Chapter 4

Finite Element Methods

Before developing the theory behind finite element analysis, let's give an overview between the finite difference and finite element methods so that we know from hence we came and where we are heading.

Finite Difference Methods

- Utilizes uniformly spaced grids.
- At each node, each derivative is approximated by an algebraic expression which references the adjacent nodes.
- A system of algebraic equations is obtained by evaluating the previous step for each node.
- The system is solved for the dependent variable.

Finite Element Methods

- Utilizes either uniformly or nonuniformly spaced grids.
- Within each element, the change of the dependent variable is approximated by an iterpolation formula.
- The original problem is replaced with some type of equivalent integral expression.
- The interpolation functions are substituted into the integral equation, integrated and combined with results for other elements which yields a set of algebraic equations.
- The system is solved for the dependent variable.

So while the last step is the same, most of the other steps are different. Given the length of the two lists, one would assume (correctly) that the finite element method is a bit more complicated than the finite difference method. What we gain for the price of complexity is the ability to model more complex domains with fewer number of elements and a solid mathematical basis from which to perform our error analysis which leads to methods for automatically adapting our mesh to increase the accuracy of our solution. The finite element method offers a more powerful (but also more costly in the computational sense) method for analyzing science and engineering field problems within complex three dimensional domains. While finite difference methods dominated computational science in the 1960s and 1970s, since the 1980s, finite element methods have been pervasive (I should note that multigrid and multi-level methods seem like they will play a large role in the coming years).

Let's get a taste of the different forms of the finite element method by walking through a simple one dimensional example. Consider the familiar first order differential equation,

$$\frac{du}{dt} + \lambda u(t) = 0 \tag{4.1}$$

which we saw earlier as examples of population growth and Newton's law of cooling. For convenience, we restrict ourselves to be interested in the interval $[0, 1]$ and let $\lambda = 1$. If we further allow the initial condition, $u(t = 0) = 1$, we have the familiar analytical solution:

$$u(t) = e^{-t} = 1 - t + \frac{1}{2!}t^2 - \frac{1}{3!}t^3 + \ldots \tag{4.2}$$

Let's consider a *trial solution* which consists of trying to solve our sample problem by using a power series,

$$\tilde{u}_n(t) = a_0 + a_1 t + a_2 t^2 + \ldots + a_n t^n \tag{4.3}$$

where we have denoted that this is just a trial solution by placing the tilde over u. To satisfy our particular initial condition, we require that $a_0 = 1$. The other coefficients will have to be dealt with later. Now let us restrict our space to the class of functions which are only up to quadratic in order. Obviously, this will cause us to incur some error when compared with the analytic solution. However, since we are only interested in the interval $[0, 1]$, we can perhaps adjust the coefficients to still obtain a good agreement. This is really the idea behind finite element analysis. We can show that by breaking up our domain into a finite subspace and then approximating how the function behaves within that finite subspace, we can approximate the continuous case to a high degree of accuracy. Let's continue with our example and illustrate some of the methods which can assure us this accuracy.

We now substitute our trial solution into our original differential equation and define the quantity, $R(t; a_1, a_2)$ to measure the difference between the trial solution and the true solution,

$$R(t; a_1, a_2) = \frac{d\tilde{u}_2}{dt} + \tilde{u}_2 = 1 + (1 + t)a_1 + (2t + t^2)a_2 \tag{4.4}$$

which is defined as the *residual*. Now the idea (even a reasonable one), is to try and find a method which picks the parameters, a_n such that the residual is as close to zero as possible, i.e. agrees with the true solution.

Perhaps the first way that comes to mind is the *collocation method*. This method requires the residual to vanish at n points $t_1, t_2, \ldots, t_n$, within the domain of interest,

$$R(t_i; \mathbf{a}) = 0 \qquad i = 1, 2, \ldots, n \tag{4.5}$$

For our sample problem, we can take any two points, t_1 and t_2 within $[0, 1]$. For example, let us choose, $t_1 = \frac{1}{3}$ and $t_2 = \frac{2}{3}$ such that we have broken up our domain into three equal sections (note: we didn't have to choose equal sections).

At $t = t_1 = \frac{1}{3}$ we obtain the residual function,

$$R(\frac{1}{3}, \mathbf{a}) = 1 + \frac{4}{3}a_1 + \frac{7}{9}a_2 \tag{4.6}$$

and similarily for $t = t_2 = \frac{2}{3}$, we obtain,

$$R(\frac{2}{3}, \mathbf{a}) = 1 + \frac{5}{3}a_1 + \frac{16}{9}a_2 \tag{4.7}$$

We now impose the condition that both $R(\frac{1}{3}, \mathbf{a})$ and $R(\frac{2}{3}, \mathbf{a})$ both vanish. This yields a pair of simulataneous equations which we can solve for a_1 and a_2,

$$12a_1 + 7a_2 = -915a_1 + 16a_2 = -9 \tag{4.8}$$

which gives us,

$$a_1 = -\frac{27}{29} \qquad a_2 = \frac{9}{29} \tag{4.9}$$

such that the n=2 approximate solution for our differntial equation obtained by the collocation method on the interval $[0, 1]$ is

$$u_2(t) = 1 - \frac{27}{29}t + \frac{9}{29}t^2 \tag{4.10}$$

Let's now consider another method, in which instead of insisting that the residual vanish at the two points, we set the average value of the residual in the two parts to vanish. That is, we demand that,

$$\frac{1}{\Delta t_i}\int_{\Delta t_i} R(t; \mathbf{a})dt = 0 \tag{4.11}$$

where the integration is carried out within the subdomain, Δt_i. This is know as the *subdomain* method. For out sample problem, let's take the two subdomains, $\Delta t_1 = [0, \frac{1}{2}]$ and $\Delta t_2 = [\frac{1}{2}, 1]$. The average residuals are,

$$\frac{1}{\Delta t_1}\int_0^{\frac{1}{2}} R(t; \mathbf{a})dt = 2[\frac{1}{2} + \frac{5}{8}a_1 + \frac{7}{24}a_2] \tag{4.12}$$

$$\frac{1}{\Delta t_2}\int_{\frac{1}{2}}^{1} R(t;\mathbf{a})dt = 2[\frac{1}{2} + \frac{7}{8}a_1 + \frac{25}{24}a_2] \tag{4.13}$$

which gives us two equations to solve for a_1 and a_2, yielding,

$$a_1 = \frac{-18}{19} \qquad a_2 = \frac{6}{19} \tag{4.14}$$

such that our subdomain solution for $n = 2$ is,

$$u_2(t) = 1 - \frac{18}{19}t + \frac{6}{19}t^2 \tag{4.15}$$

Intead of requiring the average value of the residual within each subdomain to vanish, we can also use least-squares techniques to find the optimum values of the parameters ($\mathbf{a}$) that produce the smallest average residual $R^2(t;\mathbf{a})$ in the entire domain. This yields the formulation,

$$\frac{\partial}{\partial a_i}\int_{t_b}^{t_d} R^2(t;\mathbf{a})dt = 2\int_{t_b}^{t_d} R(t;\mathbf{a})\frac{\partial R}{\partial a_i}dt \tag{4.16}$$

As you might have imagined, we are going to leave the calculations as an excerise.

While the three previous methods gave somewhat different values for the coefficients, a_1 and a_2, the way in which they determinded these coefficients are very similar. Actually, they all belong to a more general method called *weighted residual* methods. In terms of the integral,

$$\int_{t_b}^{t_d} R(t;\mathbf{a})W_i(t)dt = 0 \tag{4.17}$$

the only difference between the methods is the value they use for W_i. For the collocation method, $W_i(t) = \delta(t_i)$. For the subdomain method, $W_i(t) = 1$ if t is within Δt_i and $W_i = 0$ otherwise and for the least squares method, $W_i = \frac{\partial R(t;\mathbf{a})}{\partial a_i}$.

As you can imagine, there are many other possible weighting functions we could try. The example of weighted residuals which is used most often in finite elements methods today uses elements of the trial solution themselves as $W_i(t)$. This is called the *Galerkin* method.

To see how this works, we reformulate the trial solution, $\tilde{u}(t)$ in terms of $(n+1)$ linearly independent functions which constitute as set of *basis* functions with which to express our solutions.

$$\tilde{u}_n(t) = \psi_0 + \sum_{i=1}^{n} a_i\psi_i(t) \tag{4.18}$$

the first term is often left outside the sum because it is often associated with part or all of the intial or boundary conditions. This is equivalent to putting into ψ_0 everything that can be fixed by the initial and boundary conditions. For our $n = 2$ trial function in our example, one possible choice of functions is,

$$\psi_0 = 1 \quad \psi_1 = t \quad \psi_2 = t^2 \tag{4.19}$$

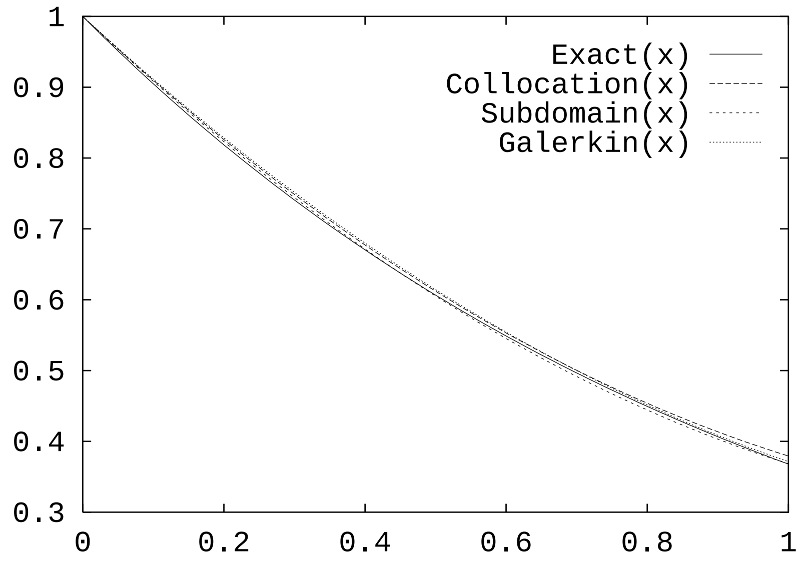

Figure 4.1: Comparison between different finite element methods.

If we substitute these basis functions into our expression as the weighting term, we obtain,

$$\int_{t_b}^{t_d} R(t;\mathbf{a})\psi_i(t)dt = 0 \quad i = 1, 2, \ldots, n \tag{4.20}$$

For our sample problem, the two Galerkin equations are,

$$\int_0^1 [1 + (1+t)a_1 + (2t + t^2)a_2]tdt = \frac{1}{2} + \frac{5}{6}a_1 + \frac{11}{12}a_2 = 0 \tag{4.21}$$

$$\int_0^1 [1 + (1+t)a_1 + (2t + t^2)a_2]t^2dt = \frac{1}{3} + \frac{7}{12}a_1 + \frac{7}{10}a_2 = 0 \tag{4.22}$$

which yields the approximate solution to our differential equation,

$$u_2(t) = 1 - \frac{32}{35}t + \frac{2}{7}t^2 \tag{4.23}$$

which is very close to the analytical solution within our domain of interest. Comparisons of the previous finite element methods are shown in figure 4.1. The approximations over the interval $[0, 1]$ are all very close. Look what happens, however, when the interval is changed to $[0, 2]$ without re-evaluating our integrals (see figure 4.2). As you can see, the solutions are only good over the very specific domain we specified $[0, 1]$.

4.0.1 Variational and Minimization Approaches

Historically, the finite element method was originally developed by the mathematician Courant in the early 1900's. No work was done on the method until the 1950's and 1960's when

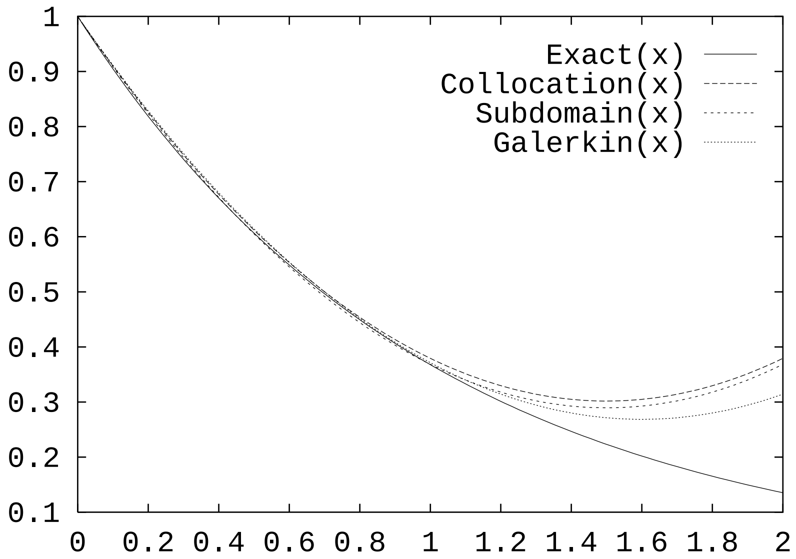

Figure 4.2: Comparison between different finite element methods over the interval $[0, 2]$.

mechanical engineers figured it out independently. What we have seen thus far are the mathematically based versions of the finite element method. Let's now look at the engineering approaches. The two classical approaches are based upon a variational approach and a minimization formulation. Consider the $1D$ Poisson equation

$$\frac{d^2u(x)}{dx^2} = -f(x) \qquad 0 < x < 1 \tag{4.24}$$

with boundary conditions,

$$u(0) = u(1) = 0 \tag{4.25}$$

Define **V** to denote the *variational* formulation and **M** to denote the *minimization* formulation.

Before defining the two approaches, we need to define some mathematical quantities. Let,

$$(v, w) = \int_0^1 v(x)w(x)dx \tag{4.26}$$

for real-valued piecewise continuous functions. Also, let's define the linear space,

$$\begin{aligned} \Omega = [u : u \text{ is a continuous function on } [0,1], \\ \frac{du}{dx} \text{ is piecewise continuous and bounded on } [0,1] \text{ and } u(0) = u(1) = 0]. \end{aligned} \tag{4.27}$$

We also need the functional:

$$F(u) = \frac{1}{2}(u', u') - (f, u) \tag{4.28}$$

which is the short hand notation for

$$F(u) = \frac{1}{2}\int_0^1 \frac{du}{dx}\frac{du}{dx}dx - \int_0^1 f(x)u(x)dx \tag{4.29}$$

where we have defined $u' = \frac{du}{dx}$.

The two forms of the finite element approximation are now defined as:

$$(\mathbf{V}) \text{ Find } u \in \Omega \text{ s.t. } (u', v') = (f, v) \quad \forall v \in \Omega \tag{4.30}$$

$$(\mathbf{M}) \text{ Find } u \in \Omega \text{ s.t. } F(u) \leq= F(v) \quad \forall v \in \Omega \tag{4.31}$$

Now let's see if we can explain them. The **V** formulation corresponds to the principle of virtual work that is often used in mechanical engineering and physics. Often times, the total energy of the system is expressed as a function of the dependent variable and its first derivative. A first *vartiation* is then performed on the energy function which yields the equation(s) of motion.

For the **M** formuation, $F(v)$ would correspond to the total potential energy of the system, therefore, **M** corresponds to the *principle of minimum potential energy* which can be stated as, "The displacement field that satifies the geometric boundary conditions and corresonds to the state of equilibrium is the one that minimizes the total potential energy." If that doesn't clarify things for you, consider the following examples of a harmonic oscillator (the motion of a mass attached to the end of a massless spring). First let's consider the variational form and define the total energy of the system:

$$E = T + V \tag{4.32}$$

where E is the total energy in the system, the first term represents the kinetic energy of the spring and the second term represents the potential energy. T and V have the familiar forms,

$$T = \frac{m}{2}(\frac{dx}{dt})^2 \qquad V = \frac{k}{2}x^2 \tag{4.33}$$

where m is the mass of the particle and $x(t)$ represents the displacement of the particle away from equilibrium. Such that the total energy is,

$$E = \frac{m}{2}(\dot{u})^2 + \frac{k}{2}x^2 \tag{4.34}$$

Since the energy is a conserved quantity, any variation of it vanishes,

$$\delta E(x, x') = \frac{\partial E}{\partial x}\frac{dx}{dt} + \frac{\partial E}{\partial \dot{x}}\frac{d\dot{x}}{dt} = 0 \tag{4.35}$$

this gives us the familiar equation of motion for a harmonic oscillator,

$$m\frac{\partial^2 x}{\partial t^2} + kx = 0 \tag{4.36}$$

or, if this were a multidimensional problem, this could be written in matrix form as, $\mathbf{F} = -k\mathbf{x}$.

Now let's consider the minimization formulation. We define a energy *functional*,

$$\Pi = \frac{1}{2}kx^2 - Fx \tag{4.37}$$

where the first term is the internal strain energy of the spring and the second term represents the work done by external forces. We now minimize the energy functional with respect to the displacement,

$$\frac{\partial \Pi}{\partial x} = kx - Fx = 0 \tag{4.38}$$

which gives us the familar equation, $\mathbf{F} = k\mathbf{x}$.

Briefly, let's show that the variational formulation is actually equivalent to the original differnetial equation. We first multiply $u'' = -f$ by an arbitrary function, $v \in \Omega$, a so-called *test function*, and then integrate of the interval $[0, 1]$

$$(u'', v) = -(f, v) \tag{4.39}$$

Integrating the left hand side by parts and using that fact that $v(0) = v(1) = 0$ we obtain,

$$-(u'', v) = -u'(1)v(1) + u'(0)v(0) + (u', v') = (u', v') \tag{4.40}$$

which yields,

$$(u', v') = (f, v) \quad \forall v \in \Omega \tag{4.41}$$

which shows that u is a solution of the variational formulation.

4.0.2 Example: Galerkin Formulation of Poisson's Equation

Now that we've given a brief overview of all the different varieties of finite element methods, let's start solving some problems and figuring out how to choose the right elements and set of basis functions.

Let's start by considering the simplest element and set of basis functions, those that are piecewise linear functions.

In our problem, the first thing that we will do is to discretize our solution domain (Ω) into a finite number of elements. Using piecewise linear functions, this means our elements will be line segments for 1D problems, triangles or quadrilaterals for 2D problems, and tetrahedra, prisms, or hexahedra for 3D problems. Let's go through the steps using 1D linear elements and then build up to higher dimensions.

Again, consider Poisson's equation.

$$\frac{d^2u(x)}{dx^2} = -f(x) \quad 0 < x < 1 \tag{4.42}$$

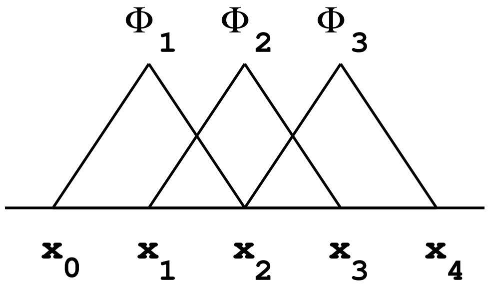

Figure 4.3: 1D Linear Basis Functions

with boundary conditions,

$$u(0) = u(1) = 0 \tag{4.43}$$

We first construct our finite dimensional subspace, Ω_h consisting of piecewise linear functions. Let

$$0 = x_0 < x_1 < x_2 < \ldots < x_N < x_{N+1} = 1 \tag{4.44}$$

be a partition on $[0, 1]$ into subintervals, $I_j = x_j, x_{j+1}$ of length, $\Delta x = h_j = x_{j+1} - x_j$, where $j = 0, 2, \ldots, N$. We will define $h = \max h_j$ as a measure of the overall fineness of the grid.

As parameters to describe how our function changes over the subintervals, we'll choose the *basis functions* as the set of triangular functions defined as

$$\Phi_j(x) = \begin{cases} \frac{1}{\Delta x}(x - x_{j-1}) & \text{for } x_{j-1} \leq x \leq x_j \\ \frac{1}{\Delta x}(x_{j+1} - x) & \text{for } x_j \leq x \leq x_{j+1} \\ 0 & \text{otherwise} \end{cases} \tag{4.45}$$

for $j = 1, 2, \ldots, N$ and $\Delta x = x_{j+1} - x_j$ and if constant, is equal to h. Note that our basis functions have the following property:

$$\Phi_j(x) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{4.46}$$

If the segments are equal, the interval $[0, 1]$ is deivded into $(N + 1)$ equal segments of length Δx. Each basis function would then be an isosceles triangle, displaced from the previous one by a distance equal to one-half of the length of the base line as shown in figure 4.3. What this means is the solution to u between any two points x_j and x_{j+1} is approximated by a piecewise linear function

$$u = u_j\left(1 - \frac{x - x_j}{x_{j+1} - x_j}\right) + u_{j+1}\frac{x - x_j}{x_{j+1} - x_j} \tag{4.47}$$

where u_j has the physical meaning of being the value of u at $x = x_j$. These u_j are the generalized coordinates of the problem - they are determined from the Galerkin formulation (in which they are represented as α_j). Thus, the finite element approximation to a typical one-dimensional problem using piecewise linear elements would look like that in figure 4.4

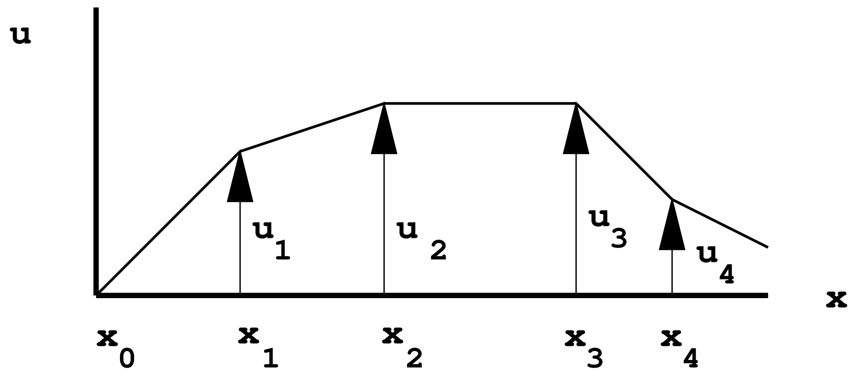

Figure 4.4: 1D finite element solution for a second-order differential equation.

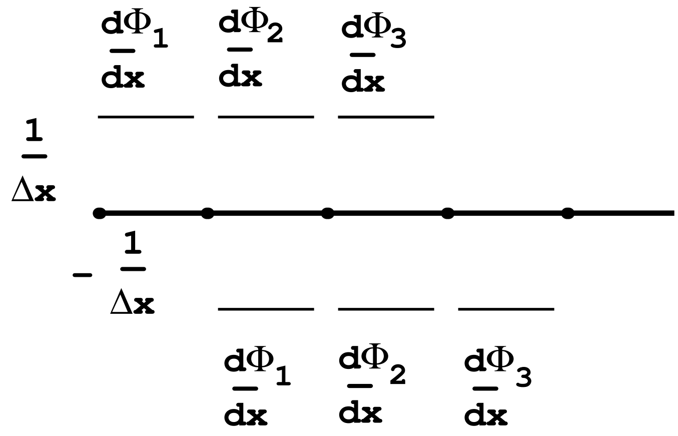

Figure 4.5: First Derivatives of the 1D Linear Basis Functions

The first-order derivatives of the basis functions are,

$$\frac{d\Phi_j(x)}{dx} = \begin{cases} \frac{1}{\Delta x} & \text{for } x_{j-1} \leq x \leq x_j \\ -\frac{1}{\Delta x} & \text{for } x_j \leq x \leq x_{j+1} \\ 0 & \text{otherwise} \end{cases} \tag{4.48}$$

which have a discontinuity at $x = x_j$, are shown in figure 4.5. If we let $x_0 = 0$ and $x_{n+1} = 1$, that is, the two endpoints of our solution domain, all the basis functions vianish on both boundaries. This means that every Φ_i satisfies the boundary conditions for the problems. Our trial solution with n degrees of freedom takes on the form:

$$u(x) \approx \tilde{u}_n(x) = \sum_{i=1}^{n} \alpha_i \Phi_i(x) \tag{4.49}$$

Now we substitute this back into our residual equation for the Galerkin method.

$$R(x;\alpha) = \sum_{i=1}^{n} \alpha_i \frac{d^2\Phi_i}{dx^2} + f(x) \tag{4.50}$$

The Galerkin equation becomes

$$\sum_{i=1}^{n} \alpha_i \int_{x_0}^{x_{n+1}} \frac{d^2\Phi_i}{dx^2}\Phi_j(x)dx + \int_{x_0}^{x_{n+1}} f(x)\Phi_j(x)dx = 0 \tag{4.51}$$

We can use integration by parts to simplify the first term on the left hand side of the equation

$$\int_{x_0}^{x_{n+1}} \frac{d^2\Phi_i}{dx^2}\Phi_j(x)dx = \frac{d\Phi_i}{dx}\Phi_j(x)\Big|_{x=x_0}^{x=x_{n+1}} - \int_{x_0}^{x_{n+1}} \frac{d\Phi_i}{dx}\frac{d\Phi_j}{dx}dx = -\int_{x_0}^{x_{n+1}} \frac{d\Phi_i}{dx}\frac{d\Phi_j}{dx}dx \tag{4.52}$$

where we have invoked the fact that all the basis functions $\Phi_j(x)$ vanish at the boundary.

Let's now compute the elements of the matrices. First we'll calcuate the elements of our global matrix (note that we really only have to compute the cases for which $|i-j| \leq 1$ because of using linear elements in 1D). For $j = 1, \ldots, N$ we have,

$$(\Phi_j', \Phi_j') = \int_{x_{j-1}}^{x_j} \frac{1}{(\Delta x)^2}dx + \int_{x_j}^{x_{j+1}} \frac{1}{(\Delta x)_{j+1}^2}dx = \frac{1}{(\Delta x)_j} + \frac{1}{(\Delta x)_{j+1}} \tag{4.53}$$

and for $j = 2, \ldots, N$ we have,

$$(\Phi_j', \Phi_{j-1}') = (\Phi_{j-1}', \Phi_j') = -\int_{x_{j-1}}^{x_j} \frac{1}{(\Delta x)_j^2}dx = -\frac{1}{(\Delta x)_j} \tag{4.54}$$

such that the components of our matrix $(\Phi_i', \Phi_j') = a_{ij}$ are,

$$a_{ij} = \begin{cases} \frac{2}{\Delta x} & \text{for } i = j \\ -\frac{1}{\Delta x} & \text{for } |i-j| = 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.55}$$

and the source term components are evaluated as:

$$b_j = \int_{x_j}^{x_{j+1}} f(x)\Phi_j(x)dx. \tag{4.56}$$

This then constitutes a linear system of equations with N equations and N unknowns α. In matrix form, we can reformulate this as,

$$\mathbf{A}\alpha = \mathbf{b} \tag{4.57}$$

$$\frac{1}{\Delta x}\begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & & -1 \\ & & -1 & 2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}. \tag{4.58}$$

where $A = (a_{ij})$ is the $N \times N$ matrix with elements $a_{ij} = (\Phi_i', \Phi_j')$ and is often called the *stiffness matrix.* The approximate solution values at the nodes are given by the vector $\alpha = (\alpha_1, \ldots, \alpha_N)$ and the vector $\mathbf{b} = (f, \Phi_i)$ is often called the *load* vector and contains contributions from sources and the boundary conditions.