Solving Nonlinear Algebraic Equations II

Nonlinear in Parameter Models: Gauss-Newton Method

In many problems the parameters appear nonlinearly in the model

$$\hat{y}_i = f\left(\mathbf{x}^{(i)} ; \theta_1, \ldots, \theta_m\right) \quad ; \quad (i = 1, 2, \ldots N) \qquad \text{---- (234)}$$

or in the vector notation

$$\hat{\mathbf{y}} = F[\mathbf{X}, \theta] \qquad \text{---- (235)}$$

where

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 & \hat{y}_2 & \ldots & \hat{y}_N \end{bmatrix}^T \qquad \text{---- (236)}$$

$$F = \begin{bmatrix} f\left(\mathbf{x}^{(1)}, \theta\right) & f\left(\mathbf{x}^{(2)}, \theta\right) & \ldots & f\left(\mathbf{x}^{(N)}, \theta\right) \end{bmatrix}^T \qquad \text{---- (237)}$$

and $\mathbf{X} = \left\{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(N)}\right\}$ represents data set. The problem is to determine vector $\hat{\theta}$ such that

$$\Psi = \mathbf{e}^T \mathbf{W} \mathbf{e} \qquad \text{---- (238)}$$

$$\mathbf{e} = \hat{\mathbf{y}} - F(\mathbf{X}, \theta) \qquad \text{---- (239)}$$

is minimized. Note that. in general, the above problem cannot be solved analytically and we have to resort to iterative procedures. There are three solution approaches:

- Approximate solution using weighted least square when the model is analytically linearizable: In many situations, it is possible to use some transformation of the original model to a linear in parameter form. For example, the non-linear in parameter model given by equation (213) was transformed to the following linear in parameter form

$$\log(-r_A) = \log(k_o) + n \log C_A - \frac{E}{R}\left(\frac{1}{T}\right)$$

  After linearizing transformation, the theory developed in the previous section can be used for parameter estimation.

- Gauss-Newton method or *successive linear least square* approach
- Use of direct optimization (nonlinear programming)

The first two approaches use the linear least square formulation as basis while the non-linear programming approaches is a separate class of algorithms. In this sub-section, we only present details of the Gauss-Newton method in detail.

This approach is iterative. Start with an initial guess vector $\theta^{(0)}$. By some process, generate improved guess $\theta^{(k)}$ from $\theta^{(k-1)}$. At k^{th} iteration let $\theta^{(k-1)}$ be the guess solution. By expanding the model as Taylor series in the neighborhood of $\theta = \theta^{(k-1)}$ and neglecting higher order terms we have

---- (240)

$$\tilde{\mathbf{y}}^{(k)} \simeq F\left(\mathbf{X}, \theta^{(k-1)}\right) + \left[\frac{\partial F}{\partial \theta}\right]_{\theta = \theta^{(k-1)}} \left(\Delta \theta^{(k)}\right)$$

where

$$\mathbf{J}^{(k-1)} = \left[\frac{\partial F}{\partial \theta}\right] \qquad \text{---- (241)}$$

is a $(N \times m)$ matrix with elements

$$\left[\frac{\partial F}{\partial \theta}\right]_{ij} = \left[\frac{\partial F\left(\mathbf{x}^{(i)}, \theta\right)}{\partial \theta_j}\right]_{\theta=\theta^{(k-1)}} \qquad \text{---- (242)}$$

$$i = 1, \ldots N \text{ and } j = 1, \ldots m \qquad \text{---- (243)}$$

Let us denote

$$\mathbf{J}^{(k-1)} = \left[\frac{\partial F}{\partial \theta}\right]_{\theta=\theta^{(k-1)}} \qquad \text{---- (244)}$$

and

$$F^{(k-1)} = F\left(\mathbf{X}, \theta^{(k-1)}\right) \qquad \text{---- (245)}$$

Then approximate error vector at k^{th} iteration can be defined as

$$\widetilde{\mathbf{e}}^{(k)} = \mathbf{y} - \widetilde{\mathbf{y}}^{(k)} = \left[\mathbf{y} - F^{(k-1)}\right] - \mathbf{J}^{(k-1)} \, \Delta\theta^{(k)} \qquad \text{---- (246)}$$

and k^{th} linear sub-problem is defined as

$$\min_{\Delta\theta^{(k)}} \left[\widetilde{\mathbf{e}}^{(k)}\right]^T \mathbf{W} \, \widetilde{\mathbf{e}}^{(k)} \qquad \text{---- (247)}$$

The least square solution to above sub problem can be obtained by solving the normal equation

$$\left(\mathbf{J}^{(k-1)}\right)^T \mathbf{W} \, \mathbf{J}^{(k-1)} \Delta\theta^{(k)} = \left(\mathbf{J}^{(k-1)}\right)^T \mathbf{W} \left[\mathbf{y} - F^{(k-1)}\right] \qquad \text{---- (248)}$$

$$\Delta\theta^{(k)} = \left[\left(\mathbf{J}^{(k-1)}\right)^T \mathbf{W} \, \mathbf{J}^{(k-1)}\right]^{-1} \left(\mathbf{J}^{(k-1)}\right)^T \mathbf{W} \left[\mathbf{y} - F^{(k-1)}\right] \qquad \text{---- (249)}$$

and an improved guess can be obtained as

$$\theta^{(k)} = \theta^{(k-1)} + \Delta\theta^{(k)} \qquad \text{---- (250)}$$

Termination criterion : Defining $\mathbf{e}^{(k)} = \mathbf{y} - F^{(k)}$ and

$$\Phi^{(k)} = \left[\mathbf{e}^{(k)}\right]^T \mathbf{W} \, \mathbf{e}^{(k)} \qquad \text{---- (251)}$$

terminate iterations when $\Phi^{(k)}$ changes only by a small amount, i.e.

$$\frac{\left| \Phi^{(k)} - \Phi^{(k-1)} \right|}{\left| \Phi^{(k)} \right|} < \varepsilon \qquad \text{---- (252)}$$

ODE-BVP / PDE Discretization using Minimum Residual Methods

In interpolation based methods, we force residuals to zero at a finite set of collocation points. Based on the least squares approach discussed in this section, one can think of constructing an approximation so that the residual becomes small (in some sense) on the entire domain. Thus, given a ODE-BVP / PDE, we seek an approximate solution as linear combination of finite number of linearly independent functions. Parameters of this approximation are determined in such a way that some norm of the residuals is minimized. There are many discretization methods that belong to this broad class. In this section, we provide a brief introduction to these discretization approaches.

Raleigh-Ritz method

To understand the motivation for developing this approach, first consider a linear system of equations

$$\mathbf{Ax} = \mathbf{b} \qquad \text{---- (253)}$$

where $\mathbf{A}$ is a $n \times n$ positive definite and symmetric matrix and it is desired to solve for vector $\mathbf{x}$. We can pose this as a minimization problem by defining an objective function of the form

$$\phi(\mathbf{x}) = (1/2)\mathbf{x}^T\mathbf{Ax} - \mathbf{x}^T\mathbf{b} \qquad \text{---- (254)}$$

$$= (1/2)\langle \mathbf{x}, \mathbf{Ax}\rangle - \langle \mathbf{x}, \mathbf{b}\rangle \qquad \text{---- (255)}$$

If $\phi(\mathbf{x})$ minimum at $\mathbf{x} = \mathbf{x}^*$, then the necessary condition for optimality requires

$$\partial\phi/\partial\mathbf{x} = \mathbf{Ax}^* - \mathbf{b} = \overline{0} \qquad \text{---- (256)}$$

which is precisely the equation we want to solve. Since the Hessian matrix

$$\partial^2\phi/\partial\mathbf{x}^2 = \mathbf{A}$$

is positive definite, the solution of $\mathbf{x} = \mathbf{x}^*$ of $\mathbf{Ax} = \mathbf{b}$ is the global minimum of objective function $\phi(\mathbf{x})$.

In the above demonstration, we were working in space R^n. Now, let us see if a similar formulation can be worked out in another space, namely $C^{(2)}[0,1]$, i.e. the set of twice differentiable continuous functions on $[0,1]$. Consider ODE-BVP

$$Lu = -\frac{d^2u}{dz^2} = f(z) \qquad \text{---- (257)}$$

$$B.C.\ 1 : u(0) = 0 \qquad \text{---- (258)}$$

$$B.C.\ 2 : u(1) = 0 \qquad \text{---- (259)}$$

Similar to the linear *operator* (matrix) $\mathbf{A}$, which operates on vector $\mathbf{x} \in R^n$ to produce another vector $\mathbf{b} \in R^n$, the linear operator $\mathbf{L} = \left[-d^2/dz^2\right]$ operates on vector $u(z) \in C^{(2)}[0,1]$ to produce $f(z) \in C[0,1]$. Note that the matrix $\mathbf{A}$ in our motivating example is symmetric and positive definite, i.e.

$$\langle \mathbf{x}, \mathbf{Ax}\rangle > 0 \text{ for all } \mathbf{x} \neq \overline{0}$$
$$\text{and } \mathbf{A}^T = \mathbf{A}$$

In order to see how the concept of symmetric matrix can be generalized to operators on infinite dimensional spaces, let us first define adjoint of a matrix.

Definition 30 (Adjoint of matrix): A matrix $\mathbf{A}^*$ is said to be adjoint of matrix $\mathbf{A}$ if it satisfies $\langle \mathbf{x}, \mathbf{Ay}\rangle = \langle \mathbf{A}^*\mathbf{x}, \mathbf{y}\rangle$. Further, the matrix $\mathbf{A}$ is called self adjoint if $\mathbf{A}^* = \mathbf{A}$.

When matrix $\mathbf{A}$ has all real elements, we have

$$\mathbf{x}^T(\mathbf{Ay}) = (\mathbf{A}^T\mathbf{x})^T\mathbf{y}$$

and it is easy to see that $\mathbf{A}^* = \mathbf{A}^T$, i.e.

$$\langle \mathbf{x}, \mathbf{A}\mathbf{y} \rangle = \left\langle \mathbf{A}^T \mathbf{x}, \mathbf{y} \right\rangle \qquad \text{---- (260)}$$

The matrix $\mathbf{A}$ is called *self-adjoint* if $\mathbf{A}^T = \mathbf{A}$. Does operator $\mathbf{L}$ defined by equations (257-259) have some similar properties of *symmetry* and *positiveness*? Analogous to the concept of adjoint of a matrix, we first introduce the concept of adjoint of an operator $\mathbf{L}$ on any inner product space.

Definition 31 (Adjoint of Operator): An operator $\mathbf{L}^*$ is said to be adjoint of operator L if it satisfies

$$\langle v, \mathbf{L}u \rangle = \langle \mathbf{L}^* v, u \rangle$$

Further, the operator L is said to be self-adjoint, if $\mathbf{L}^* = \mathbf{L}$, $B.C.1^* = B.C.1$ and $B.C.2^* = B.C.2$.

To begin with, let us check whether the operator L defined by equations (257-259) is self-adjoint.

$$\langle v, \mathbf{L}u \rangle = \int_0^1 v(z)(-d^2u/dz^2)dz$$

$$= \left[-v(z)\frac{du}{dz} \right]_0^1 + \int_0^1 \frac{dv}{dz}\frac{du}{dz}dz$$

$$= \left[-v(z)\frac{du}{dz} \right]_0^1 + \left[\frac{dv}{dz}u(z) \right]_0^1 + \int_0^1 \left(-\frac{d^2v}{dz^2} \right) u(z)dz$$

Using the boundary conditions $u(0) = u(1) = 0$, we have

$$\left[\frac{dv}{dz}u(z) \right]_0^1 = \frac{dv}{dz}u(1) - \frac{dv}{dz}u(0) = 0$$

If we set

$$B.C.1^* : v(0) = 0$$
$$B.C.2^* : v(1) = 0$$

then

$$\left[\frac{du}{dz}v(z) \right]_0^1 = 0$$

and we have

$$\langle v, \mathbf{L}u \rangle = \int_0^1 \left(-\frac{d^2v}{dz^2} \right) u(z)dz = \langle L^* v, u \rangle$$

In fact, it is easy to see that the operator $\mathbf{L}$ is self adjoint as $\mathbf{L}^* = \mathbf{L}$, $BC1^* = BC1$ and $BC2^* = BC2$. In addition to the self-adjointness of $\mathbf{L}$, we have

$$\langle u, \mathbf{L}u \rangle = \left[-u(z)\frac{du}{dz} \right]_0^1 + \int_0^1 \left(\frac{du}{dz} \right)^2 dz$$

$$= \int_0^1 \left(\frac{du}{dz} \right)^2 dz > 0 \text{ for all } u(z)$$

when $u(z)$ is a non-zero vector in $C^{(2)}[0,1]$. In other words, solving the ODE-BVP is analogous to solving $\mathbf{Ax} = \mathbf{b}$ by optimization formulation where A is symmetric and positive definite matrix, i.e.

$$\mathbf{A} \leftrightarrow \left[-d^2/dz^2 \right]; \quad \mathbf{x} \leftrightarrow u(z); \; \mathbf{b} \leftrightarrow f(z)$$

Let $u(z) = u^*(z)$ represent the true solution of the ODE-BVP. Now, taking motivation from the optimization formulation for solving $Ax = \mathbf{b}$, we can formulate a minimization problem to compute the solution

$$\phi[u(z)] = (1/2)\left\langle u(z), -d^2u/dz^2 \right\rangle - \langle u(z), f(z) \rangle \qquad \text{---- (261)}$$

$$= 1/2 \int_0^1 u(z)(-d^2u/dz^2)dz - \int_0^1 u(z)f(z)dz \qquad \text{---- (262)}$$

$$u^*(z) = \underset{u(z)}{Min} \quad \phi[u(z)] \qquad \text{---- (263)}$$

$$= \underset{u(z)}{Min} \quad (1/2)\langle u(z), Lu(z) \rangle - \langle u(z), f(z) \rangle \qquad \text{---- (264)}$$

$$u(z) \in C^{(2)}[0,1] \qquad \text{---- (265)}$$

$$\text{subject to } \; u(0) = u(1) = 0$$

Thus, solving the $ODE - BVP$ has been converted to solving a minimization problem. Integrating the first term in equation (262) by parts, we have

$$\int_0^1 u(z)\left(-\frac{d^2u}{dz^2} \right)dz = \int_0^1 \left(\frac{du}{dz} \right)^2 dz - \left[u\frac{du}{dz} \right]_0^1 \qquad \text{---- (266)}$$

Now, using boundary conditions, we have

$$\left[u\frac{du}{dz} \right]_0^1 = \left[u(0)\left(\frac{du}{dz} \right)_{z=0} - u(1)\left(\frac{du}{dz} \right)_{z=1} \right] = 0 \qquad \text{---- (267)}$$

This reduces $\phi(u)$ to

$$\phi(u) = \left[1/2 \int_0^1 \left(\frac{du}{dz} \right)^2 dz \right] - \left[\int_0^1 uf(z)dz \right] \qquad \text{---- (268)}$$

The above equation is similar to an *energy function,* where the first term is analogous to kinetic energy and the second term is analogous to potential energy. As

$$\int_0^1 \left(\frac{du}{dz}\right)^2 dz$$

is *positive and symmetric*, we are guaranteed to find the minimum. The main difficulty in performing the search is that, unlike the previous case where we were working in R^n, the search space is infinite dimensional as $u(z) \in C^{(2)}[0,1]$. One remedy to alleviate this difficulty is to reduce the infinite dimensional search problem to a finite dimensional search space by constructing an approximate solution using n trial functions. Let $v^{(1)}(z),.....,v^{(n)}(z)$ represent the trial functions. Then, the approximate solution is constructed as follows

$$\widehat{u}(z) = \alpha_0 v^{(0)}(z) + ... + \alpha_n v^{(n)}(z) \qquad \text{---- (269)}$$

where $v^{(i)}(z)$ represents *trial functions*. Using this approximation, we convert the infinite dimensional optimization problem to a finite dimensional optimization problem as follows

$$\underset{\theta}{Min} \quad \widehat{\phi}(\theta) = \left[1/2 \int_0^1 \left(\frac{d\widehat{u}}{dz}\right)^2 dz \right] - \left[\int_0^1 \widehat{u} f(z) dz \right] \qquad \text{---- (270)}$$

$$= 1/2 \int_0^1 \left[\alpha_0 \left(dv^{(0)}(z)/dz\right) + ... + \alpha_n (dv^{(n)}(z)/dz)\right]^2 dz$$

$$- \int_0^1 f(z)[\alpha_0 v^{(0)}(z) + ... + \alpha_n v^{(n)}(z)] dz \qquad \text{---- (271)}$$

The trial functions $v^{(i)}(z)$ are chosen in advance and coefficients $\alpha_1, ...\alpha_m$ are treated as unknown. Also, let us assume that these functions are selected such that $\widehat{u}(0) = \widehat{u}(1) = 0$. Then, using the necessary conditions for optimality, we get

$$\frac{\partial \widehat{\phi}}{\partial \alpha_i} = 0 \quad for \quad i = 0, 2, ... n \qquad \text{---- (272)}$$

These equations can be rearranged as follows

$$\frac{\partial \widehat{\phi}}{\partial \theta} = \mathbf{A}\theta^* - \mathbf{b} = \overline{\mathbf{0}} \qquad \text{---- (273)}$$

where

$$\theta = \begin{bmatrix} \alpha_0 & \alpha_1 & ... & \alpha_n \end{bmatrix}^T$$

$$\mathbf{A} = \begin{bmatrix} \left\langle \frac{dv^{(0)}}{dz}, \frac{dv^{(0)}}{dz} \right\rangle & .... & \left\langle \frac{dv^{(0)}}{dz}, \frac{dv^{(n)}}{dz} \right\rangle \\ .... & .... & .... \\ \left\langle \frac{dv^{(n)}}{dz}, \frac{dv^{(0)}}{dz} \right\rangle & .... & \left\langle \frac{dv^{(n)}}{dz}, \frac{dv^{(n)}}{dz} \right\rangle \end{bmatrix} \qquad \text{---- (274)}$$

---- (275)

$$\mathbf{b} = \begin{bmatrix} \langle v^{(1)}(z), f(z) \rangle \\ \ldots\ldots\ldots\ldots \\ \langle v^{(n)}(z), f(z) \rangle \end{bmatrix}$$

Thus, the optimization problem under consideration can be recast as follows

$$\underset{\theta}{Min}\ \widehat{\phi}(\theta) = \underset{\theta}{Min}\left[(1/2)\theta^T\mathbf{A}\theta - \theta^T\mathbf{b}\right] \qquad \text{---- (276)}$$

It is easy to see that matrix $\mathbf{A}$ is positive definite and symmetric and the global minimum of the above optimization problem can be found by using necessary condition for optimality i.e. $\partial\widehat{\phi}/\partial\theta = \mathbf{A}\theta^* - \mathbf{b} = \overline{\mathbf{0}}$ or $\theta^* = \mathbf{A}^{-1}\mathbf{b}$. Note the similarity of the above equation with the normal equation arising from the projection theorem. Thus, steps in the Raleigh-Ritz method can be summarized as follows

1. Choose an approximate solution.
2. Compute matrix $\mathbf{A}$ and vector $\mathbf{b}$
3. Solve for $\mathbf{A}\theta = \mathbf{b}$

Method of Least Squares

This is probably best known minimum residual method. When used for solving linear operator equations, this approach does not require self adjointness of the linear operator. To understand the method, let us first consider a linear ODE-BVP

$$\mathbf{L}[u(z)] = f(z) \qquad \text{---- (277)}$$

$$B.C.1 : u(0) = 0 \qquad \text{---- (278)}$$

$$B.C.2 : u(1) = 0 \qquad \text{---- (279)}$$

Consider an approximate solution constructed using linear combination of set of finite number of linearly independent functions as follows

$$\widehat{u}(z) = \alpha_1\widehat{u}_1(z) + \alpha_2\widehat{u}_2(z) + \ldots + \alpha_n\widehat{u}_n(z)$$

Let us assume that these basis functions are selected such that the two boundary conditions are satisfied, i.e. $\widehat{u}_i(0) = \widehat{u}_i(1) = 0$. Given this approximate solution, the residual is defined as follows

$$R(z) = \mathbf{L}[\widehat{u}(z)] - f(z) \quad \text{where} \quad 0 < z < 1$$

The idea is to determine

$$\alpha = \begin{bmatrix} \alpha_1 & \alpha_2 & \ldots & \alpha_n \end{bmatrix}^T$$

such that

$$\underset{\alpha}{Min}\ \phi(\alpha) = \langle R(z), R(z) \rangle$$

$$\langle R(z), R(z) \rangle = \int_0^1 \omega(z)R(z)^2 dz$$

where $\omega(z)$ is a positive function on $0 < z < 1$. This minimization problem leads to a generalized normal form of equation

$$\begin{bmatrix} \langle \mathbf{L}\hat{u}_1, \mathbf{L}\hat{u}_1 \rangle & \langle \mathbf{L}\hat{u}_1, \mathbf{L}\hat{u}_2 \rangle & .... & \langle \mathbf{L}\hat{u}_1, \mathbf{L}\hat{u}_n \rangle \\ \langle \mathbf{L}\hat{u}_2, \mathbf{L}\hat{u}_1 \rangle & \langle \mathbf{L}\hat{u}_2, \mathbf{L}\hat{u}_2 \rangle & .... & \langle \mathbf{L}\hat{u}_2, \mathbf{L}\hat{u}_n \rangle \\ ..... & ..... & ..... & ..... \\ \langle \mathbf{L}\hat{u}_n, \mathbf{L}\hat{u}_1 \rangle & \langle \mathbf{L}\hat{u}_n, \mathbf{L}\hat{u}_2 \rangle & ..... & \langle \mathbf{L}\hat{u}_n, \mathbf{L}\hat{u}_n \rangle \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ .... \\ \alpha_m \end{bmatrix} = \begin{bmatrix} \langle \mathbf{L}\hat{u}_1, f(z) \rangle \\ \langle \mathbf{L}\hat{u}_2, f(z) \rangle \\ .... \\ \langle \mathbf{L}\hat{u}_n, f(z) \rangle \end{bmatrix} \quad \text{---- (280)}$$

which can be solved analytically.

Example 32 Use the least squares method to find an approximate solution of the equation

$$\mathbf{L}[u(z)] = \frac{\partial^2 u}{\partial z^2} - u = 1 \quad \text{---- (281)}$$

$$B.C.\ 1 : u(0) = 0 \quad \text{--- (282)}$$

$$B.C.\ 2 : u(1) = 0 \quad \text{--- (283)}$$

Let us select the function expansion as

$$\hat{u}(z) = \alpha_1 \sin(\pi z) + \alpha_2 \sin(2\pi z)$$

It may be noted that this choice ensures that the boundary conditions are satisfied. Now,

$$\mathbf{L}[\hat{u}_1(z)] = -(\pi^2 + 1)\sin(\pi z)$$

$$\mathbf{L}[\hat{u}_2(z)] = -(4\pi^2 + 1)\sin(2\pi z)$$

With the inner product defined as

$$\langle f, g \rangle = \int_0^1 f(z) g(z) dz$$

the normal equation becomes

$$\begin{bmatrix} \frac{(\pi^2+1)^2}{2} & 0 \\ 0 & \frac{(4\pi^2+1)^2}{2} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \frac{-2(\pi^2+1)}{\pi} \\ 0 \end{bmatrix}$$

and the approximate solution is

$$\hat{u}(z) = -\frac{4}{\pi(\pi^2+1)} \sin(\pi z)$$

which agrees with the exact solution

$$u(z) = \frac{e^z + e^{1-z}}{(e+1)} - 1$$

to within 0.006.

When boundary conditions are non-homogeneous, it is some times possible to transform them to homogeneous conditions. Alternatively, the optimization problem is formulated in such a way that the boundary conditions are satisfied in the least square sense . While this method can be, in principle, extended to discretization of general ODE-BVP of type (32-34a), working with parameter vector α as minimizing argument can pose practical difficulties as the resulting minimization problem has to be solved numerically. Coming up with initial guess of α to start the iterative algorithms can prove to be a tricky task. Alternatively, one can work with trial solutions of the form (310) or (326) to make the problem computationally tractable.

Gelarkin's Method

The Gelarkin's method can be applied for any problem where differential operator is not self adjoint or symmetric. Instead of minimizing $\phi(\widehat{\mathbf{u}})$, we solve for

$$\left\langle v^{(i)}(z), \mathbf{L}\widehat{u}(z)\right\rangle = \left\langle v^{(i)}(z), f(z)\right\rangle$$
$$i = 1,2,....,n$$

where $\widehat{u}(z)$ is chosen as finite dimensional approximation to $u(z)$

$$u(z) = u_1 v^{(1)}(z) + ...+ u_n v^{(n)}(z) \quad ---- (284)$$

Rearranging above equations as

$$\left\langle v^{(i)}(z), (\mathbf{L}\widehat{u}(z) - f(z))\right\rangle = 0 \text{ for } (i = 1,2,....n)$$

we can observe that parameters $u_1,.....u_n$ are computed such that the error or residual vector

$$e(z) = (\mathbf{L}\widehat{u}(z) - f(z))$$

is orthogonal to the (n) dimensional subspace spanned by set S defined as

$$S = \left\{ v^{(i)}(z) \ : \ i = 1,2,....n \right\}$$

This results in a linear algebraic equation of the form

$$\mathbf{A}\widehat{\mathbf{u}} = \mathbf{b} \quad ---- (285)$$

where

$$\mathbf{A} = \begin{bmatrix} \left\langle v^{(1)}, L(v^{(1)})\right\rangle & .... & \left\langle v^{(1)}, L(v^{(n)})\right\rangle \\ .... & .... & .... \\ \left\langle v^{(n)}, L(v^{(1)})\right\rangle & .... & \left\langle v^{(n)}, L(v^{(n)})\right\rangle \end{bmatrix} \quad ---- (286)$$

$$\mathbf{b} = \begin{bmatrix} \left\langle v^{(1)}(z), f(z)\right\rangle \\ .... \\ \left\langle v^{(n)}(z), f(z)\right\rangle \end{bmatrix}$$

Solving for $\widehat{\mathbf{u}}$ gives approximate solution given by equation (284).When the operator is $\mathbf{L}$ self adjoint, the Gelarkin's method reduces to the Raleigh-Ritz method.

Example 33 Consider ODE-BVP

$$\mathbf{L}u = \partial^2 u/\partial z^2 - \partial u/\partial z = f(z) \quad ----(287)$$
$$in \ (0 < z < 1) \quad --- (288)$$
$$subject\ to\ u(0) = 0;\ u(1) = 0 \quad --- (289)$$

It can be shown that

$$L^*(= \partial^2/\partial z^2 + \partial/\partial z) \neq (\partial^2/\partial z^2 - \partial/\partial z) = L$$

Thus, Raleigh-Ritz method cannot be applied to generate approximate solution to this problem, however, Gelarkin's method can be applied.

It may be noted that one need not restrict to linear transformations while applying the Gelarkin's method. This approach can be used even when the ODE-BVP or PDE at hand is a nonlinear transformation. Given a general nonlinear transformation of the form

$$\mathcal{T}(u) = f(z)$$

we select a set of trial function $\left\langle v^{(i)}(z) : i = 0, 1, \ldots n \right\rangle$ and an approximate solution of the form (284) and solve for

$$\left\langle v^{(i)}(z), \mathcal{T}(u(z)) \right\rangle = \left\langle v^{(i)}(z), f(z) \right\rangle \quad \text{for} \quad i = 0, 1, 2, \ldots n$$

Example 34 Use the Gelarkin's method to find an approximate solution of the equation

$$\mathbf{L}[u(z)] = \frac{\partial^2 u}{\partial z^2} - u = 1 \qquad \text{---- (290)}$$

$$B.C.\ 1 : u(0) = 0 \qquad \text{--- (291)}$$

$$B.C.\ 2 : u(1) = 0 \qquad \text{--- (292)}$$

Let us select the function expansion as follows

$$\widehat{u}(z) = \alpha_1 \sin(\pi z) + \alpha_2 \sin(2\pi z)$$

which implies

$$\mathbf{L}[\widehat{u}(z)] = -\alpha_1(\pi^2 + 1)\sin(\pi z) - \alpha_2(4\pi^2 + 1)\sin(2\pi z)$$

With the inner product defined as

$$\langle f, g \rangle = \int_0^1 f(z) g(z) dz$$

the normal equation becomes

$$\begin{bmatrix} \frac{-(\pi^2+1)}{2} & 0 \\ 0 & \frac{-(4\pi^2+1)}{2} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \frac{2}{\pi} \\ 0 \end{bmatrix}$$

and the approximate solution is

$$\widehat{u}(z) = -\frac{4}{\pi(\pi^2 + 1)} \sin(\pi z)$$

which turns out to be identical to the least square solution.

Example 35 Consider the ODE-BVP describing steady state conditions in a tubular reactor with axial mixing (TRAM) in which an irreversible 2nd order reaction is carried out.

$$\mathcal{T}(C) = \frac{1}{Pe}\frac{d^2C}{dz^2} - \frac{dC}{dz} - DaC^2 = 0 \qquad (0 \leq z \leq 1)$$

$$\frac{dC}{dz} = Pe(C-1) \qquad at \quad z = 0;$$
$$\frac{dC}{dz} = 0 \qquad at \quad z = 1;$$

The approximate solution is chosen as

$$\widehat{C}(z) = \widehat{C}_1 \nu^{(1)}(z) + \ldots\ldots + \widehat{C}_{n+1} \nu^{(n+1)}(z) = \sum_{i=1}^{n+1} \widehat{C}_i \nu^{(i)}(z) \qquad \text{---- (293)}$$

and we then evaluate the following set of equations

$$\left\langle \nu^{(i)}(z), \frac{1}{Pe}\frac{d^2\widehat{C}(z)}{dz^2} - \frac{d\widehat{C}(z)}{dz} - Da\widehat{C}(z)^2 \right\rangle = \left\langle \nu^{(i)}(z), f(z) \right\rangle \quad for \quad i = 2, \ldots. n$$

where the inner product is defined as

$$\langle g(z), h(z) \rangle = \int_0^1 g(q)h(q)dq$$

It may be noted that evaluation of integrals, such as

$$\left\langle \nu^{(i)}(z), \widehat{C}(z)^2 \right\rangle = \int_0^1 \nu^{(i)}(q) \left(\sum_{i=0}^{n} \widehat{C}_i \nu^{(i)}(q) \right)^2 dq$$

will give rise to equations that are nonlinear in terms of unknown coefficients. Two additional equations arise from enforcing the boundary conditions. i.e.

$$\frac{d\widehat{C}(0)}{dz} = Pe(\widehat{C}(0) - 1)$$
$$\frac{d\widehat{C}(1)}{dz} = 0$$

Thus, we get (n+1) nonlinear algebraic equations in (n+1) unknowns, which have to be solved simultaneously to compute the unknown coefficients $\widehat{C}_1, \ldots \widehat{C}_{n+1}$.

Discretization of ODE-BVP / PDEs using Finite Element Method

The finite element method is a powerful tool for solving PDEs particularly when the system under consideration has complex geometry. This method is based on the least square approximation. In this section, we provide a very brief introduction to the method discretization of PDEs and ODE-BVPs using the finite element method.

Discretization of ODE-BVP using Finite Element

Similar to finite difference method, we begin by choosing $(n-1)$ equidistant internal node (grid) points as follows

$$z_i = i\Delta z \ \ (i = 0, 1, 2, \ldots. n)$$

and defining n finite elements

$$z_{i-1} \leq z \leq z_i \quad \text{for } i = 1,2,\ldots n$$

Then we formulate the approximate solution using piecewise constant polynomials on each finite element. The simplest possible choice is a line

$$\widehat{u}_i(z) = a_i + b_i z \qquad \text{---- (294)}$$

$$z_{i-1} \leq z \leq z_i \quad \text{for } i = 1,2,\ldots n \qquad \text{---- (295)}$$

With this choice, the approximate solution for the ODE-BVP can be expressed as

$$\widehat{u}(z) = \left\{ \begin{array}{ccc} a_1 + b_1 z & for & z_0 \leq z \leq z_1 \\ a_2 + b_2 z & for & z_1 \leq z \leq z_2 \\ & \ldots\ldots & \\ a_n + b_n z & for & z_{n-1} \leq z \leq z_n \end{array} \right\} \qquad \text{---- (296)}$$

In principle, we can work with this piecewise polynomial approximation. However, the resulting optimization problems has coefficients $(a_i, b_i : i = 1,2,\ldots n)$ as unknowns. If the optimization problem has to be solved numerically, it is hard to generate initial guess for these unknown coefficients. Thus, it is necessary to parameterize the polynomial in terms of unknowns for which it is relatively easy to generate the initial guess. This can be achieved as follows. Let $\widehat{u}_i$ denote the value of the approximate solution $\widehat{u}(z)$ at $z = z_i$, i.e.

$$\widehat{u}_i = \widehat{u}(z_i) \qquad \text{---- (297)}$$

Then, at the boundary points of the i'th element, we have

$$\widehat{u}(z_{i-1}) = \widehat{u}_{i-1} = a_i + b_i z_{i-1} \qquad \text{---- (298)}$$

$$\widehat{u}(z_i) = \widehat{u}_i = a_i + b_i z_i \qquad \text{--- (299)}$$

Using these equations, we can express (a_i, b_i) in terms of unknowns $(\widehat{u}_{i-1}, \widehat{u}_i)$ as follows

$$a_i = \frac{\widehat{u}_{i-1} z_i - \widehat{u}_i z_{i-1}}{\Delta z} \;;\; b_i = \frac{\widehat{u}_i - \widehat{u}_{i-1}}{\Delta z} \qquad \text{---- (300)}$$

Thus, the polynomial on the i'th segment can be written as

$$\widehat{u}_i(z) = \frac{\widehat{u}_{i-1} z_i - \widehat{u}_i z_{i-1}}{\Delta z} + \left(\frac{\widehat{u}_i - \widehat{u}_{i-1}}{\Delta z} \right) z \qquad \text{--- (301)}$$

$$z_{i-1} \leq z \leq z_i \quad \text{for } i = 1,2,\ldots n$$

and the approximate solution can be expressed as follows

$$\widehat{u}(z) = \left\{ \begin{array}{c} \frac{\widehat{u}_0 z_1}{\Delta z} + \left(\frac{\widehat{u}_1 - \widehat{u}_0}{\Delta z} \right) z \quad for \quad z_0 \leq z \leq z_1 \\ \frac{\widehat{u}_1 z_2 - \widehat{u}_2 z_1}{\Delta z} + \left(\frac{\widehat{u}_2 - \widehat{u}_1}{\Delta z} \right) z \quad for \quad z_1 \leq z \leq z_2 \\ \ldots\ldots\ldots\ldots\ldots \\ \frac{\widehat{u}_{n-1} z_n - \widehat{u}_n z_{n-1}}{\Delta z} + \left(\frac{\widehat{u}_n - \widehat{u}_{n-1}}{\Delta z} \right) z \quad for \quad z_{n-1} \leq z \leq z_n \end{array} \right\} \qquad \text{--- (302)}$$

Thus, now we can work in terms of unknown values $\{\widehat{u}_0, \widehat{u}_1, \ldots. \widehat{u}_n\}$ instead of parameters a_i and b_i.

Since unknowns $\{\hat{u}_0, \hat{u}_1, ... \hat{u}_n\}$ correspond to some physical variable, it is relatively easy to generate good guesses for these unknowns from knowledge of the underlying physics of the problem. The resulting form is still not convenient from the viewpoint of evaluating integrals involved in the computation of $\phi[\hat{u}(z)]$. A more elegant and useful form of equation (302) can be found by defining shape functions. To arrive at this representation, consider the rearrangement of the line segment equation on i'th element as follows

$$\begin{aligned} \hat{u}_i(z) &= \frac{\hat{u}_{i-1}z_i - \hat{u}_i z_{i-1}}{\Delta z} + \left(\frac{\hat{u}_i - \hat{u}_{i-1}}{\Delta z}\right)z \\ &= \frac{z_i - z}{\Delta z}\hat{u}_{i-1} + \frac{z - z_{i-1}}{\Delta z}\hat{u}_i \end{aligned} \quad \text{--- (303)}$$

Let us define two functions, $M_i(z)$ and $N_i(z)$, which are called as shape functions, as follows

$$M_i(z) = \frac{z_i - z}{\Delta z} \ ; \ N_i(z) = \frac{z - z_{i-1}}{\Delta z}$$

$$z_{i-1} \leq z \leq z_i \quad \text{for } i = 1, 2, \ldots n$$

The graphs of these shape functions are straight lines and they have fundamental properties

$$M_i(z) = \left\{ \begin{array}{ll} 1 & ; \ z = z_{i-1} \\ 0 & ; \ z = z_i \end{array} \right\} \quad \text{--- (304)}$$

$$N_i(z) = \left\{ \begin{array}{ll} 0 & ; \ z = z_{i-1} \\ 1 & ; \ z = z_i \end{array} \right\} \quad \text{--- (305)}$$

This allows us to express $\hat{u}_i(z)$ as

$$\hat{u}_i(z) = \hat{u}_{i-1} M_i(z) + \hat{u}_i N_i(z)$$

$$i = 1, 2, \ldots n$$

Note that the coefficient $\hat{u}_i$ appears in polynomials $\hat{u}_i(z)$ and $\hat{u}_{i+1}(z)$, i.e.

$$\hat{u}_i(z) = \hat{u}_{i-1} M_i(z) + \hat{u}_i N_i(z)$$

$$\hat{u}_{i+1}(z) = \hat{u}_i M_{i+1}(z) + \hat{u}_{i+1} N_{i+1}(z)$$

Thus, we can define a continuous *trial function* by combining $N_i(z)$ and $M_{i+1}(z)$ as follows

$$v^{(i)}(z) = \left\{ \begin{array}{ll} N_i(z) = \dfrac{z - z_{i-1}}{\Delta z} = 1 + \dfrac{z - z_i}{\Delta z} & ; \ z_{i-1} \leq z \leq z_i \\ M_{i+1}(z) = \dfrac{z_{i+1} - z}{\Delta z} = 1 - \dfrac{z - z_i}{\Delta z} & ; \ z_i \leq z \leq z_{i+1} \\ 0 & \text{Elsewhere} \end{array} \right\} \quad \text{--- (306)}$$

$$i = 1, 2, \ldots. n$$

This yields the simplest and most widely used *hat function*, which is shown in Figure 10 . This is a continuous linear function of z, but, it is not differentiable at $z_{i-1}, z_i,$ and z_{i+1}. Also, note that at $z = z_i$, we have

--- (307)

$$v^{(i)}(z_j) = \begin{Bmatrix} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{Bmatrix}$$

$$j = 1, 2, \ldots, n$$

Thus, plot of this function looks like a symmetric triangle. The two functions at the boundary points are defined as ramps

$$v^{(0)}(z) = \begin{Bmatrix} M_1(z) = 1 - \frac{z}{\Delta z} & ; \; 0 \leq z \leq z_1 \\ 0 & \text{Elsewhere} \end{Bmatrix} \quad \text{--- (308)}$$

$$v^{(n)}(z) = \begin{Bmatrix} N_n(z) = 1 + \frac{z - z_n}{\Delta z} & ; \; z_{n-1} \leq z \leq z_n \\ 0 & \text{Elsewhere} \end{Bmatrix} \quad \text{--- (309)}$$

Introduction of these trial functions allows us to express the approximate solution as

$$\widehat{u}(z) = \widehat{u}_0 v^{(0)}(z) + \ldots\ldots + \widehat{u}_n v^{(n)}(z) \quad \text{--- (310)}$$

$$0 \leq z \leq 1$$

and now we can work with $\widehat{\mathbf{u}} = \begin{bmatrix} \widehat{u}_0 & \widehat{u}_1 & \ldots & \widehat{u}_n \end{bmatrix}^T$ as unknowns. Now, we have two boundary conditions, i.e.

$$\widehat{u}_0 = 0 \text{ and } \widehat{u}_n = 0$$

and the set of unknowns is reduced to $\widehat{\mathbf{u}} = \begin{bmatrix} \widehat{u}_1 & \widehat{u}_2 & \ldots & \widehat{u}_{n-1} \end{bmatrix}^T$. The optimum parameters $\widehat{\mathbf{u}}$ can be computed by solving equation

$$\mathbf{A}\widehat{\mathbf{u}} - \mathbf{b} = \overline{0} \quad \text{--- (311)}$$

where

$$(\mathbf{A})_{ij} = \left\langle \frac{dv^{(i)}}{dz}, \frac{dv^{(j)}}{dz} \right\rangle \quad \text{--- (312)}$$

and

$$\frac{dv^{(i)}}{dz} = \begin{Bmatrix} 1/\Delta z & \text{on interval left of } z_i \\ -1/\Delta z & \text{on interval right of } z_i \end{Bmatrix}$$

If intervals do not overlap, then

$$\left\langle \frac{dv^{(i)}}{dz}, \frac{dv^{(j)}}{dz} \right\rangle = 0 \quad \text{--- (313)}$$

The intervals overlap when

$$i = j \; : \; \left\langle \frac{dv^{(i)}}{dz}, \frac{dv^{(i)}}{dz} \right\rangle = \int_{z_{i-1}}^{z_i} (1/\Delta z)^2 dz + \int_{z_i}^{z_{i+1}} (-1/\Delta z)^2 dz = 2/\Delta z \quad \text{--- (314)}$$

or

$$i = j + 1 : \left\langle \frac{dv^{(i)}}{dz}, \frac{dv^{(i-1)}}{dz} \right\rangle = \int_{z_{i-1}}^{z_i} (1/\Delta z).(-1/\Delta z)dz = -1/\Delta z \quad \text{--- (315)}$$

$$i = j - 1 : \left\langle \frac{dv^{(i)}}{dz}, \frac{dv^{(i+1)}}{dz} \right\rangle = \int_{z_i}^{z_{i+1}} (1/\Delta z).(-1/\Delta z)dz = -1/\Delta z \quad \text{--- (316)}$$

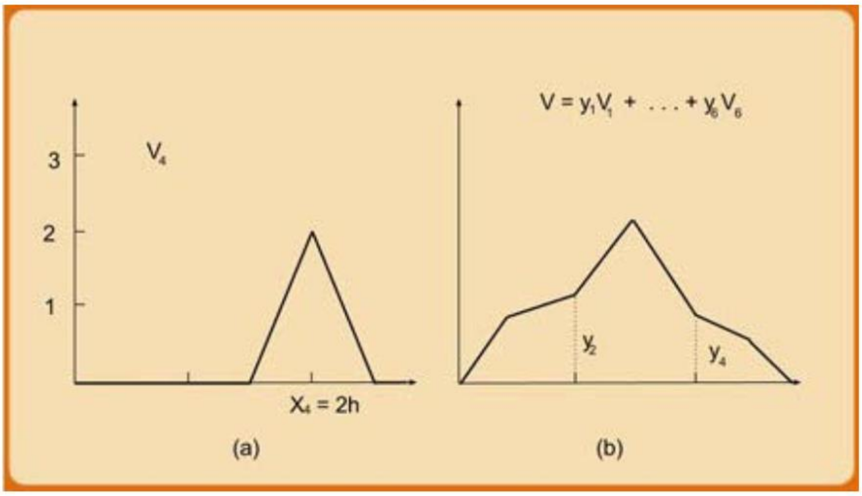

Figure 10: (a) Trial functions and (b) Piece-wise linear approximation

Thus, the matrix $\mathbf{A}$ is a tridiagonal matrix

$$\mathbf{A} = \frac{1}{\Delta z} \begin{bmatrix} 2 & -1 & .... & .... & 0 \\ -1 & 2 & -1 & .... & ... \\ .... & .... & .... & .... & ... \\ 0 & .... & .... & -1 & 2 \end{bmatrix} \quad \text{--- (317)}$$

which is similar to the matrix obtained using finite difference method. The components of vector $\mathbf{b}$ on **the** R.H.S. is computed as

$$b_i = \langle v^{(i)}, f(z) \rangle \quad \text{--- (318)}$$

$$= \int_{z_{i-1}}^{z_i} f(z)\left(1 + \frac{z - z_i}{\Delta z}\right)dz + \int_{z_i}^{z_{i+1}} f(z)\left(1 - \frac{z - z_i}{\Delta z}\right)dz \quad \text{--- (319)}$$

$$i = 1, 2, \ldots, n - 1 \quad \text{---- (320)}$$

which is a weighted average of $f(z)$ over the interval $z_{i-1} \leq z \leq z_{i+1}$. Note that the R.H.S. is significantly different from finite difference method.

In this sub-section, we have developed approximate solution using piecewise linear approximation. It is possible to develop piecewise quadratic or piecewise cubic approximations and generate better approximations.

Discretization of PDE using Finite Element Method

The Raleigh-Ritz method can be easily applied to discretize PDEs when the operators are self-adjoint. Consider Laplace / Poisson's equation

$$\mathbf{L}u = -\partial^2 u/\partial x^2 - \partial^2 u/\partial y^2 = f(x,y) \quad \text{--- (321)}$$

in open set S and $u(x,y) = 0$ on the boundary. Let the inner product on the space $C^{(2)}[0,1] \times C^{(2)}[0,1]$ be defined as

$$\langle f(x,y), g(x,y)\rangle = \int_0^1\int_0^1 f(x,y)\ g(x,y)\ dxdy \qquad \text{--- (322)}$$

We formulate an optimization problem

$$\phi(u) = 1/2\left\langle u(x,y), -\partial^2 u/\partial x^2 - \partial^2 u/\partial y^2 \right\rangle - \langle u(x,y), f(x,y)\rangle \qquad \text{--- (323)}$$

Integrating by parts, we can show that

$$\phi(u) = \int\int [1/2(\partial u/\partial x)^2 + 1/2(\partial u/\partial y)^2 - fu]dxdy \qquad \text{--- (324)}$$

$$= (1/2)\langle \partial u/\partial x, \partial u/\partial x\rangle + 1/2\langle \partial u/\partial y, \partial u/\partial y\rangle - \langle f(x,y), u(x,y)\rangle \qquad \text{--- (325)}$$

We begin by choosing $(n-1)\times(n-1)$ equidistant (with $\Delta x = \Delta y = h$) internal node (grid) points at (x_i, y_j) where

$$x_i = ih \ \ (i = 1, 2, \ldots . n-1)$$

$$y_j = jh \ \ (j = 1, 2, \ldots . n-1)$$

In two dimension, the simplest element divides region into triangles on which simple polynomials are fitted. For example, $u(x,y)$ can be approximated as

$$\widehat{u}(x,y) = a + bx + cy$$

where vertices a, b, c can be expressed in terms of values of $\widehat{u}(x,y)$ at the triangle vertices. For example, consider triangle defined by (x_i, y_j), (x_{i+1}, y_j) and (x_i, y_{j+1}). The value of the approximate solution at the corner points is denoted by

$$\widehat{u}_{i,j} = \widehat{u}(x_i, y_j)\ ; \widehat{u}_{i+1,j} = \widehat{u}(x_{i+1}, y_j)\ ;\ \widehat{u}_{i,j+1} = \widehat{u}(x_i, y_{j+1})$$

Then, $\widehat{u}(x,y)$ can be written in terms of shape functions as follows

$$\widehat{u}(x,y) = \widehat{u}_{i,j} + \frac{\widehat{u}_{i+1,j} - \widehat{u}_{i,j}}{h}(x - x_{i,j}) + \frac{\widehat{u}_{i,j+1} - \widehat{u}_{i,j}}{h}(y - y_{i,j})$$

$$= \widehat{u}_{i,j}\left[1 - \frac{(x - x_{i,j})}{h} - \frac{(y - y_{i,j})}{h}\right] \qquad \text{--- (326)}$$

$$+ \widehat{u}_{i+1,j}\left[\frac{(x - x_{i,j})}{h}\right] + \widehat{u}_{i,j+1}\left[\frac{(y - y_{i,j})}{h}\right]$$

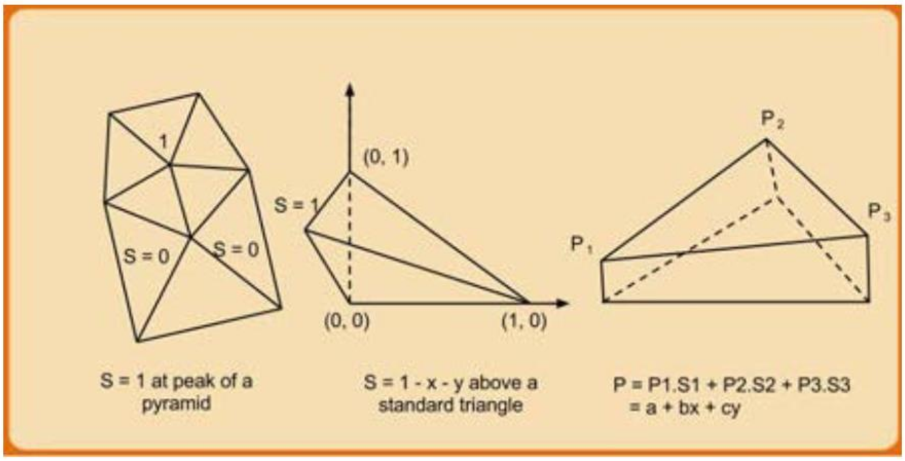

Figure 11: Trial function in two dimensions

Now, coefficient $\widehat{u}_{i,j}$ appears in the shape functions of four triangular element around (x_i, y_j). Collecting these shape functions, we can define a two dimensional trial function as follows

$$v^{(i,j)}(z) = \left\{ \begin{array}{ll} 1 - \frac{(x - x_{i,j})}{h} - \frac{(y - y_{i,j})}{h} & ; \; x_i \le x \le x_{i+1} \; ; \; y_j \le y \le y_{j+1} \\ 1 + \frac{(x - x_{i,j})}{h} - \frac{(y - y_{i,j})}{h} & ; \; x_{i-1} \le x \le x_i \; ; \; y_j \le y \le y_{j+1} \\ 1 - \frac{(x - x_{i,j})}{h} + \frac{(y - y_{i,j})}{h} & ; \; x_i \le x \le x_{i+1} \; ; \; y_{j-1} \le y \le y_j \\ 1 + \frac{(x - x_{i,j})}{h} + \frac{(y - y_{i,j})}{h} & ; \; x_{i-1} \le x \le x_i \; ; \; y_{j-1} \le y \le y_j \\ 0 & \text{Elsewhere} \end{array} \right\}$$

The shape of this trial function is like a pyramid (see Figure 11). We can define trial functions at the boundary points in a similar manner. Thus, expressing the approximate solution using trial functions and using the fact that $\widehat{u}(x,y) = 0$ at the boundary points, we get

$$\widehat{u}(x,y) = \widehat{u}_{1,1} v^{(1,1)}(x,y) + \ldots + \widehat{u}_{n-1,n-1} v^{(n-1,n-1)}(x,y)$$

where $v^{(i,j)}(z)$ represents the (i,j)'th trial function. For the sake of convenience, let us re-number these trial functions and coefficients using a new index $l = 0, 1, \ldots\ldots, N$ such that

$$l = i + (n-1)j$$
$$i = 1, \ldots n-1 \text{ and } j = 0, 1, \ldots n-1$$
$$N = (n-1) \times (n-1)$$

The approximate solution can now be expressed as

$$\widehat{u}(x,y) = \widehat{u}_0 v^0(x,y) + \ldots + \widehat{u}_N v^N(x,y)$$

The minimization problem an be reformulated as

$$\underset{\widehat{\mathbf{u}}}{Min} \; \phi(\widehat{u}) = \underset{\widehat{\mathbf{u}}}{Min} \left[\frac{1}{2} \left\langle \frac{\partial \widehat{u}}{\partial x}, \frac{\partial \widehat{u}}{\partial x} \right\rangle + \frac{1}{2} \left\langle \frac{\partial \widehat{u}}{\partial y}, \frac{\partial \widehat{u}}{\partial y} \right\rangle - \langle f(x,y), \widehat{u}(x,y) \rangle \right]$$

where

$$\widehat{\mathbf{u}} = \left[\begin{array}{cccc} \widehat{u}_0 & \widehat{u}_2 & \ldots & \widehat{u}_N \end{array} \right]^T$$

Thus, the above objective function can be reformulated as

$$\underset{\widehat{\mathbf{u}}}{Min} \; \phi(\widehat{\mathbf{u}}) = \underset{\widehat{\mathbf{u}}}{Min} \left(1/2 \widehat{\mathbf{u}}^T \mathbf{A} \widehat{\mathbf{u}} - \widehat{\mathbf{u}}^T \mathbf{b} \right) \qquad \text{--- (327)}$$

where

$$(\mathbf{A})_{ij} = (1/2) \left\langle \partial v^{(i)} / \partial x, \partial v^{(j)} / \partial x \right\rangle + (1/2) \left\langle \partial v^{(i)} / \partial y, \partial v^{(j)} / \partial y \right\rangle \qquad \text{--- (328)}$$

--- (329)

$$b_i = \langle f(x,y), v^{(i)}(x,y) \rangle$$

Again, the matrix A is symmetric and positive definite matrix and this guarantees that stationary point of $\phi(\mathbf{u})$ is the minimum. At the minimum, we have

$$\partial\phi/\partial\hat{\mathbf{u}} = \mathbf{A}\hat{\mathbf{u}} - \mathbf{b} = 0 \qquad \text{--- (330)}$$

The matrix $\mathbf{A}$ will also be a sparse matrix. The main limitation of Raleigh-Ritz method is that it works only when the operator $\mathbf{L}$ is *symmetric* or self adjoint.

References and cited materials

1. Gilbert Strang, *Linear Algebra and Its Applications* (*4th Ed.*), Wellesley Cambridge Press (2009).

2. Philips, G. M.,Taylor, P. J. ; Theory and Applications of Numerical Analysis (2nd Ed.),Academic Press, 1996.

3. Gourdin,A. and M Boumhrat; Applied Numerical Methods. Prentice Hall (2000).

4. Gupta, S. K.; Numerical Methods for Engineers. Wiley Eastern, New Delhi, 1995.