

Chapter 7

Theoretical foundations of FEM

Using finite element methods, we need to answer these questions:

- What is the right functional space V for the solution?
- What is the right *weak* or *variational* form of a differential equation.
- What kind of basis functions or finite element spaces should we choose?
- How accurate is a FEM solution?

We try briefly to answer these questions in this Chapter. Remember that finite element methods are based on integral forms, not in the point-wise sense as in finite difference methods. We need to generalize the theory corresponding to the point-wise form to integral forms.

7.1 Functional space of $C^m(\Omega)$

Functional Space is a *set* of functions with operations. For example,

$$C(\Omega) = C^0(\Omega) = \left\{ u(x), \quad u(x) \text{ is continuous on } \Omega \right\} \quad (7.1)$$

is a linear space that contains all continuous functions on Ω . It is a linear space because for any real numbers α and β , we have,

$$\text{if } u_1 \in C(\Omega), \quad u_2 \in C(\Omega), \quad \text{then } \alpha u_1 + \beta u_2 \in C(\Omega).$$

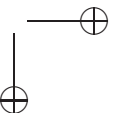
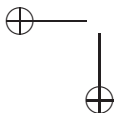
In the notation, Ω is a domain where the functions are defined, for example, $\Omega = [0, 1]$.

The functional space with first order continuous derivatives is defined as

$$C^1(\Omega) = \left\{ u(x), \quad u(x), u'(x) \text{ are continuous on } \Omega \right\}, \quad (7.2)$$

and similarly

$$C^m(\Omega) = \left\{ u(x), \quad u(x), u'(x), \dots, u^{(m)} \text{ are continuous on } \Omega \right\}. \quad (7.3)$$



Obviously, we have,

$$C^0 \supset C^1 \supset \dots \supset C^m \supset \dots \quad (7.4)$$

Let m go to the infinity, we define

$$C^\infty(\Omega) = \{u(x), \quad u(x) \text{ is indefinitely differentiable on } \Omega\}. \quad (7.5)$$

For examples, e^x , $\sin x$ and elementary functions are in C^∞ .

7.1.1 Multi-dimensional spaces and multi-index notations

Consider multi-dimensional functions $u(x_1, x_2, \dots, x_n)$, or $u(\mathbf{x})$, $\mathbf{x} \in R^n$. We use multi-index notations to simplify the expressions of partial derivatives. Let $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)$, $\alpha_i \geq 0$ be an integer vector in R^n . For example, if $n = 5$, $\alpha = (1, 2, 0, 0, 2)$ is one of α 's. The multi-index notation is used to express the following partial derivative

$$D^\alpha u(\mathbf{x}) = \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \dots \partial x_n^{\alpha_n}}, \quad |\alpha| = \alpha_1 + \alpha_2 + \dots + \alpha_n, \quad \alpha_i \geq 0. \quad (7.6)$$

Example 7.1. When $n = 2$ and $u(\mathbf{x}) = u(x_1, x_2)$, all possible $D^\alpha u$ are

$$\begin{aligned} \alpha = (2, 0), \quad D^\alpha u &= \frac{\partial^2 u}{\partial x_1^2}, \\ \alpha = (1, 1), \quad D^\alpha u &= \frac{\partial^2 u}{\partial x_1 \partial x_2}, \\ \alpha = (0, 2), \quad D^\alpha u &= \frac{\partial^2 u}{\partial x_2^2}. \end{aligned}$$

The C^m space can be defined as

$$C^m(\Omega) = \left\{ u(x_1, x_2, \dots, x_n), \quad D^{|\alpha|} u \text{ are continuous on } \Omega, \quad |\alpha| \leq m \right\}. \quad (7.7)$$

That is all possible derivatives up to order m are continuous on Ω .

Example 7.2. When $n = 2$, $m = 3$, then u , u_x , u_y , u_{xx} , u_{xy} , u_{yy} , u_{xxx} , u_{xxy} , u_{xyy} , u_{yyy} all have to be continuous on Ω if $u \in C^3(\Omega)$. Note that $C^m(\Omega)$ has infinite dimensions.

The **distance** in $C^0(\Omega)$ is defined as,

$$d(u, v) = \max_{x \in \Omega} |u(x) - v(x)| \quad (7.8)$$

A linear space with distance defined is called a *metric space*. A distance is a non-negative function of u and v that satisfies,

$$d(u, v) \geq 0, \quad d(u, v) = 0, \quad \text{iff } u \equiv v. \quad (7.9)$$

$$d(u, v + w) \leq d(u, v) + d(u, w), \quad \text{triangle inequality.} \quad (7.10)$$

In the expressions above, the word *iff* means *if and only if*.



The **Norm** in $C^0(\Omega)$ is a non-negative function of u defined by

$$\|u(x)\| = d(u, \theta) = \max_{x \in \Omega} |u(x)|, \quad (7.11)$$

where θ is the zero element. A norm needs to satisfy

$$\|u(x)\| \geq 0, \quad \|u(x)\| = 0, \quad \text{iff } u \equiv 0. \quad (7.12)$$

$$\|\alpha u(x)\| = |\alpha| \|u(x)\|, \quad \alpha \text{ is a number}, \quad (7.13)$$

$$\|u(x) + v(x)\| \leq \|u(x)\| + \|v(x)\|, \quad \text{triangle inequality}. \quad (7.14)$$

A linear space with a norm defined is called a *normed space*. In $C^m(\Omega)$, the distance and the norm are defined as

$$d(u, v) = \max_{0 \leq |\alpha| \leq m} \max_{x \in \Omega} |D^\alpha u(x) - D^\alpha v(x)|, \quad (7.15)$$

$$\|u(x)\| = \max_{0 \leq |\alpha| \leq m} \max_{x \in \Omega} |D^\alpha u|. \quad (7.16)$$

7.2 Spaces in integral forms – $L^2(\Omega)$ and $L^p(\Omega)$

In analogue to point-wise spaces $C^m(\Omega)$, we can define spaces $H^m(\Omega)$ called Sobolev spaces using integral forms. The square-integrable space $H^0(\Omega) = L^2(\Omega)$ is defined as

$$L^2(\Omega) = \left\{ u(x), \int_{\Omega} u^2(x) dx < \infty \right\}, \quad (7.17)$$

corresponding to the point-wise $C^0(\Omega)$ space. It is easy to see that

$$C(0, 1) = C^0(0, 1) \subset L^2(0, 1).$$

Example 7.3. It is easy to verify that $u(x) = 1/x^{1/4} \notin C^0$, but

$$\int_0^1 \left(\frac{1}{x^{1/4}} \right)^2 dx = \int_0^1 \frac{1}{\sqrt{x}} dx = 2 < \infty.$$

Therefore $u(x) \in L^2(0, 1)$.

The distance in $L^2(\Omega)$ is defined as

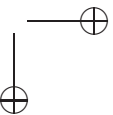
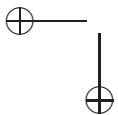
$$d(f, g) = \left\{ \int_{\Omega} |f - g|^2 dx \right\}^{1/2}. \quad (7.18)$$

Thus $L^2(\Omega)$ is a metric space. We say that $f \equiv g$ (identical) if $d(f, g) = 0$. The following two functions are identical in $L^2(-2, 2)$

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0, \end{cases} \quad g(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 & \text{if } x > 0. \end{cases}$$

The norm in the $L^2(\Omega)$ space is defined as

$$\|u\|_{L^2} = \|u\|_0 = \left\{ \int_{\Omega} |u|^2 dx \right\}^{1/2}. \quad (7.19)$$



We can prove that these definitions satisfy the requirements of the properties requirements for the distance and the norm.

$L^2(\Omega)$ is a complete space means that any Cauchy sequence $\{f_n\}$ in L^2 has a limit in L^2 . In other words, there is a $f \in L^2(\Omega)$ such that

$$\lim_{n \rightarrow \infty} \|f_n - f\|_{L^2} = 0, \quad \text{or} \quad \lim_{n \rightarrow \infty} f_n = f.$$

A Cauchy sequence is a sequence that satisfies the property: for any given positive number ϵ , there is an integer N , such that

$$\|f_n - f_m\|_{L^2} < \epsilon, \quad \text{if } m \geq N, \quad n \geq N.$$

A complete normed space is called a *Banach* space (a Cauchy sequence converges in terms of the norm). Therefore L^2 is a Banach space.

7.2.1 The inner product in L^2

In R^n , given any two vectors,

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in R^n, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in R^n,$$

we know that the inner product is defined as

$$(x, y) = x^T y = \sum_{i=1}^n x_i y_i = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n.$$

Similarly, the inner product in the L^2 space is defined as

$$(f, g) = \int_{\Omega} f(x)g(x)dx, \quad \text{if } f \in L^2(\Omega), \quad g \in L^2(\Omega). \quad (7.20)$$

It satisfies the requirements of the definition of an inner product:

$$\begin{aligned} (f, g) &= (g, f), \\ (\alpha f, g) &= (f, \alpha g) = \alpha(f, g) \quad \forall \alpha \in R, \\ (f, g + w) &= (f, g) + (f, w). \end{aligned}$$

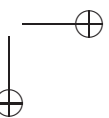
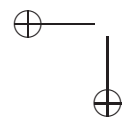
for any $f \in L^2(\Omega)$ and $g \in L^2(\Omega)$.

With the inner product, the weak form for the simple model problem

$$-u'' = f, \quad 0 < x < 1, \quad u(0) = u(1) = 0,$$

can be written as

$$(u', v') = (f, v), \quad \forall v \in H^1(0, 1),$$



and the minimization forms is

$$\min_{v \in H_0^1(0,1)} F(v) : F(v) = \frac{1}{2}(v', v') - (f, v).$$

The norm, distance, and the inner product in $L^2(\Omega)$ have the following relations

$$\|u\|_0 = \sqrt{(u, u)} = d(u, \theta) = \left\{ \int_{\Omega} |u|^2 dx \right\}^{1/2}. \quad (7.21)$$

7.2.2 The Cauchy-Schwartz inequality in $L^2(\Omega)$

For a Hilbert space with a norm (u, v) and its resulted norm $\|u\| = \sqrt{(u, u)}$, the Cauchy-Schwartz inequality is the following,

$$|(u, v)| \leq \|u\|_0 \|v\|_0. \quad (7.22)$$

Below we list some examples of the Cauchy-Schwartz inequality corresponding to inner products in R^n and L^2 spaces:

$$\begin{aligned} \left| \sum_{i=1}^n x_i y_i \right| &\leq \left\{ \sum_{i=1}^n x_i^2 \right\}^{1/2} \left\{ \sum_{i=1}^n y_i^2 \right\}^{1/2}; \\ \left| \sum_{i=1}^n x_i \right| &\leq \sqrt{n} \left\{ \sum_{i=1}^n x_i^2 \right\}^{1/2}; \\ \left| \int_{\Omega} fg dx \right| &\leq \left\{ \int_{\Omega} f^2 dx \right\}^{1/2} \left\{ \int_{\Omega} g^2 dx \right\}^{1/2}; \\ \left| \int_{\Omega} f dx \right| &\leq \left\{ \int_{\Omega} f^2 dx \right\}^{1/2} \sqrt{V}, \end{aligned}$$

where V is the volume of Ω .

A proof of the Cauchy-Schwartz inequality

Noting that $(u, u) = \|u\|^2$, we construct a quadratic function of α given u and v as follows

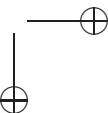
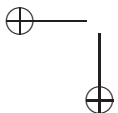
$$f(\alpha) = (u + \alpha v, u + \alpha v) = (u, u) + 2\alpha(u, v) + \alpha^2(v, v) \geq 0.$$

The quadratic function is non-negative, and therefore the discriminant of the quadratic form satisfies

$$\Delta = b^2 - 4ac \leq 0, \quad \text{i.e.} \quad 4(u, v)^2 - 4(u, u)(v, v) \leq 0, \quad \text{or} \quad (u, v)^2 \leq (u, u)(v, v).$$

Taking square root from both sides of the last inequality above, we then have the Cauchy-Schwartz inequality.

A complete Banach space with an inner product defined is called a *Hilbert space*. So $L^2(\Omega)$ is a Hilbert space (linear space, inner product, complete).



Summary of definitions of spaces

The following diagram shows relations among the definitions of spaces.

Metric Space (distance) \longrightarrow Normed Space (norm) \longrightarrow Banach space (complete)
 \longrightarrow Hilbert space (inner product).

7.2.3 The $L^p(\Omega)$ spaces

An $L^p(\Omega)$ space is defined as

$$L^p(\Omega) = \left\{ u(x), \int_{\Omega} |u(x)|^p dx < \infty \right\}. \quad (7.23)$$

The distance in $L^p(\Omega)$ is defined as

$$d(f, g) = \left\{ \int_{\Omega} |f - g|^p dx \right\}^{1/p}. \quad (7.24)$$

An $L^p(\Omega)$ space is a metric and complete space, so it is also a Banach space. But it is not a Hilbert space because there is no corresponding inner product can be defined unless $p = 2$.

7.3 The Sobolev spaces – related to the weak derivatives

Similar to $C^m(\Omega)$ spaces, we use $H^m(\Omega)$ to define function spaces with derivatives in integral forms. Such function spaces are part of the Sobolev spaces.

If there is no derivative, then we define the space

$$H^0(\Omega) = L^2(\Omega) = \left\{ v(x), \int_{\Omega} |v|^2 dx < \infty \right\}. \quad (7.25)$$

7.3.1 The definition of the weak derivatives

We know that if $u(x) \in C^1(0, 1)$, then for any function $\phi \in C^1(0, 1)$, $\phi(0) = \phi(1) = 0$, we have

$$\int_0^1 u'(x)\phi(x) dx = u\phi \Big|_0^1 - \int_0^1 u(x)\phi'(x) dx = - \int_0^1 u(x)\phi'(x) dx. \quad (7.26)$$

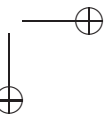
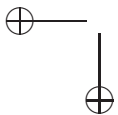
Such a $\phi(x)$ is called a testing function in $C_0^1(0, 1)$. Therefore we can define the weak derivative of $u(x) \in L^2(\Omega) = H^0(\Omega)$ as a function $v(x)$ that satisfies

$$\int_{\Omega} v(x)\phi(x) dx = - \int_{\Omega} u(x)\phi'(x) dx \quad (7.27)$$

for all $\phi(x) \in C_0^1(\Omega)$, $\phi(0) = \phi(1) = 0$. If such a function exists, it is called the first order weak derivative of $u(x)$ and it is denoted as $v(x) = u'(x)$.

Similarly, we can define the m -th order weak derivative of $u(x) \in H^0(\Omega)$ as a function $v(x)$ that satisfies

$$\int_{\Omega} v(x)\phi(x) dx = (-1)^m \int_{\Omega} u(x)\phi^{(m)}(x) dx \quad (7.28)$$



for all $\phi(x) \in C_0^m(\Omega)$. That is

$$\phi(x) \in C^m(\Omega), \phi(x) = \phi'(x) = \dots = \phi^{(m-1)}(x) = 0, \quad \text{if } x \in \partial\Omega.$$

If such a function $v(x)$ exists, it is called the m -th order weak derivative of $u(x)$ and it is denoted as $v(x) = u^{(m)}(x)$. High order weak derivatives are defined from lower order derivatives.

7.3.2 Definition of the Sobolev spaces $H^m(\Omega)$

Having defined the first partial derivatives, we define the Sobolev space $H^1(\Omega)$ as

$$H^1(\Omega) = \left\{ v(x), \quad D^{|\alpha|}v \in L^2(\Omega), \quad |\alpha| \leq 1 \right\}. \quad (7.29)$$

For example, $H^1(a, b)$ is defined as

$$H^1(a, b) = \left\{ v(x), \quad a < x < b, \quad \int_a^b v^2 dx < \infty; \quad \int_a^b (v')^2 dx < \infty \right\}.$$

In two space dimensions, $H^1(\Omega)$ is defined as

$$H^1(\Omega) = \left\{ v(x, y), \quad v \in L^2(\Omega), \quad \frac{\partial v}{\partial x} \in L^2(\Omega), \quad \frac{\partial v}{\partial y} \in L^2(\Omega) \right\}.$$

Similarly, the Sobolev space $H^m(\Omega)$ is defined as

$$H^m(\Omega) = \left\{ v(\mathbf{x}), \quad D^{|\alpha|}v \in L^2(\Omega), \quad |\alpha| \leq m \right\}. \quad (7.30)$$

7.3.3 The inner product in H^m spaces

The inner product in $H^0(\Omega)$ is the same as that in $L^2(\Omega)$,

$$(u, v)_{H^0(\Omega)} = (u, v)_0 = \int_{\Omega} uv dx.$$

The inner product in $H^1(a, b)$ is defined as

$$(u, v)_{H^1(a, b)} = (u, v)_1 = \int_a^b (uv + u'v') dx.$$

The inner product in $H^1(\Omega)$ of two variables is defined as

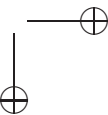
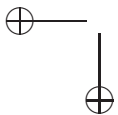
$$(u, v)_{H^1(\Omega)} = (u, v)_1 = \iint_{\Omega} \left(uv + \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \frac{\partial v}{\partial y} \right) dx dy.$$

The inner product in $H^m(\Omega)$ of general dimensions is

$$(u, v)_{H^m(\Omega)} = (u, v)_m = \iint_{\Omega} \sum_{|\alpha| \leq m} (D^{|\alpha|}u(\mathbf{x})) (D^{\alpha}v(\mathbf{x})) d\mathbf{x}. \quad (7.31)$$

The norm in $H^m(\Omega)$ of general dimensions is

$$\|u\|_{H^m(\Omega)} = \|u\|_m = \left\{ \iint_{\Omega} \sum_{|\alpha| \leq m} |D^{\alpha}u(\mathbf{x})|^2 d\mathbf{x} \right\}^{1/2}. \quad (7.32)$$



Therefore $H^m(\Omega)$ is a Hilbert space. A norm can be defined from the inner product. For example, in $H^1(a, b)$, the H^1 norm is defined by

$$\|u\|_1 = \left\{ \int_a^b (u^2 + u'^2) dx \right\}^{1/2}.$$

The *distance* in $H^m(\Omega)$ of general dimensions is then defined as

$$d(u, v)_m = \|u - v\|_m. \quad (7.33)$$

7.3.4 Relations between $C^m(\Omega)$ and $H^m(\Omega)$ –Sobolev embedding theorem

In one dimensional space, we have

$$H^1(\Omega) \subset C^0(\Omega), \quad H^2(\Omega) \subset C^1(\Omega), \quad \dots, \quad H^{1+j}(\Omega) \subset C^j(\Omega).$$

Theorem 7.1. *The Sobolev embedding theorem: If $2m > n$, then*

$$H^{m+j} \subset C^j, \quad j = 0, 1, \dots, \quad (7.34)$$

where n is the dimension of the independent variables of the elements in the Sobolev space.

Example 7.4. *In two dimensional space, we have $n = 2$. The condition $2m > n$ means that $m > 1$. From the embedding theorem, we have*

$$H^{2+j} \subset C^j, \quad j = 0, \longrightarrow H^2 \subset C^0, \quad j = 1, \longrightarrow H^3 \subset C^1, \dots \quad (7.35)$$

If $u(x, y) \in H^2$, which means $u, u_x, u_y, u_{xx}, u_{xy}$, and u_{yy} all belong to L^2 , we can conclude that $u(x, y)$ is continuous, but u_x and u_y may not be continuous!

Example 7.5. *In three dimensional space, we have $n = 3$. The condition $2m > n$ means that $m > 3/2$. We have the same result as that in 2D:*

$$H^{2+j} \subset C^j, \quad j = 0, \longrightarrow H^2 \subset C^0, \quad j = 1, \longrightarrow H^3 \subset C^1, \dots \quad (7.36)$$

We regard the *regularity* of solutions as the degree of smoothness or lack of it of a class of problems measured in C^m or H^m spaces. If $u(x) \in H^m$ or C^m spaces, then the larger m is, the smoother of the function is.

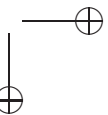
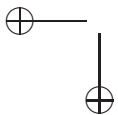
7.4 The FEM analysis for the 1D model problem

For the simple 1D model problem

$$-u'' = f, \quad 0 < x < 1, \quad u(0) = u(1) = 0,$$

we know that the weak form is

$$\int_0^1 u'v' dx = \int_0^1 fvd x \quad \text{or} \quad (u', v') = (f, v).$$



Since v is arbitrary, intuitively, we can take $v = f$ or $v = u$. Then we get

$$\int_0^1 u'v' dx = \int_0^1 u'^2 dx, \quad \int_0^1 fvd x = \int_0^1 f^2 dx.$$

So both u, u', f, v , and v' should belong to $L^2(0, 1)$. That is $u \in H^1(0, 1)$ and $v \in H^1(0, 1)$. So the solution is in the $H^1(0, 1)$ Sobolev space. We should also take v in the same space. From the Sobolev embedding theorem, we also know that $H^1 \subset C^0$, which means the solution is continuous.

7.4.1 Conforming finite element methods

Definition 7.2. *If the finite element space is a subspace of the solution space, then the finite element method is called a conforming finite space.*

For example, the piecewise linear function over a given a triangulation is a conforming finite element space for the model problem. The finite element method is called a *conforming finite element method*. We mainly discuss conforming finite element methods in this book.

If we put the boundary condition together, then we define the solution space as

$$H_0^1(0, 1) = \{v(x), \quad v(0) = v(1) = 0, \quad v \in H^1(0, 1)\}. \quad (7.37)$$

When we look for a finite element solution in a finite dimensional space V_h , V_h should be a subspace of $H_0^1(0, 1)$ for conforming finite element methods. For example, given a mesh for the 1D model problem, we can define a finite dimensional space using piecewise continuous linear functions over the mesh

$$V_h = \{v_h, \quad v_h(0) = v_h(1) = 0, \quad v_h \text{ is continuous piecewise linear}, \quad v_h \in H^1(0, 1)\}.$$

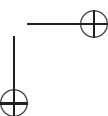
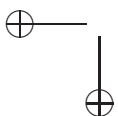
The finite element solution would be chosen from the finite dimensional space V_h , a subspace of $H_0^1(0, 1)$. If the solution of the weak form is in $H_0^1(0, 1)$ but not in V_h space, then an error is introduced as the result of substituting the solution space with the finite dimensional space. Nevertheless, the finite element solution is the best approximation in V_h in some norm as we will see later.

7.4.2 The FEM analysis for the one dimensional Sturm-Liouville problems

A one-dimensional Sturm-Liouville problem with Dirichlet boundary conditions at two ends has the following form,

$$\begin{aligned} -(p(x)u'(x))' + q(x)u(x) &= f(x), \quad x_l < x < x_r \\ u(x_l) &= 0, \quad u(x_r) = 0, \\ p(x) &\geq p_{min} > 0, \quad q(x) \geq q_{min} \geq 0. \end{aligned} \quad (7.38)$$

The condition on $p(x)$ and $q(x)$ is to guarantee the problem is well-posed, that is, the weak form has a unique solution. It is convenient to assume that $p(x) \in C(x_l, x_r)$ and



$q(x) \in C(x_l, x_r)$. Later we will see that with these conditions along with $f(x) \in L^2(x_l, x_r)$, we can guarantee a unique solution to the weak form of the problem. To derive the weak form, we multiply a testing function $v(x)$, $v(x_l) = v(x_r) = 0$ to the both sides of the equations above and integrate from x_l to x_r to get

$$\begin{aligned} \int_{x_l}^{x_r} (-(p(x)u')' + qu) v dx &= \int_{x_l}^{x_r} f v dx \\ -pu'v \Big|_{x_l}^{x_r} + \int_{x_l}^{x_r} (pu'v' + quv) dx &= \int_{x_l}^{x_r} f v dx \\ \rightarrow \int_{x_l}^{x_r} (pu'v' + quv) dx &= \int_{x_l}^{x_r} f v dx, \quad \forall v \in H_0^1(x_l, x_r). \end{aligned}$$

7.4.3 The bilinear form

We define the bilinear form as

$$a(u, v) = \int_{x_l}^{x_r} (pu'v' + quv) dx. \quad (7.39)$$

We call it the bilinear form because it is linear for both u and v , that is

$$\begin{aligned} a(\alpha u + \beta w, v) &= \int_{x_l}^{x_r} (p(\alpha u' + \beta w')v' + q(\alpha u + \beta w)v) dx \\ &= \alpha \int_{x_l}^{x_r} (pu'v' + quv) dx + \beta \int_{x_l}^{x_r} (pw'v' + qwv) dx \\ &= \alpha a(u, v) + \beta a(w, v), \end{aligned}$$

where α and β are scalars. Similarly

$$a(u, \alpha v + \beta w) = \alpha a(u, v) + \beta a(u, w).$$

Note that the bilinear form is an inner product different from the L^2 and H^1 inner products in general. If $p \equiv 1$ and $q \equiv 1$, then

$$a(u, v) = (u, v)_1.$$

Since $a(u, v)$ itself is an inner product, we can define a corresponding norm under the conditions ($p(x) \geq p_{min} > 0$, $q(x) \geq 0$). Such a norm is called the *energy norm*:

$$\|u\|_a = \sqrt{a(u, u)} = \left\{ \int_{x_l}^{x_r} (pu'^2 + qu^2) dx \right\}^{\frac{1}{2}}, \quad (7.40)$$

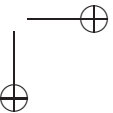
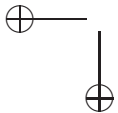
where the first term is called *the kinetic energy*, and the second term is called *the potential energy* in some literature. The Cauchy-Schwartz inequality implies $|a(u, v)| \leq \|u\|_a \|v\|_a$.

Use the bilinear form, the notations are much simplified. The weak form for the Sturm-Liouville problem is

$$a(u, v) = f(v), \quad \forall v \in H_0^1(x_l, x_r). \quad (7.41)$$

The minimization form is

$$\min_{v \in H_0^1(x_l, x_r)} F(v) = \min_{v \in H_0^1(x_l, x_r)} \frac{1}{2} a(v, v) - (f, v). \quad (7.42)$$



Later we will see that all self-adjoint elliptic differential equations have this kind of weak form and minimization from. The finite element method using the Ritz form is the same as the Galerkin form.

7.4.4 The FEM for the 1D S-L problems using piecewise linear basis functions in H^1 .

Given *any* finite dimensional space $V_s \subset H_0^1(x_l, x_r)$ whose basis are

$$\phi_1(x) \in H_0^1, \quad \phi_2(x) \in H_0^1, \quad \dots, \quad \phi_M(x) \in H_0^1(x_l, x_r),$$

that is,

$$\begin{aligned} V_s &= \text{span} \{ \phi_1, \phi_2, \dots, \phi_M \} \\ &= \left\{ v_s, \quad v_s = \sum_{i=1}^M \alpha_i \phi_i \right\} \subset H_0^1(x_l, x_r). \end{aligned}$$

The Galerkin method (using the weak form) is the following. Let the approximate solution be

$$u_s = \sum_{j=1}^M \alpha_j \phi_j. \quad (7.43)$$

The coefficients are chosen such that the weak form is satisfied

$$a(u_s, v_s) = (f, v_s), \quad \forall v_s \in V_h.$$

In other words, we enforce the weak form on V_s not in $H_0^1(a, b)$. This is where the error comes in.

In a finite element method, we enforce the weak form in the finite dimensional space V_s instead of the solution space ($H_0^1(x_l, x_r)$). This can be done by enforcing the weak for the basis functions since *any* element in the space is a linear combination of the basis functions. Therefore we have

$$a(u_s, \phi_i) = (f, \phi_i), \quad i = 1, 2, \dots, M,$$

or

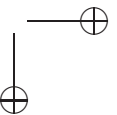
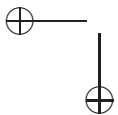
$$a \left(\sum_{j=1}^M \alpha_j \phi_j, \phi_i \right) = (f, \phi_i), \quad i = 1, 2, \dots, M,$$

or

$$\sum_{j=1}^M a(\phi_j, \phi_i) \alpha_j = (f, \phi_i), \quad i = 1, 2, \dots, M.$$

In the matrix-vector form, $AU = F$, the system of equations for the coefficients is

$$\begin{bmatrix} a(\phi_1, \phi_1) & a(\phi_1, \phi_2) & \cdots & a(\phi_1, \phi_M) \\ a(\phi_2, \phi_1) & a(\phi_2, \phi_2) & \cdots & a(\phi_2, \phi_M) \\ \vdots & \vdots & \ddots & \vdots \\ a(\phi_M, \phi_1) & a(\phi_M, \phi_2) & \cdots & a(\phi_M, \phi_M) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_M \end{bmatrix} = \begin{bmatrix} (f, \phi_1) \\ (f, \phi_2) \\ \vdots \\ (f, \phi_M) \end{bmatrix}.$$



The linear system of equations has the following nice properties:

- A is symmetric. That is $\{a_{ij}\} = \{a_{ji}\}$, or $A = A^T$ since $a(\phi_i, \phi_j) = a(\phi_j, \phi_i)$. Note that this is only true for self-adjoint elliptic problems. In one dimensions, the second order ODE has the form

$$-(pu')' + qu = f.$$

The following equation

$$-u'' + xu' = f$$

is not self-adjoint, the weak form for this equation is

$$(u', v') + (u', v) = (f, v), \quad \text{or} \quad (u', v') + (u, v') = (f, v).$$

We still can use the Galerkin finite element method, but we will have terms like (ϕ'_i, ϕ_j) which is different from (ϕ'_j, ϕ_i) . So the coefficient matrix for the none self-adjoint one is not symmetric anymore.

- A is positive definite, that is,
 1. $x^T Ax > 0$, if $x \neq 0$.
 2. All eigenvalues of A are positive.

Proof: For any $\eta \neq 0$, we show that $\eta^T A \eta > 0$.

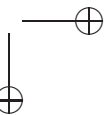
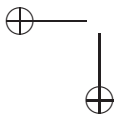
$$\begin{aligned} \eta^T A \eta &= \eta^T (A \eta) = \sum_{i=1}^M \eta_i \sum_{j=1}^M a_{ij} \eta_j \\ &= \sum_{i=1}^M \eta_i \sum_{j=1}^M a(\phi_i, \phi_j) \eta_j \\ &= \sum_{i=1}^M \eta_i \sum_{j=1}^M a(\phi_i, \eta_j \phi_j) \\ &= \sum_{i=1}^M \eta_i a \left(\phi_i, \sum_{j=1}^M \eta_j \phi_j \right) \\ &= a \left(\sum_{i=1}^M \eta_i \phi_i, \sum_{j=1}^M \eta_j \phi_j \right) \\ &= a(v_s, v_s) = \|v_s\|_a^2 > 0 \end{aligned}$$

since $v_s = \sum_{i=1}^M \eta_i \phi_i \neq 0$ because η is a nonzero vector and ϕ_i are linear independent.

7.4.5 The local stiffness matrix and load vector using the hat basis functions

The local stiffness matrix using hat basis function is still 2 by 2 matrix that has the form

$$K_i^e = \begin{bmatrix} \int_{x_i}^{x_{i+1}} p(\phi'_i)^2 dx & \int_{x_i}^{x_{i+1}} p\phi'_i\phi'_{i+1} dx \\ \int_{x_i}^{x_{i+1}} p\phi'_{i+1}\phi'_i dx & \int_{x_i}^{x_{i+1}} p(\phi'_{i+1})^2 dx \end{bmatrix} + \begin{bmatrix} \int_{x_i}^{x_{i+1}} q\phi_i^2 dx & \int_{x_i}^{x_{i+1}} q\phi_i\phi_{i+1} dx \\ \int_{x_i}^{x_{i+1}} q\phi_{i+1}\phi_i dx & \int_{x_i}^{x_{i+1}} q\phi_{i+1}^2 dx \end{bmatrix}$$



and the load vector is

$$F_i^e = \begin{bmatrix} \int_{x_i}^{x_{i+1}} f \phi_i dx \\ \int_{x_i}^{x_{i+1}} f \phi_{i+1} dx \end{bmatrix}.$$

The global stiffness matrix and load vector can be assembled by element by element approach.

7.5 Error analysis for FEM

Error analysis for finite element methods usually include two parts. One is error estimates for a given finite element space. The other one is the convergence analysis (a limiting process) that shows the finite element solution converges the true solution to the weak form in some norms as the mesh size h approaches zero. We first recall some notations and set-ups.

1. We are given a weak form $a(u, v) = L(v)$ and a space V , which usually has infinite dimensions. The problem is to find a u , $u \in V$, such that the weak form is satisfied for any $v \in V$. We call u the solution of the weak form.
2. We denote V_h as a *finite dimensional* and subspace of V , i.e. $V_h \subset V$ (condition for a conforming FEM). It does not have to depend on h though.
3. We denote u_h as the solution of the weak form in the subspace V_h . That is, $a(u_h, v_h) = L(v_h)$ is true for any $v_h \in V_h$.
4. We define $e_h = u(x) - u_h(x)$ as the global error. We want to find a sharp upper bound for $\|e_h\|$ using certain norms.

Note that error arises when we substitute the solution space with a finite dimensional space. In other words, the weak form is satisfied only in the sub-space V_h , not in the solution space V . However, we can prove that the solution that satisfies the weak form in the sub-space V_h is the best approximation to the exact solution u in the finite element space in the energy norm.

Theorem 7.3.

1. u_h is the projection of u onto V_h in terms of energy inner product, that is

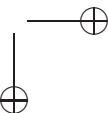
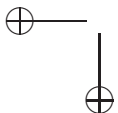
$$u - u_h \perp V_h, \quad \text{or} \quad u - u_h \perp \phi_i, \quad i = 1, 2, \dots, M, \quad (7.44)$$

$$a(u - u_h, v_h) = 0, \quad \forall v_h \in V_h, \quad \text{or} \quad a(u - u_h, \phi_i) = 0, \quad i = 1, 2, \dots, M \quad (7.45)$$

assuming that $\{\phi_i\}$ are the basis functions.

2. Best approximation in the energy norm:

$$\|u - u_h\|_a \leq \|u - v_h\|_a, \quad \forall v_h \in V_h.$$



Proof:

$$\begin{aligned}
 a(u, v) &= (f, v), \quad \forall v \in V, \\
 \rightarrow a(u, v_h) &= (f, v_h), \quad \forall v_h \in V_h \quad \text{since } V_h \subset V, \\
 a(u_h, v_h) &= (f, v_h), \quad \forall v_h \in V_h, \quad \text{since } u_h \text{ is the solution in } V_h, \\
 \text{subtract} \rightarrow a(u - u_h, v_h) &= 0, \quad \text{or } a(e_h, v_h) = 0, \quad \forall v_h \in V_h.
 \end{aligned}$$

Now we prove that u_h is the best approximation in V_h .

$$\begin{aligned}
 \|u - v_h\|_a^2 &= a(u - v_h, u - v_h) \\
 &= a(u - u_h + u_h - v_h, u - u_h + u_h - v_h), \quad \text{Let } w_h = u_h - v_h \in V_h, \\
 &= a(u - u_h + w_h, u - u_h + w_h) \\
 &= a(u - u_h, u - u_h + w_h) + a(w_h, u - u_h + w_h) \\
 &= a(u - u_h, u - u_h) + a(u - u_h, w_h) + a(w_h, u - u_h) + a(w_h, w_h) \\
 &= \|u - u_h\|_a^2 + 0 + 0 + \|w_h\|_a^2, \quad \text{since } a(e_h, u_h) = 0 \\
 &\geq \|u - u_h\|_a^2.
 \end{aligned}$$

That is

$$\|u - u_h\|_a \leq \|u - v_h\|_a.$$

Example: For the Sturm-Liouville problem, we have

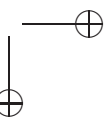
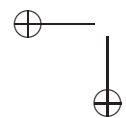
$$\begin{aligned}
 \|u - u_h\|_a^2 &\leq \int_a^b (p(u' - u_h')^2 + q(u - u_h)^2) dx \\
 &\leq p_{max} \int_a^b (u' - u_h')^2 dx + q_{max} \int_a^b (u - u_h)^2 dx \\
 &\leq \max\{p_{max}, q_{max}\} \int_a^b ((u' - u_h')^2 + (u - u_h)^2) dx \\
 &= C \|u - u_h\|_1^2,
 \end{aligned}$$

where $C = \max\{p_{max}, q_{max}\}$. So we obtain

$$\begin{aligned}
 \|u - u_h\|_a &\leq \hat{C} \|u - u_h\|_1, \\
 \|u - u_h\|_a &\leq \|u - v_h\|_a \leq \bar{C} \|u - v_h\|_1.
 \end{aligned}$$

7.5.1 Interpolation functions and error estimates

Usually the solution is unknown, in order to get the error estimate, we need to choose a special $v_h^* \in V_h$, which is very close to the solution. Then we use the error estimate $\|u - u_h\|_a \leq \|u - v_h^*\|_a$. Usually we choose the *piecewise interpolation function*. That is another reason we choose the finite element space as piecewise linear, quadratic, or cubic functions over the given triangulation.



Linear piecewise interpolation function of one dimension

Given a triangulation $x_0, x_1, x_2, \dots, x_M$, the linear piecewise interpolation function is defined as $u_I(x)$:

$$u_I(x) = \frac{x - x_i}{x_{i-1} - x_i} u(x_{i-1}) + \frac{x - x_{i-1}}{x_i - x_{i-1}} u(x_i), \quad x_{i-1} \leq x \leq x_i.$$

It is obvious that $u_I(x) \in V_h$, where $V_h \subset H^1$, the set of continuous piecewise linear functions. Therefore

$$\|u - u_h\|_a \leq \|u - u_I\|_a.$$

Since $u(x)$ is unknown, $u_I(x)$ is unknown as well. However we know the *upper bound* of the interpolation functions.

Theorem 7.4. *Given a function $u(x) \in C^2[a, b]$ and a triangulation $x_0, x_1, x_2, \dots, x_M$. The continuous piecewise linear function u_I has the following error estimates*

$$\|u - u_I\|_\infty = \max_{x \in [a, b]} |u(x) - u_I(x)| \leq \frac{h^2}{8} \|u''\|_\infty, \quad (7.46)$$

$$\|u'(x) - u'_I(x)\|_{L^2(a, b)} \leq h \sqrt{b-a} \|u''\|_\infty. \quad (7.47)$$

Proof: Let $\tilde{e}_h = u(x) - u_I(x)$, then

$$\tilde{e}_h(x_{i-1}) = \tilde{e}_h(x_i) = 0.$$

From the Rolle's theorem, there must be at least one point z_i between x_{i-1} and x_i , that is

$$\tilde{e}_h'(z_i) = 0.$$

Therefore

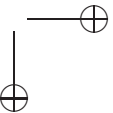
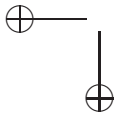
$$\begin{aligned} \tilde{e}_h'(x) &= \int_{z_i}^x \tilde{e}_h''(t) dt \\ &= \int_{z_i}^x (u''(t) - u''_I(t)) dt \\ &= \int_{z_i}^x u''(t) dt. \end{aligned}$$

We obtain the following error estimate

$$|\tilde{e}_h'(x)| \leq \int_{z_i}^x |u''(t)| dt \leq \|u''\|_\infty \int_{z_i}^x dt \leq \|u''\|_\infty h,$$

and

$$\begin{aligned} \|\tilde{e}_h'\|_{L(a, b)} &= \|\tilde{e}_h'\|_0 \leq \left\{ \|u''\|_\infty^2 \int_a^b h^2 dt \right\}^{\frac{1}{2}} \\ &\leq \sqrt{b-a} \|u''\|_\infty h. \end{aligned}$$



Thus we have proved the second equality. To prove the first one, we can assume that $x_{i-1} + h/2 \leq z_i \leq x_i$, otherwise we can use the other half interval. We use the Taylor expansions to get:

$$\begin{aligned}\tilde{e}_h(x) &= \tilde{e}_h(z_i + x - z_i), \quad \text{assume } x_{i-1} \leq x \leq x_i, \\ &= \tilde{e}_h(z_i) + \tilde{e}_h'(z_i)(x - z_i) + \frac{1}{2}\tilde{e}_h''(\xi)(x - z_i)^2, \quad x_{i-1} \leq \xi \leq x_i, \\ &= \tilde{e}_h(z_i) + \frac{1}{2}\tilde{e}_h''(\xi)(x - z_i)^2,\end{aligned}$$

Take $x = x_i$, we have

$$\begin{aligned}0 &= \tilde{e}_h(x_i) = \tilde{e}_h(z_i) + \frac{1}{2}\tilde{e}_h''(\xi)(x_i - z_i)^2, \\ \tilde{e}_h(z_i) &= -\frac{1}{2}\tilde{e}_h''(\xi)(x_i - z_i)^2, \\ |\tilde{e}_h(z_i)| &\leq \frac{1}{2}\|u''\|_\infty(x - z_i)^2 \leq \frac{h^2}{8}\|u''\|_\infty.\end{aligned}$$

Note that the largest value of $\tilde{e}_h(x)$ has to be one of z_i 's where the derivative is zero.

7.5.2 Error estimates of the finite element methods using the interpolation function.

For one dimensional Sturm-Liouville problem, we have

Theorem 7.5.

$$\|u - u_h\|_a \leq Ch\|u''\|_\infty \quad (7.48)$$

$$\|u - u_h\|_1 \leq \hat{C}h\|u''\|_\infty, \quad (7.49)$$

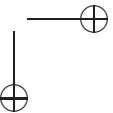
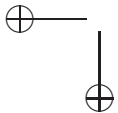
where C and \hat{C} are two constants.

Proof:

$$\begin{aligned}\|u - u_h\|_a^2 &\leq \|u - u_I\|_a^2 \\ &\leq \int_a^b (p(u' - u_I')^2 + q(u - u_I)^2) dx \\ &\leq \max\{p_{max}, q_{max}\} \int_a^b ((u' - u_I')^2 + (u - u_I)^2) dx \\ &\leq \max\{p_{max}, q_{max}\} \|u''\|_\infty^2 \int_a^b (h^2 + h^4/64) dx \\ &\leq Ch^2\end{aligned}$$

The second inequality is obtained because $\| \cdot \|_a$ and $\| \cdot \|_1$ are equivalent meaning that

$$c\|v\|_a \leq \|v\|_1 \leq C\|v\|_a, \quad \hat{c}\|v\|_1 \leq \|v\|_a \leq \hat{C}\|v\|_1.$$



7.5.3 Error estimates in pointwise norm

For one dimensional Sturm-Liouville problem, we can easily prove the following (over-estimate though).

Theorem 7.6.

$$\|u - u_h\|_\infty \leq Ch \|u''\|_\infty \quad (7.50)$$

$$\|u' - u'_h\|_\infty \quad \text{Does not make sense} \quad (7.51)$$

where C is a constants. Since u'_h is discontinuous at nodal points, the infinity norm makes no sense because it can only be defined for continuous functions.

Proof:

$$\begin{aligned} e_h(x) &= u(x) - u_h(x) = \int_a^x e'_h(t) dt \\ |e_h(x)| &\leq \int_a^b |e'_h(t)| dt \\ &\leq \left\{ \int_a^b |e'_h|^2 dt \right\}^{1/2} \left\{ \int_a^b 1 dt \right\}^{1/2} \\ &\leq \sqrt{b-a} \left\{ \int_a^b \frac{p}{p_{min}} |e'_h|^2 dt \right\}^{1/2} \\ &\leq \frac{\sqrt{b-a}}{p_{min}} \|e_h\|_a \\ &\leq \frac{\sqrt{b-a}}{p_{min}} \|\tilde{e}_h\|_a \\ &\leq Ch \|u''\|_\infty. \end{aligned}$$

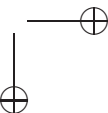
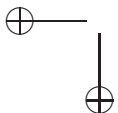
Remark 7.1. Actually we can prove a better inequality:

$$\|u - u_h\|_\infty \leq Ch^2 \|u''\|_\infty.$$

The finite element method is second order accurate.

7.6 Exercises

1. Take $n = 3$, the number of variables, describe the Sobolev space $H^3(\Omega)$, i.e. $m = 3$, in terms of $L^2(\Omega)$ using all the terms but not the multi-index notation. Also explain the inner product, the norm, the Schwartz inequality, the distance, and the Sobolev embedding theorem in this space using the multi-index notation when applicable.
2. Consider the function $v(x) = |x|^\alpha$ on $\Omega = (-1, 1)$ with $\alpha \in \mathcal{R}$. For what values of α is $v \in H^0(\Omega)$? (Consider negative α as well). For what values is $v \in H^1(\Omega)$? in $H^m(\Omega)$? For what values of α is $v \in C^m(\Omega)$?



Hint: Generally

$$|x|^\alpha = \begin{cases} x^\alpha & \text{if } x \geq 0 \\ (-x)^\alpha & \text{if } x < 0. \end{cases}$$

However, if $\alpha = 2k$, where k is a non-negative integer, then

$$|x|^\alpha = \begin{cases} x^{2k} & \text{if } \alpha = 2k, k > 0, \\ 1 & \text{if } \alpha = 0. \end{cases}$$

Also

$$\lim_{x \rightarrow 0} |x|^\alpha = \begin{cases} 0 & \text{if } \alpha > 0 \\ 1 & \text{if } \alpha = 0 \\ \infty & \text{if } \alpha < 0, \end{cases} \quad \int_{-1}^1 |x|^\alpha dx = \begin{cases} \frac{2}{\alpha + 1} & \text{if } \alpha > -1, \\ \infty & \text{if } \alpha \leq -1. \end{cases}$$

3. Are each of the following statements true or false? Justify your answers.

- If $u \in H^2(0, 1)$ then u' and u'' are both continuous functions.
- If $u(x, y) \in H^2(\Omega)$, then $u(x, y)$ may not have continuous partial derivatives $\frac{\partial u}{\partial x}$ and $\frac{\partial u}{\partial y}$? Does $u(x, y)$ have first and second order *weak* derivatives? Is $u(x, y)$ continuous in Ω ?

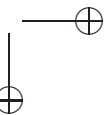
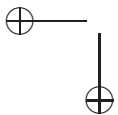
4. Consider the Sturm-Liouville problem:

$$\begin{aligned} -(p(x)u(x)')' + q(x)u(x) &= f(x), \quad 0 < x < \pi, \\ \alpha u(0) + \beta u'(0) &= \gamma, \quad u'(\pi) = u_b, \end{aligned}$$

where

$$0 < p_{\min} \leq p(x) \leq p_{\max} < \infty, \quad 0 \leq q_{\min} \leq q(x) \leq q_{\max} < q\infty.$$

- Derive the weak form for the problem. Define a bilinear form $a(u, v)$ and a linear form $L(v)$ to simplify the weak form. What is the energy norm?
- What kind of restrictions should we have for α , β , and γ in order that the weak form has a solution?
- Determine the space where the solution resides under the weak form.
- If we look for a finite element solution in a finite dimensional space V_h using a conforming finite element method, should V_h be a subspace of C^0 , C^1 , C^2 ?
- Given a triangulation $x_0 = 0 < x_1 < x_2 \cdots < x_{M-1} < x_M = \pi$, if the finite dimensional space is generated by the hat functions, what kind of structure do the local and global stiffness matrix and the load vector have? Is the resulting linear system of equations formed by the global stiffness matrix and the load vector symmetric, positive definite, and banded?



5. Extra Credit: Consider the two-point boundary value problem

$$-u''(x) = f(x), \quad a < x < b, \quad u(a) = u(b) = 0.$$

Let $u_h(x)$ be the finite element solution using the piecewise linear space (in $H_0^1(a, b)$) spanned by a mesh $\{x_i\}$. Show that

$$\|u - u_h\|_\infty \leq Ch^2,$$

where C is a constant. **Hint:** First show that $\|u_h - u_I\|_a = 0$, where $u_I(x)$ is the interpolation function in V_h .

