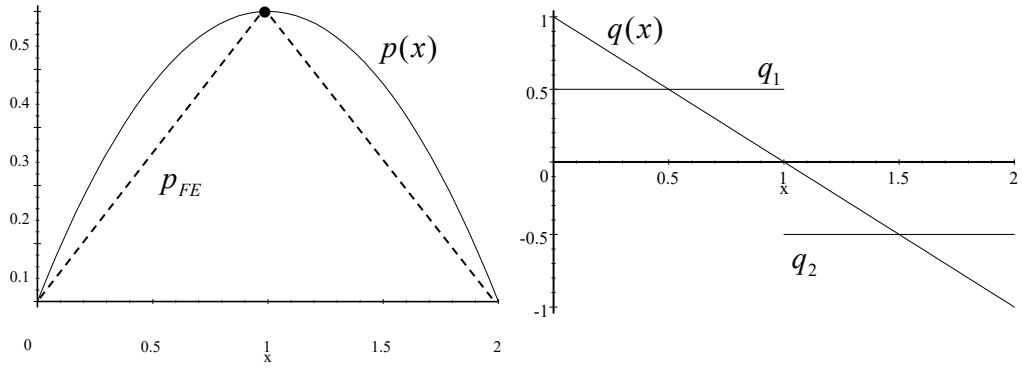


Figure 3.1: Standard GFEM solution to Eqn. 3.14

Solution II (Standard Mixed FEM)

In the (standard) Mixed Method, one can take q to vary linearly over an element and p to be constant over an element, and replace the second order Eqn. 3.14 by the two separate first order equations:

$$\frac{\partial q}{\partial x} = -1, \quad q = \frac{\partial p}{\partial x} \tag{3.16}$$

This allows one to solve for p and q simultaneously. Further, the first of these equations, the **conservation equation**, so called because it often arises in practical problems as an expression of conservation of some property such as mass, will now hold over an element, and this is often important from a physical point of view – it will be noted that, in the standard FEM with linear elements, q is a constant and its derivative is zero; thus this conservation condition is satisfied using the standard FEM *only* in the special case that the governing equation is homogeneous, i.e. $\partial^2 p / \partial x^2 = 0$.

The equations are now discretised in the usual way, with $q = N_i q_i + N_{i+1} q_{i+1}$, Fig. 3.2.

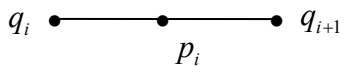


Figure 3.2: Element with *p* constant, *q* varying linearly

The equations on the left here have a constant weight function z (=1), equivalent to a finite difference scheme, those on the right have the standard linear shape/weight functions:

$$\frac{\partial q}{\partial x} = -1 \qquad\qquad q = \frac{\partial p}{\partial x}$$

$$\int_{x_i}^{x_{i+1}} \frac{\partial q}{\partial x} z dx = -\int_{x_i}^{x_{i+1}} z dx \qquad\qquad \int_{x_i}^{x_{i+1}} q w dx = \int_{x_i}^{x_{i+1}} \frac{\partial p}{\partial x} w dx$$

$$q_i \int_{x_i}^{x_{i+1}} \frac{\partial N_1}{\partial x} dx + q_{i+1} \int_{x_i}^{x_{i+1}} \frac{\partial N_2}{\partial x} dx = -(x_{i+1} - x_i)$$

$$q_i \left[\int_{-1}^{+1} \frac{\partial N_1}{\partial \xi} d\xi \right] + q_{i+1} \left[\int_{-1}^{+1} \frac{\partial N_2}{\partial \xi} d\xi \right] = -(x_{i+1} - x_i)$$

$$q_i[-1] + q_{i+1}[+1] = -L$$

$$q_i \int_{x_i}^{x_{i+1}} N_1 N_j dx + q_{i+1} \int_{x_i}^{x_{i+1}} N_2 N_j dx = \left[pN_j \right]_{x_i}^{x_{i+1}} - p_i \int_{x_i}^{x_{i+1}} \frac{\partial N_j}{\partial x} dx$$

$$q_i \left[\frac{L}{2} \int_{-1}^{+1} N_1 N_j d\xi \right] + q_{i+1} \left[\frac{L}{2} \int_{-1}^{+1} N_2 N_j d\xi \right] = \left[pN_j \right]_{x_i}^{x_{i+1}} - p_i \int_{x_i}^{x_{i+1}} \frac{\partial N_j}{\partial x} dx$$

$$q_i \begin{bmatrix} L/3 \\ L/6 \end{bmatrix} + q_{i+1} \begin{bmatrix} L/6 \\ L/3 \end{bmatrix} + p_i \begin{bmatrix} -1 \\ +1 \end{bmatrix} = \begin{bmatrix} -p(x_i) \\ +p(x_{i+!}) \end{bmatrix}$$

(3.17)

Using a single element then gives

$$\begin{bmatrix} L/3 & L/6 & -1 \\ L/6 & L/3 & +1 \\ -1 & +1 & 0 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ p_1 \end{bmatrix} = \begin{bmatrix} -p(x_1) \\ +p(x_2) \\ -L \end{bmatrix} \rightarrow \begin{bmatrix} L/3 & +1 \\ +1 & 0 \end{bmatrix} \begin{bmatrix} q_2 \\ p_1 \end{bmatrix} = \begin{bmatrix} -L/6 \\ -L+1 \end{bmatrix}$$
$$\rightarrow \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = \begin{bmatrix} +1 \\ -1 \end{bmatrix}, \quad p_1 = \frac{1}{3} \tag{3.18}$$

This is the exact solution for q (since it is linear). Note that the coefficient matrix is symmetric. With two elements, one arrives at {▲Problem 2} $p_1 = p_2 = 1/3$ and with four elements one obtains a better solution for the p: $p = \begin{bmatrix} 0.2083 & 0.4583 & 0.4583 & 0.2083 \end{bmatrix}^T$.

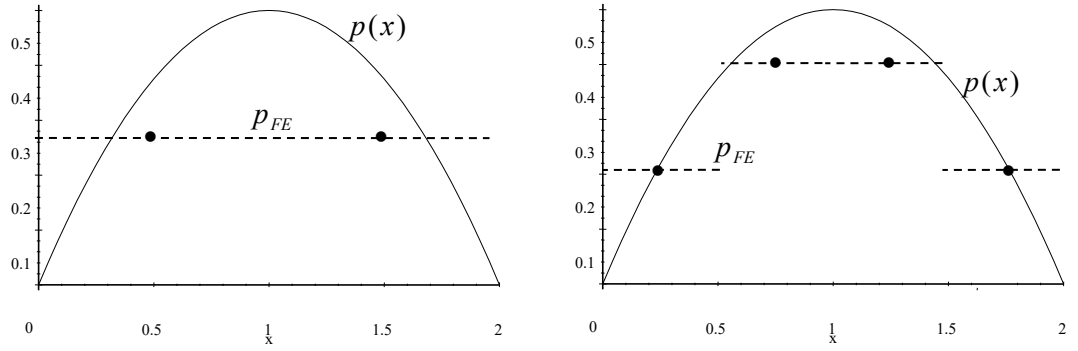


Figure 3.3: Standard Mixed Method solution to Eqn. 3.14 (2 & 4 elements)

Consider now a problem with a cubic solution:

$$\frac{\partial^2 p}{\partial x^2} = x \text{ subject to } p(0) = 2, \quad p(2) = 0 \text{, where } q = \frac{\partial p}{\partial x} \tag{3.19}$$

[exact solution: $2 - \frac{5}{3}x + \frac{1}{6}x^3$]

Using the mixed method with two elements leads to {▲Problem 3}

$$\begin{bmatrix} L/3 & L/6 & 0 & -1 & 0 \\ L/6 & 2L/3 & L/6 & +1 & -1 \\ 0 & L/6 & L/3 & 0 & +1 \\ -1 & +1 & 0 & 0 & 0 \\ 0 & -1 & +1 & 0 & 0 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} -p(x_1) \\ 0 \\ +p(x_{i+1}) \\ L(x_i + L/2) \\ L(x_i + L/2) \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} = \begin{bmatrix} -7/4 \\ -5/4 \\ +1/4 \end{bmatrix}, \qquad \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} 1.2083 \\ 0.1250 \end{bmatrix} \tag{3.20}$$

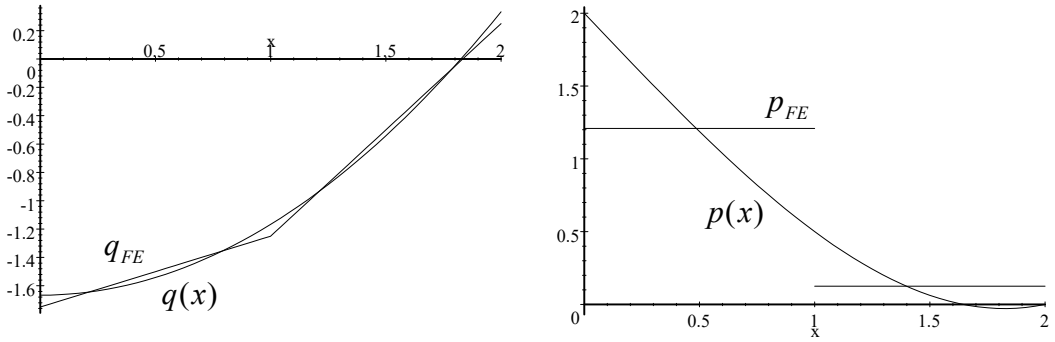


Figure 3.4: Standard Mixed Method solution to Eqn. 3.20

Note that the FE solution here ensures that q is *continuous* across element boundaries.

Higher Order Accuracy

One cannot increase the accuracy of p by letting it vary also linearly over an element, with two unknowns at the node points (element end-points). This is because one cannot have unknown

p's and *q*'s at the *same* node – otherwise the resulting coefficient matrix will be singular, even after application of the boundary conditions.

One cannot circumvent this by having the unknown p's at two interior nodes, i.e. not at the end-points, since then there will be more unknowns than equations.

3.3 Mixed Finite Volume / Covolume Methods

Control Volume methods are widely used in the numerical solution of flow problems. These types of problem have traditionally been solved using the Finite Difference method, but new FEM methods, such as the one described here, are now also being used.

Here problem (3.14) is re-visited using a so-called Control Volume Mixed Finite Element Method[1]. Shown in Fig. 3.5 are three "control volumes" Q_i, $Q_{i+1/2}$ and Q_{i+1}. One can consider Q_i and Q_{i+1} to be "elements" of a primary mesh/grid. The functions *p* and *q* are interpolated over these elements. A secondary or dual grid consists of the overlapping volumes $Q_{i+1/2}$, etc.

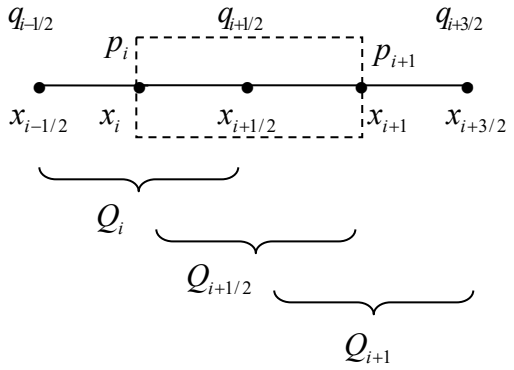


Figure 3.5: Dual grid and Control Volumes

First, consider the equation $q = \dfrac{\partial p}{\partial x}$; integrating over the control volume $Q_{i+1/2}$ gives

[1] *ref*: Cai Z, Jones JE, McCormick SF, Russell TF, "control-volume mixed finite element methods", Computational Geosciences, 1997;1:289-315

$$\int_{x_i}^{x_{i+1}} q dx - \int_{x_i}^{x_{i+1}} \frac{\partial p}{\partial x} dx = 0$$

$$\rightarrow \int_{x_i}^{x_{i+1}} q dx - \left[p(x_{i+1}) - p(x_i) \right] = 0$$

$$\rightarrow q_{i-1/2} \int_{x_i}^{x_{i+1/2}} N_1 dx + q_{i+1/2} \int_{x_i}^{x_{i+1/2}} N_2 dx + q_{i+1/2} \int_{x_{i+1/2}}^{x_{i+1}} N_1 dx + q_{i+3/2} \int_{x_{i+1/2}}^{x_{i+1}} N_2 dx - \left[p_{i+1} - p_i \right] = 0$$

$$\rightarrow q_{i-1/2} \int_{x_i}^{x_{i+1/2}} \left(1 - \frac{x - x_{i-1/2}}{L_i} \right) dx + q_{i+1/2} \int_{x_i}^{x_{i+1/2}} \left(\frac{x - x_{i-1/2}}{L_i} \right) dx$$

$$+ q_{i+1/2} \int_{x_{i+1/2}}^{x_{i+1}} \left(1 - \frac{x - x_{i+1/2}}{L_{i+1}} \right) dx + q_{i+3/2} \int_{x_{i+1/2}}^{x_{i+1}} \left(\frac{x - x_{i+1/2}}{L_{i+1}} \right) dx - \left[p_{i+1} - p_i \right] = 0$$

$$\rightarrow q_{i-1/2} \left[\frac{L_i}{8} \right] + q_{i+1/2} \left[\frac{3L_i}{8} \right] + q_{i+1/2} \left[\frac{3L_{i+1}}{8} \right] + q_{i+3/2} \left[\frac{L_{i+1}}{8} \right] - \left[p_{i+1} - p_i \right] = 0 \tag{3.21}$$

Next consider the conservation equation $\partial q / \partial x = -1$: as with the standard mixed method, integrating over the elements Q_i and Q_{i+1} gives

$$\int_{x_{i-1/2}}^{x_{i+1/2}} \frac{\partial q}{\partial x} dx = -\int_{x_{i-1/2}}^{x_{i+1/2}} dx \qquad \int_{x_{i+1/2}}^{x_{i+3/2}} \frac{\partial q}{\partial x} dx = -\int_{x_{i+1/2}}^{x_{i+3/2}} dx$$

$$\rightarrow q_{i+1/2} - q_{i-1/2} = -L_i \qquad \rightarrow q_{i+3/2} - q_{i+/2} = -L_{i+1} \tag{3.22}$$

For these two elements under consideration, integrating the equation $q = \partial p / \partial x$ over the two half-sized control volumes at either end, which involve pressure values at the boundaries, give rise to

$$
\begin{aligned}
&\int_{x_{i-1/2}}^{x_i} q\,dx - \int_{x_{i-1/2}}^{x_i} \frac{\partial p}{\partial x}\,dx = 0 \\
&\quad \rightarrow q_{i-1/2} \int_{x_{i-1/2}}^{x_i} N_1\,dx + q_{i+1/2} \int_{x_{i-1/2}}^{x_i} N_2\,dx - p_i = -p(x_{i-1/2}) \\
&\quad \rightarrow q_{i-1/2}\left[\frac{3L_i}{8}\right] + q_{i+1/2}\left[\frac{L_i}{8}\right] - p_i = -p(x_{i-1/2}) \\
&\int_{x_{i+1}}^{x_{i+3/2}} q\,dx - \int_{x_{i+1}}^{x_{i+3/2}} \frac{\partial p}{\partial x}\,dx = 0 \\
&\quad \rightarrow q_{i+1/2} \int_{x_{i+1}}^{x_{i+3/2}} N_1\,dx + q_{i+3/2} \int_{x_{i+1}}^{x_{i+3/2}} N_2\,dx + p_{i+1} = +p(x_{i-3/2}) \\
&\quad \rightarrow q_{i+1}\left[\frac{L_{i+1}}{8}\right] + q_{i+3/2}\left[\frac{3L_{i+1}}{8}\right] + p_{i+1} = +p(x_{i-3/2})
\end{aligned}
\tag{3.23}
$$

The (symmetric) system of equations for two elements is then

$$
\begin{bmatrix}
\frac{3L_i}{8} & \frac{L_i}{8} & 0 & -1 & 0 \\
\frac{L_i}{8} & \frac{3(L_i + L_{i+1})}{8} & \frac{L_{i+1}}{8} & +1 & -1 \\
0 & \frac{L_{i+1}}{8} & \frac{3L_{i+1}}{8} & 0 & +1 \\
-1 & +1 & 0 & 0 & 0 \\
0 & -1 & +1 & 0 & 0
\end{bmatrix}
\begin{bmatrix} q_{i-1/2} \\ q_{i+1/2} \\ q_{i+3/2} \\ p_i \\ p_{i+1} \end{bmatrix}
=
\begin{bmatrix} -p(x_{i-1/2}) \\ 0 \\ +p(x_{i-3/2}) \\ -L_i \\ -L_{i+1} \end{bmatrix}
\tag{3.24}
$$

With $L_i = L_{i+1}$ these are (compare with the slightly different system of equations resulting from the standard mixed method, Eqn 3.20)

$$\begin{bmatrix} 3L/8 & L/8 & 0 & -1 & 0 \\ L/8 & 3L/4 & L/8 & +1 & -1 \\ 0 & L/8 & 3L/8 & 0 & +1 \\ -1 & +1 & 0 & 0 & 0 \\ 0 & -1 & +1 & 0 & 0 \end{bmatrix} \begin{bmatrix} q_{i-1/2} \\ q_{i+1/2} \\ q_{i+3/2} \\ p_i \\ p_{i+1} \end{bmatrix} = \begin{bmatrix} -p(x_{i-1/2}) \\ 0 \\ +p(x_{i-3/2}) \\ -L \\ -L \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ L/8 & 3L/4 & L/8 & +1 & -1 \\ 0 & L/8 & 3L/8 & 0 & +1 \\ -1 & +1 & 0 & 0 & 0 \\ 0 & -1 & +1 & 0 & 0 \end{bmatrix} \begin{bmatrix} q_{i-1/2} \\ q_{i+1/2} \\ q_{i+3/2} \\ p_i \\ p_{i+1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ -L \\ -L \end{bmatrix} \rightarrow \begin{bmatrix} q_{i-1/2} \\ q_{i+1/2} \\ q_{i+3/2} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \quad \begin{bmatrix} p_i \\ p_{i+1} \end{bmatrix} = \begin{bmatrix} 3/8 \\ 3/8 \end{bmatrix} \tag{3.25}$$

which is similar to the solution from the standard mixed method.

3.4 Non-Mixed Method for Conservative Elements

As mentioned, in the standard FEM the conservation relation $q' = f(x)$ (for example $f(x) = -1$ in problem 1 above) does not usually hold over an element. The mixed methods discussed above ensure that this relation does hold, but the variable p is not evaluated accurately.

In the method outlined here, both p and q are evaluated accurately. Further, control volumes are not used, so all calculations are carried out for each element – there is only one grid.

First[2], suppose that p_h, the FE approximation to p, is evaluated using any method, for example the standard GFEM. One can then take a Taylor's series of q_h about the centre of the element x_c:

[2] *ref*: Chou S-H, Tang S, "Conservative *p1* conforming and non-conforming galerkin FEMs: effective flux evaluation via a nonmixed method approach", SIAM J. Numeric. Anal., 2000;38(2):660-680

$$
\begin{aligned}
q_h(x) &= q_h(x_c) + (x - x_c)\left.\frac{\partial q}{\partial x}\right|_{x=x_c} \\
&= q_h(x_c) + (x - x_c) f(x_c) \\
&= \frac{\partial p_h}{\partial x} + (x - x_c) f(x_c)
\end{aligned}
\tag{3.26}
$$

In other words, q is evaluated by taking the derivative of p, as in the standard GFEM, and then by adding a *correction* term to make it linear over the element. The accuracy of this depends on how accurate $f(x_c)$ is evaluated, and on how accurate is $\partial q_h(x)/\partial x = f(x_c)$ in any element. There are a number of different ways of evaluating $f(x_c)$, e.g. taking the average of f over an element or using various interpolation schemes (see below). If q is linear, so that $q'' = 0$, etc., and $q'(x_c)$ is evaluated exactly, then $q_h(x)$ thus evaluated will be exact.

Consider the following problem:

$$
\frac{\partial q}{\partial x} = 12x^2 \text{ subject to } p(0) = 2, \quad p(2) = 0 \text{, where } q = \frac{\partial p}{\partial x} \tag{3.27}
$$

$$
\text{[exact solution: } \begin{matrix} p(x) = 2 - 9x + x^4 \\ q(x) = -9 + 4x^3 \end{matrix} \text{]}
$$

First evaluate p using the standard GFEM. Note that in practical codes, it is often convenient to be able to change the term $12x^2$ easily. Thus, instead of inputting $f(x) = 12x^2$ directly and evaluating the weighted integral $12\int_{x_i}^{x_{i+1}} x^2 N_j dx$, one can be more general and interpolate $f(x)$ linearly as in $f(x) = f_i N_1 + f_{i+1} N_2$. This doesn't result in much loss of accuracy, since p is only accurate to this order in any case. Thus, assuming that $f(x)$ is known at the nodes (the f_i's),

$$
\int_{x_i}^{x_{i+1}} \frac{\partial^2 p}{\partial x^2} w dx = 12 \int_{x_i}^{x_{i+1}} x^2 w dx
$$

$$
\rightarrow \quad p_i \int_{x_i}^{x_{i+1}} \frac{\partial N_1}{\partial x}\frac{\partial N_j}{\partial x} dx + p_{i+1} \int_{x_i}^{x_{i+1}} \frac{\partial N_2}{\partial x}\frac{\partial N_j}{\partial x} dx + f_i \int_{x_i}^{x_{i+1}} N_1 N_j dx + f_{i+1} \int_{x_i}^{x_{i+1}} N_2 N_j dx = \left[\frac{\partial p}{\partial x} N_j\right]_{x_i}^{x_{i+1}}
$$

$$
\rightarrow \quad p_i \frac{1}{L}\begin{bmatrix} +1 \\ -1 \end{bmatrix} + p_{i+1} \frac{1}{L}\begin{bmatrix} -1 \\ +1 \end{bmatrix} = \begin{bmatrix} -p'(x_i) \\ +p'(x_{i+1}) \end{bmatrix} - \frac{L}{6}\begin{bmatrix} 2f_i + f_{i+1} \\ f_i + 2f_{i+1} \end{bmatrix}
$$

(3.28)

In the formula

$$q_h(x) = \frac{\partial p_h}{\partial x} + (x - x_c) f(x_c), \tag{3.29}$$

$f(x_c)$ can be evaluated from the linear interpolation of $f(x)$, i.e. simply the average of the nodal values, $(f_i + f_{i+1})/2$. One could also use the more specific expression $f(x_c) = 12x_c^2$.

Results are shown below for 3 elements. Note that the FE solution for the derivative q is discontinuous (very slightly so here) across the element boundaries.

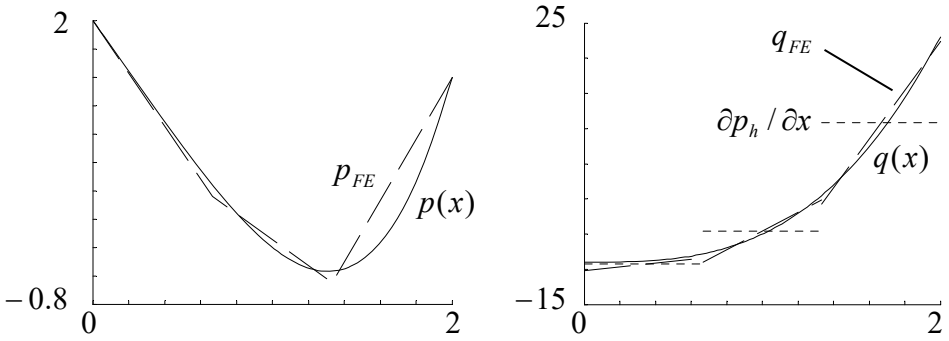


Figure 3.6: Non-Mixed Method solution to Eqn. 3.27

3.5 Problems

1. For the Penalty Method, show how the boundary term $[(\partial\omega/\partial x)p]$ leads to the final symmetric system (3.10). What is the accuracy obtained at the boundary, that is, what are the values of p_1, p_n in this case, in terms of $C = L\eta$, L, $\bar{p}$ and f_i?
2. For the standard mixed method, use (3.18) to write out the system of equations for the two-element mesh for the problem (3.16), the solution of which is $p_1 = p_2 = 1/3$.
3. Derive the system of equations (3.20).

4 Finite Element Methods for Partial Differential Equations

Ordinary Differential Equations (ODEs) have been considered in the previous two Chapters. Here, Partial Differential Equations (PDEs) are examined. Taking x and t to be the independent variables, a general second-order PDE is

$$a\frac{\partial^2 u}{\partial x^2} + b\frac{\partial^2 u}{\partial x \partial t} + c\frac{\partial^2 u}{\partial t^2} + d\frac{\partial u}{\partial x} + e\frac{\partial u}{\partial t} + fu = g \tag{4.1}$$

PDEs are classified according to the value of $b^2 - ac$:

$$b^2 - ac = \begin{cases} > 0 & \text{hyperbolic} \\ = 0 & \text{parabolic} \\ < 0 & \text{elliptic} \end{cases} \tag{4.2}$$

Two special cases of the PDE (4.1) will be examined here, the most commonly encountered ones in applications; these are the first order (in t) parabolic system

$$a\frac{\partial^2 u}{\partial x^2} + d\frac{\partial u}{\partial t} + fu = g \tag{4.3}$$

and the second order (in t) hyperbolic system

$$a\frac{\partial^2 u}{\partial x^2} + c\frac{\partial^2 u}{\partial t^2} + fu = g, \quad ac < 0 \tag{4.4}$$

In applications, x will usually represent a spatial coordinate and t will represent time. This terminology is used below for these variables.

The Galerkin Finite Element Method is used to reduce these PDEs to a system of ODEs, which can then be solved using standard ODE solver algorithms.

4.1 First Order Systems

Here, the first order parabolic equation (4.3) is discussed. In particular, consider the following problem:

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0, \quad \text{B.C.} \quad \begin{matrix} u(0,t) = 0 \\ u(1,t) = 1 \end{matrix}, \quad \text{I.C.} \quad u(x,0) = \sin \pi x + x \tag{4.5}$$

[exact solution: $u(x,t) = \sin \pi x \exp(-\pi^2 t) + x$]

4.1.1 FE equations for First Order Systems

The weighted residual form of (4.5) is

$$\int_0^1 \frac{\partial u}{\partial t} w dx + \int_0^1 \frac{\partial u}{\partial x} \frac{\partial w}{\partial x} dx = \left[\frac{\partial u}{\partial x} w \right]_0^1 \tag{4.6}$$

The Galerkin procedure and shape functions are used to discretise the *space* variable in (4.6) only; the nodal values are functions of t. For a linear element, Fig. 4.1, let

$$u(x,t) = u_i(t) N_1(x) + u_{i+1}(t) N_2(x)$$

$$\frac{\partial u(x,t)}{\partial x} = u_i(t) \frac{\partial N_1(x)}{\partial x} + u_{i+1}(t) \frac{\partial N_2(x)}{\partial x} \tag{4.7}$$

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial u_i(t)}{\partial t} N_1(x) + \frac{\partial u_{i+1}(t)}{\partial t} N_2(x)$$

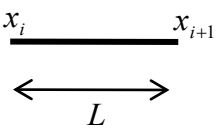


Figure 4.1: A Linear Element

These lead to two equations, one for each weight N_j,

$$\frac{\partial u_i}{\partial t}\int_{x_i}^{x_{i+1}} N_1 N_j dx + \frac{\partial u_{i+1}}{\partial t}\int_{x_i}^{x_{i+1}} N_2 N_j dx + u_i \int_{x_i}^{x_{i+1}} \frac{\partial N_1}{\partial x}\frac{\partial N_j}{\partial x} dx + u_{i+1}\int_{x_i}^{x_{i+1}} \frac{\partial N_2}{\partial x}\frac{\partial N_j}{\partial x} dx = \left[\frac{\partial u}{\partial x} N_j\right]_{x_i}^{x_{i+1}} \qquad j = 1,2 \tag{4.8}$$

Evaluating the integrals leads to the element equations

$$\frac{L}{6}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} \dot{u}_i \\ \dot{u}_{i+1} \end{bmatrix} + \frac{1}{L}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} u_i \\ u_{i+1} \end{bmatrix} = \begin{bmatrix} -\partial u/\partial x(x_i) \\ +\partial u/\partial x(x_{i+1}) \end{bmatrix} \tag{4.9}$$

element capacitance matrix

Element stiffness matrix

The difference between the FE equations for a first order system and those for the standard linear (ODE) system, is the appearance of the capacitance matrix[1] **C**.

After assembly, one has the system of *ordinary* differential equations

$$\mathbf{C}\dot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{F} \tag{4.10}$$

These equations can be solved in a number of different ways (see below).

4 linear elements

Assembling the global 5×5 matrices for the case of four linear elements, applying the boundary conditions $u(0) = u_1 = 0,\ u(1) = u_5 = 1$, noting that $\dot{u}_1 = 0, \dot{u}_5 = 0$, and eliminating the first and last equations leads to {▲Problem 1}

$$\frac{1}{24}\begin{bmatrix} 4 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 4 \end{bmatrix}\begin{bmatrix} \dot{u}_2 \\ \dot{u}_3 \\ \dot{u}_4 \end{bmatrix} + 4\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}\begin{bmatrix} u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 4 \end{bmatrix} \tag{4.11}$$

[1] so-called because this first order system arises in heat conduction problems, and this **C** matrix involves the specific heat capacity of materials

which is a system of three coupled first order ODEs, which can be solved subject to the initial conditions $u(x,0) = \sin \pi x + x$.

Example

As another example, consider the differential equation

$$\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left(\alpha \frac{\partial u}{\partial x} \right) = f(x) \tag{4.12}$$

which is a general form of a first order in time equation which arises in many important problems, including transient heat conduction, diffusion, flow through channels and other applications.

Forming the weighted residual integral and using linear shape functions, $\tilde{u}(x,t) = N_1(x)u_1(t) + N_2(x)u_2(t)$,

$$\begin{aligned}
&\frac{du_1}{dt}\int_0^l N_1 N_1 dx + \frac{du_2}{dt}\int_0^l N_2 N_1 dx \\
&\quad + u_1 \int_0^l \alpha \frac{dN_1}{dx}\frac{dN_1}{dx}dx + u_2 \int_0^l \alpha \frac{dN_2}{dx}\frac{dN_1}{dx}dx = \left[\alpha \frac{\partial u}{\partial x} N_1 \right]_0^l + \int_0^l f\, N_1 dx \\
&\frac{du_1}{dt}\int_0^l N_1 N_2 dx + \frac{du_2}{dt}\int_0^l N_2 N_2 dx \\
&\quad + u_1 \int_0^l \alpha \frac{dN_1}{dx}\frac{dN_2}{dx}dx + u_2 \int_0^l \alpha \frac{dN_2}{dx}\frac{dN_2}{dx}dx = \left[\alpha \frac{\partial u}{\partial x} N_2 \right]_0^l + \int_0^l f\, N_2 dx
\end{aligned} \tag{4.13}$$

Transforming to local coordinates, from $x = [x_i, x_{i+1}]$ to $\xi = [-1, +1]$, and also approximating the "loading" function $f(x)$ by a linear interpolation, $f(x) = f_1 N_1(x) + f_2 N_2(x)$, and taking α to eb a constant for the sake of illustration,

$$\begin{aligned}
&\dot{u}_1 \left[\frac{L}{2}\int_{-1}^{+1} N_1 N_j d\xi \right] + \dot{u}_2 \left[\frac{L}{2}\int_{-1}^{+1} N_2 N_j d\xi \right] \\
&\quad + u_1 \alpha \left[\frac{2}{L}\int_{-1}^{+1} \frac{dN_1}{d\xi}\frac{dN_j}{d\xi}d\xi \right] + u_2 \alpha \left[\frac{2}{L}\int_{-1}^{+1} \frac{dN_2}{d\xi}\frac{dN_j}{d\xi}d\xi \right] \qquad , \qquad j = 1,2 \\
&\quad = \left[\alpha \frac{\partial u}{\partial x} N_j \right]_{-1}^{+1} + f_1 \left[\frac{L}{2}\int_{-1}^{+1} N_1\, N_j d\xi \right] + f_2 \left[\frac{L}{2}\int_{-1}^{+1} N_2\, N_j d\xi \right]
\end{aligned} \tag{4.14}$$

Evaluating all the integrals using the results of the Appendix to Chapter 2, section 2.12.1, leads to the element equations

$$\frac{L}{6}\begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix}\begin{bmatrix}\dot{u}_1\\ \dot{u}_2\end{bmatrix}+\alpha\frac{1}{L}\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}\begin{bmatrix}u_1\\ u_2\end{bmatrix}=\alpha\begin{bmatrix}-u'(-1)\\ +u'(+1)\end{bmatrix}+\frac{L}{6}\begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix}\begin{bmatrix}f_1\\ f_2\end{bmatrix} \tag{4.15}$$

The final global system of equations is then

$$\begin{aligned}\frac{L}{6}\begin{bmatrix}2 & 1 & 0 & \cdots & 0 & 0\\ 1 & 4 & 1 & \cdots & 0 & 0\\ 0 & 1 & 4 & \cdots & 0 & 0\\ \vdots & & & & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & 1 & 2\end{bmatrix}\begin{bmatrix}\dot{u}_1\\ \dot{u}_2\\ \dot{u}_3\\ \vdots\\ \dot{u}_{n+1}\end{bmatrix}+\frac{\alpha}{L}\begin{bmatrix}1 & -1 & 0 & \cdots & 0 & 0\\ -1 & 2 & -1 & \cdots & 0 & 0\\ 0 & -1 & 2 & \cdots & 0 & 0\\ \vdots & & & & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & -1 & 1\end{bmatrix}\begin{bmatrix}u_1\\ u_2\\ u_3\\ \vdots\\ u_{n+1}\end{bmatrix}\\ =\alpha\begin{bmatrix}-u'(-1)\\ 0\\ 0\\ \vdots\\ +u'(+1)\end{bmatrix}+\frac{L}{6}\begin{bmatrix}2 & 1 & 0 & \cdots & 0 & 0\\ 1 & 4 & 1 & \cdots & 0 & 0\\ 0 & 1 & 4 & \cdots & 0 & 0\\ \vdots & & & & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & 1 & 2\end{bmatrix}\begin{bmatrix}f_1\\ f_2\\ f_3\\ \vdots\\ f_{n+1}\end{bmatrix}\end{aligned} \tag{4.16}$$

Let us assume that the boundary conditions are

$$u_1=\overline{u}_1,\quad \left.\frac{\partial u}{\partial x}\right|_{x_{n+1}}=\overline{u}'_{n+1} \tag{4.17}$$

The natural boundary condition can be applied by directly replacing the term $u'(+1)$ in the right-hand side vector. The essential boundary condition can be applied by replacing the first row as follows:

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ \frac{L}{6} & 4\frac{L}{6} & \frac{L}{6} & \cdots & 0 & 0 \\ 0 & \frac{L}{6} & 4\frac{L}{6} & \cdots & 0 & 0 \\ \vdots & & & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{L}{6} & 2\frac{L}{6} \end{bmatrix} \begin{bmatrix} \dot{u}_1 \\ \dot{u}_2 \\ \dot{u}_3 \\ \vdots \\ \dot{u}_{n+1} \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -\frac{\alpha}{L} & 2\frac{\alpha}{L} & -\frac{\alpha}{L} & \cdots & 0 & 0 \\ 0 & -\frac{\alpha}{L} & 2\frac{\alpha}{L} & \cdots & 0 & 0 \\ \vdots & & & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\frac{\alpha}{L} & \frac{\alpha}{L} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{n+1} \end{bmatrix}$$
$$= \begin{bmatrix} \overline{u}_1 \\ 0 \\ 0 \\ \vdots \\ +\alpha\overline{u}'_{n+1} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ \frac{L}{6} & 4\frac{L}{6} & \frac{L}{6} & \cdots & 0 & 0 \\ 0 & \frac{L}{6} & 4\frac{L}{6} & \cdots & 0 & 0 \\ \vdots & & & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{L}{6} & \frac{L}{3} \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_{n+1} \end{bmatrix} \tag{4.18}$$

Note that we have retained a "1" in the C_{11} element of the capacitance matrix **C**, otherwise it will be singular. The first row states that $\dot{u}_1 + u_1 = \overline{u}_1$. The general solution to this differential equation is $u_1 = Ae^{-t} + \overline{u}_1$; with the initial condition that $u_1(0) = \overline{u}_1$, it in effect states that $u_1 = \overline{u}_1$ for all time.

4.1.2 Eigenvalues and Eigenvectors

Before going on to discuss the various possible ways of solving the system 4.11, 4.18, in §4.1.3 below, it is worthwhile discussing the associated eignevalue problem. An eigenvalue analysis of the PDE (4.5) *involves the boundary conditions but disregards the initial conditions*. Although the initial conditions are not considered, and so the full solution is not obtained, nevertheless the analysis can furnish much useful information. For example, eigenvalues and eigenvectors often have a real physical significance for the problem at hand, and eigenvalues are related to the stability of numerical solution procedures for the associated FE equations (see below).

Consider first a single degree of freedom of (4.10): $c\dot{u} + ku = f$, which has the solution $u(t) = f/k + e^{-\lambda t}$, where $\lambda = k/c$. The transient solution decays and after a sufficient amount of time the solution approaches the steady-state solution $u_s = f/k$. A solution to the complete system can be obtained by assuming a similar behaviour:

$$\mathbf{u}(t) = \mathbf{u}_s + \overline{\mathbf{u}}e^{-\lambda t} \tag{4.19}$$

Here, $\overline{\mathbf{u}} = \left[\overline{u}_1 \;\; \overline{u}_2 \;\; \cdots \;\; \overline{u}_n\right]^{\mathrm{T}}$ is called an *eigenvector* and λ an *eigenvalue*. Substituting into the system of equations (4.10) gives

$$\left[\mathbf{K} - \lambda\mathbf{C}\right]\overline{\mathbf{u}} = 0 \tag{4.20}$$

This is a system of $n \times n$ equations in the n nodal values of $\overline{\mathbf{u}}$. From Linear Algebra, such a system of homogeneous equations only has a (non-zero) solution if the determinant of the coefficient matrix is zero, that is

$$\left|\mathbf{K} - \lambda\mathbf{C}\right| = 0 \tag{4.21}$$

Eqn. 4.21 is a polynomial of the nth order and so has n solutions for the eigenvalue[2] λ; there is one eigenvalue for each degree of freedom of the system. Corresponding to each of the n eigenvalues $\lambda^{(j)}$ there is an eigenvector $\overline{\mathbf{u}}^{(j)}$. Each pair $\lambda^{(j)}, \overline{\mathbf{u}}^{(j)}$, corresponds to a certain *mode* of the system. The complete solution is a linear combination of these modes:

$$\begin{aligned} \mathbf{u}(t) &= \mathbf{u}_s + \sum_{j=1}^{n} \beta_j \overline{\mathbf{u}}^{(j)} e^{-\lambda^{(j)} t} \\ u_i(t) &= u_{is} + \sum_{j=1}^{n} \beta_j \overline{u}_i^{(j)} e^{-\lambda^{(j)} t} \end{aligned} \tag{4.22}$$

for the nodes $i = 1, 2, \cdots, n$; the coefficients β_j depend on the initial conditions.

Mesh Size

In first-order linear problems, it can be shown that $\lambda_{\max} = \mathrm{O}\left(1/h^2\right)$, where h is a mesh length parameter (for example element-length). For example, consider the FE equations for a single linear element, Eqns (4.9). Then

$$\left|\mathbf{K} - \lambda\mathbf{C}\right| = \begin{vmatrix} \dfrac{1}{L} - \lambda\dfrac{L}{3} & -\dfrac{1}{L} - \lambda\dfrac{L}{6} \\ -\dfrac{1}{L} - \lambda\dfrac{L}{6} & \dfrac{1}{L} - \lambda\dfrac{L}{3} \end{vmatrix} = -\lambda + \frac{L^2}{12}\lambda^2 = 0 \quad \rightarrow \quad \lambda = 0, \frac{12}{L^2} \tag{4.23}$$

[2] it can be proved that these eigenvalues are also the eigenvalues of the matrix $\mathbf{C}^{-1}\mathbf{K}$

It can be seen that $\lambda_{\max} \propto 1/L^2$ so that, as the mesh gets very fine, the maximum eigenvalue gets very large. The consequences of this fact will be discussed further below.

4.1.3 Direct Integration for First Order Systems

A number of different direct integration methods are available for the integration of the first order system (4.10), for example,

1. Explicit Euler's method
2. Implicit Euler's method
3. Semi-implicit Euler's method
4. Predictor-Corrector method
5. Methods based on Runga-Kutta formulae

1. Explicit Euler's method

To derive the explicit Euler's method, first expand $u_i(t)$ in a Taylor series, *i* referrring to a particular node:

$$u_i(t+\Delta t) = u_i(t) + \Delta t \dot{u}_i(t) + \frac{1}{2}(\Delta t)^2 \ddot{u}_i(t) + \cdots \tag{4.24}$$

The time derivative at time *t* can then be approximated by the forward difference approximation[3]

$$\dot{u}_t = \frac{u_{t+\Delta t} - u_t}{\Delta t} \tag{4.25}$$

In (4.25), terms of order $\mathrm{O}(\Delta t)$ have been neglected from (4.22), that is, the *truncation error* is proportional to Δt. The FE equations are now written at time *t*, $\mathbf{C}\dot{\mathbf{u}}(t) + \mathbf{K}\mathbf{u}(t) = \mathbf{F}(t)$, and then rearranged as

[3] note that there are many other explicit formulae, each derived from different finite difference Taylor expansion formulae; some of these are discussed further on

$$\mathbf{C}\mathbf{u}(t+\Delta t) = \mathbf{C}\mathbf{u}(t) + \Delta t\left[\mathbf{F}(t) - \mathbf{K}\mathbf{u}(t)\right] \tag{4.26}$$

For the purpose of coding, the equation can be rewritten using $\mathbf{u}(t+\Delta t) = \mathbf{u}(t) + \Delta\mathbf{u}$:

Explicit Euler Algorithm:

$$\begin{aligned} &\mathbf{C}\Delta\mathbf{u} = \mathbf{R}(t) \\ &\quad \text{where } \mathbf{R}(t) = \Delta t\left[\mathbf{F}(t) - \mathbf{K}\mathbf{u}(t)\right] \\ &\mathbf{u}(t+\Delta t) = \mathbf{u}(t) + \Delta\mathbf{u} \end{aligned} \tag{4.27}$$

The right-hand side here is known: Δt is chosen by the user, **K** is constant for all time, **F** is a known "loading" term which is specified, and **u** is known at time *t*. The left-hand side **C** is also a constant for all time. The algorithm is started by specifying initial conditions at all the nodes: $\mathbf{u}(0)$.

Note that the cost of the integration, that is, the number of operations required, is directly proportional to the number of time steps required for solution. It follows that the selection of an appropriate time step in direct integration is of much importance.

Considering a one-dimensional case for illustrative purposes, consider the ODE

$$\frac{du}{dt} + \lambda u = f, \quad u(0) = \overline{u}_0 \tag{4.28}$$

Substituting $\dot{u}_t = \left(u_{t+\Delta t} - u_t\right)/\Delta t$ into Eqn. 4.28 gives

$$u(t+\Delta t) = (1-\lambda\Delta t)u(t) + f\Delta t \tag{4.29}$$

which leads to, summing the geometric series,

$$\begin{aligned} u(n\Delta t) &= (1-\lambda\Delta t)^n u(0) + f\Delta t\sum_{i=0}^{n-1}(1-\lambda\Delta t)^i \\ &= (1-\lambda\Delta t)^n u(0) + \frac{f}{\lambda}\left[1-(1-\lambda\Delta t)^n\right] \end{aligned} \tag{4.30}$$

This is plotted in Fig. 4.2 for $\lambda = 2$, $f = 1$, $\overline{u}_0 = 1$; for $\Delta t = 1/3$ and $\Delta t = 1.05$, together with the exact solution $u(t) = \frac{1}{2}\left(1+e^{-2t}\right)$.

The solution is fairly accurate for $\Delta t = 1/3$ (the solution is very close to the exact solution for $\Delta t < 0.1$). On the other hand, when the time step is as large as $\Delta t = 1.05$, the solution is highly inaccurate; this issue is explained further below.

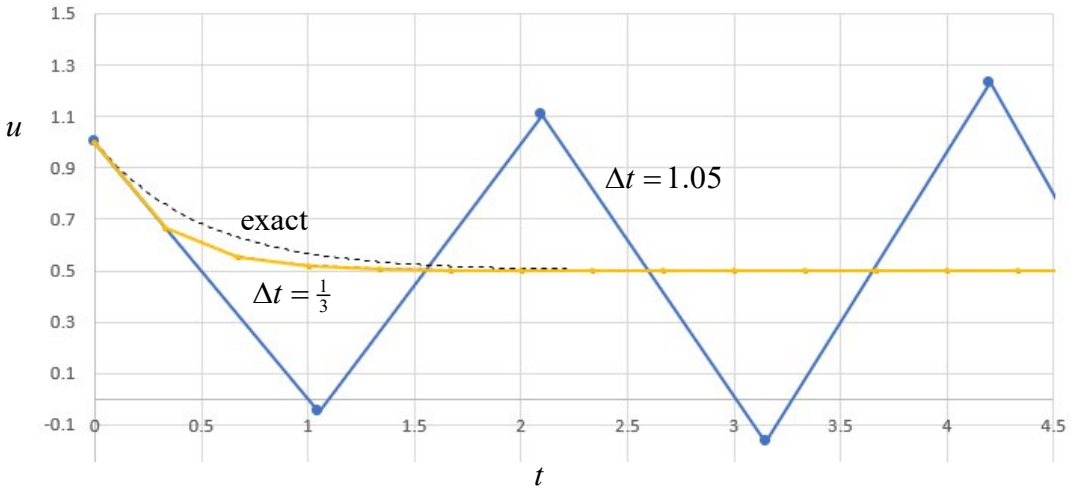


Figure 4.2: Explicit Euler scheme for the solution of an ODE

Matrix Lumping

The inversion of the explicit Euler equations can be greatly speeded up by having **C** diagonal. Altering **C** so that it is diagonal is called **matrix lumping**. There is no one generally accepted method, or theory, of matrix lumping – rather it is an *ad hoc* procedure, which happens not to introduce too significant an error. As an example, considering the earlier example, Eqns (4.11), the usual way to lump the global **C** matrix is as follows:

$$\frac{1}{24}\begin{bmatrix} 4 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 4 \end{bmatrix} \rightarrow \frac{1}{24}\begin{bmatrix} 5 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 5 \end{bmatrix} \tag{4.31}$$

The matrix on the left is called the **consistent matrix**, that on the right the **lumped matrix**.

Stability

The FE equations derived above are exact in the sense that if they are solved exactly, they will give the correct (approximate, FE) solution. However, the equations must be solved numerically, using for example the explicit Euler approximation of the time derivative, Eqn. 4.25. This and other approximations will lead to numerical errors (along with the

inevitable and rounding errors) in the various terms and equations. The question then arises: if there are some numerical/rounding errors in our calculations, will we still get (approximately) the correct solution?

A solution algorithm which is **stable** is one which remains close to the correct solution, i.e. errors in the result at one time step are damped down into the following time steps. An **unstable** algorithm, on the other hand, is one where errors at one time step are magnified in subsequent time-steps, causing the solution to diverge catastrophically.

Examining the one-dimensional case, Eqn. 4.28, $\dot{u} + \lambda u = f$, the general solution is of exponential form. In the case of constant *f*, the solution is $u(t) = \left[u(0) - f/\lambda\right]e^{-\lambda t} + f/\lambda$. We will assume at the outset that $\lambda > 0$; otherwise the solution will grow exponentially and we will usually be interested in practical problems involving physical systems which decay. In that case, the exact solution decays towards the steady-state $u_s = f/\lambda$. The explicit Euler numerical approximation of this exact solution is, on the other hand, given by Eqn. 4.30, $u(n\Delta t) = (1-\lambda\Delta t)^n u(0) + \frac{f}{\lambda}\left[1-(1-\lambda\Delta t)^n\right]$. It can be seen that the term $(1-\lambda\Delta t)^n$ is critical in the sense that, if $|1-\lambda\Delta t| > 1$, this term will grow without bound with successive time steps. Thus it appears that one requires that $|1-\lambda\Delta t| < 1$ for the solution to decay as required, i.e. $-1 < 1-\lambda\Delta t < 1$. With $\lambda > 0$, this implies that we must have

$$\Delta t < \frac{2}{\lambda} \tag{4.32}$$

for the solution to decay "correctly". It is for this reason that the solution diverged in Fig. 4.2 for the case of $\Delta t = 1.05 > 2/\lambda$ with $\lambda = 2$.

To examine the stability of algorithms associated with the general first order in time partial differential equation 4.12, here are listed the FE equations for various internal nodes in the mesh resulting from the use of linear elements (see Eqns. 4.16), neglecting the forcing vector $f(x)$, which does not affect the stability:

$$
\begin{aligned}
\frac{L}{6}\left\{\dot{u}_{i-3}+4\dot{u}_{i-2}+\dot{u}_{i-1}\right\}+\frac{\alpha}{L}\left\{-u_{i-3}+2u_{i-2}-u_{i-1}\right\}=0\\
\frac{L}{6}\left\{\dot{u}_{i-2}+4\dot{u}_{i-1}+\dot{u}_{i}\right\}+\frac{\alpha}{L}\left\{-u_{i-2}+2u_{i-1}-u_{i}\right\}=0\\
\frac{L}{6}\left\{\dot{u}_{i-1}+4\dot{u}_{i}+\dot{u}_{i+1}\right\}+\frac{\alpha}{L}\left\{-u_{i-1}+2u_{i}-u_{i+1}\right\}=0\\
\frac{L}{6}\left\{\dot{u}_{i}+4\dot{u}_{i+1}+\dot{u}_{i+2}\right\}+\frac{\alpha}{L}\left\{-u_{i}+2u_{i+1}-u_{i+2}\right\}=0\\
\frac{L}{6}\left\{\dot{u}_{i+1}+4\dot{u}_{i+2}+\dot{u}_{i+3}\right\}+\frac{\alpha}{L}\left\{-u_{i+1}+2u_{i+2}-u_{i+3}\right\}=0
\end{aligned}
\tag{4.33}
$$

Examining the lumped capacitance matrix, and the explicit Euler representation 4.25:

$$
\begin{aligned}
u_{i-2}\left(t+\Delta t\right)&=ru_{i-3}\left(t\right)+\left(1-2r\right)u_{i-2}\left(t\right)+ru_{i-1}\left(t\right)\\
u_{i-1}\left(t+\Delta t\right)&=ru_{i-2}\left(t\right)+\left(1-2r\right)u_{i-1}\left(t\right)+ru_{i}\left(t\right)\\
u_{i}\left(t+\Delta t\right)&=ru_{i-1}\left(t\right)+\left(1-2r\right)u_{i}\left(t\right)+ru_{i+1}\left(t\right)\\
u_{i+1}\left(t+\Delta t\right)&=ru_{i}\left(t\right)+\left(1-2r\right)u_{i+1}\left(t\right)+ru_{i+2}\left(t\right)\\
u_{i+2}\left(t+\Delta t\right)&=ru_{i+1}\left(t\right)+\left(1-2r\right)u_{i+2}\left(t\right)+ru_{i+3}\left(t\right)
\end{aligned}
\tag{4.34}
$$

where

$$
r=\frac{\alpha\Delta t}{L^2}
\tag{4.35}
$$

Now suppose that the boundary conditions are that the nodal values are all zero. Suppose also that the algorithm begins with all nodes having a value of zero. In that case, one would expect the nodal values to remain at zero for all time. However, let us suppose that we perturb one of the nodes, node *i* say, so that it has a small non-zero value ε. From the nature of the problem, we would expect this nodal value to decay back towards the steady-state solution of zero, and this is what a stable solution will do. From Eqns. 4.33,

$$
\begin{aligned}
u_{i-2}\left(\Delta t\right)&=0 & u_{i-2}\left(2\Delta t\right)&=r^2\varepsilon\\
u_{i-1}\left(\Delta t\right)&=r\varepsilon & u_{i-1}\left(2\Delta t\right)&=2r\left(1-2r\right)\varepsilon\\
u_{i}\left(\Delta t\right)&=\left(1-2r\right)\varepsilon, & u_{i}\left(2\Delta t\right)&=\left[2r^2+\left(1-2r\right)^2\right]\varepsilon,\quad\ldots\\
u_{i+1}\left(\Delta t\right)&=r\varepsilon & u_{i+1}\left(2\Delta t\right)&=2r\left(1-2r\right)\varepsilon\\
u_{i+2}\left(\Delta t\right)&=0 & u_{i+2}\left(2\Delta t\right)&=r^2\varepsilon
\end{aligned}
\tag{4.36}
$$

As can be seen, the error becomes of the order $r^n\varepsilon$ at the nth time step, $t = n\Delta t$. Thus if $r > 1$, the initial small error will magnify without bound as time proceeds. If, on the other hand, $r \le 1$, the initial error will not grow. If $r < 1$, the error will diminish as time proceeds. Even if $r < 1$, there is still the possibility that the solution will oscillate in sign between negative and positive values, because of the $1 - 2r$ term; if $r < \frac{1}{2}$, the error will decrease without oscillation.

One says that the explicit Euler scheme with linear elements is unstable if $\Delta t > L^2 / \alpha$, and stable if

$$\Delta t < \frac{L^2}{\alpha} \tag{4.37}$$

The explicit scheme is **conditionally stable**, since it is only stable provided the time step is less some **critical time step** (or **stability limit**).

The above analysis was done for linear elements with a lumped mass matrix. A similar analysis can be carried our for any type of element or system. It is easier in the general case to examine the stability in terms of the eigenvalues of the system. It will be shown in §4.1.5 below that the Euler-Explicit scheme is more generally stable provided

Stability Requirement for Explicit-Euler:

$$\Delta t < \frac{2}{\lambda_{\max}} \tag{4.38}$$

where $\lambda_{\max}$ is the largest eigenvalue of the system. Further, the solution is non-oscillatory provided $\Delta t \le 1/\lambda_{\max}$. It seems that one has to evaluate the largest eigenvalue of the complete system to determine the critical time-step, but there is a powerful theorem of Linear Algebra which states that *the largest eigenvalue of an assembled system is less than the largest eigenvalue of any of the individual elements* in the model. Thus one need only determine the eigenvalues of the individual elements and use the maximum of these in the stability criterion.

It was mentioned above that as the mesh gets finer, $\lambda_{\max} = \mathrm{O}\left(1/h^2\right)$, where h is a mesh/element length parameter. This puts severe restrictions on the allowable time-step for very fine meshes.

Note the following:

- If one element of **K** is too large or one element of **C** is very small, then the maximum eigenvalue of the system will be increased and hence the critical time-step will be reduced. For this reason it is usual to *keep the FE mesh as uniform as possible*.
- If one uses higher order elements, the entries of **K** and **C** are more varied. It is usual to avoid this variation for the reason stated above, and hence it is typical to *use many lower-order elements* in an FE explicit analysis, rather than fewer higher-order elements.
- For the linear element, Eqn (4.23), $\lambda_{\max} = 12/L^2$. For the lumped **C** matrix one finds that $\lambda_{\max} = 4/L^2$, which allows for a larger time step.

2. Implicit Euler's method

In the implicit Euler scheme, approximate the derivative at time $t + \Delta t$ by the backward difference approximation

$$\dot{u}_{t+\Delta t} = \frac{u_{t+\Delta t} - u_t}{\Delta t} \tag{4.39}$$

In the implicit schemes, the FE equations are written at time $t + \Delta t$,

$$\mathbf{C}\dot{\mathbf{u}}(t+\Delta t) + \mathbf{K}\mathbf{u}(t+\Delta t) = \mathbf{F}(t+\Delta t) \tag{4.40}$$

and then rewritten as

Implicit Euler Algorithm:

$$\begin{aligned} &\overline{\mathbf{K}}\Delta\mathbf{u} = \mathbf{R} \\ &\quad \text{where } \overline{\mathbf{K}} = \mathbf{C} + \Delta t\mathbf{K} \\ &\qquad\quad \mathbf{R} = \Delta t\left[\mathbf{F}(t+\Delta t) - \mathbf{K}\mathbf{u}(t)\right] \\ &\mathbf{u}(t+\Delta t) = \mathbf{u}(t) + \Delta\mathbf{u} \end{aligned} \tag{4.41}$$

It can be shown that this scheme is stable provided $\Delta t\lambda_i \geq 0$. Thus the scheme is **unconditionally stable** provided the eigenvalues are all positive.

3. Semi-Implicit Euler's method

In the semi-implicit method, let

$$\dot{u}_{t+\alpha\Delta t} = \frac{u_{t+\Delta t} - u_t}{\Delta t} \tag{4.42}$$

with $0 \leq \alpha \leq 1$. This is equivalent to taking a linear variation of u between t and $t + \Delta t$, as illustrated below. The FE equations are now written at time $t + \alpha\Delta t$:

$$\mathbf{C}\dot{\mathbf{u}}(t + \alpha\Delta t) + \mathbf{K}\mathbf{u}(t + \alpha\Delta t) = \mathbf{F}(t + \alpha\Delta t) \tag{4.43}$$

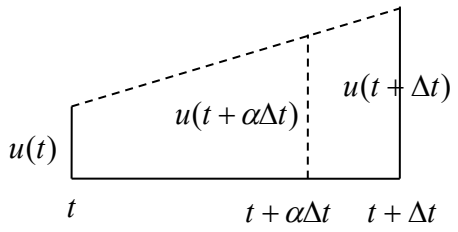


Figure 4.3: semi-implicit definition

Also,

$$\mathbf{u}(t + \alpha\Delta t) = \alpha\mathbf{u}(t + \Delta t) + (1 - \alpha)\mathbf{u}(t) \tag{4.44}$$

and the term $\mathbf{F}(t + \alpha\Delta t)$ is dealt with in a similar manner. This results in the scheme {▲Problem 3}

Semi-Implicit Euler Algorithm:

$$\overline{\mathbf{K}}\Delta\mathbf{u} = \mathbf{R}$$
$$\text{where } \overline{\mathbf{K}} = \mathbf{C} + \alpha\Delta t\mathbf{K}$$
$$\mathbf{R} = \Delta t\{\alpha\mathbf{F}(t+\Delta t) + (1-\alpha)\mathbf{F}(t) - \mathbf{K}\mathbf{u}(t)\} \tag{4.45}$$
$$\mathbf{u}(t+\Delta t) = \mathbf{u}(t) + \Delta\mathbf{u}$$

Note that for {▲Problem 4}

$\alpha = 0$	…	Explicit Euler	truncation error $0(\Delta t)$
$\alpha = \frac{1}{2}$	…	Crank-Nicholson scheme	truncation error $0(\Delta t^2)$
$\alpha = 1$	…	Implicit Euler	truncation error $0(\Delta t)$

For positive definite **C** and **K**, the stability criterion is $\Delta t \le 2/[(1-2\alpha)l_{\max}]$ for $0 \le \alpha < 0.5$; the scheme is unconditionally stable[4] for $\alpha \ge 0.5$. For a stable solution without numerical oscillation, the critical time step is half this value.

4. Predictor-Corrector method

In the predictor-corrector methods, one does the following:
(a) use an explicit formula to predict the first value of $\mathbf{u}(t+\Delta t)$
(b) use an implicit formula to improve that value by an iteration in place

4.1.4 Mode Superposition

A number of direct methods for the integration of the first order system (4.10) have been described above. An alternative solution procedure is the **mode superposition** method. The choice between these two methods is merely one of numerical effectiveness; the solutions obtained using either scheme are identical (if the same integration procedure is used in both). The mode superposition method has the advantage of providing information about the stability of the system (see later).

[4] by which is meant the scheme is stable for any time step. This does not mean that the scheme is accurate for large time-steps, merely that the solution will not diverge dramatically

The basic idea behind mode superposition is this: the FE equations $\mathbf{C}\dot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{F}$ are coupled equations, and to obtain a solution, all *n* equations need to be solved simultaneously. It is possible, however, to rewrite these equations in the form

$$\begin{aligned}
&\dot{z}^{(1)} + \lambda^{(1)} z^{(1)} = f^{(1)}, \qquad && f^{(1)} = \overline{u}_1^{(1)} F_1 + \overline{u}_2^{(1)} F_2 + \cdots \\
&\dot{z}^{(2)} + \lambda^{(2)} z^{(2)} = f^{(2)}, \qquad && f^{(2)} = \overline{u}_1^{(2)} F_1 + \overline{u}_2^{(2)} F_2 + \cdots \\
&\dot{z}^{(3)} + \lambda^{(3)} z^{(3)} = f^{(3)}, \qquad && f^{(3)} = \overline{u}_1^{(3)} F_1 + \overline{u}_2^{(3)} F_2 + \cdots \\
&\cdots
\end{aligned} \tag{4.46}$$

which are *n uncoupled* equations involving the *n* eigenvalues $\lambda^{(j)}$ and eigenvectors $\overline{u}^{(j)}$; each of these equations can be solved *independently* of the others. Once the equations have been solved for the so-called **generalised coordinates** $z^{(j)}$, u can be evaluated through (see below)

$$u_i = \sum_{j=1}^{n} \overline{u}_i^{(j)} z^{(j)} \tag{4.47}$$

that is, by summing up the contributions from all *n* eigenvectors/modes for that node.

The great advantage of the modal superposition method is that not all the equations need to be solved in order to obtain a solution. For example, one might solve the first three equations to obtain $z^{(1)}, z^{(2)}, z^{(3)}$ in which case

$$u_i \approx \overline{u}_i^{(1)} z^{(1)} + \overline{u}_i^{(2)} z^{(2)} + \overline{u}_i^{(3)} z^{(3)} \tag{4.48}$$

In other words, an approximate solution is found which only accounts for a limited number of modes, and it usually the first, limited, number of modes which dominate a solution.

The uncoupled differential equations (4.46) can be solved analytically when **F** is simple, for example when it is a constant or harmonic. For more complicated **F** the equations must be integrated using a numerical procedure, for example one of the direct numerical integration methods discussed earlier.

The modal equations in terms of the generalised coordinates are derived next. This is followed by a detailed example of the mode superposition method.

Derivation of the Modal Equations

Normalise the eigenvectors according to:

$$\overline{\mathbf{u}}^{(i)\mathrm{T}}\mathbf{C}\overline{\mathbf{u}}^{(j)} = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases} \tag{4.49}$$

for $i, j = 1 \cdots n$, the number of degrees of freedom. Define the matrix $\mathbf{\Phi}$ whose columns are the eigenvectors $\overline{\mathbf{u}}^{(i)}$ and the diagonal matrix $\mathbf{\Omega}$ whose elements are the *n* eigenvalues:

$$\mathbf{\Phi} = \begin{bmatrix} \overline{u}_1^{(1)} & \overline{u}_1^{(2)} & \cdots & \overline{u}_1^{(n)} \\ \overline{u}_2^{(1)} & \overline{u}_2^{(2)} & \cdots & \overline{u}_2^{(n)} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{u}_n^{(1)} & \overline{u}_n^{(2)} & \cdots & \overline{u}_n^{(n)} \end{bmatrix}, \qquad \mathbf{\Omega} = \begin{bmatrix} \lambda^{(1)} & 0 & \cdots & 0 \\ 0 & \lambda^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda^{(n)} \end{bmatrix} \tag{4.50}$$

To be clear, the subscripts here refer to the nodal locations, the superscripts refer to a particular mode. The *u* at any node *i* is a linear combination of the individual modal values for that node (*c.f.* Eqn. 4.27)

$$u_i(t) = \beta_1 \overline{u}_i^{(1)} \exp(-\lambda^{(1)} t) + \cdots + \beta_n \overline{u}_i^{(n)} \exp(-\lambda^{(n)} t), \qquad i = 1, 2, \cdots, n \tag{4.51}$$

The *n* solutions to the eigenvalue problem $\left[\mathbf{K} - \lambda^{(j)}\mathbf{C}\right]\overline{\mathbf{u}}^{(j)} = 0$ can be rewritten in the form

$$\mathbf{K\Phi} = \mathbf{C\Phi\Omega} \tag{4.52}$$

With the eigenvectors **C**-orthonormalised as in (4.49), one has $\mathbf{\Phi}^{\mathrm{T}}\mathbf{C\Phi} = \mathbf{I}$ and so, pre-multiplying the above equation by $\mathbf{\Phi}^{\mathrm{T}}$,

$$\mathbf{\Phi}^{\mathrm{T}}\mathbf{K\Phi} = \mathbf{\Omega} \tag{4.53}$$

Introduce now new generalised coordinates **z** such that

$$\mathbf{u}(t) = \mathbf{\Phi}\mathbf{z}(t), \qquad \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} \overline{u}_1^{(1)} & \overline{u}_1^{(2)} & \cdots & \overline{u}_1^{(n)} \\ \overline{u}_2^{(1)} & \overline{u}_2^{(2)} & \cdots & \overline{u}_2^{(n)} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{u}_n^{(1)} & \overline{u}_n^{(2)} & \cdots & \overline{u}_n^{(n)} \end{bmatrix} \begin{bmatrix} z^{(1)} \\ z^{(2)} \\ \vdots \\ z^{(n)} \end{bmatrix} \tag{4.54}$$

Then, pre-multiplying the equations (4.10) by $\mathbf{\Phi}^{\mathrm{T}}$ and using (4.54) gives

$$\begin{aligned} &\mathbf{\Phi}^{\mathrm{T}}\mathbf{C}\mathbf{\Phi}\dot{\mathbf{z}} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{K}\mathbf{\Phi}\mathbf{z} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{F} \\ &\rightarrow \\ &\dot{\mathbf{z}} + \mathbf{\Omega}\mathbf{z} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{F} \end{aligned} \tag{4.55}$$

These equations are the uncoupled equations (4.46) given at the beginning of this subsection. Each equation can be integrated in turn to evaluate the coordinates $z^{(j)}$, whence the u_i can be evaluated through (4.54). For this purpose one needs the initial conditions on $\mathbf{z}(t)$. Since $\mathbf{\Phi}^{\mathrm{T}}\mathbf{C}\mathbf{\Phi} = \mathbf{I}$, then $\mathbf{u}(t) = \mathbf{\Phi}\,\mathbf{z}(t)$ becomes $\mathbf{\Phi}^{\mathrm{T}}\mathbf{C}\mathbf{u}(t) = \mathbf{z}(t)$ so that

$$\mathbf{z}(0) = \mathbf{\Phi}^{\mathrm{T}}\mathbf{C}\mathbf{u}(0) \tag{4.56}$$

Example

Consider the following problem

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0, \quad \text{B.C.} \quad \begin{matrix} u(0,t) = 0 \\ \partial u/\partial x(1,t) = 1 \end{matrix}, \quad \text{I.C.} \quad u(x,0) = \sin(-2.074x) \tag{4.57}$$

Using two linear elements, with $L = 1/2$, and applying the essential BC at $x = 0$, leads to the eigenvalues and eigenvectors:

$$\begin{aligned} &\mathbf{C} = \frac{1}{12}\begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}, \quad \mathbf{K} = 2\begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}, \quad \left|\mathbf{K} - \lambda\mathbf{C}\right| = 4 - \tfrac{5}{3}\lambda + \tfrac{7}{144}\lambda^2 = 0 \\ &\rightarrow \lambda^{(1)} = 2.597, \qquad \lambda^{(2)} = 31.689 \\ &\overline{\mathbf{u}}^{(1)} : \left[\mathbf{K} - 2.597\mathbf{C}\right]\overline{\mathbf{u}} = 0 \rightarrow \overline{\mathbf{u}}^{(1)} = \begin{bmatrix} 1 \\ \sqrt{2} \end{bmatrix} \\ &\overline{\mathbf{u}}^{(2)} : \left[\mathbf{K} - 31.689\mathbf{C}\right]\overline{\mathbf{u}} = 0 \rightarrow \overline{\mathbf{u}}^{(2)} = \begin{bmatrix} 1 \\ -\sqrt{2} \end{bmatrix} \end{aligned} \tag{4.58}$$

Write the eigenvectors as

$$\overline{\mathbf{u}}^{(1)} = \eta^{(1)}\begin{bmatrix} 1 \\ \sqrt{2} \end{bmatrix}, \qquad \overline{\mathbf{u}}^{(2)} = \eta^{(2)}\begin{bmatrix} 1 \\ -\sqrt{2} \end{bmatrix} \tag{4.59}$$

with $\eta^{(1)}, \eta^{(2)}$ to be determined. Normalising according to (4.54) leads to four equations which can be used to obtain {▲Problem 5}

$$\eta^{(1)} = \sqrt{\frac{6}{4+\sqrt{2}}} \approx 1.053, \qquad \eta^{(2)} = \sqrt{\frac{6}{4-\sqrt{2}}} \approx 1.523 \qquad \text{and} \qquad \eta^{(1)}\eta^{(2)} = 0 \quad (4.60)$$

Form the matrices

$$\mathbf{\Phi} = \begin{bmatrix} 1.053 & 1.523 \\ 1.489 & -2.154 \end{bmatrix}, \quad \mathbf{\Omega} = \begin{bmatrix} 2.597 & 0 \\ 0 & 31.689 \end{bmatrix} \tag{4.61}$$

With the natural BC at $x = 1$, constant over time, $\partial u / \partial x(1,t) = 1$, the **F** vector is

$$\mathbf{F} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{4.62}$$

The modal equations are

$$\begin{aligned} &\dot{\mathbf{z}} + \mathbf{\Omega z} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{F} \\ &\rightarrow \begin{bmatrix} \dot{z}^{(1)} \\ \dot{z}^{(2)} \end{bmatrix} + \begin{bmatrix} 2.597 & 0 \\ 0 & 31.689 \end{bmatrix} \begin{bmatrix} z^{(1)} \\ z^{(2)} \end{bmatrix} = \begin{bmatrix} 1.053 & 1.523 \\ 1.489 & -2.154 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \end{aligned} \tag{4.63}$$

or

$$\begin{aligned} \dot{z}^{(1)} + 2.597 z^{(1)} &= 1.523 \\ \dot{z}^{(2)} + 31.689 z^{(1)} &= -2.154 \end{aligned} \tag{4.64}$$

These are first order ODEs which can be solved for $z^{(i)}$:

$$\begin{aligned} z^{(1)} &= +0.586 + Ae^{-2.597t} \\ z^{(2)} &= -0.068 + Be^{-31.689t} \end{aligned} \tag{4.65}$$

From the initial condition $u(x,0) = \sin(-2.074x)$:

$$\mathbf{z}(0)=\begin{bmatrix} z^{(1)}(0) \\ z^{(2)}(0) \end{bmatrix}=\mathbf{\Phi}^{\mathrm{T}}\mathbf{C}\mathbf{u}(0)=\begin{bmatrix} 1.053 & 1.489 \\ 1.523 & -2.154 \end{bmatrix}\frac{1}{12}\begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} u_1(\frac{1}{2},0) \\ u_2(1,0) \end{bmatrix}$$
$$=\begin{bmatrix} 0.475 & 0.336 \\ 0.328 & -0.232 \end{bmatrix}\begin{bmatrix} -0.861 \\ -0.876 \end{bmatrix} \tag{4.66}$$
$$=\begin{bmatrix} -0.703 \\ -0.079 \end{bmatrix}$$

Using these initial conditions leads to evaluation of the constants A and B:

$$\begin{aligned} z^{(1)} &= +0.586-1.289e^{-2.597t} \\ z^{(2)} &= -0.068-0.011e^{-31.689t} \end{aligned} \tag{4.67}$$

Finally, the values of u_i are obtained through $\mathbf{u}(t)=\mathbf{\Phi}\,\mathbf{z}(t)$:

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}=\begin{bmatrix} 1.053 & 1.523 \\ 1.489 & -2.154 \end{bmatrix}\begin{bmatrix} z^{(1)} \\ z^{(2)} \end{bmatrix}=\begin{bmatrix} 0.513-1.357e^{-2.597t}-0.017e^{-31.689t} \\ 1.019-1.919e^{-2.597t}+0.024e^{-31.689t} \end{bmatrix} \tag{4.68}$$

4.1.5 Stability

In the following, the stability criterion Eqn. 4.38 for the Explicit Euler scheme, Eqn. 4.27, is derived (but the same methods may be used to analyse any numerical integration scheme).

The mode superposition and direct integration methods both involve the integration of differential equations. They are two slightly different ways of solving the same problem; if one supposes that an FE problem is solved using both methods, and (4.10, 4.46) are both solved using the same numerical scheme, each with the same time step Δt, *both methods are completely equivalent*. Therefore, to study the accuracy of direct integration, one may focus on and estimate the accuracy and stability of integration of the modal equations, $\dot{\mathbf{z}}+\mathbf{\Omega}\mathbf{z}=\mathbf{\Phi}^{\mathrm{T}}\mathbf{F}$, which is an easier task. Furthermore, since all the modal equations are similar one need only examine one typical equation, which may be written as

$$\dot{z}+\lambda z=f \tag{4.69}$$

In fact, this is just the one-dimensional equation considered earlier, Eqn. 4.28, and the explicit Euler scheme was examined in relation to this equation in Eqns. 4.29-4.30. Nevertheless, although the following is repetition to a large extent, we will examine it again anew in the current context.

Stability is determined by examining the numerical solution for arbitrary initial conditions. One may consider the case of $f = 0$, and one sees that the stability and accuracy depends on the eigenvalue λ and whatever time-step is used. Thus, considering the homogeneous modal equation

$$\dot{z} + \lambda z = 0 \tag{4.70}$$

Separating variables and solving gives the general solution

$$z = Ae^{-\lambda t} \tag{4.71}$$

Starting with initial condition $z(t)$ at time t, one has $z(t+\Delta t) = z(t)\exp\{-\lambda \Delta t\}$. Regardless of the time-stepping algorithm used, then, one requires for a stable solution:

$$\begin{aligned} |z(t+\Delta t)| &< |z(t)| \qquad \lambda > 0 \\ z(t+\Delta t) &= z(t) \qquad \lambda = 0 \end{aligned} \tag{4.72}$$

with instability for $\lambda < 0$.

Examining now the explicit Euler scheme, replace the $\dot{z}$ in Eqn. 4.70 with $\left[z(t+\Delta t) - z(t)\right]/\Delta t$, leading to

$$z(t+\Delta t) = A\, z(t) \tag{4.73}$$

where A is the **amplification factor** (so called, since any errors at one time step will be magnified by this amount into the next time step)

$$A = 1 - \lambda \Delta t \tag{4.74}$$

When $\lambda = 0$, $z(t+\Delta t) = z(t)$ as before and, when $\lambda > 0$, in order that $|z(t+\Delta t)| < |z(t)|$, it is required that $|A| < 1$, or $-1 < 1 - \lambda \Delta t < 1$. The inequality on the right is always satisfied; the left-hand inequality leads to the condition

$$\Delta t < \frac{2}{\lambda} \tag{4.75}$$

This stability condition must hold for all modes in the system. The largest eigenvalue $\lambda_{\max}$ imposes the greatest restriction, leading to the criterion (4.38).

4.2 Second-Order Systems

Here, the hyperbolic second order system (4.4) is examined. In particular, consider the following problem:

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0, \qquad \text{B.C.} \quad \begin{matrix} u(0,t) = 0 \\ \dfrac{\partial u}{\partial x}(l,t) = 0 \end{matrix}, \qquad \text{I.C.} \quad \begin{matrix} u(x,0) = 0 \\ \dfrac{\partial u}{\partial t}(x,0) = \dfrac{2x}{l} \end{matrix} \tag{4.76}$$

4.2.1 FE equations for Second Order Systems

Using the Galerkin procedure to discretise the space variable, one has for a linear element,

$$\frac{\partial^2 u_i}{\partial t^2} \int_{x_i}^{x_{i+1}} N_1 N_j dx + \frac{\partial^2 u_{i+1}}{\partial t^2} \int_{x_i}^{x_{i+1}} N_2 N_j dx + u_i c^2 \int_{x_i}^{x_{i+1}} \frac{\partial N_1}{\partial x}\frac{\partial N_j}{\partial x} dx + u_{i+1} c^2 \int_{x_i}^{x_{i+1}} \frac{\partial N_2}{\partial x}\frac{\partial N_j}{\partial x} dx = c^2 \left[\frac{\partial u}{\partial x} N_j \right]_{x_i}^{x_{i+1}}$$

$$j = 1,2 \tag{4.77}$$

Evaluating the integrals leads to the element equations[5]

$$\frac{L}{6}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} \ddot{u}_i \\ \ddot{u}_{i+1} \end{bmatrix} + c^2 \frac{1}{L}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} u_i \\ u_{i+1} \end{bmatrix} = c^2 \begin{bmatrix} -u'(x_i) \\ +u'(x_{i+1}) \end{bmatrix} \tag{4.78}$$

element mass matrix **M_e**

element stiffness matrix **K_e**

After assembly, one has the system of second order ODEs of the form

[5] the first matrix here is called the *mass* matrix, so-called because of its physical relevance in elastodynamic problems (see later)

$$\frac{L}{6}\begin{bmatrix} 2 & 1 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & 1 & 2 \end{bmatrix}\begin{bmatrix} \ddot{u}_1 \\ \ddot{u}_2 \\ \ddot{u}_3 \\ \vdots \\ \ddot{u}_{n+1} \end{bmatrix} + \frac{c^2}{L}\begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{n+1} \end{bmatrix} = c^2\begin{bmatrix} -u'(0) \\ 0 \\ 0 \\ \vdots \\ +u'(l) \end{bmatrix} \tag{4.79}$$

or, in short,

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{F} \tag{4.80}$$

which can be solved in a number of different ways (see below).

4.2.2 Eigenvalues and Eigenvectors

As with the first order system, an eignenvalue analysis can be carried out for the system $\mathbf{M}\ddot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{F}$, and will tell us much useful information. First, consider the single degree of freedom model, $m\ddot{u} + ku = f$, with f constant (see the Appendix to his Chapter), which has the solution

$$u(t) = A\sin(\omega t + \phi) + \frac{f}{k} \tag{4.81}$$

This is an oscillation at **natural frequency** ω about the mean position f/k (which is the solution to the time independent "static" equation $ku = f$). A solution can be obtained for the complete system by assuming that it also oscillates about some mean configuration $\mathbf{u}_m = \mathbf{K}^{-1}\mathbf{F}$,

$$\mathbf{u}(t) = \mathbf{u}_m + \bar{\mathbf{u}}\sin(\omega t + \phi) \tag{4.82}$$

Substitution into the FE equations (4.76) gives

$$\left[\mathbf{K} - \omega^2\mathbf{M}\right]\bar{\mathbf{u}} = 0 \tag{4.83}$$

and this system of $n \times n$ equations in the n entries of $\bar{\mathbf{u}} = \begin{bmatrix}\bar{u}_1 & \bar{u}_2 & \cdots & \bar{u}_n\end{bmatrix}^{\mathrm{T}}$ has a solution only if the determinant of the coefficient matrix is zero:

$$\left|\mathbf{K}-\omega^2\mathbf{M}\right|=0 \tag{4.84}$$

This equation can be solved for the *n* eigenvalues ω^2.

Using two linear elements for the example problem (4.76), applying the boundary condition $u_1=u(0,t)=0$ ($\ddot{u}_1=\ddot{u}(0,t)=0$), and eliminating the first row and column, leads to the equations

$$\mathbf{M}\ddot{\mathbf{u}}+\mathbf{K}\mathbf{u}=\mathbf{F},\qquad \mathbf{M}=\frac{l}{12}\begin{bmatrix}4 & 1\\ 1 & 2\end{bmatrix},\quad \mathbf{u}=\begin{bmatrix}u_2\\ u_3\end{bmatrix},\quad \mathbf{K}=\frac{2c^2}{l}\begin{bmatrix}2 & -1\\ -1 & 1\end{bmatrix},\quad \mathbf{F}=\begin{bmatrix}0\\ 0\end{bmatrix} \tag{4.85}$$

where $l=2L$ and so

$$\left|\mathbf{K}-\omega^2\mathbf{M}\right|=\mathbf{0}\quad\rightarrow\quad\frac{2c^2}{l}\begin{vmatrix}2-4\alpha & -1-\alpha\\ -1-\alpha & 1-2\alpha\end{vmatrix}=0,\quad \alpha=\frac{\omega^2l^2}{24c^2} \tag{4.86}$$

Thus $7\alpha^2-10\alpha+1=0$ which yields $\alpha=\left(5\pm\sqrt{18}\right)/7$ and the square-roots of the eigenvalues, the natural frequencies, are then

$$\begin{aligned}\omega^{(1)}&=1.61142\frac{c}{l},\quad \omega^{(2)}=5.62930\frac{c}{l}\\ &=0.8057\frac{c}{L},\qquad\quad =2.8147\frac{c}{L}\end{aligned} \tag{4.87}$$

Note that the same eigenvalues would be obtained from the 3×3 global system (after applying the essential BC but *not* eliminating a row and column); the third eigenvalue would be $\omega=1$.

This eigenvalue analysis will be continued further below in section 4.2.5, in the context of the elastodynamic problem, where the eigenvalues and eigenvectors have a specific physical meaning.

4.2.3 Direct Integration for Second Order Systems

As with first order systems, one can solve (4.80) using either one of many direct integration methods or through mode superposition. The direct integration methods are discussed here.

As with first order systems, a number of different direct integration methods are available for the integration of the second order system (4.79), for example,

1. Explicit Central Difference Scheme
2. Linear Acceleration Scheme (Implicit)
3. Wilson θ Scheme (Implicit)
4. Newmark Scheme (Implicit)
5. Trapezoidal Scheme (Implicit)

1. Explicit Central Difference Scheme

In the explicit central difference scheme, one expands the unknown nodal functions $u_i(t)$ in Taylor series:

$$\begin{aligned} u_i(t+\Delta t) &= u_i(t) + \Delta t\dot{u}_i(t) + \frac{1}{2}(\Delta t)^2\ddot{u}_i(t) + \cdots \\ u_i(t-\Delta t) &= u_i(t) - \Delta t\dot{u}_i(t) + \frac{1}{2}(\Delta t)^2\ddot{u}_i(t) + \cdots \end{aligned} \tag{4.88}$$

Adding and subtracting these expressions then lead to the following approximations for the derivatives:

$$\begin{aligned} \ddot{u}_i(t) &= \frac{u_i(t+\Delta t) - 2u_i(t) + u_i(t-\Delta t)}{(\Delta t)^2} \\ \dot{u}_i(t) &= \frac{u_i(t+\Delta t) - u_i(t-\Delta t)}{2\Delta t} \end{aligned} \tag{4.89}$$

Being an explicit scheme, the FE equations are considered at time *t*,

$$\mathbf{M}\ddot{\mathbf{u}}(t) + \mathbf{K}\mathbf{u}(t) = \mathbf{F}(t) \tag{4.90}$$

Substituting in the approximate expressions for the derivatives leads to

Explicit Central Difference Scheme:

$$\left(\frac{1}{(\Delta t)^2}\mathbf{M}\right)\mathbf{u}(t+\Delta t) = \hat{\mathbf{F}}(t)$$
$$\hat{\mathbf{F}}(t) = \mathbf{F}(t) - \left(\mathbf{K} - \frac{2}{(\Delta t)^2}\mathbf{M}\right)\mathbf{u}(t) - \left(\frac{1}{(\Delta t)^2}\mathbf{M}\right)\mathbf{u}(t-\Delta t) \tag{4.91}$$

To start the scheme, one needs the value of $\mathbf{u}(-\Delta t)$. To obtain this value, note that $\mathbf{u}(0), \dot{\mathbf{u}}(0)$ are known from the initial conditions, and one can hence obtain $\ddot{\mathbf{u}}(0)$ from the equations $\mathbf{M}\ddot{\mathbf{u}}(0) + \mathbf{K}\mathbf{u}(0) = \mathbf{F}(0)$. One can then re-arrange the approximate expressions for $\dot{\mathbf{u}}(t), \ddot{\mathbf{u}}(t)$ above to obtain

$$\begin{aligned}\mathbf{u}(-\Delta t) &= \mathbf{u}(0) - \Delta t\dot{\mathbf{u}}(0) + \frac{1}{2}(\Delta t)^2\ddot{\mathbf{u}}(0) \\ &= \mathbf{u}(0) - \Delta t\dot{\mathbf{u}}(0) + \frac{1}{2}(\Delta t)^2\mathbf{M}^{-1}\left[\mathbf{F}(0) - \mathbf{K}\mathbf{u}(0)\right]\end{aligned} \tag{4.92}$$

Considering a one-dimensional case for illustrative purposes, consider the ODE

$$\frac{d^2u}{dt^2} + \omega^2 u = f, \quad \dot{u}(0) = \overline{v}_0, \quad u(0) = \overline{u}_0 \tag{4.93}$$

Substituting $\ddot{u}_t = \left[u_{t+\Delta t} - 2u_t + u_{t-\Delta t}\right]/(\Delta t)^2$ into Eqn. 4.93 gives

$$u(t+\Delta t) = \left[2 - \omega^2(\Delta t)^2\right]u(t) - u(t-\Delta t) + f(\Delta t)^2 \tag{4.94}$$

It is convenient to express the relationship between the values at the different time-steps in the form of the matrix recursive algorithm:

$$\begin{bmatrix} u(t+\Delta t) \\ u(t) \end{bmatrix} = \begin{bmatrix} 2-\omega^2\Delta t^2 & -1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} u(t) \\ u(t-\Delta t) \end{bmatrix} + \begin{bmatrix} \Delta t^2 \\ 0 \end{bmatrix} f(t) \tag{4.95}$$

To keep things simple, let $f = 0$, so that at any time $t = n\Delta t$, the solution is given by

$$\begin{bmatrix} u(n\Delta t) \\ u((n-1)\Delta t) \end{bmatrix} = \mathbf{A}\begin{bmatrix} u((n-1)\Delta t) \\ u((n-2)\Delta t) \end{bmatrix} = \mathbf{A}^n \begin{bmatrix} u(0) \\ u(-\Delta t) \end{bmatrix} \tag{4.96}$$

where

$$\mathbf{A} = \begin{bmatrix} 2-\omega^2\Delta t^2 & -1 \\ 1 & 0 \end{bmatrix} \tag{4.97}$$

The value of $u(-\Delta t)$ can be obtained in the same way as was Eqn. 4.92:

$$u(-\Delta t) = \left[1-\tfrac{1}{2}\omega^2(\Delta t)^2\right]u(0) - \Delta t\dot{u}(0) \tag{4.98}$$

This solution is plotted in Fig. 4.4 for $\omega = 2$, $\dot{u}(0) = 3/2$; for $\Delta t = 0.5$ and $\Delta t = 1.05$, together with the exact solution $u(t) = \frac{3}{4}\sin(2t)$. The solution is quite accurate for $\Delta t = 0.5$. On the other hand, when the time step is as large as $\Delta t = 1$, the solution is highly inaccurate; this issue is discussed further below.

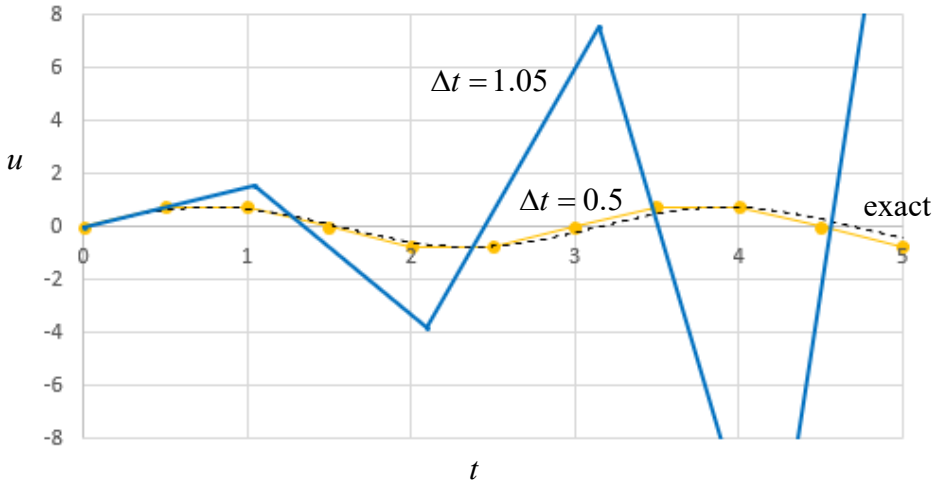


Figure 4.4: Explicit Central Difference scheme for the solution of an ODE

Matrix Lumping

Analogous to the first order system, the inversion of the explicit central difference equations can be greatly speeded up by having **M** diagonal, that is by lumping the **M** matrix.

Stability

Examining the one-dimensional case, Eqn. 4.93, $\ddot{u} + \omega^2 u = f$, the general solution is periodic in form. In the case of constant *f*, the solution is

$$u(t) = \left[u(0) - f/\omega^2\right]\cos(\omega t) + \left[\dot{u}(0)/\omega\right]\sin(\omega t) + f/\omega^2 \tag{4.99}$$

The solution is seen to oscillate about $u = f/\omega^2$. The explicit Central Difference numerical approximation of this exact solution is, on the other hand, given by Eqn. 4.95. In the case of $f = 0$, it is given by Eqn. 4.96. The matrix **A**, Eqn. 4.97, is clearly critical to the stability of the numerical scheme. It is helpful now to decompose the matrix **A** into its **eigendecomposition** (**spectral decomposition**): $\mathbf{A} = \mathbf{PJP}^{-1}$. Here, **J** is the diagonal matrix of eigenvalues and **P** is the matrix of eigenvectors (columns of **P** are the eigenvectors). This decomposition has the special property that $\mathbf{A}^n = \mathbf{PJ}^n\mathbf{P}^{-1}$. Evaluating the eigenvalues and eigenvectors of **A**, one has

$$\mathbf{A} = \begin{bmatrix} 2-\alpha & -1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} \frac{2-\alpha}{2} + i\Delta & \frac{2-\alpha}{2} - i\Delta \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{2-\alpha}{2} + i\Delta & 0 \\ 0 & \frac{2-\alpha}{2} - i\Delta \end{bmatrix} \begin{bmatrix} -i\frac{1}{2\Delta} & \frac{1}{2} + i\frac{2-\alpha}{4\Delta} \\ +i\frac{1}{2\Delta} & \frac{1}{2} - i\frac{2-\alpha}{4\Delta} \end{bmatrix} \tag{4.100}$$

where

$$\Delta = \sqrt{1 - \frac{(2-\alpha)^2}{4}}, \qquad \alpha = \omega^2(\Delta t)^2 \tag{4.101}$$

Thus, from Eqn. 4.96,

$$\begin{bmatrix} u(n\Delta t) \\ u((n-1)\Delta t) \end{bmatrix} = \begin{bmatrix} \frac{2-\alpha}{2} + i\Delta & \frac{2-\alpha}{2} - i\Delta \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{2-\alpha}{2} + i\Delta & 0 \\ 0 & \frac{2-\alpha}{2} - i\Delta \end{bmatrix}^n \begin{bmatrix} -i\frac{1}{2\Delta} & \frac{1}{2} + i\frac{2-\alpha}{4\Delta} \\ +i\frac{1}{2\Delta} & \frac{1}{2} - i\frac{2-\alpha}{4\Delta} \end{bmatrix} \begin{bmatrix} u(0) \\ u(-\Delta t) \end{bmatrix} \tag{4.102}$$

The eigenvalues are complex for $\alpha < 4$. In fact, for $\alpha < 4$, the absolute value of the eigenvalues is always 1:

$$\left| \frac{2-\alpha}{2} \pm i\sqrt{1 - \frac{(2-\alpha)^2}{4}} \right| = 1 \tag{4.103}$$

in which case $\mathbf{A}^n$ is bounded as n increases. In the case of $\alpha > 4$, the eigenvalues are real and the magnitude is always greater than 1, in which case $\mathbf{A}^n$ becomes unbounded. The criterion for stability is therefore that $\alpha < 4$, or

$$\Delta t < \frac{2}{\omega} \tag{4.104}$$

This criterion is seen to be satisfied in Fig. 4.4, for which $\omega = 2$, so that the stability criterion is $\Delta t < 1$.

More generally, as with the explicit Euler scheme for the first order system discussed in section 4.1.3, the stability of the explicit Central Difference scheme can be examined by considering the system of equations for a general differential equation of the form 4.76. The FE equations for various internal nodes in the mesh resulting from the use of linear elements and a lumped mass matrix are:

$$\begin{aligned}
\ddot{u}_{i-2} + \frac{c^2}{L^2}\{-u_{i-3} + 2u_{i-2} - u_{i-1}\} &= 0 \\
\ddot{u}_{i-1} + \frac{c^2}{L^2}\{-u_{i-2} + 2u_{i-1} - u_i\} &= 0 \\
\ddot{u}_i + \frac{c^2}{L^2}\{-u_{i-1} + 2u_i - u_{i+1}\} &= 0 \\
\ddot{u}_{i+1} + \frac{c^2}{L^2}\{-u_i + 2u_{i+1} - u_{i+2}\} &= 0 \\
\ddot{u}_{i+2} + \frac{c^2}{L^2}\{-u_{i+1} + 2u_{i+2} - u_{i+3}\} &= 0
\end{aligned} \tag{4.105}$$

Using the Central Difference approximation, Eqn. 4.89,

$$
\begin{aligned}
u_{i-2}(t+\Delta t) &= -u_{i-2}(t-\Delta t) + ru_{i-3}(t) + 2(1-r)u_{i-2}(t) + ru_{i-1}(t) \\
u_{i-1}(t+\Delta t) &= -u_{i-1}(t-\Delta t) + ru_{i-2}(t) + 2(1-r)u_{i-1}(t) + ru_{i}(t) \\
u_{i}(t+\Delta t) &= -u_{i}(t-\Delta t) + ru_{i-1}(t) + 2(1-r)u_{i}(t) + ru_{i+1}(t) \\
u_{i+1}(t+\Delta t) &= -u_{i+1}(t-\Delta t) + ru_{i}(t) + 2(1-r)u_{i+1}(t) + ru_{i+2}(t) \\
u_{i+2}(t+\Delta t) &= -u_{i+2}(t-\Delta t) + ru_{i+1}(t) + 2(1-r)u_{i+2}(t) + ru_{i+3}(t)
\end{aligned}
\tag{4.106}
$$

where

$$
r = \frac{c^2(\Delta t)^2}{L^2} \tag{4.107}
$$

Suppose now that we begin the algorithm with all nodal values and initial conditions zero except for node *i*, which is given a small non-zero value ε. From Eqns. 4.106,

$$
\begin{array}{lll}
u_{i-2}(\Delta t) = 0 & u_{i-2}(2\Delta t) = r^2\varepsilon & u_{i-2}(3\Delta t) = \left(-6r^3 + 6r^2\right)\varepsilon \\
u_{i-1}(\Delta t) = r\varepsilon & u_{i-1}(2\Delta t) = \left(-4r^2 + 4r\right)\varepsilon & u_{i-1}(3\Delta t) = \left(15r^3 - 24r^2 + 10r\right)\varepsilon \\
u_{i}(\Delta t) = 2(1-r)\varepsilon, & u_{i}(2\Delta t) = \left(6r^2 - 8r + 3\right)\varepsilon, & u_{i}(3\Delta t) = \left(-28r^3 + 44r^2 - 12r - 4\right)\varepsilon \\
u_{i+1}(\Delta t) = r\varepsilon & u_{i+1}(2\Delta t) = \left(-4r^2 + 4r\right)\varepsilon & u_{i+1}(3\Delta t) = \left(15r^3 - 24r^2 + 10r\right)\varepsilon \\
u_{i+2}(\Delta t) = 0 & u_{i+2}(2\Delta t) = r^2\varepsilon & u_{i+2}(3\Delta t) = \left(-6r^3 + 6r^2\right)\varepsilon
\end{array}
\tag{4.108}
$$

As can be seen, the error becomes of the order $r^n\varepsilon$ at the *n*th time step, $t = n\Delta t$. In the same way as with the explicit Euler scheme earlier, it can be seen that the explicit Central Difference scheme for linear elements is conditionally stable, with stability for

$$
\Delta t < \frac{L}{c} \tag{4.109}
$$

The above analysis was done for linear elements with a lumped mass matrix. More generally, as proved below, the Explicit Central Difference scheme is stable provided

Stability Requirement for Explicit Central Difference Scheme:

$$\Delta t < \frac{2}{\omega_{\max}} \tag{4.110}$$

where $\omega_{\max}$ is the largest natural frequency of the system. As for the explicit Euler scheme, one need only determine the natural frequency of the individual elements and use the maximum of these in the stability criterion.

For the same reasons given regarding the first order system, when using the Central Difference explicit scheme, the mesh should be kept as regular as possible and low-order (linear) elements should be used if possible.

2. Linear Acceleration Scheme (Implicit)

Here, suppose that the quantities $u, \dot{u}, \ddot{u}$ at time t are known. To find the values of these quantities a time Δt later, assume a linearly varying "acceleration"[6] $\ddot{u}$ in the time step. Let τ be the increase in time starting at time t. From Fig. 4.5,

$$\ddot{u}(t+\tau) = \ddot{u}(t) + \frac{\tau}{\Delta t}\left(\ddot{u}(t+\Delta t) - \ddot{u}(t)\right) \tag{4.11}$$

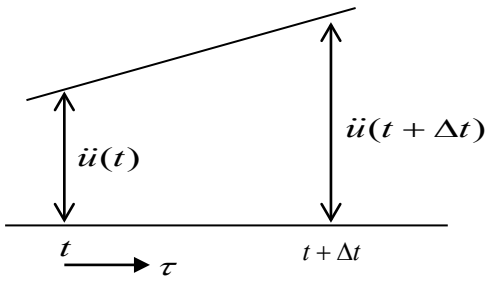


Figure 4.5: Linear Acceleration Scheme

Integration with respect to τ then gives (the $\dot{u}(t), u(t)$ terms are constants of integration)

[6] this terminology assumes that u represents a "displacement"

$$\begin{aligned}\dot{u}(t+\tau) &= \dot{u}(t) + \tau\ddot{u}(t) + \frac{\tau^2}{2\Delta t}\left(\ddot{u}(t+\Delta t) - \ddot{u}(t)\right) \\ u(t+\tau) &= u(t) + \tau\dot{u}(t) + \frac{\tau^2}{2}\ddot{u}(t) + \frac{\tau^3}{6\Delta t}\left(\ddot{u}(t+\Delta t) - \ddot{u}(t)\right)\end{aligned} \tag{4.112}$$

Substituting $\tau = \Delta t$ into these equations and rearranging gives

$$\begin{aligned}\dot{u}(t+\Delta t) &= \dot{u}(t) + \frac{\Delta t}{2}\left(\ddot{u}(t+\Delta t) + \ddot{u}(t)\right) \\ \ddot{u}(t+\Delta t) &= \frac{6}{(\Delta t)^2}\left(u(t+\Delta t) - u(t)\right) - \frac{6}{\Delta t}\dot{u}(t) - 2\ddot{u}(t)\end{aligned} \tag{4.113}$$

Substituting the latter equation into the former finally leads to the expressions

$$\begin{aligned}\dot{u}(t+\Delta t) &= \frac{3}{\Delta t}\left(u(t+\Delta t) - u(t)\right) - 2\dot{u}(t) - \frac{\Delta t}{2}\ddot{u}(t) \\ \ddot{u}(t+\Delta t) &= \frac{6}{(\Delta t)^2}\left(u(t+\Delta t) - u(t)\right)\frac{6}{\Delta t}\dot{u}(t) - 2\ddot{u}(t)\end{aligned} \tag{4.114}$$

The FE equations 4.80 are written at time $t + \Delta t$,

$$\mathbf{M}\ddot{\mathbf{u}}(t+\Delta t) + \mathbf{K}\mathbf{u}(t+\Delta t) = \mathbf{F}(t+\Delta t) \tag{4.115}$$

The use of Eqn. 4.114 then leads to

Linear Acceleration Scheme:

$$\left[\mathbf{M}\frac{6}{(\Delta t)^2} + \mathbf{K}\right]\mathbf{u}(t+\Delta t) = \mathbf{F}(t+\Delta t) + \mathbf{M}\left[\frac{6}{(\Delta t)^2}\mathbf{u}(t) + \frac{6}{\Delta t}\dot{\mathbf{u}}(t) + 2\ddot{\mathbf{u}}(t)\right] \tag{4.116}$$

Once $\mathbf{u}(t+\Delta t)$ is obtained, then $\dot{\mathbf{u}}(t+\Delta t)$ and $\ddot{\mathbf{u}}(t+\Delta t)$ can be obtained from (4.116). The scheme is unconditionally stable.

3. The Wilson θ Scheme (Implicit)

Here, assume a linearly varying $\ddot{u}$ in the time step, but now extrapolate the hypothetical solution out to time $t+\theta\Delta t$, where θ is some parameter ($\theta \geq 1$), as illustrated in Fig. 4.6. When $\theta = 1$ the method reduces to the linear acceleration scheme. Then

$$\ddot{u}(t+\tau) = \ddot{u}(t) + \frac{\tau}{\theta\Delta t}\left(\ddot{u}(t+\theta\Delta t) - \ddot{u}(t)\right) \tag{4.117}$$

Integration with respect to τ, the substitution $\tau = \theta\Delta t$, and some rearranging then leads to {▲Problem 9}

$$\begin{aligned}
\dot{u}(t+\theta\Delta t) &= \frac{3}{\theta\Delta t}\left(u(t+\theta\Delta t) - u(t)\right) - 2\dot{u}(t) - \frac{\theta\Delta t}{2}\ddot{u}(t) \\
\ddot{u}(t+\theta\Delta t) &= \frac{6}{(\theta\Delta t)^2}\left(u(t+\theta\Delta t) - u(t)\right) - \frac{6}{\theta\Delta t}\dot{u}(t) - 2\ddot{u}(t)
\end{aligned} \tag{4.118}$$

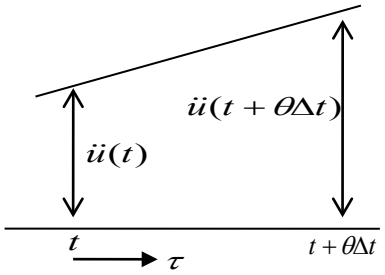


Figure 4.6: The Wilson θ Scheme

The FE equations are now written at time $t+\theta\Delta t$:

$$\mathbf{M}\ddot{\mathbf{u}}(t+\theta\Delta t) + \mathbf{K}\mathbf{u}(t+\theta\Delta t) = \mathbf{F}(t+\theta\Delta t) \tag{4.119}$$

As with the linearly varying $\ddot{u}$, the $\mathbf{F}$ vector is approximated by $\overline{\mathbf{F}}$, a linear extrapolation:

$$\overline{\mathbf{F}}(t+\theta\Delta t) = \mathbf{F}(t) + \theta\left(\mathbf{F}(t+\Delta t) - \mathbf{F}(t)\right) \tag{4.120}$$

These expressions lead to

Wilson θ Scheme:

$$\left[\mathbf{M}\frac{6}{(\theta\Delta t)^2}+\mathbf{K}\right]\mathbf{u}(t+\theta\Delta t)=\overline{\mathbf{F}}(t+\theta\Delta t)+\mathbf{M}\left[\frac{6}{(\theta\Delta t)^2}\mathbf{u}(t)+\frac{6}{\theta\Delta t}\dot{\mathbf{u}}(t)+2\ddot{\mathbf{u}}(t)\right] \quad (4.121)$$

To obtain the solution at time $t+\Delta t$, the solution for $\mathbf{u}(t+\theta\Delta t)$ is substituted into (4.116b). This is then used in (4.117) and its two integrated equations, and τ is set to Δt. This leads to

$$\begin{aligned}
\ddot{\mathbf{u}}(t+\Delta t)&=\frac{6}{\theta(\theta\Delta t)^2}\left(\mathbf{u}(t+\theta\Delta t)-\mathbf{u}(t)\right)-\frac{6}{\theta(\theta\Delta t)}\dot{\mathbf{u}}(t)+\left(1-\frac{3}{\theta}\right)\ddot{\mathbf{u}}(t)\\
\dot{\mathbf{u}}(t+\Delta t)&=\dot{\mathbf{u}}(t)+\frac{\Delta t}{2}\left(\ddot{\mathbf{u}}(t+\Delta t)+\ddot{\mathbf{u}}(t)\right)\\
\mathbf{u}(t+\Delta t)&=\mathbf{u}(t)+\Delta t\dot{\mathbf{u}}(t)+\frac{(\Delta t)^2}{6}\left(\ddot{\mathbf{u}}(t+\Delta t)+2\ddot{\mathbf{u}}(t)\right)
\end{aligned} \quad (4.122)$$

This scheme is unconditionally stable for $\theta\geq 1.37$.

4. Newmark Scheme (Implicit)

In the Newmark integration scheme, first expand as a Taylor series

$$u(t+\Delta t)=u(t)+\Delta t\dot{u}(t)+\frac{1}{2}(\Delta t)^2\ddot{u}(t)+\frac{1}{6}(\Delta t)^3\dddot{u}(\xi),\qquad t\leq\xi\leq t+\Delta t \quad (4.123)$$

Using a linear approximation for the $\ddot{u}$ term,

$$\begin{aligned}
\ddot{u}(t+\Delta t)&=\ddot{u}(t)+\Delta t\dddot{u}(t)\quad+O(\Delta t)^2\\
&=\ddot{u}(t)+\Delta t\dddot{u}(\xi)
\end{aligned} \quad (4.124)$$

Thus the error term in the Taylor series can be written in terms of some parameter α:

$$\begin{aligned}
u(t+\Delta t)&=u(t)+\Delta t\dot{u}(t)+\frac{1}{2}(\Delta t)^2\ddot{u}(t)+\alpha(\Delta t)^2\left(\ddot{u}(t+\Delta t)-\ddot{u}(t)\right)\\
&=u(t)+\Delta t\dot{u}(t)+(\Delta t)^2\left(\alpha\ddot{u}(t+\Delta t)+(\tfrac{1}{2}-\alpha)\ddot{u}(t)\right)
\end{aligned} \quad (4.125)$$

When $\alpha = 1/6$ the linear acceleration scheme expression (4.112b) is recovered. Similarly, the following assumption is made regarding the $\dot{u}$ term:

$$\dot{u}(t+\Delta t) = \dot{u}(t) + \Delta t\left(\delta\ddot{u}(t+\Delta t) + (1-\delta)\ddot{u}(t)\right) \tag{4.126}$$

When $\delta = 1/2$, the expression (4.114a) from the linear acceleration scheme is recovered.

Solving (4.125) for $\ddot{\mathbf{u}}(t+\Delta t)$ and substituting into (4.126) gives

$$\begin{aligned}
\ddot{\mathbf{u}}(t+\Delta t) &= \frac{1}{\alpha}\left\{\frac{1}{(\Delta t)^2}\left(\mathbf{u}(t+\Delta t) - \mathbf{u}(t)\right) - \frac{1}{\Delta t}\dot{\mathbf{u}}(t) - (\tfrac{1}{2}-\alpha)\ddot{\mathbf{u}}(t)\right\} \\
\dot{\mathbf{u}}(t+\Delta t) &= \dot{\mathbf{u}}(t) + \frac{\delta}{\alpha}\left\{\frac{1}{(\Delta t)}\left(\mathbf{u}(t+\Delta t) - \mathbf{u}(t)\right) - \dot{\mathbf{u}}(t) - \Delta t(\tfrac{1}{2}-\alpha)\ddot{\mathbf{u}}(t)\right\} \\
&\qquad + \Delta t(1-\delta)\ddot{\mathbf{u}}(t)
\end{aligned} \tag{4.127}$$

Substituting (4.127a) into the FE equations (4.92) written at time $t+\Delta t$ then leads to

Newmark Scheme:

$$\left(\frac{1}{\alpha(\Delta t)^2}\mathbf{M} + \mathbf{K}\right)\mathbf{u}(t+\Delta t) = \mathbf{F}(t+\Delta t) + \mathbf{M}\left[\frac{1}{\alpha(\Delta t)^2}\mathbf{u}(t) + \frac{1}{\alpha\Delta t}\dot{\mathbf{u}}(t) + \frac{1}{\alpha}(\tfrac{1}{2}-\alpha)\ddot{\mathbf{u}}(t)\right] \tag{4.128}$$

The scheme is unconditionally stable for $\delta \geq 1/2, \alpha \geq (\delta + \tfrac{1}{2})^2/4$.

5. The Trapezoidal Scheme (Implicit)

This is the Newmark scheme with $\delta = 1/2, \alpha = 1/4$, which are the parameters which generally give the best accuracy. This is also called the constant-average-acceleration method because the expression for *u* becomes

$$u(t+\Delta t) = u(t) + \Delta t\dot{u}(t) + \frac{(\Delta t)^2}{2}\left(\frac{1}{2}\left(\ddot{u}(t+\Delta t) + \ddot{u}(t)\right)\right) \tag{4.129}$$

This is a Taylor series with, instead of the usual $\ddot{u}(t)$, the average over the interval, $\left(\ddot{u}(t+\Delta t)+\ddot{u}(t)\right)/2$.

4.2.4 Mode Superposition

As with the first-order system, the system of coupled ODEs $\mathbf{M}\ddot{\mathbf{u}}+\mathbf{K}\mathbf{u}=\mathbf{F}$ can be rewritten in terms of generalised coordinates $\mathbf{z}$ and $\dot{\mathbf{z}}$, so that the equations become uncoupled, and each can be solved independently of the others.

The analysis is essentially the same as for the first order system. Assuming that the eigenvalues and eigenvectors have been calculated, the eigenvectors are normalised through the equation

$$\overline{\mathbf{u}}^{(i)\mathrm{T}}\mathbf{M}\overline{\mathbf{u}}^{(j)}=\begin{cases}1, i=j\\ 0, i\neq j\end{cases} \tag{4.130}$$

for $i,j=1\cdots n$. Next, define the matrix $\mathbf{\Phi}$ whose columns are the eigenvectors $\overline{\mathbf{u}}^{(i)}$ and the diagonal matrix $\mathbf{\Omega}^2$ whose elements are the *n* eigenvalues:

$$\mathbf{\Phi}=\begin{bmatrix}\overline{u}_1^{(1)} & \overline{u}_1^{(2)} & \cdots & \overline{u}_1^{(n)}\\ \overline{u}_2^{(1)} & \overline{u}_2^{(2)} & \cdots & \overline{u}_2^{(n)}\\ \vdots & \vdots & \ddots & \vdots\\ \overline{u}_n^{(1)} & \overline{u}_n^{(2)} & \cdots & \overline{u}_n^{(n)}\end{bmatrix},\qquad \mathbf{\Omega}^2=\begin{bmatrix}\omega^{(1)2} & 0 & \cdots & 0\\ 0 & \omega^{(2)2} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \omega^{(n)2}\end{bmatrix} \tag{4.131}$$

The *n* solutions to the eigenvalue problem $\left[\mathbf{K}-\omega^{(i)2}\mathbf{M}\right]\overline{\mathbf{u}}^{(i)}=0$ can then be written in the form $\mathbf{K}\mathbf{\Phi}=\mathbf{M}\mathbf{\Phi}\mathbf{\Omega}^2$. With the eigenvectors **M**-orthonormalised as above, one has $\mathbf{\Phi}^{\mathrm{T}}\mathbf{M}\mathbf{\Phi}=\mathbf{I}$ and so, pre-multiplying the above equation by $\mathbf{\Phi}^{\mathrm{T}}$, $\mathbf{\Phi}^{\mathrm{T}}\mathbf{K}\mathbf{\Phi}=\mathbf{\Omega}^2$. Introduce next new generalised coordinates $\mathbf{z}$ and transform the original equations through $\mathbf{u}(t)=\mathbf{\Phi}\mathbf{z}(t)$. Thus, multiplying the equations by $\mathbf{\Phi}^{\mathrm{T}}$ gives

$$\dot{\mathbf{z}}+\mathbf{\Omega}^2\mathbf{z}=\mathbf{\Phi}^{\mathrm{T}}\mathbf{F} \tag{4.132}$$

These equations are now uncoupled and each equation can be integrated in turn to evaluate the coordinates $z^{(i)}$, whence the nodal values of $u(t)$ can be evaluated, through $\mathbf{u}(t)=\mathbf{\Phi}\mathbf{z}(t)$. For this purpose one needs the initial conditions on $\mathbf{z}(t)$. Since $\mathbf{\Phi}^{\mathrm{T}}\mathbf{M}\mathbf{\Phi}=\mathbf{I}$, then $\mathbf{u}(t)=\mathbf{\Phi}\mathbf{z}(t)$ becomes $\mathbf{\Phi}^{\mathrm{T}}\mathbf{M}\mathbf{u}(t)=\mathbf{z}(t)$ so that

$$\begin{aligned}\mathbf{z}(0) &= \mathbf{\Phi}^{\mathrm{T}}\mathbf{M}\mathbf{u}(0)\\ \dot{\mathbf{z}}(0) &= \mathbf{\Phi}^{\mathrm{T}}\mathbf{M}\dot{\mathbf{u}}(0)\end{aligned} \tag{4.133}$$

4.2.5 Stability

Here, the stability of the explicit central difference scheme is examined. As with the explicit Euler analysis, it is only necessary to consider the homogeneous modal equation

$$\ddot{z} + \omega^2 z = 0 \tag{4.134}$$

This equation has already been examined in detail in section 4.2.3 above (see Eqn. 4.93 and Eqns. 4.99-104). That analysis now leads directly to the stability criterion $\Delta t \le 2/\omega$. The critical time step for stability will depend on the largest natural frequency in the system, giving the criterion (4.98).

4.3 Application: Elastodynamics

Here the problem of an elastic material subject to arbitrary loading and initial conditions is examined. This problem is examined in detail in Solid Mechanics, Part II, section 2.2, but the main points are discussed again here.

The geometry of the problem is as shown in Fig. 4.7, a (one dimensional) rod of length *l*, cross section A, subjected to a given displacement or stress/force at its ends. The Young's modulus of the rod is *E*.

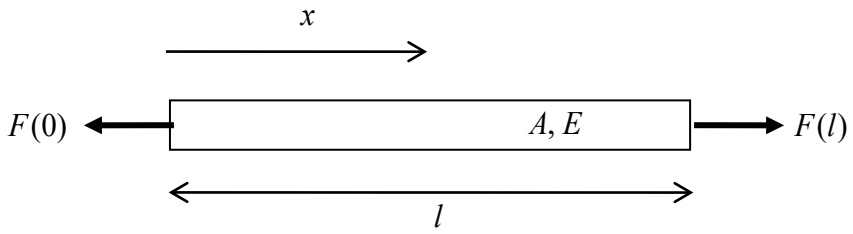


Figure 4.7: The Elastic Rod

4.3.1 Governing Differential Equation

The equations governing the response of the rod are:

Governing Equations for Elastodynamics:

Equation of Motion:

$$\frac{\partial \sigma}{\partial x} = \rho \frac{\partial^2 u}{\partial t^2} \tag{4.135}$$

Strain-Displacement Relation:

$$\varepsilon = \frac{du}{dx} \tag{4.136}$$

Constitutive Relation:

$$\sigma = E\varepsilon \tag{4.137}$$

The first two of these are derived in the Appendix to Chapter 2, §2.12.2 (where the body force is neglected[7]). The third is Hooke's law for elastic materials.

In these equations, σ is the stress, ε is the small strain (change in length per original length), u is the displacement and E is the Young's modulus of the material.

[7] the body force is usually much smaller than the other terms in dynamic problems

The strain-displacement relation and the constitutive equation can be substituted into the equation of equilibrium to obtain

1D Governing Equation for Dynamic Elasticity:

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}, \qquad c = \sqrt{\frac{E}{\rho}} \tag{4.138}$$

This is the **one-dimensional wave equation,** and is the second order equation considered in Eqn. 4.76. The solution predicts that a wave emanates from a struck end of the rod with speed c. As the wave passes a certain point in the material, the material particles undergo a small displacement u and suffer a consequent stress [8].

Material	$\rho\ (\mathrm{kg/m^3})$	$E\ (\mathrm{GPa})$	$c\ (\mathrm{m/s})$
Aluminium Alloy	2700	70	5092
Brass	8300	95	3383
Copper	8500	114	3662
Lead	11300	17.5	1244
Steel	7800	210	5189
Glass	1870	55	5300
Granite	2700		3120
Limestone	2600		4920
Perspex			2260

Table 4.1: Elastic Wave Speeds for Several Materials

The stressed material undergoes longitudinal vibrations, with the particles oscillating about some equilibrium position. One should be clear about the distinction between the velocity of the oscillating particles, say $v = du/dt$, and the speed of the travelling stress wave, c.

For example, suppose that the bar is given a sudden displacement u_0 at time $t = 0$, Fig. 4.8.

[8] it was assumed that the density ρ in this analysis is constant. In fact, as the wave passes, the material gets compressed and the density of a constant mass of material increases. It can be shown that these fluctuations in density are, however, second-order effects and can be neglected

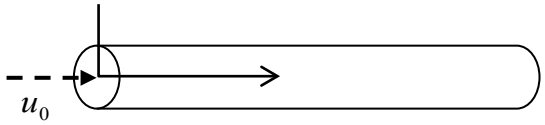


Figure 4.8: A sudden displacement prescribed at one end of an elastic rod

In that case the wave will travel from left to right, Fig. 4.9. As it passes a point, the material there will experience a sudden stress – the stress is discontinuous at the wave front. Eventually the wave will reach the other end of the bar and get reflected – there is then reinforcement and cancellation of waves as they meet each other in opposite directions.

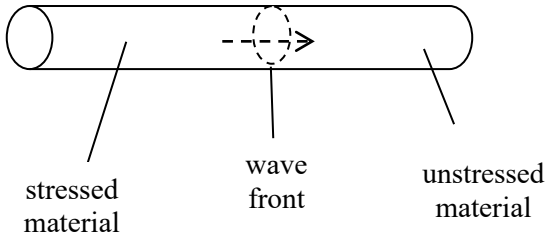


Figure 4.9: wave propagation along an elastic rod

Boundary and Initial Conditions

The case of a static rod was examined in §2.10. As in the static case, one must

$$\begin{aligned} &\text{specify } u(0,t) \quad \text{or} \quad \frac{\partial u}{\partial x}(0,t) && \text{B.C. at } x=0 \\ &\text{specify } u(L,t) \quad \text{or} \quad \frac{\partial u}{\partial x}(L,t) && \text{B.C. at } x=L \end{aligned} \tag{4.139}$$

Also, one must

$$\begin{aligned} &\text{Specify } u(x,0) && \text{I.C. for displacement} \\ &\text{Specify } \frac{\partial u}{\partial t}(x,0) && \text{I.C. for velocity} \end{aligned} \tag{4.140}$$

References to some exact solutions to the wave equation are given in the Appendix to this Chapter, as is a review of the dynamics of a single degree of freedom.

4.3.2 The FEM Solution

The only difference between this case and the static case is the inclusion of the acceleration term

$$\frac{\partial^2 u}{\partial t^2} \rightarrow \int_0^L \frac{\partial^2 u}{\partial t^2}\omega(x)dx \rightarrow \left[\int_0^L \omega_i \omega_j(x)dx\right]\ddot{u}_i \tag{4.141}$$

which leads to the mass matrix (the term inside the square brackets), so-called since the complete term, M times the nodal acceleration $\ddot{u}_i$ gives a force.

Eigenvalues (Natural Frequencies) and Eigenvectors (Mode Shapes)

The natural frequencies for the two-linear-element FE model of §4.2.2, are (as in Eqn. 4.87),

$$\omega^{(1)} = 1.61142\frac{c}{l}, \quad \omega^{(2)} = 5.62930\frac{c}{l} \tag{4.142}$$

The number of natural frequencies in a system will equal the number of degrees of freedom in the system. This compares with the real physical system, which has an infinite number of degrees of freedom and natural frequencies associated with the infinite number of material particles in the rod. To obtain a solution for the higher frequencies $\omega^{(3)}, \omega^{(4)}, \cdots$, it is necessary to include more degrees of freedom, i.e. elements, into the FE mesh. The FE solution for the natural frequencies, solved for 1, 2, 3 and 4 elements, is as tabulated below.

The exact solution for the frequencies is also tabulated; these are given by (see the Appendix for a reference to this)

$$\omega^{(n)} = \frac{(2n-1)\pi}{2}\frac{c}{l} \quad (n = 1,2,\ldots) \qquad = \frac{\pi}{2}\frac{c}{l}, \frac{3\pi}{2}\frac{c}{l}, \frac{5\pi}{2}\frac{c}{l}, \cdots \tag{4.143}$$

No. of elements	1	2	3	4	Exact
ω_1	1.7321 c/l	1.6114 c/l	1.5888 c/l	1.5809 c/l	1.5708 c/l
ω_2		5.6293 c/l	5.1962 c/l	4.9872 c/l	4.7124 c/l
ω_3			9.4266 c/l	9.0594 c/l	7.8539 c/l
ω_4				13.1007 c/l	10.9956 c/l
ω_5					14.1372 c/l
ω_6					17.279 c/l

Table 4.2: Natural Frequencies for 2-noded Linear Elements ($\mathbf{u} = \mathbf{0}$ at one end)

Note the following:

- the FE results for the *lower frequencies are more accurate* than those for the higher frequencies. This will be explained below.
- the FE model yields natural frequencies which are *higher than the true values*. This is because the FE model is a constrained version of the real system – it is not allowed the same degree of freedom as the real system – the FE model of the material is stiffer (see the case of a single degree of freedom, Eqn. 4.80, and the Appendix to this Chapter, §4.5.1, where $\omega = \sqrt{k/m}$, k being the stiffness).

Mode Shapes

Continuing again the above two element example, the eigenvectors or *modes* $\overline{\mathbf{u}}$ are now obtained from

$$\left(\mathbf{K} - \omega^2 \mathbf{M}\right)\overline{\mathbf{u}} = \mathbf{0}, \qquad \frac{2c^2}{l}\begin{bmatrix} 2-4\alpha & -1-\alpha \\ -1-\alpha & 1-2\alpha \end{bmatrix}\begin{bmatrix} \overline{u}_2 \\ \overline{u}_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{4.144}$$

one for each frequency:

$$\omega^{(1)}: \quad \overline{\mathbf{u}}^{(1)} = \begin{bmatrix} \overline{\mathbf{u}}_2 \\ \overline{\mathbf{u}}_3 \end{bmatrix} = \begin{bmatrix} 1 \\ \sqrt{2} \end{bmatrix}, \qquad \omega^{(2)}: \quad \overline{\mathbf{u}}^{(2)} = \begin{bmatrix} \overline{\mathbf{u}}_2 \\ \overline{\mathbf{u}}_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -\sqrt{2} \end{bmatrix} \tag{4.145}$$

The modes give the character of the system of equations, whose general solution is $\mathbf{u} = C_1\overline{\mathbf{u}}^{(1)}\sin\left(\omega^{(1)}t + \phi^{(1)}\right) + C_2\overline{\mathbf{u}}^{(2)}\sin\left(\omega^{(2)}t + \phi^{(2)}\right)$. The mode shapes for this particular problem are plotted below in Fig. 4.10 (solid lines).

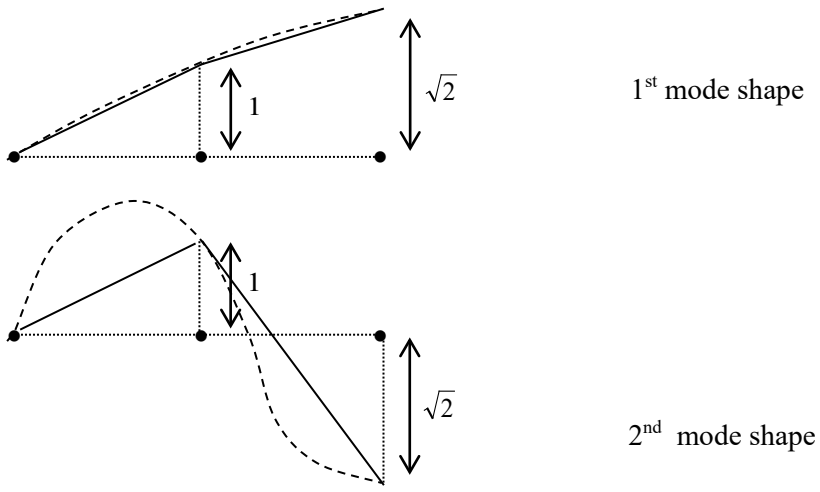


Figure 4.10: the first two mode shapes

These mode shapes can be compared to the exact shapes (see reference to these in the Appendix §4.5.2, and which are plotted in dotted lines:

$$\sin\left(\omega^{(n)}\frac{x}{c}\right)=\sin\left(\frac{(2n-1)\pi}{2}\frac{x}{l}\right)(n=1,2,\ldots)\quad=\sin\left(\frac{\pi}{2}\frac{x}{l}\right),\sin\left(\frac{3\pi}{2}\frac{x}{l}\right),\sin\left(\frac{5\pi}{2}\frac{x}{l}\right),\cdots \tag{4.146}$$

The simple linear two-element solution is not so bad an approximation for the first mode but, as with the natural frequencies, the higher, second, mode is not as well represented[9]. Note that the ratios of the amplitudes at the nodal points 2 and 3 are as the ratios of the exact solution.

The reason why the higher frequencies and corresponding mode shapes cannot be obtained with great accuracy is now clear. The higher modes contain many "waves" and one would need many elements to capture the features of this wave. These higher modes contain much more curvature than the lower modes and are difficult to model. For example, one would probably need five elements of equal length to capture the third mode with any real accuracy.

[9] the exact mode shapes here have been multiplied by $\sqrt{2}$ to fit the FE solution; the amplitudes of these shapes are unimportant as they depend on the initial conditions

Vibration Analysis

The above is a **vibration analysis**, where the natural frequencies and modes of the system are evaluated without regard to which of them might be important in an application and without regard to how the vibration is initiated. The exact combination of the modes for a particular problem is determined from the initial conditions (see below). The vibration is **free** if the "load vector" **F** is zero or constant (as in our case); **forced vibration** occurs when the load vector itself oscillates (is sinusoidal).

Complete Solution

Although the primary interest in this section was the determination and discussion of the frequencies and mode shapes, it is instructive to continue and solve the problem completely. The FE equations can only be solved exactly here because of the simplicity of this two-element problem.

To apply the initial conditions, it is best to rewrite the solution in the form

$$\mathbf{u} = \overline{\mathbf{u}}^{(1)}\left[A\cos\left(\omega^{(1)}t\right)+B\sin\left(\omega^{(1)}t\right)\right]+\overline{\mathbf{u}}^{(2)}\left[C\cos\left(\omega^{(2)}t\right)+D\sin\left(\omega^{(2)}t\right)\right] \tag{4.147}$$

Then, from $\mathbf{u}(0)=0$ and $\dot{\mathbf{u}}(0)=2x/l$, so that $\dot{u}_2(0)=1, \dot{u}_3(0)=2$, $A=C=0$,

$$\mathbf{u}(t)=\begin{bmatrix}u_2(t)\\u_3(t)\end{bmatrix}=\frac{1+\sqrt{2}}{2\omega^{(1)}}\begin{bmatrix}1\\\sqrt{2}\end{bmatrix}\sin\left(\omega^{(1)}t\right)+\frac{1-\sqrt{2}}{2\omega^{(2)}}\begin{bmatrix}1\\-\sqrt{2}\end{bmatrix}\sin\left(\omega^{(2)}t\right) \tag{4.148}$$

Using the shape functions, the complete solution is

$$u(x,t)=\frac{x}{l}\left\{\frac{1+\sqrt{2}}{\omega^{(1)}}\sin\left(\omega^{(1)}t\right)+\frac{1-\sqrt{2}}{\omega^{(2)}}\sin\left(\omega^{(2)}t\right)\right\} \qquad 1^{\text{st}} \text{ element}$$

$$u(x,t)=\frac{x/l+\sqrt{2}-1/\sqrt{2}}{\omega^{(1)}}\sin\left(\omega^{(1)}t\right)+\frac{x/l-\sqrt{2}+1/\sqrt{2}}{\omega^{(2)}}\sin\left(\omega^{(2)}t\right) \qquad 2^{\text{nd}} \text{ element} \tag{4.149}$$

with x here measured from the left-hand end. This can be compared to the exact solution (see reference to this in the Appendix to this Chapter),

$$u(x,t) = \frac{32l}{\pi^3 c}\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{(2n-1)^3}\sin(\lambda_n x)\sin(\lambda_n ct), \qquad \omega^{(n)} = \lambda_n c = \frac{(2n-1)\pi c}{2l}, \qquad n = 1,2,\ldots \tag{4.150}$$

The table below compares the FE and exact solutions (30 terms) for $x = l/4$ (in the first element) , with $l = 1, c = 5$ and $\Delta t = l/10c$. Considering the wave equation to represent the propagation of a wave through an elastic material at speed *c*, this time step Δt is one tenth the time a wave would take to travel the length of the bar.

	Δt	$2\Delta t$	$3\Delta t$	$4\Delta t$	$5\Delta t$	$6\Delta t$	$7\Delta t$	$8\Delta t$	$9\Delta t$	$10\Delta t$
Exact	0.310	0.620	0.930	1.240	1.550	1.860	2.170	2.465	2.651	2.715
FE	0.312	0.633	0.966	1.307	1.634	1.937	2.179	2.340	2.409	2.386

Table 4.3: Comparison of 2-element linear FE model with Exact Solution (for *u*(*l*/4))

Also shown, in Fig. 4.11, are the deformed shapes of the bar for the same 10 time intervals (using the exact solution, but plotted linearly through five points) – the 10th case is the maximum deformation, after which the displacement begins to decrease again, and then down to negative values.

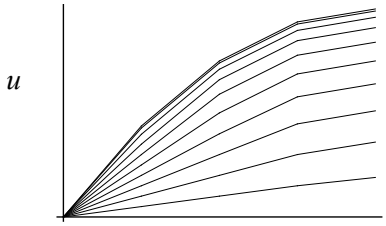


Figure 4.11: deformed shapes of the elastic bar (Note: this pricture gives the impression that the bar is swaying up and down (like a beam); actually, these displacements are along the direction of the rod – it is all one-dimensional)

Note the following:

- when a material is loaded or displaced, only a certain range of its natural frequencies are excited. For this reason it is not actually necessary to evaluate many of them in order to determine the material's response. When the loading itself is harmonic with frequency ω_f, then a general rule of thumb is that all the natural frequencies up to about $4\omega_f$ should be evaluated. By the same token, if the frequency of the loading function is very low, say one quarter of the lowest natural frequency or lower, then a *static* solution should yield an accurate result.

- in practice, in a model with hundreds or thousands of nodes, use of the standard method of solution for the eigenvalues and eigenvectors, expanding the determinant and solving the resulting polynomial, is not practical. Special techniques have been developed for this purpose (see advanced texts on FE and computational techniques).

Damping

The above solution for $u(x,t)$ continues to oscillate about $u = 0$ and does not decay with time. This is a characteristic of ideally elastic materials, for which there is no energy loss. In any real material, there will be *damping*, which dissipates energy and causes the amplitude of free vibration to decay with time[10]. A simple, somewhat artificial, way of introducing damping into the current model is to define a viscous damping matrix

$$\mathbf{C} = \alpha\mathbf{M} + \beta\mathbf{K} \tag{4.151}$$

Here, α, β are constants to be determined experimentally (the former damps the lower modes whereas the latter has more of an effect on the higher modes). The FE equations are now

$$\mathbf{M\ddot{u}} + \mathbf{C\dot{u}} + \mathbf{Ku} = \mathbf{F} \tag{4.152}$$

Note the following

- some FE software is capable of calculating damped natural frequencies. These frequencies are often only slightly smaller than the undamped natural frequencies (see Appendix 1 for the case of a single degree of freedom).
- in real transient problems, FE models will often incorporate some damping to eliminate resonance problems and oscillatory noise

Direct Integration

In most of the above, the free vibration model was analysed. The transient (or **dynamic**) response can be evaluated using one of the direct integration methods discussed earlier

In the implicit schemes, it is usual to take as the time-step the "element length" divided by the wave speed c, $\Delta t = L/c$, since this is the time taken for the wave to pass through the

[10] damping is a feature of many material models, for example of viscoelasticity and plasticity

element, but no further. Non-uniform and low- or high- order elements can be used, and when high-order elements are employed, a consistent mass matrix is usually appropriate.

Explicit schemes, with their smaller time-steps, are more appropriate for systems which are changing rapidly, for example for systems describing the sudden impact of materials. Implicit schemes are more appropriate for more slowly evolving systems, for example for systems describing the moderately paced flow of fluid through a porous medium.

Example

In the following graphs, Figs. 4.12-14, are plotted the solution to the wave equation with $u(0,t) = 0$, $u'(0,t) = F$. The explicit Euler scheme is used. The solution is compared with the "exact" modal equations solution (that is, obtaining a solution by first solving for the eigenvalues and eigenvectors, as done in the above).

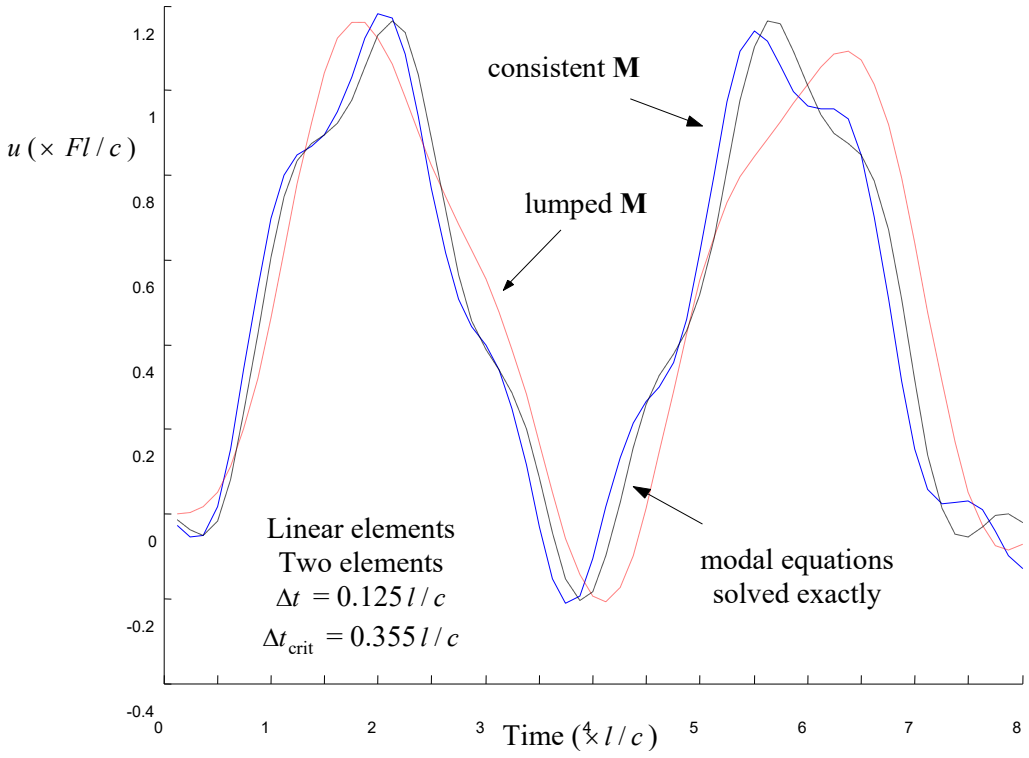


Figure 4.12: displacement at a material particle

unstable solution using the critical time step for two linear elements $\Delta t = 0.355\, l / c$

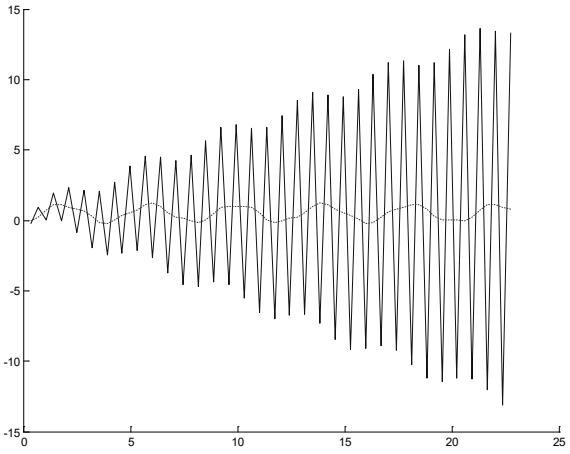


Figure 4.13: unstable displacement solution

solution using 20 linear elements

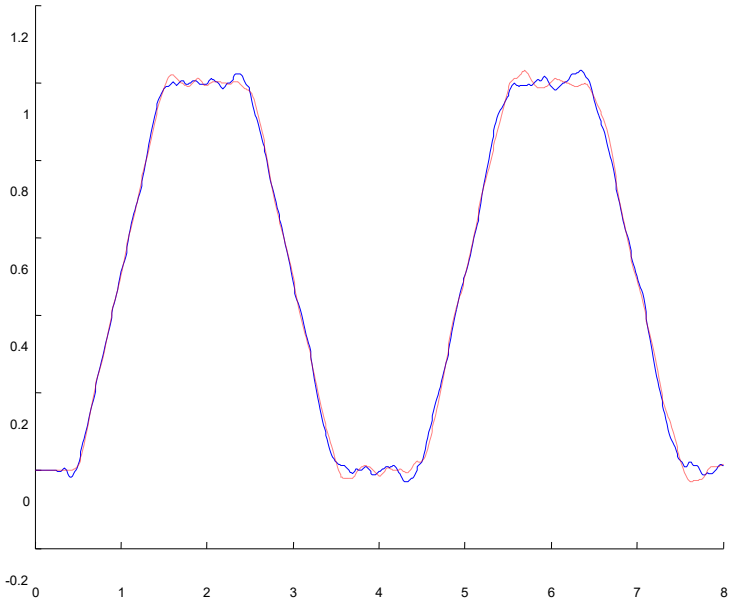


Figure 4.14: displacement at a material particle

Mode Superposition

As with the first-order system, the system of coupled ODEs (4.80) can be rewritten in terms of generalised coordinates $\mathbf{z}$ and $\dot{\mathbf{z}}$, so that the equations become uncoupled, and each can be solved independently of the others.

The mode superposition method is well-suited to problems which are dominated by the lower modes, and where the response of the higher modes is unimportant and can be neglected. This occurs for example with earthquake loading, where only the lowest 10 modes or so need to be considered, even though the order of the system may be quite large. On the other hand, for blast or shock loading, many more modes generally need to be included, perhaps about two-thirds of them. If this is the case then direct integration may be a more suitable solution procedure.

4.4 Problems

1. Derive the system of equations (4.11) for the 4-element model of the first order equation (4.5).
2. Consider the equation

$$q\frac{\partial^2 p}{\partial x^2} = \frac{\partial p}{\partial t}, \qquad 0 \le x \le l$$

   a) derive the **C** matrix (linear element)
   b) consider the boundary conditions $p(0) = 0$, $p(l) = A$. How many eigenvalues/modes would there be in a two-element FE model of this? Evaluate them.
   c) is it true that $\lambda_{\max} \propto 1/L^2$?
3. Use equations (4.42-44) to derive the Implicit-Euler algorithm (4.45).
4. Considering the Semi-Implict Algorithm for first order systems, show that the Crank-Nicholson scheme, $\alpha = 1/2$, leads to a truncation error proportional to $(\Delta t)^2$ [hint: expand $u(t_0 \pm \frac{1}{2}\Delta t)$ in Taylor series and subtract.]
5. Derive the relations (4.60) for the *C*-normalised eigenvectors (4.58).
6. What is matrix lumping? When and why is it done?
7. What is mode superposition and when might it be used to advantage?
8. Expand $\ddot{u}(t+\tau)$, $\ddot{u}(t+\Delta t)$ in Taylor series and hence derive the linear acceleration scheme formula (4.111) and deduce the truncation error involved.
9. Derive the Wilson θ equations (4.118).
10. When would you use an explicit scheme and when an implicit scheme? Why?
11. In a FE solution for the natural frequencies of the elastodynamic problem, which frequencies are more accurate? Why? What about the corresponding eigenvectors?

4.5 Appendix to Chapter 4

4.5.1 Review of the Dynamics of a Single Degree of Freedom

Free Vibration: No Damping

Consider a mass *m* attached to a freely oscillating spring, at initial position x_0 and with initial velocity $\dot{x}_0$. From Newton's Law

$$m\ddot{x} = -kx \tag{4A.1}$$

where *k* is the spring constant. This 2nd order ODE can solved to obtain

$$x(t) = x_0 \cos \omega t + (\dot{x}_0 / \omega) \sin \omega t \tag{4A.2}$$

where the frequency is $\omega = \sqrt{k/m}$ and the period of vibration is $T = 2\pi / \omega$. Different initial conditions simply shift the oscillations along the *t* axis, which can be seen by rewriting the displacement as

$$x(t) = A \sin(\omega t + \phi) \tag{4A.3}$$

where $\tan \phi = \omega x_0 / \dot{x}_0$, $A = \sqrt{x_0^2 + (\dot{x}_0^2 / \omega)^2}$. Shown in Fig. 4A.1 is a plot with $x_0 = 1, \dot{x}_0 = 7$ and $\omega = 3$ (so that there is one complete cycle every $T = 2\pi / 3 \approx 2.1\text{s}$.

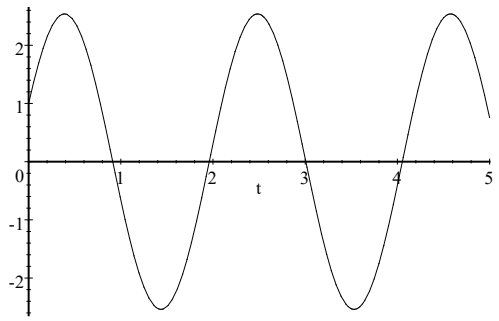


Figure 4A.1: Free Vibration

Consider now a constant force *P* applied to the oscillating mass, so that

$$m\ddot{x} = -kx + P \tag{4A.4}$$

This can be solved to obtain

$$x(t) = A\sin(\omega t + \phi) + \frac{P}{k} \tag{4A.5}$$

where $\tan\phi = \omega(x_0 - P/k)/\dot{x}_0$, $A = \sqrt{(x_0 - P/k)^2 + (\dot{x}_0^2/\omega)^2}$. It can be seen that the frequency is the same as in the unforced case and the mass oscillates about a mean position $x = P/k$, which is the static solution, that is, the position the mass would occupy if the force was applied very slowly and gradually from zero up to *P*.

Forced Vibration: No Damping

When an oscillatory force is applied, say $P = P_0\sin(\Omega t + \Phi)$, the solution to the non-homogeneous ODE is

$$x(t) = A\sin(\omega t + \phi) + \frac{P_0/k}{1 - (\Omega/\omega)^2}\sin(\Omega t + \Phi) \tag{4A.6}$$

and A, ϕ depend on the initial conditions (but are lengthy in this case). This is a superposition of two harmonic oscillations. Note that the amplitude becomes very large as $\Omega \to \omega$, a situation known as **resonance**.

Free Vibration: Damping

If one now also has a viscous damper with force $c\dot{x}$, then

$$m\ddot{x} = -c\dot{x} - kx \tag{4A.7}$$

When the damping is very large, $c^2 > 4mk$, the solution is of the form $x = Ae^{\beta_1 t} + Ae^{\beta_2 t}$ where $\beta_1, \beta_2 < 0$ and so the displacement falls quickly to the equilibrium position. If, on the other hand, the damping is not so high, then

$$x = e^{-\frac{c}{2m}t}\{A\cos(\omega_d t) + B\sin(\omega_d t)\}, \qquad \omega_d = \omega\sqrt{1-\xi^2}, \qquad \xi = \frac{c}{2m\omega} \tag{4A.8}$$

Here, ω is the undamped frequency and ξ is called the damping ratio.

Forced Vibration: Damping

Now we consider the system

$$m\ddot{x} = -c\dot{x} - kx + P_0 \sin(\Omega t + \Phi) \tag{4A.9}$$

The solution to the corresponding homogeneous equation is of the form $x(t) = \exp(-ct/2m)\{A\cos(\omega_d t) + B\sin(\omega_d t)\}$. This part of the solution dies away after a sufficient amount of time and is known as the **transient solution**. What remains is the particular solution,

$$x(t) = A\sin(\Omega t + \Phi - \alpha), \qquad \tan\alpha = \frac{2(\Omega/\omega)\xi}{1-(\Omega/\omega)^2} \tag{4A.10}$$

with

$$A = \frac{P_0/k}{\sqrt{\left(1-(\Omega/\omega)^2\right)^2 + \left(2(\Omega/\omega)\xi\right)^2}} \tag{4A.11}$$

4.5.2 Exact Solution to the 1-D Wave Equation

As mentioned above, the exact solution to the 1-D wave equation is detailed in Solid Mechanics, Part II, section 2.2 (see section 2.2.6).

5 Non-Linear Differential Equations

Application of the Finite Element Method to the solution of linear differential equations leads to a system of linear algebraic equations of the form $\mathbf{Ax} = \mathbf{b}$; with non-linear differential equations one arrives at a system of non-linear equations, which cannot be solved by elementary elimination methods. Thus, much of the focus here is on methods of solving the resulting systems of FE non-linear equations.

5.1 Methods for the Solution of Non-Linear Equations

There are a number of basic techniques for solving non-linear equations. For example, there are the

1. Substitution method
2. Newton-Raphson method
3. Incremental (step by step) method
   - Initial Stress Method
   - Modified Newton-Raphson method

The Substitution Method is not commonly used, but is given in the Appendix to this Chapter. The Newton-Raphson method is the primary solution scheme for the non-linear equations which arise in the FEM and will be discussed in detail.

5.1.1 The Newton-Raphson Method

Consider first the one-dimensional case: the non-linear equation

$$R(u) = 0\,, \tag{5.1}$$

whose exact solution is $u^{(e)}$. Suppose one has an initial estimate of the solution, $u^{(0)}$. Using a Taylor expansion and dropping higher order terms,

$$R(u^{(e)}) = R(u^{(0)}) + \Delta u \left.\frac{\partial R}{\partial u}\right|_{u^{(0)}}, \tag{5.2}$$

where $\Delta u = u^{(e)} - u^{(0)}$. Using $R(u^{(e)}) = 0$ then leads to the approximation

$$u^{(e)} \approx u^{(0)} - \frac{R(u^{(0)})}{\partial R / \partial u\big|_{u^{(0)}}} \tag{5.3}$$

which is an expression for the next approximation.

Consider the following one-dimensional example: solve the non-linear equation

$$-2u^2 + \tfrac{8}{3}u = -\tfrac{2}{3} \tag{5.4}$$

One has

$$R(u) = -2u^2 + \tfrac{8}{3}u + \tfrac{2}{3}, \qquad \frac{\partial R}{\partial u} = -4u + \tfrac{8}{3} \tag{5.5}$$

and the algorithm

$$\Delta u^i = -\frac{R\big|_{u^{i-1}}}{\left(\partial R / \partial u\right)_{u^{i-1}}}, \qquad u^i = u^{i-1} + \Delta u^i \tag{5.6}$$

The function $R(u)$ is called a **residual** function, in the sense that the objective is to drive this function to zero.

Using an initial estimate of $u = 1$ gives the sequence

		error
$u^{(0)}$	1.00000000	0.54858377
$u^{(1)}$	2.00000000	0.45141623
$u^{(2)}$	1.62500000	0.07641623
$u^{(3)}$	1.55163043	0.00304666
$u^{(4)}$	1.54858901	0.00000524
$u^{(5)}$	1.54858377	0.00000000
exact	**1.54858377**	

Table 5.1: Newton-Raphson solution

If one had started with an initial estimate of -1, one would have converged to the other root, -0.21525044.

A geometrical interpretation of the Newton-Raphson method is shown below for this example. In deriving the Newton-Raphson method, the process of *linearization* has been used. The linearised model is the tangent to the nonlinear residual function (and hence the name "tangent matrix").

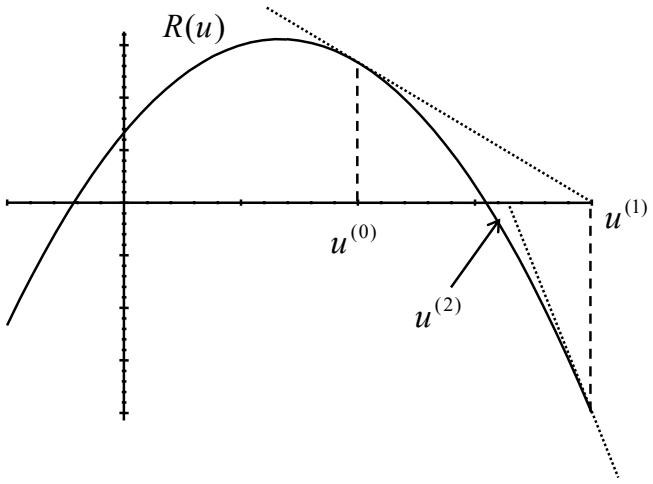


Figure 5.1: Geometric interpretation of the Newton-Raphson Solution Scheme

Notes on Convergence and Accuracy

(1) if the derivative $\partial R / \partial u$ is continuous in a neighbourhood of the solution, and if $u^{(i-1)}$ lies in that neighbourhood, then the next iterated solution $u^{(i)}$ *will be closer to the solution* than $u^{(i-1)}$ and the scheme will converge towards the solution.

(2) if $\partial R / \partial u\big|_{u^{(i-1)}} = 0$, the method will fail – geometrically, this occurs when the tangent to *R* is horizontal

(3) The convergence is quadratic, that is, if the error after iteration $i-1$ is ε, the error after iteration i is ε^2. This can be seen as follows: write

$$u^{(i)} = u^{(i-1)} - \frac{R(u^{(i-1)})}{R'(u^{(i-1)})} \equiv g(u^{(i-1)})$$

Let the true solution be $u^{(e)}$, so that $u^{(i-1)} = u^{(e)} - \varepsilon_{i-1}$, where ε_{i-1} is the error of $u^{(i-1)}$. By a Taylor series,

$$u^{(i)} = g(u^{(i-1)}) = g(u^{(e)} - \varepsilon_{i-1}) = g(u^{(e)}) - \varepsilon_{i-1} g'(u^{(e)}) + \tfrac{1}{2}\varepsilon_{i-1}^2 g''(u^{(e)}) - \cdots$$

Now

$$g(u^{(e)}) = u^{(e)}, \qquad g'(u^{(e)}) = \frac{R(u^{(e)})R''(u^{(e)})}{\left[R'(u^{(e)})\right]^2} = 0, \qquad g''(u^{(e)}) = \frac{R''(u^{(e)})}{R'(u^{(e)})} \neq 0,$$

since $R(u^{(e)}) = 0$. Thus the error in the next estimate is

$$\varepsilon_i = u^{(e)} - u^{(i)} \approx -\tfrac{1}{2}\varepsilon_{i-1}^2 g''(u^{(e)})$$

and so the method is of second-order (quadratic) – provided $R'(u^{(e)}) \neq 0$

Multi-dimensional Equations

Consider the following system of non-linear equations,

$$\mathbf{R}(\mathbf{u}) = \begin{bmatrix} R_1(\mathbf{u}) \\ R_2(\mathbf{u}) \\ \vdots \\ R_n(\mathbf{u}) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}, \tag{5.7}$$

whose exact solution is $\mathbf{u}^{(e)}$. Suppose one has an initial estimate of the solution, $\mathbf{u}^{(0)}$. Using a Taylor expansion and dropping higher order terms,

$$\mathbf{R}(\mathbf{u}^{(e)}) = \mathbf{R}(\mathbf{u}^{(0)}) + \Delta\mathbf{u} \left.\frac{\partial \mathbf{R}}{\partial \mathbf{u}}\right|_{\mathbf{u}^{(0)}}, \tag{5.8}$$

where $\Delta\mathbf{u} = \mathbf{u}^{(e)} - \mathbf{u}^{(0)}$. Using $\mathbf{R}(\mathbf{u}^{(e)}) = 0$ then leads to

$$\left.\frac{\partial \mathbf{R}}{\partial \mathbf{u}}\right|_{\mathbf{u}^{(0)}} \Delta\mathbf{u} = -\mathbf{R}(\mathbf{u}^{(0)}) \tag{5.9}$$

Solving this system of linear equations leads to a new approximation $\mathbf{u}^{(1)} = \mathbf{u}^{(0)} + \Delta\mathbf{u}$. An algorithm is then

Newton-Raphson Algorithm:

$$\left[\mathbf{K}_T(\mathbf{u}^{i-1})\right]\Delta\mathbf{u}^i = -\mathbf{R}(\mathbf{u}^{i-1}), \qquad \mathbf{K}_T(\mathbf{u}^{i-1}) = \left.\frac{\partial \mathbf{R}}{\partial \mathbf{u}}\right|_{\mathbf{u}^{(i-1)}} \tag{5.10}$$

$$\mathbf{u}^i = \mathbf{u}^{i-1} + \Delta\mathbf{u}^i$$

The matrix $\mathbf{K}_T$ is called the *tangent matrix*.

Tangent Matrix:

$$\mathbf{K}_T = \frac{\partial \mathbf{R}}{\partial \mathbf{u}} = \begin{bmatrix} \frac{\partial R_1}{\partial u_1} & \frac{\partial R_1}{\partial u_2} & \cdots & \frac{\partial R_1}{\partial u_n} \\ \frac{\partial R_2}{\partial u_1} & \frac{\partial R_2}{\partial u_2} & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial R_2}{\partial u_n} & \cdots & \cdots & \frac{\partial R_n}{\partial u_n} \end{bmatrix} \tag{5.11}$$

A Two-Dimensional Problem

Consider the following system of non-linear equations

$$\begin{array}{l} \frac{2}{3} + \frac{8}{3}u_1 + \frac{7}{3}u_2 - 2u_1^2 - \frac{16}{5}u_1u_2 - \frac{7}{5}u_2^2 = 0 \\ \frac{7}{12} + \frac{7}{3}u_1 + \frac{34}{15}u_2 - \frac{8}{5}u_1^2 - \frac{14}{5}u_1u_2 - \frac{47}{35}u_2^2 = 0 \end{array} \tag{5.12}$$

One has

$$\mathbf{R}(\mathbf{u}) = \begin{bmatrix} \mathbf{R}_1(\mathbf{u}) \\ \mathbf{R}_2(\mathbf{u}) \end{bmatrix} = \begin{bmatrix} \frac{2}{3} + \frac{8}{3}u_1 + \frac{7}{3}u_2 - 2u_1^2 - \frac{16}{5}u_1u_2 - \frac{7}{5}u_2^2 \\ \frac{7}{12} + \frac{7}{3}u_1 + \frac{34}{15}u_2 - \frac{8}{5}u_1^2 - \frac{14}{5}u_1u_2 - \frac{47}{35}u_2^2 \end{bmatrix} \tag{5.13}$$

and

$$\mathbf{K}_T = \frac{\partial \mathbf{R}}{\partial \mathbf{u}} = \begin{bmatrix} \frac{\partial R_1}{\partial u_1} & \frac{\partial R_1}{\partial u_2} \\ \frac{\partial R_2}{\partial u_1} & \frac{\partial R_2}{\partial u_2} \end{bmatrix} = \begin{bmatrix} \frac{8}{3} - 4u_1 - \frac{16}{5}u_2 & \frac{7}{3} - \frac{16}{5}u_1 - \frac{14}{5}u_2 \\ \frac{7}{3} - \frac{16}{5}u_1 - \frac{14}{5}u_2 & \frac{34}{15} - \frac{14}{5}u_1 - \frac{94}{35}u_2 \end{bmatrix} \tag{5.14}$$

The Newton-Raphson scheme then converges to the exact solutions $(u_1, u_2) = (-0.19401, -0.02477)$, $(0.00541, 1.90911)$, $(3.09001, -2.659787)$ or $(-3.03474, 4.508781)$.

5.1.2 The Incremental Method (with the Newton-Raphson Method)

Consider the general residual function

$$R(u) = K(u) - F = 0 \tag{5.15}$$

Here, u is the unknown and F will in general be some known. In a mechanics problem, for example, u would be the unknown displacement and F would be due to the known applied loads. In a linear problem, one can specify F and one can solve for u., Fig. 5.2a. In a non-linear problem, however, this is not so straight-forward. As an extreme example, the true relationship between F and u might look like that illustrated in Fig. 5.2b; there will in general be more than one solution for u corresponding to any given F. Referring to Fig. 5.2b, it would be unlikely that one could find $\bar{u}$ given $\bar{F}$, unless one began the algorithm close to $\bar{u}$. For this reason, in a practical FE problem, one does not try to solve a non-linear problem "in one hit". There is a danger that, if the initial prediction $u^{(0)}$ is inaccurate, the solution will not be found.

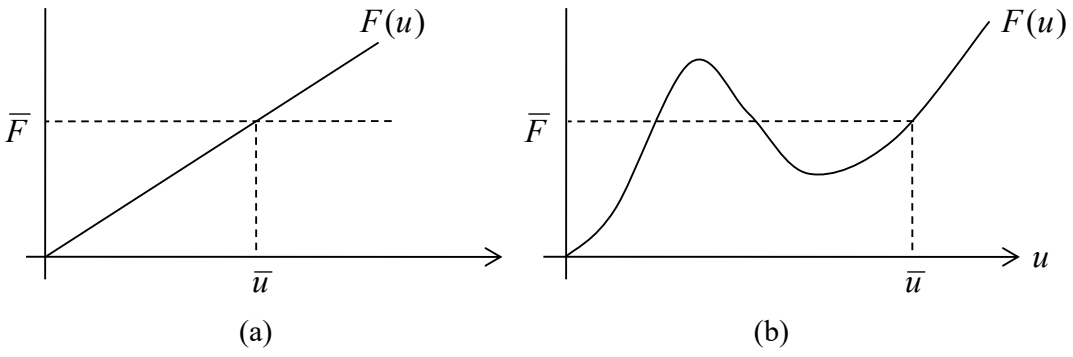


Figure 5.2: Example relationship between *F* and *u* in the residual function: (a) linear problem, (b) non-linear problem

For this reason, the problem is split up into a number of **increments**: the known "loading" function F is split into the increments $F_0, F_1, F_2, \ldots$ One first solves the equation

$$K(u) = F_0 \tag{5.16}$$

using the Newton-Raphson method until the solution converges. Once the solution is found, another increment in F is made, and the equation to be solved is

$$K(u) = F_1, \tag{5.17}$$

using the previous found value of u as the new prediction. F is incremented in this fashion until the final value is reached and the solution is obtained. The procedure is illustrated in Fig. 5.3.

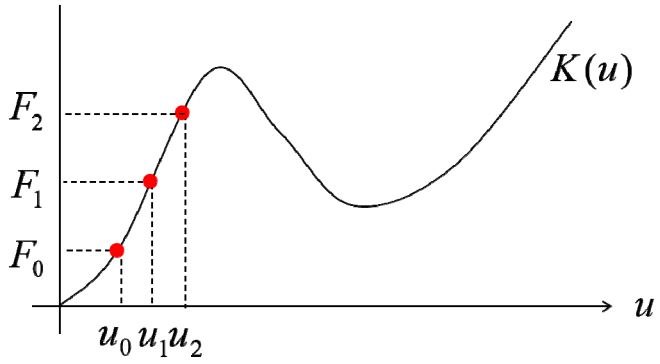


Figure 5.3: Solving non-linear equations using a number of stepped increments in the known function *F*

To be clear, the term **increment** is used to mean a change in F, whereas an **iteration** is used to mean a step in the Newton-Raphson algorithm (denoted by the "i" of Eqns. 5.10). Thus, at each increment, there are a number of iterations to convergence.

Assuming that one has equal increments in F, the algorithm can be written as (letting $F = (j-1)\hat{F}$, $j = 1, 2, \ldots$, with increment j and iteration i)

$$\begin{gathered}\left[K_T(a_j^{i-1})\right]\Delta a_j^i = -R(a_j^{i-1}) \qquad R(a_j^{i-1}) = K(a_j^{i-1}) - (j-1)\hat{F} \\ a_j^i = a_j^{i-1} + \Delta a_j^i\end{gathered} \tag{5.18}$$

Results are shown below for the one-dimensional example considered earlier, Eqn. 5.4, with four equal increments $\hat{F} = -1/6$, and three iterations per increment.

increment	Fj	iteration	a
0	0	0	1.00000000
		1	1.50000000
		2	1.35000000
		3	1.33353659
1	-1/6	0	1.33353659
		1	1.39581432
		2	1.39315469
		3	1.39314982
2	-1/3	0	1.39314982
		1	1.45050376
		2	1.44840544
		3	1.44840263
3	-1/2	0	1.44840263
		1	1.50170281
		2	1.50000174
		3	1.50000000
4	-2/3	0	1.50000000
		1	1.55000000
		2	1.54858491
		3	1.54858377

Initial stress method: use this K_T throughout the complete

Modified Newton-Raphson method: use this K_T for all iterations during this increment. Update K_T at the start of the

There are a number of variations of the incremental method:

1. The **initial stress method**: here the tangent matrix used at the start of the solution process is used throughout the analysis. The effort required with this method is greatly reduced, since only one evaluation of K_T is required. This reduced effort is offset by the fact that the scheme will inevitable converge more slowly (or even diverge).

2. The **modified Newton-Raphson method**: here the tangent matrix used at the start of an increment, $K_T(u_j^0)$, is used for all iterations during that increment – it is an approach somewhat in between the initial stress and full Newton-Raphson methods.

Quasi-Newton or **matrix update** methods, for example the BFGS and DFP schemes, are a compromise between the full Newton-Raphson method and the other schemes which use a tangent matrix method from a previous configuration. These schemes often involve secants to the curve.

All the above solution methods have their advantages, whether they be "constant *K*" or "variable *K*" methods. The precise choice of the optimal methodology is problem dependent and although many comparative solution cost studies have been published, the differences are often marginal. There is little doubt, however, that the full Newton-Raphson process has to be used when convergence is difficult to achieve. An automatic procedure that self-adaptively chooses an effective technique is most attractive.

Note: small increments reduce the total number of iterations required per increment and in many finite element software programs automatic guidance on the size of an increment to preserve a (nearly) constant number of iterations is provided.

Comparison with a simple explicit solution

The equations $\mathbf{K}(\mathbf{u}) = \mathbf{F}$ can also be solved incrementally using a simple explicit procedure. In this case one can write

$$\Delta\mathbf{K}(\mathbf{u}) = \Delta\mathbf{F} \tag{5.19}$$

or

$$\frac{\partial\mathbf{K}}{\partial\mathbf{u}}\Delta\mathbf{u} = \Delta\mathbf{F} \tag{5.20}$$

For example, consider again the one-dimensional equation $-2u^2 + \frac{8}{3}u = F$. Then $K_T = \partial K / \partial u = -4u + \frac{8}{3}$, and the algorithm is $K_T(u_k)\Delta u_k = \Delta F_k$. This can now be solved by specifying the "step size" ΔF_k. The solution is shown in Fig. 5.4 for $\Delta F_k = 0.1$ and $\Delta F_k = 0.2$. One can observe the drift away from the exact solution with successive steps, typical of explicit methods.

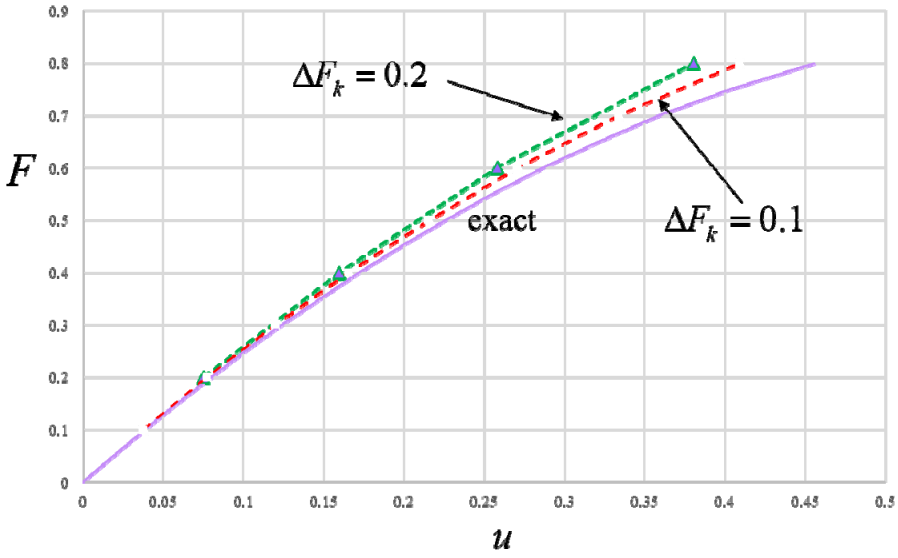


Figure 5.4: Explicit solution to one-dimensional equation

In contrast to the above explicit method, the Newton-Raphson scheme "pulls" the solution back towards the exact solution at each increment. This is illustrated in Fig. 5.5.

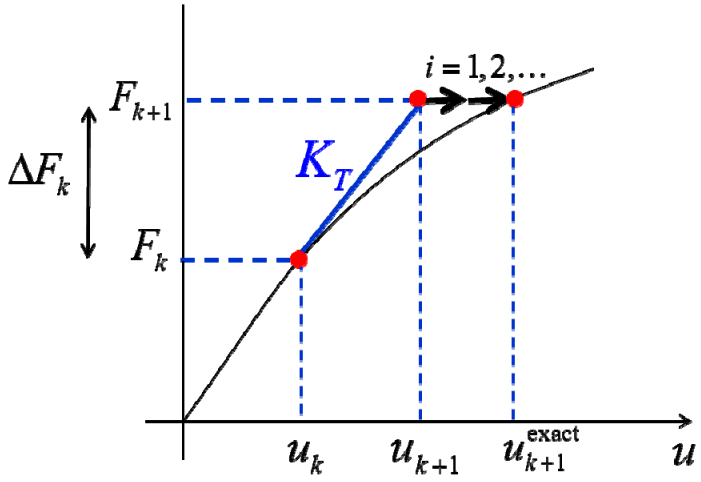


Figure 5.5: Newton-Raphson method pulling solution back towards exact solution at each increment

5.2 The FEM for the Solution of Non-Linear ODEs

Here, the FEM solution for non-linear ODEs is outlined by considering again the non-linear differential equation considered in Chapter 1, Eqn. 1.70,

$$2\frac{du}{dx}\frac{d^2u}{dx^2}+1=0, \qquad u(0)=0, \quad \left.\frac{\partial u}{\partial x}\right|_{x=1}=1 \tag{5.21}$$

[the exact solution is $u(x)=\frac{2}{3}\left(2^{3/2}-(2-x)^{3/2}\right)$]

Formation of the weighted residual, integration by parts, and substituting in the shape functions for linear elements gives

$$-u_1^2\int_0^1\left(\frac{dN_1}{dx}\right)^2\frac{dN_j}{dx}dx-2u_1u_2\int_0^1\frac{dN_1}{dx}\frac{dN_2}{dx}\frac{dN_j}{dx}dx-u_2^2\int_0^1\left(\frac{dN_2}{dx}\right)^2\frac{dN_j}{dx}dx+\int_0^1 N_j dx=-\left[\left(\frac{du}{dx}\right)^2\omega\right]_0^1 \tag{5.22}$$

Converting to local coordinates and evaluating the integrals:

$$\begin{aligned}
&\int_{x_i}^{x_{i+1}}\left(\frac{dN_i}{dx}\right)^2\frac{dN_j}{dx}dx=\left(\frac{d\xi}{dx}\right)^2\int_{-1}^{+1}\left(\frac{dN_i}{d\xi}\right)^2\frac{dN_j}{d\xi}d\xi=\frac{4}{L^2}\begin{bmatrix}-\frac{1}{4} & +\frac{1}{4}\\ -\frac{1}{4} & +\frac{1}{4}\end{bmatrix}\\
&\int_{x_i}^{x_{i+1}}\frac{dN_1}{dx}\frac{dN_2}{dx}\frac{dN_j}{dx}dx=\left(\frac{d\xi}{dx}\right)^2\int_{-1}^{+1}\frac{dN_1}{d\xi}\frac{dN_2}{d\xi}\frac{dN_j}{d\xi}d\xi=\frac{4}{L^2}\begin{bmatrix}+\frac{1}{4}\\ -\frac{1}{4}\end{bmatrix}
\end{aligned} \tag{5.23}$$

leads to the element equations

$$\begin{aligned}
&+u_1^2\frac{1}{L^2}-2u_1u_2\frac{1}{L^2}+u_2^2\frac{1}{L^2}+\frac{L}{2}=+\left(\frac{du}{dx}\right)^2\Bigg|_{x_i}\\
&-u_1^2\frac{1}{L^2}+2u_1u_2\frac{1}{L^2}-u_2^2\frac{1}{L^2}+\frac{L}{2}=-\left(\frac{du}{dx}\right)^2\Bigg|_{x_{i+1}}
\end{aligned} \tag{5.24}$$

Rather than form the global matrix, and thence derive the tangent matrix through differentiation of the global matrix, it is more efficient to form the elemental tangent matrix and then from there to form the global tangent matrix. The element tangent matrix is, differentiating the element matrix,

$$\mathbf{K}_T^{(el)} = \frac{2}{L^2}\begin{bmatrix} +u_1 - u_2 & -u_1 + u_2 \\ -u_1 + u_2 & +u_1 - u_2 \end{bmatrix} \tag{5.25}$$

Using two elements leads to the system of three non-linear equations (with $L = 1/2$)

$$\begin{aligned} +4u_1^2 - 8u_1u_2 + 4u_2^2 + \frac{1}{4} &= +\left(\frac{du}{dx}\right)^2\Bigg|_0 \\ -4u_1^2 + 8u_1u_2 - 8u_2u_3 + 4u_3^2 + \frac{1}{2} &= 0 \\ -4u_2^2 + 8u_2u_3 - 4u_3^2 + \frac{1}{4} &= -\left(\frac{du}{dx}\right)^2\Bigg|_1 \end{aligned} \tag{5.26}$$

Applying the boundary conditions $u_1 = u(0) = 0, \quad u'(1) = 1$ leads to

$$\begin{aligned} -8u_2u_3 + 4u_3^2 + \tfrac{1}{2} &= 0 \\ -4u_2^2 + 8u_2u_3 - 4u_3^2 + \tfrac{5}{4} &= 0 \end{aligned} \tag{5.27}$$

The equations are now solved using the Newton-Raphson method, for which

$$\mathbf{R}(\mathbf{u}) = 8\begin{bmatrix} -u_2u_3 + \frac{1}{2}u_3^2 + \frac{1}{16} \\ -\frac{1}{2}u_2^2 + u_2u_3 - \frac{1}{2}u_3^2 + \frac{5}{32} \end{bmatrix} \tag{5.28}$$

$$\mathbf{K}_T(\mathbf{u}) = 8\begin{bmatrix} -u_3 & -u_2 + u_3 \\ -u_2 + u_3 & u_2 - u_3 \end{bmatrix} \tag{5.29}$$

Solving the equations $\mathbf{K}_T(\mathbf{u}^{(i-1)})\Delta\mathbf{u}^{(i)} = -\mathbf{R}(\mathbf{u}^{(i-1)})$ iteratively, with the initial conditions $\mathbf{u}^{(0)} = \begin{bmatrix} 0 & 0.1 & 0.2 \end{bmatrix}^{\mathrm{T}}$, results in the solution sequence

$u_2^{(0)}$	0.10000000	$u_3^{(0)}$	0.20000000
$u_2^{(1)}$	2.23750000	$u_3^{(1)}$	3.85000000
$u_2^{(2)}$	1.21651536	$u_3^{(2)}$	2.11966459
$u_2^{(3)}$	0.78807456	$u_3^{(3)}$	1.41265492
$u_2^{(4)}$	0.67161254	$u_3^{(4)}$	1.23407069
$u_2^{(5)}$	0.66151490	$u_3^{(5)}$	1.22054242
$u_2^{(6)}$	0.66143783	$u_3^{(6)}$	1.22045483
$u_2^{(7)}$	0.66143783	$u_3^{(7)}$	1.22045482
exact	**0.66087321**	exact	**1.21895142**

Table 5.2: Solution of non-linear equations

Using three elements and applying the boundary conditions leads to

$$\begin{aligned} -18u_2u_3 + 9u_3^2 + \tfrac{1}{3} &= 0 \\ -9u_2^2 + 18u_2u_3 - 18u_3u_4 + 9u_4^2 + \tfrac{1}{3} &= 0 \\ -9u_3^2 + 18u_3u_4 - 9u_4^2 + \tfrac{7}{6} &= 0 \end{aligned} \tag{5.30}$$

so that

$$\mathbf{R}(\mathbf{u}) = 18\begin{bmatrix} -u_2u_3 + \frac{1}{2}u_3^2 + \frac{1}{54} \\ -\frac{1}{2}u_2^2 + u_2u_3 - u_3u_4 + \frac{1}{2}u_4^2 + \frac{1}{54} \\ -\frac{1}{2}u_3^2 + u_3u_4 - \frac{1}{2}u_4^2 + \frac{7}{108} \end{bmatrix} \tag{5.31}$$

$$\mathbf{K}_T(\mathbf{u}) = 18\begin{bmatrix} -u_3 & -u_2 + u_3 & 0 \\ -u_2 + u_3 & u_2 - u_4 & -u_3 + u_4 \\ 0 & -u_3 + u_4 & u_3 - u_4 \end{bmatrix} \tag{5.32}$$

which can be used to obtain better accuracy.

5.3 The FEM for the Solution of Non-Linear PDEs

Here, the FEM solution for non-linear PDEs is outlined by considering again Eqn. 5.21, but now with a first order term in *t*,

$$\frac{\partial u}{\partial t}+2\frac{\partial u}{\partial x}\frac{\partial^2 u}{\partial x^2}+1=0, \qquad u(0,t)=0, \quad \tfrac{\partial u}{\partial x}(1,t)=1, \quad u(x,0)=\sin(-2.074x) \tag{5.33}$$

The formulation of the FE problem is the same as before, only now one has a capacitance matrix **C**:

$$\int_{x_i}^{x_{i+1}} \frac{\partial u}{\partial t}\omega_j\,dx = \dot{u}_1\int_{x_i}^{x_{i+1}} N_1 N_j\,dx + \dot{u}_2\int_{x_i}^{x_{i+1}} N_2 N_j\,dx \quad\rightarrow\quad \mathbf{C}^{(el)}=\frac{L}{6}\begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix} \tag{5.34}$$

The time-independent part of the element equations can be written as $\mathbf{K}^{(el)}(\mathbf{u})=\mathbf{F}^{(el)}$,

$$\mathbf{K}^{(el)}(\mathbf{u})=\frac{1}{L^2}\begin{bmatrix}+u_1^2-2u_1u_2+u_2^2\\ -u_1^2+2u_1u_2-u_2^2\end{bmatrix}, \qquad \mathbf{F}^{(el)}=\begin{bmatrix}-\dfrac{L}{2}+(u')^2\Big|_{x_i}\\ -\dfrac{L}{2}-(u')^2\Big|_{x_{i+1}}\end{bmatrix} \tag{5.35}$$

with, as before,

$$\mathbf{K}_T^{(el)}=\frac{2}{L^2}\begin{bmatrix}+u_1-u_2 & -u_1+u_2\\ -u_1+u_2 & +u_1-u_2\end{bmatrix} \tag{5.36}$$

Using two elements and applying the boundary conditions leads to

$$\mathbf{C}\dot{\mathbf{u}}(t)+\mathbf{K}\big(\mathbf{u}(t)\big)=\mathbf{F}(t) \tag{5.37}$$

where

$$\mathbf{C}=\frac{1}{12}\begin{bmatrix}4 & 1\\ 1 & 2\end{bmatrix}, \quad \mathbf{K}(\mathbf{u}(t))=4\begin{bmatrix}-2u_2u_3+u_3^2\\ -u_2^2+2u_2u_3-u_3^2\end{bmatrix}, \quad \mathbf{F}=\begin{bmatrix}-\frac{1}{2}\\ -\frac{5}{4}\end{bmatrix}, \quad \mathbf{u}=\begin{bmatrix}u_2\\ u_3\end{bmatrix} \tag{5.38}$$

An explicit and an implicit means of solving these equations are discussed next.

5.3.1 Explicit formula

In the explicit scheme, the governing equations are written at time t, $\mathbf{C}\dot{\mathbf{u}}(t) + \mathbf{K}(\mathbf{u}(t)) = \mathbf{F}(t)$, and then the time derivative is replaced by the forward difference approximation, leading to

$$\mathbf{C}\mathbf{u}(t+\Delta t) = \mathbf{C}\mathbf{u}(t) + \Delta t\left[\mathbf{F}(t) - \mathbf{K}(\mathbf{u}(t))\right] \tag{5.39}$$

or, in incremental form,

Explicit Algorithm:

$$\begin{aligned} &\mathbf{C}\Delta\mathbf{u} = \mathbf{R}(t) \\ &\quad \text{where } \mathbf{R}(t) = \Delta t\left[\mathbf{F}(t) - \mathbf{K}(\mathbf{u}(t))\right] \\ &\mathbf{u}(t+\Delta t) = \mathbf{u}(t) + \Delta\mathbf{u} \end{aligned} \tag{5.40}$$

There is little difficulty in implementing this strategy – the only difference between this scheme and the corresponding linear one is that the linear equations $\mathbf{K}\mathbf{u}$ are replaced by the non-linear equations $\mathbf{K}(\mathbf{u})$. Note that there are *no* non-linear equations to be solved here, so *the Newton-Raphson scheme is not necessary*.

5.3.2 Implicit formula

The governing equations are written at time $t + \Delta t$ and then rewritten using the backward difference formula as

$$\mathbf{C}\left[\mathbf{u}(t+\Delta t) - \mathbf{u}(t)\right] = \Delta t\left[\mathbf{F}(t+\Delta t) - \mathbf{K}(\mathbf{u}(t+\Delta t))\right] \tag{5.41}$$

Introduce a residual function

$$\mathbf{R}(\mathbf{u}(t+\Delta t)) = \mathbf{C}\left[\mathbf{u}(t+\Delta t) - \mathbf{u}(t)\right] - \Delta t\mathbf{F}(t+\Delta t) + \Delta t\mathbf{K}(\mathbf{u}(t+\Delta t)) \tag{5.42}$$

Suppose that one already has an estimate of $\mathbf{u}(t+\Delta t)$, say $\mathbf{u}^{(0)}(t+\Delta t)$ - one can use the value $\mathbf{u}(t)$ as the initial prediction $\mathbf{u}^{(0)}(t+\Delta t)$ - or one could perhaps obtain an initial prediction by extrapolation through $\cdots, \mathbf{u}(t-\Delta t), \mathbf{u}(t)$.

One has $\mathbf{R}(\mathbf{u}^{(0)}(t+\Delta t)) \neq 0$ and it is required that the expression $\mathbf{R}(\mathbf{u}^{(1)}(t+\Delta t)) = 0$ holds. Following the Newton-Raphson procedure, expand in a Taylor series (note that the following is *for fixed time* $t+\Delta t$)

$$\mathbf{R}\left(\mathbf{u}^{(1)}(t+\Delta t)\right) = \mathbf{R}\left(\mathbf{u}^{(0)}(t+\Delta t)\right) + \Delta\mathbf{u}\frac{\partial \mathbf{R}}{\partial \mathbf{u}}\bigg|_{\mathbf{u}^{(0)}(t+\Delta t)} + \cdots = \mathbf{0} \tag{5.43}$$

where $\Delta\mathbf{u} = \mathbf{u}^{(1)}(t+\Delta t) - \mathbf{u}^{(0)}(t+\Delta t)$. Dropping the higher order terms gives the implicit algorithm

Implicit Algorithm:

$$\begin{gathered}
\left[\mathbf{K}_T\left(\mathbf{u}^{i-1}(t+\Delta t)\right)\right]\Delta\mathbf{u}^i(t+\Delta t) = -\mathbf{R}\left(\mathbf{u}^{i-1}(t+\Delta t)\right), \qquad \mathbf{K}_T = \frac{\partial \mathbf{R}}{\partial \mathbf{u}}\bigg|_{\mathbf{u}^{i-1}(t+\Delta t)} \\
\mathbf{R}\left(\mathbf{u}^{i-1}(t+\Delta t)\right) = \mathbf{C}\left[\mathbf{u}^{i-1}(t+\Delta t) - \mathbf{u}(t)\right] - \Delta t\mathbf{F}(t+\Delta t) + \Delta t\mathbf{K}\left(\mathbf{u}^{i-1}(t+\Delta t)\right) \\
\mathbf{K}_T = \mathbf{C} + \Delta t\frac{\partial \mathbf{K}}{\partial \mathbf{u}}\bigg|_{\mathbf{u}^{i-1}(t+\Delta t)}
\end{gathered} \tag{5.44}$$

Results using two elements are shown in Fig. 5.6.

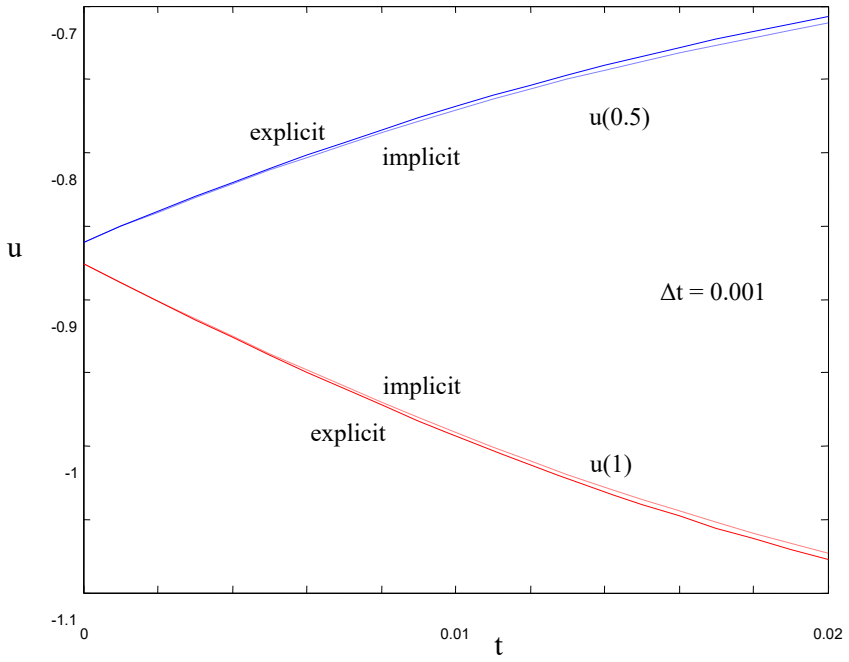


Figure 5.6: FE Solution to the PDE (5.30)

5.4 Problems

1. Solve the following system of two non-linear equations using the Newton-Raphson scheme

$$5 - 3a_1 + 2a_1 a_2 = 0$$
$$2 + a_2 - a_1^2 = 0$$

[the exact solution is $(1,-1)$ and $(1.1583,-0.6583)$]

5.5 Appendix to Chapter 5

5.5.1 The Substitution Method

Write the non-linear equation as

$$R = [K(a)]a - F, \qquad K(a) = -2a + \tfrac{8}{3}, \quad F(a) = -\tfrac{2}{3} \tag{5.A1}$$

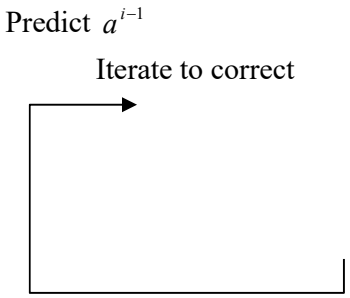
and so the residual for the approximation a^{i-1} is

$$R^i = [K(a^{i-1})]a^{i-1} - F \neq 0 \tag{5.A2}$$

The next approximation a^i is determined by solving the linear system $[K(a^{i-1})]a^i = F$, where a^{i-1} is the previous approximate value. In incremental form, one has, with $a^i = a^{i-1} + \Delta a^i$,

$$\begin{gathered}[K(a^{i-1})]\Delta a^i = -R^i, \qquad R^i = [K(a^{i-1})]a^{i-1} - F \\ a^i = a^{i-1} + \Delta a^i\end{gathered} \tag{5.A3}$$

One can now use the following algorithm:



for $i = 1, 2, ...$
solve $K(a^{i-1})\Delta a^i = R^i$
$a^i = a^{i-1} + \Delta a^i$
test convergence

The convergence test is usually of the form: is $|\Delta a^i| < \varepsilon$?, is $|R^i| < \varepsilon$?, is $\left|\frac{\Delta a^i}{a^i}\right| < \varepsilon$?

Using an initial estimate of $a = 1$ gives the sequence

$a^{(0)}$	1.00000000
$a^{(1)}$	-1.00000005

$a^{(2)}$	-0.14285715
$a^{(3)}$	-0.22580646
$a^{(4)}$	-0.21379311
$a^{(5)}$	-0.21545320
exact	**-0.21525044**

A geometrical interpretation of the iteration from $a^{(0)} \rightarrow a^{(1)}$ is shown here. The method applied to this equation cannot locate the root at $a = 1.549$. Further, the solution may diverge for certain functions (which undergo two or more changes in curvature within the search area).