CHAPTER -2

A Discussion Of Spline Collocation And Finite Element Method

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Introduction To Splines</td>
</tr>
<tr>
<td>2.2</td>
<td>Survey Of Advances In Spline Theory</td>
</tr>
<tr>
<td>2.3</td>
<td>Spline Solutions Of Linear Boundary Value Problems</td>
</tr>
<tr>
<td>2.4</td>
<td>Spline In Terms Of Derivations</td>
</tr>
<tr>
<td>2.5</td>
<td>Comparison With Finite Difference Scheme</td>
</tr>
<tr>
<td>2.6</td>
<td>Introduction Of Finite Element Method</td>
</tr>
<tr>
<td>2.7</td>
<td>History Of Finite Element Method</td>
</tr>
<tr>
<td>2.8</td>
<td>The Basic Concept Of The Finite Element Method</td>
</tr>
<tr>
<td>2.9</td>
<td>Ancillary Concepts And Formulas</td>
</tr>
<tr>
<td>2.10</td>
<td>Functionals</td>
</tr>
<tr>
<td>2.11</td>
<td>Variational Method For Approximation</td>
</tr>
<tr>
<td>2.12</td>
<td>Time Dependent Problem</td>
</tr>
<tr>
<td>2.13</td>
<td>Finite Element Method For One Dimensional Second Order Equation</td>
</tr>
<tr>
<td>2.14</td>
<td>Finite Element Method For Two Dimensional Second Order Equation</td>
</tr>
</tbody>
</table>
2.1 INTRODUCTION TO SPLINES:

For tracing the graphs often we use flexible thin strips or rods and make them to pass through the plotted points by attaching weights or ducts at these points. Draftsman calls this mechanical device as spline and the weights as the ducts of the splines. The term ‘spline function’ is intended to suggest that the graph of such a function is similar to a curve smoothly drawn with the help of this device spline. The points at which weights are attached or through which the curve is to be fitted are said to be knots of the spline, some people call them nodes too.

Polynomials have long been the functions widely used to approximate other functions, mainly because they possess the simplest mathematical properties such as continuity, differentiability and they are easy to be expressed. It is quite evident from the above stated meaning of spline functions that they are also piecewise polynomials. Being polynomials, spline functions possess all the basic mathematical properties. That is to say, the spline functions along with their successive derivatives are continuous, differentiable and more generally they are analytic functions. It is a common observation that a polynomial of moderately high degree fitted to a fairly large number of given data points tends to exhibit more numerous and severe undulations than a curve drawn with a spline or a French curve.

There is considerable evidence that in many circumstances, a spline function is more acceptable as an
approximate function than a polynomial involving a comparable number of parameters. It will be justified later on through an actual numerical experience and mathematical demonstrations that the solutions of a variety of problems of best approximation really turn out to be spline functions. As per our interest we regard the spline functions as an approximate tool to solve differential equations.

2.2 SURVEY OF ADVANCES IN SPLINE THEORY:

The concept of mathematical splines was first introduced by Schoenberg [1946] who suggested their use for interpolation in many ways. Whitney [1949, 53], a student of Schoenberg working on spline functions, derived a criterion that only certain splines exist for the purpose of interpolation through a given data. The splines of even order interpolating the data at junctions, appear in a very simple fashion and their existence criterion was developed by Ahlberg et al [1961, 62, 63, 64, 65] and they established also the best approximation and convergence properties of splines. Extension of splines to several dimensions was an important step to expand the theory of splines to many directions. This fact was illustrated by Brikhoff and Garabedian [1960]. Brikhoff et al [1964] worked on investigation of error bounds for spline interpolation. Almost during the same period Greville [1964, 64a, 64b] reported that in numerical procedures spline functions can be utilized. Carl de Boor [1961, 62b] was also able to establish
the idea of existence and uniqueness of particular bicubic splines.

General spline functions with their minimum norm properties were discussed by Lynch et al [1964, 66]. Ahlberg et al [1964a, 1965b] extended the integral relation to splines in more than one dimensions. The period of 1960 to 1972 was significant in the field of spline theory, as a remarkable research on existence, uniqueness, minimum norm property and best approximation properties was done. During this period Ahlberg et al [abs.1966], Schoenberg [1964d], Ziegler [abs.1965, 69], Carl de Boor [1963], Sharma et al [abs.1964, 66], Loscalzo et al [1967], Brikhoff [abs.1964], Kershaw [1971], Golomb [1969] and many others were actively engaged in developing this area. In spline function theory, various type of splines are found like generalized splines, B-splines, cardinal splines, Lg-splines, natural splines, polynomial splines, parabolic splines, trigonometric splines. These functions were found as well as studied during the years 1964 to 1969. Kershaw [1971] presented a list of various end conditions mostly occurring in the spline fit problems. Albasiny et al [1969] tried to use spline functions as interpolants of the solution to linear differential equations. Callender [1971] presented the procedure for obtaining low order, high accuracy Spline approximations of solutions to initial value problem in ordinary differential equations. The use of B-splines was made to obtain the solution of the equations by Sincovec [1972] arising from collocation.
The error estimates of cubic splines were obtained for the first derivative by Kershaw [1972]. Barrodale et al [1966] have shown as to how to use standard numerical procedures without a change to calculate the best spline approximation. A modification in the algorithm where end point derivatives are not known to the user was proposed by Seidman et al [1972]. Ferguson et al [1973] demonstrated the use of least squares technique to find a cubic spline fit to a finite set of approximate plahar data.

Some results in the theory of best approximation by spline functions were put up by Schumaker [1969]. A fundamental theorem of algebra for Tchebycheffian monosplines is found in the work of Karlin [1969]. Rice [1969] characterized the Chebechev approximation by splines. Jerome [1969] developed Lg-splines whereas the study of Lg-splines was presented by Schultz [1969]. The spline functions are also useful in the methods developed in mathematical programming. This was suggested by Ritter [1969]. The generalized splines regarded as solutions to the constrained variational problems were investigated by Mangasarian [1969] with a view point of optimal control. The development of approximation theory by multivariate spline subspaces of Sobolev spaces is due to Schultz [1969]. The work regarding construction of spline functions in a convex set is found in a paper due to Laurent [1969].

It was the year 1968 in which Bickley [1968] brought forward an useful aspect of spline functions in light that can
be employed to solve a linear two point boundary value problem approximately. The spline presented by him is expressed in terms of infinite series or in truncated powers. This work was supported by Fife [1969]. In fact, a substantial work regarding the error estimate in cubic spline approximations was presented there. Again Fyfe [1970] extended the use of same cubic spline function to the solution of a fourth order linear two point boundary value problem and this was analysed by Hoskins [1973] with a view point of error. The applicability of spline functions to non-linear differential equations took place due to Blue [1971]. However, this work was quite dependent on the literature available in the comprehensive note by Ahlberg et al [1967]. This literature suggests the use of cardinal splines. Of course, both of these aspects of spline collocation method do not cross the restricted domain of the problem. Especially the algorithm recommended by Blue [1971] fails if the domain of the problem is extended beyond the interval [0, 1]. It will be remarkable to mention that spline functions are found useful in the solution of partial differential equations and integral equations too.

Viewing the above survey, especially emphasizing the developments and advances in theory of splines it can be said that spline functions cover many of the approximation aspects. In spite of this, there is a very long way to go and much remains to be done. The application of spline functions in the field of differential equations is a quite recent topics and this needs more analysis from the point of
view of existence, uniqueness and error bounds. This type of research will certainly explore the field of spline functions more rigorously so that they can be used to solve problems in many areas.

2.2 SPLINE SOLUTIONS OF LINEAR BOUNDARY VALUE PROBLEMS:

DEFINITION OF A POLYNOMIAL SPLINE:

Let \( \{x_i\}_{i=0}^{n} \) be a strictly increasing sequence of points such that \( a = x_0 < x_1 < x_2 .... < x_n = b \). Consider that \( f(x) \) is a continuous function over an interval \([a, b]\). A polynomial spline \( S(x) \) of degree \( m \), is an interpolant to \( f(x) \), such that

(i) \( S(x) \) is a polynomial of degree \( m \) over each subinterval \([x_i, x_{i+1}]\), \( i = 0, 1, 2, \ldots, n-1 \).

(ii) \( S(x_i) = f(x_i), \ i = 0, 1, 2, \ldots, n-1, n \).

(iii) \( S(x) \in C^{m-2}[a,b], \) i.e. \( S(x) \) and its successive derivatives of order \((m-2)\) are continuous over the entire interval \([a, b]\).

At this stage let us confine our interest for \( m = 3 \) i.e. to a cubic spline. Consider a function \( S(x) \) defined by

\[
S(x) = a_0 + b_0 (x - x_0) + \frac{1}{2} c_0 (x - x_0)^2 + \frac{1}{6} \sum_{k=0}^{n-1} d_k (x - x_k)^3
\]

...(2.3.1)

where \( n \) is the number of subintervals of \([a, b]\) and \( a_0, b_0, c_0, d_0, \ldots, d_{n-1} \) are \((n+3)\) unknowns. Here the power function \( (x - x_k)_+ \) has a property that
\[(x - x_k)_+ = x - x_k \quad \text{if } x > x_k\]
\[= 0 \quad \text{if } x \leq x_k\]

The equation (2.3.1) represents a cubic in each subinterval. Moreover, \(S(x)\) represents different cubic in distinct subintervals \([x_i, x_{i+1}]\) of the interval \([a, b]\). Being a cubic \(S(x)\), is such that

\[S(x) \in C^2 [a, b]\]

Thus equation (2.3.1) belongs to the class of cubic spline functions. Taking a general two point boundary value problem of second order

\[y''(x) + p(x)y'(x) + q(x)y(x) = r(x) \quad \text{for } a \leq x \leq b \quad \text{...(2.3.2)}\]

subject to the boundary conditions

\[\alpha_0 y_0 + \beta_0 y_0' = \gamma_0 \quad \text{at } x = a \quad \text{...(2.3.3)}\]
\[\alpha_n y_n - \beta_n y_n' = \gamma_n \quad \text{at } x = b \quad \text{...(2.3.4)}\]

Our aim is to interpolate this function \(y(x)\) by the cubic spline given in equation (2.3.1). The condition of interpolation suggests that

\[S(x_i) = y(x_i) \quad i=0,1,2,...,n \quad \text{...(2.3.5)}\]

The relation (2.3.5) and (2.3.1) with its subsequent expressions by virtue of equation (2.3.2), (2.3.3) and (2.3.4) result in a system of equations as given below:

\[\alpha_0 a_0 + \beta_0 b_0 = \gamma_0 \quad \text{at } x = a \quad \text{...(2.3.6)}\]
\[ c_0 + \sum_{k=0}^{n-1} d_k (x_i - x_k)_+ + p_i \left[ b_0 + c_0 (x - x_0) + \frac{1}{2} \sum_{k=0}^{n-1} d_k (x_i - x_k)^2 \right] \\
+ q_i \left[ a_0 + b_0 (x_i - x_0) + \frac{1}{2} c_0 (x_i - x_0)^2 + \frac{1}{6} \sum_{k=0}^{n-1} d_k (x_i - x_k)^3 \right] = r_i \]

...(2.3.7)

and at \( x = b \)

\[ \alpha_n \left[ a_0 + b_0 (x_n - x_0) + \frac{1}{2} c_0 (x_n - x_0)^2 + \frac{1}{6} \sum_{k=0}^{n-1} d_k (x_n - x_k)^3 \right] \\
- q \beta_n \left[ b_0 + c_0 (x_n - x_0) + \frac{1}{2} \sum_{k=0}^{n-1} d_k (x_n - x_k)^2 \right] = \gamma_n \]

...(2.3.8)

where \( p_i = p(x_i), \ q_i = q(x_i), \ r_i = r(x_i). \)

Equations (2.3.6), (2.3.7) with \( i = 0(1)n \) and (2.3.8) are \((n+3)\) linear algebraic equations in \((n+3)\) unknowns \( a_0, b_0, c_0, d_0, d_1, ..., d_{n-1}. \) The matrix form of this system is given by

\[ A X = B \quad \text{...(2.3.9)} \]

where \( X = [a_0, b_0, c_0, d_0, d_1, ..., d_{n-1}]^T \)

and \( B = [\gamma_0, r_0, r_1, r_2, ..., r_n, \gamma_n]^T \)

The coefficient matrix \( A \) is an upper triangular Hessenberg matrix with single lower subdiagonal, principal and upper diagonals having nonzero elements. Because of this nature of matrix \( A \), the determination of the required quantities becomes simple and less time consuming. The values of these constants ultimately yield the cubic spline \( S(x) \) in equation (2.3.1).
Since \( S(x) \) is an approximation to \( y(x) \), the error term \( e(x) \) in \( y(x) \) due to \( S(x) \) is defined as \( y(x) = S(x) + e(x) \). This leads to the transformation of equation (2.3.2) into
\[
S''(x) + e''(x) + p(x) S'(x) + e'(x) + q(x) S(x) + e(x) = r(x) \quad \text{...(2.3.10)}
\]
At \( x = x_i \) this is read as
\[
S''_i + p_i S'_i + q_i S_i = r_i - e''_i + p_i e'_i + q_i e_i \quad \text{...(2.3.11)}
\]
i = 0, 1, 2, ..., n.

The associated boundary conditions are
\[
\alpha_0 S_0 + \beta_0 S'_0 = \gamma_0 - [\alpha_0 e_0 + \beta_0 e'_0] \quad \text{...(2.3.12)}
\]
and
\[
\alpha_n S_n - \beta_n S'_n = \gamma_n - [\alpha_n e_n - \beta_n e'_n] \quad \text{...(2.3.13)}
\]
Now by virtue of definition of a cubic spline and interpolation conditions it is seen that
\[
S(x) = y(x) \quad \text{at } x = x_i \quad e(x) = 0 \quad \text{at } x = x_i
\]
\[
\Rightarrow e_i = 0 \quad \text{...(2.3.14)}
\]

The relations (2.3.1) and (2.3.11) provide
\[
c_0 + \sum_{k=0}^{n-1} d_k (x_i - x_k) + p_i \left[ b_0 + c_0 (x_i - x_0) + \frac{1}{2} \sum_{k=0}^{n-1} d_k (x_i - x_k)^2 \right]
\]
\[
+ q_i \left[ a_0 + b_0 (x_i - x_0) + \frac{1}{2} c_0 (x_i - x_0)^2 + \frac{1}{6} \sum_{k=0}^{n-1} d_k (x_i - x_k)^3 \right]
\]
\[
= r_i - [e''_i + p_i e'_i] \quad \text{...(2.3.15)}
\]
The associated boundary conditions (2.13.12) and (2.3.13) are now reduced to
\[
\alpha_0 a_0 + \beta_0 b_0 = \gamma_0 - \beta_0 e'_0 \quad \text{...(2.3.16)}
\]
and
\[
\alpha_n \left[ a_0 + b_0 (x_n - x_0) + \frac{1}{2} c_0 (x_n - x_0)^2 + \frac{1}{6} \sum_{k=0}^{n-1} d_k (x_n - x_k)^3 \right] \\
- \beta_n \left[ b_0 + c_0 (x_n - x_0) + \frac{1}{2} \sum_{k=0}^{n-1} d_k (x_n - x_k)^2 \right] = \gamma_n - \beta_n \varepsilon_n
\]

...(2.3.17)

This show that a better approximation of \( y(x) \) can be obtained by calculating constants \( a_0, b_0, c_0, d_0, d_1, ..., d_{n-1} \) with the help of equations (2.3.15) with \( i = 0 \) (1) \( n \), (2.3.16) and (2.3.17). It can be observed that the coefficient matrix is the same matrix A. Ahlberg et al \[1967\] have shown that if \( S^{(0)}(x) \) is any spline on a mesh \( \Delta \) and \( S(x) \) is a spline interpolating \( y(x) \) over \( \Delta \), then \( S(x) \) is the best approximations spline to \( y(x) \) meaning that

\[
\int_\Delta [y''(x) - S''(x)]^2 \, dx \leq \int_\Delta [y'(x) - S'^{(0)}(x)]^2 \, dx.
\]

Hence for an improvement is \( S(x) \) assume that

\[
S(x) = S^{(0)}(x) + \varepsilon(x)
\]

...(2.3.18)

so that \( \varepsilon(x) \) is also a cubic spline denoting the error term omitted while determining \( S(x) \) and

\[
y(x) = S^{(0)}(x) + \varepsilon(x) + e(x)
\]

...(2.3.19)

This transforms the equation (2.3.2) to the relation

\[
[S^{(0)}''(x) + \varepsilon''(x) + e''(x)] + p(x) [S^{(0)}''(x) + \varepsilon'(x) + e'(x)] \\
+ q(x) [S^{(0)}(x) + \varepsilon(x) + e(x)] = r(x)
\]

...(2.3.20)

But since \( [S^{(0)}''(x) + p(x) S^{(0)}'(x) + q(x) S^{(0)}(x) = r(x) \), one gets

\[
\varepsilon''(x) + p(x) \varepsilon(x) + q(x) \varepsilon(x) = -[e''(x) + p(x) e'(x) + q(x) e(x)]
\]

which at the nodes \( x = x_i \) becomes
\[ \varepsilon_i'' + p_i \varepsilon_i' + q_i \varepsilon_i = -[\varepsilon_i'' + p_i \varepsilon_i'] \quad \ldots(2.3.21) \]

and at the end points,
\[ \alpha_0 \varepsilon_0 + \beta_0 \varepsilon_0' = -\beta_0 \varepsilon_0' \text{ at } x = a \quad \ldots(2.3.22) \]

and
\[ \alpha_n \varepsilon_n - \beta_n \varepsilon_n' = -\beta_n \varepsilon_n' \text{ at } x = b \quad \ldots(2.3.23) \]

For the convenience in actual work out, the length of every subinterval \( x_{i+1} - x_i \) is taken as \( h \). The continuity of \( S'(x) \) and \( S''(x) \) leads to their comparison at \( x = x_i + \) and \( x = x_i - \). This gives
\[ 3(y_{i+1} - y_{i-1}) = h(S'_{i+1} + 4S'_i + S'_{i-1}) \quad \ldots(2.3.24) \]

\[ S''_i = \frac{6}{h^2} (S_{i+1} - S_i) - \frac{2}{h} (S'_i + S'_{i+1}) \quad \ldots(2.3.25) \]

and
\[ S'''_i = \frac{12}{h^3} (S_i - S_{i+1}) + \frac{6}{h^2} (S'_i + S'_{i+1}) \quad \ldots(2.3.26) \]

where
\[ x \in [x_i, x_{i+1}] \].

The equation (2.3.24) in the operator notation is written as
\[ 3(E - E^{-1}) y_i = h(E^{-1} + 4 + E) S'_i \]

where \( E = e^{hD} \)

Expanding this equation in powers of \( hD \), one gets
\[ S'_i = y'_i - \frac{1}{180} h^4 y_i^v + o(h^6) \quad \ldots(2.3.27) \]

\[ S''_i = y''_i - \frac{1}{12} h^2 y_i^{iv} + \frac{1}{360} h^4 y_i^{vi} + o(h^6) \quad \ldots(2.3.28) \]

\[ S'''_i = y'''_i + \frac{1}{2} hy_i^{iv} + \frac{1}{12} h^2 y_i^{vi} - \frac{1}{360} h^4 y_i^{vii} \]
\[ -\frac{1}{1440} h^5 y_i^{viii} + o(h^6) \quad \ldots(2.3.29) \]
where \( S'(x_+)=S'(x) \) at \( x = x_i \) for \( x \in [x_{i-1}, x_i] \) and \( S'(x_-)=S'(x) \) at \( x = x_i \) for \( x \in [x_i, x_{i+1}] \). Similarly obtaining \( S''_i^- \) and subtracting it from equation (2.3.29) it is found that

\[
S''_{i+} - S''_{i-} = h y_i^{iv} - \frac{1}{720} h^5 y_i^{vii} + o(h^7) \quad \cdots (2.3.30)
\]

Now the relation (2.3.1) gives

\[
S''_{i+} - S''_{i-} = d_i, \quad i=1(1)n-1
\]

hence from (2.3.30)

\[
h y_i^{iv} + o(h^7) = d_i, \quad i=1(1)n-1 \quad \cdots (2.3.32)
\]

Also the equations (2.3.27) and (2.3.28) give \( e_i' \) and \( e_i'' \) as

\[
e_i' = \frac{1}{180} h^4 y_i^v + o(h^6) \quad \cdots (2.3.33)
\]

\[
e_i'' = \frac{1}{12} h^2 y_i^{iv} - \frac{1}{360} h^4 y_i^{vi} + o(h^6) \quad \cdots (2.3.34)
\]

Ultimately the expression on the right hand side of equation (2.3.21) simplifies to

\[
e_i'' + p_i e_i' + q_i e_i = - \frac{1}{12} h^2 y_i^{iv} + \frac{1}{360} h^4 y_i^{vi} - \frac{1}{180} p_i h^4 y_i^v + o(h^6)
\]

\[
= - \frac{1}{12} h d_i + o(h^4)
\]

\[
\cdots (2.3.35)
\]

for \( i = 1, 2, 3, \ldots, (n-1) \)

This relation gives equations at the internal nodes only. For the end points, using linear extrapolation, at \( x = a, \)

\[
- \frac{1}{12} h^2 y_0^{iv} = - \frac{1}{12} h^2 (2y_i^{iv} - y_0^{iv}) + \frac{1}{12} h^4 y_0^{vi}
\]

\[
= - \frac{1}{12} h (2d_1 - d_0) + o(h^4)
\]

\[
\cdots (2.3.36)
\]
and similarly at $x = b$,
\[
- \frac{1}{12} h^2 y_n^{iv} = - \frac{1}{12} h (2d_{n-1} - d_{n-2}) \quad \ldots (2.3.37)
\]

The equations (2.3.22) and (2.3.23) with the aid of equations (2.3.36) and (2.3.37) are now written as
\[
\alpha_0 \epsilon_0 + \beta_0 \epsilon_0' = - \frac{1}{180} h^4 p_0 y_0^\nu + o(h^6)
\]
and
\[
\alpha_n \epsilon_n - \beta_n \epsilon_n' = - \frac{1}{180} h^4 p_0 y_0^\nu + o(h^6)
\]

The above equation in light of a cubic spline $(x)$ and are reduced to
\[
\alpha_0 \epsilon_0 + \beta_0 \epsilon_0' = 0 \quad \ldots (2.3.38)
\]
\[
\alpha_n \epsilon_n - \beta_n \epsilon_n' = 0 \quad \ldots (2.3.39)
\]

The equations (2.3.35), (2.3.38) and (2.3.39) together constitute a system of $(n+3)$ equations having $(n+3)$ unknowns $\alpha_0, \beta_0, c_0, d_0, d_1, \ldots, d_{n-1}$ to be determined. However, these constants are related to a cubic spline $\epsilon(x)$, an error in $S(x)$, so computing these values $\epsilon(x)$ is obtained in a form similar to the function $S(x)$ by equation (2.3.1). This determines a corrected spline $S(x)$ defined by $S^{(0)}(x) + \epsilon(x)$. A very little additional computation is required in getting $\epsilon(x)$ as the coefficient matrix for the system of equations remains the same by putting equations (2.3.35), (2.3.36), (2.3.37), (2.3.38) and (2.3.39) together as shown
\[ \varepsilon_i'' + \beta_i \varepsilon_i' + \gamma_i \varepsilon_i = \begin{cases} \frac{-h}{12} (2d_i - d_0) & i = 0 \\ \frac{-h}{12} d_i & i = 1(1)n - 1 \\ \frac{-h}{12} (d_{n-1} - d_{n-2}) & i = n \end{cases} \]

...(2.3.40)

However, the variation only in right hand side of equation (2.3.9) occurs. The iterations for \( \varepsilon(x) \) are continued till the desired accuracy is achieved or two successive approximate values agree reasonably.

The case of non-uniform mesh spacing is also possible to be tackled. It is found to be similar to the manner presented by Curtis et al [1965]. However, this case is dealt with and explained by Fyfe [1970]. Also the actual numerical experience does not recommend for taking unequal partitions of the domain as it has not been found fruitful compared to labour and time consumed for determining the desired spline through equal intervals.

### 2.4 SPLINE IN TERMS OF DERIVATIVES:

While discussing the meaning of mathematical splines it was mentioned that spline theory has an analogy with beam theory in engineering. Hence in the similar fashion, referring to the second derivatives of spline functions as moments Ahlberg et al [1967] represented a cubic spline in terms of its corresponding moments. Use of spline function so expressed to interpolate a function over a given set of data is suggested in this comprehensive collocation of literature about spline theory. Albasiny et al [1971] tried to
solve a linear two point boundary value problem by using such type of cubic splines.

Here for the first instance, let us derive this new approach of a cubic spline which is different from that expressed in terms of truncated power function. Consider a linear two point value boundary value problem

\[ y''(x) + p(x)y'(x) + q(x)y(x) = r(x) \quad \text{...(2.4.1)} \]

for \( a \leq x \leq b \)

subject to the boundary conditions

\[ G_1 \left[ y(a), y'(a) \right] = 0 \quad \text{at} \quad x = a \quad \text{...(2.4.2)} \]
\[ G_2 \left[ y(b), y'(b) \right] = 0 \quad \text{at} \quad x = b \quad \text{...(2.4.3)} \]

Since \( S(x) \) is a cubic spline interpolating \( Y(x) \) given by equation (2.4.1) we have \( S(x_i) = y(x_i) \) and also \( S''(x) \) is a linear function. Let us define this \( S''(x) \) in the subinterval \([x_i, x_{i+1}]\) of \([a, b]\) as follows:

\[ S''(x) = y''_{i+1} \frac{x - x_i}{h_{i+1}} + y''_i \frac{x_{i+1} - x}{h_i} \quad \text{...(2.4.4)} \]

where \( h_{i+1} = x_{i+1} - x_i \). Two successive integrations produce a cubic \( S(x) \) in \([x_i, x_{i+1}]\) which has the form

\[ S(x) = y''_{i+1} \frac{(x - x_i)^3}{6h_{i+1}} + y''_i \frac{(x_{i+1} - x)^3}{6h_i} + A(x_{i+1} - x) \frac{1}{h_{i+1}} + B(x - x_i) \frac{1}{h_{i+1}} \quad \text{...(2.4.5)} \]

with \( A \) and \( B \) as constants to be determined. Interpolation condition provides their evaluation. They are

\[ A = y_i - \frac{h_{i+1}^2}{6} y''_i \quad \text{and} \quad B = Y_{i+1} - \frac{h_i^2}{6} Y''_{i+1}. \]
Equation (2.4.5) is given due to these constants as follows:

\[
S(x) = y''_{i+1} \frac{(x - x_i)^3}{6h_{i+1}} + y''_i \frac{(x_{i+1} - x_i)^3}{6h_i} + (y_i - \frac{h_{i+1}^2}{6} y''_i) \frac{(x_{i+1} - x_i)}{h_{i+1}} \\
+ (y_{i+1} - \frac{h_i^2}{6} y''_{i+1}) \frac{(x - x_i)}{h_{i+1}}
\]

\[\ldots(2.4.6)\]

Similarly, \( S(x) \) can be found in the interval \([x_{i-1}, x_i]\) as

\[
S(x) = y_{i-1}'' \frac{(x - x_{i-1})^3}{6h_i} + y_{i-1}'' \frac{(x_{i} - x)_{i}}{6h_i} + (y_i - \frac{h_i^2}{6} y_{i-1}''_i) \frac{(x_{i} - x)}{h_i} \\
+ (y_{i-1} - \frac{h_{i-1}^2}{6} y''_{i-1}) \frac{(x - x_{i-1})}{h_i}
\]

\[\ldots(2.4.7)\]

Deriving \( S'(x) \) at \( x = x_i \) from the equation (2.4.6) which is denoted by \( S'(x_i+) \), we have

\[
S'(x_i+) = -\frac{h_{i+1}}{3} y''_i - \frac{h_{i+1}}{6} y''_{i+1} + \frac{y_{i+1} - y_i}{h_{i+1}} \]

\[i = 0, 1, 2, \ldots, n-1\]

and in the same way,

\[
S'(x_i-) = \frac{h_i}{6} y''_{i-1} + \frac{h_i}{3} y''_i + \frac{y_i - y_{i-1}}{h_i} \]

\[i = 0, 1, 2, \ldots, n\]

Continuity of \( S''(x) \) at \( x = x_i \) requires that,

\[
S''(x_i+) = S''(x_i-)
\]

so that

\[
h_i y_{i+1} - (h_i + h_{i+1}) y_i + h_{i+1} y_{i-1} \\
= h_i h_{i+1} (\frac{h_i}{6} y''_{i+1} + \frac{(h_i + h_{i+1})}{3} y''_i + \frac{h_{i+1}}{6} y''_{i+1})
\]

\[\ldots(2.4.10)\]

\[i = 1, 2, 3, \ldots, n-1.\]
This equation gives a system of \((n-1)\) equations in \((n+1)\) variables \(y_i, \quad i = 0, 1, 2, \ldots, n\) to be determined. Hence for evaluating these quantities two more conditions are required so that the system of equations due to (2.4.10) is complete. Ahlberg et al [1967] have shown that the moments \(y_i'', i = 0, 1, 2, \ldots, n\) can be obtained from relation (2.4.10) if a curve is initially fitted to the data and two extra conditions are encountered to complete the system (2.4.10).

Here our objective is to solve a differential equation (2.4.1) with the help of relations (2.4.10). Let us express the equation (2.4.1) in the form

\[
y''(x) = f(x, y, y')
\]  \hspace{1cm} ...(2.4.11)

subject to the boundary conditions (2.4.2) and (2.4.3). From these boundary conditions, it is seen that there are four pairs of boundary conditions as possible combinations Viz.

(i) \(y(a) = K; \ y(b) = L\),

(ii) \(y(a) = K; \ y'(b) = L\),

(iii) \(y'(a) = K; \ y(b) = L\)

and

(iv) \(y'(a) = K; \ y'(b) = L\).

The relation (2.4.10) will take the form for case (i),
\[
\begin{aligned}
- y_1(h_1 + h_2) + h_1 y_2 &= \frac{h_1 h_2}{6} \left( \frac{h_1}{6} y_0'' + \frac{h_1 + h_2}{3} y_1'' + \frac{h_2}{6} y_2'' \right) - h_1 y_0 \\

h_i y_{i+1} - (h_i + h_{i+1}) y_i + h_{i+1} y_{i+1} &= \frac{h_i h_{i+1}}{6} \left( \frac{h_i}{6} y_i'' + \frac{h_i + h_{i+1}}{3} y_i'' \right) \\
&\quad \text{for } i = 2, 3, \ldots, n-1 + \frac{h_{i+1}}{6} y_{i+1}'

- (h_n + h_{n-1}) y_{n-1} + h_n y_{n-2} &= \frac{h_n h_{n+1}}{6} \left( \frac{h_n}{6} y_{n-2}'' + \frac{h_{n-1} + h_n}{3} y_{n-1}'' \right) \\
&\quad + \frac{h_{n-1}}{6} y_{n-1}' - h_n y_n
\end{aligned}
\]

\(...(2.4.12)\)

Case (iv),
\[
\begin{aligned}
y_0 - y_1 &= - \frac{h_1^2}{6} (2y_0'' + y_1'') - h_1 y_0' \\

h_i y_{i+1} - (h_i + h_{i+1}) y_i + h_{i+1} y_{i+1} &= \frac{h_i h_{i+1}}{6} \left( \frac{h_i}{6} y_i'' + \frac{h_i + h_{i+1}}{3} y_i'' \right) \\
&\quad + \frac{h_{i+1}}{6} y_i'' \\
&\quad \text{for } i = 2, 3, \ldots, n-1

y_n - y_{n-1} &= \frac{h_n^2}{6} (y_{n-1}'' + 2y_n'') + \frac{6}{h_n} y_n'
\end{aligned}
\]

\(...(2.4.13)\)

and similarly we will be able to deal with remaining cases. In these remaining cases either first equation or the last equation of system (2.4.12) or (2.4.13) will be there. It is observed for any case that on the left hand side the equations are obtained for which the coefficient matrix, for in the matrix form is upper tri-diagonal one.

Now in the process for obtaining a solution to equation (2.4.1) we start with a straight line \( Y(x) = mx + c \) passing
through the boundary points. Considering this straight line as an initial guess, the moments $y_i''$ are calculated from the relation (2.4.11) at the nodal points $x = x_i$ as

$$y_i'' = f(x_i, y_i, y'_i) \quad \ldots(2.4.14)$$

for $i = 0, 1, 2, \ldots, n$. These moments are now utilized to evaluate $y_i, i = 0, 1, 2, \ldots, n$ through the relation (2.4.10) along with two additional equations sought from boundary conditions (2.4.2) and (2.4.3). This can be furnished by solving merely a tri-diagonal system of equations. The results so obtained can again be improved by continuing the same processes till the desired solutions are found or two successive iterations produce the same values.

Before discussing an error analysis for such cubic spline approximation of $y(x)$ given by equation (2.4.1) let us consider that all the nodal points are equally spaced. This assumption leads to equal partitioning of the domain $[a, b]$, say $h_{i+1} = h_i = h$ is the length of each partition. In such a situation relation (2.4.10) simplifies to

$$y_{i-1} - 2y_i + y_{i+1} = \frac{h^2}{6} (y_i'' + y_{i-1}'' + y_{i+1}'') \quad \ldots(2.4.15)$$

Also above equation can be written as

$$S''(x_{i-1}) + 4S''(x_i) + S''(x_{i+1}) = \frac{6}{h^2}[f(x_{i-1}) - 2f(x_i) + f(x_{i+1})]$$

$$\ldots(2.4.15a)$$

and the consequently the derivatives given in equations (2.4.8) and (2.4.9) correspond to
\[ S'_i = \frac{y_{i+1} - y_i}{h} - \frac{h}{6} (2y_i'' + y_{i+1}'') \quad \text{ ...(2.4.16)} \]

\[ i = 0, 1, 2, \ldots, n-1 \]

and

\[ S'_{i-1} = \frac{y_i - y_{i-1}}{h} + \frac{h}{6} (2y_i'' + y_{i-1}'') \quad \text{ ...(2.4.17)} \]

\[ i = 0, 1, 2, \ldots, n \]

respectively. Define an error term \( e(x) \) in \( y(x) \) due to approximate function \( S(x) \) by

\[ e(x) = y(x) - S(x) \quad \text{ ...(2.4.18)} \]

In particular, when \( y'(x) \) is absent \( p(x) = 0 \), equation (2.1.1) will take the form

\[ y''(x) + q(x) y(x) = r(x) \quad \text{ ...(2.4.19)} \]

Now as \( S(x) \) interpolates \( y(x) \) at the nodes

\[ S(x_i) = y(x_i) \Rightarrow e(x_i) = e_i = 0 \quad \text{ ...(2.4.20)} \]

So that from equations (2.4.18) and (2.4.19)

\[ y''_i + q_i y_i = r_i \Rightarrow S''_i = e''_i + q_i (S_i + e_i) = r_i \quad \text{ ...(2.4.21)} \]

\[ \Rightarrow S''_i = r_i - q_i S_i - e''_i \]

Keeping in mind that \( y_i'' = S''_i \) and eliminating \( S''_i \) one gets

\[ \frac{h^2}{6} (r_{i-1} - q_{i-1} S_{i-1}) + 4 (r_i - q_i S_i) + (r_{i+1} - q_{i+1} S_{i+1}) \]

\[ - \frac{h^2}{6} (e''_{i-1} + 4 e''_i + e''_{i+1}) = y_{i+1} - 2y_i + y_{i-1} \]

i.e

\[ (1 + \frac{1}{6} h^2 q_{i-1}) y_{i-1} - (2 - \frac{2}{3} h^2 q_i) y_i + (1 + \frac{1}{6} h^2 q_{i+1}) y_{i+1} \]

\[ = \frac{1}{6} h^2 (r_{i-1} + 4 r_i + r_{i+1}) - \frac{1}{6} h^2 (e''_{i-1} + 4 e''_i + e''_{i+1}) \quad \text{ ...(2.4.22)} \]
Hence an approximation \( S_i^{(0)} = S^{(0)}(x_i) \) to \( S_i \) can be obtained by solving equation (2.4.22) with the error terms omitted. The coefficients matrix for the system of these equations is a tri-diagonal. It determines \( S^{(0)}(x) \) at the knots. The quantities \( y_i^{(0)*} \) can be calculated from

\[
y_i^{(0)*} = r_i - q_i S_i^{(0)} \tag{2.4.23}
\]

in consequence of relation (2.4.1) with \( p(x) = 0 \). Complete cubic spline approximation \( S^{(0)}(x) \) thus can be obtained from equation (2.4.6) with equally spaced knots in \([x_i, x_{i+1}]\) as below:

\[
S(x) = y''_i \frac{(x_{i+1} - x)^3}{6h} + y''_{i-1} \frac{(x - x_i)^3}{6h} + \left( y''_i - \frac{1}{6} h^2 y_i'' \right) \left( \frac{x_{i+1} - x}{h} \right) \\
+ \left( y''_{i-1} - \frac{1}{6} h^2 y_{i-1}'' \right) \left( \frac{x - x_i}{h} \right) \\
\tag{2.4.24}
\]

in which \( y_i'' \) and \( y_i \) are to be considered as \( y_i^{(0)*} \) and \( y_i^{(0)} \) respectively. A better approximation can be obtained if the error terms are encountered. For this let us introduce

\[
E_i = \frac{1}{6} h^2 (e''_{i-1} + 4 e''_i + e''_{i+1})
\]

Hence by virtue of equations (2.3.23) and (2.3.24) this becomes
\[ E_i = \frac{1}{6} h^2 \left( \frac{1}{12} h^2 y_{i-1}^{iv} - \frac{4}{12} h^2 y_{i}^{iv} + \frac{1}{12} h^2 y_{i+1}^{iv} \right) + o(h^6) \]
\[ = \frac{h^3}{72} \left( d_{i-1} + 4d_i + d_{i+1} \right) + o(h^6) \]
\[ \text{for} \quad i = 1, 2, 3, \ldots, n-1 \]

...(2.4.25)

and at the end points
\[ E_1 = \frac{1}{72} h^3 6d_1 + o(h^6) \quad \text{and} \quad E_{n-1} = \frac{1}{72} h^3 6d_{n-1} + o(h^6). \]

Now the second difference of \( y_i'' \) is given by
\[ y_{i-1}'' - 2y_i'' + y_{i+1}'' = h^2 y_i^{iv} \]
so that
\[ y_{i-1}'' - 2y_i'' + y_{i+1}'' = h d_i \]
\[ \quad \text{for} \quad i = 1, 2, 3, \ldots, n-1 \]

...(2.4.26)

\[ E_i = \frac{1}{12} h^2 (y_0'' - 2y_i'' + y_n'') + o(h^6) \]
\[ \quad \text{for} \quad i = 1, 2, 3, \ldots, n-1 \]

...(2.4.27)

and
\[ E_{n-1} = \frac{1}{12} h^2 (y_{n-2}'' - 2y_{n-1}'' + y_n'') \]
\[ \quad \text{for} \quad i = 1, 2, 3, \ldots, n-1 \]

...(2.4.28)

Letting \( S(x) = S^{(0)}(x) + \varepsilon(x) \), an approximation \( \widetilde{\varepsilon}(x) \) of \( \varepsilon(x) \) can be determined by finding \( \widetilde{\varepsilon}_i \) which satisfies
\[ (1 + \frac{1}{6} h^2 q_{i+1})\widetilde{\varepsilon}_{i+1} - \left( 2 - \frac{2}{3} h^2 q_i \right)\widetilde{\varepsilon}_i + (1 + \frac{1}{6} h^2 q_{i-1})\widetilde{\varepsilon}_{i-1} = -E_i \]
\[ \quad \text{for} \quad i = 1, 2, 3, \ldots, n-1 \]

...(2.4.29)

Substituting \( S^{(0)}_i \) and \( y_i^{(0)} \) into equations (2.4.25), (2.4.27) and (2.4.28), where the terms of \( o(h^6) \) are ignored, the quantities \( \widetilde{\varepsilon}_i \) can be approximated. The coefficient matrix for this system of equations is the same as that is
found for system (2.4.22) and an additional computation required for evaluation of \( \tilde{\varepsilon}_i \) is sufficiently less. Then the quantities

\[
\tilde{S}_i = \tilde{S}_i^{(0)} + \tilde{\varepsilon}_i
\]

and hence \( \tilde{y}_i'' \) can be found from the relation (2.4.23) in the form

\[
y_i'' = r_i - q_i \tilde{S}_i - \frac{1}{12} h d_i^{(0)} \quad i = 1, 2, 3, \ldots, mn-1
\]

\[
y_0'' = r_0 - q_0 \tilde{S}_0 - \frac{1}{12} h (2d_1^{(0)} - d_2^{(0)})
\]

\[
y_n'' = r_n - q_n \tilde{S}_n - \frac{1}{12} h (2d_{n-1}^{(0)} - d_{n-2}^{(0)})
\]

The complete spline \( S(x) \) can now be derived from equation (2.4.24) with \( S_i \) and \( y_i'' \) replaced by \( \tilde{S}_i \) and \( \tilde{y}_i'' \) respectively.

We now turn to the general case of equation (2.4.1) where the term \( y'(x) \) is present. The equations (2.4.16) and (2.4.17) are substituted in equation (2.4.1) which gives

\[
y_i'' + e_i'' + p_i \left( \frac{h}{3} y_i'' - \frac{h}{6} y_{i+1}'' + \frac{S_i - S_{i+1}}{h} + e_i \right) + q_i S_i = r_i
\]

\[\ldots(2.4.30)\]

\[i = 1, 2, 3, \ldots, n\]

and

\[
y_i'' + e_i'' + p_i \left( -\frac{h}{3} y_i'' - \frac{h}{6} y_{i-1}'' + \frac{S_{i-1} - S_i}{h} + e_i \right) + q_i S_i = r_i
\]

\[\ldots(2.4.31)\]

Using equation (2.4.30) with \( i \) replaced by \( i-1 \), \( y_i'' \) is found to be
\[ A_i y''_i = \left(1 - \frac{h}{3} p_{i+1}\right) \left[(r_i - q_i S_i) - (e''_i + p_i e'_i)\right] \]
\[ - \frac{h}{6} p_i \left[(r_{i+1} - q_{i+1} S_{i+1}) - (e''_{i+1} + p_{i+1} e'_{i+1})\right] \]
\[ - \frac{1}{h} p_i \left(1 - \frac{h}{2} p_{i+1}\right) (S_i - S_{i+1}) \]

\[ \ldots(2.4.32) \]

where \[ A_i = \left(1 + \frac{h}{3} p_i\right) \left(1 - \frac{h}{3} p_{i-1}\right) + \frac{1}{36} h^2 p_i p_{i+1} \cdot \]

Also in the equation (2.4.31) if \( i \) is replaced by \( i + 1 \) then,

\[ B_i y''_i = \frac{h}{6} p_i \left[(r_{i+1} - q_{i+1} S_{i+1}) - (e''_{i+1} + p_{i+1} e'_{i+1})\right] \]
\[ + \left(1 + \frac{h}{3} p_{i+1}\right) \left[(r_i - q_i S_i) - (e''_i + p_i e'_i)\right] \]
\[ - \frac{1}{h} p_i \left(1 + \frac{h}{2} p_{i+1}\right) (S_{i+1} - S_i) \]

\[ \ldots(2.4.33) \]

where \[ B_i = \left(1 + \frac{h}{3} p_{i+1}\right) \left(1 - \frac{h}{3} p_i\right) + \frac{1}{36} h^2 p_i p_{i+1} = A_{i+1}. \]

Multiplying equations (2.4.32) by \( B_i \) and (2.4.33) by \( A_i \) and then equating them, one gets

\[ - \left(1 - \frac{h}{2} p_{i-1} + \frac{1}{6} h^2 q_{i-1}\right) B_i S_{i-1} + \left[(1 + \frac{h}{2} p_{i+1}) A_i + \left(1 - \frac{h}{2} p_{i-1}\right) B_i - \frac{2}{3} h^2 q_i C_i\right] S_i \]
\[ - \left(1 + \frac{h}{2} p_{i+1} + \frac{1}{6} h^2 q_{i+1}\right) A_i S_{i+1} \]
\[ = - \frac{1}{6} h^2 \left(B_i r_{i-1} + 4 C_i r_i + A_i r_{i+1}\right) - D_i \]

\[ \ldots(2.4.34) \]

\[ i = 1, 2, 3, \ldots, n-1. \]

where \[ C_i = 1 + \frac{h}{24} (p_{i+1} - p_{i-1}) - \frac{1}{12} h^2 p_{i-1} p_{i+1}. \]
Now due to equation (2.4.16) the boundary conditions expressed by a two term recurrence relation, namely, $\alpha_0 y_0 + \beta_0 y'_0 = \gamma_0$ gets transformed into

$$\alpha_0 S_0 + \beta_0 \left[ -\frac{h}{3} y''_0 - \frac{h}{6} y''_i + \frac{S_i - S_0}{h} + c'_0 \right] = \gamma_0 \quad \text{...(2.4.35)}$$

If we use equation (2.4.33) with $i = 0$ to eliminate $y''_0$ and equation (2.4.32) with $i = 0$ to eliminate $y''_i$, the relation (2.4.35) is left with $S_0$ and $S_1$ only. In a similar way, the second end condition can be expressed with the terms $S_{n-1}$ and $S_n$. Relation (2.4.34) along with these two term equations for boundary conditions constitutes a tridiagonal system of $(n+1)$ equations in $(n+1)$ unknowns $S_0, S_1, ..., S_n$. A cubic spline $S^{(0)}(x)$ can be determined through these quantities, where the error terms $E_i$ ignored. With a similar argument, here also $E_i, i=1, 2, ..., n-1$ are given by

$$E_i = \begin{cases} 
\frac{1}{72} h^3 (B_i d_{i-1} + 4C_i d_{i-1} + A_i d_{i+1}) + o(h^6) & i = 1, 2, 3, ..., n-2 \\
\frac{1}{72} h^3 \left[ (2B_i - C_i) d_i + (A_i - B_i) d_2 \right] + o(h^6) & i = 1 \\
\frac{1}{72} h^3 \left[ (B_{n-1} - A_{n-1}) d_{n-2} + (3A_{n-1} - 2B_{n-1}) d_{n-1} \right] + o(h^6) & i = n-1 
\end{cases} \quad \text{...(2.4.36)}$$
The values of \(d_i\) can be obtained from the relation (2.4.26). An approximation of \(E_i\) is given by substituting \(y'_i(0)\) into equation (2.4.26) and using \(d_i\) so obtained into equation (2.4.36) with \(o(h^6)\) terms omitted. A better approximation to \(S(x)\) is \(S^{(0)}(x) + \tilde{\epsilon}(x)\) where \(\tilde{\epsilon}(x)\) is a cubic spline defined by

\[
\begin{align*}
- (1 - \frac{h}{2} p_{i-1} + \frac{1}{6} h^2 q_{i-1}) B_i \tilde{\epsilon}_{i-1} \\
+ \left[ (1 + \frac{h}{2} p_{i+1}) A_i - (1 - \frac{h}{2} p_{i+1}) B_i - \frac{2}{3} h^2 q_i C_i \right] \tilde{\epsilon}_i \\
- (1 + \frac{h}{2} p_{i+1} - \frac{1}{6} h^2 q_{i+1}) A_i \tilde{\epsilon}_{i+1} = -E_i \\
i = 1, 2, 3, ..., n-1
\end{align*}
\]

...(2.4.37)

and the two equations involving \(\tilde{\epsilon}_0, \tilde{\epsilon}_1, \tilde{\epsilon}_{n-1}\) and \(\tilde{\epsilon}_n\) derived from the boundary conditions. Again as before the calculation of \(\tilde{\epsilon}_i\)'s involves the same tri-diagonal system of equations which was for \(S^{(0)}(x)\). Hence a complete spline \(S(x) = S^{(0)}(x) + \tilde{\epsilon}(x)\) is found by calculating \(y'_i\) from the relation (2.4.32) or (2.4.33) and relation (2.4.24) for \(S(x)\), in which the quantities \(y'_i\) and \(y_i\) are referred to as \(\tilde{y}'_i\) and \(\tilde{y}_i\) respectively.

### 2.5 COMPARISON WITH FINITE DIFFERENCE SCHEME:

Let us assume that a parabola \(y(x)\) passes through the points \(x_{i-1}, x_i\) and \(x_{i+1}\) say,

\[
y(x) = a(x - x_i)^2 + b(x - x_i)^2 + c
\]

...(2.5.1)
Then
\[ y(x) = \frac{y_{i+1} - 2y_i + y_{i-1}}{2h^2} (x - x_i)^2 + \frac{(y_{i+1} - y_{i-1})}{2h} (x - x_i) + y_i \]
\[ \quad \ldots(2.5.2) \]

where \( h \) is the length of an interval \([x_i, x_{i+1}]\).

Hence,
\[ y'(x) = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} (x - x_i) + \frac{y_{i+1} - y_{i-1}}{2h} \]
\[ \quad \ldots(2.5.3) \]

and
\[ y''(x) = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} \]
\[ \quad \ldots(2.5.4) \]

Equations (2.5.3) and (2.5.4) are given at the points \( x = x_i \) by
\[ y'(x_i) = y'_i = \frac{y_{i+1} - y_{i-1}}{2h} \]
\[ \quad \ldots(2.5.5) \]

and
\[ y''(x_i) = y''_i = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} \]
\[ \quad \ldots(2.5.6) \]

Now in the finite difference scheme the expression \( y_{i+1} - 2y_i + y_{i-1} \) is known as a second difference of \( y_i \). Here in equation (2.5.6) the same expression is approximately equal to the second derivative of \( y(x) \) at \( x = x_i \).

Mathematically,
\[ y_{i+1} - 2y_i + y_{i-1} = h^2 y''_i \]

Now for a cubic spline function, a recurrence relation for equally spaced
\[ y_{i+1} - 2y_i + y_{i-1} = \frac{h^2}{6} (y''_{i-1} + 4y''_i + y''_{i+1}) \]
\[ \quad \ldots(2.5.7) \]

Thus it is seen that
\[ y''_i = \frac{h^2}{6} (y''_{i-1} + 4y''_i + y''_{i+1}) \]
\[ \quad \ldots(2.5.8) \]
Hence, a second derivative of \( y(x) \) at the point \( x = x_i \) in the finite difference scheme is approximated by a relation (2.5.7) if spline approximation is employed. Now from the relation (2.5.7) and (2.4.14) the following relation is derived:

\[
f(X_i, y_i, y'_i) = \frac{h^2}{6} (y''_{i} + 4y''_{i+1} + y''_{i+2}) \quad \ldots(2.5.9)
\]

and now by virtue of equation (2.4.1) we have the relation given below:

\[
y_{i+1} - 2y_i + y_{i-1} + p_i \frac{y_{i+1} - y_{i-1}}{2h} + q_i y_i
\]

\[
= \frac{h^2}{6} (y''_{i} + 4y''_{i+1} + y''_{i+2}) + p_i y'_i + q_i y_i
\]

so that

\[
y_{i+1}(1 + \frac{1}{2h}p_i) - y_i(2 + q_i) + y_{i-1}(1 - \frac{1}{2h}p_i)
\]

\[
= \frac{h^2}{6} (y''_{i} + 4y''_{i+1} + y''_{i+2}) + p_i y'_i + q_i y_i
\]

\[
\ldots(2.5.10)
\]

In order to determine \( y_i \)'s it is seen that a tri-diagonal system of equations in \( y_i \)'s is obtained in the method of finite difference as well as spline collocation. Thus spline collocation method has one type of analogy with finite difference technique. The relation (2.5.10) is regarded as a linearized version of equation (2.5.7) for a cubic spline approximation of the solution to a differential equation (2.4.1). The effects of this linearized version in the convergence of spline approximations will be discussed later on by taking its actual application to a problem.
2.6 INTRODUCTION OF FINITE ELEMENT METHOD:

The finite element method (FEM) is a numerical method for finding approximate solutions of PDE as well as of integral equations. The solution approach is based either on eliminating the differential equation. Completely (steady state problem), or rendering the PDE into an approximating system of ordinary differential equations, which are then numerically integrated using standard techniques such as Euler’s method, Runge-Kutta, etc.

The basic idea behind the finite element method is to replace a continuous function by means of piecewise polynomials. Such an approximation, called the piecewise polynomial approximation. We know the importance of polynomial approximations in numerical analysis. Theses are used in numerical solution of practical problems where the exact functions are difficult to obtain. The idea of piecewise polynomial approximation is also not new.

In engineering applications, several approximate methods of solution are used. e.g. the method of least squares, the method of collocation etc. Here we discuss two important methods of approximation, Viz the Rayleigh-Ritz method and the Galerkin technique. Rayleigh developed the method to solve certain vibration problems and Ritz provided a mathematical basis for it and also applied it to more general problems. Whereas the Rayleigh-Ritz method is based on the existence of a functional, the Galerkin technique uses the governing equation of the problem and minimizes the error of the approximate solution. A
disadvantage of both methods is that higher order polynomials have to be used to obtain reasonable accuracy.

The FEM is one of the most important numerical applications of Rayleigh-Ritz and Galerkin methods. Its mathematical software is quite popular and used extensively in the solution of many practical problems of engineering and applied science. In FEM, the domain of integration is subdivided into a number of smaller regions called elements and over each of these elements the continuous function is approximated by a suitable piecewise polynomial. To obtain a better approximation one need not use higher order polynomials but only use a finer subdivision i.e. increases the number of elements.

Several types of elements are in use, the type used being largely dependent upon the geometrical shape of the region under consideration. In two-dimensional problems the elements used are triangles, rectangular and quadrilaterals. For three dimensional problems tetrahedral, hexahedra and parallelepiped elements are used. We are discusses here the rectangular element in the solution of two dimensional problems. The typical finite elements are shown in fig (2.6.1)
2.7 HISTORY OF FINITE ELEMENT METHOD:

The finite element method originated from the need for solving complex elasticity and structural analysis problems in Civil and Aeronautical engineering. The idea of representing a given domain as a collocation of discrete elements is not novel with finite element method. It was recorded that ancient mathematicians estimated the value of $\pi$ by noting that the perimeter of a polygon inscribed in a circle approximate the circumference of a circle. They predicted the value of $\pi$ to accuracies of almost 40 significant digits by representing the circle as a polygon of a finitely large number of sides. In modern times the idea found a home in
aircraft structural analysis, where, for example, wings and fuselages are treated as assemblages of stringers, skins, and shear panels. Hrenikoff [1941] introduced the so-called framework method, in which a plane elastic medium was represented as a collection of bars and beems. The use of piecewise continuous functions define over a subdomain to approximate the unknown function dates back to the work of Courant[1943], who used an assemblage of triangular elements and the principle of minimum potential energy to study the St. Venant torsion problem. Although certain key features of the finite element method can be found in the works of Hrenikoff [1941] and Courant[1943], the formal presentation of the finite element method is attributed to Argyris and Kelsey [1960] and to Turner, Clough, Martin, and TOPP [1956]. Courant’s contribution was evolutionary drawing on a large of earlier result for Partial differential equations (PDEs) developed by Rayleigh, Ritz and Galerkin.

However, the term “finite element” was first used by Clough in 1960. Since its inception, the literature on finite element applications has grow exponentially, and today there are numerous journals which are primarily devoted to the theory and application of the finite element method. By late 1950’s, the key concepts of stiffness matrix and element assembly existed essentially in the form used today. NASA issued request for proposals for the development of the finite element software NASTRAN in 1965. The method was provided with a rigorous mathematical foundation in 1973 with the publication of Strang and Fix’s An Analysis of the
finite element method has since been generalized into branch of applied mathematics for numerical modeling of physical system in a wide variety of engineering disciplines, e.g. electromagnetism and fluid dynamics.

2.8 THE BASIC CONCEPT OF THE FINITE ELEMENT METHOD:

In the finite element method the region of interest is divided into a finite number of sub regions, called the element and over each element the variational formulation of the given differential equation is constructed using simple function for the whole problem are formed by a piecewise application of the variational method. For better accuracy it will not be necessary to increase the order of the functions used, but it would be sufficient to use a finer mesh. In this way, the difficulties encountered in the direct application of the variational method are overcome. The basic steps involved in the finite element method are as follows:

(1) **Discretization:**

The given domain is divided into a number of finite elements. The points of intersection are called nodes. The nodes and the elements are both numbered.

(2) **Derivation Of Element Equations:**

For the given differential equation, a variational formulation is constructed over a typical element. The
element equation are obtained by substituting a typical
dependent variable, say

\[ u = \sum_{i=1}^{n} u_i \psi_i \]

into the variational formulation. After choosing \( \psi_i \), the
interpolation functions, the element matrices are computed.

(3) **Assembly :**

The next step is the assembly of the element
equations. So the total solution is continuous. When this is
done, the entire system takes the matrix form

\[ k u' = F' \]

where \( k \) = assemblage property matrix

\( u' = \) column vectors containing unknowns

\( F' = \) external forces.

(4) **Boundary Conditions :**

The above system of equations is modified using the
boundary conditions of the problem.

(5) **Solutions Of The Equations :**

After incorporating the boundary conditions, the
system is solved by any standard technique, for example the
LU decomposition.

2.9 **ANCILLARY CONCEPTS AND FORMULAS :**

The finite element method is a piecewise application of
a variational method. There are two basis concepts in the
variational solution of differential equation.

(1)To cast a given differential equation in variational
form.
To determine the approximate solution using a variational method, such as the Ritz method, the Galerkin method or other methods.

The term “variational formulation” means the weak formulation in which given differential is recast in an equivalent integral form by trading the differentiation between a test function and the dependent variable. For most linear problems the weak formulation is equivalent to the minimization of a quadratic function $I(u)$ called the total potential energy in solid mechanics problems. Analogues to the necessary condition for the minimum of an ordinary function, the necessary condition for a quadratic functional is that its first derivative with respect to the dependent variable be zero from calculus of variations. One knows that the minimizing function is the true solution of the differential equation.

In a variational method the dependent variable of a given problem is approximated by a linear combination of appropriately chosen function

$$u = \sum c_j \phi_j$$

The parameter $c_j$ are determine such that the function $u$ minimizes the functional $I(u)$.

Consider the problem of finding the solution $w$ to the differential equation

$$\frac{d^2}{dx^2} \left[ b(x) \frac{d^2w}{dx^2} \right] + f(x) = 0 \quad \text{...(2.9.1)}$$

for $0 < x < L$

subject to the end condition
The equation arises in the study of the elastic bending of beams. In this case $w$ denotes the transverse deflection of beam. $L$ is the total length of the beam, $b(x) > 0$ is the flexural rigidity (i.e. the product of modulus of elasticity and moment of inertia) of the beam, $f(x)$ is the transverse distributed load, and $M_0$ is the bending moment in figure (2.9.1a). The solution $w$ is called the dependent variable of the problem and all other quantities ($L$, $b$, $f$, $M_0$) which are known in advance are called the data of the problem.

\[
\begin{align*}
\frac{dw}{dx}(0) &= 0 \\
\left( b \frac{d^2 w}{dx^2} \right)_{x=L} &= M_0 \\
\left[ \frac{d}{dx} \left( b \frac{d^2 w}{dx^2} \right) \right]_{x=L} &= 0
\end{align*}
\]...

(2.9.2)

Figure (2.9.1)
(a) A cantilever beam with continuous load
(b) A cantilever beam with discontinuous load
When \( b(x) \) and \( f(x) \) are continuous functions of \( x \) in \((0, L)\), the data are said to be smooth, for which case the solution \( w \) to the problem exist and satisfies the differential equation (2.9.1) at every point \( x \) in \((0, L)\) as well as the boundary conditions (2.9.2) at the boundary points. As an example consider the case in which \( b \) and \( f \) are non-zero constants. Then the exact solution of equation (2.9.1) and (2.9.2) is given by

\[
w(x) = \frac{2M_0 - f L^2}{4b} x^2 + \frac{f L}{6b} x^3 - \frac{f}{24b} x^4 \quad \ldots(2.9.3)
\]

Thus the solution \( w \) and its derivatives up to fourth order are well defined at every point of the domain \((0, L)\).

In most practical situations the data given a problem are not smooth because the flexural rigidity may be discontinuous or the transverse loading \( f \) may be discontinuous.

Suppose that (fig. (2.9.1b))

\[
f(x) = f_0 \cdot H(a - x) \quad \ldots(2.9.4)
\]

and \( b(x) \) is continuous. Here \( H(a - x) \) denotes the Heaviside step function,

\[
H(a - x) = \begin{cases} 
1 & \text{for } x < a \\
0 & \text{for } x > a 
\end{cases} \quad \ldots(2.9.5)
\]

In this case the fourth derivatives of the solution (i.e. \( w \)) does not exist (i.e. not single valued) at \( x = a \). Therefore, the exact solution \( w \) to equations (2.9.1) and (2.9.2) does not exist in classical sense. [i.e. \( w \) must satisfy the differential equation (2.9.1) at all points of the domain]. Similar
difficulties are encountered when $w$ and its derivatives are specified at points between the end points $x = 0$ and $x = L$.

Multiplying equation (2.9.1) with a function $v$, called the test function that is twice differentiable and satisfies the conditions.

$$v(0) = \frac{dv}{dx}(0) = 0$$

integrating the first term twice by parts, and using the boundary conditions (2.9.2).

\[
0 = \int_0^L v \left[ \frac{d^2}{dx^2} \left( b \frac{d^2w}{dx^2} \right) + f \right] \, dx
\]

\[
= \int_0^L \left[ \left( -\frac{dv}{dx} \right) \frac{d}{dx} \left( b \frac{d^2w}{dx^2} \right) + vf \right] \, dx + \left[ v \frac{d}{dx} \left( b \frac{d^2w}{dx^2} \right) \right]_0^L
\]

\[
= \int_0^L \left[ b \frac{d^2v}{dx^2} \frac{d^2w}{dx^2} + vf \right] \, dx + \left[ v \frac{d}{dx} \left( b \frac{d^2w}{dx^2} \right) - \frac{dv}{dx} \frac{b^2w}{dx^2} \right]_0^L
\]

\[
= \int_0^L b \frac{d^2v}{dx^2} \frac{d^2w}{dx^2} \, dx + \int_0^L vf \, dx - \frac{dv}{dx}(L)(M_0)
\]

\[\cdots\text{(2.9.6)}\]

The test function $v$ can be viewed as a variational in $w$, consistent with the boundary conditions (2.9.2), equation (2.9.6) is called the weak, generalized, or variational equation associated with equations (2.9.1) and (2.9.2) and $w$ is called the weak, generalized or variational solution of equation (2.9.1) whenever classical solution exists, it coincides with the weak solution of the problem. In other word, the variational equation (2.9.6) is equivalent to equation (2.9.1). By formulating the problem variationally,
the continuity requirement on the solution is weak ended, and the second pair of boundary condition in equation (2.9.2) are included in variational problem (2.9.6).

The quadratic functional for the problem is obtained by multiplying the expression involving both v and w by one half and setting v = w in equation (2.9.6)

$$I(w) = \int_{0}^{L} \left[ \frac{1}{2} b \left( \frac{d^2w}{dx^2} \right) + w f \right] dx - \frac{1}{2} \frac{dw}{dx} (L)(M_0) \ldots \text{(2.9.7)}$$

The functional represents the total potential energy of the beam. It should be noted that not all differential equation admit functional formulation. On the other hand, one does not need a quadratic functional to develop a finite element model.

### 2.10 FUNCTIONALS :

In the sequel we shall consider integral expression of the form $I(w)$ in equation (2.9.7)

$$I(u) = \int_{a}^{b} F(x, u, u') dx$$

where the integrand $F(x, u, u')$ is a given functional of the arguments $x$, $u$ and $\frac{du}{dx}$. The value $I(u)$ of the integral depends on $u$, hence the notation $I(u)$ is appropriate. However to describe functional defined by integrals whose arguments themselves are functions. Functional is a “function of functions”, mathematically, a functional is an operator $I$ mapping $u$ into a scalar $I(u)$. The set of all
functions $u(x)$ for which $I(u)$ makes sense is called the
domain space of the functional. The set of images of all
functions $u$ under the mapping $I$ is called the range of
functions $I(u)$. By definition, the range of a functional is a
subset of the real number field.

A functional $I(u)$ is said to be linear in $u$, if and only if
it satisfies the relation

$$I(\alpha u + \beta v) = \alpha I(u) + \beta I(v)$$

for any scalar $\alpha$ and $\beta$ and dependent variables $u$ and $v$. A
functional $B(u, v)$ is said to be bilinear if it is linear in each
of its arguments $u$ and $v$. The linear form is given by the last
two terms in equation (2.9.6),

$$I(v) = -\int_0^L v f \, dx + \frac{dv}{dx} (L)(M_0)$$

and bilinear form is given by the first term in equation
(2.9.6)

$$B(v, w) = \int_0^L b \frac{dv}{dx} \frac{dw}{dx} \, dx$$

Now by above definitions, we can rewrite the weak
form in equation (2.9.6) in the form

$$B(v, w) = I(v) \quad \text{for any } v$$

2.11 VARIATIONAL METHOD FOR APPROXIMATION:

Variational method of approximation includes the Ritz
method, the Galerkin method, the Petrov–Galerkin method,
the least square method and the collocation method. All of
these methods seek an approximation solution in the form
of a linear combination of suitable approximation functions.
The parameters in the linear combination are determined such that the approximate solution satisfies the weak form or minimizes the quadratic functional of the equation. Various methods differ from each other in the choice of the approximate functions.

The finite element method makes use of the variational method to formulate the discrete equation for sub domain, called element. The choice of the approximation function in the finite element method is different from that in the classical variational method. Here we discuss the Ritz method and Galerkin method.

(1) **The Ritz Method**

Consider the variational problem of finding the solution \( u \) such that

\[
B(v, u) = l(v) \quad \text{(2.11.1)}
\]

for all sufficiently differentiable \( v \) that satisfy the homogeneous form of essential boundary condition on \( u \). When the functional \( B \) is bilinear and symmetric and \( l \) is linear, then the problem in equation (2.11.1) is equivalent to minimizing the quadratic functional

\[
I(u) = \frac{1}{2} B(v,u) - l(u) \quad \text{(2.11.2)}
\]

The Ritz method seeks an approximate solution to equation (2.11.1) in the form of a finite series

\[
u_N = \sum_{j=1}^{N} c_j \phi_j + \phi_0 \quad \text{(2.11.3)}
\]
where the constants \( c_j \), called the Ritz coefficients, are chosen such that equation (2.11.1) holds for \( v = \phi_0 \ (i = 1, 2, ..., N) \)

\[
B \left( \phi_i \sum_{j=1}^{N} c_j \phi_j + \phi_0 \right) = l(\phi_i) \quad i = 1, 2, ..., N \quad \text{(2.11.4)}
\]

If \( B \) is bilinear, we have

\[
\sum_{j=1}^{N} B(\phi_i, \phi_j) c_j = l(\phi_i) - B(\phi_i, \phi_0) \quad \text{...(2.11.5)}
\]

which represents a system of \( N \) linear algebraic equation in \( N \) constants \( c_j \). The columns of matrix coefficient \( b_{ij} = B(\phi_i, \phi_j) \) must be linearly independent in order that the coefficient matrix in equation (2.11.5) can be inverted.

For symmetric bilinear forms, the Ritz method can be viewed as one that seeks solution of the from in equation (2.11.3) in which the parameters are determined by minimizing the quadratic functional corresponding to the symmetric bilinear form, that is, functional \( I(u) \) in equation (2.11.2). After substituting \( u_N \) from equation (2.11.3) for \( u \) into equation (2.11.2) and integrating, the functional \( I(u) \) becomes an ordinary function of the parameters \( c_1, c_2, ..., c_N \). Then the necessary condition for the minimum of \( I(c_1, c_2, ..., c_N) \) is that its partial derivatives with respect to each of the parameters be zero.

\[
\frac{\partial I(c_j)}{\partial c_1} = 0, \quad \frac{\partial I(c_j)}{\partial c_2} = 0, ..., \frac{\partial I(c_j)}{\partial c_N} = 0 \quad \text{...(2.11.6)}
\]

Thus there are \( N \) linear algebraic equations in \( N \) unknowns. \( (c_j (j = 1, 2, ..., N)) \). These equations are exactly the
same as those in equations (2.11.5) for linear symmetric bilinear forms. Of course, when \( B(., .) \) is not symmetric, we do not have a quadratic functional. In other words, equation (2.11.5) is more general than equation (2.11.6) and they are same when \( B(., .) \) is bilinear any symmetric. In most problems of interest in our discussion, we will have a symmetric bilinear form.

Returning to the Ritz approximation in equation (2.11.3), we list the properties required of the approximation functions \( \phi_i (i=1,2,...,N) \) and \( \phi_0 \). The function \( \phi_0 \) is selected to satisfy the specified essential boundary conditions are all homogeneous, then \( \phi_0 = 0 \). Since \( \phi_0 \) satisfies specified essential boundary conditions, we require \( \phi_i (i=1,2,...,N) \) satisfying the homogeneous form of the essential boundary conditions. So that \( u_N = \phi_0 \) at the points at which the essential boundary conditions are specified. Since \( \phi_i \) satisfy the homogeneous essential boundary conditions, the choice \( v = \phi_i \) is consistent with the requirements of a test function. In addition to the above requirements, we require \( \phi_i \) to satisfy the following conditions.

(1a) \( \phi_i \) should be such that \( B(\phi_i, \phi_j) \) is well defined and non zero \[ \text{i.e. sufficiently differentiable as required by the linear form } B(., .) \].

(1b) \( \phi_i \) must satisfy at least the homogeneous form of the essential boundary conditions of the problem.

(2) For any \( N \), the set \( \{\phi_i\}_{i=1}^{N} \) along with the column (and rows) of \( B(\phi_i, \phi_j) \) are linearly independent.
(3) \( \{ \phi_i \} \) is complete. 

...\( (2.11.7) \)

The above requirement of the approximation function guarantee, for linear problem, convergence of the Ritz solution to the exact solution as the value of \( N \) is increased. The convergence is understood to be in the following sense

\[
I(u_N) \geq I(u_M) \quad \text{for} \quad N \leq M \quad \ldots (2.11.8)
\]

For any value of \( N \), the previously computed elements of the matrix coefficients \( b_{ij} \) and the column vector \( F_i = (\phi_i) - B(\phi_i, \phi_b) \) remain unchanged, and one must add to existing coefficients newly computed rows and columns.

(2) The Galerkin Method:

For \( \psi_i = \phi_i \), the weighted residual method is better known as the Galerkin method. When the operator is a linear differential operator of even order, the Galerkin method reduces to the Ritz method. Because half of the differentiation can be transferred to the weight functions, the resulting coefficient matrix will be symmetric.

2.12 TIME DEPENDENT PROBLEM:

In time dependent (unsteady) problems, the undetermined parameters \( c_i \) in equation (2.11.3) are assumed to be functions of time, while \( \phi_i \) are assumed to depend on spatial coodinated. This leads to two stages of solution, both of which employ approximate is considered first and the time approximation next. Such a procedure is commonly known as semi discrete approximation. Semi
discrete variational approximation is space, as discussed in the preceding sections, results in a set of ordinary differential equations in time, which must be further approximated to obtain a set of algebraic equations. The spatial approximation of time-dependent problems leads to a matrix differential equation of the form

\[
[ A ] \left\{ \frac{\partial \{ c \}}{\partial t} \right\} + [ B ] \{ c \} - \{ p \} \quad \ldots(2.12.1)
\]

for equations involving first order time derivatives and

\[
[ A ] \left\{ \frac{\partial^2 \{ c \}}{\partial t^2} \right\} + [ B ] \{ c \} - \{ p \} \quad \ldots(2.12.2)
\]

for equations containing second order time derivatives, where

\[
A_{ij} = \int_{\Omega} \phi_i \phi_j \ dx \ dy \quad \ldots(2.12.3)
\]

we discuss approximation schemes for first and second order time derivatives.

**APPROXIMATION OF FIRST ORDER TIME DERIVATIVES**

Consider the matrix equation of the form

\[
[ A ] \{ \dot{c} \} + [ B ] \{ c \} = \{ p \} \quad \text{for} \ 0 < t \leq T_0 \quad \ldots(2.12.4)
\]

where [A], [B] and \{p\} are known matrices and \{c\} is the column matrix of the undetermined parameters. A superposed dot on \{c\} indicates differentiation with respect to t. Equation (2.12.4) is valid for any \( t > 0 \).
We introduce a $\theta$ family of approximation which approximates a weighted average of time derivative of a dependent variable at two consecutive time steps by linear interpolation of the values of the variables at the two time steps:

$$
\theta \{\dot{c}\}_{n+1} + (1 - \theta) \{\dot{c}\}_n = \frac{\{c\}_{n+1} - \{c\}_n}{\Delta t_{n+1}}
$$

...(2.12.5)

for $0 \leq \theta \leq 1$

where $\{\cdot\}_n$ refers to the value of the enclosed quantity at time $t = t_n = \sum_{i=1}^{n} \Delta t_i$ and $\Delta t_n = t_n - t_{n-1}$ is $n^{th}$ time step. If the time interval $[0, T_0]$ is divided into equal time steps, then $t_n = n \Delta t$. From equation (2.12.5) we can obtain a number of well-known difference schemes by choosing the value of $\theta$.

$$
\theta = \left\{ \begin{array}{ll}
0 & \text{forward difference conditionally stable} \\
1/2 & \text{Crank - Nicolson Scheme} \\
2/3 & \text{Galerkin method} \\
1 & \text{backward difference Scheme conditionally stable}
\end{array} \right.
$$

Using the approximation (2.12.5) for time $t_n$ and $t_{n+1}$ in equation (2.12.1) we obtain,

$$
[A] \{c\}_{n+1} = [A] \{c\}_n + \theta \Delta t_{n+1} (\{p\}_{n+1} - [B] \{c\}_{n+1}) \\
+ (1 - \theta) \Delta t_{n+1} (\{p\}_n - [B] \{c\}_n)
$$

Rearranging the terms to write $\{c\}_{n+1}$ in terms of $\{c\}_n$, we obtain
\[
\left( [A] + \theta \Delta t_{n+1} [B] \right) \{c\}_{n+1} = \left( [A] - (1-\theta) \Delta t_{n+1} [B] \right) \{c\}_n
\]

\[
+ \Delta t_{n+1} \left[ \theta \{p\}_{n+1} + (1-\theta) \{p\}_n \right]
\]

\[
[\hat{A}] \{c\}_{n+1} = [\hat{B}] \{c\}_n + \{p\}_{n,n+1} = \{\hat{p}\}_{n,n+1} \quad \ldots(2.12.6)
\]

where

\[
[\hat{A}] = [A] + \theta \Delta t_{n+1} [B]
\]

\[
[\hat{B}] = [A] - (1-\theta) \Delta t_{n+1} [B] \quad \ldots(2.12.7)
\]

The solution at time \( t = t_{n+1} \) is obtained in terms of the solution at time \( t_n \) by inverting the matrix \([\hat{A}]\). At \( t = 0 \), the solution is known from the initial conditions of the problem and therefore, equation (2.12.6) can be used to obtain the solution at \( t = \Delta t_1 \). Since the column vector \( \{p\} \) is known at all times \( \{p\}_{n+1} \) is known in advance.

It must be pointed out that one can expect better results if smaller time steps are used. In practice, however, one wishes to take as large a time step as possible to cut down the computational expense. Larger time steps, in addition to decreasing the accuracy of the solution, can introduce some unwanted, numerically induced oscillations into the solution. Thus an estimate of an upper bound on the time step proves to be very useful. A stability analysis shows that the numerical scheme (2.12.6) is stable if the minimum eigenvalue \( \lambda \) of the equation

\[
\det \left( [\hat{B}] - \lambda [\hat{A}] \right) = 0 \quad \ldots(2.12.8)
\]

is non-negative. More specifically, we have

Stable without oscillations \( 0 < \lambda < 1 \)
Stable with oscillations \(-1 < \lambda < 0\)

Unstable \(\lambda < -1\)

The Crank – Nicolson and Galerkin methods are to be stable methods.

**APPROXIMATIONS FOR SECOND ORDER TIME DERIVATIVES :**

In structural dynamics problems the equations of motion involve the second order time derivatives of the dependent variables. The semi discrete approximation of the equations results in a matrix differential equation of the form

\[
[A] \{\ddot{c}\} + [B]\{c\} = \{F\} \quad 0 < f < f_0 \quad \ldots(2.12.9)
\]

There are several approximation schemes available for time derivatives. The most commonly used one is the “Newmark direct integration method”. In the Newmark direct integration method, the first time derivative \(\{\dot{c}\}\) and the function \(\{c\}\) itself are approximated at the \((n+1)\)th time step. (\(\Delta f_1 = \Delta f_2 = \ldots = \Delta f\)) by the following expressions :

\[
\{\ddot{c}\}_{n+1} = \{\ddot{c}\}_n + \left[(1-\alpha)\{\ddot{c}\}_n + \alpha\{\ddot{c}\}_{n+1}\right]\Delta t
\]

\[
\{c\}_{n+1} = \{c\}_n + \{\dot{c}\}_n \Delta t + \left[\left(\frac{1}{2} - \beta\right)\{\ddot{c}\}_n + \beta\{\ddot{c}\}_{n+1}\right](\Delta t)^2
\]

\ldots (2.12.10)

where \(\alpha\) and \(\beta\) are parameters that control the accuracy and the stability of the scheme, and the subscript \(n\) indicates that the solution is evaluated at the \(n^{\text{th}}\) time step. (i.e at time \(t = t_n\)). The choice \(\alpha = \frac{1}{2}\) and \(\beta = \frac{1}{4}\) is known to give an
unconditionally stable (in linear problems) scheme, which corresponds to the constant – average acceleration method.

The case \( \alpha = \frac{1}{2} \) and \( \beta = \frac{1}{6} \) corresponds to the linear acceleration method.

Rearranging equation (2.12.9) and (2.12.10), we arrive at,

\[
[\mathbf{A}]{\mathbf{c}}_{n+1} = \{\mathbf{F}\}_{n,n+1}
\]

where

\[
[\mathbf{A}] = [\mathbf{B}] + a_0 [\mathbf{A}]
\]

\[
\{\mathbf{F}\} = \{\mathbf{F}\}_{n,n+1} + [\mathbf{A}]\{\mathbf{c}\}_{n} + a_1 \{\mathbf{c}\}_{n} + a_2 \{\mathbf{c}\}_{n}
\]

...(2.12.11)

Once the solution \( \{\mathbf{c}\} \) is known at \( t_{n+1} = (n+1) \Delta t \), the first and second derivatives of \( \{\mathbf{c}\} \) at \( t_{n+1} \) can be computed from equation (2.12.10)

\[
\{\mathbf{c}\}_{n+1} = a_0 (\{\mathbf{c}\}_{n+1} - \{\mathbf{c}\}_{n}) - a_1 \{\mathbf{c}\}_{n} - a_2 \{\mathbf{c}\}_{n}
\]

\[
\{\mathbf{c}\}_{n+1} = \{\mathbf{c}\}_{n} + a_3 \{\mathbf{c}\}_{n} + a_4 \{\mathbf{c}\}_{n}
\]

...(2.12.13)

\[
a_0 = \frac{1}{\beta (\Delta t)^2} \quad a_1 = a_0 \Delta t \quad a_2 = \frac{1}{2 \beta} \quad a_3 = (1 - \alpha) \Delta t \quad a_4 = \alpha \Delta t
\]

For a given set of initial condition \( \{\mathbf{c}\} \), \( \{\mathbf{c}\}_0 \) and \( \{\mathbf{c}\}_0 \), we can solve equation (2.12.11) repeatedly, marching forward in time, for the column vector \( \{\mathbf{c}\} \) and its time derivatives at any time \( t > 0 \).

A couple of comments are in order on the selection of time step and the computation of initial conditions. Although, the Newmark method is unconditionally stable
(i.e. the solution is stable for any value of $\Delta t$, however, it may be inaccurate, it is helpful to have a means to determine the value of $\Delta t$ for which the solution is also accurate. The following gives an estimate for the time increment;

$$
\Delta t = \frac{T_{\text{min}}}{\pi}
$$

...(2.12.14)

where $T_{\text{min}}$ is the smallest period of natural vibration associated with the approximate problem. An estimate for $\Delta t$ can also be obtained from the condition that the smallest eigenvalue of the eigenvalue problem.

$$
(a_0 [ A ] - \lambda [ \hat{A} ])\{u\} = 0
$$

...(2.12.15)

is less than 1.

Initial values of $\{\hat{c}\}$ are generally not known from the problem description. In that case one can make use of equation (2.12.11) at time $t = 0$ to compute $\{\hat{c}\}$.

### 2.13 FINITE ELEMENT METHOD FOR ONE DIMENSIONAL SECOND ORDER EQUATION:

We consider the two point boundary value problem defined by

$$
\frac{d}{dx} \left[ a \frac{dy}{dx} \right] = -f(x) \quad 0 < x < 1
$$

...(2.13.1)

with boundary conditions

$$
y(0) = 0, \quad \left[ a \frac{dy}{dx} \right]_{x=1} = 0
$$

...(2.13.2)

The basic steps involved in the finite element method are now elaborated and explained.
Step - 1: Discretization Of Domain :

In the present problem the region of interest is the x-axis from $x = 0$ to $x = 1$. Suppose that this is divided into a set of subintervals called element, of unequal length in general. The intersection points are called nodes. Let these be given by $x_0, x_1, x_2, \ldots, x_{n-1}, x_n$, where $x_0 = 0$ and $x_n = 1$. The elements are numbered as (1), (2), (3), ..., (n), a typical element being the $e^{th}$ element of length $h_e$ from node $e$ to node $e+1$. Let $x_e$ and $x_{e+1}$ be the values of $x$ at the nodes $e+1$ and $e$ and let $y^e$ and $y^{e+1}$ be the values of $y$ at these nodes, respectively. In general, $y^{(e)}$ satisfies the condition that outside $e+1$.

$$y^{(e+1)}(x) = 0 \quad \text{for all elements } (e+1) \quad \ldots(2.13.3)$$

\begin{align*}
\text{(a)} & \quad f(x) \quad p = \frac{a \, dy}{dx} \\
\text{(b)} & \quad u = 0 \quad \frac{a \, dy}{dx} \equiv p
\end{align*}
Using (2.13.3), it follows that the global approximate solution, \( y(x) \) can be written as

\[
y(x) = \sum_{e} y^{(e+1)}(x) \tag{2.13.4}
\]

where the summation is taken over all the elements.

This completes the discretization process and in the next step, we choose a particular element \( e+1 \) and formulate a variational principle for it.

**Step – 2: Variational Formulation Over The Element \( e \):**

From (2.13.1), we obtain

\[
\int_{x_{e}}^{x_{e+1}} v \frac{d}{dx} \left[ a \frac{dy}{dx} \right] dx = - \int_{x_{e}}^{x_{e+1}} v f dx
\]
which is written as

\[ 0 = \int_{x_e}^{x_{e+1}} \left\{ v \frac{d}{dx} \left[ a \frac{dy}{dx} \right] + v f \right\} dx \]

\[ = \left[ v \frac{dy}{dx} \right]_{x_e}^{x_{e+1}} - \int_{x_e}^{x_{e+1}} a v' \frac{dy}{dx} dx + \int_{x_e}^{x_{e+1}} v f dx \]

\[ = - \int_{x_e}^{x_{e+1}} \left[ a v' y' - v f \right] dx + v (x_{e+1}) P_2^{(e)} + v (x_e) P_1^{(e)} \]

...(2.13.5)

The associated quadratic form for the element is given by

\[ I_e(y) = \int_{x_e}^{x_{e+1}} \left[ \frac{1}{2} \left( \frac{dy}{dx} \right)^2 - yf \right] dx - P_1^{(e)} y(x_e) - P_2^{(e)} y(x_{e+1}) \] ... (2.13.5 a)

where

\[ P_1^{(e)} = \left[ - a \frac{dy}{dx} \right]_{x_e}^{x_{e+1}} \text{ and } P_2^{(e)} = \left[ a \frac{dy}{dx} \right]_{x_e}^{x_{e+1}} \] ... (2.13.6)

In the next step, we use a variational method to approximate equation (2.13.5). We demonstrate this by using the Ritz method.

**Step – 3: Ritz Approximation Over The Element**:

Let \( y_e(x) \) be an approximation to \( y(x) \) over the element (e). So that

\[ y_e(x) = \sum_{j=1}^{n} \alpha_j \psi_j^{(e)} (x) \] ... (2.13.7)

where \( \alpha_j \) are parameters to be determined and \( \psi_j^{(e)} (x) \) are approximation functions to be chosen. Substituting (2.13.7) in (2.13.5), we obtain
Step – 4: Derivation Of The Approximation Function

For An Element:
In the Ritz and Galerkin methods, the system of equation is obtained in terms of arbitrary parameters \( \alpha_j \). In the finite element method, on the other hand, the unknown values of the dependent variable \( y \) at the node are taken as parameters. This is done in following way. Let

\[
y(x) = \alpha_1 + \alpha_2 \, x
\]

be an approximation in the element \( e \). We have

\[
\begin{align*}
y(x_e) &= \alpha_1 + \alpha_2 \, x_e = y_1^e \\
y(x_{e+1}) &= \alpha_1 + \alpha_2 \, x_{e+1} = y_2^{e+1}
\end{align*}
\]

solving the equation in (2.13.13), we obtain

\[
\sum_{j=1}^{n} \alpha_j^e \left[ \int_{x_e}^{x_{e+1}} a \psi_j'(x) \psi_i'(x) \, dx \right] = \int_{x_e}^{x_{e+1}} f \psi_i(x) \, dx + \psi_i(x_e) P_2^{(e)} + \psi_i(x_{e+1}) P_1^{(e)}
\]

\[
i = 1, 2, 3, \ldots, n
\]

... (2.13.8)

Equation (2.13.8) can be written in the matrix form

\[
\begin{bmatrix} k_{ij}^{(e)} \end{bmatrix} \{ \alpha_j^e \} = F_i^{(e)}
\]

... (2.13.9)

where \( k_{ij} \) and \( F_i \) are called stiffness matrix and force vector respectively and are given by

\[
k_{ij}^{(e)} = \int_{x_e}^{x_{e+1}} a \frac{d\psi_i}{dx} \frac{d\psi_j}{dx} \, dx
\]

... (2.13.10)

and

\[
F_i^{(e)} = \int_{x_e}^{x_{e+1}} f \psi_i(x) \, dx + \psi_i(x_e) P_2^{(e)} + \psi_i(x_{e+1}) P_1^{(e)}
\]

... (2.13.11)
\[ \alpha_1 = \frac{y_1^e x_{e+1} - y_2^e x_e}{x_{e+1} - x_e} \] ...
(2.13.14)

and

\[ \alpha_2 = \frac{y_2^e - y_1^e}{x_{e+1} - x_e} \] ...
(2.13.15)

equation (2.13.12) now becomes,

\[
y(x) = \frac{y_1^e x_{e+1} - y_2^e x_e}{x_{e+1} - x_e} + \frac{y_2^e - y_1^e}{x_{e+1} - x_e}
= \frac{x_{e+1} - x}{x_{e+1} - x_{e-1}} y_1^e + \frac{x - x_e}{x_{e+1} - x_e} y_2^e
= \sum_{i=1}^{2} y_i^{(e)} \psi_i^{(e)}(x)
\] ...
(2.13.16)

where

\[
\psi_1^{(e)}(x) = \frac{x_{e+1} - x}{x_{e+1} - x_{e-1}} \quad \text{and} \quad \psi_2^{(e)}(x) = \frac{x - x_e}{x_{e+1} - x_e} \] ...
(2.13.17)

with \( x_1 = x_e \) and \( x_2 = x_{e+1} \), the functions \( \psi_i^{(e)} \) have the property

\[
\psi_i^{(e)}(x_j) = \begin{cases} 
0 & i \neq j \\
1 & i = j 
\end{cases}
\] ...
(2.13.18)

Instead of equation (2.13.9), we now have

\[
\begin{bmatrix} k_{ij}^{(e)} \end{bmatrix} y_j^{(e)} = F_i^{(e)}
\] ...
(2.13.19)

Where \( k_{ij}^{(e)} \) and \( F_i^{(e)} \) are given by (2.13.10) and (2.13.11)

with the choice of \( \psi_i^{(e)}(x) \) as in equation (2.13.17), we now demonstrate the computation of \( k^{(e)} \) and \( F^{(e)} \). In particular, \( h_e = x_{e+1} - x_e \), we obtain
\[
\frac{d\psi_1^{(e)}}{dx} = -\frac{1}{h_e} \quad \text{and} \quad \frac{d\psi_2^{(e)}}{dx} = \frac{1}{h_e} \quad \ldots (2.13.20)
\]

Where
\[
k_{11} = \int_{x_e}^{x_e+1} \left( -\frac{1}{h_e} \right)^2 \, dx = \frac{1}{h_e}
\]
\[
k_{12} = \int_{x_e}^{x_e+1} -\frac{1}{h_e^2} \, dx = -\frac{1}{h_e} = k_{21}
\]
\[
k_{22} = \int_{x_e}^{x_e+1} \frac{1}{h_e^2} \, dx = \frac{1}{h_e}
\]

And
\[
F_1^{(e)} = \int_{x_e}^{x_e+1} \frac{x_e+1-x_e}{h_e} \, dx + p_1^{(e)} = \frac{h_e}{2} + p_1^{(e)}
\]
\[
F_2^{(e)} = \int_{x_e}^{x_e+1} \frac{x-x_e}{h_e} \, dx + p_2^{(e)} = \frac{h_e}{2} + p_2^{(e)}
\]

from equation (2.13.9) for an element, we have from equation (2.13.16) and (2.13.18) that
\[
[k^e] \{ y^{(e)} \} = \{ F^{(e)} \} \quad \ldots (2.13.21)
\]

for linear element
\[
k_{ij}^{(e)} = \int_{x_e}^{x_e+1} a \frac{d\psi_i^{(e)}}{dx} \frac{d\psi_j^{(e)}}{dx} \, dx \quad \ldots (2.13.22)
\]
\[
F_1^{(e)} = \int_{x_e}^{x_e+1} \psi_i f \, dx + p_1^{(e)}
\]

**Step – 5: Assembly Of Element Equations:**

We express the element equation (2.13.21) in terms of the global nodal values \( y_i \) for each element. Since equation (2.13.21) is derived for an arbitrary typical element, it holds for any element from the finite element mesh. For the sake of discussion, suppose that the domain of problem
$\Omega = (0,1)$ is divided into three elements of possibly unequal lengths. Since these elements are connected at nodes 2 and 3 and $y$ is continuous, $y_2$ of element $\Omega^e$ should be the same as $y_1$ of element $\Omega^{e+1}$ for $e = 1, 2$. To express this correspondence mathematically, we label the value of $y$ at the global nodes with $y_i$ ($i=1,2,...,N$) where $N$ is the total number of global nodes. Then we have the following correspondence between the local nodal values and the global nodal values (see figure 2.13.2(a) and 2.13.2(b)).

![Diagram](image-url)
Correspondence of local and global nodal values and representation of the finite element solution by global interpolation functions for the model problem

(a) Correspondence of element and global nodal values
(b) Representation of the solution by global interpolation functions

\[
\begin{align*}
y_1^{(1)} &= y_1 & y_2^{(1)} &= y_2 = y_2^{(2)} \\
y_3^{(1)} &= y_3 = y_3^{(3)} & y_2^{(3)} &= y_4 & \ldots (2.13.23)
\end{align*}
\]

We shall these relations inter element continuity conditions.

We express element equation (2.13.21) in terms of the global nodal values $y_i$ for each element.
Equations (2.13.24), (2.13.25) and (2.13.26) indicate the contribution of each element to overall problem. Note that none of the element equations is solvable because the element matrices are singular and because parts of the right hand sides are not known. By superimposing above equation, we obtain the equation of the system
equation (2.13.27) represent global finite element model of equation (2.13.1).

An alternative derivation of the assembly of element equations is also presented here. The procedure to be described is based on the idea that the quadratic functional associated with the problem is equal to sum of the quadratic functional $I_e$ of equation (2.13.5 a) of all elements.

$$I(y_i) = \sum_{e=1}^{N} I_e(y_i^{(e)})$$

or in matrix form

$$0 = \left[ \delta I(y_i) \right] \{ \delta (y_i) \} = \sum_{e=1}^{N} \left[ \delta I_e(y_i^{(e)}) \right] \{ \delta y_i^{(e)} \}$$

where $\delta y_i^{(e)}$ can be viewed as the test function values $y_i^{(e)}$.

$$I_e(y^{(e)}) = \frac{1}{2} \{ u^{(e)} \}^T \{ k^e \} \{ y^{(e)} \} - \{ y^{(e)} \}^T \{ f^{(e)} \}$$

$N$ denotes the number of elements in the mesh and $\delta$ the variational operator. Since (2.13.29) is more general than (2.13.28), we shall work with the former. Carrying out the algebra in equation (2.13.29), we obtain
0 = \sum_{e=1}^{N} \sum_{i=1}^{2} \left[ \sum_{j=1}^{2} k_{ij}^{(e)} y_{j}^{(e)} - F_{i}^{(e)} \right] \delta y_{i}^{(e)}

= \left( k_{11}^{(1)} y_{1}^{(1)} + k_{12}^{(1)} y_{2}^{(1)} - F_{1}^{(1)} \right) \delta y_{1}^{(1)} + \left( k_{21}^{(1)} y_{1}^{(1)} + k_{22}^{(1)} y_{2}^{(1)} - F_{2}^{(1)} \right) \delta y_{2}^{(1)}

+ \left( k_{11}^{(2)} y_{1}^{(2)} + k_{12}^{(2)} y_{2}^{(2)} - F_{1}^{(2)} \right) \delta y_{1}^{(2)} + \left( k_{21}^{(2)} y_{1}^{(2)} + k_{22}^{(2)} y_{2}^{(2)} - F_{2}^{(2)} \right) \delta y_{2}^{(2)}

+ \left( k_{11}^{(3)} y_{1}^{(3)} + k_{12}^{(3)} y_{2}^{(3)} - F_{1}^{(3)} \right) \delta y_{1}^{(3)} + \left( k_{21}^{(3)} y_{1}^{(3)} + k_{22}^{(3)} y_{2}^{(3)} - F_{2}^{(3)} \right) \delta y_{2}^{(3)}

\cdots(2.13.31)

using the correspondence in equation (2.13.23) in the above equation and collecting the co-efficient of \delta y_{i} (i=1,2,\ldots,N+1) , we obtain

\begin{align*}
0 &= \delta y_{1} \left[ k_{11}^{(1)} y_{1} + k_{12}^{(1)} y_{2} - F_{1}^{(1)} \right] \\
&+ \delta y_{2} \left[ k_{21}^{(1)} y_{1} + (k_{22}^{(1)} + k_{11}^{(2)}) y_{2} + k_{12}^{(2)} y_{3} - (F_{2}^{(1)} + F_{1}^{(2)}) \right] \\
&+ \delta y_{3} \left[ k_{21}^{(2)} y_{2} + (k_{22}^{(2)} + k_{11}^{(3)}) y_{3} + k_{12}^{(3)} y_{4} - (F_{2}^{(2)} + F_{1}^{(3)}) \right] \\
&+ \delta y_{3} \left[ k_{21}^{(3)} y_{3} + k_{22}^{(3)} y_{4} - F_{2}^{(3)} \right]
\end{align*}

Since variation in \( y_{i} \) are arbitrary, the above equation implies that the co-efficient of each \( \delta y_{i} (i=1,2,\ldots,N+1) \), where \( N + 1 \) is the total number of global nodes, should be equal to zero separately. The result express in matrix form is,

\[
\begin{bmatrix}
k_{11}^{(1)} & k_{12}^{(1)} & 0 & 0 \\
k_{21}^{(1)} & k_{22}^{(1)} + k_{11}^{(2)} & k_{21}^{(2)} & 0 \\
0 & k_{21}^{(2)} & k_{22}^{(2)} + k_{11}^{(3)} & k_{12}^{(3)} \\
0 & 0 & k_{21}^{(3)} & k_{22}^{(3)}
\end{bmatrix}
\begin{bmatrix}y_{1} \\ y_{2} \\ y_{3} \\ y_{4}\end{bmatrix} =
\begin{bmatrix}F_{1}^{(1)} \\ F_{2}^{(1)} + F_{1}^{(2)} \\ F_{2}^{(2)} + F_{1}^{(3)} \\ F_{2}^{(3)}\end{bmatrix}
\cdots(2.13.32)

**Step : 6  Imposition Of Boundary Condition :**

The co efficient matrix (2.13.27) is singular prior to the imposition of the essential boundary condition. Upon the imposition of suitable boundary conditions of the problem,
we obtain a non singular matrix that can be inverted. To this end, we note first that the nature boundary conditions are included in the column vector \( \{ F^{(e)} \} \) through \( \{ p_j^{(e)} \} \) at all global nodes between the boundary nodes, the sum of the conditions of the natural boundary conditions from node 2 of element \( e \) and node 1 of element \( e+1 \) should equal the specified value of the secondary variable \( \left( a \frac{dy}{dx} \right) \).

\[
p_2^{(e)} + p_1^{(e+1)} = \text{Specified value of } \left( a \frac{dy}{dx} \right)_{x = x_{e+1}} \quad \ldots(2.13.33)
\]

for the problem on hand \( \left( a \frac{dy}{dx} \right) \) at \( x = h_1 \) and \( x = h_1 + h_2 \) are specified to be zero. We have

\[
\begin{align*}
p_2^{(1)} + p_1^{(2)} &= 0 \\
p_2^{(2)} + p_1^{(3)} &= 0
\end{align*}
\quad \ldots(2.13.34)
\]

For the longitudinal deformation of a bar, these equation can be interpreted as the equilibrium conditions for the internal forces. Mathematically, this amounts to assuming that the difference of the value \( \left( a \frac{dy}{dx} \right) \) from elements \( e \) and \( e+1 \) at global node \( e+1 \) equal the specified value of \( \left( a \frac{dy}{dx} \right) \) at node \( e+1 \).

\[
\left( a \frac{dy}{dx} \right)_{x = x_{e+1}} - \left( a \frac{dy}{dx} \right)_{x = x_{e+1}}^{(e+1)} = p_2^{(e)} + p_1^{(e+1)} = 0 \quad \ldots(2.13.35)
\]
The difference accounts for any point, source applied at the point \( x = x_{e+1} \). If no point source is applied, the difference is assumed to be zero. Note that \( \left( \frac{dy}{dx} \right) \) is known at \( x = 0 \) and it is equal to put \( x = 1 \).

Next, we discuss the imposition of essential boundary conditions of the problem. In the present case, the only known essential boundary condition is

\[
U_i = u_i^{(1)} = 0 \quad \text{...(2.13.36)}
\]

The modified columns of the unknowns and \( p_i^e \)’s are given by

\[
\{\Delta\} = \begin{bmatrix} 0 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}, \quad \{p\} = \begin{bmatrix} \left( \frac{dy}{dx} \right)_{x=0} = p_i^{(1)} \\ 0 \\ p \\ p \end{bmatrix} \quad \text{...(2.13.37)}
\]

Step : 7 Solution Of Equations :

The global finite element equation (2.13.32) can be partitioned conveniently into the following form

\[
\begin{bmatrix} [k^{11}] & [k^{12}] \\ [k^{21}] & [k^{22}] \end{bmatrix} \begin{bmatrix} \{\Delta^1\} \\ \{\Delta^2\} \end{bmatrix} = \begin{bmatrix} (F^1) \\ (F^2) \end{bmatrix} \quad \text{...(2.13.38)}
\]

where \( \{\Delta^1\} \) is the column of known displacements, \( (y_1), \{\Delta^2\} \) is the column of the unknown displacements \( (y_2, y_3, y_4) \), \( \{F^1\} \) is the column of the unknown forces \( p_i^{(1)} \), and \( \{F^2\} \) is the column of the known forces \( (0, 0, \rho) \). Writing equation (2.13.38) as two matrix equation, we obtain
\[
[k^{11}] \{\Delta^1\} + [k^{12}] \{\Delta^2\} = \{F^1\} \quad \text{...(2.13.39)}
\]
\[
[k^{21}] \{\Delta^1\} + [k^{22}] \{\Delta^2\} = \{F^2\} \quad \text{...(2.13.40)}
\]

From equation (2.13.40), we have
\[
\{\Delta^2\} = [k^{22}] \left( \{F^2\} - [k^{21}] \{\Delta^1\} \right) \quad \text{...(2.13.41)}
\]

Once \{\Delta^2\} is known, \{F^1\} can be computed from equation (3.13.39). In the present case, we have
\[
[k^{11}] = k_{11}^{(1)} \quad [k^{12}] = k_{12}^{(1)} \begin{bmatrix} 0 & 0 \end{bmatrix}
\]
\[
[k^{21}] = \begin{bmatrix} k_{21}^{(1)} \\ 0 \\ 0 \end{bmatrix}
\]
\[
[k^{22}] = \begin{bmatrix} k_{22}^{(1)} + k_{11}^{(2)} & k_{12}^{(2)} & 0 \\ k_{21}^{(2)} & k_{22}^{(2)} + k_{11}^{(3)} & k_{12}^{(3)} \\ 0 & k_{21}^{(3)} & k_{22}^{(3)} \end{bmatrix} \quad \text{...(2.13.42)}
\]
\[
\{F^1\} = F_1^{(1)} \quad \{F^2\} = \begin{bmatrix} F_2^{(1)} + F_1^{(2)} \\ F_2^{(2)} + F_1^{(3)} \\ F_2^{(3)} \end{bmatrix}
\]

For \(a = \text{constant}, \ h_1 = h_2 = h_3 = \frac{1}{3} \) and \(f = \text{constant} \), we obtain
\[
[k^{22}]^{-1} = \frac{1}{3a} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix} \quad \text{...(2.13.43a)}
\]
\[
\{\Delta^2\} = \begin{bmatrix} y_2 \\ y_3 \\ y_4 \end{bmatrix} = \frac{fL^2}{18a} \begin{bmatrix} 5 \\ 8 \\ 9 \end{bmatrix} + \frac{pL}{3a} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad \text{...(2.13.43b)}
\]
and the unknown natural boundary condition at \( x = 0 \)

\[
p_{f}^{(1)} = \left. a \frac{dy}{dx} \right|_{x=0} = -(fL + p) \quad \ldots(2.13.44)
\]

Since \( \psi_{i}^{(e)} (e=1,2,3) \) is zero in any element \( \Omega^{f} \) for \( e \neq f \) (see figure 2.13.1(a)), the finite element solution for the entire domain is given by

\[
y(x) = \sum_{i=1}^{3} \left( \sum_{i=1}^{2} Y_{i}^{(e)} \psi_{i}^{(e)} \right) = \sum_{i=1}^{4} Y_{i} \Phi_{1}(x) \quad \ldots(2.13.45)
\]

where \( \Phi_{1}(x), \, I = 1, 2, \ldots, N+1 \) are the piecewise continuous global interpolation functions.

\[
\Phi_{1}(x) = \begin{cases} 
\psi_{2}^{(1-1)}(x) & x_{1-1} \leq x \leq x_{1} \\
\psi_{1}^{(I)}(x) & x_{1} \leq x \leq x_{1+1}
\end{cases} \quad \ldots(2.13.46)
\]

for computational purpose equation (2.13.45) can be written in the form

\[
y(x) = \begin{cases} 
y_{1} \psi_{1}^{(1)}(x) + y_{2} \psi_{2}^{(1)}(x) & 0 \leq x \leq 1/3 \\
y_{2} \psi_{1}^{(2)}(x) + y_{3} \psi_{2}^{(2)}(x) & 1/3 \leq x \leq 2/3 \\
y_{3} \psi_{1}^{(3)}(x) + y_{4} \psi_{2}^{(3)}(x) & 2/3 \leq x \leq 1
\end{cases} \quad \ldots(2.13.47)
\]

**Step : 8 Post Processing Of Solution :**

The solution of the finite element equation gives the nodal values of the primary unknown or unknowns (y). Post processing of the result includes or more of following:

2. Interrogation of the results to check whether the solution makes sense.
3. Tabular of graphic presentation of the result.
2.14 FINITE ELEMENT METHOD FOR TWO DIMENSIONAL SECOND ORDER EQUATION:

2.14.1 Description Of The Model Equation:

Consider the problem of finding the solution $u$ of the second-order partial differential equation:

$$
\begin{array}{c}
-\frac{\partial}{\partial x} \left( a_{11} \frac{\partial u}{\partial x} + a_{12} \frac{\partial u}{\partial y} \right) - \frac{\partial}{\partial y} \left( a_{21} \frac{\partial u}{\partial x} + a_{22} \frac{\partial u}{\partial y} \right) + a_\infty u - f = 0 \\
\end{array}
$$

...(2.14.1)

for given data $a_{ij}$ (i, j=1, 2), $a_\infty$ and $f$ and specified boundary conditions. The form of the boundary conditions will be apparent from the variational formulation. As a special case, one can obtain the Poisson equation from Equation (2.14.1) by setting $a_{11} = a_{22} = a$ and $a_{12} = a_{21} = a_\infty = 0$.

$$
- \nabla . (a \nabla u) = f \quad \text{in} \quad \Omega 
$$

...(2.14.2)

where $\nabla$ is the gradient operator. If $\hat{i}$ and $\hat{j}$ denote the unit vectors directed along the x and y axes, respectively, the gradient operator can be expressed as

$$
\nabla = \hat{i} \frac{\partial}{\partial x} + \hat{j} \frac{\partial}{\partial y} 
$$

...(2.14.3)

In the following, we shall discuss the variational formulation and the finite element formulation of the model equation (2.14.1).

2.14.2 Variational Formulation:

In two dimensions there is more than one geometric shape that can be used as a finite element (see figure
2.14.1). As we shall see shortly, the interpolation functions depend not only on the number of nodes in the element, but also on the shape of the element. At this stage of the development, we do not have to confine our discussion to any particular element. We assume that $\Omega^e$ is a typical element.

Now by same procedure as we discussed in section (2.13) we can develop variational form of equation (2.14.1).

The variational form of equation (2.14.1) is given by

$$0 = \int_{\Omega^e} \left[ \frac{\partial v}{\partial x} \left( a_{11} \frac{\partial u}{\partial x} + a_{12} \frac{\partial u}{\partial y} \right) + \frac{\partial v}{\partial y} \left( a_{21} \frac{\partial u}{\partial x} + a_{22} \frac{\partial u}{\partial y} \right) + a_{xv} v u - vy \right] \text{dx dy}$$

$$- \int_{\Gamma^e} v q_n \text{ds} \quad \ldots(2.14.4)$$

**(Figure 2.14.1)**

**Finite element discretization of an irregular domain**

(a) Discretization of a domain by triangular and quadrilateral element

(b) A typical triangular element (boundary $\Gamma^e$, the unit normal $\hat{n}$ on the boundary of the element)
2.14.3 Finite Element Formulation:

The variational form in equation (2.14.4) indicates that the approximation chosen for \( u \) should be at least bilinear in \( x \) and \( y \), so that the first two terms in equation (2.14.4) and \( q_n \) are non-zero. Suppose that \( u \) is approximated by the expression

\[
u = \sum_{j=1}^{n} u_j \psi_j \quad \text{...(2.14.5)}
\]

where \( u_j \) are the values of \( u \) at the point \( (x_j, y_j) \) and \( \psi_j \) are linear interpolation functions with the property

\[
\psi_i(x_j, y_j) = \delta_{ij} \quad \text{...(2.14.6)}
\]

The specific form of \( \psi_i \) will be derived for linear triangular and rectangular elements in section (2.14.4).

Substituting equation (2.14.5) for \( u \) and \( \psi_i \) for \( v \) into the variational form (2.14.4), we obtain

\[
0 = \sum_{j=1}^{n} \left\{ \int_{\Omega} \left[ \frac{\partial \psi_i}{\partial x} \left( a_{11} \frac{\partial \psi_j}{\partial x} + a_{12} \frac{\partial \psi_j}{\partial y} \right) + \frac{\partial \psi_j}{\partial y} \left( a_{21} \frac{\partial \psi_j}{\partial x} + a_{22} \frac{\partial \psi_j}{\partial y} \right) + a_{\infty} \psi_i \psi_j \right] dx \, dy \right\} u_j
\]

or

\[
\sum_{j=1}^{n} k_{ij}^{(e)} u_j^{(e)} = F_i^{(e)} \quad \text{...(2.14.7)}
\]

where

\[
k_{ij}^{(e)} = \int_{\Omega} \left[ \frac{\partial \psi_i}{\partial x} \left( a_{11} \frac{\partial \psi_j}{\partial x} + a_{12} \frac{\partial \psi_j}{\partial y} \right) + \frac{\partial \psi_j}{\partial y} \left( a_{21} \frac{\partial \psi_j}{\partial x} + a_{22} \frac{\partial \psi_j}{\partial y} \right) + a_{\infty} \psi_i \psi_j \right] dx \, dy
\]

\[
F_i^{(e)} = \int_{\Omega} f \psi_i dx \, dy + \int_{\Gamma_e} q_n \psi_i ds
\]
Note that $k_{ij}^{(e)} = k_{ji}^{(e)}$ (i.e. $k$ is symmetric) only when $a_{12} = a_{21}$. Equation (2.14.7) represents the finite element model of equation (2.14.1). In the next section, we shall discuss the derivation of interpolation functions.

2.14.4 **Interpolation Functions:**

An examination of the variational form (2.14.4) and the finite element matrices in equations (2.14.8) shows that $\psi_i$ should be at least bilinear functions of $x$ and $y$. As pointed out before, there is a correspondence between both the number and location of nodal points and the number of primary unknowns per node in a finite element and the number of terms used in the polynomial approximations of a dependent variable over an element. In one-dimensional second-order problems, the number of nodes $n$ in an element uniquely defines the degree $r$ of the polynomial, with the correspondence between $n$ and $r$ being $n = r + 1$. In two-dimensional second-order problems, the correspondence between the number of nodes (which is equal to the number of terms in the approximating polynomial) and the degree of the polynomial is not unique. For example, the polynomial

$$u(x, y) = c_1 + c_2 x + c_3 y$$

...(2.14.9)

contains three (linearly independent) terms, and it is linear in both $x$ and $y$. On the other hand, the polynomial

$$u(x, y) = c_1 + c_2 x + c_3 y + c_4 xy$$

...(2.14.10)

contains four (linearly independent) terms, but it is also linear in both $x$ and $y$. The former requires an element with three nodes (with one primary unknown per node), and the
latter requires an element with four nodes. A two-dimensional element with three nodes is a triangle with the nodes at the vertices of the triangle. When the number of nodes is equal to 4, one can choose a triangle with the fourth node at the center (or centroid) of the triangle or a rectangle (or quadrilateral) with the nodes at the vertices of the rectangle. A polynomial with five constants is the (incomplete) quadratic polynomial. Here we derive interpolation function for four node rectangular element.

**LINEAR INTERPOLATION FUNCTIONS FOR THE FOUR-NODE RECTANGULAR ELEMENT:**

Here we consider approximation of the form (2.4.10) and use a rectangular element with sides a and b (see Fig. 2.14.4). For the sake of convenience, we choose a local coordinate system \((\xi, \eta)\) to derive the interpolation functions. We assume that

![Diagram](attachment:image.png)
Figure (2.14.4)

Linear rectangular element and associated interpolation functions

(a) A four-node rectangular element
(b) Linear interpolation functions

\[ u(\xi(\eta)) = c_1 + c_2 \xi + c_3 \eta + c_4 \xi \eta \]  
\[ ... (2.14.11) \]

and require

\[ u_1 = u(0, 0) = c_1 \]
\[ u_2 = u(a, 0) = c_1 + c_2 a \]
\[ u_3 = u(a, b) = c_1 + c_2 a + c_3 b + c_4 a b \]
\[ u_4 = u(0, b) = c_1 + c_3 b \]

Solving for \( c_i \) (\( i = 1, 2, 3, 4 \)), we obtain

\[
\begin{bmatrix}
  c_1 \\
  c_2 \\
  c_3 \\
  c_4 \\
\end{bmatrix} = \begin{bmatrix}
  1 & 0 & 0 & 0 \\
  1 & a & 0 & 0 \\
  1 & a & b & ab \\
  1 & 0 & b & 0 \\
\end{bmatrix}^{-1} \begin{bmatrix}
  u_1 \\
  u_2 \\
  u_3 \\
  u_4 \\
\end{bmatrix} = \begin{bmatrix}
  ab & 0 & 0 & 0 \\
  1 & -b & b & 0 \\
  1 & -a & 0 & 0 \\
  1 & 1 & 1 & 1 \\
\end{bmatrix} \begin{bmatrix}
  u_1 \\
  u_2 \\
  u_3 \\
  u_4 \\
\end{bmatrix} 
\]

\[ ... (2.14.13) \]
Substituting equation (2.14.13) into (2.14.11), we obtain

\[
u(\xi, \eta) = \{1 \quad \xi \quad \eta \quad \xi \eta\} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \{\psi_1 \quad \psi_2 \quad \psi_3 \quad \psi_4\} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix}
\]

\[
= \sum_{i=1}^{4} u_i \psi_i (\xi, \eta)
\]

...(2.14.14)

where

\[
\psi_1 (\xi, \eta) = \left(1 - \frac{\xi}{a}\right) \left(1 - \frac{\eta}{b}\right)
\]

\[
\psi_2 (\xi, \eta) = \frac{\xi}{a} \left(1 - \frac{\eta}{b}\right)
\]

\[
\psi_3 (\xi, \eta) = \frac{\xi}{a} \frac{\eta}{b}
\]

\[
\psi_4 (\xi, \eta) = \left(1 - \frac{\xi}{a}\right) \frac{\eta}{b}
\]

...(2.14.15)

The interpolation functions are shown in Fig. (2.14.4b). We have again

\[
\psi_i (\xi_j, \eta_j) = \delta_{ij} \quad i, j = 1, 2, 3, 4
\]

\[
\sum_{i=1}^{4} \psi_i = 1
\]

...(2.14.16)

Note that the linear interpolation functions for the four-node rectangular element can also be obtained by taking the tensor product of the linear interpolation functions associated with sides 1-2 and 2-3 (which can be viewed as line elements):
\[
\begin{bmatrix}
1 - \frac{\xi}{a} \\
\frac{\xi}{b}
\end{bmatrix}
\begin{bmatrix}
1 - \frac{\eta}{b} \\
\frac{\eta}{b}
\end{bmatrix} =
\begin{bmatrix}
\psi_1 & \psi_4 \\
\psi_2 & \psi_3
\end{bmatrix}
\]
\[
\text{(2.14.17)}
\]

The conventional procedure given above for the construction of the interpolation functions involves the inversion of an \( n \times n \) matrix [see equation (2.14.13)], where \( n \) is the number of nodes in the element. When \( n \) is large, the inversion becomes very tedious.

### 2.14.5 Computation Of Element Matrices:

Computation of the element matrices \([ K^{(e)} ]\) and \({ F^{(e)} }\) in equation (2.14.8) by the conventional methods (i.e., by exact integration) is, in general, not easy. However, when \( a_{ij}, a_{00} \) and \( f \) are constants, it is possible to evaluate the integrals exactly over the triangular and rectangular elements discussed in the previous section. The boundary integral in \({ F^{(e)} }\) of equations (2.14.8) can be evaluated whenever \( q_n \) is known. For an interior element (i.e., an element that does not have any of its sides on the boundary of the problem) the contribution from the boundary integral cancels with similar contributions from adjoining elements of the mesh (analogous to the \( P_i^{(e)} \) and \( Q_i^{(e)} \) in the one-dimensional problems). A more detailed discussion is given below.

For the sake of brevity, we rewrite \([ K^{(e)} ]\) in equations (2.14.8) as the sum of four basic matrices \([ S^{\alpha \beta} ](\alpha, \beta = 1, 2)\) and \([ S ]\),
\[
[K^{(e)}] = a_{11}[S^{11}] + a_{12}[S^{12}]^T + a_{21}[S^{21}]^T + a_{22}[S^{22}] + a_{00}[S]
\]

...(2.14.18a)

where

\[
S_{ij}^{11} = \int_{\Omega^e} \frac{\partial \psi_i}{\partial x} \frac{\partial \psi_j}{\partial x} \, dx \, dy \\
S_{ij}^{12} = \int_{\Omega^e} \frac{\partial \psi_i}{\partial x} \frac{\partial \psi_j}{\partial y} \, dx \, dy \\
S_{ij}^{22} = \int_{\Omega^e} \frac{\partial \psi_i}{\partial y} \frac{\partial \psi_j}{\partial y} \, dx \, dy \\
S_{ij} = \int_{\Omega^e} \psi_i \psi_j \, dx \, dy
\]

...(2.14.18b)

Also we let

\[
f_i^{(e)}(x) = \int_{\Omega^e} f(x) \psi_i \, dx \\
Q_i^{(e)} = \int_{\Gamma e} q_n \psi_i \, ds
\]

...(2.14.19)

We now proceed to compute the matrices in equations (2.14.18) and (2.14.19) using the linear interpolation functions derived in the previous section.

**ELEMENT MATRICES FOR A LINEAR RECTANGULAR ELEMENT:**

Since \(a_{00}, a_{ij}\) and \(f\) are constants, we can use the interpolation functions of equations (2.14.15) with \(\xi\) and \(\eta\) replaced by \(x\) and \(y\) respectively. We have

\[
[S^{11}] = \frac{b}{6a} \begin{bmatrix}
2 & -2 & -1 & 1 \\
-2 & 2 & 1 & -1 \\
-1 & 1 & 2 & -2 \\
1 & -1 & 2 & 2 \\
\end{bmatrix} \\
[S^{12}] = \frac{1}{4} \begin{bmatrix}
1 & 1 & -1 & -1 \\
-1 & -1 & 1 & 1 \\
-1 & -1 & 1 & 1 \\
1 & 1 & -1 & -1 \\
\end{bmatrix} \\
[S^{22}] = \frac{a}{6a} \begin{bmatrix}
2 & 1 & -1 & -2 \\
1 & 2 & -2 & -1 \\
-1 & -2 & 2 & 1 \\
-2 & -1 & 1 & 2 \\
\end{bmatrix} \\
[S] = \frac{ab}{36} \begin{bmatrix}
4 & 2 & 1 & 2 \\
2 & 4 & 2 & 1 \\
1 & 2 & 4 & 2 \\
2 & 1 & 2 & 4 \\
\end{bmatrix} \\
\{f\} = \frac{fab}{4} \{1 \ 1 \ 1 \ 1\}^T
\]

...(2.14.20)
Evaluation Of The Boundary Integrals:

Here we consider the evaluation of boundary integrals of the type

$$Q_i^{(e)} = \int_{\Gamma^e} q_n^{(e)} \psi_i^{(e)}(s) \, ds \quad \ldots(2.14.21)$$

when $q_n^{(e)}$ is a known function of the distance $s$ along the boundary $\Gamma^e$. It is not necessary to compute such integrals when a portion of $\Gamma^e$ does not coincide with the boundary $\Gamma$ of the domain $\Omega$.

The boundary $\Gamma^e$ of linear two-dimensional elements is a set of linear one-dimensional elements. Therefore, the evaluation of the boundary integral amounts to evaluating line integrals. It should not be surprising to the reader that when two-dimensional interpolation functions are restricted to (i.e. evaluated on) the boundary, we get the corresponding one-dimensional interpolation functions. To fix the ideas, consider a finite element that has a portion of its boundary on the boundary of the domain and assume that $q_n$ is known here. Then

$$\int_0^h q_n(s) \psi_i(s) \, ds \equiv Q_i \quad \ldots(2.14.22)$$

gives the contribution of $q_n$ to node $i$. Here $h$ denotes the length of the side that is subjected to the force $q_n$, and $\psi_i(s)$ are the one-dimensional interpolation functions. When $\psi_i(x,y)$ are linear [then $\psi_i(s)$ are linear], $i$ takes the values of 1 and 2, and when $\psi_i(x,y)$ are quadratic [then $\psi_i(s)$ are
quadratic, \( i \) takes the values of 1, 2, and 3. For example, when \( \psi_i \) are linear (\( \psi_1 = 1 - s/h, \psi_2 = s/h \)), we have

\[
Q_i = \begin{cases} 
\int_0^h q_n(s) \psi_i(s) \, ds & \text{for any } q_n \\ 
\frac{q_n h}{2} & \text{for } q_n \text{ constant}
\end{cases}
\]...(2.14.23)

2.14.6 Assembly Of Element Matrices:

The assembly of finite element equation is based on the same principle as that we discuss in section (2.13), one dimensional problem.

![Assembly of Element Matrices](image)

**Figure (2.14.6)**

**Assembly of Element Matrices**

Here assembled co-efficient matrix for the four-element is given by
The assembled force vector are given by

\[
\{ F \} = \begin{bmatrix}
F_1^{(1)} \\
F_2^{(1)} + F_1^{(2)} \\
F_2^{(2)} \\
F_4^{(1)} + F_1^{(3)} \\
F_3^{(1)} + F_4^{(2)} + F_2^{(3)} + F_1^{(4)} \\
F_3^{(2)} + F_2^{(4)} \\
F_4^{(3)} \\
F_3^{(3)} + F_4^{(4)} \\
F_3^{(4)}
\end{bmatrix}
\]

The remaining procedure stays same as one dimensional problem.