Jim Lambers
ENERGY 281
Spring Quarter 2007-08
Lecture 11 Notes

These notes were originally written by Tara LaForce.

1 Derivation of the Finite Element Method

In order to derive the fundamental concepts of finite element methods (FEM) we will start by looking at an extremely simple ODE and approximate it using FEM.

1.1 The Model Problem

The model problem is:

$$\begin{array}{cc} -u'' + u = x & 0 < x < 1 \\ u(0) = 0 & u(1) = 0 \end{array} \tag{1}$$

and this problem can be solved analytically: $u(x) = x - \sinh x / \sinh 1$. The purpose of starting with this problem is to demonstrate the fundamental concepts and pitfalls in FEM in a situation where we know the correct answer, so that we will know where our approximation is good and where it is poor. In cases of practical interest we will look at ODEs and PDEs that are too complex to be solved analytically.

FEM doesn't actually approximate the original equation, but rather the *weak form* of the original equation. The purpose of the weak form is to satisfy the equation in the "average sense," so that we can approximate solutions that are discontinuous or otherwise poorly behaved. If a function $u(x)$ is a solution to the original form of the ODE, then it also satisfies the weak form of the ODE. The weak form of Eq. 1 is

$$\int_0^1 \left(-u'' + u\right) v dx = \int_0^1 x v dx \tag{2}$$

The function $v(x)$ is called the weight function or *test function*. $v(x)$ can be any function of x that is sufficiently well behaved for the integrals to exist.

The set of all functions v that also have $v(0) = 0$, $v(1) = 0$ are denoted by H. (We will put many more constraints on v shortly.)

The new problem is to find u so that

$$\begin{array}{cc} \int_0^1 (-u'' + u - x)\, v dx = 0 & \text{for all } v \in H \\ u(0) = 0 & u(1) = 0 \end{array} \tag{3}$$

Once the problem is written in this way we can say that the solution u belongs to the class of *trial functions* which are denoted $\tilde{H}$. When the problem is written in this way the classes of test functions H and trial functions $\tilde{H}$ are not the same. For example, u must be twice differentiable and have the property that $\int_0^1 u'' v dx < \infty$, while v doesn't even have to be continuous as long as the integral in Eq. 3 exists and is finite. It is possible to approximate u in this way, but having to work with two different classes of functions unnecessarily complicates the problem. In order to make sure that H and $\tilde{H}$ are the same we can observe that if v is sufficiently smooth then

$$\int_0^1 -u'' v dx = \int_0^1 u'v' dx - u'v|_0^1 \tag{4}$$

This formulation must be valid since u must be twice differentiable and v was arbitrary. This puts another constraint on v that it must be differentiable and that those derivatives must be well-enough behaved to ensure that the integral $\int_0^1 u'v' dx$ exists. Moreover, since we decided from the outset that $v(0) = 0$ and $v(1) = 0$, the second term in Eq. 4 is zero regardless of the behavior of u' at these points. The new problem is

$$\int_0^1 \left(u'v' + uv - xv\right) dx = 0. \tag{5}$$

Notice that by performing the integration by parts we restricted the class of test functions H by introducing v' into the equation. We have simultaneously expanded the class of trial functions $\tilde{H}$, since u is no longer required to have a second derivative in Eq. 5. The weak formulation defined in Eq. 5 is called a variational boundary-value problem.

In Eq. 5 u and v have exactly the same constraints on them:

1. u and v must be square integrable, that is: $\int_0^1 uv dx \approx \int_0^1 u^2 dx < \infty$

2. The first derivatives of u and v must be square integrable, that is: $\int_0^1 u'v'dx \approx \int_0^1 (u')^2 dx < \infty$ (this actually guarantees the first property)

3. We had already assumed that $v(0) = 0$ and $v(1) = 0$ and we know from the original statement of the problem that $u(0) = 0$ and $u(1) = 0$.

Now we have that $\tilde{H} = H = H_0^1$. Any function w is a member of H_0^1 if $\int_0^1 (u')^2 dx < \infty$ and $w(0) = w(1) = 0$. H_0^1 is the space of admissible functions for the variational boundary-value problem (ie. all admissible test *and* trial functions are in H_0^1)

We will consider the variational form Eq. 5 to be the equation that we would like to approximate, rather than the original statement in Eq. 1. Once we have found a solution to Eq. 5 in this way we can ask the question whether this formulation is also a solution to Eq. 1: That is, whether this solution is a function satisfying Eq. 1 at every x in $0 < x < 1$, or whether we have found a solution that satisfies only the weak form of the equation. In the case that we can only find a solution to the weak form, no "classical" solution exists.

1.2 Galerkin Approximations

We now have the problem re-stated so that we are looking for $u \in H_0^1$ such that

$$\int_0^1 \left(u'v' + uv\right) dx = \int_0^1 xv dx \tag{6}$$

for all $v \in H_0^1$. In order to narrow down the number of functions we will consider in our approximate solutions we will make two more assumptions about H_0^1. First, we will assume that H_0^1 is a linear space of functions (that is if $v_1, v_2 \in H_0^1$ and a, b are constants then $av_1 + bv_2 \in H_0^1$.)

The second assumption is that H_0^1 is infinite dimensional. For example if we have the sine series $\psi_n(x) = \sqrt{2}\sin(n\pi x)$ for $n = 1, 2, 3, ...$ and $v \in H_0^1$ then v can be represented by $v(x) = \sum_{n=1}^{\infty} a_n \psi_n(x)$. The scalar coefficients a_n are given by $a_n = \int_0^1 v(x)\,\psi_n(x)\,dx$, just like usual. Hence infinititely many coefficients a_n must be found to define v exactly. As in Fourier analysis, many of these coefficients will be zero. We will also truncate the series in order to have managable length series, just like in discrete Fourier analysis.

Unlike in Fourier analysis, though the basis functions do not have to be sines and cosines, much less smooth functions can be used. In fact our set of basis functions do not even have to be smooth and can contain discontinuities in the derivatives, but they must be continuous. We will assume that the infinite series converges so that we can consider only the first N basis functions and get a good approximation v_N of the original test (or trial) function:

$$v \cong v_N = \sum_{i=1}^{N} \beta_i \phi_i(x) \tag{7}$$

where ϕ_i are as-yet unspecified basis functions. This subspace of functions is denoted $H_0^{(N)}$ and is a *subspace* of H_0^1. Galerkin's method consists of finding an approximate solution to Eq. 6 in a finite-dimensional subspace $H_0^{(N)}$ of $H_0^{(1}$ of admissible functions rather than in the whole space H_0^1. Now we are looking for $u_N = \sum_{i=1}^{N} \alpha_i \phi_i(x)$. The new approximate problem we have is to find $u_N \in H_0^{(N)}$ such that

$$\int_0^1 \left(u_N' v_N' + u_N v_N\right) dx = \int_0^1 x v_N dx \tag{8}$$

for all $v_N \in H_0^{(N)}$. Since the ϕ_i are known (in principle) u_N will be completely determined once the coefficients α_i have been found.
In order to find that α_n we put $\sum_{i=1}^{N} \alpha_i \phi_i(x)$ and $\sum_{i=1}^{N} \beta_i \phi_i(x)$ into Eq. 8.

$$\int_0^1 \left\{ \begin{array}{l} \frac{d}{dx}\left[\sum_{i=1}^{N} \beta_i \phi_i(x)\right] \frac{d}{dx}\left[\sum_{j=1}^{N} \alpha_j \phi_i(x)\right] + \\ \left[\sum_{i=1}^{N} \beta_i \phi_i(x)\right]\left[\sum_{j=1}^{N} \alpha_j \phi_i(x)\right] - \\ x \sum_{i=1}^{N} \beta_i \phi_i(x) \end{array} \right\} dx = 0 \tag{9}$$

for all N independent sets of β_i.
This can be expanded and factored to give

$$\sum_{i=1}^{N} \beta_i \left(\sum_{j=1}^{N} \left\{ \int_0^1 \left[\phi_j'(x)\phi_i'(x) + \phi_j(x)\phi_i(x)\right] dx \right\} \alpha_j - \int_0^1 x \phi_i(x)\, dx \right) = 0 \tag{10}$$

for all N independent sets of β_i. The structure of Eq. 10 is easier to see if it is re-written as

$$\sum_{i=1}^{N} \beta_i \left(\sum_{j=1}^{N} K_{ij}\alpha_j - F_i \right) = 0 \tag{11}$$

for all β_i. Where

$$K_{ij} = \int_0^1 \left[\phi_j'(x)\,\phi_i'(x) + \phi_j(x)\,\phi_i(x) \right] dx \quad F = \int_0^1 x\phi_i(x)\,dx \tag{12}$$

and where $i,\ j = 1, ..., N$. The $N \times N$ matrix of K_{ij} is called the stiffness matrix and the vector F is the load vector. Since the β_i are known K_{ij} and F can be calculated directly. But the β_i were arbitrary so we can choose each element β_i for each equation. For the first equation choose $\beta_1 = 1$ and $\beta_n = 0$ for $n \neq 1$. Now $\sum_{j=1}^{N} K_{1j}\alpha_j = F_1$. Similarly for the second equation choose $\beta_2 = 1$ and $\beta_n = 0$ for $n \neq 2$ so that $\sum_{j=1}^{N} K_{2j}\alpha_j = F_2$. In this way we have chosen N independent equations that can be used to find the N unknowns α_i. Moreover the N coefficients α_i can be found from $\alpha_j = \sum_{j=1}^{N} \left(K^{-1}\right)_{ji} F_i$ where $\left(K^{-1}\right)_{ji}$ are the elements of the inverse of K. The stiffness matrix K is symmetric for this simple problem, which makes the computation of the matrix faster since we don't have to compute all of the elements, symmetric matricies are also much faster to invert.

1.3 Finite Elements Basis Functions

Now we have done a great deal of work, but it may not seem like we are much closer to finding a solution to the original ODE since we still know nothing about ϕ_i. The purpose of using such a general formulation is that any set of linearly independent functions will work to solve the ODE. Now we are finally going to talk about what kind of functions we will want to use as basis functions. The finite element method is a general and systematic technique for constructing basis functions for Galerkin approximations. In FEM the basis functions ϕ_i are defined piecewise over subregions. Over any subdomain the ϕ_i will be chosen to be polynomials of low degree, though other possibilities do exist.

- *finite elements* are the subregions of the domain over which each basis function is defined. Hence each basis function has compact support over an element. Each element has length h. The lengths of the elements do NOT need to be the same (but generally we will assume that they are.)

- *nodes* or *nodal points* are defined within each element. In Figure 1 the five nodes are the endpoints of each element (numbered 0 to 4).
- the *finite element mesh* is the collection of elements and nodal points that make up the domain and is shown in Figure 1. An element i is denoted by Ω_i.

Now we need to construct the actual basis functions using the three criteria defined before: 1) The basis functions are simple functions defined piecewise over the finite element mesh, 2) the basis functions must be in the class of test functions H_0^1, and 3) The basis functions are chosen so that the parameters α_i are the values of $u_N(x)$ at the nodal points.

The simplest set of basis functions are the "hat functions" on elements $i = 1, 2, 3$.

$$\phi_i(x) = \left\{ \begin{array}{ccc} \frac{x-x_{i-1}}{h_i} & \text{for} & x_{i-1} \leq x \leq x_i \\ \frac{x_{i+1}-x}{h_{i+1}} & \text{for} & x_i \leq x \leq x_{i+1} \\ 0 & \text{for} & x < x_{i-1}, x > x_{i+1} \end{array} \right\} \tag{13}$$

where $h_i = x_i - x_{i-1}$ is the length of element i. The derivatives are

$$\phi_i'(x) = \left\{ \begin{array}{ccc} \frac{1}{h_i} & \text{for} & x_{i-1} \leq x \leq x_i \\ \frac{-1}{h_{i+1}} & \text{for} & x_i \leq x \leq x_{i+1} \\ 0 & \text{for} & x < x_{i-1}, x > x_{i+1} \end{array} \right\} \tag{14}$$

The equations for elements 0 and 4 have been left out since we decided that $u(0) = u(1) = 1$, so no basis functions are required. In general the basis functions for the first and last elements are half of the functions since there is no $i-1$ or $i+1$ node, respectively. The hat functions are shown in Figure 2. The mathematical term for hat functions is *piece-wise linear basis functions*

Looking at the three criteria above, clearly the functions in Eq. 13 are simple and defined element-wise. It is easy to show that they are in H_0^1, since they have square-integrable first derivatives. They also satisfy the third criteria since $\phi_i(x_j) = 1$ if $i = j$ and 0 otherwise. Hence each function contributes to the value of u_N at exactly one node and $\alpha_i = u_N(x_i)$.

It is less clear that the hat functions will give a continuous representation of v_N and u_N. Let v be the sine function with period 2 shown in Figure 3. At the nodes (0, 1, 2, 3, 4) sine has the values (0,0.7071,1,0.7071,0). The representation v_N on the finite element mesh is $v_N = 0.7071\phi_1(x) +$

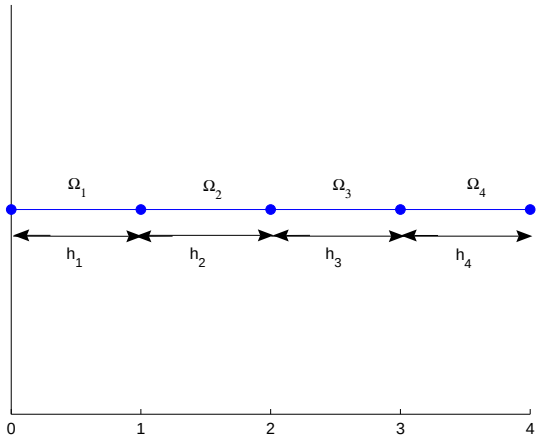

Figure 1: Four finite elements on the interval [0 1].

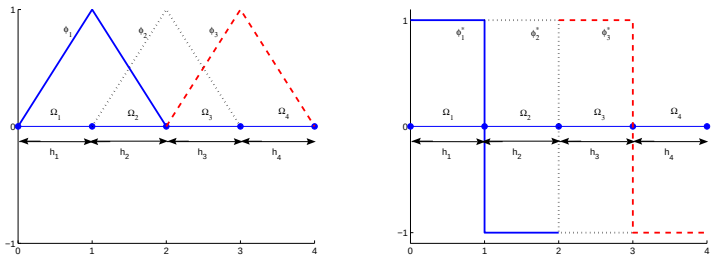

Figure 2: Four hat functions (top) and their derivatives (bottom) on the interval [0 1].

$\phi_2(x) + 0.7071\phi_3(x)$. When the elements are summed up the sine wave is approximated by piecewise linear functions between each of the nodes, and is exactly represented at each node. When more nodes are used the approximation improves and in the limit of $N \to \infty$ the sine wave would be exactly represented. In FEM we will never proceed all the way to the limit, so the interval size h will always have finite size h. This is why the term *finite elements* is used.

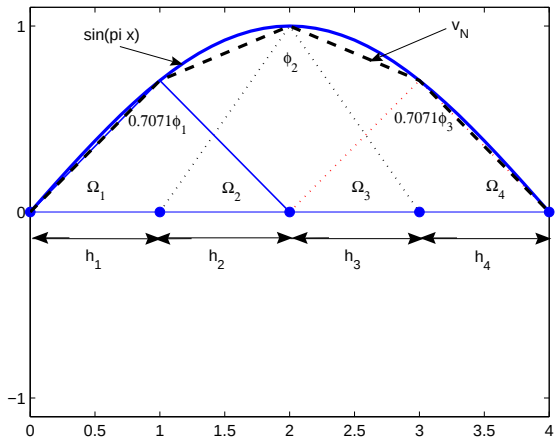

Figure 3: The finite element approximation of $sin(\pi x)$ using five nodes on the interval [0 1].

1.4 The Stiffness Matrix K and the Load Vector F for Hat Functions

Recall from Eq. 12 that each element of the stiffness matrix K is given by

$$\begin{aligned} K_{ij} &= \int_0^1 \left(\phi_i'(x)\,\phi_j'(x) + \phi_i(x)\,\phi_j(x)\right)dx \\ &= \sum_{e=1}^4 \int_{\Omega_e} \left(\phi_i'(x)\,\phi_j'(x) + \phi_i(x)\,\phi_j(x)\right)dx \\ &= \sum_{e=1}^4 K_{ij}^e \end{aligned} \tag{15}$$

similarly

$$F_i = \int_0^1 x\phi_i(x)dx = \sum_{e=1}^4 \int_{\Omega_e} x\phi_i(x)dx = \sum_{e=1}^4 F_i^e \tag{16}$$

where we have used the property that $\phi(x)$ are defined piecewise on each element 1 through 4. In order to compute an approximation of the solution to the model ODE it is necessary to compute nine elements for K_{ij} from $i, j = 1, 2, 3$ and three elements for F. But since each of the functions $\phi(x)$ are defined in the same way it is possible to compute K^e and F^e for a generic element and then to construct the matrix using the sums above. Consider a generic interior element Ω_e on the interval x_A to x_B. We will use a change of variables and rewrite this in terms of ξ, a dummy variable for x. We will have $\xi = (0, h)$. On this element exactly two of the hat functions are nonzero: $\psi_A(\xi) = 1 - \frac{\xi}{h}$ and $\psi_B(\xi) = \frac{\xi}{h}$. Convince yourself that this definition is equivalent to the previous definition of the hat function, but with the origin shifted to the start of one of the interior elements. The two hat functions have derivatives $\psi'_A(\xi) = -\frac{1}{h}$ and $\psi'_B(\xi) = \frac{1}{h}$.

It is also important to notice that for the hat functions $\phi_i(x) \neq 0$ on only the elements Ω_i and Ω_{i+1}. This results in a tridiagonal sparse matrix K for any number of elements in the mesh as will be shown below. Using Eq. 15 you can see that there are three integrals that contribute to K_{ij}:

$$\begin{array}{l} k_{AA} = \int_0^h \left(\left[\psi_A^e{}' (\xi) \right]^2 + \left[\psi_A^e (\xi) \right]^2 \right) d\xi \\ = \int_0^h \left(\left[1/h \right]^2 + \left[1 - \xi/h \right]^2 \right) d\xi = 1/h + h/3 \\ k_{AB} = \int_0^h \left(\psi_A^e{}' (\xi) \, \psi_B^e{}' (\xi) + \psi_A^e (\xi) \, \psi_B^e (\xi) \right) d\xi \\ = \int_0^h \left((-1/h) (1/h) + (1 - \xi/h) (\xi/h) \right) d\xi = -1/h + h/6 \\ k_{BB} = \int_0^h \left(\left[\psi_B^e{}' (\xi) \right]^2 + \left[\psi_B^e (\xi) \right]^2 \right) d\xi \\ = \int_0^h \left(\left[-1/h \right]^2 + \left[\xi/h \right]^2 \right) d\xi = 1/h + h/3 \end{array} \tag{17}$$

Similarly the components that contribute to the load vector are:

$$\begin{array}{l} F_A^e = \int_0^h (x_A + \xi) (1 - \xi/h) \, d\xi = \frac{h}{6} (2x_A + x_B) \\ F_B^e = \int_0^h (x_A + \xi) (\xi/h) \, d\xi = \frac{h}{6} (x_A + 2x_B) \end{array} \tag{18}$$

where the x_A and x_B terms come from evaluating the forcing function $f(x) = x$ at the endpoints of the generic element.

Thus each generic interior element contributes to the stiffness matrix a 2×2 submatrix

$$k^e = \begin{bmatrix} 1/h + h/3 & -1/h + h/6 \\ -1/h + h/6 & 1/h + h/3 \end{bmatrix} \tag{19}$$

and two entries to the load vector

$$f^e = h/6 \begin{bmatrix} 2x_A + x_B \\ x_A + 2x_B \end{bmatrix} \tag{20}$$

For the 4 element mesh we have derived the contributions to the overall stiffness matrix K from each node is given by:

$$K^1 = \begin{bmatrix} 1/h + h/3 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} K^2 = \begin{bmatrix} 1/h + h/3 & -1/h + h/6 & 0 \\ -1/h + h/6 & 1/h + h/3 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$
$$K^3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1/h + h/3 & -1/h + h/6 \\ 0 & -1/h + h/6 & 1/h + h/3 \end{bmatrix} K^4 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1/h + h/3 \end{bmatrix} \tag{21}$$

where the contributions from elements 1 and 4 have only one entry because only half of the hat function exists on these elements. Similarly the contributions to the load vector are

$$F^1 = h/6 \begin{bmatrix} 2h \\ 0 \\ 0 \end{bmatrix} F^2 = h/6 \begin{bmatrix} 2h + 2h \\ h + 4h \\ 0 \end{bmatrix} \tag{22}$$

$$F^3 = h/6 \begin{bmatrix} 0 \\ 4h + 3h \\ 2h + 6h \end{bmatrix} F^4 = h/6 \begin{bmatrix} 0 \\ 0 \\ 6h + 4h \end{bmatrix} \tag{23}$$

where $h = 0.25$ for the model problem. Now $K = K^1 + K^2 + K^3 + K^4$ and $F = F^1 + F^2 + F^3 + F^4$. The final system of equations has symmetric and diagonally dominant stiffness matrix K, which is very nice to work with mathematically. The values of u_N at each node is given by $\tilde{\alpha} = K^{-1}F$ and $u_N = \sum_{i=1}^{3} \alpha_i \phi_i(x)$.

Using this we get that the approximation to the model problem is $u = 0.0353\phi_1(x) + 0.0569\phi_2(x) + 0.0505\phi_3(x)$. This is not a very accurate answer, since only four elements were used. A more accurate approximation can be obtained by using more elements, but at the cost of building and inverting a larger stiffness matrix K. The usual way of estimating the error of an FEM approximation using linear basis functions (the hat functions we derived) using the L_2 or mean-square norm is that $||e||_0 < C_2 h^2$. This is an a-priori

error estimate and in general a worst-case scenario, the actual error may be substantially smaller.

References

[1] Becker, E. B., G. F. Carey, and J. T. Oden, *Finite Elements an Introduction*, Texas Institute for Computational Mechanics, UT Austin, 1981.