A brief introduction to finite element methods

1. Two-point boundary value problem and the variational formulation

1.1. The model problem. Consider the two-point boundary value problem:

Given a constant \( a \geq 0 \) and a function \( f(x) \), find \( u(x) \) such that

\[
\begin{aligned}
-u'' + au &= f(x), \quad 0 < x < 1, \\
u(0) &= 0, \quad u'(1) = 0.
\end{aligned}
\]  

(0.1)

1.2. The variational formulation. If \( u \) is the solution to (0.1) and \( v(x) \) is any (sufficiently regular) function such that \( v(0) = 0 \), then integration by parts yields

\[
\int_0^1 -u''v \, dx + \int_0^1 auv \, dx = -u'(1)v(1) + u'(0)v(0) + \int_0^1 u'(x)v'(x) \, dx + \int_0^1 auv \, dx
\]

Let us introduce the bilinear form

\[
A(u, v) = \int_0^1 (u'v' + auv) \, dx,
\]

and define

\[
V = \{ v \in L^2([0, 1]) : A(v, v) < \infty \text{ and } v(0) = 0 \}.
\]

Then we can see that the solution \( u \) to (0.1) is characterized by

\[
u \in V \text{ such that } A(u, v) = \int_0^1 f(x)v(x) \, dx \quad \forall v \in V.
\]  

(0.2)

which is called the variational or weak formulation of (0.1).

**Theorem 1.1.** Suppose \( f \in C^0([0, 1]) \) and \( u \in C^2([0, 1]) \) satisfies (0.2). Then \( u \) solves (0.1).

**Proof.** Let \( v \in V \cap C^1([0, 1]) \). Then integration by parts gives

\[
\int_0^1 f v \, dx = A(u, v) = \int_0^1 -u''v \, dx + \int_0^1 auv \, dx + u'(1)v(1).
\]  

(0.3)

Thus, \( \int_0^1 (f + u'' - au) v \, dx = 0 \) for all \( v \in V \cap C^1([0, 1]) \) such that \( v(1) = 0 \). Let \( w = f + u'' - au \in C^0([0, 1]) \). If \( w \neq 0 \), then \( w(x) \) is of one sign in some interval \([b, c]\) \( \subset [0, 1] \), with \( b < c \). Choose \( v(x) = (x - b)^2(x - c)^2 \) in \([b, c]\) and \( v \equiv 0 \) outside.
2. A BRIEF INTRODUCTION TO FINITE ELEMENT METHODS

[0, c]. But then \( \int_0^1 wv \, dx \neq 0 \) which is a contradiction. Thus \(-u'' + au = f\). Now apply (0.3) with \( v(x) = x \) to find \( u'(1) = 0 \). So \( u \) solves (0.1). \( \square \)

We remark that the boundary condition \( u(0) = 0 \) is called \textit{essential} as it appears in the variational formulation explicitly, i.e., in the definition of \( V \). This type of boundary condition is also called “Dirichlet” boundary condition. The boundary condition \( u'(0) = 0 \) is called \textit{natural} as it is incorporated implicitly. This type of boundary condition is often referred to by the name “Neumann”.

2. The finite element method

2.1. Meshes. Let \( M_h \) be a partition of \([0, 1]\):

\[
0 = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = 1.
\]

The points \( \{x_i\} \) are called \textit{nodes}. Let \( h_i = x_i - x_{i-1} \) be the length of the \( i \)-th subinterval \([x_{i-1}, x_i]\). Define \( h = \max_{1 \leq i \leq n} h_i \).

2.2. Finite element spaces. We shall approximate the solution \( u(x) \) by using the continuous piecewise linear functions over \( M_h \). Introduce the linear space of functions

\[
V_h = \left\{ v \in C^0([0, 1]) : v(0) = 0, \quad v|_{[x_{i-1}, x_i]} \text{ is a linear polynomial, } i = 1, \cdots, n \right\}. \quad (0.4)
\]

It is clear that \( V_h \subset V \).

2.3. The finite element method. The finite element discretization of (0.2) reads as:

Find \( u_h \in V_h \) such that

\[
A(u_h, v_h) = \int_0^1 f(x)v_h(x) \, dx \quad \forall v_h \in V_h. \quad (0.5)
\]

2.4. A nodal basis. For \( i = 1, \cdots, n \), define \( \phi_i \in V_h \) by the requirement that \( \phi_i(x_j) = \delta_{ij} \) = the Kronecker delta, as shown in Fig. 1.

\[
\phi_i(x) \quad \phi_n(x)
\]

\[
\begin{array}{cccc}
0 & x_i & x_{i+1} & 1 \\
0 & x_i & x_{i+1} & 1 \\
\hline
\hline
0 & x_i & x_{i+1} & 1 \\
0 & x_i & x_{i+1} & 1 \\
\end{array}
\]

\[\text{Figure 1. piecewise linear basis function } \phi_i.\]
For any \( v \in V_h \), let \( v_i \) be the value of \( v \) at the node \( x_i \), i.e.,

\[
v_i = v_h(x_i), \quad i = 1, 2, \ldots, n,
\]

then

\[
v_h = v_1 \phi_1(x) + v_2 \phi_2(x) + \cdots + v_n \phi_n(x).
\]

### 2.5. The finite element equations.

Let

\[
u_h = u_1 \phi_1 + u_2 \phi_2 + \cdots + u_n \phi_n, \quad u_1, \ldots, u_n \in \mathbb{R},
\]

where \( u_i = u_h(x_i) \).

Let \( v_h = \phi_i, \quad i = 1, \ldots, n \) in (0.5), then we obtain an algebraic linear system in unknowns \( u_1, u_2, \ldots, u_n \):

\[
A(\phi_1, \phi_1)u_1 + A(\phi_2, \phi_1)u_2 + \cdots + A(\phi_n, \phi_1)u_n = \int_0^1 f(x)\phi_1 \, dx, \\
i = 1, \ldots, n.
\]

(0.6)

Denote by

\[
k_{ij} = A(\phi_j, \phi_i) = \int_0^1 \phi_j' \phi_i' + a \phi_j \phi_i \, dx, \quad f_i = \int_0^1 f(x)\phi_i \, dx,
\]

and

\[
K = (k_{ij})_{n \times n}, \quad F = (f_i)_{n \times 1}, \quad U = (u_i)_{n \times 1},
\]

then (0.6) can be rewritten as:

\[
KU = F
\]

(0.7)

Here \( K \) is called the stiffness matrix.

It is clear that \( k_{ij} = 0 \) if \( x_i \) and \( x_j \) are not adjacent to each other. Therefore \( K \) is sparse.

Next we compute \( k_{ij} = A(\phi_j, \phi_i) \). Note that

\[
k_{ij} = A(\phi_j, \phi_i) = \sum_{l=1}^n \int_{x_{l-1}}^{x_l} \phi_j' \phi_i' + a \phi_j \phi_i \, dx.
\]

On the interval \([x_{l-1}, x_l]\), the following integrals are involved.

\[
k_{11}^l := \int_{x_{l-1}}^{x_l} \phi_{l-1}' \phi_{l-1} + a \phi_{l-1} \phi_{l-1} \, dx, \quad k_{21}^l := \int_{x_{l-1}}^{x_l} \phi_{l-1}' \phi_l + a \phi_{l-1} \phi_l \, dx,
\]

\[
k_{12}^l := \int_{x_{l-1}}^{x_l} \phi_{l-1}' \phi_l + a \phi_{l-1} \phi_l \, dx, \quad k_{22}^l := \int_{x_{l-1}}^{x_l} \phi_l' \phi_l + a \phi_l \phi_l \, dx.
\]

Some simple calculations yield

\[
k_{11}^l = k_{22}^l = \frac{1}{h_l} + \frac{a}{3} h_l, \quad k_{12}^l = k_{21}^l = -\frac{1}{h_l} + \frac{a}{6} h_l.
\]
The matrix $K^i := (k_{ij})_{2 \times 2}$ is called the element stiffness matrix on $[x_{i-1}, x_i]$. The stiffness matrix $K$ can be assembled by using the element stiffness matrices as follows.

$$k_{ii} = \begin{cases} k_{22}^i + k_{11}^{i+1} = \frac{1}{h_i} + \frac{1}{h_{i+1}} + \frac{a}{3} (h_i + h_{i+1}), & i = 1, \ldots, n - 1, \\ k_{22}^n = \frac{1}{h_n} + \frac{a}{3} h_n, & i = n, \\ k_{i-1} = k_{i, i-1} = k_{12} = \frac{1}{h_i} + \frac{a}{6} h_i, & i = 2, \ldots, n. \end{cases}$$

Combining the above equations and (0.6) yields

$$\begin{cases} \left[ \frac{a(h_i + h_j)}{3} + \frac{1}{h_i} + \frac{1}{h_j} \right] u_1 + \left( \frac{ah_i}{6} - \frac{h_j}{h_i} \right) u_2 = f_1, \\ \left( \frac{ah_i}{6} - \frac{1}{h_i} \right) u_{i-1} + \left[ \frac{a(h_i + h_{i+1})}{3} + \frac{1}{h_i} + \frac{1}{h_{i+1}} \right] u_i + \left( \frac{ah_{i+1}}{6} - \frac{1}{h_{i+1}} \right) u_{i+1} = f_i & i = 2, \ldots, n - 1, \\ \left( \frac{ah_n}{6} - \frac{1}{h_n} \right) u_{n-1} + \left( \frac{ah_n}{3} + \frac{1}{h_n} \right) u_n = f_n. \end{cases}$$

### 2.6. The interpolant.

Given $u \in C^0([0, 1])$, the interpolant $u_I \in V_h$ of $u$ is determined by

$$u_I = \sum_{i=1}^n u(x_i) \phi_i.$$ 

It is clear that $u_I(x_i) = u(x_i), i = 0, 1, \ldots, n$, and

$$u_I(x) = \frac{x_i - x}{h_i} u(x_{i-1}) + \frac{x - x_{i-1}}{h_i} u(x_i) \quad \text{for } x \in [x_{i-1}, x_i].$$

Denote by $\tau_i = [x_{i-1}, x_i]$ and by $\|g\|_{L^2(\tau_i)} = \left( \int_{x_{i-1}}^{x_i} g^2 \, dx \right)^{1/2}$.

**Theorem 2.1.**

\[
\begin{align*}
\|u - u_I\|_{L^2(\tau_i)} & \leq \frac{1}{\pi^2} h_i \|u'\|_{L^2(\tau_i)} , \\
\|u - u_I\|_{L^2(\tau_i)} & \leq \frac{1}{\pi^2} h_i^2 \|u''\|_{L^2(\tau_i)} , \\
\|u' - u_I'\|_{L^2(\tau_i)} & \leq \frac{1}{\pi} h_i \|u''\|_{L^2(\tau_i)} .
\end{align*}
\]  

**Proof.** We only prove (0.8) and leave the others as an exercise. We first change (0.5) to the reference interval $[0, 1]$. Let $\tilde{x} = (x - x_{i-1}) / h_i$ and let $\hat{u}(\tilde{x}) = u(x) - u_I(x)$. Note that $\hat{u}(0) = \hat{u}'(1) = 0$ and $k = u'_I$ is a constant. The inequality (0.5) is equivalent to

$$\|\hat{u}\|_{L^2([0, 1])} = \frac{1}{h_i} \|u - u_I\|_{L^2(\tau_i)} \leq \frac{1}{\pi^2} \|u'\|_{L^2(\tau_i)} = \frac{1}{\pi^2} \|\hat{u}' + kh_i\|_{L^2([0, 1])} ,$$

that is

$$\|\hat{u}\|_{L^2([0, 1])} \leq \frac{1}{\pi^2} \|\hat{u}'\|_{L^2([0, 1])} + \frac{1}{\pi^2} \|kh_i\|_{L^2([0, 1])} .$$

Since $\hat{u}(0) = \hat{u}'(1) = 0$, it follows from the Fourier expansion $\hat{u}(\tilde{x}) = \sum_{n=1}^{\infty} a_n \sin n\pi\tilde{x}$ and the Parseval’s identity that $\|\hat{u}\|_{L^2([0, 1])} \leq \frac{1}{\pi^2} \|\hat{u}'\|_{L^2([0, 1])}$, which implies (0.11). This completes the proof of (0.8). □
2.7. **A priori error estimate.** Introduce the energy norm
\[ \|v\| = A(v, v)^{1/2}. \]
From the Cauchy inequality,
\[ A(u, v) \leq \|u\|\|v\|. \]
By taking \( v = v_h \in V_h \) in (0.2) and subtracting it from (0.5), we have the following fundamental orthogonality
\[ A(u - u_h, v_h) = 0 \quad \forall v_h \in V_h. \] (0.12)
Therefore
\[ \|u - u_h\|^2 = A(u - u_h, u - u_h) = A(u - u_h, u - u_I) \leq \|u - u_h\|\|u - u_I\|, \]
It follows from Theorem 2.1 that
\[ \|u - u_h\| \leq \|u - u_I\| = \left[ \sum_{i=1}^{n} \left( \|u' - u'_I\|^2_{L^2(\tau_i)} + a \|u''\|^2_{L^2(\tau_i)} \right) \right]^{1/2} \]
\[ \leq \frac{h}{\pi} \left[ 1 + a \left( \frac{h}{\pi} \right)^2 \int_0^1 (u'')^2 \, dx \right]^{1/2}. \]
We have proved the following error estimate.

**Theorem 2.2.**
\[ \|u - u_h\| \leq \frac{h}{\pi} \left( 1 + a \left( \frac{h}{\pi} \right)^2 \right)^{1/2} \|u''\|_{L^2([0,1])}. \]
Since the above estimate depends on the unknown solution \( u \), it is called the a priori error estimate.

2.8. **A posteriori error estimates.** We will derive error estimates independent of the unknown solution \( u \).
Let \( e = u - u_h \). Then
\[ A(e, e) = A(u - u_h, e - e_I) \]
\[ = \int_0^1 f \cdot (e - e_I) \, dx - \int_0^1 u'_h(e - e_I)' \, dx - \int_0^1 au_h(e - e_I) \, dx \]
\[ = \int_0^1 (f - au_h)(e - e_I) \, dx - \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} u'_h(e - e_I)' \, dx \]
Since \( u'_h \) is constant on each interval \((x_{i-1}, x_i)\),
\[ A(e, e) = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} (f - au_h)(e - e_I) \, dx \]
\[ \leq \sum_{i=1}^{n} \|f - au_h\|_{L^2(\tau_i)} \|e - e_I\|_{L^2(\tau_i)} \]
\[ \leq \sum_{i=1}^{n} \frac{h_i}{\pi} \|f - au_h\|_{L^2(\tau_i)} \|e'\|_{L^2(\tau_i)}. \]
Here we have used Theorem 2.3 to derive the last inequality.

Define the local error estimator on the element \( \tau_i = [x_{i-1}, x_i] \) as follows

\[
\eta_i = \frac{1}{\pi} h_i \| f - au_h \|_{L^2(\tau_i)}. \tag{0.13}
\]

Then

\[
\| e \|^2 \leq \left( \sum_{i=1}^{n} \eta_i^2 \right)^{1/2} \| e' \|^2 \leq \left( \sum_{i=1}^{n} \eta_i^2 \right)^{1/2} \| e \|.
\]

That is, we have the following a posteriori error estimate.

**Theorem 2.3** (Upper bound).

\[
\| u - u_h \| \leq \left( \sum_{i=1}^{n} \eta_i^2 \right)^{1/2}. \tag{0.14}
\]

Now a question is if the above upper bound overestimates the true error. To answer this question we introduce the following theorem that gives a lower bound of the true error.

**Theorem 2.4** (Lower bound). Define \( \| \phi \|_{\tau_i} = \left( \int_{x_{i-1}}^{x_i} ((\phi')^2 + a\phi^2) \, dx \right)^{1/2} \). Let \((f - au_h)_i = \frac{1}{h_i} \int_{x_{i-1}}^{x_i} (f - au_h) \, dx \) and \( osc_i = \frac{1}{h_i} \| f - au_h - (f - au_h)_i \|_{L^2(\tau_i)} \). Then

\[
\eta_i - \frac{5 + \sqrt{30}}{5} osc_i \leq \left( \frac{60 + 6ah_i^2}{5\pi^2} \right)^{1/2} \| u - u_h \|_{\tau_i}. \tag{0.15}
\]

As a consequence

\[
\| u - u_h \| \geq \left( \sum_{i=1}^{n} \frac{5\pi^2}{60 + 6ah_i^2} \left( \max \left( \eta_i - \frac{5 + \sqrt{30}}{5} osc_i, 0 \right) \right)^2 \right)^{1/2}. \tag{0.16}
\]

**Proof.** For simplicity, denote by \( R = f - au_h \) and \( R_i = (f - au_h)_i \). For any \( \psi \in V \), it is clear that

\[
A(e, \psi) = \int_0^1 R\psi \, dx - \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} u_h' \psi' \, dx \tag{0.17}
\]

Define \( \psi_i(x) = \phi_{i-1}(x) \phi_i(x) \) if \( x \in \tau_i \) and \( \psi_i(x) = 0 \) otherwise. Choose \( \psi = \psi_i R_i \).

By simple calculations, we obtain

\[
\int_0^1 R_i \psi \, dx = \frac{1}{6} \| R_i \|_{L^2(\tau_i)}^2, \quad \sqrt{30} \| \psi \|_{L^2(\tau_i)} = \sqrt{30} h_i \| \psi' \|_{L^2(\tau_i)} = \| R_i \|_{L^2(\tau_i)}.
\]

From (0.17),

\[
A(e, \psi) = \int_0^1 R\psi \, dx = \int_{x_{i-1}}^{x_i} (R - R_i)\psi \, dx + \frac{1}{6} \| R_i \|_{L^2(\tau_i)}^2.
\]

We have,

\[
\| R_i \|_{L^2(\tau_i)}^2 \leq 6 \| e \|_{\tau_i} \| \psi \|_{\tau_i} + 6\pi osc_i \| \psi \|_{L^2(\tau_i)}
\]

\[
= \left( \left( 12 + \frac{6ah_i^2}{5} \right)^{1/2} \| e \|_{\tau_i} + \frac{\pi \sqrt{30}}{5} osc_i \right) h_i^{-1} \| R_i \|_{L^2(\tau_i)},
\]

which implies

\[
h_i \| R_i \|_{L^2(\tau_i)} \leq \left( 12 + \frac{6ah_i^2}{5} \right)^{1/2} \| e \|_{\tau_i} + \frac{\pi \sqrt{30}}{5} osc_i.
\]
3. EXERCISES

Now the proof is completed by using $\eta_i \leq \frac{1}{\pi} h_i \| R_i \|_{L^2(\tau_i)} + \text{osc}_i$. □

We remark that the term $\text{osc}_i$ is of high order compared to $\eta_i$ if $f$ and $a$ are smooth enough on $\tau_i$.

**Example 2.5.** We solve the following problem by the linear finite element method.

\[- u'' + 10000u = 1, \quad 0 < x < 1, \]
\[u(0) = u(1) = 0.\]

The true solution (see Fig. 2) is

\[u = \frac{1}{10000} \left( 1 - \frac{e^{100x} + e^{100(1-x)}}{1 + e^{100}} \right).\]

If we use the uniform mesh obtained by dividing the interval $[0, 1]$ into 1008 subintervals of equal length, then the error $\| u - u_h \| \approx 2.86 \times 10^{-5}$. On the other hand, if we use a non-uniform mesh as shown in Fig. 2 which also contains 1008 subintervals, then the error $\| u - u_h \| \approx 1.59 \times 10^{-6}$ is smaller than that obtained by using the uniform mesh.

![Figure 2. Example 2.5. The finite element solution and the mesh.](image)

3. Exercises

**Exercise 0.1.** Prove (0.9) and (0.10).

**Exercise 0.2.** Use Example 2.5 to verify numerically the a posteriori error estimates in Theorem 2.3 and (0.16) in Theorem 2.4.