An Introduction to the Finite Element Method
SUPCOM 2008

Emmanuel A. Cabral
Department of Mathematics
Ateneo de Manila University
cabral@mathsci.math.admu.edu.ph

27 May 2008

1 Introduction and Overview

Consider the two-point boundary value problem $(\mathcal{D})$:

$$(\mathcal{D}) \begin{cases} \text{Find } u \text{ such that} \\ -u''(x) = f(x), \;\; 0 < x < 1; \\ u(0) = u(1) = 0; \\ f \;\text{ is continuous on }\; [0,1]. \end{cases}$$

Problem $(\mathcal{D})$ arises in application problems in physics and engineering. For our purposes, the case when f is continuous is simple enough to illustrate the finite element method. Theoretically, if f is continuous, then the above BVP can be solved by integrating f twice (Riemann integration is sufficient here) and solving for the constants of integration using the given boundary conditions. However, f may not have a closed form integral. In such a case, we may resort to numerical methods such as the finite element method. There are a lot of other numerical methods such as numerical integration, finite difference method etc. but as the title suggests, we will focus on the finite element method.

The procedure in solving the above problem numerically is as follows:

1. Obtain the *variational formulation* $(\mathcal{V})$ of problem $(\mathcal{D})$. That is,

$$(\mathcal{V}) \begin{cases} \text{Find } \;\; u \in V \text{such that} \\ \int_0^1 u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx \;\; \forall v \in V; \\ f \;\text{ is continuous on }\; [0,1]. \end{cases}$$

   where

$$V = \{v : v \text{ continuous on } [0,1], v' \text{ piecewise continuous, bounded on } [0,1], v(0) = v(1) = 0\}$$

2. Discretize the variational formulation. This means that for a chosen $M \in \mathbb{N}$, we subdivide $[0,1]$ into $M+1$ subintervals each of length $h = \frac{1}{M+1}$ and get the formulation $(\mathcal{V}_M)$:

$$(\mathcal{V}_M) \begin{cases} \text{Find } \ u_M \in V_M \text{ such that} \\ \int_0^1 u_M'(x)v'(x)dx = \int_0^1 f(x)v(x)dx \ \ \forall v \in V_M; \\ f \ \text{ is continuous on } \ [0,1]. \end{cases}$$

where V_M is the span of the set of hat functions $\{\phi_1, \phi_2, \ldots, \phi_M\}$.

3. From the discrete variational formulation obtained previously, obtain the matrix equation $A\xi = b$ where $A = [a_{ij}]$ with $a_{ij} = \int_0^1 \phi_i'(x)\phi_j'(x)dx$ and $b = [b_i]$ with $b_i = \int_0^1 f(x)\phi_i(x)dx$.

4. Solve the matrix equation $A\xi = b$. If $\xi = [\xi_1 \ \xi_2 \ \ldots \ \xi_M]^T$, then the approximate solution $u_M = \xi_1\phi_1 + \xi_2\phi_2 + \ldots + \xi_M\phi_M$.

As a specific example, we consider the problem $\mathcal{D}_1$:

$$(\mathcal{D}_1) \begin{cases} -u''(x) = x, \ \ 0 < x < 1; \\ u(0) = u(1) = 0. \end{cases}$$

This example is simple enough for us to get the analytic solution by integrating f twice. We will use this analytic solution to compare with the numerical solution. The analytic solution is $u(x) = -\frac{1}{6}x^3 + \frac{1}{6}x$. For the numerical solution, the *stiffness matrix* A is the tridiagonal $M \times M$ matrix

$$A = \frac{1}{h}\begin{bmatrix} 2 & -1 & 0 & 0 & 0 & \ldots & 0 \\ -1 & 2 & -1 & 0 & 0 & \ldots & 0 \\ 0 & -1 & 2 & -1 & 0 & \ldots & 0 \\ 0 & 0 & -1 & 2 & -1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{bmatrix}$$

and the *load vector* b has components

$$b_i = \frac{2i^3}{3(M+1)^2} - \frac{i^3}{(M+1)^2} + \frac{(i-1)^3}{6(M+1)^2} + \frac{(i+1)^3}{6(M+1)^2}.$$

Using the software *Scilab*, the computer program based on the above procedure is shown after the list of references. The corresponding graphs of the numerical and analytical solution together with the error are also shown in the pages after bibliography. Note how the numerical solution successfully approximates the analytical solution.

At this point, one may feel lost in the above discussion. That is understandable. We will go back to this procedure later. The succeeding discussions will help us. This lecture aims to illustrate the finite element method for simple systems such as the one given above and more importantly, it aims to show why the given procedure works. Hence, we will look into the mathematics involved to arrive at the above procedure. After this, the participants will be asked to solve a similar but slightly more difficult boundary value problem than $(\mathcal{D}_1)$. An appendix of needed definitions and results has been included for the reader's quick reference. The necessary background for the reader should include at least elementary calculus and linear algebra.

2 The Variational Formulation

This section aims to show the equivalence of problem $(\mathcal{D})$ and the corresponding *variational formulation* $(\mathcal{V})$. By *equivalence*, we mean that any solution of $(\mathcal{D})$ is also a solution of $(\mathcal{V})$ and vice-versa. We note here that the family V of functions defined in the previous section is a *vector space*.

Suppose u is a solution of $(\mathcal{D})$. Then

$$-u''(x) = f(x).$$

Now take any $v \in V$ and multiply it to both sides of the previous equation. We call v a *test function*. Integrate the resulting equation over the interval $[0, 1]$. So we get

$$\int_0^1 -u''(x)v(x)dx = \int_0^1 f(x)v(x)dx.$$

Now integrate the left-hand side by parts to get the equation

$$-u'(x)v(x)|_0^1 + \int_0^1 u'(x)v'(x) = \int_0^1 f(x)v(x)dx.$$

The given boundary conditions lead to

$$\int_0^1 u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx.$$

Since v is an arbitrary element of V, we conclude that any solution u of $(\mathcal{D})$ is also a solution of $(\mathcal{V})$.

At this point we introduce the convenient notation $(u, v) = \int_0^1 u(x)v(x)dx$. So the last equation above can be written as

$$(u', v') = (f, v) \ \ \forall v \in V.$$

This is a more convenient way of writing the variational formulation.
Now, let us prove the reverse. Suppose that u is a solution of $(\mathcal{V})$. Then

$$(u', v') = (f, v) \ \ \forall v \in V.$$

So

$$\int_0^1 u'(x)v'(x)dx - \int_0^1 f(x)v(x)dx = 0 \ \ \forall v \in V.$$

If we assume that u'' is continuous and bounded in the open interval $(0, 1)$ then we can do the integration in the first term above by parts. We get

$$u'(x)v(x)|_0^1 - \int_0^1 u''(x)v(x)dx - \int_0^1 f(x)v(x)dx = 0$$

Since $v \in V$, $v(0) = v(1) = 0$, we have

$$-\int_0^1 [u''(x) + f(x)]v(x)dx = 0 \ \ \forall v \in V.$$

If u'' and f are continuous in the open interval $(0,1)$ then $u''+f$ is also continuous in $(0,1)$. So $u''(x)+f(x)=0 \ \ \forall x \in (0,1)$. So $-u''(x)=f(x)$ for all x in the open interval $(0,1)$. We have shown that u is also a solution of $(\mathcal{D})$.

Thus the problems $(\mathcal{D})$ and $(\mathcal{V})$ are equivalent. This equivalence now allows us to work on the variational formulation instead of the original problem.

3 Uniqueness of Solution

Since f is given to be continuous, there is no doubt that a solution u exists. We will prove here that the solution is unique.

If u_1 and u_2 are two solutions of the variational formulation $(\mathcal{V})$, then for any $v \in V$,

$$\begin{aligned}(u_1', v') &= (f, v)\\ (u_2', v') &= (f, v)\end{aligned}$$

Subtracting these two equations, we get

$$\int_0^1 [u_1'(x) - u_2'(x)]v'(x)dx = 0.$$

Since this equation is true for any $v \in V$, it is true $v = u_1 - u_2$. Note that $u_1 - u_2 \in V$ since V is a vector space. So

$$\int_0^1 [u_1'(x) - u_2'(x)]^2 dx = 0.$$

It follows that $u_1'(x) - u_2'(x) = 0$. Now it is also true that $u_1''(x) = f(x)$ and $u_2''(x) = f(x)$ in the open interval $(0,1)$ where f is continuous on $[0,1]$. So we may safely assume that $u_1' - u_2'$ is continuous in the open interval $(0,1)$. Thus $u_1'(x) - u_2'(x) = 0$ for every x in the open interval $(0,1)$ and so $u_1 - u_2$ is constant in the open interval $(0,1)$. Since $u_1 \in V$ and $u_2 \in V$ then u_1, u_2 and $u_1 - u_2$ are all continuous on the closed interval $[0,1]$. But $u_1(0) - u_2(0) = 0$ and $u_1(1) - u_2(1) = 0$ so $u_1(x) - u_2(x) = 0$ for every x in the open interval $(0,1)$. This finally allows us to conclude that $u_1 = u_2$.

We have shown that there is one and only one solution to the boundary value problem.

4 The Hat Functions

Consider the closed interval $[0,1]$. For a chosen $M \in \mathbb{N}$, we subdivide $[0,1]$ into $M+1$ subintervals. For simplicity, we may choose the subintervals to be of the same length $h = \frac{1}{M+1}$. Including the endpoints 0 and 1, we consider the node points $x_0, x_1, x_2, \ldots, x_M, x_{M+1}$. Here, $x_j = j * h = j * \frac{1}{M+1}$. For $j = 1, \ldots, M$, we define the *hat function* ϕ_j to be linear in the intervals (x_{j-1}, x_j) and (x_j, x_{j+1}) with $\phi_j(x_j) = 1$ but $\phi_j(x_i) = 0$ for $j \neq i$. The hat function ϕ_j is also defined to be zero outside the open interval (x_{j-1}, x_{j+1}).

4.1 Something to Do

1. Sketch the graph of ϕ_1, ϕ_2 and ϕ_3. Describe the shape of the graph of any ϕ_j.

2. Give algebraic expressions in terms of x that define ϕ_1, ϕ_2 and ϕ_3. Generalize these expressions for ϕ_j.

5 The Subspace V_M of V

Let us define the subset V_M of V to be the collection of all functions v in V such that v is linear on each subinterval (x_{j-1}, x_j). Consider the nodes $x_1, x_2, \ldots, x_M$. Let $\eta_j = v(x_j)$.

5.1 Something to Do

1. Consider V_2. Sketch the graph of each $v \in V_2$ described below:

   (a) $\eta_1 = 2, \eta_2 = 3$.

   (b) $\eta_1 = -4, \eta_2 = 2$.

2. For each v given in the previous item, find a linear combination of ϕ_1 and ϕ_2 that is equal to v. In other words, find constants a_1 and a_2 such that $v = a_1\phi_1 + a_2\phi_2$.

3. Verify your answer in the previous item by doing the following: Using the values of a_1 and a_2 in (2) above and the algebraic expressions for ϕ_1, ϕ_2 and v, show that

   $$v(x) = a_1\phi_1(x) + a_2\phi_2(x)$$

   for every x in the closed interval $[0, 1]$.

4. Make a generalization. Given $v \in V_M$ and its values $\eta_1, \eta_2, \ldots, , \eta_M$ at the nodes, write v as a linear combination of $\phi_1, \phi_2, \ldots, \phi_M$.

5. Show that V_M is a subspace of V.

The above questions have guided us to conclude that any function $v \in V_M$ is uniquely determined by its values at the nodes $x_1, x_2, \ldots, x_M$. We also conclude that any $v \in V_M$ is a unique linear combination of the hat functions $\phi_1, \phi_2, \ldots, \phi_M$.

We now consider the collection of hat functions $H_M = \{\phi_1, \phi_2, \ldots, \phi_M\}$. Recall the *span* of H_M to be the set of all possible linear combinations of hat functions in H_M. But H_M is also contained in the vector space V_M. So *span* $H_M = V_M$.

5.2 Something to Do

1. Prove your generalization in item (4) above.

2. Prove that the set of hat functions H_M is a linearly independent set.

3. Is H_M a *basis* for V_M?

6 Discretization of the Variational Formulation

We have shown earlier that the problems $(\mathcal{D})$ and $(\mathcal{V})$ are equivalent. When we have solved one, then we have also solved the other. The finite element method solves the variational formulation numerically. To solve the variational problem numerically is to solve its discretized form $(\mathcal{V}_M)$:

$$(\mathcal{V}_M)\begin{cases} \text{Find } u_M \in V_M \text{ such that} \\ (u'_M, v') = (f, v) \ \ \forall v \in V_M; \\ f \ \text{ is continuous on } \ [0,1]. \end{cases}$$

Now, we have shown earlier that

$$u_M = \xi_1\phi_1 + \xi_2\phi_2 + \ldots + \xi_M\phi_M$$

for some vector

$$\xi = [\ \xi_1 \quad \xi_2 \quad \ldots \quad \xi_M\ ]^T$$

The equation $(u'_M, v') = (f, v)$ holds if v is the hat function $\phi_j \in V_M$ so for $j = 1, 2, \ldots, M$, we have

$$(\xi_1\phi'_1 + \xi_2\phi'_2 + \ldots + \xi_M\phi'_M, \phi'_j) = (f, \phi_j).$$

Then

$$(\xi_1\phi'_1, \phi'_j) + (\xi_2\phi'_2, \phi'_j) + \ldots + (\xi_M\phi'_M, \phi'_j) = (f, \phi_j),$$

which can be written as

$$\xi_1(\phi'_1, \phi'_j) + \xi_2(\phi'_2, \phi'_j) + \ldots + \xi_M(\phi'_M, \phi'_j) = (f, \phi_j).$$

This yields a system of M linear equations with M unknowns $\xi_1, \xi_2, \ldots, \xi_M$. The ξ_j are precisely the values of u_M at the nodes. The system is as follows:

$$\left\{\begin{array}{ccccccccc} \xi_1(\phi'_1,\phi'_1) & + & \xi_2(\phi'_2,\phi'_1) & + & \ldots & + & \xi_M(\phi'_M,\phi'_1) & = & (f,\phi_1) \\ \xi_1(\phi'_1,\phi'_2) & + & \xi_2(\phi'_2,\phi'_2) & + & \ldots & + & \xi_M(\phi'_M,\phi'_2) & = & (f,\phi_2) \\ \vdots & & \vdots & & \vdots & & \vdots & & \\ \xi_1(\phi'_1,\phi'_M) & + & \xi_2(\phi'_2,\phi'_M) & + & \ldots & + & \xi_M(\phi'_M,\phi'_M) & = & (f,\phi_M) \end{array}\right\}$$

which can also be written as

$$\left\{\begin{array}{ccccccccc} \xi_1(\phi'_1,\phi'_1) & + & \xi_2(\phi'_1,\phi'_2) & + & \ldots & + & \xi_M(\phi'_1,\phi'_M) & = & (f,\phi_1) \\ \xi_1(\phi'_2,\phi'_1) & + & \xi_2(\phi'_2,\phi'_2) & + & \ldots & + & \xi_M(\phi'_2,\phi'_M) & = & (f,\phi_2) \\ \vdots & & \vdots & & \vdots & & \vdots & & \\ \xi_1(\phi'_M,\phi'_1) & + & \xi_2(\phi'_M,\phi'_2) & + & \ldots & + & \xi_M(\phi'_M,\phi'_M) & = & (f,\phi_M) \end{array}\right\}.$$

In matrix form, we write

$$\begin{bmatrix} (\phi'_1,\phi'_1) & (\phi'_1,\phi'_2) & \ldots & (\phi'_1,\phi'_M) \\ (\phi'_2,\phi'_1) & (\phi'_2,\phi'_2) & \ldots & (\phi'_2,\phi'_M) \\ \vdots & \vdots & \ddots & \vdots \\ (\phi'_M,\phi'_1) & (\phi'_M,\phi'_2) & \ldots & (\phi'_M,\phi'_M) \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_M \end{bmatrix} = \begin{bmatrix} (f,\phi_1) \\ (f,\phi_2) \\ \vdots \\ (f,\phi_M) \end{bmatrix}.$$

Here, the *stiffness matrix* A has entries $a_{ij} = (\phi'_i, \phi'_j)$ and the *load vector* b has components $b_i = (f, \phi_i)$.

6.1 Something to Do

1. Compute a general formula for the entries of the stiffness matrix A in problem $(\mathcal{D}_1)$. Show that A is the tridiagonal $M \times M$ matrix shown earlier in the introduction.

2. Compute a general formula for the components of the load vector b in problem $(\mathcal{D}_1)$. Show that this formula can be written in the form shown earlier in the introduction.

7 The Existence of A^{-1}

Here, we answer the important question why the matrix equation $A\xi = b$ has a unique solution and hence, the numerical solution exists.

We note that A is a *symmetric matrix* since $(\phi_i', \phi_j') = (\phi_j', \phi_i')$. To show that A is nonsingular, we will show that A is *positive definite*. In other words, we will show that $\xi^T A\xi > 0$ for every nonzero vector ξ in $\mathbb{R}^M$.

Let $\xi \neq 0$, where $0 = [\ 0 \quad 0 \quad \ldots \quad 0\ ]^T$, the zero vector in $\mathbb{R}^M$. It is possible for ξ to have some components that are zero but not all. Then

$$
\begin{aligned}
\xi^T A\xi &= [\ \xi_1 \quad \xi_2 \quad \ldots \quad \xi_M\ ] \begin{bmatrix} (\phi_1', \phi_1') & (\phi_1', \phi_2') & \ldots & (\phi_1', \phi_M') \\ (\phi_2', \phi_1') & (\phi_2', \phi_2') & \ldots & (\phi_2', \phi_M') \\ \vdots & \vdots & \ddots & \vdots \\ (\phi_M', \phi_1') & (\phi_M', \phi_2') & \ldots & (\phi_M', \phi_M') \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_M \end{bmatrix} \\
&= [\ \xi_1 \quad \xi_2 \quad \ldots \quad \xi_M\ ] \begin{bmatrix} \xi_1(\phi_1', \phi_1') + \xi_2(\phi_1', \phi_2') + \ldots + \xi_M(\phi_1', \phi_M') \\ \xi_1(\phi_2', \phi_1') + \xi_2(\phi_2', \phi_2') + \ldots + \xi_M(\phi_2', \phi_M') \\ \vdots \quad \vdots \quad \vdots \\ \xi_1(\phi_M', \phi_1') + \xi_2(\phi_M', \phi_2') + \ldots + \xi_M(\phi_M', \phi_M') \end{bmatrix} \\
&= [\ \xi_1 \quad \xi_2 \quad \ldots \quad \xi_M\ ] \begin{bmatrix} (\phi_1', \xi_1\phi_1') + (\phi_1', \xi_2\phi_2') + \ldots + (\phi_1', \xi_M\phi_M') \\ (\phi_2', \xi_1\phi_1') + (\phi_2', \xi_2\phi_2') + \ldots + (\phi_2', \xi_M\phi_M') \\ \vdots \quad \vdots \quad \vdots \\ (\phi_M', \xi_1\phi_1') + (\phi_M', \xi_2\phi_2') + \ldots + (\phi_M', \xi_M\phi_M') \end{bmatrix} \\
&= [\ \xi_1 \quad \xi_2 \quad \ldots \quad \xi_M\ ] \begin{bmatrix} (\phi_1', [\xi_1\phi_1' + \xi_2\phi_2' + \ldots + \xi_M\phi_M']) \\ (\phi_2', [\xi_1\phi_1' + \xi_2\phi_2' + \ldots + \xi_M\phi_M']) \\ \vdots \quad \vdots \quad \vdots \\ (\phi_M', [\xi_1\phi_1' + \xi_2\phi_2' + \ldots + \xi_M\phi_M']) \end{bmatrix} \\
&= \left(\xi_1\phi_1, \sum_{j=1}^{M} \xi_j\phi_j'\right) + \left(\xi_2\phi_2, \sum_{j=1}^{M} \xi_j\phi_j'\right) + \ldots + \left(\xi_M\phi_M, \sum_{j=1}^{M} \xi_j\phi_j'\right) \\
&= \left(\sum_{i=1}^{M} \xi_i\phi_i', \sum_{j=1}^{M} \xi_j\phi_j'\right) \\
&= (u_M', u_M') \\
&\geq 0.
\end{aligned}
$$

Thus for any nonzero vector $\xi \in \mathbb{R}^M$, we have $\xi^T A\xi \geq 0$. But we need $\xi^T A\xi$ to be strictly positive to prove that A is positive definite. So we proceed further by noting that some component ξ_k of ξ is nonzero. So

$$\begin{aligned} 0 &< (\xi_k\phi_k', \xi_k\phi_k') \\ &\leq (v', v'). \end{aligned}$$

We have shown that A is positive definite, hence A is nonsingular.

8 Convergence of the Approximate Solution u_M to the Exact Solution u

The idea of the finite element method is that the exact solution is approximated by some piecewise linear function u_M in V_M. In every subspace V_M of V there exists a unique u_M, which is, in some sense, the best approximation of u in V_M. Our numerical procedure yields the values $\xi_1, \xi_2, \ldots, \xi_M$ of u_M at the nodes. These node values of u_M may differ from the node values of the exact solution u. The more nodes we have (i.e. the larger the M), the closer u_M gets to being a *smooth* function,which is what we want our solution u to be. Now a natural question that arises is whether u_M becomes a better approximation for u as M gets larger. In other words, will the magnitude of the difference (the *error*) between the node values of u_M and u become smaller as M grows larger? Getting an affirmative answer to this question is crucial to the success of the finite element method in approximating the solution. The answer that we want is given by the following theorem from [**?**], which we simply state without proof.

Theorem 8.1 *If $u_M \in V_M$ is an approximate solution of $\mathcal{D}$ then for every $x \in [0,1]$, we have*

$$|u(x) - u_M(x)| \leq \frac{m}{M+1}$$

where m is the maximum value of $f(y)$ over the whole closed interval $[0,1]$.

Note that m exists since f is continuous on $[0,1]$ so that the *Extreme -Value Theorem* applies. So as M grows bigger, we can expect the error to shrink to zero.

9 A Second Example

Our first example is problem $(\mathcal{D}_1)$. Earlier, we found a general formula for the stiffness matrix A and the load vector b that correspond to this problem. The components of the load vector b were obtained using straightforward integration.

Our second example is problem $(\mathcal{D}_2)$:

$$(\mathcal{D}_2) \begin{cases} -u''(x) = \sin(\pi x)^2, \quad 0 < x < 1; \\ u(0) = u(1) = 0. \end{cases}$$

9.1 Something to Do

1. Find the stiffness matrix for problem $(\mathcal{D}_2)$.

Finding the value of b_j is not straightforward since the integral $\int_0^1 f(x)\phi_j(x)dx$ has no closed form. It is not easy to find an antiderivative expression for $f(x)\phi_j(x)$. What we will do is approximate the value of $\int_0^1 f(x)\phi_j(x)dx$ using known theorems in calculus.

9.2 Something to Do

1. As an application of the generalization of the *Mean-Value Theorem* in calculus, there exist $c_1 \in (x_{j-1}, x_j)$ and $c_2 \in (x_j, x_{j+1})$ such that

$$\int_{x_{j-1}}^{x_{j+1}} f(x)\phi_j(x)dx = f(c_1)\int_{x_{j-1}}^{x_j} \phi_j(x)dx + f(c_2)\int_{x_j}^{x_{j+1}} \phi_j(x)dx$$

   Show that

$$(f, \phi_j) = \left[\sin(\pi c_2)^2 + \sin(\pi c_1)^2\right]\left[\frac{1}{2(M+1)}\right].$$

2. The preceding item indicates that (f, ϕ_j) can still be computed if c_1 and c_2 can be found. Unfortunately, c_1 and c_2 are not easy to obtain. Suppose we try to approximate c_1 and c_2 by the midpoints of the open intervals (x_{j-1}, x_j) and (x_j, x_{j+1}). Find an expression for each of these midpoints in terms of j and M.

3. Using the expressions obtained in (2) above, give an expression that might approximate (f, ϕ_j).

4. Indeed the expression to be obtained in (3) above is good enough approximation for (f, ϕ_j) for a sufficiently large M. What property of the function f makes this possible?

10 Summary

The basic ideas of the finite element method can be understood with only a background in elementary calculus and linear algebra. In this lecture, we have seen how problem $(\mathcal{D})$ can be solved numerically by following the procedure outlined in section 1. The mathematics involved in arriving at this procedure was discussed. Two illustrative examples were given with the second example slightly more difficult than the first. The working computer programs in Scilab and their output graphs can be found in the pages after the list of references. We have only done the mathematics leading to the procedure. The actual computer programming is another story since the time is not enough. It could be the topic of another lecture or seminar in the future.

11 Beyond the Fundamentals

The finite element method is a very broad subject. There is a lot more about this method that involves deep and highly technical mathematics. There are many possible extensions that can be

developed. With the appropriate adjustments,the method can be applied to more complicated ordinary and partial differential equations involving time and space variables. The method has become a dominating technique in computational mathematics with important applications in many areas of science, engineering and finance.

12 Appendix

Here are some useful definitions and results for quick reference as we go through the above discussions.

12.1 Linear Algebra

The following definitions and results can be found in [**?**]

Definition 12.1 *A real vector space is a set V of elements on which we have two operations $\oplus$ and $\odot$ defined with the following properties:*

(a) If u and v are any elements in V, then $u \oplus v$ is in V. (We say that V is closed under the operation $\oplus$.)

(1) $u \oplus v = v \oplus u$ for all u, v in V.

(2) $u \oplus (v \oplus w) = (u \oplus v) \oplus w$ for all u, v, w in V.

(3) There exists an element 0 in V such that $u \oplus 0 = 0 \oplus u = u$ for any u in V.

(4) For each u in V there exists an element $-u$ in V such that $u \oplus -u = -u \oplus u = 0$.

(b) If u is any element in V and c is any real number, then $c \odot u$ is in V (i.e. V is closed under the operation $\odot$).

(5) $c \odot (u \oplus v) = c \odot u \oplus c \odot v$ for any u, v in V and any real number c.

(6) $(c + d) \odot u = c \odot u \oplus d \odot u$ for any u in V and any real numbers c and d.

(7) $c \odot (d \odot u) = (cd) \odot u$ for any u in V and any real numbers c and d.

(8) $1 \odot u = u$ for any u in V.

The elements of V are called vectors; the elements of the set of real numbers $\mathbb{R}$ are called scalars. The operation $\oplus$ is called vector addition; the operation $\odot$ is called scalar multiplication. The vector 0 in property (3) is called the zero vector. The vector $-u$ in property (4) is called the negative of u. In V, the vector 0 is unique in the sense that it is the only element of V that satisfies (3). Similarly, $-u$ is unique in the sense that no other vector in V when added to u gives the zero vector.

For our own purposes, our vector space V will be a set of functions satisfying certain conditions; the operation $\oplus$ will be the ordinary function addition $+$, and the operation $\odot$ will simply be the usual multiplication.

Definition 12.2 *Let V be a vector space and W a nonempty subset of V. If W is a vector space with respect to the operations in V, then W is called a subspace of V.*

Theorem 12.1 *Let V be a vector space with operations $\oplus$ and $\odot$, and let W be a nonempty subset of V. Then W is a subspace of V if and only if the following conditions hold:*

(a) If u and v are any vectors in W, then $u \oplus v$ is in W.

(b) If c is any real number and u is any vector in W, then $c \odot u$ is in W.

Suppose $v_1, v_2, \ldots, v_k$ are vectors in a vector space V, and $a_1, a_2, \ldots, a_k$ are real numbers, then $a_1v_1 + a_2v_2 + \cdots + a_kv_k$ is called a *linear combination* of the vectors $v_1, v_2, \ldots, v_k$.

Definition 12.3 *If $S = \{v_1, v_2, \ldots, v_k\}$ is a set of vectors in a vector space V, then the set of all vectors in V that are linear combinations of the vectors in S is denoted by*

$$span\ S \quad or \quad span\{v_1, v_2, \ldots, v_k\}$$

Theorem 12.2 *Let $S = \{v_1, v_2, \ldots, v_k\}$ be a set of vectors in a vector space V. Then span S is a subspace of V.*

Definition 12.4 *A set of vectors $S = \{v_1, v_2, \ldots, v_k\}$ in a vector space V is said to span V if every vector in V is a linear combination of $v_1, v_2, \ldots, v_k$. We also say that S spans V or V is spanned by S or span $S = V$.*

Definition 12.5 *The vectors $v_1, v_2, \ldots, v_k$ in a vector space V are said to be linearly independent if whenever*

$$a_1v_1 + a_2v_2 + \cdots + a_kv_k = 0$$

then

$$a_1 = a_2 = \cdots = a_k = 0.$$

Otherwise, the vectors are said to be linearly dependent. A set S of vectors is said to be linearly independent or linearly dependent if the vectors have the corresponding property.

Definition 12.6 *The vectors $v_1, v_2, \ldots, v_k$ in a vector space V are said to form a basis for V if*

(a) $v_1, v_2, \ldots, v_k$ span V and

(b) $v_1, v_2, \ldots, v_k$ are linearly independent.

Definition 12.7 *If $A = [a_{ij}]$ is an $m \times n$ matrix, then the transpose A^T of A, written $a^T = [a_{ij}^T]$, is the $n \times m$ matrix defined by $a_{ij}^T = a_{ji}$. Thus the transpose of A is obtained by interchanging rows and columns of A.*

Definition 12.8 *A matrix A is called symmetric if $A^T = A$. Thus for a symmetric matrix, $a_{ij} = a_{ji}$.*

The identity matrix I_n is the $n \times n$ matrix whose entries on the main diagonal are all 1 and zero elsewhere.

Definition 12.9 *An $n \times n$ matrix is nonsingular if A^{-1} exists, where A^{-1} is the unique matrix such that $AA^{-1} = A^{-1}A = I_n$*

Theorem 12.3 *An $n \times n$ matrix A is nonsingular if and only if $Ax = 0$ has only the trivial solution.*

Theorem 12.4 *An $n \times n$ matrix A is nonsingular if and only if $Ax = b$ has a unique solution for any $n \times 1$ matrix b. (We also call b a vector in $\mathbb{R}^n$.)*

Theorem 12.5 *An symmetric matrix A is said to be positive definite if $x^T Ax > 0$ for every nonzero vector x in $\mathbb{R}^n$.*

Theorem 12.6 *A positive definite matrix A is nonsingular.*

12.2 Elementary Calculus

The following definitions and results can be found in [**?**]

Definition 12.10 *Let f be a function defined at every number in some open interval containing a, except possibly at the number a itself. The limit of $f(x)$ as x approaches a is L, written as*

$$\lim_{x \to a} f(x) = L$$

if the following statement is true: Given any $\varepsilon > 0$, however small, there exists a $\delta > 0$ such that if $0 < |x - a| < \delta$ then $|f(x) - L| < \varepsilon$.

Definition 12.11 *The function f is said to be continuous at the number a if and only if the following three conditions are satisfied:*

(i) $f(a)$ *exists;*

(ii) $\lim_{x \to a} f(x)$ *exists;*

(iii) $\lim_{x \to a} f(x) = f(a)$.

If one or more of these three conditions fails to hold at a, the function f is said to be discontinuous at a.

Definition 12.12 *Let f be a function defined at every number in some open interval (a, c). Then the limit of $f(x)$, as x approaches a from the right, is L, written*

$$\lim_{x \to a^+} f(x) = L$$

if for any $\varepsilon > 0$, however small, there exists a $\delta > 0$ such that if $0 < x - a < \delta$ then $|f(x) - L| < \varepsilon$.

Definition 12.13 *Let f be a function defined at every number in some open interval (d, a). Then the limit of $f(x)$, as x approaches a from the left, is L, written*

$$\lim_{x \to a^-} f(x) = L$$

if for any $\varepsilon > 0$, however small, there exists a $\delta > 0$ such that if $0 < a - x < \delta$ then $|f(x) - L| < \varepsilon$.

Theorem 12.7 $\lim_{x\to a} f(x)$ *exists and is equal to* L *if and only if* $\lim_{x\to a^+} f(x)$ *and* $\lim_{x\to a^-} f(x)$ *both exist and both are equal to* L.

Definition 12.14 *The function* f *is said to be continuous from the right at the number* a *if and only if the following three conditions are satisfied:*

(i) $f(a)$ *exists;*

(ii) $\lim_{x\to a^+} f(x)$ *exists;*

(iii) $\lim_{x\to a^+} f(x) = f(a)$.

If one or more of these three conditions fails to hold at a*, the function* f *is said to be discontinuous from the right at* a.

Definition 12.15 *The function* f *is said to be continuous from the left at the number* a *if and only if the following three conditions are satisfied:*

(i) $f(a)$ *exists;*

(ii) $\lim_{x\to a^-} f(x)$ *exists;*

(iii) $\lim_{x\to a^-} f(x) = f(a)$.

If one or more of these three conditions fails to hold at a*, the function* f *is said to be discontinuous from the left at* a.

Theorem 12.8 *A function* f *is continuous at the number* a *if and only if it is both continuous from the right and continuous from the left at* a.

Definition 12.16 *A function* f *is continuous in an open interval* (a, b) *if it is continuous at every number in* (a, b).

Definition 12.17 *A function* f *is continuous in the closed interval* $[a, b]$ *if all of the following hold:*

(i) f *is continuous in the open interval* (a, b).

(ii) f *is continuous from the right at* a.

(iii) f *is continuous from the left at* b.

Theorem 12.9 *If* f *and* g *are two functions that are continuous at the number* a*, then*

(i) $f + g$ *is continuous at* a*;*

(ii) $f - g$ *is continuous at* a*;*

(iii) fg *is continuous at* a*;*

(iv) f/g *is continuous at* a*, provided* $g(a) \neq 0$*;*

(v) cf is continuous at a, where c is a constant.

The above theorem holds when '*continuous*' is replaced by '*continuous from the right*' or '*continuous from the left*'

Definition 12.18 *A function F is called an antiderivative of the function f on an interval I if $F'(x) = f(x)$ for every value of x in I.*

Theorem 12.10 *If f is continuous on $[a, b]$, then it has an antiderivative.*

Definition 12.19 *Consider the points $x_0, x_1, x_2, \ldots, x_n$ with $a = x_0 < x_1 < x_2 < \cdots < x_k = b$. The collection of subintervals $\{[x_0, x_1], [x_1, x_2], \ldots, [x_{n-1}, x_n]\}$ is called a partition Δ of $[a, b]$. The length of the interval $[x_{i-1}, x_i]$ is denoted by $\Delta_i x$. The length of the longest subinterval in the Δ is called the norm $\|\Delta\|$ of the partition. If the subintervals are all of the same length, then the partition is called a uniform partition.*

Definition 12.20 *Let f be a function whose domain includes the closed interval $[a, b]$. We say that f is Riemann integrable on $[a, b]$ if there is a number L satisfying the condition that, for every $\varepsilon > 0$ there exists a $\delta > 0$ such that for every partition Δ for which $\|\Delta\| < \delta$, and for any w_i in the closed interval $[x_{i-1}, x_i], i = 1, 2, \ldots, n$, then*

$$\left| \sum_{i=1}^{n} f(w_i)\Delta_i x - L \right| < \varepsilon.$$

For such a situation, we write

$$\lim_{\|\Delta\| \to 0} \sum_{i=1}^{n} f(w_i)\Delta_i x = L$$

Definition 12.21 *If f is a function defined on the closed interval $[a, b]$, then the definite (Riemann) integral of f from a to b, denoted by $\int_a^b f(x)dx$, is given by*

$$\int_a^b f(x)dx = \lim_{\|\Delta\| \to 0} \sum_{i=1}^{n} f(w_i)\Delta_i x.$$

Theorem 12.11 *If a function f is continuous on the closed interval $[a, b]$, then it is (Riemann) integrable on $[a, b]$.*

Theorem 12.12 *If the function f is Riemann integrable on $[a, b]$ and $c \in (a, b)$, then f is Riemann integrable on $[a, c]$ and $[c, b]$ with*

$$\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx.$$

Theorem 12.13 *If $\lim_{x \to c} f(x)$ exists and is positive, then there is an open interval containing c such that $f(x) > 0$ for every x distinct from c in the open interval.*

Theorem 12.14 *If* $\lim_{x \to c} f(x)$ *exists and is negative, then there is an open interval containing* c *such that* $f(x) < 0$ *for every* x *distinct from* c *in the open interval.*

Theorem 12.15 (The Mean-Value Theorem for Derivatives) *Let* f *be a function such that*

(i) it is continuous on the closed interval $[a, b]$*;*

(ii) it is differentiable on the open interval (a, b)*.*

Then there exists a number c *in the open interval* (a, b) *such that*

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Theorem 12.16 (The Mean-Value Theorem for Integrals) *If the function* f *is continuous on the closed interval* $[a, b]$*, there exists a number* c *in* $[a, b]$ *such that*

$$\int_a^b f(x)dx = f(c)(b - c).$$

Theorem 12.17 (The Intermediate-Value Theorem) *If the function* f *is continuous on the closed interval* $[a, b]$*, and if* $f(a) \neq f(b)$*, then for any number* k *between* $f(a)$ *and* $f(b)$*, there exists a number* c *between* a *and* b *such that* $f(c) = k$*.*

Theorem 12.18 (Generalization of the Mean-Value Theorem) *If* f *and* g *are two functions continuous on the closed interval* $[a, b]$ *and* $g(x) > 0$ *for all* x *in the open interval* (a, b)*, then there exists a number* c *in* $[a, b]$ *such that*

$$\int_a^b f(x)g(x)dx = f(c) \int_a^b g(x)dx.$$

Theorem 12.19 *If* f *is a function such that* $f'(x) = 0$ *for all values of* x *in the interval* I*, then* f *is constant on* I*.*

The results found below can be shown using those given above.

Theorem 12.20 *Suppose* f *is continuous on* $[a, b]$ *with* $f(x) > 0$ *for every* x *in the open interval* (a, b)*, then*

$$\int_a^b f(x)dx > 0$$

Theorem 12.21 *Suppose* $\int_a^b [f(x)]^2 dx = 0$ *where* f *is a continuous function in the open interval* (a, b)*, then* $f(x) = 0$ *for every* x *in the open interval* (a, b)*.*

Definition 12.22 *A function f is said to be piecewise continuous on the closed interval $[0,1]$ if there are points $x_1, x_2, \ldots, x_k$ with $0 < x_1 < x_2 < \cdots < x_k < 1$ such that f is continuous in each of the open intervals $(0, x_1), (x_1, x_2), \ldots, (x_{k-1}, x_k), (x_k, 1)$.*

Definition 12.23 *A function f is said to be bounded on $[0,1]$ if there is a positive number m such that $-m \leq f(x) \leq m$ for every x in $[0,1]$.*

Theorem 12.22 *Suppose w is a continuous function in the open interval $(0,1)$ and*

$$\int_0^1 w(x)v(x)dx = 0$$

for every $v \in V$ where

$$V = \{v : v \text{ continuous on } [0,1], v' \text{ piecewise continuous, bounded on } [0,1], v(0) = v(1) = 0\},$$

then $w(x) = 0$ for every $x \in (0,1)$.

References

[1] Johnson, Claes, *Numerical Solutions of Partial Differential Equations by the Finite Element Method*, Cambridge University Press, Cambridege, 1987.

[2] Leithold, Louis, *The Calculus 7*, Harper Collins College Publishers, 1996.

[3] Kolman Bernard, *Elementary Linear Algebra, Sixth Edition*, Prentice-Hall International Inc., New Jersey, 1998.

Implementation of the Method Using Computers

Shown in the last few attached pages are the Scilab programs for problems $(\mathcal{D}_1)$ and $(\mathcal{D}_2)$, respectively. The corresponding outputs are also shown.

Note: Scilab is a freeware and you may download it from the site

`www.scilab.org`

.

Scilab tutorials may be downloaded from, for example,

`http://comptlsci.anu.edu.au/Numerical-Methods/tutorial-all.pdf`