FINITE ELEMENT METHOD

Abdusamad A. Salih
Department of Aerospace Engineering
Indian Institute of Space Science and Technology
Thiruvananthapuram - 695547, India.
salih@iist.ac.in

Contents

Chapter 1

Introduction

The finite element method usually abbreviated as FEM is a numerical technique to obtain approximate solution to physical problems. FEM was originally developed to study stresses in complex aircraft structures; it has since been extended and applied to the broad field of continuum mechanics, including fluid mechanics and heat transfer. Because of its capability to handle complex problems and its flexibility as a analysis tool, FEM has gained a prominent role in engineering analysis and design.

It must be emphasized that the FEM can only give you an approximate solution. So it is not the most desired way to solve a physical problem. The best way to solve a physical problem governed by a differential equation is to obtain a closed form analytical solution. Unfortunately, there are many practical situations where the analytical solution is difficult to obtain, or an analytical does not exist. For example, we may want to determine the drag force acting on an arbitrary shaped body kept in a viscous flow field. To obtain analytical solution, the shape of the body must be known in mathematical form. This is necessary to apply proper boundary conditions. If the shape of the body is irregular, so that no mathematical representation can be made, then it is impossible to solve the problem using analytical method. Even if the body has a regular shape, the governing differential equation of the problem may be nonlinear. There is no general analytical method available for the solution of nonlinear partial differential equations. However, for certain class of problems the troublesome nonlinear terms may naturally drops out from the equation, so that analytical solution can be attempted. But for most of the practical problems of interest, the governing equations are nonlinear. In such situations we have to resort to approximate numerical techniques for solving the problem.

There are several procedures to obtain a numerical solution to a differential equation. If the governing differential equation is a first-order ordinary differential equation, we have well-known methods such as Euler method, a variety of Runge-Kutta methods, or multi-step methods like Adam-Bashforth and Adam-Moulten methods to obtain numerical solution.

If the governing equation is a higher-order ordinary differential equation, it is possible to transform into a system of coupled first-order equations and then use any of the standard method developed for first-order equations. Not all physical problems are governed by ordinary differential

equation; in fact many problems in engineering and science requires the solution of partial differential equations.

There are several techniques to obtain the approximate solution of PDEs. Some of the popular methods are:

1. Finite Difference Method (FDM)
2. Finite Volume Method (FVM)
3. Finite Element Method (FEM)
4. Boundary Element Method (BEM)
5. Spectral Method
6. Perturbation Method (especially useful if the equation contains a small parameter)

1.1 Finite Difference Method

The finite difference method is the easiest method to understand and apply. To solve a differential equation using finite difference method, first a mesh or grid will be laid over the domain of interest. This process is called the discretization. A typical grid point in the mesh may be designated as i. The next step is to replace all derivatives present in the differential equation by suitable algebraic difference quotients. For example, the derivative

$$\frac{d\phi}{dx}$$

may be approximated as a first-order accurate forward difference quotient

$$\left.\frac{d\phi}{dx}\right|_i \approx \frac{\phi_{i+1} - \phi_i}{\Delta x}$$

or as a second-order accurate central difference quotient

$$\left.\frac{d\phi}{dx}\right|_i \approx \frac{\phi_{i+1} - \phi_{i-1}}{2\Delta x}$$

where Δx is the grid size and ϕ_i is the value of ϕ at at i^{th} grid point and is an unknown. This process yield an algebraic equation for the typical grid point i. The application of the algebraic equation to all interior grid point will generate a system of algebraic equation in which the grid point values of ϕ are unknowns. After the introduction of proper boundary conditions, the number of unknowns in the equation will be equal to the number of interior nodes in the mesh. The system (of equations) is typically solved using iterative methods such as Jacobi method, Gauss-Seidel method, or any of the advanced techniques.

We note that the finite difference method gives point-wise approximation to the differential equation and hence it gives the values of dependent variables at discrete points.

Using finite difference approach we can solve fairly difficult problems. It works well when the boundaries of the domain are parallel to the coordinate axes. But, we find that the method becomes harder to use when irregular boundaries are encountered. It is also difficult to write general purpose computer codes for FDM.

1.2 Finite Element Method

As mentioned earlier, the finite element method is a very versatile numerical technique and is a general purpose tool used to solve a variety of physical problems. It can be used to solve both field problems (governed by differential equations) and non-field problems.

There are several advantages of FEM over FDM. Among them, the most important advantage is that FEM is well suited for problem with complex geometries, because no special difficulties are encountered when the physical domain has a complex geometry. The other important advantage is that it is easier to write general purpose computer codes for FEM formulations.

Three different approaches are being used when formulating an FEM problem. They are:

1. Direct Approach
2. Variational Approach
3. Weighted Residual Method

1.2.1 Direct Approach

The direct approach is related to the "direct stiffness method" of structural analysis and it is the easiest to understand when meeting FEM for the first time. The main advantage of this approach is that you can get a feel of basic techniques and the essential concept involved in the FEM formulation without using much of mathematics. However, by direct approach we can solve only simple problems.

1.2.2 Variational Approach

In variational approach the physical problem has to be restated using some variational principle such as principle of minimum potential energy. It is widely used for deriving finite element equations whenever classical variational statement is available for the given problem. A basic knowledge of calculus of variations is required to use variational approach. The major disadvantage of the variational approach is that there exist many physical problems for which classical variational statement may not be available. This is the case with most of the nonlinear problems. In such cases variational approach is not useful. The Rayleigh-Ritz method is an approximate method based on the variational formulation.

1.2.3 Weighted Residual Method

Weighted residual method (WRM) is a class of method used to obtain the approximate solution to the differential equations of the form

$$\mathscr{L}(\phi) + f = 0 \quad \text{in} \quad D$$

where $\phi(\mathbf{x})$ is an unknown function and $f(\mathbf{x})$ is a known function of $\mathbf{x}$. In WRM, we directly work on differential equation of the problem without relying on any variational principle. It is equally

suited for linear and nonlinear differential equations. Weighted residual method involves two major steps. In the first step, we assume an approximate solution based on the general behavior of the dependent variable. The approximate solution is so selected that it satisfies the boundary conditions for ϕ. The assumed solution is then substituted in the differential equation. Since the assumed solution is only approximate, it does not satisfy the differential equation resulting in an error or what we call a *residual*. The residual is then made to vanish in some average sense over the *entire* solution domain. This procedure results in a system of algebraic equations. The second step is to solve the system of equations resulting from the first step subject to the prescribed boundary condition to yield the approximate solution sought.

In the later chapters we will discuss various weighted residual methods in detail after introducing the direct approach to FEM.

Chapter 2

Direct Approach to Finite Element Method

2.1 Introduction

The direct approach is related to the "direct stiffness method" of structural analysis and it is the easiest to understand when meeting FEM for the first time. The main advantage of this approach is that you can get a feel of basic techniques and the essential concept involved in the FEM formulation without using much of mathematics. However, by direct approach we can solve only simple problems.

The first step in this approach is to replace the system under consideration by an equivalent idealized system consisting of individual elements. These elements are assumed to be connected to each other at specified points called nodes. Once the elements in the system have been defined, one can use direct physical reasoning to establish the element equations in terms of pertinent variables. In the next step, the individual element equations are combined to form the equations for the complete system and solve the system of equations for the unknown nodal variables.

Since the fundamental idea of the discretization of the system (solution region) comes from structural analysis, we shall begin our discussion of finite element concept by considering a simple example from this area.

2.2 Linear Spring System

One of the most elementary systems that we can examine from an FEM point of view is the linear spring system. Let us consider a system of two springs connected in series in x-direction. One of the ends of the spring is rigidly attached to the wall, while the spring on the other end is free to move. Here forces, displacements, and spring stiffness are the only parameters in the system. We define each spring to be an element. So, our system consists of two elements and three nodes.

To determine the properties of an element, in this case the force-displacement equations, we isolate an element and draw its free body diagram. For the isolated spring element, F_i and F_j

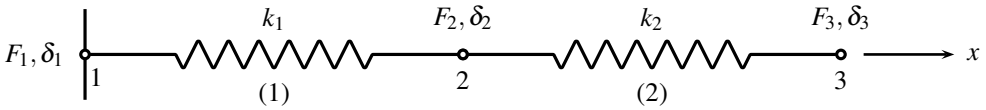

Figure 2.1: Linear springs in series

Figure 2.2: An isolated spring element

are nodal forces and δ_i and δ_j are the nodal displacements. The field (unknown) variable in this case is the displacement. Here we do not have to select an approximate solution (interpolation polynomial) to represent the variation of the field variable over the element, because an exact representation of force-displacement relation is available. By physical reasoning, we can establish the such an element equation. Here, the simple Hook's law gives the required force-displacement relation:

$$F = k\delta$$

for a single spring fixed at one end, where k is the spring stiffness.

Returning to the isolated spring, we allow the element to adopt each independent mode of displacement and apply the Hook's law. The sign convention is illustrated below:

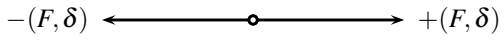

Figure 2.3: Sign convention

We have different cases here. In the first, it is assumed that only node i can deflect while the end j is being fixed. So, we have $F_{ia} = k\delta_i$. Equilibrium of forces acting on the spring requires that

$$F_{ia} + F_{ja} = 0$$
$$F_{ia} = -F_{ja} = k\delta_i$$

It should be noted that the continuity requirements of displacement is automatically satisfied for this simple spring. In the next case, we reverse the situation by fixing node i in its initial position and allowing node j to deflect under the action of force F_{jb}. So, we have $F_{jb} = k\delta_j$ and

$$F_{ib} + F_{jb} = 0$$
$$F_{jb} = -F_{ib} = k\delta_j$$

Now, if both the nodes are allowed to deflect at the same time, the relationship between nodal forces and nodal displacements can be obtained by the principle of superposition of first

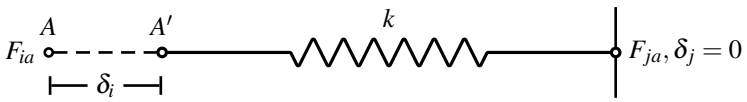

Figure 2.4: Case (a): node j is fixed and node i deflects.

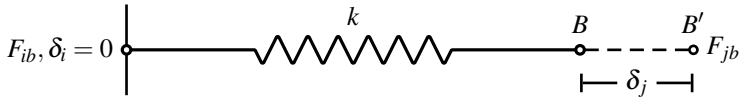

Figure 2.5: Case (b): node i is fixed and node j deflects.

two cases. Thus, the total force at node i,

$$F_i = F_{ia} + F_{ib} = k\delta_i - k\delta_j$$

and the total force at node j,

$$F_j = F_{ja} + F_{jb} = -k\delta_i + k\delta_j$$

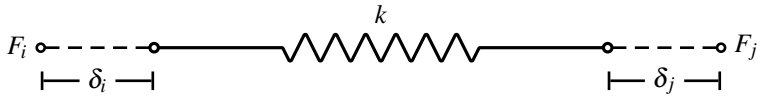

Figure 2.6: Node i and node j deflect.

Using matrix notation, the above set of equations can be combined and written in compact form

$$\begin{bmatrix} k & -k \\ -k & k \end{bmatrix} \begin{bmatrix} \delta_i \\ \delta_j \end{bmatrix} = \begin{bmatrix} F_i \\ F_j \end{bmatrix} \qquad \Longrightarrow \qquad \begin{bmatrix} k^e \end{bmatrix} \begin{bmatrix} \delta \end{bmatrix} = \begin{bmatrix} F \end{bmatrix} \tag{2.1}$$

where the square matrix $[k^e]$ is known as the element stiffness matrix, column vector $[\delta]$ is the nodal displacement vector, and the column vector $[F]$ is the nodal force vector for the element. Although the above element equation is derived for a simple system of finite elements, the general form of the element equation remains the same, regardless of the type of the problem and the complexity of the element. The form of the equation is also the same irrespective of the way in which the element properties are derived.

Having derived the element equation for a single element, our next objective is to obtain an equation for complete system. To do this, we proceed in the same manner as we did in the case of individual element.

Referring to figure 2.8, we set δ_2 and δ_3 equal to zero, allowing only node '1' to deflect. Considering the left spring, the laws of statics gives

$$F_2 = -F_1 \qquad \text{and} \qquad F_1 = -k_a\delta_1$$

Since δ_2 and δ_3 are specified as zero, no force can exists at node '3'. So,

$$F_3 = 0$$

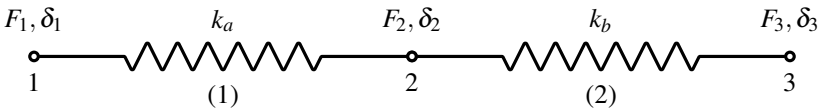

Figure 2.7: Combined case (a) and case (b).

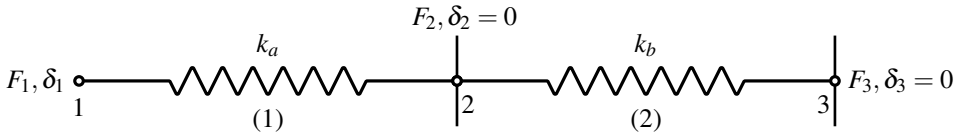

Figure 2.8: Node 2 and node 3 fixed.

Next, referring to figure 2.9, we set δ_1 and δ_3 equal to zero. In this case, continuity of displacement requires that both springs deflects by the same amount; thus force at node '2' consists of two components, $k_a\delta_2$ and $k_b\delta_2$. Therefore,

$$\begin{aligned} F_2 &= (k_a + k_b)\delta_2 \\ F_1 &= -k_a\delta_2 \\ F_3 &= -k_b\delta_2 \end{aligned}$$

and

$$F_1 + F_2 + F_3 = 0$$

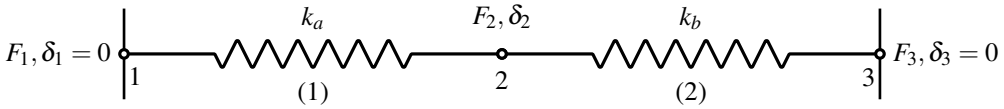

Figure 2.9: Node 1 and node 3 fixed.

Finally, referring to figure 2.10, we set δ_1 and δ_2 equal to zero to obtain

$$\begin{aligned} F_3 &= k_b\delta_3 \\ F_2 &= -F_3 = -k_b\delta_3 \\ F_1 &= 0 \end{aligned}$$

Now, using the principle of superposition, we combine all all the three cases to obtain the stiffness matrix for the system. The total forces at three nodes are given by

$$\begin{aligned} F_1 &= k_a\delta_1 - k_a\delta_2 + 0 \\ F_2 &= -k_a\delta_1 + k_a\delta_2 + k_b\delta_2 - k_b\delta_3 \\ F_3 &= 0 - k_b\delta_2 + k_b\delta_3 \end{aligned}$$

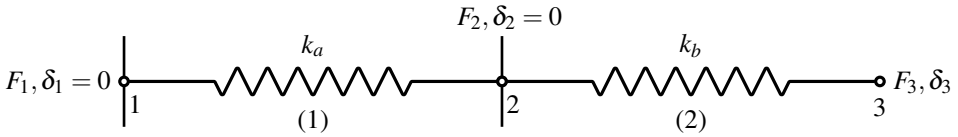

Figure 2.10: Node 1 and node 2 fixed.

In matrix form:

$$\begin{bmatrix} k_a & -k_a & 0 \\ -k_a & k_a+k_b & -k_b \\ 0 & -k_b & k_b \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix} = \begin{bmatrix} F_1 \\ F_2 \\ F_3 \end{bmatrix} \quad \Longrightarrow \quad [\ K\ ][\ \delta\ ] = [\ F\ ] \tag{2.2}$$

where $[K]$ is the complete stiffness matrix of the system.

The assembly of the stiffness matrix is not difficult in this simple case, but this method of constructing stiffness matrix for the system would be extremely tedious if the system comprised of large number of elements (springs). We have simple and straightforward way of assembling stiffness matrix if individual element matrix are known. The element matrix for individual elements 1 and 2 are given respectively by

$$\begin{bmatrix} k_a & -k_a \\ -k_a & k_a \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} = \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} k_b & -k_b \\ -k_b & k_b \end{bmatrix} \begin{bmatrix} \delta_2 \\ \delta_3 \end{bmatrix} = \begin{bmatrix} F_2 \\ F_3 \end{bmatrix}$$

Although the two elements matrices are of the same order, they may not be added directly, since they relate to different sets of displacements. By inserting rows and columns with zeroes, both matrices can be expanded as follows:

$$\begin{bmatrix} k_a & -k_a & 0 \\ -k_a & k_a & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix} = \begin{bmatrix} F_1 \\ F_2 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & 0 & 0 \\ 0 & k_b & -k_b \\ 0 & -k_b & k_b \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ F_2 \\ F_3 \end{bmatrix}$$

Now, both the stiffness matrix can be added to obtain

$$\begin{bmatrix} k_a & -k_a & 0 \\ -k_a & k_a+k_b & -k_b \\ 0 & -k_b & k_b \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix} = \begin{bmatrix} F_1 \\ F_2 \\ F_3 \end{bmatrix} \quad \Longrightarrow \quad [\ K\ ][\ \delta\ ] = [\ F\ ]$$

This sequence of operation is identical to superposition principle used earlier to obtain the complete stiffness matrix of the system. So, stiffness matrices of individual elements can be suitably added to obtain the complete system matrix of the system.

2.3 Solution of System of Equations

The system of equation (2.2) may be rewritten as

$$[\ \delta\ ] = [\ K\ ]^{-1}[\ F\ ] \tag{2.3}$$

The system matrix $[K]$ in (2.3) is singular, since the determinant is zero, so the inverse of $[K]$ does not exist. This means that the system of equations cannot be solved for δ! However, a perfectly simple explanation exists for this dilemma; the structure has not been secured to the ground (wall). Therefore, application of any external force on the structure would result in the system moving as a rigid body. This situation can be remedied, if we secure any of the node to the ground, thereby that node is constrained to have zero displacement. This constraint becomes the boundary condition. Assume node '1' to be fixed, $(\delta_1 = 0)$, then

$$\begin{bmatrix} k_a & -k_a & 0 \\ -k_a & k_a+k_b & -k_b \\ 0 & -k_b & k_b \end{bmatrix} \begin{bmatrix} \delta_1 = 0 \\ \delta_2 \\ \delta_3 \end{bmatrix} = \begin{bmatrix} F_1 \\ F_2 \\ F_3 \end{bmatrix} \tag{2.4}$$

The system of equation (2.4) contains an unknown reaction F_1 and two unknown displacements δ_2 and δ_3. F_2 and F_3 are known applied forces. The matrix equation (2.4) can be broken into two:

$$\begin{bmatrix} -k_a & 0 \end{bmatrix} \begin{bmatrix} \delta_2 \\ \delta_3 \end{bmatrix} = \begin{bmatrix} F_1 \end{bmatrix} \tag{2.5a}$$

$$\begin{bmatrix} k_a+k_b & -k_b \\ -k_b & k_b \end{bmatrix} \begin{bmatrix} \delta_2 \\ \delta_3 \end{bmatrix} = \begin{bmatrix} F_2 \\ F_3 \end{bmatrix} \tag{2.5b}$$

Equation (2.5b) can be solved for δ_2 and δ_3 and their values can be substituted in (2.5a) for finding the value of unknown reaction force F_1.

Once the displacements are obtained, the internal forces in the elements may be computed as follows:

$$P_1 = k_a(\delta_2 - \delta_1) = k_a\delta_2$$
$$P_2 = k_b(\delta_3 - \delta_2)$$

This completes the solution process.

Example 2.1

Obtain the system stiffness matrix for the linear spring system below. Also find the force in spring '3'.

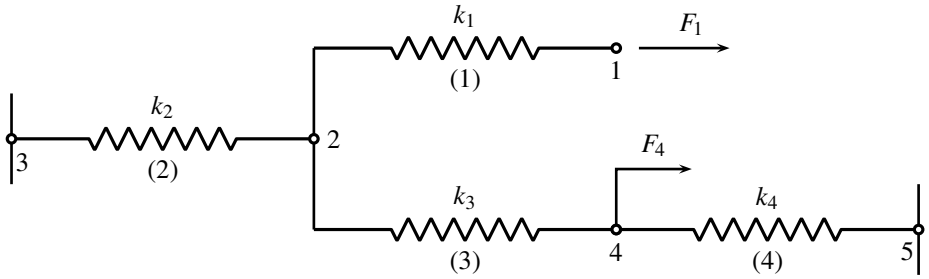

Figure 2.11: Example 1.

2.4 Direct Approach to Steady-Sate Heat Conduction Problem

We shall now take up a problem in heat conduction to show how element properties can be established by direct physical reasoning. Consider the problem of one-dimensional heat flow through composite wall. Here, we have a section of layered material through which heat is conducted in x-direction. To simplify the analysis let us assume that there is no internal heat generation present. The left-hand side of the wall is held at a higher temperature than that of the right-hand

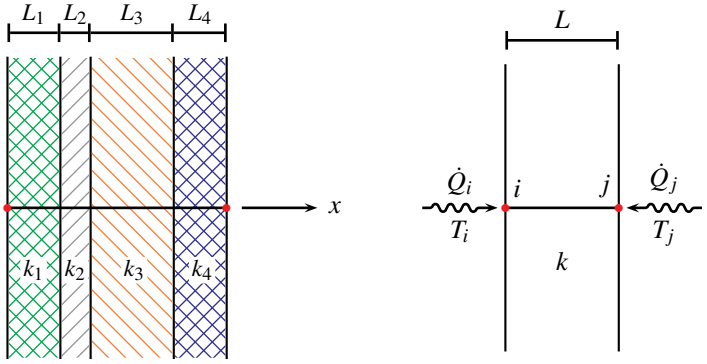

Figure 2.12: Heat flow through composite wall.

side, and each layer is a homogeneous material whose thermal conductivities are known. The pertinent parameters of the problem are heat flux, temperature, thermal conductivity, and layer thickness. The field variable for this problem is the temperature. The problem can be split into a series of simple ones, if we consider each layer as a finite element whose properties can be determined by the basic law of heat conduction. The nodes of the element here are not points but the boundary planes of the layer and each node will have a definite temperature. So, our system consists of four elements with five nodes. An isolated element is shown in figure. Here we can obtain exact heat flow behavior of an element by using *Fourier law of conduction*. So again, we do not have to assume an interpolation function over the elements. In the present case of one-dimensional conduction, the heat flow rate is given by

$$\dot{Q} = -kA\frac{dT}{dx}$$

where k is the thermal conductivity of the material and A is the area normal to heat flow direction.

For a typical element,

$$\dot{Q} = -kA\frac{\Delta T}{L}$$

where ΔT is the temperature drop across the element whose thickness is L. We can express the nodal heat flows, in terms of element nodal temperatures:

$$\dot{Q}_i = \frac{kA}{L}(T_i - T_j)$$

Since the conservation of energy requires that $\dot{Q}_i + \dot{Q}_j = 0$, we have

$$\dot{Q}_j = -\frac{kA}{L}(T_i - T_j)$$

In matrix form,

$$\frac{kA}{L}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} T_i \\ T_j \end{bmatrix} = \begin{bmatrix} \dot{Q}_i \\ \dot{Q}_j \end{bmatrix} \qquad \Longrightarrow \qquad \begin{bmatrix} k^e \end{bmatrix}\begin{bmatrix} T \end{bmatrix} = \begin{bmatrix} \dot{Q} \end{bmatrix} \tag{2.6}$$

where the square matrix $[k^e]$ is known as the element matrix of thermal conductance, column vector $[T]$ is the nodal temperature vector, and the column vector $\dot{Q}$ is the nodal heat flow vector for the element.

Equation (2.6) is in standard form and it completely defines the heat conduction properties of the simple thermal element. Later when we we consider the general heat conduction problem, we will see that the element properties will again be expressed in standard form, only difference will be the dimension of the matrix $[k]$ and complexity of its terms.

After finding the necessary algebraic equations describing the characteristics of each element in the system, we can combine (assemble) all the element equations to form a complete set of equations governing the system. The procedure for constructing the system equations is same regardless of the type of problem and complexity of the element. Even if the system is modeled with a mixture of several different kinds of elements, the system equations are assemble from the element equations in the same way. So we can simply adopt the procedure discussed in the case of spring system for assembling the individual element matrix. By putting $kA/L = K$, the element equation (2.6) becomes

$$\begin{bmatrix} K & -K \\ -K & K \end{bmatrix}\begin{bmatrix} T_i \\ T_j \end{bmatrix} = \begin{bmatrix} \dot{Q}_i \\ \dot{Q}_j \end{bmatrix}$$

Adding the individual matrices, we obtain

$$\begin{bmatrix} K_1 & -K_1 & 0 & 0 & 0 \\ -K_1 & K_1+K_2 & -K_2 & 0 & 0 \\ 0 & -K_2 & K_2+K_3 & -K_3 & 0 \\ 0 & 0 & -K_3 & K_3+K_4 & -K_4 \\ 0 & 0 & 0 & -K_4 & K_4 \end{bmatrix}\begin{bmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5 \end{bmatrix} = \begin{bmatrix} \dot{Q}_1 \\ \dot{Q}_2 \\ \dot{Q}_3 \\ \dot{Q}_4 \\ \dot{Q}_5 \end{bmatrix}$$

It must be noted that before solving the above set of equations, we must substitute the boundary conditions, for example, $T_1 = T_{\text{hot}}$ and $T_5 = T_{\text{cold}}$.

Chapter 3

Calculus of Variations

3.1 Introduction

The calculus of variations deals with functionals, which are functions of a function, to put it simply. For example, the methods of calculus of variations can be used to find an unknown function that minimizes or maximizes a functional. Many of its methods were developed over two hundred years ago by Euler (1701-1783), Lagrange (1736-1813), and others. It continues to the present day to bring important techniques to many branches of engineering and physics.

3.2 Functionals

As we have seen in the last section, there exist a great variety of physical problems that deals with functionals, which are functions of a function. We are familiar with the definition of a function. A function can be regarded as a rule that maps one number (or a set of numbers) to another value. For example,

$$f(x) = x^2 + 2x$$

is a function, which maps $x = 2$ to $f(x) = 8$, and $x = 3$ to $f(x) = 15$, etc. On the other hand, a functional is a mapping from a function (or a set of functions) to a value. That is, a functional is a rule that assigns a real number to each function $y(x)$ in a well-defined class. Like a function, a functional is a rule, but its domain is some set of functions rather than a set of real numbers. We can consider $F[y(x)]$ as a functional for the fixed values of x. For example,

$$F[y(x)] = 3y^2 - y + 10$$

where

$$y(x) = e^x + \cos x - x \qquad \text{for} \qquad x = \pi$$

is a functional. Another class of functional has the form

$$J[y] = \int_a^b y(x)\,dx$$

Here J gives the area under the curve $y = y(x)$. Hence J is not a function of x and its value will be a number. However, this number depends on the particular form of $y(x)$ and hence $J[y]$ is a functional. For $a = 0$ and $b = \pi$, the value of the functional when $y(x) = x$ is

$$J[y] = \int_0^\pi x\,dx = \frac{\pi^2}{2} \approx 4.93$$

and when $y(x) = \sin x$,

$$J[y] = \int_0^\pi \sin x\,dx = 2$$

Therefore the given functional $J[y]$ maps $y(x) = x$ to $\pi^2/2$ and maps $y(x) = \sin x$ to 2. Because an integral maps a function to a number, a functional usually involves an integral. The following form of functional often appears in the calculus of variations,

$$J[y] = \int_a^b F(x, y, y')\,dx \tag{3.1}$$

The fundamental problem of the calculus of variations is to find the extremum (maximum or minimum) of the functional (3.1).

3.3 First Variation of Functionals

Consider a function $y = f(x)$. When the independent variable x changes to $x + \Delta x$, then the dependent variable y changes to $f(x + \Delta x) = f(x) + \Delta f(x)$, where Δf is the total change in the function. Δf can be computed by expanding $f(x + \Delta x)$ using Taylor series. Thus,

$$\begin{aligned} f(x+\Delta x) &= f(x) + \frac{df}{dx}\Delta x + \frac{d^2f}{dx^2}\frac{\Delta x^2}{2!} + \frac{d^3f}{dx^3}\frac{\Delta x^3}{3!} + \dots \\ \Delta f \equiv f(x+\Delta x) - f(x) &= \frac{df}{dx}\Delta x + \frac{d^2f}{dx^2}\frac{\Delta x^2}{2!} + \frac{d^3f}{dx^3}\frac{\Delta x^3}{3!} + \dots \end{aligned} \tag{3.2}$$

By definition, the differential df of the function $f(x)$ is how much f changes if its argument, x, changes by an infinitesimal amount Δx. That is

$$df = \lim_{\Delta x \to 0} \Delta f = \frac{df}{dx}\Delta x \tag{3.3}$$

Comparing (3.2) and (3.3), we see that the differential df is the linear part of the total change Δf. That is

$$\Delta f = df + \text{higher-order terms in } \Delta x \tag{3.4}$$

In line with definition of differential of a function $f(x)$, we now introduce the concept of the variation (or differential) of a functional $F[y(x)]$. Let $y(x)$ is changed to $y(x) + \delta y(x)$, where $\delta y(x)$ is the vertical displacement of the curve $y(x)$. It is known as the variation of y and is denoted by δy. We introduce an alternative function of the form

$$Y(x) = y(x) + \delta y(x) \tag{3.5}$$

This is illustrated in figure 3.1, where $y(x)$ is shown in red color and $Y(x)$ is shown in blue color.

By definition, the total change in the functional is given by

$$\Delta F[y] = F[y(x) + \delta y(x)] - F[y(x)] = F[Y(x)] - F[y(x)] \tag{3.6}$$

If $\eta(x)$ is an arbitrary differentiable function that vanishes at the boundaries of the domain, i.e.,

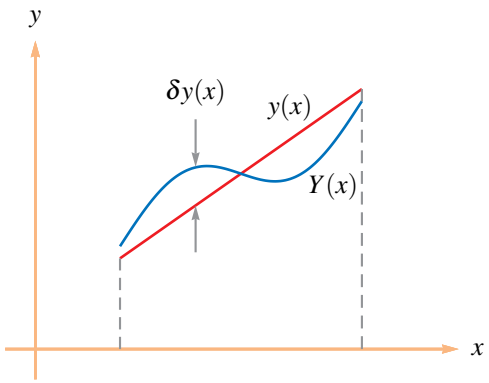

Figure 3.1: Plot of $y(x)$ and a small variation from it.

$\eta(a) = 0$ and $\eta(b) = 0$, then the variation $\delta y(x)$ can be represented as

$$\delta y(x) = \varepsilon \eta(x) \qquad y, \eta \in A \tag{3.7}$$

where ε is an arbitrary parameter independent of x. This definition enable us to write equation (3.5) in the following form,

$$Y = y + \varepsilon \eta \tag{3.8}$$

Now from (3.6), the total change in functional F is given by

$$\Delta F = F[y + \varepsilon \eta] - F[y] \tag{3.9}$$

Using Taylor series, we can expand the first term on R.H.S. as

$$F[Y] = F[y + \varepsilon \eta] = F[y] + \frac{dF}{dy}\eta\varepsilon + \frac{d^2F}{dy^2}\eta^2\frac{\varepsilon^2}{2!} + \frac{d^3F}{dy^3}\eta^3\frac{\varepsilon^3}{3!} + \dots \tag{3.10}$$

Rearranging equation (3.10) to obtain the change in functional F:

$$\Delta F = F[y + \varepsilon \eta] - F[y] = \frac{dF}{dy}\eta\varepsilon + \text{higher-order terms} \tag{3.11}$$

By definition, first variation of a functional $F[y]$, denoted by δF, is how much F changes if its argument, y, changes by an infinitesimal amount δy. Therefore,

$$\delta F = \lim_{\varepsilon \to 0} \Delta F = \lim_{\varepsilon \to 0} (F[y + \varepsilon \eta] - F[y]) = \frac{dF}{dy}\eta\varepsilon = \frac{dF}{dy}\delta y \tag{3.12}$$

which shows that δF is given by the linear part of the equation (3.11). Thus, the change in functional $F[y]$ and its first variation is related by the equation

$$\Delta F = \delta F + \text{higher-order terms} \tag{3.13}$$

Let us now define what is called the *Gâteaux derivative* or *Gâteaux variation* in the direction of $\eta(x)$. It is denoted by $\delta F[y;\eta]$ and is defined as

$$\delta F[y;\eta] = \lim_{\varepsilon\to 0}\frac{\Delta F}{\varepsilon} = \lim_{\varepsilon\to 0}\frac{F[y+\varepsilon\eta]-F[y]}{\varepsilon} = \frac{d}{d\varepsilon}F[y+\varepsilon\eta]\bigg|_{\varepsilon=0} \tag{3.14}$$

Note that the first variation and the Gâteaux variation are related through the parameter ε, i.e., $\delta F_{(fv)} = \varepsilon\delta F_{(gv)}$ where we have denoted first variation by $\delta F_{(fv)}$ and Gâteaux variation by $\delta F_{(gv)}$. Unfortunately, in the literature, these two variations are denoted by the same symbol δF.

Let us look at the meaning of η and ε geometrically. Since y is the unknown function to be found so as to extremize a functional, we want to see what happens to the functional $F[y]$ when we perturb this function slightly. For this, we take another function η and multiply it by a small number ε. We add $\varepsilon\eta$ to y and look at the value of $F[y+\varepsilon\eta]$. That is, we look at the perturbed value of the functional due to perturbation $\varepsilon\eta$. This is the shaded area shown in figure 3.2. Now as $\varepsilon\to 0$, we consider the limit of the shaded area divided by ε. If this limit exists, such a limit is called the Gâteaux variation of $F[y]$ at y for an arbitrary but fixed function η.

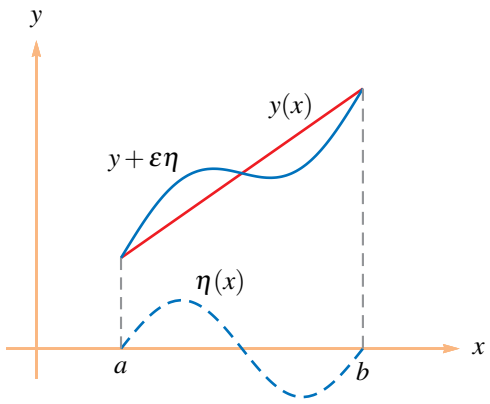

Figure 3.2: Plot of $y(x)$ and its variation.

Note that choosing a different η gives a different set of varied curves and hence a different variation. Hence $\delta F[y;\eta]$ depends on which function η is chosen to define the increment δy and this dependence is explicitly shown in the notation.

First variation of functional $F[x,y,y',y'']$

We now consider the first variation of the functional

$$F[x, y, y', y'']$$

for fixed values of x. If y changes to $y+\varepsilon\eta$, then y' changes to $y'+\varepsilon\eta'$ and y'' changes to $y''+\varepsilon\eta''$. From equation (3.8), we have

$$\begin{aligned} Y &= y+\varepsilon\eta \\ Y' &= y'+\varepsilon\eta' \qquad \text{and} \\ Y'' &= y''+\varepsilon\eta'' \end{aligned}$$

The new value of the functional is then

$$F[x,Y,Y',Y''] = F[x,y+\varepsilon\eta,y'+\varepsilon\eta',y''+\varepsilon\eta'']$$

where $\varepsilon\eta'$ is known as the variation of y' and is denoted by $\delta y'$. Similarly, $\varepsilon\eta''$ is known as the variation of y'' and is denoted by $\delta y''$. The change in the functional F is then defined as

$$\Delta F = F[x,y+\varepsilon\eta,y'+\varepsilon\eta',y''+\varepsilon\eta''] - F[x,y,y',y''] \tag{3.15}$$

Using Taylor series, we can expand the first term on R.H.S. as

$$\begin{aligned} F[x,y+\varepsilon\eta,y'+\varepsilon\eta',y''+\varepsilon\eta''] &= F[x,y,y',y''] + \left(\frac{\partial F}{\partial y}\eta+\frac{\partial F}{\partial y'}\eta'+\frac{\partial F}{\partial y''}\eta''\right)\varepsilon \\ &+\left(\frac{\partial^2 F}{\partial y^2}\eta^2+\frac{\partial^2 F}{\partial y'^2}\eta'^2+\frac{\partial^2 F}{\partial y''^2}\eta''^2+2\frac{\partial^2 F}{\partial y\partial y'}\eta\eta'+2\frac{\partial^2 F}{\partial y\partial y''}\eta\eta''+2\frac{\partial^2 F}{\partial y'\partial y''}\eta'\eta''\right)\frac{\varepsilon^2}{2!}+\dots \end{aligned}$$

Rearranging the above Taylor series expansion, we obtain the change in functional F:

$$\begin{aligned} \Delta F = \left(\frac{\partial F}{\partial y}\eta+\frac{\partial F}{\partial y'}\eta'+\frac{\partial F}{\partial y''}\eta''\right)\varepsilon + \Bigg(&\frac{\partial^2 F}{\partial y^2}\eta^2+\frac{\partial^2 F}{\partial y'^2}\eta'^2+\frac{\partial^2 F}{\partial y''^2}\eta''^2 \\ &+2\frac{\partial^2 F}{\partial y\partial y'}\eta\eta'+2\frac{\partial^2 F}{\partial y\partial y''}\eta\eta''+2\frac{\partial^2 F}{\partial y'\partial y''}\eta'\eta''\Bigg)\frac{\varepsilon^2}{2!}+\dots \end{aligned}$$

In analogy with the definition of a function, the sum of the linear part in the ΔF is called the first variation of the functional F. Therefore,

$$\delta F = \frac{\partial F}{\partial y}\eta\varepsilon+\frac{\partial F}{\partial y'}\eta'\varepsilon+\frac{\partial F}{\partial y''}\eta''\varepsilon \tag{3.16}$$

Since

$$\delta y = \varepsilon\eta\,, \qquad \delta y' = \varepsilon\eta'\,, \qquad \delta y'' = \varepsilon\eta''$$

The variation of F can be written as

$$\delta F = \frac{\partial F}{\partial y}\delta y+\frac{\partial F}{\partial y'}\delta y'+\frac{\partial F}{\partial y''}\delta y'' \tag{3.17}$$

Now, the total differential dF of a function $F(x,y,y',y'')$, when x is considered fixed, is given by

$$dF = \frac{\partial F}{\partial y}dy+\frac{\partial F}{\partial y'}dy'+\frac{\partial F}{\partial y''}dy''$$

Formula (3.17) for δF has the same form as the above formula for dF. Thus the variation of F is given by the same formula as differential of F, if x is considered to be fixed.

It is to be noted that the differential of a function is the first-order approximation to the change in that function, along a particular curve while the variation of a functional is the first-order approximation to the change in the functional from one curve to other.

We mention here that the sum of terms in ε and ε^2 is called the second variation of F and the sum of terms in ε, ε^2, and ε^3 is called the third variation of F. However, when the term variation is used alone, the first variation is meant.

Some rules of variational calculus

The variational operator δ follows the rules of differential operator d of calculus. Let F_1 and F_2 be any continuous and differentiable functionals. Then we have the following results:

- $\delta F^n = nF^{n-1}\delta F$
- $\delta(F_1 + F_2) = \delta F_1 + \delta F_2$
- $\delta(F_1 F_2) = F_1\delta F_2 + F_2\delta F_1$
- $\delta\left(\frac{F_1}{F_2}\right) = \frac{F_2\delta F_1 - F_1\delta F_2}{F_2^2}$

It is easy to show that the operators $\frac{d}{dx}$ and δ are commutative. The commutative property may be written mathematically as

$$\frac{d}{dx}(\delta y) = \delta\frac{dy}{dx}$$

The proof is as follows:

$$\frac{d}{dx}(\delta y) = \frac{d}{dx}(\varepsilon\eta) = \varepsilon\frac{d\eta}{dx} = \varepsilon\eta' = \delta y' = \delta\frac{dy}{dx}$$

That is, the differential of the variation of a function is identical to the variation of the differential of the same function.

Another commutative property is the one that states that the variation of the integral of a functional F is the same as the integral of the variation of the same functional, or mathematically

$$\delta\int F dx = \int \delta F dx$$

Note that the two integrals must be evaluated between the same two limits.

First variation of functional $\int_a^b F(x, y, y', y'')\,dx$

Next we consider the first variation of the functional defined by

$$J[y] = \int_a^b F(x, y, y', y'')\,dx$$

If y changes to $Y = y + \varepsilon\eta$, then y' changes to $Y' = y' + \varepsilon\eta'$ and y'' changes to $Y'' = y'' + \varepsilon\eta''$. The change in functional, ΔJ, is given by

$$\Delta J = J[Y] - J[y] = J[y + \varepsilon\eta] - J[y] \tag{3.18}$$

where

$$J[y + \varepsilon\eta] = \int_a^b F[x, y + \varepsilon\eta, y' + \varepsilon\eta', y'' + \varepsilon\eta'']\, dx$$

Therefore, the change in functional is given by

$$\Delta J = \int_a^b F[x, y + \varepsilon\eta, y' + \varepsilon\eta', y'' + \varepsilon\eta'']\, dx - \int_a^b F(x, y, y', y'')\, dx \tag{3.19}$$

As previously defined, the Gâteaux derivative or Gâteaux variation in the direction of $\eta(x)$ is given by

$$\delta J[y;\eta] = \lim_{\varepsilon\to 0} \frac{\Delta J}{\varepsilon} = \lim_{\varepsilon\to 0} \frac{J[y + \varepsilon\eta] - J[y]}{\varepsilon} = \frac{d}{d\varepsilon} J[y + \varepsilon\eta]\bigg|_{\varepsilon=0} \tag{3.20}$$

Example 3.1

Consider the functional

$$J[y] = \int_0^1 \left(x^2 - y^2 + y'^2\right) dx$$

with $y(0) = 0$ and $y(1) = 1$. Calculate ΔJ and $\delta J[y;\eta]$ when $y(x) = x$ and $\eta(x) = x^2$.

We first evaluate $J[y]$,

$$\begin{aligned} J[y] &= \int_0^1 \left(x^2 - y^2 + y'^2\right) dx \\ &= \int_0^1 \left(x^2 - x^2 + 1\right) dx = \int_0^1 dx = 1 \end{aligned}$$

The family of curves $y + \varepsilon\eta$ is given by $x + \varepsilon x^2$. We next evaluate J on the family $y + \varepsilon\eta$ to get

$$\begin{aligned} J[y + \varepsilon\eta] &= \int_0^1 \left[x^2 - (y + \varepsilon\eta)^2 + (y' + \varepsilon\eta')^2\right] dx \\ &= \int_0^1 \left[x^2 - (x + \varepsilon x^2)^2 + (1 + 2\varepsilon x)^2\right] dx \\ &= 1 + \frac{3}{2}\varepsilon + \frac{17}{15}\varepsilon^2 \end{aligned}$$

Hence, the change in the functional

$$\Delta J = J[y + \varepsilon\eta] - J[y] = \frac{3}{2}\varepsilon + \frac{17}{15}\varepsilon^2$$

The derivative of the functional

$$\frac{d}{d\varepsilon} J[y + \varepsilon\eta] = \frac{3}{2} + \frac{34}{15}\varepsilon$$

Evaluating this derivative at $\varepsilon = 0$ gives the Gâteaux derivative

$$\frac{d}{d\varepsilon} J[y + \varepsilon\eta]\bigg|_{\varepsilon=0} = \frac{3}{2}$$

Hence we conclude that variation $\delta J = 1.5$ in the direction $\eta(x) = x^2$.

3.4 The Fundamental Problem

A fundamental problem of the calculus of variations can be stated as follows: Given a functional J and a well-defined set of function A, determine which function in A afford a minimum (or maximum) value to J. The word minimum can be interpreted as a local minimum or an absolute minimum – a minimum relative to all elements in A. The well-defined set A is called the set of *admissible functions*. It is those functions that are the competing functions for extremizing J. For example, the set of admissible functions might be the set of all continuous functions on an interval $[a,b]$, the set of all continuously differentiable functions on $[a,b]$ satisfying the conditions such as $f(a) = 0$.

Classical calculus of variations restricts itself to functionals that are defined by certain integrals and to the determination of both necessary and sufficient conditions for extrema. The problem of extremizing a functional J over the set A is called a variational problem. To a certain degree the calculus of variations could be termed as the calculus of functionals. In the present discussion we restrict ourselves to an analysis of necessary conditions for extrema. An elementary treatment of sufficient conditions can be found in Gelfand and Fomin.

Let us concentrate on the simplest class of variational problems, in which the unknown is a continuously differentiable scalar function, and the functional to be minimized depends upon at most its second derivative. As already mentioned, the basic minimization problem, then, is to determine a suitable function $y = y(x)$ that minimizes the objective functional

$$J[y] = \int_a^b F(x, y, y' y'')\, dx, \qquad y \in A \tag{3.21}$$

where $F(x, y, y', y'')$ is some given function and A is a admissible class of functions. The integrand F is known as the Lagrangian for the variational problem. We assume that the Lagrangian is continuously differentiable in each of its four arguments x, y, y', and y''.

Very often, we encounter variational problems in which the integrand F takes the simple form $F(x, y, y')$ and hence have the functional in the form

$$J[y] = \int_a^b F(x, y, y')\, dx, \qquad y \in A \tag{3.22}$$

3.5 Maxima and Minima

One of the central problems in the calculus is to maximize or minimize a given real valued function of a single variable. If f is a given function defined in an open interval (a,b), then f

has a local minimum at a point $x = x_0$ in (a,b) if $f(x_0) < f(x)$ for all x near x_0 on both sides of $x = x_0$. In other words, f has a local minimum at a point $x = x_0$ in (a,b) if $f(x_0) < f(x)$ for all x, satisfying $|x - x_0| < \delta$ for some δ. If f has a local minimum at x_0 in (a,b) and f is differentiable in (a,b), then it is well known that

$$f'(x_0) = 0 \tag{3.23a}$$

Similar statements can be made if f has a local maximum at x_0. The aforementioned condition (3.23a) is called a necessary condition for a local minimum; that is, if f has a local minimum at x_0, then (3.23a) necessarily follows. Equation (3.23a) is not sufficient for a local minimum, however; that is, if (3.23a) holds, it does not guarantee that x_0 provides an actual minimum. The following conditions are sufficient conditions for f to have a local minimum at x_0

$$f'(x_0) = 0 \qquad \text{and} \qquad f''(x_0) > 0 \tag{3.23b}$$

provided f'' exists. Again, similar conditions can be formulated for local maxima. If (3.23b) holds, we say f is stationary at x_0 and that x_0 is an extreme point for f.

3.5.1 Maxima and minima of functionals

Instead of extremizing functions in calculus, the calculus of variations deals with extremizing functionals. The necessary condition for the functional $J[y]$ to have an extremum at $y(x) = \hat{y}(x)$ is that its variation vanishes for $y = \hat{y}$. That is,

$$\delta J[\hat{y}; \eta] = 0 \tag{3.24}$$

for $y = \hat{y}$ and for all admissible variations η.

The fact that the condition (3.24) holds for all admissible variations η often allows us to eliminate η from the condition and obtain an equation just in terms of $\hat{y}$, which can then be solved for $\hat{y}$. Generally the equation for $\hat{y}$ is a differential equation. Since (3.24) is a necessary condition we are not guaranteed that solutions $\hat{y}$ actually will provide a minimum. Therefore the solutions $\hat{y}$ to (3.24) are called (local) extremals or stationary functions, and are the candidates for maxima and minima. If $\delta J[\hat{y}; \eta] = 0$, we say J is stationary at $\hat{y}$ in the direction η.

Based on the variations δy and $\delta y'$, we distinguish between the following cases, i.e., strong extremum and weak extremum. Strong extremum occurs when δy is small, however, $\delta y'$ is large, while weak extremum occurs when both δy and $\delta y'$ are small.

Example 3.2

Consider the functional

$$J[y] = \int_0^1 \left(1 + y'(x)^2\right) dx$$

with $y(0) = 0$ and $y(1) = 1$. Let $\hat{y}(x) = x$ and $\eta(x) = x(1-x)$. The family of curves $\hat{y} + \varepsilon\eta$ is given by $x + \varepsilon x(1-x)$ and a few members are sketched in figure 3.3. We evaluate J on the family $\hat{y} + \varepsilon\eta$ to get

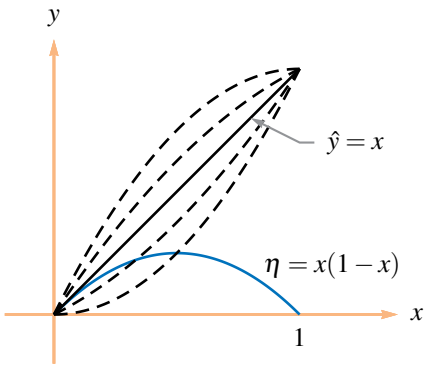

Figure 3.3: The one parameter family of curves $(x + \varepsilon x(1-x))$.

$$\begin{aligned} J[\hat{y} + \varepsilon\eta] &= \int_0^1 \left[1 + \left(\hat{y}'(x) + \varepsilon\eta'(x)\right)^2\right] dx \\ &= \int_0^1 \left[1 + (1 + \varepsilon(1-2x))^2\right] dx \\ &= 2 + \frac{\varepsilon^2}{3} \end{aligned}$$

Then the derivative of the functional

$$\frac{d}{d\varepsilon} J[\hat{y} + \varepsilon\eta] = \frac{2\varepsilon}{3}$$

Evaluating this derivative at $\varepsilon = 0$ gives the Gâteaux derivative

$$\delta J[\hat{y}; \eta] = \frac{d}{d\varepsilon} J[y + \varepsilon\eta]\bigg|_{\varepsilon=0} = 0$$

Hence we conclude that variation $\delta J[y; \eta] = 0$ and J is stationary at $\hat{y} = x$ in the direction $\eta = x(1-x)$.

Example 3.3

Consider the functional

$$J[y] = \int_0^{2\pi} \left(1 + y'(x)^2\right) dx$$

with $y(0) = 0$ and $y(2\pi) = 2\pi$. Let $\hat{y}(x) = x$ and $\eta(x) = \sin x$. The family of curves $\hat{y} + \varepsilon\eta$ is given by $x + \varepsilon \sin x$ and a few members are sketched in figure 3.2. We evaluate J on the family $\hat{y} + \varepsilon\eta$ to get

$$\begin{aligned} J[\hat{y} + \varepsilon\eta] &= \int_0^{2\pi} \left[1 + \left(\hat{y}'(x) + \varepsilon\eta'(x)\right)^2\right] dx \\ &= \int_0^{2\pi} \left[1 + (1 + \varepsilon\cos x)^2\right] dx \\ &= \pi(4 + \varepsilon^2) \end{aligned}$$

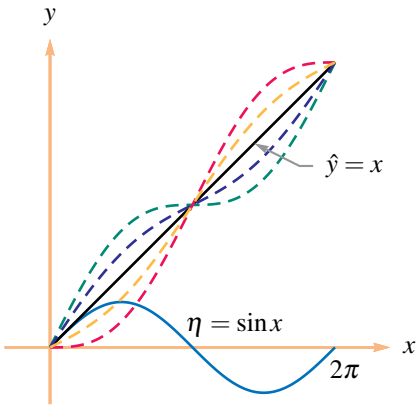

Figure 3.4: The one parameter family of curves $(x + \varepsilon \sin x)$.

Then the derivative of the functional

$$\frac{d}{d\varepsilon} J[y + \varepsilon\eta] = 2\pi\varepsilon$$

Evaluating this derivative at $\varepsilon = 0$ gives the Gâteaux derivative

$$\frac{d}{d\varepsilon} J[y + \varepsilon\eta]\Big|_{\varepsilon=0} = 0$$

Hence we conclude that variation $\delta J[y;\eta] = 0$ and J is stationary at $\hat{y} = x$ in the direction $\eta = \sin x$.

3.6 The Simplest Problem

The simplest problem of calculus of variations is to determine a function $y(x)$ for which the value of the following functional

$$J[y] = \int_a^b F(x, y, y')\, dx \tag{3.25}$$

is a minimum. Here $y \in C^2[a,b]$.[1] and F is a given function that is twice continuously differentiable on $[a,b] \times R^2$. In order to uniquely specify a minimizing function, we must impose suitable boundary conditions. Any type of boundary conditions including, Dirichlet (essential) and Neumann (natural) boundary conditions may be prescribed. In the interests of brevity, we shall impose the Dirichlet boundary conditions of the form

$$y(a) = \alpha, \qquad y(b) = \beta$$

That is, the graphs of the admissible functions pass through the end points (a,α) and (b,β).

We seek a necessary condition for the functional $J[y]$ to be a minimum. For this, we need to compute the Gâteaux variation of δJ. Let $y(x)$ be a local minimum and $\eta(x)$ a twice continuously

[1] $C^2[a,b]$ is the set of all continuous functions on an interval $[a,b]$ whose second derivative is also continuous. If $y \in C^2[a,b]$, we say y is a function of class C^2 on $[a,b]$.

differentiable function satisfying $\eta(a) = \eta(b) = 0$. Then, $Y = y + \varepsilon\eta$ is an admissible function and the new functional becomes

$$J[Y] = \int_a^b F[x, Y, Y']\,dx = \int_a^b F[x, y+\varepsilon\eta, y'+\varepsilon\eta']\,dx \tag{3.26}$$

Its derivative with respect to the parameter ε is

$$\begin{aligned}\frac{d}{d\varepsilon}J[Y] &= \int_a^b \frac{\partial}{\partial\varepsilon}F[x, Y, Y']\,dx \\ &= \int_a^b \left(\frac{\partial F}{\partial Y}\frac{\partial Y}{\partial\varepsilon} + \frac{\partial F}{\partial Y'}\frac{\partial Y'}{\partial\varepsilon}\right)dx = \int_a^b \left(\frac{\partial F}{\partial Y}\eta + \frac{\partial F}{\partial Y'}\eta'\right)dx\end{aligned}$$

Evaluating the above integral at $\varepsilon = 0$, we obtain

$$\left.\frac{d}{d\varepsilon}J[y+\varepsilon\eta]\right|_{\varepsilon=0} = \int_a^b \left(\frac{\partial F}{\partial y}\eta + \frac{\partial F}{\partial y'}\eta'\right)dx \tag{3.27}$$

As we have seen earlier, the necessary condition for the functional $J[y]$ to have an extremum at y is that its variation vanishes for y. That is,

$$\delta J[y;\eta] = \left.\frac{d}{d\varepsilon}J[y+\varepsilon\eta]\right|_{\varepsilon=0} = 0 \tag{3.28}$$

Therefore, from (3.27) the necessary condition for the functional $J[y]$ to have an extremum at y is given by

$$\int_a^b \left(\frac{\partial F}{\partial y}\eta + \frac{\partial F}{\partial y'}\eta'\right)dx = 0 \tag{3.29}$$

for all $\eta \in C^2[a,b]$ with $\eta(a) = \eta(b) = 0$.

An alternate approach for the derivation of equation (3.29)

Since first variation and Gâteaux variation are linearly related through the parameter ε, the Gâteaux variation in equation (3.28) may be replaced by the first variation. Thus the necessary condition given by equation (3.28) becomes

$$\delta J = \delta\int_a^b F(x, y, y')\,dx = \int_a^b \delta F\,dx = 0$$

Hence, using equation (3.16), we can write

$$\begin{aligned}\int_a^b \delta F\,dx &= \int_a^b \left(\frac{\partial F}{\partial y}\delta y + \frac{\partial F}{\partial y'}\delta y'\right)dx \\ &= \int_a^b \left(\frac{\partial F}{\partial y}\eta\varepsilon + \frac{\partial F}{\partial y'}\eta'\varepsilon\right)dx = 0\end{aligned}$$

Dividing this by ε, we have

$$\int_a^b \left(\frac{\partial F}{\partial y}\eta + \frac{\partial F}{\partial y'}\eta'\right)dx = 0$$

which is same as (3.29).

Condition (3.29) is not useful as it stands for determining $y(x)$. Using the fact that it must hold for all η, however, we can thus eliminate η and η' and thereby obtain a condition for y alone. First we integrate the second term in (3.29) by parts[2] to obtain

$$\int_a^b \underbrace{\frac{\partial F}{\partial y'}}_{u} \underbrace{\eta'}_{v'} dx = \left[\frac{\partial F}{\partial y'}\eta\right]_a^b - \int_a^b \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right)\eta\, dx$$

Thus, condition (3.29) can be written as

$$\int_a^b \left[\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right)\right]\eta\, dx + \left[\frac{\partial F}{\partial y'}\eta\right]_a^b = 0 \tag{3.30}$$

Since, $\eta(a) = \eta(b) = 0$, the last term on right-hand side vanishes and thus the condition (3.30) becomes

$$\int_a^b \left[\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right)\right]\eta\, dx = 0 \tag{3.31}$$

The above equation must hold for any arbitrary limits. This is possible only if the integrand is identically zero (Dubois–Reymond lemma). Therefore, we have

$$\left[\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right)\right]\eta = 0$$

Since $\eta(x)$ is an arbitrary admissible function, equation (3.31) holds good only if

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) = 0$$

We will state this result in the form of a theorem.

Theorem: If a function y provides a local minimum to the functional

$$J[y] = \int_a^b F(x, y, y')\, dx$$

where $y \in C^2[a,b]$ and

$$y(a) = \alpha, \qquad y(b) = \beta$$

then y must satisfy the equation

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) = 0, \qquad x \in [a,b] \tag{3.32a}$$

Equation (3.32a) is called the *Euler–Lagrange equation* or simply *Euler equation*. There are two important aspects of the derivation of the Euler–Lagrange equation that deserve close inspection. First, it provides a necessary condition for a local minimum but not a sufficient one. It is analogous to the derivative condition $f'(x) = 0$ in differential calculus. Therefore its

[2] $\int uv'\,dx = uv - \int u'v\,dx$

solutions are not necessarily local minima. It is a second-order ordinary differential equation with a solution that is required to satisfy two conditions at the boundaries of the domain of solution. Such boundary value problems may have no solution, one unique solution, or multiple solutions depending on the situation. A case with multiple solutions will imply that more than one paths from point (a,α) to point (b,β) satisfy the Euler–Lagrange equation. However, not all of these paths will necessarily minimize the functional $J[y]$. A second important aspect of the Euler–Lagrange equation is related to our assumption that the curve $y(x) \in C^2[a,b]$. Indeed, our considerations focused only on such smooth functions. However, the actual path that extremizes an integral might be one with a corner or a kink. Such paths are not relevant for the use of the Euler–Lagrange equation in Newtonian mechanics. However, they are often the true solutions in other problems in the calculus of variations, as we have seen in the case of physics of soap films.

It may be worthwhile to note that if y is treated as independent variable and x is dependent variable, then the Euler–Lagrange equation (3.32a) will takes the form

$$\frac{\partial F}{\partial x} - \frac{d}{dy}\left(\frac{\partial F}{\partial x'}\right) = 0, \qquad y \in [\alpha,\beta] \tag{3.32b}$$

3.6.1 Essential and natural boundary conditions

In the derivation of the Euler–Lagrange equation, we used the conditions that $\eta(a) = \eta(b) = 0$, which means that the variations $\delta y(a) = \delta y(b) = 0$. These conditions are a consequence of our imposition of fixed values of $y(x)$ at the endpoints a and b. That is

$$y(a) = \alpha, \qquad y(b) = \beta$$

where α and β are constants. This is called the *essential* (or *Dirichlet*) boundary condition. In some applications, we may need to apply other types of boundary conditions to the function $y(x)$.

If we still want the last term in equation (3.30) to vanish (so that we obtain the familiar Euler–Lagrange equation), but allowing $\delta y(a)$ and $\delta y(b)$ to be non-zero, then we need to have,

$$\left.\frac{\partial F}{\partial y'}\right|_{x=a} = 0, \qquad \left.\frac{\partial F}{\partial y'}\right|_{x=b} = 0$$

This is called a *natural* (or *Neumann*) boundary condition. A system may also have a natural boundary condition at one end $(x = a)$ and an essential boundary condition at the other end $(x = b)$.

3.6.2 Other forms of Euler–Lagrange equation

The functional F in the Euler–Lagrange equation is a function of x, y, and y'. Therefore,

$$\begin{aligned} \frac{dF}{dx} &= \frac{\partial F}{\partial x} + \frac{\partial F}{\partial y}\frac{dy}{dx} + \frac{\partial F}{\partial y'}\frac{dy'}{dx} \\ \frac{dF}{dx} &= \frac{\partial F}{\partial x} + y'\frac{\partial F}{\partial y} + y''\frac{\partial F}{\partial y'} \end{aligned} \tag{3.33}$$

But, we have

$$\frac{d}{dx}\left(y'\frac{\partial F}{\partial y'}\right) = y''\frac{\partial F}{\partial y'} + y'\frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) \tag{3.34}$$

Subtracting (3.34) from (3.33), we have

$$\frac{dF}{dx} - \frac{d}{dx}\left(y'\frac{\partial F}{\partial y'}\right) = \frac{\partial F}{\partial x} + y'\frac{\partial F}{\partial y} - y'\frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right)$$

Rewriting the above equation to give

$$\frac{d}{dx}\left[F - y'\frac{\partial F}{\partial y'}\right] - \frac{\partial F}{\partial x} = y'\left[\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right)\right]$$

By the Euler–Lagrange equation (3.32a) we see that the right-hand side of the above equation is zero. Thus,

$$\frac{d}{dx}\left[F - y'\frac{\partial F}{\partial y'}\right] - \frac{\partial F}{\partial x} = 0 \tag{3.35}$$

Equation (3.35) is another useful form of the Euler–Lagrange equation.

3.6.3 Special cases

Case I. Often in applications, the functional F does not depend directly on x and the Euler–Lagrange equation, in this case, takes a particularly nice form. Here we have $\partial F/\partial x = 0$ and the corresponding form of Euler–Lagrange equation (3.35) becomes

$$\frac{d}{dx}\left[F - y'\frac{\partial F}{\partial y'}\right] = 0$$

Integrating, we get the first integral of Euler–Lagrange equation

$$F - y'\frac{\partial F}{\partial y'} = C \tag{3.36}$$

Thus, the extremizing function y is obtained as the solution of a first-order differential equation (3.36) involving y and y' only. This simplified form of Euler–Lagrange equation (3.36) is known as the *Beltrami identity*. The combination $F - y'F_{y'}$ that appears on the left of the Beltrami identity is sometimes referred to as Hamiltonian.

Case II. If F is independent of y, then $\partial F/\partial y = 0$ and the form of Euler–Lagrange equation (3.32a) becomes

$$\frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) = 0$$

Integrating, we get the first integral of the Euler–Lagrange equation as,

$$\frac{\partial F}{\partial y'} = k \tag{3.37}$$

where k is a constant. Note that equation (3.37) is a first order differential equation involving x and y'.

Case III. If F is independent of y', then $\partial F/\partial y' = 0$ and the form of Euler–Lagrange equation (3.32a) becomes

$$\frac{\partial F}{\partial y} = 0$$

integrating, we get $F = F(x)$, a function of x alone.

3.7 Advanced Variational Problems

3.7.1 Variational problems with high-order derivatives

Here we will consider the problem of finding the function $y(x)$ that extremizes the integral

$$J[y] = \int_a^b F(x, y, y' y'')\, dx \tag{3.38}$$

with prescribed Dirichlet (essential) boundary conditions

$$\begin{aligned} y(a) &= \alpha, \qquad & y'(a) &= \alpha' \\ y(b) &= \beta, \qquad & y'(b) &= \beta' \end{aligned}$$

Here $y \in C^4[a,b]$ and F is a given function that is twice continuously differentiable on $[a,b] \times R^2$.

The necessary condition for the functional $J[y]$ to be a minimum is that the function $y(x)$ satisfies the following Euler–Lagrange equation

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) + \frac{d^2}{dx^2}\left(\frac{\partial F}{\partial y''}\right) = 0 \tag{3.39}$$

Instead of the Dirichlet-type boundary conditions we may also prescribe a Neumann -type (natural) boundary conditions of the form

$$\begin{aligned} \frac{\partial F}{\partial y'} - \frac{d}{dx}\left(\frac{\partial F}{\partial y''}\right)\bigg|_{x=a} &= 0, \qquad & \frac{\partial F}{\partial y''}\bigg|_{x=a} &= 0 \\ \frac{\partial F}{\partial y'} - \frac{d}{dx}\left(\frac{\partial F}{\partial y''}\right)\bigg|_{x=b} &= 0, \qquad & \frac{\partial F}{\partial y''}\bigg|_{x=b} &= 0 \end{aligned}$$

In general, when the functional contains higher derivatives of $y(x)$, which extremizes the functional

$$J[y] = \int_a^b F(x, y, y' y'', \cdots, y^{(n)})\, dx \tag{3.40}$$

must be a solution of the equation

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) + \frac{d^2}{dx^2}\left(\frac{\partial F}{\partial y''}\right) - \cdots\cdots + (-1)^n \frac{d^n}{dx^n}\left(\frac{\partial F}{\partial y^{(n)}}\right) = 0 \tag{3.41}$$

Equation (3.41) is differential equation of order $2n$ and is called *Euler–Poisson equation*. The general solution of this contains $2n$ arbitrary constants, which may be determined from the $2n$ boundary conditions.

3.7.2 Variational problems with several independent variables

If the extremal function u is a function on two independent variables x & y and the functional to be extremized is of the form

$$J[u] = \iint_R F(x, y, u, u_x, u_y)\, dx\, dy \tag{3.42}$$

then the $u(x,y)$ must be a solution of the equation

$$\frac{\partial F}{\partial u} - \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_x}\right) - \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_y}\right) = 0 \tag{3.43}$$

This second-order partial differential equation that must be satisfied by the extremizing function $u(x,y)$ is called the *Ostrogradsky equation* after the Russian mathematician M. Ostrogradsky.

3.8 Application of EL Equation: Minimal Path Problems

This section deals with few classical problems to illustrate the methodology to solve the variational problems with Euler-Lagrange equation. Problems of determining shortest distances furnish a useful introduction to the theory of the calculus of variations because the properties characterizing their solutions are familiar ones which illustrate many of the general principles common to all of the problems suggested above.

3.8.1 Shortest distance

Let us begin with the simplest case of all, the problem of determining the shortest distance joining two given points. Let $P(x_1, y_1)$ and $Q(x_2, y_2)$ be two fixed points in a space. Then we want to find the shortest distance between these two points. The length of the curve using the arc-length expression is

$$L = J[y(x)] = \int_P^Q ds = \int_{x_1}^{x_2} \sqrt{1 + y'(x)^2}\, dx$$

The variational problem is to find the plane curve whose length is shortest i.e., to determine the function $y(x)$ which minimizes the functional $J[y]$. The curve $y(x)$ which minimizes the functional $J[y]$ is be determined by solving the Euler–Lagrange equation (3.32a)

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right) = 0$$

In the present problem

$$F = \sqrt{1 + y'(x)^2}$$

and is a special case in which F independent of x and y. Then according to (3.17) EL equation reduces to

$$\frac{\partial F}{\partial y'} = k$$

where k is a constant. The derivative

$$\frac{\partial F}{\partial y'} = \frac{1}{2}\frac{2y'}{\sqrt{1+y'(x)^2}} = k$$

Therefore,

$$y' = k\sqrt{1+y'^2}$$

Solving for y' to obtain

$$y' = \sqrt{\frac{k^2}{1-k^2}} = m$$

Integrating, $y = mx + c$, where constants m and c are to be found using the boundary conditions $y(x_1) = y_1$ and $y(x_2) = y_2$. Thus, the straight line joining the two points $P(x_1, y_1)$ and $Q(x_2, y_2)$,

$$y = \frac{y_2 - y_1}{x_2 - x_1}x + \frac{x_2 y_1 - x_1 y_2}{x_2 - x_1}$$

is the curve with shortest length.

3.8.2 The brachistochrone problem

Let $P(x_1, y_1)$ and $Q(x_2, y_2)$ be two points on a vertical plane. Consider a curved path connecting these points. We allow a particle, without friction, to slide down this path under the influence of gravity. The question here is what is the shape of curve that allows the particle to complete the journey in the shortest possible time. Clearly, the shortest path from point P to point Q is the straight line that connects the two points. However, along the straight line, the acceleration is constant and not necessarily optimal. Naive guesses for the paths's optimal shape, including a straight line, a circular arc, a parabola, or a catenary are wrong.

In order to calculate the optimal curve we set up a two-dimensional Cartesian coordinate system on the vertical plane that contains the two points P and Q as shown in figure 3.5. Our goal is to find the path that minimizes the time it takes for an object to move from point P to point Q.

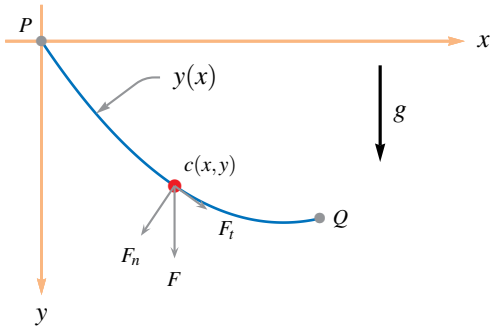

Figure 3.5: A particle sliding down a curved path.

From figure 3.5 we see that at any point $c(x,y)$ on curve $y(x)$, the gravitational force vector F decomposes into a component F_t tangent and F_n normal to curve at P. The component F_n does nothing to move the particle along the path, only the component F_t has any effect. The vector $\overline{F}$ is a constant at each point on the curve of ($\overline{F} = m\overline{g}$, where m is the mass of the particle and $\overline{g}$ is the gravitational acceleration), but F_n and F_t depend on the steepness of the curve at c. The steeper the curve, the larger F_t is, and the faster the particle moves. So it would be better if the path close to point P is more steeper so that the velocity of the object increases rapidly and then flattens towards point c. Definitely this sort of curve is longer than the straight line connecting the end points. But the extra speed that the particle develops just as it is released will more than make up for the extra distance that it must travel, and it will arrive at Q in less time than it takes along a straight line. The curve along which the particle takes the least time to go from P to Q is called the *Brachistochrone* (from the Greek words for shortest time). This famous problem, known as the Brachistochrone Problem, was posed by Johann Bernoulli (1667-1748) in 1696. The problem was solved by Johann Bernoulli, his older brother Jakob Bernoulli, Newton, and L'Hospital.

Let us begin our own study of the problem by deriving a formula relating the choice of the curve y to the time required for a particle to fall from P to Q. The instantaneous velocity of the ball along the curve is $v = \frac{ds}{dt}$, where s denotes the arc-length. Therefore,

$$dt = \frac{ds}{v} = \frac{\sqrt{dx^2+dy^2}}{v} = \frac{1}{v}\sqrt{1+y'(x)^2}\,dx \tag{3.44}$$

Let τ be the time of descent from A to B along the curve $y = y(x)$. Then,

$$\tau = \int_0^\tau dt = \int_0^S \frac{ds}{v} \tag{3.45}$$

where S is the total arc-length of the curve. If the origin of the coordinate system is taken as the staring point A, we have, using (3.44)

$$\tau = \int_0^{x_2} \frac{\sqrt{1+y'(x)^2}}{v}\,dx \tag{3.46}$$

To obtain an expression for v we use the fact that energy is conserved through the motion. Thus, the total energy at any time t must be the same as the total energy at time zero (corresponding to location P), which we may take to be zero; that is

$$\frac{1}{2}mv^2 + mg(-y) = 0$$

Solving for v gives $v = \sqrt{2gy}$. Therefore the time required for the particle to descend is

$$\tau[y] = \frac{1}{\sqrt{2g}}\int_0^{x_2} \sqrt{\frac{1+y'(x)^2}{y(x)}}\,dx \tag{3.47}$$

where we have explicitly noted that τ depends on the curve $y(x)$. Equation (3.47) defines a functional.

The Brachistochrone problem can be stated as: find the function $y(x)$ that minimizes the functional

$$\tau = J[y] = \frac{1}{\sqrt{2g}} \int_0^{x_2} \sqrt{\frac{1+y'(x)^2}{y(x)}}\, dx \tag{3.48}$$

subject to the conditions $y(0) = 0$ and $y(x_2) = y_2 > 0$. We could experiment with formula (3.48) to determine the the shortest time. Clearly it would be tedious to choose $y(x)$ one after another and look for the shortest time.

First of all we note that

$$F = \sqrt{\frac{1+y'^2}{y}}$$

which is independent of x and therefore we can apply the Beltrami identity (3.36)

$$F - y' \frac{\partial F}{\partial y'} = B$$

where B is a constant. Now

$$\frac{\partial F}{\partial y'} = \frac{1}{\sqrt{y}} \cdot \frac{1}{2\sqrt{1+y'^2}} \cdot 2y'$$

Therefore the Beltrami identity becomes

$$\frac{\sqrt{1+y'^2}}{\sqrt{y}} - \frac{y'^2}{\sqrt{y}\sqrt{1+y'^2}} = B$$

Creating a common denominator on the left-hand side produces

$$\frac{\sqrt{1+y'^2}\sqrt{1+y'^2} - y'^2}{\sqrt{y}\sqrt{1+y'^2}} = B$$

The above equation simplifies to

$$y\left(1+y'^2\right) = C$$

where C is another constant.

$$y\left[1+\left(\frac{dy}{dx}\right)^2\right] = C$$

That is, the solution to the brachistochrone problem is the solution $y = y(x)$ of the above ordinary differential equation. To solve this differential equation, we first rewrite it in the following form:

$$y = \frac{C}{1+y'^2}$$

Substitute $y' = \cot\theta$ (where θ is a parameter) in the differential equation to obtain

$$y = \frac{C}{1+\cot^2\theta} = C\sin^2\theta = \frac{C}{2}(1-\cos 2\theta)$$

Now the dx can be expressed as follows

$$dx = \frac{dy}{y'} = \frac{\frac{C}{2}(2\sin 2\theta)\,d\theta}{\cot\theta} = \frac{C2\sin\theta\cos\theta\,d\theta}{\cot\theta} = 2C\sin^2\theta d\theta$$
$$dx = C(1-\cos 2\theta)d\theta$$

Integrating the above differential equation to obtain

$$x = C\left(\theta - \frac{\sin 2\theta}{2}\right) + D$$

where the constant of integration D can be determined from the condition $y(0) = 0$, we get $D = 0$. Putting $2\theta = \phi$, we can write

$$x = \frac{C}{2}(\phi - \sin\phi) \qquad \text{and} \qquad y = \frac{C}{2}(1-\cos\phi)$$

This is the parametric equation for cycloid. A cycloid is the locus of a point fixed on the circumference of a circle as the circle rolls on a flat horizontal surface, see figure 3.6. We can show that there is one and only one cycloid passing through points P and Q. The parametric equation of cycloid may be written the following standard form:

$$\begin{aligned} x(\phi) &= a(\phi - \sin\phi) \\ y(\phi) &= a(1-\cos\phi) \end{aligned} \tag{3.49}$$

where $a = C/2$ is the radius of the rolling circle and ϕ is the angle of rotation. Using the condition that the curve (cycloid) passes through $Q(x_2, y_2)$, the value of the constant a can be determined.

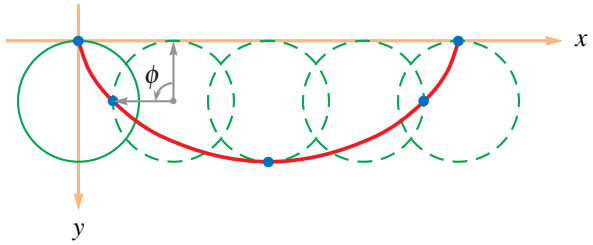

Figure 3.6: The cycloid acts as a brachistochrone.

Another remarkable characteristic of the brachistochrone particle is that when two particles *at rest* are simultaneously released from two different points M and N of the curve they will reach the terminal point of the curve at the same time, if the terminal point is the lowest point on the path (see figure 3.7). Such a curve is called an *isochrone* or a *tautochrone*. This is also counterintuitive, since clearly they have different geometric distances to cover; however, since they are acting under the gravity and the slope of the curve is different at the two locations, the particle starting from a higher location gathers much bigger speed than the particle starting at a lower location. Hence the brachistochrone problem may also be posed with a specified terminal point and a variable starting point, leading to the class of variational problems with open boundary.

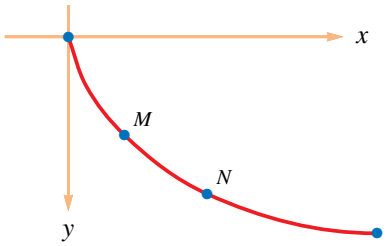

Figure 3.7: The tautochrone

3.8.3 Deflection of beam – variational formulation

Consider a simply supported beam subjected to concentrated moments at both the ends. From Euler-Bernoulli beam theory, the governing differential equation for deflection y can be derived. It is is a one-dimensional Poisson-type equation of the form

$$EI\frac{d^2y}{dx^2} - M(x) = 0 \tag{3.50}$$

with the fixed boundary conditions

$$y(0) = 0 \qquad \text{and} \qquad y(L) = 0$$

where E is the Young's modulus I is the second moment of area of the cross-section of the beam, and L is the span of the beam. The product EI, called the flexural rigidity, represents the resistance offered by the beam to deflection and $M(x)$ is the bending moment. In the present problem $M = M_0$ is a constant. Therefore,

$$EIy'' - M_0 = 0$$

This standard differential equation can be readily integrated to obtain the deflection curve. The solution is given by

$$y(x) = \frac{M_0}{2EI}x(x-L) \tag{3.51}$$

This beam deflection problem can also be solved by using the variational methods. To do this we

Figure 3.8: Simply supported beam

need to recast the problem as a variational problem using an appropriate variational statement. Here we use the principle of minimum potential energy which states that

"For conservative structural systems, of all the kinematically admissible deformations, those corresponding to the stable equilibrium state has the minimum total potential energy."

The potential energy Π in a structural system is the sum of strain energy (SE) and the work potential (WP). The potential energy of the beam under consideration is given by the following integral

$$\Pi(y) = \int_0^L \left[\frac{EI}{2} \left(\frac{dy}{dx} \right)^2 + M_0 y \right] dx \tag{3.52}$$

Here the Lagrangian F is given by

$$F = \frac{EI}{2} \left(\frac{dy}{dx} \right)^2 + M_0 y = \frac{EI}{2} y'^2 + M_0 y$$

which is independent of x. To minimize the $\Pi(y)$, we use the EL equation (3.32a)

$$\frac{\partial F}{\partial y} - \frac{d}{dx} \left(\frac{\partial F}{\partial y'} \right) = 0$$

We compute

$$\frac{\partial F}{\partial y} = M_0 \qquad \text{and} \qquad \frac{\partial F}{\partial y'} = EIy'$$

Substitute the above results in the EL equation to obtain

$$M_0 - \frac{d}{dx}(EIy') = 0$$

Separate the variable and integrating

$$EIy' = M_0 x + c_1$$

Integrating again

$$EIy = M_0 \frac{x^2}{2} + c_1 x + c_2$$

where constants c_1 and c_2 are to be found using the boundary conditions $y(0) = 0$ and $y(L) = 0$. Thus, we have

$$c_1 = -\frac{M_0 L}{2} \qquad \text{and} \qquad c_2 = 0$$

Substitution of these values in the general solution gives the equation of the deflection curve

$$y = \frac{M_0}{2EI} x(x - L)$$

which is same as the solution of the differential equation.

3.9 Construction of Functionals from PDEs

We have noticed that EL equation produces the governing differential equation corresponding to a given functional or variational principle. Here we seek the inverse procedure of constructing a variational principle for a given differential equation, $\mathscr{L}(y) = 0$. The procedure for finding the functional associated with the differential equation involves four basic steps:

- Multiply the left-hand side of the differential equation $\mathscr{L}(y)$ with the variational δy of the dependent variable y and integrate over the domain of the problem.
- Use integration by parts to transfer the derivatives to variation δy.
- Express the boundary integrals in terms of the specified boundary conditions.
- Bring the variational operator δ outside the integrals.

The procedure is best illustrated with an example. We will take the problem of the deflection of beam governed by the equation (3.50). Since the differential equation holds good for all points within the system, we can write

$$\left(EI\frac{d^2y}{dx^2} - M_0 \right) \delta y = 0$$

where δy is an arbitrary variation on y with $\delta y|_{x=0} = 0$. Integrating over the domain of the problem,

$$\delta J = \int_0^L \left(EI\frac{d^2y}{dx^2} - M_0 \right) \delta y dx = 0$$

$$\delta J = \int_0^L EI\frac{d^2y}{dx^2}\,\delta y dx - \int_0^L M_0\,\delta y dx$$

Now, the first integral on the right-hand side can be integrated by parts[3] by letting $u = \delta y$ and $v' = EI\frac{d^2y}{dx^2}$. Thus

$$\delta J = \left. \delta y EI\frac{dy}{dx} \right|_0^L - \int_0^L \frac{d(\delta y)}{dx} EI\frac{dy}{dx} dx - \int_0^L M_0\,\delta y dx$$

The first term vanish if we assume either the homogeneous Dirichlet or Neumann conditions at the boundaries. That is,

$$y(0) = y(L) = 0 \qquad \Rightarrow \qquad \delta y(0) = \delta y(L) = 0$$

or

$$\left. \frac{dy}{dx} \right|_L = \left. \frac{dy}{dx} \right|_0 = 0$$

Hence

$$\delta J = \delta \int_0^L \frac{EI}{2} \left(\frac{dy}{dx} \right)^2 dx - \delta \int_0^L M_0 y dx$$

[3] $\int uv'\,dx = uv - \int u'v\,dx$

Therefore,

$$J[y] = \int_0^L \left[\frac{EI}{2} \left(\frac{dy}{dx} \right)^2 + M_0 y \right] dx$$

Some standard differential equations and their functional are given below.

If the differential equation is of the form

$$D\frac{d^2\phi}{dx^2} + P(x)\phi + Q(x) = 0, \qquad x \in [a,b] \tag{3.53a}$$

the corresponding variational principle is given by

$$J[\phi] = \frac{1}{2}\int_a^b \left[D\left(\frac{d\phi}{dx} \right)^2 - P(x)\phi^2 - 2Q(x)\phi \right] dx \tag{3.53b}$$

and if the differential equation is of the form

$$\nabla^2\phi + p^2\phi = q, \qquad x \in \mathscr{D} \tag{3.54a}$$

the corresponding variational principle is given by

$$J[\phi] = \frac{1}{2}\int_{\mathscr{D}} \left[|\nabla\phi|^2 - p^2\phi^2 + 2q\phi \right] d\mathscr{D} \tag{3.54b}$$

where

$$|\nabla\phi|^2 = \nabla\phi \cdot \nabla\phi = \left(\frac{\partial\phi}{\partial x} \right)^2 + \left(\frac{\partial\phi}{\partial y} \right)^2$$

Example 3.4

Find the functional for the ordinary differential equation

$$\frac{d^2y}{dx^2} + 3y + x = 0, \qquad 0 < x < 1$$

subject to $y(0) = y(1) = 0$.

This equation is of the form (3.53a). Therefore, the corresponding functional is given by

$$J[y] = \frac{1}{2}\int_0^1 \left[\left(\frac{dy}{dx} \right)^2 - 3y^2 - 2xy \right] dx$$

As a check we will use the EL equation (3.32a)

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\left(\frac{\partial F}{\partial y'} \right) = 0$$

for the above functional to recover the original differential equation. That is,

$$-6y - 2x - \frac{d}{dx}\left(2\frac{dy}{dx} \right) = 0 \qquad \Rightarrow \qquad \frac{d^2y}{dx^2} + 3y + x = 0$$

3.10 Rayleigh–Ritz Method

Rayleigh–Ritz method is a direct method for minimizing a given functional. It is direct in the sense that it yields a solution to the variational problem without solving the associated Euler-Lagrange Equation. It may be noted that, for most of the physical problems, the functional we get from the variational principle is not simple and thus the solution using the EL equation will be difficult to obtain. The Rayleigh–Ritz method is an approximate method where the given functional is directly minimized without recourse to the associated EL equation.

To illustrate the method let us consider the following functional

$$J[\phi] = \int_S F(x, y, \phi, \phi_x, \phi_y)\, dS \tag{3.55}$$

Our objective is to minimize this integral. In the Rayleigh-Ritz method, we select a linearly independent set of functions called basis functions u_n and construct an approximate solution to equation (3.55), satisfying some prescribed boundary conditions. The solution is in the form of a finite series

$$\tilde{\phi} = u_0 + \sum_{n=1}^{N} a_n u_n \tag{3.56}$$

where u_0 meets the nonhomogeneous boundary conditions if any, and u_n satisfies homogeneous boundary conditions. The unknown coefficients a_n are to be determined and $\tilde{\phi}$ is an approximate solution to the exact solution ϕ. Substitution of the approximate solution into equation (3.55) results in the function with N coefficients $a_1, a_2, \cdots a_N$. That is,

$$J(\tilde{\phi}) = J(a_1, a_2, \cdots a_N)$$

The minimum of this function is obtained when its partial derivatives with respect to each coefficient is zero. That is,

$$\frac{\partial J}{\partial a_1} = 0, \quad \frac{\partial J}{\partial a_2} = 0, \quad \cdots \quad \frac{\partial J}{\partial a_N} = 0$$

or

$$\frac{\partial J}{\partial a_n} = 0, \quad n = 1, 2, \cdots N \tag{3.57}$$

Thus we obtain a system of N linear algebraic equations which can be solved to obtain a_n. These a_n are then substituted into the approximate solution (3.56). Now, if $\tilde{\phi} \to \phi$ as $N \to \infty$ in some sense, then the procedure is said to converge to the exact solution.

The basis functions are selected to satisfy the prescribed boundary conditions of the problem. u_0 is chosen to satisfy the inhomogeneous boundary conditions, while $u_n (n = 1, 2, \cdots N)$ are selected to satisfy the homogeneous boundary conditions. It may be noted that $u_0 = 0$ if the prescribed boundary conditions are all homogeneous (Dirichlet conditions). The Rayleigh-Ritz method has two major limitations. First, the variational principle in equation (3.55) may not exist in some problems such as in nonself-adjoint equations (odd order derivatives). Second, it is difficult, if not impossible, to find the functions u_0 satisfying the global boundary conditions for the domains with complicated geometries.

Example 3.5

Use the Rayleigh-Ritz method to solve the beam deflection problem given by the variational principle (3.52):

$$\Pi[y] = \int_0^L \left[\frac{EI}{2} \left(\frac{dy}{dx} \right)^2 + M_0 y \right] dx$$

with the boundary conditions $y(0) = 0 = y(L)$. The exact solution of this minimization problem is

$$y(x) = \frac{M_0}{2EI} x(x-L)$$

We let the approximate solution be

$$\tilde{y} = u_0 + \sum_{n=1}^{N} a_n u_n$$

where $u_0 = 0$. Some of the possible choices for base function are polynomial of the form

$$\tilde{y} = \sum_{n=1}^{N} a_n x^n$$

and trigonometric functions of the form

$$\tilde{y} = \sum_{n=1}^{N} a_n \sin k_n \pi x$$

Trigonometric approximation. We will first explore the case of trigonometric function with $N = 1$. That is, we have

$$\tilde{y} = a \sin k\pi x$$

The assumed solution should satisfy both the boundary conditions. If we set $k = 1/L$, we have a solution which satisfies the boundary conditions. Thus, we have

$$\tilde{y} = a \sin \frac{\pi x}{L}$$

Here a is the undetermined parameter to be found out. We have to select a such that the functional $\Pi[y]$ is a minimum. Substituting the above approximate solution into the functional gives

$$\Pi(a) = \int_0^L \left[\frac{EI}{2} \left(\frac{a\pi}{L} \cos \frac{\pi x}{L} \right)^2 + M_0 a \sin \frac{\pi x}{L} \right] dx$$

Evaluating the integral to yield

$$\Pi(a) = \left(\frac{EI\pi^2}{4L} \right) a^2 + \left(\frac{2M_0 L}{\pi} \right) a$$

At this point observe that $\Pi(a)$ is an ordinary function of the unknown a. The function $\Pi(a)$ is minimum when

$$\frac{\partial \Pi}{\partial a} = 0 \qquad \rightarrow \qquad 2\left(\frac{EI\pi^2}{4L} \right) a + \frac{2M_0 L}{\pi} = 0 \qquad \text{or} \qquad a = -\frac{4M_0 L^2}{\pi^3 EI}$$

Hence the approximate solution is

$$\tilde{y} = -\frac{4M_0L^2}{\pi^3 EI}\sin\pi\frac{x}{L}$$

Figure 3.9 shows that the approximate solution $\tilde{y}(x)$ agrees well with the exact solution $y(x)$

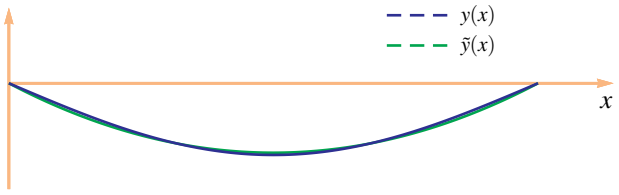

Figure 3.9: Beam deflection problem with trigonometric approximation.

over the interval $[0, L]$.

Polynomial approximation. Next, we will try with the polynomial function with $N = 2$. That is, we have

$$\tilde{y} = a_1 x + a_2 x^2$$

The assumed solution satisfy the boundary condition $y(0) = 0$. Application of the second boundary condition yields

$$0 = a_1 L + a_2 L^2 \qquad \rightarrow \qquad a_1 = -a_2 L$$

Hence the approximate solution which satisfies both the BCs is given by

$$\tilde{y} = ax(x - L)$$

where we have dropped the subscript of a. Substituting the above approximate solution into the functional gives

$$\begin{aligned}
\Pi(a) &= \int_0^L \left[\frac{EI}{2}\left[a(2x-L)\right]^2 + M_0 ax(x-L)\right] dx \\
&= \int_0^L \left[\frac{EI}{2}\left(4a^2x^2 - 4a^2Lx + a^2L^2\right)^2\right] dx + \int_0^L M_0\left[ax^2 - aLx\right] dx \\
&= \frac{EI}{2}\left[\frac{4}{3}a^2L^3 - 2a^2L^3 + a^2L^3\right] + M_0\left[\frac{aL^3}{3} - \frac{aL^3}{2}\right] \\
&= \frac{EI}{2}\left[\frac{a^2L^3}{3}\right] + M_0\left[-\frac{aL^3}{6}\right]
\end{aligned}$$

The function $\Pi(a)$ is minimum when

$$\frac{\partial \Pi}{\partial a} = 0 \qquad \rightarrow \qquad \frac{EI}{2}\frac{2aL^3}{3} - \frac{M_0L^3}{6} = 0 \qquad \text{or} \qquad a = \frac{M_0}{2EI}$$

Hence the approximate solution is

$$\tilde{y} = \frac{M_0}{2EI}x(x-L)$$

We see here that this is the exact solution of the problem. This has happened because the selected approximate solution (polynomial) represents the exact behaviour of the deflection curve.

Chapter 4

Weighted Residual Methods

4.1 Introduction

Weighted residual method is a generic class of method developed to obtain approximate solution to the differential equations of the form

$$\mathscr{L}(\phi) + f = 0 \quad \text{in} \quad D \tag{4.1}$$

where $\phi(\mathbf{x})$ is the dependent variable and is unknown and $f(\mathbf{x})$ is a known function of $\mathbf{x}$. $\mathscr{L}$ denotes the differential operator involving spatial derivative of ϕ, which specifies the actual form of the differential equation.

Weighted residual method involves two major steps. In the first step, an approximate solution based on the general behavior of the dependent variable is assumed. The assumed solution is often selected so as to satisfy the boundary conditions for ϕ. This assumed solution is then substituted in the differential equation. Since the assumed solution is only approximate, it does not in general satisfy the differential equation and hence results in an error or what we call a *residual*. The residual is then made to vanish in some average sense over the *entire* solution domain to produce a system of algebraic equations. The second step is to solve the system of equations resulting from the first step subject to the prescribed boundary condition to yield the approximate solution sought.

Let $\psi(\mathbf{x}) \approx \phi(\mathbf{x})$, is an approximate solution to the differential equation (4.1). When $\psi(\mathbf{x})$ is substituted in the differential equation (4.1), it is unlikely that the equation is satisfied. That is, we have

$$\mathscr{L}(\psi) + f \neq 0.$$

Or we may write

$$\mathscr{L}(\psi) + f = R \tag{4.2}$$

where $R(\mathbf{x})$ is a measure of error commonly referred to as the residual.

Multiply equation (4.1) by an arbitrary *weight function* $w(\mathbf{x})$ and integrating over the domain D to obtain

$$\int_D w[\mathscr{L}(\phi) + f]\, dD = 0. \tag{4.3}$$

Equations (4.1) and (4.3) are equivalent. Replacing ϕ by ψ in equation (4.3) results in

$$\int_D w(\mathbf{x}) \left[\mathscr{L}(\psi) + f\right] dD = \int_D w(\mathbf{x}) R(\mathbf{x}) \, dD \neq 0. \tag{4.4}$$

The integral in (4.4) gives the weighted average of the residual over the solution domain. In weighted residual method we force this integral (i.e., the inner product (w,R)) to vanish over the solution domain. That is,

$$(w,R) = \int_D w(\mathbf{x}) R(\mathbf{x}) \, dD = 0. \tag{4.5}$$

We now seek the approximate solution in the form a generalized Fourier series, say

$$\psi(\mathbf{x}) = \sum_{i=1}^{n} c_i N_i(\mathbf{x}) = c_1 N_1(\mathbf{x}) + c_2 N_2(\mathbf{x}) + \cdots\cdots + c_n N_n(\mathbf{x}). \tag{4.6a}$$

In vector form

$$\psi(\mathbf{x}) = \mathbf{C}^T \mathbf{N}^T = (\mathbf{N}\mathbf{C})^T = \mathbf{N}\mathbf{C} \tag{4.6b}$$

where $\mathbf{N}$ is a row vector

$$\mathbf{N} = \begin{bmatrix} N_1 & N_2 & \cdots\cdots & N_n \end{bmatrix},$$

$\mathbf{C}$ is a column vector

$$\mathbf{C} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}.$$

and $\mathbf{N}\mathbf{C}$ is a 1×1 matrix.

Here c_i's are unknown coefficients called *fitting coefficients* and n is the number of fitting coefficients. $N_i(\mathbf{x})$'s are assumed to be linearly independent functions of $\mathbf{x}$ and are called *trial functions*. The trial functions can be polynomials, trigonometric functions etc. The trial functions are usually chosen in such a way that the assumed function $\psi(\mathbf{x})$ satisfies the global boundary conditions for $\phi(\mathbf{x})$, although this not strictly necessary and certainly not always possible.

Polynomial Approximation. One of the simplest choices for a trial function is a polynomial, for a one-dimensional problem which can be obtained by taking $N_i(x) = x^i$. The result is

$$\psi(x) = \sum_{i=0}^{n} c_i x^i = c_0 + c_1 x + \cdots + c_n x^n.$$

This produces a smooth solution, but it suffers the same limitations as Lagrange interpolation. A particularly significant flaw is that this choice need not converge to $\phi(x)$ as n increases.

Trigonometric Approximation. Another often used set of trial function is trigonometric approximation based on Fourier series. An example is a Fourier sine series obtained by taking $N_k(x) = \sin\frac{k\pi x}{L}$. For a one-dimensional problem,

$$\psi(x) = \sum_{k=1}^{n} c_k \sin\frac{k\pi x}{L}.$$

Because $\sin(k\pi x/L)$ at $x=0$ and $x=L$ are zero, this expansion requires the boundary conditions $y(0)=y(L)=0$. This is not much of a restriction, because one can always make the change of variables so that the boundary conditions become homogeneous.

With the selection of $\psi(x)$ as the series expansion (4.6), it is evident that the residual R depends on the unknown parameters c_i's in the expansion:

$$R = R(\mathbf{x};\mathbf{C}).$$

If the number of trial functions n is sufficiently large, then in principle, the unknown parameters c_i's can be chosen so that the residual R is small over the domain.

Weight functions. In general the weight function $w(\mathbf{x})$ may be written as

$$w(\mathbf{x}) = \sum_{i=1}^{n} a_i w_i = a_1 w_1 + a_2 w_2 + \cdots\cdots + a_n w_n = \mathbf{aw} \tag{4.7}$$

where $\mathbf{a}$ and $\mathbf{w}$ are row and column vector given respectively by

$$\mathbf{a} = \begin{bmatrix} a_1 & a_2 & \cdots\cdots & a_n \end{bmatrix}, \qquad \mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}.$$

Here w_i's are known functions of $\mathbf{x}$ and a_i's are constant parameters. Substituting $w(x)=\mathbf{aw}$ in the weighted residual equation (4.5) to yield

$$\mathbf{a}\int_D \mathbf{w} R\, dD = 0.$$

Since $\mathbf{a}$ is a constant vector, we have

$$\int_D \mathbf{w} R\, dD = 0 \tag{4.8a}$$

or

$$\begin{aligned} \int_D w_1 R\, dD &= 0 \\ \vdots \qquad & \quad \vdots \\ \int_D w_n R\, dD &= 0 \end{aligned} \tag{4.8b}$$

Now we have n equations to determine unknown coefficients c_i's. Finally, inserting $\psi = \mathbf{NC}$ in equation (4.2) yields

$$R = \mathscr{L}(\mathbf{NC}) + f = \mathscr{L}(\mathbf{N})\,\mathbf{C} + f \tag{4.9}$$

and hence the condition (4.8a) becomes

$$\int_D \mathbf{w}\left[\mathscr{L}(\mathbf{N})\,\mathbf{C} + f\right] dD = 0$$

or

$$\left[\int_D \mathbf{w}\,\mathscr{L}(\mathbf{N})\;dD\right]\mathbf{C} = -\int_D \mathbf{w}\,f\;dD. \tag{4.10a}$$

Defining matrix $\mathbf{K}$ and column vector $\mathbf{f}$ as

$$\mathbf{K} = \int_D \mathbf{w}\,\mathscr{L}(\mathbf{N})\;dD \qquad \text{and} \qquad \mathbf{f} = -\int_D \mathbf{w}\,f\;dD$$

allows us to write equation (4.8) in compact form:

$$\mathbf{K}\mathbf{C} = \mathbf{f} \tag{4.10b}$$

which may be expanded as

$$\begin{bmatrix} \int_D w_1\mathscr{L}(N_1)\,dD & \int_D w_1\mathscr{L}(N_2)\,dD & \cdots & \int_D w_1\mathscr{L}(N_n)\,dD \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \int_D w_n\mathscr{L}(N_1)\,dD & \int_D w_n\mathscr{L}(N_2)\,dD & \cdots & \int_D w_n\mathscr{L}(N_n)\,dD \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = -\begin{bmatrix} \int_D w_1 f\,dD \\ \int_D w_2 f\,dD \\ \vdots \\ \int_D w_n f\,dD \end{bmatrix}. \tag{4.10c}$$

The system of equation given by (4.10) can be solved for n unknown coefficients c_i's provided that a suitable weight function w is selected.

With regards to the selection of weight function, we have several choices. Hence, depending upon nature of weight function, we have different types of weighted residual methods. Some of the standard methods are:

1. Point Collocation Method
2. Subdomain Collocation Method
3. Least Square Method
4. Galerkin Method

4.2 Point Collocation Method

In point collocation method, the weight functions are selected in such a way that the residual can be set equal to zero at n distinct points in the domain. This can be achieved by choosing weight function as the displaced Dirac delta function. So, for one-dimensional case,

$$w_i = \delta(x-x_i) = \begin{cases} \infty, & \text{if } x = x_i \\ 0, & \text{else} \end{cases} \tag{4.11}$$

where the fixed points $x_i \in [a,b]$, $(i = 1,2,\cdots,n)$ are called collocation points. The number of collocation points selected must be equal to the number of unknown coefficients c_i's in the definition of approximate function, $u(x)$.

The displaced Dirac delta function has the property that

$$(w_i, R) = \int_a^b \delta(x-x_i)\,R(x)\;dx = R(x_i).$$

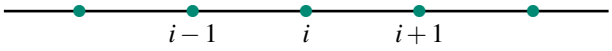

Figure 4.1: Collocation points in a one-dimensional domain.

Thus, from equation (4.8) we have

$$R(x_i) = 0, \qquad i = 1,2,\cdots,n \tag{4.12}$$

or

$$c_1\mathscr{L}(N_1(x_i)) + c_2\mathscr{L}(N_2(x_i)) + \cdots + c_n\mathscr{L}(N_n(x_i)) + f(x_i) = 0, \qquad i = 1,2,\cdots,n$$

i.e., the residual $R(x)$ is forced to be zero at n collocation points. So, for the point collocation method the linear system of equation (4.10) takes the form

$$\begin{bmatrix} \mathscr{L}(N_1(x_1)) & \mathscr{L}(N_2(x_1)) & \cdots & \mathscr{L}(N_n(x_1)) \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \mathscr{L}(N_1(x_n)) & \mathscr{L}(N_2(x_n)) & \cdots & \mathscr{L}(N_n(x_n)) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = - \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{bmatrix}. \tag{4.13}$$

Note: It can be shown that the point collocation method is equivalent to the classical finite difference method.

Example 4.1

Let us illustrate the application of point collocation method using a simple physical problem. We consider a simply supported beam subjected to concentrated moments at both ends. The problem is governed by the following differential equation

$$EI\frac{d^2y}{dx^2} - M_0 = 0, \qquad x \in [0,L] \tag{4.14}$$

with boundary conditions (support condition in this case)

$$y(0) = 0 \qquad \& \qquad y(L) = 0.$$

Here, the coefficient EI represents the resistance of the beam to deflection called *flexural rigidity* (or beam stiffness), M_0 is the applied moment, and L is the length of the beam. The analytical solution of the problem in the interval $[0,L]$ is

$$y(x) = -\frac{M_0}{2EI}x(L-x). \tag{4.15}$$

The negative sign in the expression shows that the displacement is negative for positive values of bending moment, M_0.

Trigonometric approximation to deflection curve. Let us pretend that we do not know the solution and select the approximating function $u(x)$ as a sinusoidal function of the form

$$u(x) = A\sin Bx$$

where A and B are constants. The function which satisfies the prescribed boundary conditions can be obtained by the application of boundary conditions to the chosen approximating function. Thus, we have

$$u(x) = A\sin\frac{\pi x}{L} = c_1 N_1 \tag{4.16}$$

where $c_1 = A$ and $N_1 = \sin\frac{\pi x}{L}$. The second derivative of the assumed function,

$$\frac{d^2u}{dx^2} = c_1\frac{d^2N_1}{dx^2} = -\frac{A\pi^2}{L^2}\sin\frac{\pi x}{L}$$

Substitution the above expression for the second derivative into the (4.14) gives the residual R. That is,

$$R(x;A) = -EI\frac{A\pi^2}{L^2}\sin\frac{\pi x}{L} - M_0.$$

Since the approximating function contains just one fitting coefficient, we need to select only one collocation point in the domain $[0,L]$ and force residual to zero there. We do not know which point will be the best choice, so we arbitrarily select collocation point at $x_1 = L/2$. By equation (4.12), we have

$$R(L/2) = -EI\frac{A\pi^2}{L^2}\sin\frac{\pi}{2} - M_0 = 0.$$

Solving for the unknown coefficient A, we obtain

$$A = -\frac{M_0L^2}{EI\pi^2}.$$

Thus, the approximate solution in the interval $[0,L]$ is

$$u(x) = -\frac{M_0L^2}{EI\pi^2}\sin\frac{\pi x}{L}. \tag{4.17}$$

Figure 4.3 shows that the approximate solution $u(x)$ agrees well with the exact solution $y(x)$ over the interval $[0,L]$. Note that if we had selected the collocation point other than at $x_1 = L/2$, a different approximate solution would have been obtained.

Figure 4.2: Simply supported beam subject to bending moments

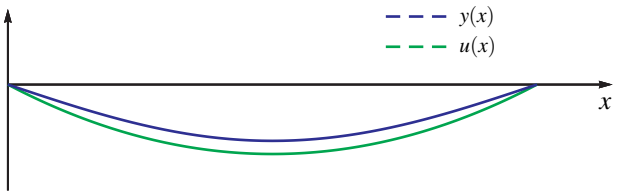

Figure 4.3: Beam deflection problem – result of point collocation method.

Polynomial approximation to deflection curve. Here we select a second degree polynomial of the form

$$u(x) = a + bx + cx^2.$$

The function which satisfies the prescribed boundary conditions can be obtained by the application of boundary conditions to the chosen approximating function. Thus, we have

$$u(x) = cx(x-L) = c_1 N_1 \tag{4.18}$$

where $c_1 = c$ and $N_1 = x(x-L)$. The second derivative,

$$\frac{d^2u}{dx^2} = 2c$$

The residual R is then given by

$$R(x;c) = EI \times 2c - M_0.$$

Here $R(x)$ is independent of x, so that the residual can be set to zero at every point in the interval automatically. Therefore,

$$EI \times 2c - M_0 = 0.$$

Solving for the unknown coefficient c, we get

$$c = \frac{M_0}{2EI}.$$

Thus, the approximate solution is

$$u(x) = -\frac{M_0}{2EI}x(L-x). \tag{4.19}$$

It may be noted that selection of a second degree polynomial yields exact solution since the selected polynomial represents the exact behaviour of the deflection curve.

Example 4.2

It is interesting to note that one-dimensional steady state heat conduction problem with uniform heat generation is similar to the beam deflection problem discussed above. The governing differential equation for the heat conduction problem is given by

$$k\frac{d^2T}{dx^2} + S = 0, \qquad x \in [0,L] \tag{4.20}$$

with the boundary conditions

$$T(0) = T(L) = 0$$

where S is the uniform the rate of heat generation per unit volume of the material with thermal constant conductivity k. The exact solution of the problem is

$$T(x) = \frac{S}{2k} x(L-x) \tag{4.21}$$

Exact solution will be obtained if second degree polynomial is selected as the trial function.

Example 4.3

We will now take a fluid mechanics problem which is governed by a second-order linear ordinary differential equation similar to that of beam deflection problem and steady state heat conduction problem discussed earlier. Consider the fully developed flow between infinite parallel plates. The plates are separated by a distance h, as shown in figure. The length of the plates in z-direction is assumed to be very large compared to h, with no variation of any fluid property in this direction. With this assumption, we have $\partial/\partial z = 0$. The flow is assumed to be steady, incompressible, and unidirectional with velocity components $v = w = 0$. Since the flow under consideration is unidirectional it satisfies the condition for parallel flows. The continuity and x-momentum equation are given by

$$\frac{\partial u}{\partial x} = 0$$

$$\rho \frac{\partial u}{\partial t} = \rho g_x - \frac{\partial p}{\partial x} + \mu \left(\frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right)$$

$$0 = \rho g_y - \frac{\partial p}{\partial y}$$

$$0 = \rho g_z - \frac{\partial p}{\partial z}$$

For steady flow in the absence of gravitational force, the system reduces to

$$0 = -\frac{\partial p}{\partial x} + \mu \frac{\partial^2 u}{\partial y^2}$$

$$0 = \frac{\partial p}{\partial y}$$

$$0 = \frac{\partial p}{\partial z}$$

From the continuity equation we can infer that the velocity u is not a function of stream-wise direction, x. In other words, the flow is same in any x-location. The phrase *fully developed flow* is often used to describe this situation. Thus, in the fully developed flow, u is function of only y; i.e., $u = u(y)$.

The y and z-momentum equations show that the pressure is independent of y and z coordinates. Thus, pressure could be a function of x alone, i.e.,

$$p = p(x)$$

The x-momentum equation can be written as

$$\frac{d^2u}{dy^2} - \frac{1}{\mu}\frac{dp}{dx} = 0 \qquad x \in [0,h] \tag{4.22}$$

Since the left-hand side varies only with y and the right-hand side varies only with x, it follows that both sides must be equal to the same constant. Hence, the pressure gradient dp/dx is a constant. This equation can be integrated twice and no-slip boundary conditions can then be applied to obtain the analytical solution

$$u(y) = -\frac{1}{2\mu}\frac{dp}{dx}y(h-y) \tag{4.23}$$

Figure shows the parabolic velocity profile. Exact solution will be obtained if second degree polynomial is selected as the trial function.

Example 4.4

Solve the differential equation

$$\frac{d^2y}{dx^2} + y = x, \qquad x \in [0,2]$$

with the boundary conditions

$$y(0) = 0, \qquad y(2) = 5$$

using point collocation method. The exact solution of the problem is

$$y(x) = \frac{3}{\sin 2}\sin x + x$$

over the interval $[0,2]$.

To solve the problem using point collocation method, we use a polynomial trial function $u(x)$ of degree 3 in the form

$$u(x) = 2.5x + c_2x(x-2) + c_3x^2(x-2) = 2.5N_1 + c_2N_2 + c_3N_3.$$

Here we have three linearly independent trial functions $N_1 = x$, $N_2 = x(x-2)$, and $N_3 = x^2(x-2)$. The boundary conditions are met by the first term, and other terms are so selected that they are equal to zero at the boundaries so that $u(x)$ also meets the boundary conditions.[1]

The residual is obtained after substituting $u(x)$ for $y(x)$ in the differential equation,

$$R(x) = \frac{d^2u}{dx^2} + u - x.$$

[1] It is customary to match the boundary conditions with the initial term(s) of $u(x)$ and then make the succeeding terms equal to zero at the boundaries.

From the $u(x)$ defined, we have

$$\frac{d^2u}{dx^2} = 2c_2 + c_3(6x-4).$$

Therefore, the residual becomes

$$R(x) = 2c_2 + c_3(6x-4) + 2.5x + c_2x(x-2) + c_3x^2(x-2) - x.$$

Since the trial function contains two unknown fitting coefficients, we can force the residual to be zero at two distinct points in $[0,2]$. We do not know which two points will be the best choices, so we arbitrarily select collocation points at $x = 0.7$ and $x = 1.3$. (Note that these points are more or less equally spaced in the interval). Setting the residual zero at these points gives a pair of equation for the constants c_2 and c_3:

$$1090c_2 - 437c_3 + 1050 = 0,$$
$$1090c_2 + 2617c_3 + 1950 = 0.$$

or in matrix form

$$\begin{pmatrix} 1.09 & -0.437 \\ 1.09 & 2.617 \end{pmatrix} \begin{bmatrix} c_2 \\ c_3 \end{bmatrix} = - \begin{bmatrix} 1.05 \\ 1.95 \end{bmatrix}.$$

Solving the above set of equations for c_2 and c_3 and substitute in the assumed trial function to obtain

$$\begin{aligned} u(x) &= \left(\frac{5}{2}\right)x - \left(\frac{60000}{55481}\right)x(x-2) - \left(\frac{900}{3054}\right)x^2(x-2) \\ &= -\left(\frac{900}{3054}\right)x^3 - \left(\frac{13895700}{28239829}\right)x^2 + \left(\frac{517405}{110962}\right)x. \end{aligned}$$

Figure 4.4 shows that the approximate solution $u(x)$ agrees well with the exact solution $y(x)$

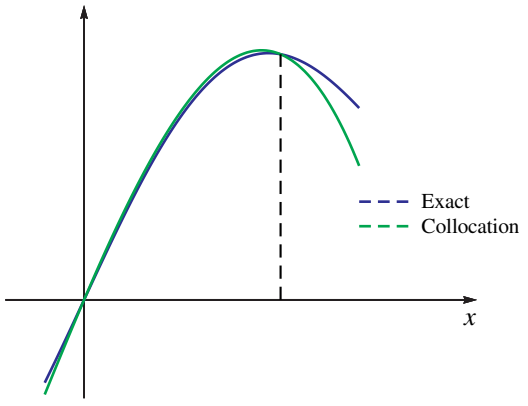

Figure 4.4: Comparison of point collocation and exact solutions of problem #4.

over the interval $[0,2]$.

It is instructive to note the following points about point collocation method:

- Point collocation method does not automatically produce symmetric coefficient matrix which is a desirable property when the solution of the equation is sought. Also, symmetry has nothing to do with the type of approximate solution ϕ selected.
- Setting the residual to zero at discrete points does not mean that the errors in those points are actually zero.
- Computational effort required in the point collocation method is minimal.

4.3 Subdomain Collocation Method

In the subdomain collocation method, we divide the physical domain into a number of non-overlapping subdomains. Number of subdomain n is taken as equal to the number of unknown coefficients in the approximating function. Now, each weight function is selected as unity over a specific subdomain and set equal to zero over other the other parts. That is, for one-dimensional problems,

$$w_i = \begin{cases} 1, & \text{if } x_i \le x \le x_{i+1} \\ 0, & \text{else} \end{cases} \qquad (i = 1,2,\cdots,n) \tag{4.24}$$

Thus, equation (4.10) may be written as

$$\int_a^b w_i R(x)\,dx = \int_{x_i}^{x_{i+1}} R(x)\,dx = 0, \quad (i = 1,2,\cdots,n). \tag{4.25}$$

This means that the average of the residual over each of n subdomains is forced to be zero. Or, in other words, differential equation is satisfied on the average in each of the n subdomains. For the subdomain collocation method the linear system of equation (4.8) takes the form

$$\begin{bmatrix} \int_{x_1}^{x_2} \mathscr{L}(N_1)dx & \int_{x_1}^{x_2} \mathscr{L}(N_2)dx & \cdots & \int_{x_1}^{x_2} \mathscr{L}(N_n)dx \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \int_{x_n}^{x_{n+1}} \mathscr{L}(N_1)dx & \int_{x_n}^{x_{n+1}} \mathscr{L}(N_2)dx & \cdots & \int_{x_n}^{x_{n+1}} \mathscr{L}(N_n)dx \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = - \begin{bmatrix} \int_{x_1}^{x_2} f\,dx \\ \int_{x_2}^{x_3} f\,dx \\ \vdots \\ \int_{x_n}^{x_{n+1}} f\,dx \end{bmatrix}. \tag{4.26}$$

Note: It can be shown that the subdomain collocation method is equivalent to the widely used *finite volume method* in computational fluid dynamics.

Example 4.5

Now, let us illustrate the application of subdomain collocation method using the beam deflection problem considered earlier. The governing differential equation is given by

$$EI\frac{d^2y}{dx^2} - M_0 = 0$$

with boundary conditions

$$y(0) = 0 \qquad \& \qquad y(L) = 0.$$

Trigonometric approximation to deflection curve. The sinusoidal trial function that satisfies the specified boundary conditions is given by

$$u(x) = A\sin\frac{\pi x}{L} = c_1 N_1$$

where $N_1 = \sin\frac{\pi x}{L}$ and the residual

$$R(x;A) = -EI\frac{A\pi^2}{L^2}\sin\frac{\pi x}{L} - M_0.$$

Since there is just one unknown coefficient in the approximating function, we have only one subdomain which is the domain itself. Thus, equation (4.25) becomes

$$\int_0^L R(x)dx = \int_0^L \left(-EI\frac{A\pi^2}{L^2}\sin\frac{\pi x}{L} - M_0\right)dx = 0.$$

The integration yields the following equation

$$-\left(\frac{2EI\pi}{L}\right)A - M_0 L = 0$$

which can be solved for A to obtain

$$A = -\frac{M_0 L^2}{2\pi EI}.$$

Thus, the approximate solution is

$$u(x) = -\frac{M_0 L^2}{2\pi EI}\sin\frac{\pi x}{L}.$$

This approximate solution is also found to be in close agreement with the exact solution. However, a comparison of the above results with that of point collocation method shows that the approximate solutions are different.

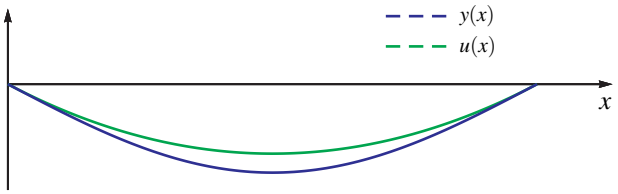

Figure 4.5: Beam deflection problem – result of subdomain collocation method.

Polynomial approximation to deflection curve. The second degree polynomial trial function that satisfies the specified boundary conditions is given by

$$u(x) = cx(x-L) = c_1 N_1$$

where $N_1 = x(x-L)$ and the residual

$$R(x;c) = EI \times 2c - M_0.$$

Integrating the residual over $[0,L]$

$$\int_0^L R(x)dx = \int_0^L (2EIc - M_0)\,dx = 0$$

which yields

$$(2EIc - M_0)\,L = 0.$$

Solving for c, we have

$$c = \frac{M_0}{2EI}.$$

and thus, the approximate solution is

$$u(x) = -\frac{M_0}{2EI}x(L-x).$$

As in the case of point collocation method, selection of a second degree polynomial as approximating function results in exact solution.

4.4 Least Square Method

In the least square weighted residual method, the weight functions are chosen to be the derivatives of residual with respect to unknown fitting coefficients c_i's of the approximate solution. So, we set

$$w_i = \frac{\partial R}{\partial c_i}, \qquad (i = 1,2,\cdots,n). \tag{4.27}$$

Thus, for a one-dimensional problem in the interval $[a,b]$, the weighted residual integral given by equation (4.8) becomes

$$\int_a^b w_i R(x)\,dx = \int_a^b \frac{\partial R}{\partial c_i} R(x)\,dx = 0, \quad (i = 1,2,\cdots,n). \tag{4.28}$$

The motivation for this choice of weight function is that we have the following equation

$$\frac{\partial}{\partial c_i}\int_a^b R^2(x)\,dx = 0$$

which implies that the 'average squared residual' in the interval $[a,b]$ is to be minimized with respect to fitting coefficients c_i. Driving the average squared residual to zero will drive the residual R to zero. Since, we have from equation (4.9), $\partial R/\partial c_i = \mathscr{L}(N_i)$,

$$\frac{\partial R}{\partial c_i} = \mathscr{L}(N_i)$$

for the least square method the linear system of equation (4.10) takes the form

$$\begin{bmatrix} \int_a^b \mathscr{L}(N_1)\mathscr{L}(N_1)dx & \int_a^b \mathscr{L}(N_1)\mathscr{L}(N_2)dx & \cdots & \int_a^b \mathscr{L}(N_1)\mathscr{L}(N_n)dx \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \int_a^b \mathscr{L}(N_n)\mathscr{L}(N_1)dx & \int_a^b \mathscr{L}(N_n)\mathscr{L}(N_2)dx & \cdots & \int_a^b \mathscr{L}(N_n)\mathscr{L}(N_n)dx \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = -\begin{bmatrix} \int_a^b \mathscr{L}(N_1) f\,dx \\ \int_a^b \mathscr{L}(N_2) f\,dx \\ \vdots \\ \int_a^b \mathscr{L}(N_n) f\,dx \end{bmatrix}. \tag{4.29}$$

Example 4.6

Trigonometric approximation to deflection curve. We again consider the beam deflection problem. The trigonometric trial function is given by

$$u(x) = A\sin\frac{\pi x}{L} = c_1 N_1$$

where $N_1 = \sin\frac{\pi x}{L}$ and the residual

$$R(x;A) = -EI\frac{A\pi^2}{L^2}\sin\frac{\pi x}{L} - M_0$$

and its derivative,

$$\frac{\partial R}{\partial A} = -EI\frac{\pi^2}{L^2}\sin\frac{\pi x}{L}.$$

The weighted residual equation (4.24) can now be written as

$$\int_0^L \frac{\partial R}{\partial A}R(x)dx = \int_0^L -EI\frac{\pi^2}{L^2}\sin\frac{\pi x}{L}\left(-EI\frac{A\pi^2}{L^2}\sin\frac{\pi x}{L} - M_0\right)dx = 0.$$

The integration yields the following equation

$$\left(\frac{EI\pi^2}{2L}\right)A + \frac{2M_0L}{\pi} = 0.$$

Solving for A, we have

$$A = -\frac{4M_0L^2}{\pi^3 EI}$$

and thus, the approximate solution is

$$u(x) = -\frac{4M_0L^2}{\pi^3 EI}\sin\frac{\pi x}{L}.$$

Figure 4.6 shows that the approximate solution $u(x)$ agrees well with the exact solution $y(x)$ over the interval $[0,L]$ and is found to be more accurate than the solution using point collocation and subdomain collocation methods.

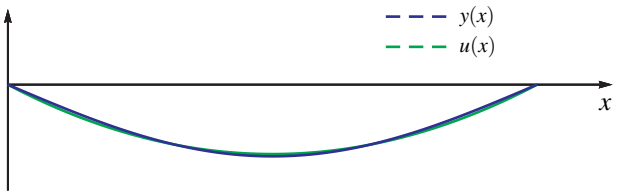

Figure 4.6: Beam deflection problem – result of least square method.

Polynomial approximation to deflection curve. We use the second degree polynomial trial function

$$u(x) = cx(x-L) = c_1 N_1$$

where $N_1 = x(x-L)$ and the residual

$$R(x;c) = EI \times 2c - M_0$$

and its derivative,

$$\frac{\partial R}{\partial c} = 2EI.$$

The weighted residual equation (4.24) can now be written as

$$\int_0^L \frac{\partial R}{\partial c} R(x)dx = \int_0^L 2EI\,(2EIc - M_0)\,dx = 0.$$

The integration yields the following equation

$$2EI\,(2EIc - M_0)\,L = 0.$$

Solving for c, we have

$$c = \frac{M_0}{2EI}$$

and thus, the approximate solution is

$$u(x) = -\frac{M_0}{2EI} x(L-x).$$

As in the case of other two methods, selection of a second-order polynomial as approximating function results in exact solution.

Following points about least square method may be noted:

- Least square method always produces symmetric coefficient matrix regardless of the differential operator $\mathscr{L}$ and approximate solution ϕ. Further, this method also produces positive definite matrix since diagonal entries are always positive.
- Least square method is often computationally expensive.

4.5 Galerkin Method

In Galerkin version of weighted residual method, the weight functions are chosen to be the trial functions themselves. This is the method we usually used for developing finite element equations for field problems. So, in Galerkin method we set

$$w_i = N_i, \qquad (i = 1,2,\cdots,n). \tag{4.30}$$

The unknown coefficients in the approximate solution are determined by setting the integral over D of the weighted residual to zero. For one-dimensional problem in the interval $[a,b]$, this procedure will results

$$\int_a^b w_i R(x)\,dx = \int_a^b N_i R(x)\,dx = 0, \quad (i = 1,2,\cdots,n). \tag{4.31}$$

For the Galerkin method the linear system of equation (4.8) takes the form

$$\begin{bmatrix} \int_a^b N_1\mathscr{L}(N_1)dx & \int_a^b N_1\mathscr{L}(N_2)dx & \cdots & \int_a^b N_1\mathscr{L}(N_n)dx \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \int_a^b N_n\mathscr{L}(N_1)dx & \int_a^b N_n\mathscr{L}(N_2)dx & \cdots & \int_a^b N_n\mathscr{L}(N_n)dx \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = - \begin{bmatrix} \int_a^b N_1 f\,dx \\ \int_a^b N_2 f\,dx \\ \vdots \\ \int_a^b N_n f\,dx \end{bmatrix}. \tag{4.32}$$

Following points about Galerkin method may be noted:

- Galerkin method produces symmetric positive definite coefficient matrix if the differential operator is self-adjoint.
- Galerkin method requires less computational effort compared to the least square method.

Example 4.7

Trigonometric approximation to deflection curve. Yet again we consider the now familiar beam deflection problem. The trigonometric trial function is given by

$$u(x) = A\sin\frac{\pi x}{L} = c_1 N_1$$

where $N_1 = \sin\frac{\pi x}{L}$ and the residual

$$R(x;A) = -EI\frac{A\pi^2}{L^2}\sin\frac{\pi x}{L} - M_0.$$

The unknown coefficients in the approximate solution are determined by setting the integral over $[0,L]$ of the weighted residual to zero. The weighted residual equation give by (4.31) can now be written as

$$\int_0^L N_1 R(x)dx = \int_0^L \sin\frac{\pi x}{L}\left(-EI\frac{A\pi^2}{L^2}\sin\frac{\pi x}{L} - M_0\right)dx = 0.$$

The integration yields the following equation

$$\left(\frac{EI\pi^2}{2L}\right)A + \frac{2M_0L}{\pi} = 0.$$

Solving for A, we have

$$A = -\frac{4M_0L^2}{\pi^3 EI}$$

and thus, the approximate solution is

$$u(x) = -\frac{4M_0L^2}{\pi^3 EI}\sin\frac{\pi x}{L}.$$

Figure 4.7 shows that the approximate solution $u(x)$ agrees well with the exact solution $y(x)$ over the interval $[0,L]$ and is found to be slightly more accurate than the solution using point collocation method.

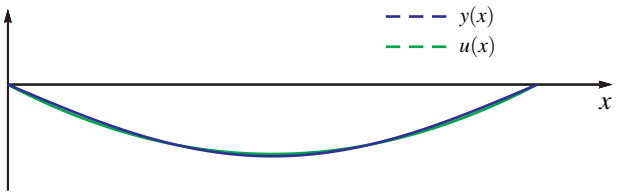

Figure 4.7: Beam deflection problem – result of Galerkin method.

Polynomial approximation to deflection curve. We use the second degree polynomial trial function

$$u(x) = cx(x-L) = c_1 N_1$$

where $N_1 = x(x-L)$ and the residual

$$R(x;c) = EI \times 2c - M_0.$$

Integrating the weighted residual over $[0,L]$

$$\int_0^L N_1 R(x)dx = \int_0^L x(x-L)\,(2EIc2c - M_0)\,dx = 0$$

to obtain the following equation

$$(2EIc - M_0)\left(\frac{L^3}{3} - \frac{L^3}{2}\right) = 0.$$

Solving for c, we have

$$c = \frac{M_0}{2EI}$$

and thus, the approximate solution is

$$u(x) = -\frac{M_0}{2EI}x(L-x).$$

As in the case of other methods, selection of a second-order polynomial as approximating function results in exact solution.

Example 4.8

Solve the differential equation

$$\frac{d^2y}{dx^2} + y = x, \qquad x \in [0,2]$$

with the boundary conditions

$$y(0) = 0, \qquad y(2) = 5$$

using Galerkin method.

We use the same trial function $u(x)$ as with the point collocation method:

$$u(x) = 2.5x + c_2 x(x-2) + c_3 x^2(x-2) = 2.5N_1 + c_2 N_2 + c_3 N_3$$

so that $N_2 = x(x-2)$ and $N_3 = x^2(x-2)$. The residual of the differential equation is given by

$$R(x) = \frac{d^2u}{dx^2} + u - x.$$

After duly substituting u and u'' in the above residual equation, we get

$$R(x) = 2c_2 + c_3(6x-4) + 2.5x + c_2 x(x-2) + c_3 x^2(x-2) - x.$$

The unknown coefficients in the approximate solution are determined by using equation (4.31):

$$\int_0^2 x(x-2)\,R(x)\,dx = 0$$
$$\int_0^2 x^2(x-2)\,R(x)\,dx = 0$$

which gives the two algebraic equations for c_2 and c_3:

$$4c_2 + 4c_3 = -5$$
$$2c_2 + 4c_3 = -3$$

Solving the above set of equations for c_2 and c_3 and substitute in the assumed trial function to obtain

$$u(x) = \left(\frac{5}{2}\right)x - x(x-2) - \left(\frac{1}{4}\right)x^2(x-2)$$
$$= -\left(\frac{1}{4}\right)x^3 - \left(\frac{1}{2}\right)x^2 + \left(\frac{9}{2}\right)x.$$

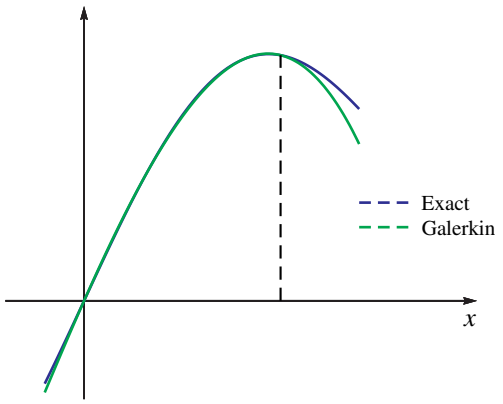

Figure 4.8: Comparison of Galerkin and exact solutions of problem #8.

Figure 4.8 shows that the approximate solution $u(x)$ agrees very well with the exact solution $y(x)$ over the interval $[0,2]$.

So we have used several types of weighted residual method for solving boundary value problems. It can be seen that, for the beam deflection problem, the application of all the method yield the exact solution, if second or higher degree polynomial is selected as the approximating function. This is because, the actual behavior of the deflection curve is parabolic, i.e., a second degree polynomial. However, the selection of a sinusoidal function as approximating function yield different solutions for different method except for leat square and Galerkin methods. Now, the question naturally arises is which method gives the most accurate results. Unfortunately, there is no conclusive answer for this. The error depend on the approximating function and the differential equation to be solved. However, for most problems, the Galerkin method gives the best results.

Before we close this discussion, we will develop the Galerkin formulation for the boundary-value problem governed by the generic second-order linear ordinary differential equation. Those differential equations which we have already considered are all could be viewed as special cases of this generic equation.

Example 4.9

Consider the following linear boundary value problem governed by the following generic second-order linear ordinary differential equation:

$$\frac{d^2y}{dx^2} + Q(x)y = F(x), \qquad x \in [0,1] \tag{4.33}$$

with the Dirichlet boundary conditions

$$y(0) = 0, \qquad y(1) = Y$$

To use Galerkin method to solve the above boundary value problem, we use a polynomial trial function $\phi(x)$ of degree 3 in the form

$$\phi(x) = c_1N_1(x) + c_2N_2(x) + c_3N_3(x) = c_1x + c_2x(x-1) + c_3x^2(x-1)$$

The trial functions $N_1 = x$, $N_2 = x(x-1)$, and $N_3 = x^2(x-1)$ are linearly independent. Applying the boundary conditions yields $c_1 = Y$. Thus, the approximate solution is given by

$$\phi(x) = Yx + c_2x(x-1) + c_3x^2(x-1) = \phi(x; c_2, c_3) \tag{4.34}$$

The residual is obtained after substituting $\phi(x)$ for $y(x)$ in the differential equation (4.19)

$$R(x) = \frac{d^2\phi}{dx^2} + Q(x)\phi - F(x) \tag{4.35}$$

The second derivative ϕ'' is obtained from equation (4.20):

$$\frac{d^2\phi}{dx^2} = 2c_2 + c_3(6x-2)$$

Therefore, the residual becomes

$$R(x) = 2c_2 + c_3(6x-2) + Q\left[Yx + c_2x(x-1) + c_3x^2(x-1)\right] - F \tag{4.36}$$

In Galerkin method, we choose the weighting function as the trial functions, thus:

$$w_2 = N_2 = x(x-1) \qquad \text{and} \qquad w_3 = N_3 = x^2(x-1)$$

The unknown coefficients in the approximate solution are determined by setting the integral of the weighted residual to zero.

$$\int_0^1 x(x-1)\left\{2c_2 + c_3(6x-2) + Q\left[Yx + c_2x(x-1) + c_3x^2(x-1)\right] - F\right\} dx = 0 \tag{4.37a}$$

$$\int_0^1 x^2(x-1)\left\{2c_2 + c_3(6x-2) + Q\left[Yx + c_2x(x-1) + c_3x^2(x-1)\right] - F\right\} dx = 0 \tag{4.37b}$$

Integration can be performed after substituting the functions $Q(x)$ and $F(x)$ to obtain the algebraic equations for unknowns c_2 and c_2. If Q and F are constants, it is easy to carry out the integration. The result is:

$$c_2\left(\frac{1}{3} - \frac{Q}{30}\right) + c_3\left(\frac{1}{6} - \frac{Q}{60}\right) = -\frac{QY}{12} + \frac{F}{6} \tag{4.38a}$$

$$c_2\left(\frac{1}{6} - \frac{Q}{60}\right) + c_3\left(\frac{2}{15} - \frac{Q}{105}\right) = -\frac{QY}{20} + \frac{F}{12} \tag{4.38b}$$

Note: It must be emphasized that the Galerkin method is not the finite element method. In fact, Galerkin method was available much before the concept of FEM is introduced. The essential difference between the Galerkin method and FEM is that unlike in Galerkin method, the approximating function in FEM is not defined over the whole physical domain; it is only defined over the individual elements which constitutes the physical domain. In standard FEM, the Galerkin method is often used to derive the element equations.

Chapter 5

Finite Element Method

5.1 Finite Element Formulation

The main drawback of the weighted residual method is that it is difficult to find good trial functions because one may not have any prior knowledge of the behaviour of the solution $y(x)$. Polynomials are often selected as trial function in such cases and might do a poor job of interpolation (we can think of the trail function $\phi(x)$ as an interpolation function between the boundary conditions that also satisfies the differential equation), especially when the interval $[a,b]$ is large. Higher degree polynomial interpolation may leads to *Runge phenomenon* (oscillation at the edges of an interval that occurs when using polynomial interpolation with polynomials of high degree over a set of equispaced interpolation points) for certain functions.

It is our experience that low-degree polynomial can capture the behaviour of solution if the interval $[a,b]$ is very short. So, we hope to successfully apply the weighted residual methods using low-degree polynomials by subdividing the interval $[a,b]$ into smaller subintervals. That is, we use piece-wise lower degree polynomials in smaller subintervals rather than going for a higher degree polynomial for the entire domain. This is the strategy used in the *finite element method*.

5.1.1 Steps in FEM

The major steps involved in the solution of a problem using FEM are:

1. The solution interval (domain) is discretized (subdivided) into number of small nonoverlapping subintervals (subregions) referred to as *finite elements*. These elements are interconnected through points known as the *nodes* (of the elements).

   The process of discretization is essentially an exercise of engineering judgment. The shape, size, and number of elements have to be chosen carefully in such a way that the original domain is approximated as closely as possible without increasing the computational efforts.

2. Select an approximating function known as interpolation polynomial for $\phi(x)$ to represent the variation of the dependent variable $y(x)$ over the elements.

3. Apply the Galerkin method to each element separately to interpolate (subject to the differential equation) between the end nodal values, $\phi(x_i)$ and $\phi(x_j)$, where these $\phi(x_i)$'s are approximations to the $y(x_i)$'s that are the true solution to the differential equation. [These nodal values are actually the fitting coefficients c's in the equation (4.6) for approximate solution $\psi(x)$.]

4. The result of applying Galerkin method to an element (e) is a pair of equations in which the unknowns are the nodal values at the ends of element (e), the c's. When we have done this for each element, we have equations that involve all the nodal values. Assembly of the element equation to form the global equation for the problem. It produce a system of algebraic equations - one equation for each element.

5. These equations are adjusted for the boundary conditions and solved to get approximations to $y(x)$ at the nodes; we get intermediate values for $y(x)$ by linear interpolation.

6. The system of algebraic equations are then solved to get the approximate solution $\phi(x)$ of the problem.

5.1.2 Selection of Elements

One of the most important step in finite element analysis is the selection of the particular type of finite elements and the definition of appropriate approximating function within the element. The approximating function is referred to as *interpolation polynomial*. Each element is characterized by several features. So, when someone asks 'what type of element you are using' for a particular problem, they are really asking for four distinct pieces of information.

1. The geometric shape of the element; whether it is a line segment, triangle, rectangle, tetrahedron, etc.

2. The number and types of nodes in each element; whether the element contains two nodes or three nodes etc. By type of nodes we mean whether the nodes are interior or exterior. Exterior nodes are the nodes that lie on the boundaries of the element and they represent the point of connection between bordering elements. Interior nodes are the nodes that do not connect to the neighboring elements.

3. The type of the nodal variable. Depending upon the problem, the nodal variable may have single degree of freedom or several degrees of freedom.

4. The type of the approximating function. Whether the approximating function is polynomial, trigonometric functions etc. Polynomial approximating functions have found wide spread acceptance because they are easy to manipulate mathematically.

If any one these characteristic features is missing, the description of the element is incomplete.

5.1.3 One-dimensional Linear Element

We now begin the formal development of the FEM procedure. Although it involves several steps, each step is straightforward. The differential equation that we will solve is

$$\frac{d^2y}{dx^2} + Q(x)y = F(x), \qquad x \in [a,b] \tag{5.1}$$

As we have already seen, this is the model equation for many simple physical problems. One of them is the one-dimensional steady-state heat conduction problem. Consider a fin attached to a solid wall. We would like to determine the steady-state temperature variation along the length of the fin (with insulated tip condition) using FEM. The problem is governed by a second-order ODE

$$\frac{d^2\theta}{dx^2} - m^2\theta = 0, \qquad x \in [0,L] \tag{5.2a}$$

with the boundary conditions

$$\theta(0) = T(0) - T_\infty = T_0 - T_\infty \qquad \text{and} \qquad \left.\frac{d\theta}{dx}\right|_{x=L} = 0$$

where T_0 is the temperature at the root of the fin, T_∞ is the ambient temperature, $\theta = T - T_\infty$, and

$$m = \sqrt{\frac{hP}{kA_c}}$$

where h is the heat transfer coefficient, P is the perimeter of the fin, A_c is the cross-section area of the fin, and k is the thermal conductivity of the fin material.

The analytical solution for the temperature distribution is given by

$$\frac{\theta}{\theta_0} = \frac{T - T_\infty}{T_0 - T_\infty} = \frac{\cosh m(L-x)}{\cosh mL} \tag{5.2b}$$

To apply the finite element method, the first step is to discretize the domain. We use one-dimensional *linear element* for this purpose. A linear element is an element for which the assumed function varies linearly within the element. Since two points are needed to uniquely specify a linear function, a linear element should have exactly two nodes. The nodes of the elements are numbered from left to right. We cannot place the nodes arbitrarily. The are certain rules to be followed. These are:

1. Place the nodes closer in the region where you expect the variables to change rapidly and further apart where the variable have less changes.
2. Place a node wherever a stepped or abrupt change in the material or geometric properties of the domain occurs.
3. Place a node wherever the numerical value of the unknown variable is desired.

These rules requires the user to have some knowledge of the behavior of the unknown variable in the domain. This is where the engineering knowledge of the user comes handy.

Since we are using the linear element, the interpolation polynomial for an isolated element (e) may be written as

$$\phi(x) = a_1 + a_2 x \tag{5.3}$$

where a_1 and a_2 are two constants whose value can be expressed in terms of nodal unknowns. Let the nodes of the isolated element is designated by i and j. The corresponding nodal values of the unknowns are denoted by ϕ_i and ϕ_j. With reference to a reference coordinate system the the nodal coordinates are x_i and x_j so that the length of the element, $L_i = x_j - x_i$. The expression for the interpolation polynomial can also be written in matrix for as

$$\phi(x) = [\mathsf{p}][\mathsf{a}]$$

where the row vector [p] is

$$[\mathsf{p}] = \begin{bmatrix} 1 & x \end{bmatrix}$$

and the column vector [a] is

$$[\mathsf{a}] = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

Equation (5.3) when applied to each of the nodes provides the following set of equation

$$\phi_i = a_1 + a_2 x_i$$
$$\phi_j = a_1 + a_2 x_j$$

or

$$[\phi] = [\mathsf{G}][\mathsf{a}]$$

where

$$[\phi] = \begin{bmatrix} \phi_i \\ \phi_j \end{bmatrix}$$

and

$$[\mathsf{G}] = \begin{bmatrix} 1 & x_i \\ 1 & x_j \end{bmatrix}$$

Thus,

$$[\mathsf{a}] = [\mathsf{G}]^{-1}[\phi]$$

These equations can be solved (say, using Cramer's rule) for

$$a_1 = \frac{\phi_i x_j - \phi_j x_i}{L_i} \qquad \text{and} \qquad a_2 = \frac{\phi_j - \phi_i}{L_i}$$

Substitution of a_1 and a_2 in equation (5.3) which, after collection of terms, gives the following expression for the interpolation polynomial

$$\begin{aligned}\phi(x) &= \phi_i + \frac{\phi_j - \phi_i}{L_i}(x - x_i) \\ &= \left(\frac{x_j - x}{L_i}\right)\phi_i + \left(\frac{x - x_i}{L_i}\right)\phi_j \\ &= N_i\phi_i + N_j\phi_j \\ &= [\mathrm{N}][\phi]\end{aligned} \tag{5.4}$$

where

$$N_i = \frac{x_j - x}{L_i}, \qquad N_j = \frac{x - x_i}{L_i} \tag{5.5}$$

and

$$[\mathrm{N}] = \begin{bmatrix} N_i & N_j \end{bmatrix}, \qquad [\phi] = \begin{bmatrix} \phi_i \\ \phi_j \end{bmatrix}$$

The [N] is the row vector of shape functions and $[\phi]$ is the column vector of element nodal values. In the interpolation polynomial, the function which is being multiplied by the nodal values is called the *shape function* or *interpolation function*. These are the functions selected to represents the behavior of the unknown variable within an element. Recognize that the N's in the above equations are really first-degree Lagrangian polynomials.

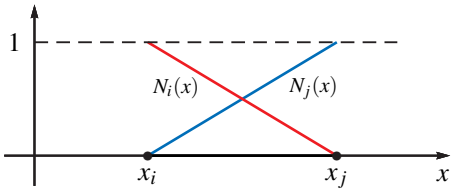

Figure 5.1: Linear shape functions N_i and N_j within element.

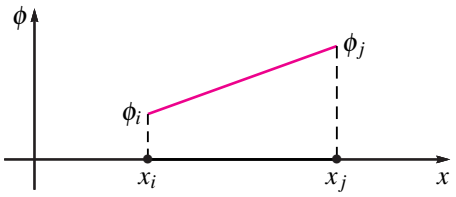

Figure 5.2: Variation of ϕ within a linear element.

Properties of Shape Function

- Number of shape functions of an element depends on the number of nodes in the element. Each shape function is associated with a unique node.

- Each shape function has a value of one at its own node and zero at the other nodes.
- The sum of the shape function is equal to one everywhere within the element.
- The shape functions and the interpolation polynomial of an element are of same types. If the interpolation polynomial is quadratic the resulting shape functions are also quadratic.
- The derivative of the shape functions with respect to the independent variable sums to zero.

These properties are valid for all types of element, whether it is one-dimensional, two-dimensional, or three-dimensional element with polynomial as interpolation polynomial.

5.1.4 One-dimensional Quadratic Element

The interpolation polynomial for a quadratic element is defined as

$$\phi(x) = a_1 + a_2 x + a_3 x^2 \tag{5.6}$$

where a_1, a_2, and a_3 are constants whose value are to be expressed in terms of nodal unknowns. Since there are three constants in the interpolation polynomial, the quadratic element should have three nodes. Let the nodes of the isolated element is designated by i, j, and k. The corresponding nodal values of the unknowns are denoted by ϕ_i, ϕ_j, and ϕ_k. With reference to a reference coordinate system the the nodal coordinates are x_i, x_j, and x_k and the length of the element $L_i = x_k - x_i$. The expression for the interpolation polynomial can also be written in matrix for as

$$\phi(x) = [\mathsf{p}][\mathsf{a}]$$

where the row vector [p] is

$$[\mathsf{p}] = \begin{bmatrix} 1 & x & x^2 \end{bmatrix}$$

and the column vector [a] is

$$[\mathsf{a}] = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

Thus, the column vector of element nodal values may be written as

$$[\phi] = [\mathsf{G}][\mathsf{a}]$$

That is

$$\begin{bmatrix} \phi_i \\ \phi_j \\ \phi_k \end{bmatrix} = \begin{bmatrix} 1 & x_i & x_i^2 \\ 1 & x_j & x_j^2 \\ 1 & x_k & x_k^2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

The vector of undetermined parameters [a] can be determined by premultiplying the above equation with $[\mathsf{G}]^{-1}$.

$$\begin{aligned} \phi(x) &= \frac{2}{L_i^2}(x-x_j)(x-x_k)\phi_i + -\frac{4}{L_i^2}(x-x_i)(x-x_k)\phi_j + \frac{2}{L_i^2}(x-x_i)(x-x_j)\phi_k \\ &= N_i\phi_i + N_j\phi_j + N_k\phi_k \\ &= [\mathsf{N}][\phi] \end{aligned} \tag{5.7}$$

The one-dimensional quadratic shape functions are thus given by

$$N_i = \frac{2}{L_i^2}(x-x_j)(x-x_k) \tag{5.8a}$$

$$N_j = -\frac{4}{L_i^2}(x-x_i)(x-x_k) \tag{5.8b}$$

$$N_k = \frac{2}{L_i^2}(x-x_i)(x-x_j) \tag{5.8c}$$

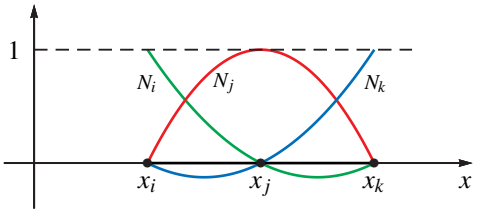

Figure 5.3: Quadratic shape functions N_i, N_j, and N_k within element.

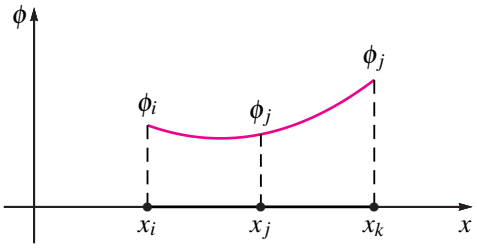

Figure 5.4: Variation of ϕ within a quadratic element.

5.2 Two-dimensional Elements

So far we have discussed some of the one-dimensional elements which is suitable for one-dimensional problems governed by ordinary differential equations. For two-dimensional problem governed by PDE's requires the determination of dependent variable in two-dimensional domains. For the discretization of a two-dimensional space we use two-dimensional elements. Two basic types of two-dimensional elements are the linear triangular element and bilinear rectangular element. Despite the simplicity of these elements they are extensively used in finite element analysis of heat transfer and solid mechanics problems.

5.2.1 Linear Triangular Element

The linear triangular element has straight edges and a node at each corner. The interpolation polynomial for the element is given by

$$\phi(x,y) = a_1 + a_2 x + a_3 y = [\mathsf{p}][\mathsf{a}] \tag{5.9}$$

which is a complete linear polynomial in x and y, because it contains a constant term and all possible terms in x and y. As a result, the triangular element can take any orientation in the domain. Here, a_1, a_2, and a_3 are constants whose value are to be expressed in terms of nodal

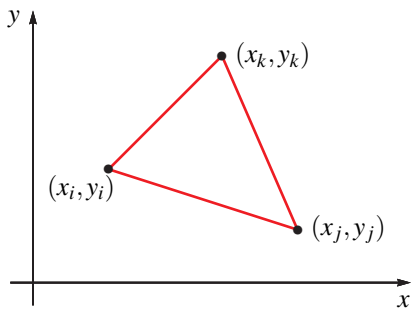

Figure 5.5: A linear triangular element.

unknowns.

$$[\mathsf{p}] = \begin{bmatrix} 1 & x & y \end{bmatrix}, \qquad [\mathsf{a}] = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

Application equation (5.9) for all the three nodes produces the following equations for the unknown vector [a]

$$\begin{bmatrix} \phi_i \\ \phi_j \\ \phi_k \end{bmatrix} = \begin{bmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

That is

$$[\phi] = [\mathsf{G}][\mathsf{a}]$$

Equation (5.9) can now be written as

$$\phi(x,y) = [\mathsf{p}][\mathsf{G}]^{-1}[\phi] = [N][\phi] = N_i\phi_i + N_j\phi_j + N_k\phi_k$$

The shape functions are given by

$$N_i = \frac{1}{2A_i}(a_i + b_i x + c_i y) \tag{5.10a}$$

$$N_j = \frac{1}{2A_i}(a_j + b_j x + c_j y) \tag{5.10b}$$

$$N_k = \frac{1}{2A_i}(a_k + b_k x + c_k y) \tag{5.10c}$$

where

$$\begin{array}{lll} a_i = x_j y_k - x_k y_j & b_i = y_j - y_k & c_i = x_k - x_j \\ a_j = x_k y_i - x_i y_k & b_j = y_k - y_i & c_j = x_i - x_k \\ a_k = x_i y_j - x_j y_i & b_k = y_i - y_j & c_k = x_j - x_i \end{array}$$

and A_i is the area of the triangle and hence

$$2A_i = \begin{vmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{vmatrix}$$

Here the scalar function $\phi(x,y)$ (interpolation polynomial) is related to the nodal values of ϕ through a set of shape function that are linear in x and y.

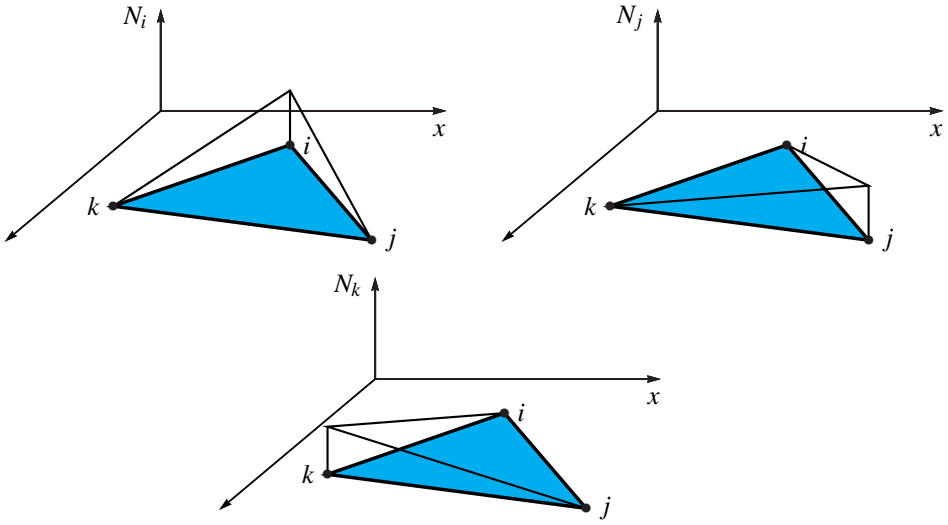

Figure 5.6: Shape functions N_i, N_j, and N_k within a linear triangular element.

5.2.2 Bilinear Rectangular Element

In local coordinate system (s,t) the interpolation polynomial for a bilinear rectangular element is given by

$$\phi = a_1 + a_2 s + a_3 t + a_4 st \tag{5.11}$$

Let $2b$ is the length of side of rectangle in the s direction and $2a$ is the length of side of rectangle in the t direction.

$$\phi = [\mathrm{N}][\phi] = N_i\phi_i + N_j\phi_j + N_k\phi_k + N_m\phi_m$$

where the shape functions are given by

$$N_i = \left(1 - \frac{s}{2b}\right)\left(1 - \frac{t}{2a}\right) \tag{5.12a}$$

$$N_j = \frac{s}{2b}\left(1 - \frac{t}{2a}\right) \tag{5.12b}$$

$$N_k = \frac{st}{4ab} \tag{5.12c}$$

$$N_i = \frac{t}{2a}\left(1 - \frac{s}{2b}\right) \tag{5.12d}$$

5.3 Finite Element Equations

Here we proceed to develop finite element equation for a one-dimensional problem using Galerkin method. Consider the differential equation of the form

$$\frac{d^2y}{dx^2} + Q(x)y = F(x), \qquad x \in [a,b] \tag{5.1}$$

The residual for the element (e) is obtained by substituting the approximate solution $\phi(x)$ for $y(x)$ in the differential equation (5.1)

$$R^{(e)}(x) = \frac{d^2\phi}{dx^2} + Q(x)\phi - F(x) \qquad x \in [x_i, x_j] \tag{5.13}$$

where $\phi(x)$ is the interpolation polynomial for the element (e) is given by

$$\phi^{(e)}(x) = N_i\phi_i + N_j\phi_j = [\mathsf{N}][\phi^{(e)}] \tag{5.14}$$

The Galerkin method sets the integral of residual $R^{(e)}$ weighted with each of the N's (over the length of the element) to zero:

$$\int_{x_i}^{x_j} N_i R^{(e)}(x)\,dx = 0 \tag{5.15a}$$

$$\int_{x_i}^{x_j} N_j R(x)^{(e)}\,dx = 0 \tag{5.15b}$$

Substituting the residual from (5.13) into (5.15), we get the weighted residual equation for the element (e):

$$\int_{x_i}^{x_j} N_i \left(\frac{d^2\phi}{dx^2} + Q(x)\phi - F(x)\right) dx = 0 \tag{5.16a}$$

$$\int_{x_i}^{x_j} N_j \left(\frac{d^2\phi}{dx^2} + Q(x)\phi - F(x)\right) dx = 0 \tag{5.16b}$$

which may be expanded as

$$\int_{x_i}^{x_j} N_i \frac{d^2\phi}{dx^2}\,dx + \int_{x_i}^{x_j} N_i Q(x)\phi\,dx - \int_{x_i}^{x_j} N_i F(x)\,dx = 0 \tag{5.17a}$$

$$\int_{x_i}^{x_j} N_j \frac{d^2\phi}{dx^2}\,dx + \int_{x_i}^{x_j} N_j Q(x)\phi\,dx - \int_{x_i}^{x_j} N_j F(x)\,dx = 0 \tag{5.17b}$$

The first integral in (5.17a) can be transformed by applying integration by parts[1] to yield

$$\int_{x_i}^{x_j} N_i \frac{d^2\phi}{dx^2}\,dx = \left[N_i \frac{d\phi}{dx}\right]_{x_i}^{x_j} - \int_{x_i}^{x_j} \frac{dN_i}{dx}\frac{d\phi}{dx}\,dx$$

Thus, we have taken the significant step of lowering the second-order derivative in the formulation to a first-order derivative. Next, in the second integral, we will take Q out from the integrand as $\overline{Q}$, an average value within the element. We also take F outside the third integral by defining the average value $\overline{F}$:

$$\int_{x_i}^{x_j} N_i Q(x)\phi\,dx = \overline{Q}\int_{x_i}^{x_j} N_i\,\phi\,dx \qquad \text{and} \qquad \int_{x_i}^{x_j} N_i F(x)\,dx = \overline{F}\int_{x_i}^{x_j} N_i\,dx$$

With these results, equation (5.17a) becomes

$$-\int_{x_i}^{x_j} \frac{dN_i}{dx}\frac{d\phi}{dx}\,dx + \overline{Q}\int_{x_i}^{x_j} N_i\,\phi\,dx - \overline{F}\int_{x_i}^{x_j} N_i\,dx + \left[N_i\frac{d\phi}{dx}\right]_{x_i}^{x_j} = 0$$

The last term of the above equation can be simplified as follows:

$$\left[N_i\frac{d\phi}{dx}\right]_{x_i}^{x_j} = N_i(x_j)\left.\frac{d\phi}{dx}\right|_{x_j} - N_i(x_i)\left.\frac{d\phi}{dx}\right|_{x_i}$$

However, recall that $N_i(x_i) = 1$ and $N_i(x_j) = 0$, and therefore

$$\left[N_i\frac{d\phi}{dx}\right]_{x_i}^{x_j} = -\left.\frac{d\phi}{dx}\right|_{x_i}$$

With this simplification and after changing the sign, we have

$$\int_{x_i}^{x_j} \frac{dN_i}{dx}\frac{d\phi}{dx}\,dx - \overline{Q}\int_{x_i}^{x_j} N_i\,\phi\,dx + \overline{F}\int_{x_i}^{x_j} N_i\,dx + \left.\frac{d\phi}{dx}\right|_{x_i} = 0 \tag{5.18a}$$

Doing the similar exercise with equation (5.17b) gives

$$\int_{x_i}^{x_j} \frac{dN_j}{dx}\frac{d\phi}{dx}\,dx - \overline{Q}\int_{x_i}^{x_j} N_j\,\phi\,dx + \overline{F}\int_{x_i}^{x_j} N_j\,dx - \left.\frac{d\phi}{dx}\right|_{x_j} = 0 \tag{5.18b}$$

Notice that the integration by parts has led to two significant outcomes. First, it has incorporated the *natural* (or *Neumann*) boundary conditions:

$$\left.\frac{d\phi}{dx}\right|_{x_i} \qquad \text{and} \qquad \left.\frac{d\phi}{dx}\right|_{x_j}$$

directly into the element equations. Second, it has lowered the second derivative to a first derivative. This latter outcome yields the significant result that the approximation functions need to preserve continuity of $\phi(x)$ but not slope $(d\phi/dx)$ at the nodes.

[1] $\int uv'\,dx = uv - \int u'v\,dx$

Combining the integral equations (5.18a) and (5.18b) to obtain

$$\int_{x_i}^{x_j} \frac{d[\mathsf{N}]^T}{dx}\frac{d\phi}{dx}\,dx - \overline{Q}\int_{x_i}^{x_j}[\mathsf{N}]^T\,\phi\,dx + \overline{F}\int_{x_i}^{x_j}[\mathsf{N}]^T\,dx - \left[[\mathsf{N}]^T\frac{d\phi}{dx}\right]_{x_i}^{x_j} = [0] \tag{5.19}$$

where

$$[\mathsf{N}] = \begin{bmatrix} N_i & N_j \end{bmatrix}$$

Next, the element interpolation polynomial

$$\phi^{(e)}(x) = [\mathsf{N}][\phi^{(e)}]$$

can be substituted in the equation (5.19) to yield equation for element (e):

$$\left(\int_{x_i}^{x_j} \frac{d[\mathsf{N}]^T}{dx}\frac{d[\mathsf{N}]}{dx}\,dx - \overline{Q}\int_{x_i}^{x_j}[\mathsf{N}]^T\,[\mathsf{N}]\,dx\right)[\phi^{(e)}] = -\overline{F}\int_{x_i}^{x_j}[\mathsf{N}]^T\,dx + \left[[\mathsf{N}]^T\frac{d\phi}{dx}\right]_{x_i}^{x_j} \tag{5.20}$$

where

$$[\phi^{(e)}] = \begin{bmatrix} \phi_i \\ \phi_j \end{bmatrix}$$

is the *element nodal vector of unknowns*. The equation (5.20) for the element (e) can be written in the following generic form:

$$[\mathsf{K}^{(e)}][\phi^{(e)}] = [\mathsf{f}^{(e)}] + [\mathsf{I}^{(e)}] \tag{5.21}$$

where the square matrix $[\mathsf{K}^{(e)}]$ is the *element stiffness matrix*,

$$[\mathsf{K}^{(e)}] = \int_{x_i}^{x_j} \frac{d[\mathsf{N}]^T}{dx}\frac{d[\mathsf{N}]}{dx}\,dx - \overline{Q}\int_{x_i}^{x_j}[N]^T\,[\mathsf{N}]\,dx$$

the vector $[\mathsf{f}^{(e)}]$ is the *element force vector*,

$$[\mathsf{f}^{(e)}] = -\overline{F}\int_{x_i}^{x_j}[\mathsf{N}]^T\,dx$$

and the vector $[\mathsf{I}^{(e)}]$ is the *inter-element requirement*

$$[\mathsf{I}^{(e)}] = \left[[\mathsf{N}]^T\frac{d\phi}{dx}\right]_{x_i}^{x_j}$$

When the element equations are assembled to get the global equation the inter-element requirement terms will cancel each other except for the boundary elements. For the boundary element the derivative type boundary conditions (natural boundary conditions in FEM terminology) are implemented using the inter-element requirement.

The next major exercise is the computation of integral involved in the element equation. For this, we can either use individual nodal equations (5.18a) and (5.18b) or the combined form of (5.20). We use individual form of equations (5.18a) and (5.18b). Here we need to evaluate several derivatives. As we are using linear element defined by the interpolation polynomial (5.4), we have

$$\frac{dN_i}{dx} = -\frac{1}{L_i}, \qquad \frac{dN_j}{dx} = \frac{1}{L_i}$$

and therefore, the derivative of ϕ is

$$\frac{d\phi}{dx} = \frac{dN_i}{dx}\phi_i + \frac{dN_j}{dx}\phi_j = \frac{\phi_j - \phi_i}{L_i}$$

Note that it represents the slope of the straight line connecting the nodes.

We will now take terms in equation (5.18a) one by one and perform the task of integrations:

$$\begin{aligned}
\int_{x_i}^{x_j} \frac{dN_i}{dx}\frac{d\phi}{dx}\,dx &= \int_{x_i}^{x_j} \left(\frac{-1}{L_i}\right)\frac{\phi_j - \phi_i}{L_i}\,dx = \frac{\phi_i - \phi_j}{L_i^2}\int_{x_i}^{x_j} dx \\
&= \left(\frac{1}{L_i}\right)\phi_i - \left(\frac{1}{L_i}\right)\phi_j \\
\overline{Q}\int_{x_i}^{x_j} N_i\,\phi\,dx = \overline{Q}\int_{x_i}^{x_j} N_i\left(N_i\phi_i + N_j\phi_j\right)dx &= \phi_i\overline{Q}\int_{x_i}^{x_j} N_i^2\,dx + \phi_j\overline{Q}\int_{x_i}^{x_j} N_i N_j\,dx \\
&= \phi_i\overline{Q}\int_{x_i}^{x_j}\left(\frac{x_j - x}{L_i}\right)^2 dx + \phi_j\overline{Q}\int_{x_i}^{x_j}\left(\frac{x_j - x}{L_i}\right)\left(\frac{x - x_i}{L_i}\right)dx \\
&= \left(\frac{\overline{Q}L_i}{3}\right)\phi_i + \left(\frac{\overline{Q}L_i}{6}\right)\phi_j \\
\overline{F}\int_{x_i}^{x_j} N_i\,dx = \overline{F}\int_{x_i}^{x_j}\frac{x_j - x}{L_i}\,dx &= \frac{\overline{f}L_i}{2}
\end{aligned}$$

Similar exercise is performed with equation (5.18b):

$$\begin{aligned}
\int_{x_i}^{x_j} \frac{dN_j}{dx}\frac{d\phi}{dx}\,dx &= -\left(\frac{1}{L_i}\right)\phi_i + \left(\frac{1}{L_i}\right)\phi_j \\
\overline{Q}\int_{x_i}^{x_j} N_j\,\phi\,dx &= \left(\frac{\overline{Q}L_i}{6}\right)\phi_i + \left(\frac{\overline{Q}L_i}{3}\right)\phi_j \\
\overline{F}\int_{x_i}^{x_j} N_j\,dx &= \frac{\overline{f}L_i}{2}
\end{aligned}$$

Substitute theses result into equations (5.18a) and (5.18b) to obtain:

$$\left(\frac{1}{L_i}\right)\phi_i - \left(\frac{1}{L_i}\right)\phi_j - \left(\frac{\overline{Q}L_i}{3}\right)\phi_i - \left(\frac{\overline{Q}L_i}{6}\right)\phi_j + \frac{\overline{f}L_i}{2} + \left.\frac{d\phi}{dx}\right|_{x_i} = 0 \tag{5.22a}$$

$$-\left(\frac{1}{L_i}\right)\phi_i + \left(\frac{1}{L_i}\right)\phi_j - \left(\frac{\overline{Q}L_i}{6}\right)\phi_i - \left(\frac{\overline{Q}L_i}{3}\right)\phi_j + \frac{\overline{f}L_i}{2} - \left.\frac{d\phi}{dx}\right|_{x_j} = 0 \tag{5.22b}$$

which can be rearranged to give two linear equations for the nodal unknown values ϕ_i and ϕ_j

$$\left(\frac{1}{L_i} - \frac{\overline{Q}L_i}{3}\right)\phi_i + \left(\frac{-1}{L_i} - \frac{\overline{Q}L_i}{6}\right)\phi_j = \frac{-\overline{f}L_i}{2} - \left.\frac{d\phi}{dx}\right|_{x_i} \tag{5.23a}$$

$$\left(\frac{-1}{L_i} - \frac{\overline{Q}L_i}{6}\right)\phi_i + \left(\frac{1}{L_i} - \frac{\overline{Q}L_i}{3}\right)\phi_j = \frac{-\overline{f}L_i}{2} + \left.\frac{d\phi}{dx}\right|_{x_j} \tag{5.23b}$$

The matrix form of equation (5.23) is given by

$$\begin{bmatrix} \left(\frac{1}{L_i} - \frac{\overline{Q}L_i}{3}\right) & \left(\frac{-1}{L_i} - \frac{\overline{Q}L_i}{6}\right) \\ \left(\frac{-1}{L_i} - \frac{\overline{Q}L_i}{6}\right) & \left(\frac{1}{L_i} - \frac{\overline{Q}L_i}{3}\right) \end{bmatrix} \begin{bmatrix} \phi_i \\ \phi_j \end{bmatrix} = \begin{bmatrix} \frac{-\overline{f}L_i}{2} \\ \frac{-\overline{f}L_i}{2} \end{bmatrix} + \begin{bmatrix} -\frac{d\phi}{dx}\Big|_{x_i} \\ \frac{d\phi}{dx}\Big|_{x_j} \end{bmatrix} \tag{5.24}$$

This is of the form (5.21) where

$$[\mathsf{K}^{(e)}] = \begin{bmatrix} \left(\dfrac{1}{L_i} - \dfrac{\overline{Q}L_i}{3}\right) & \left(\dfrac{-1}{L_i} - \dfrac{\overline{Q}L_i}{6}\right) \\ \left(\dfrac{-1}{L_i} - \dfrac{\overline{Q}L_i}{6}\right) & \left(\dfrac{1}{L_i} - \dfrac{\overline{Q}L_i}{3}\right) \end{bmatrix},$$

$$[\mathsf{f}^{(e)}] = \begin{bmatrix} \dfrac{-\overline{f}L_i}{2} \\ \dfrac{-\overline{f}L_i}{2} \end{bmatrix}, \qquad [\mathsf{I}^{(e)}] = \begin{bmatrix} -\dfrac{d\phi}{dx}\Big|_{x_i} \\ \dfrac{d\phi}{dx}\Big|_{x_j} \end{bmatrix}$$

After the individual element equations are derived, they must be linked together or assembled to characterize the unified behavior of the entire system. The assembly process is governed by the concept of continuity. That is, the solutions for contiguous elements are matched so that the unknown values (and sometimes the derivatives) at their common nodes are equivalent. Thus, the total solution will be continuous. The global finite element equation will have the form

$$[\mathsf{K}][\phi] = [\mathsf{f}] \tag{5.25}$$

where [K] is the global stiffness matrix and [f] is the global force vector.

Bibliography

[1] Bathe, K. J., *Finite Element Procedures*, 2^{nd} ed., Prentice Hall (2007).

[2] Gockenbach, M. S., *Understanding and Implementing the Finite Element Method*, SIAM (2006).

[3] Chen, Z., *The Finite Element Method: Its Fundamentals and Applications in Engineering*, World Scientific (2011).

[4] Desai, C. S. and Kundu. T., *Introductory Finite Element Method*, CRC Press (2001).

[5] Fenner, R. T., *Finite Element Methods for Engineers*, Imperial College Press (1996).

[6] Fish, J. and Belytschko, T., *A First Course in Finite Elements*, Wiley (2007).

[7] Heinrich, J. C. and Pepper, D. W., *The Intermediate Finite Element Method: Fluid Flow And Heat Transfer Applications*, CRC Press (1999).

[8] Henwood, D. and Bonet, J., *Finite Elements - A Gentle Introduction*, Macmillan (1996).

[9] Huebner, K. H., Dewhirst, D. L., Smith, D. E., and Byrom, T. G., *The Finite Element Method for Engineers*, 4^{th} ed., Wiley Interscience (2004).

[10] Hughes, T. J. R., *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis*, Dover Publications (2000).

[11] Hutton, D. V. *Fundamentals Of Finite Element Analysis*, Tata McGraw-Hill (2005).

[12] Johnson, C., *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Dover (2009).

[13] Kwon, Y. W. and Bang, H., *The Finite Element Method Using MATLAB*, 2^{nd} ed., CRC Press (2000).

[14] Larson, M. G. and Bengzon, F., *The Finite Element Method: Theory, Implementation, and Applications*, Springer (2013).

[15] Lewis, R. W., Nithiarasu, P., and Seetharamu K. N., *Fundamentals of the Finite Element Method for Heat and Fluid Flow*, Wiley (2004).

[16] Liu, G. R. and Quek, S. S., *The Finite Element Method: A Practical Course*, 2^{nd} ed., Butterworth-Heinemann (2013).

[17] Logan, D. L., *A First Course in the Finite Element Method*, 5^{th} ed., Cengage Learning (2012).

[18] Madenci, E. and Guven, I., *The Finite Element Method and Applications in Engineering Using ANSYS*, 2^{nd} ed., Springer (2015).

[19] Pepper, D. W. and Heinrich, J. C., *The Finite Element Method: Basic Concepts and Applications*, 2^{nd} ed., Taylor & Francis (2006).

[20] Reddy, J. N., *Introduction to the Finite Element Method*, 3^{rd} ed., McGraw-Hill (2005).

[21] Reddy, J. N. Gartling, D. K., *The Finite Element Method in Heat Transfer and Fluid Dynamics*, 3^{rd} ed., CRC Press (2010).

[22] Rocky, K. C., Evans, H. R., Nethercot, D. A., and Griffiths, D. W., *The Finite Element Method*, Halsted Press (1983).

[23] Segerlind, L. J., *Applied Finite Element Analysis*, 2^{nd} ed., John Wiley (1984).

[24] Seshu, P., *Textbook of Finite Element Analysis*, PHI Learning (2012).

[25] Smith, I. M., Griffiths, D. V., and Margetts, L., *Programming the Finite Element Method*, Wiley (2014).

[26] Zienkiewicz, O. C. and Taylor, R. L., *The Finite Element Method: The Basis*, 5^{th} ed., Butterworth-Heinemann (2000).

[27] Zienkiewicz, O. C., Taylor, R. L., and Nithiarasu P., *The Finite Element Method for Fluid Dynamics*, 7^{th} ed., Butterworth-Heinemann (2013).

[28] Zienkiewicz, O. C., Taylor, R. L., and Zhu, J. Z., *The Finite Element Method: Its Basis and Fundamentals*, 7^{th} ed., Butterworth-Heinemann (2013).