Lecture notes: Intro to Finite Element Methods

A. C. Quillen

July 31, 2025

Contents

1 Quick introduction to the ideas of finite element methods!

I am mostly following the zeroth chapter of the book *The Mathematical Theory of Finite Element Methods* by Susanne C. Brenner and L. Ridgway Scott.

Many books and software packages on finite element methods start with the following problem: Find $u(\mathbf{x})$ such that

$$\Delta u = f \qquad \text{on } \Omega \tag{1}$$

where $u()$ is a function on some domain in $\mathbb{R}^n$ that is called Ω. There would also be a boundary condition. This would be a condition on u or its derivative on the boundary of

Ω which is called $\partial\Omega$. Here Δ is the Laplacian operator which in 2 dimensions is $\partial_x^2 + \partial_y^2$ where $\partial_x = \frac{\partial}{\partial x}$.

To place this generic example context we will start with a 1 dimensional example, with domain Ω equal the unit interval, $x \in [0, 1]$. The boundary of the unit interval consists of two points $x = 0$ and $x = 1$. Our example is

$$-\frac{d^2u}{d^2x} = f \text{ on } (0, 1) \tag{2}$$

$$u(0) = 0,\ u'(1) = 0 \tag{3}$$

where $u' = \frac{du}{dx}$. Equations 3 are the boundary conditions. The goal is to find $u(x)$ given function $f(x)$.

In the conventional theory of partial differential equations (PDE)s, we would use a set of orthogonal functions (such as sine functions $\sin(kx)$ for suitably chosen values of k), and would write the solution in terms of a sum of coefficients times these functions. The solution could be divided into a sum of homogeneous and inhomogeneous pieces. The goal would be to solve for coefficients of the homogeneous terms.

With a **finite element method**, instead we work with an **integral form** of the PDE. We multiply equation 2 with a test function $v(x) \in V$ and integrate over Ω;

$$-\int_0^1 u''(x)v(x)\ dx = \int_0^1 f(x)v(x)\ dx. \tag{4}$$

Here $u'' = \frac{d^2u}{dx^2}$. The solution for u should satisfy equation 4 for all $v \in V$, with a nice space V to be determined.

Integrating by parts is frequently done in finite element methods.

$$\frac{d}{dx}\left[u'v\right] = u''v + u'v'$$

$$u''v = \frac{d}{dx}\left[u'v\right] - u'v'. \tag{5}$$

We replace $u''v$ in equation 4 using equation 5

$$-\int_0^1 \frac{d}{dx}\left[u'v\right] dx + \int_0^1 u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx$$

$$-u'(1)v(1) + u'(0)v(0) + \int_0^1 u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx. \tag{6}$$

We use the boundary condition $u'(1) = 0$ and set a requirement on the test function v. We require v to obey $v(0) = 0$. Both boundary terms vanish giving

$$\int_0^1 u'(x)v'(x)dx = \int_0^1 f(x)dx. \tag{7}$$

Notice that equation 2 we chose to put a minus sign on the second derivative. However, because we integrated by parts equation 7 lacks a minus sign. I have noticed that there is ambiguity in the sign chosen for an operator that is called a Laplacian.

Let

$$(f, v) = \int_0^1 f(x)v(x)dx \tag{8}$$

$$a(u, v) = \int_0^1 u'(x)v'(x)dx. \tag{9}$$

Notice that $a(u, v) = a(v, u)$ is symmetric. Using the definitions in equation 9, equation 7 can be written

$$(f, v) = a(u, v). \tag{10}$$

So far we did not put any conditions on the space V other than functions in it should satisfy $v(0) = 0$.

Let $v \in V$ with V

$$V = \left\{ v \in L^2(0, 1) : \quad a(v, v) < \infty, v(0) = 0 \right\}. \tag{11}$$

The space L^p is a space of measurable functions for which the pth power of the the absolute value of the function is Lebesgue integrable. In the above equation the requirement that $v \in L^2$ means that $v \in V$ should have a finite Lebesgue integral $\int_0^1 dx|v(x)|^2$. The requirement that $a(v, v)$ is finite means that v'^2 also has a finite Lebesgue integral.

The solution u to equations 2, 3 should be a function $u \in V$ such that

$$(f, v) = a(u, v) \qquad \text{for all } v \in V \tag{12}$$

where V is the space given in equation 11. Equation 12 (equivalent to equation 7) is called the **weak form** of the problem given in equations 2, 3 and it is called **variational** because the test function v is allowed to vary.

1.1 Boundary conditions

Notice that the boundary condition $u(0) = 0$ led us to restrict the space V so that $v \in V$ satisfies $v(0) = 0$ (see equation 11). This type of boundary condition is called **Dirichlet** in the context of PDEs and in the context of weak or variational formulation it is called **essential** as it appears in the variational formulation explicitly, i.e., in the definition of the space V for test and trial solutions.

The boundary condition $u'(1) = 0$ is called a **Neumann** boundary condition. In the context of the variational formulation, it is called **natural** because it is incorporated implicitly. Neumann boundary conditions don't affect the space V for trial and test solutions but can give additional terms in the weak formulation.

1.2 Ritz-Galerkin Approximation

In a finite element method, we approximate the solution. Let

$$S \subset V \tag{13}$$

be a finite dimensional subspace of V. Recall V contains functions on the unit interval that are zero at $x = 0$. We take function u_S

$$u_S \in S \text{ such that } a(u_S, v) = (f, v) \quad \forall v \in S. \tag{14}$$

The function u_S is not the solution of equation 2 but it might be close to it.

We create a **basis**[1] for functions on S, the functions $\{\phi_1, \phi_2, \ldots, \phi_n\}$ where n is the dimension of S.

Using this basis, we can write a function $v \in S$ in terms of the basis $v = \sum_i V_i \phi_i$. We could consider the list $\mathbf{V} = (V_1, V_2, \ldots, V_n)$ to be a vector.

Using the function f, (given in the original problem of equation 2) we compute the coefficients

$$F_i = (f, \phi_i) = \int_0^1 f(x)\phi_i(x)dx. \tag{15}$$

The coefficients $\mathbf{F} = (F_1, F_2, \ldots, F_n)$ also form a vector!

We create a matrix $\mathbf{K}$, called the **stiffness** matrix, with coefficients $K_{ij} = a(\phi_i, \phi_j)$.

Suppose we find a function $\tilde{u} = \sum_i U_i \phi_i \in S$ with coefficients U_i, giving a vector $\mathbf{U} = (U_1, U_2, \ldots, U_n)$, that satisfies the matrix equation

$$\mathbf{KU} = \mathbf{F}. \tag{16}$$

Choose index j of the above matrix equation

$$\begin{aligned}
\sum_i K_{ji} U_i &= F_j \text{ index form of matrix equation} \\
\sum_i K_{ji} U_i &= (f, \phi_j) \text{ definition of } \mathbf{F} \\
\sum_i a(\phi_j, \phi_i) U_i &= (f, \phi_j) \text{ definition of } \mathbf{K} \\
a(\phi_j, \tilde{u}) &= (f, \phi_j) \text{ linearity of } a() \text{ and definition of } \tilde{u} \\
a(\tilde{u}, \phi_j) &= (f, \phi_j) \quad a \text{ is symmetric}
\end{aligned}$$

[1] A basis is a set of elements in a vector space V that are *linearly independent* and *span* the vector space. That means every element $v \in V$ can be written uniquely in terms of sum $v = \sum_i a_i \phi_i$ where a_i are in the field, and ϕ_i are basis elements. Linear independence means there is no non-zero combination of coefficients $\{a_i\}$ such that $\sum_i a_i \phi_i = 0$.

This is satisfied for all j and because both sides are linear, for any $v \in S$. Hence $\tilde{u} = u_S$ is a solution of equation 14.

We find that solving equation 14 for $u_S \in S$ is equivalent to solving for a vector $\mathbf{U}$ that satisfies

$$\mathbf{KU} = \mathbf{F}. \tag{17}$$

Note that we did not require an inner product on the finite dimensional vector space S though we did require a basis for this finite dimensional vector space. In the above example we have been careful to specify when we have a function and when we have something that we can write as a vector or matrix. While $F(x) = \sum_i F_i \phi_i(x)$ is a function, the function $F(x)$ is not equal to $f(x)$.

Key to the finite element approach is that functions are approximated by vectors in a finite dimensional vector space[2]. Thus solving a PDE is equivalent to solving a linear matrix equation. The matrix is likely to be sparse and so the number of required computations is not necessarily large. Hence finite element methods are potentially both accurate and computationally efficient.

Theorem 1.1. *The solution to equation 14, if it exists, is unique.*

Proof. Suppose $u_1, u_2 \in S$ are solutions to equation 14. We define vectors $\mathbf{U}_1, \mathbf{U}_2$ using the basis functions $\{\phi_i\}$. These vectors must satisfy $\mathbf{KU}_1 = \mathbf{KU}_2 = \mathbf{F}$. Hence there is a nonzero vector $\mathbf{W} = \mathbf{U}_1 - \mathbf{U}_2$ that satisfies $\mathbf{KW} = 0$. Let $w = \sum_i W_i \phi_i$.

$$\begin{aligned}
\sum_j K_{ij} W_j &= 0 \text{ index form of matrix equation } \mathbf{KW} = 0 \\
\sum_j a(\phi_i, \phi_j) W_j &= 0 \text{ definition of } \mathbf{K} \\
a(\phi_i, \sum_j W_j \phi_j) &= 0 \text{ linearity of } a \\
a(\phi_i, w) &= 0 \text{ definition of } w \\
\sum_i W_i a(\phi_i, w) &= 0 \text{ multiply by } W_i \text{ and sum} \\
a(w, w) &= 0 \text{ via linearity of } a \text{ and definition of } w \\
\int_0^1 [w'(x)]^2 dx &= 0 \text{ via definition of } a.
\end{aligned}$$

This implies that $w(x)$ is a constant. However the zero boundary condition at $x = 0$ implies that the constant must be zero and this contradicts our assumption that $\mathbf{W} \neq 0$. Hence if a solution to equation 14 exists, it is unique. □

[2] A *vector space* over a field F is a space that contains elements that are in the form $v = \sum_i a_i v_i$ where coefficient $a_i \in F$ are in the field and v_i are in a particular set.

Note that going from $a(w,w) = 0$ to $w(x) = 0$ (the first implies the other) requires constraints on the nature of V w.r.t. Lebesgue integration. Our definition of V should rule out pathological functions like Cantor sets. The space V should be restricted to be a **Sobolev** space[3], as we will discuss later.

Theorem 1.2. *The function $a(u,v)$ for $u, v \in S$ gives an* **inner product** *on S.*

Proof. An inner product should satisfy the following:

An inner product (on a vector space over the real numbers) must be symmetric. Based on its definition $a(u,v) = a(v,u)$ and is symmetric. So a satisfies this condition.

An inner product should be linear in the first argument. Based on its definition a also satisfies this condition.

An inner product should be **positive definite**. In other words, We require that $a(v,v) \geq 0$ with $a(v,v) = 0$ only if $v = 0$. We compute

$$a(v,v) = \int_0^1 dx[v'(x)]^2 \geq 0 \tag{18}$$

which must be positive. It's only zero if w is constant. For non-zero w this possibility is ruled out because of the boundary condition at $x = 0$. Hence $a()$ is positive definite and also satisfies this condition.

These three conditions (that are required for an inner product), are met by a so $a(u,v)$ is an inner product on the vector space S. □

As a consequence the matrix $\mathbf{K}$ is both symmetric and positive definite. This means that it is invertible. We can multiply $\mathbf{KU} = \mathbf{F}$ by $\mathbf{K}^{-1}$ to find $\mathbf{U} = \mathbf{K}^{-1}\mathbf{F}$. As long as $\mathbf{F} \neq 0$, the matrix equation must have a non-zero solution. Hence a solution both exists and is unique.

1.3 An example of a polynomial basis on the unit interval

We illustrate an example finite subspace $S \subset V$ and a basis for it. We take

$$0 = x_0 < x_1 < x_2 < \ldots < x_n = 1$$

to be a set of $n+1$ points in the unit interval. The ordered set of positions $\{x_i\}$ can be called a *partition* of the unit interval.

We can define the subspace $S \subset V$ to be the set of functions v

(i) $v \in C^0[0,1]$, the space of continuous functions on the unit interval.

(ii) In each interval $[x_{i-1}, x_i]$ with $i \in \{1, \ldots, n\}$, the function $v(x)$ is a linear polynomial. In other words, it is a line segment as a line segment can be written in the form of a linear polynomial; $a + bx$ with appropriately chosen coefficients a, b.

[3]Lebesque integrals of a specific derivative is are finite

(iii) $v(0) = 0$, so that S is consistent with our Dirichlet boundary condition at $x = 0$.

The set of points $\{x_i\}$ are called **nodes**. One of the basis functions is shown in Figure 1.

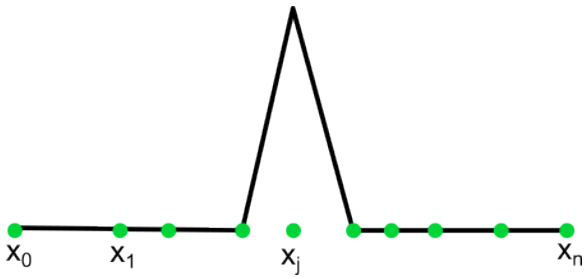

Figure 1: An example of the basis function ϕ_j is shown with the black line. The vertical height of the peak is 1. The green dots show the nodes $\{x_0, x_1, \ldots, x_n\}$ on the unit interval.

We give a basis $\{\phi_1, \ldots, \phi_n\}$. Let $\phi_i(x)$ be a piece wise linear function that satisfies

$$\phi_i(x_j) = \delta_{ij} \tag{19}$$

This condition implies that $\phi_i(x)$ is zero at all nodes except one where $\phi_i(x_i) = 1$.

The set of basis functions $\{\phi_i\}$ is linearly independent. A function $v \in S$ can be written as $v = \sum_i c_i \phi_i$. Each coefficient is uniquely set by the value of $v(x_i)$ at node x_i. The condition $\sum c_i \phi_i(x_j) = 0$ implies that $c_j = 0$. Since there is no way to write any particular ϕ_i in terms of the basis functions $\phi_j, j \neq i$, the basis is linearly independent.

An example of a function in S is shown in Figure 2.

Note that even though the set $\{\phi_i\}$ form a basis for S, they are not orthogonal when integrated:

$$\int_0^1 \phi_i \phi_{i+1} dx \neq 0 \qquad \text{and} \qquad \int_0^1 \phi_i \phi_j dx \neq \delta_{ij}. \tag{20}$$

The setting is different than when conventionally solving a boundary value PDE problem where you might be expanding in terms of a set of orthogonal functions, which are orthogonal when integrated across the domain.

1.4 The interpolant

Definition Consider $v \in C^0([0,1])$ and $v \notin S$. The **interpolant** of v is the function $v_I \in S$

$$v_I(x) = \sum_{i=1}^{n} v(x_i)\phi_i(x) \tag{21}$$

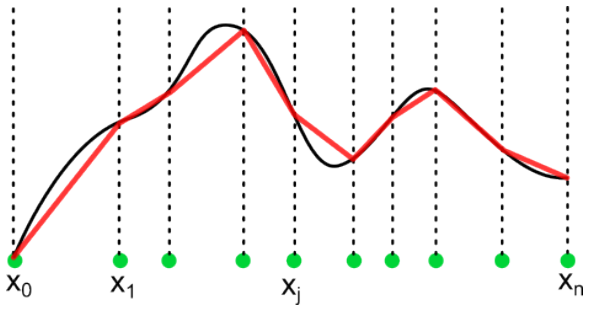

Figure 2: An example of a function $v \in C^0[0,1]$ is shown with a black line. The interpolant $v_I \in S$ of v is shown with the thicker red line.

The indexing starts at 1 so that $v_I(0) = 0$ and obeys the Dirichlet boundary condition.

See Figure 2 for an illustration of a function $v \in C^0([0,1])$ and its interpolant v_I. It should be clear from the figure that v_I is comprised of linear segments. Functions in S look like the red line shown in Figure 2.

Furthermore $v \in S$ implies that $v = v_I$.

The interpolant is a map from functions in $C^0([1,0])$ to functions in S. If we call this map $\mathcal{I}$, for $v \in C^0([1,0])$, $\mathcal{I}(v) = v_I$ with v_I equal to the interpolant of v. The map $\mathcal{I}$ is a projection as $\mathcal{I}(\mathcal{I}(v)) = \mathcal{I}(v) = v_I \ \forall v$.

Suppose we have a function in S, that we call v_I. What functions $v \in V$ have v_I as their interpolant? There are many functions in V or in $C^0([1,0])$ that have the same interpolant v_I. The space $v \in V$ such that $\mathcal{I}(v) = v_I$ is the space of functions that have the same values of v_I at all the points in the nodal set $\{x_i\}$. Moreover none of these functions are all that far away from v_I.

1.4.1 The Cauchy-Schwartz inequality

The Cauchy-Schwartz inequality states that states that for all vectors $\mathbf{u}, \mathbf{v}$ of an inner product space

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle} \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}. \tag{22}$$

On the left is an absolute value. Equivalently with norm $\|\mathbf{u}\| = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$,

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \|\mathbf{v}\|. \tag{23}$$

An example! Consider the vector dot product on the real plane (which is an inner product). We compute the dot product of two vectors $\mathbf{u}, \mathbf{v}$

$$\mathbf{u} \cdot \mathbf{v} = |\mathbf{u}||\mathbf{v}| \cos\theta \tag{24}$$

where θ is the angle between the two vectors. Because $|\cos\theta| \leq 1$

$$|\mathbf{u} \cdot \mathbf{v}| \leq |\mathbf{u}||\mathbf{v}| \tag{25}$$

which is equivalent to the Cauchy-Schwartz inequality.

1.5 Error estimates

We use two different norms[4] to quantitatively estimate distances between functions. For $v \in V$, the $L^2(0,1)$ norm

$$\|v\| = \sqrt{(v,v)} \equiv \sqrt{\int_0^1 dx|v(x)|^2}. \tag{26}$$

Because $a(v,v)$ is an inner product (theorem 1.2) we also have the *energy* norm

$$\|v\|_E \equiv \sqrt{a(v,v)}. \tag{27}$$

A relationship between an inner-product and two norms is Schwarz' inequality:

$$|a(v,w)| \leq \|v\|_E \, \|w\|_E \quad \forall v, w \in V. \tag{28}$$

We compare two solutions u, u_S,

$$\begin{aligned} u \in V \text{ such that } a(u,v) = (f,v) \qquad & \forall v \in V \\ u_S \in S \text{ such that } a(u_S,w) = (f,w) \qquad & \forall w \in S \end{aligned} \tag{29}$$

Since $S \subset V$, we can subtract these two equations to find that

$$a(u - u_S, w) = 0 \qquad \forall w \in S. \tag{30}$$

This implies that the true solution u is pretty close to the approximate solution u_S found in the discrete subspace S.

Theorem 1.3. *For $u \in V$ a solution to $a(u,w') = f(u,w')\ \forall w' \in V$ and $u_S \in S$ a solution to $a(u_S,w) = f(u,w)\ \forall w \in S$*

$$\|u - u_S\|_E = \min\{\|u - v\|_E \ : \ v \in S\}. \tag{31}$$

[4] A norm (denoted $|.|$ in this definition) is a function on a vector space giving a real number $\to \mathbb{R}$ that is positive definite ($|u| \geq 0$, and only equal to 0 if $u = 0$), obeys the triangle inequality ($|u+v| \geq |u| + |v|$), and obeys $|su| = |s||u|$ for scalars s.

Proof.

$$\begin{aligned}
\|u-u_S\|_E^2 &= a(u-u_S, u-u_S) \\
&= a(u-u_S, u-u_S+v-v)\ \forall v \in S \\
&= a(u-u_S, u-v) + a(u-u_S, v-u_S) \qquad \text{by linearity of } a() \\
&= a(u-u_S, u-v) \qquad \text{right term vanishes due to equation 30} \\
&\leq \|u-u_S\|_E\ \|u-v\|_E \qquad \text{Schwarz inequality}
\end{aligned}$$

Ignoring the boring possibility that $u = u_S$, this means that

$$\|u-u_S\|_E \leq \|u-v\|_E\ \forall v \in S \tag{32}$$

Since $u_S \in S$, the minimum value of the expression on the right must equal to that on the left. Hence

$$\|u-u_S\|_E\ = \min_{v\in S} \|u-v\|_E\ . \tag{33}$$

If $u = u_S$, then there is a $v = u_S \in S$ that is the solution and the inequality is still satisfied. □

We now attempt to place a constraint on $\|u-u_S\|$ using the L^2 norm instead of the norm $\|\|_E$ generated with $a()$.

Theorem 1.4. *There is a small* ϵ

$$\|u-u_S\| \leq \epsilon \|u-u_S\|_E \leq \epsilon^2 \|u''\|\ = \epsilon^2 \|f\|. \tag{34}$$

We notice that $\|u-u_S\|_E$ is of order ϵ whereas $\|u-u_S\|$ is even smaller, of order ϵ^2.

Proof. We define $w \in V$ such that $-w'' = u - u_S$. We require that $w'(1) = 0$ (satisfying the right boundary condition of our problem 3). If $w \in V$ it must satisfy $w(0) = 0$.

$$\begin{aligned}
\|u-u_S\|^2 &= \int_0^1 dx |u-u_S|^2 = (u-u_S, u-u_S) \\
&= (u-u_S, -w'') \qquad \text{definition of } w'' \\
&= \int_0^1 (u-u_S)(-w'')dx = \int_0^1 (u'-u_S')w'dx \qquad \text{integrate by parts} \\
&\qquad \text{and use boundary conditions for } w \\
&= a(u-u_S, w) \qquad \text{via definition of } a() \\
&= a(u-u_S, w-v) \text{ using equation 30 and } \forall v \in S \\
&\leq \|u-u_S\|_E\ \|w-v\|_E \qquad \text{Schwartz inequality}
\end{aligned}$$

$$\begin{aligned}\|u - u_S\| &\leq \|u - u_S\|_E \, \|w - v\|_E \, \frac{1}{\|u - u_S\|} \\ &\leq \|u - u_S\|_E \, \|w - v\|_E \, \frac{1}{\|w''\|} \qquad \forall v \in S. \end{aligned} \tag{35}$$

Taking the infimum[5] of the set on the right hand side of equation 35, equation 35 can be written as

$$\|u - u_S\| \leq \|u - u_S\|_E \, \inf_{v \in S} \|w - v\|_E \, \frac{1}{\|w''\|}. \tag{36}$$

We assume that there will be a $v \in S$ that is close to w, letting $\inf_{v \in S} \|w - v\|_E / \|w''\|$ be small. We make an assumption that there is a small ϵ giving

$$\inf_{v \in S} \|w - v\|_E \leq \epsilon \|w''\| \text{ assumption.} \tag{37}$$

Inserting this assumption into equation 36 gives

$$\|u - u_S\| \leq \epsilon \|u - u_S\|_E \, . \tag{38}$$

If we can approximate w with a $v \in S$ then we can probably also approximate u with another function in S. Hence our assumption of equation 37 also gives

$$\inf_{v \in S} \|u - v\|_E \leq \epsilon \|u''\| \text{ assumption.} \tag{39}$$

Using theorem 1.3 equation 38 becomes

$$\begin{aligned}\|u - u_S\| &\leq \epsilon \min_{v \in S} \|u - v\|_E \\ &\leq \epsilon^2 \|u''\| \qquad \text{using equation 39} \end{aligned} \tag{40}$$

Putting together equations 38 and 40

$$\|u - u_S\|_E \leq \epsilon \|u - u_S\|_E \leq \epsilon^2 \|u''\| \tag{41}$$

which is what we wanted to show.

□

This proof relied on the our ability to find functions in the finite dimensional space S that could well approximate functions in V. Looking again at Figure 2 we see that the interpolant $u_I \in S$ of a function u is a good approximation for the function u. We now work on showing that using the basis that we constructed in section 1.3 and with sufficient numbers of finite elements we can improve the approximation to an arbitrary level of precision (and show that our assumptions of equations 37 and 39 can be true).

[5]Infimum of x over a set S is the greatest element less than or equal to x in the set S.

1.6 Galerkin approximation

Galerkin methods are those that convert a continuous operator problem, such as a differential equation, commonly in a weak formulation, to a discrete problem by applying linear operators that are constructed from finite sets of basis functions.

Suppose we solve the problem on a finite element space and we have an approximate solution which is in a discrete subspace S. How far away could the solution be from the actual solution? We first ask the question: how far away can any function $v \in V$ be from its interpolant $v_I \in S$. The idea being that when we find a solution in the subspace S we may have actually found the interpolant of the actual solution. (Recall that many functions can have the same interpolant!).

Theorem 1.5. *Let $h = \max_{1 \le i \le n} |x_i - x_{i-1}|$ (the maximum distance between the nodal points). Then*

$$\|u - u_I\|_E \ \le Ch\|u''\| \tag{42}$$

for all $u \in V$, where constant C is independent of h and u.

This gives a limit on how far any function can be from its interpolant. Notice that the distance depends on the maximum element size. This makes sense as if you subdivide the elements in your finite element space, you expect a more accurate solution. If we set $\epsilon = Ch$, equation 42 implies that the assumptions we made in equations 37 and 39 are valid.

Proof. Recall the definition of the norm from equation 27 and setting an error function $e(x) = u(x) - u_I(x)$ for any $u \in V$ and u_I its interpolant. The definition of the interpolant implies that the error $e(x_i) = 0$ at all nodes, where the set of nodes is $\{x_i\}$.

$$\|u - u_I\|_E^2 \ = a(e, e) = \int_0^1 dx \ e'(x)^2.$$

We can consider each interval $[x_i, x_{i+1}]$ separately. Inside the interval u_I' is linear, so $u_I'' = 0$. Consequently $e'' = u''$ within each interval.

$$e'(x) = \int_{x_i}^{x} dx \ e''(x) = \int_{x_i}^{x} dx \ u''(x) \qquad \text{for } x \in [x_i, x_{i+1}]. \tag{43}$$

The right hand side can be written as a product of two functions

$$\begin{aligned}
\int_{x_i}^{x} dx\ u''(x) &= \int_{x_i}^{x} dx\ 1 \times u''(x) \\
&= (1, u'') \qquad L_2 \text{ inner product over interval } [x_i, x] \\
\left| \int_{x_i}^{x} dx\ u''(x) \right| &\leq \sqrt{\int_{x_i}^{x} dx} \sqrt{\int_{x_i}^{x} dx\ [u''(x)]^2} \qquad \text{Schwarz inequality} \\
&\leq h^{1/2} \sqrt{\int_{x_i}^{x} dx\ [u''(x)]^2} \ \text{ for } x \in [x_i, x_{i+1}].
\end{aligned}$$

Using equation 43

$$\begin{aligned}
|e'(x)|^2 &\leq h \int_{x_i}^{x} dx\ [u''(x)]^2 \ \text{ for } x \in [x_i, x_{i+1}] \\
&\leq h \int_{x_i}^{x_{i+1}} dx\ [u''(x)]^2 \ \text{ for } x \in [x_i, x_{i+1}]. \qquad (44)
\end{aligned}$$

Integrate both sides over the entire interval. Integrating the right side is like a sum of each piece multiplied by its interval width.

$$\int [e'(x)]^2 dx \leq h^2 \|u''\|^2. \qquad (45)$$

This is super rough (and might not be 100% correct), but is approximately consistent with what we wanted to show with coefficient $C \approx 1$.

□

1.7 Connection to a finite differencing numerical approach

Many numerical integrations for PDEs work on an evenly spaced grid. What does the operator $a(\phi_i, \phi_j)$ look like if we use the first order polynomial basis (shown in Figure 1) on an evenly spaced grid?

We take h to be the grid spacing.

For basis function ϕ_i, the function is equal to 1 at x_i and is linear and increasing between x_{i-1} and x_i and linear and decreasing between x_i and x_{i+1}. The distance between the grid points is h, so

$$\phi_i(x) = \begin{cases} 0 & \text{for } x < x_i - h \\ 1 + \frac{x - x_i}{h} & \text{for } x_i - h < x < x_i \\ 1 - \frac{x - x_i}{h} & \text{for } x_i < x < x_i + h \\ 0 & \text{for } x > x_i = h \end{cases}. \qquad (46)$$

We compute $\phi_i'(x)$

$$\phi_i'(x) = \begin{cases} 0 & \text{for } x < x_i - h \\ \frac{1}{h} & \text{for } x_i - h < x < x_i \\ -\frac{1}{h} & \text{for } x_i < x < x_i + h \\ 0 & \text{for } x > x_i = h \end{cases}. \tag{47}$$

We compute

$$\begin{aligned} a(\phi_i, \phi_i) &= \int_0^1 dx\ [\phi_i'(x)]^2 \\ &= \int_{x_i-h}^{x_i+h} dx\ \frac{1}{h^2} = \frac{2}{h}. \end{aligned} \tag{48}$$

$$a(\phi_i, \phi_{i+1}) = \int_{x_i}^{x_i+h} dx\ \left(-\frac{1}{h} \times \frac{1}{h}\right) = -\frac{1}{h} \tag{49}$$

$$a(\phi_i, \phi_{i-1}) = \int_{x_i-h}^{x_i} dx\ \left(-\frac{1}{h} \times \frac{1}{h}\right) = -\frac{1}{h} \tag{50}$$

As long as i, j are not near the end points, the matrix

$$K_{ij} = a(\phi_i, \phi_j) = \frac{2}{h}\delta_{ij} - \frac{1}{h}\left(\delta_{i,j+1} + \delta_{i,j-1}\right) \tag{51}$$

contains the factor $2/h$ on the diagonal and $-1/h$ on the off diagonals (that are right next to the diagonal).

$$\begin{aligned} F_j = (f, \phi_j) &= \int_0^1 dx\ f(x)\phi_j(x) \\ &= \int_{x_j-h}^{x_j} dx\ f(x)\left(1 + \frac{x - x_j}{h}\right) + \int_{x_j}^{x^j+h} dx\ f(x)\left(1 - \frac{x - x_j}{h}\right) \end{aligned} \tag{52}$$

We use a Taylor expansion for $f(x)$ near $f(x_j)$

$$f(x) = f(x_j) + f'(x_j)(x - x_j) + \frac{1}{2}f''(x_j)(x - x_j)^2 + \dots \tag{53}$$

$$\begin{aligned} F_j = &\int_{x_{j-1}}^{x^j} dx\ \left(f(x_j) + f'(x_j)(x - x_j) + f''(x_j)(x - x_j)^2\right)\left(1 + \frac{x - x_j}{h}\right) \\ &+ \int_{x_j}^{x^{j+1}} dx\ \left(f(x_j) + f'(x_j)(x - x_j) + f''(x_j)(x - x_j)^2\right)\left(1 - \frac{x - x_j}{h}\right) \end{aligned} \tag{54}$$

Let $y = x - x_j$

$$F_j = hf(x_j) + \int_{-h}^{0} dy\ f'(x_j)\frac{y^2}{h} - \int_{0}^{h} dy\ f'(x_j)\frac{y^2}{h} + \ldots$$
$$= hf(x_j) + \mathcal{O}(h^2). \tag{55}$$

The equation $\mathbf{KU} = \mathbf{F}$ becomes

$$\frac{-U_{j+1} + 2U_j - U_{j-1}}{h^2} = F_j \tag{56}$$

which can be recognized as $-u'' = f$.

We now show this! We expand u in a Taylor series at positions x_j,

$$u(x) = u(x_j) + u'(x_j)(x - x_j) + \frac{1}{2}u''(x_j)(x - x_j)^2 + \mathcal{O}(h^3). \tag{57}$$

Using the Taylor series,

$$u(x_j + h) = u(x_{j+1}) = U_{j+1} = u(x_j) + u'(x_j)h + u''(x_j)\frac{h^2}{2} + \mathcal{O}(h^3)$$
$$u(x_j - h) = u(x_{j-1}) = U_{j-1} = u(x_j) - u'(x_j)h + u''(x_j)\frac{h^2}{2} + \mathcal{O}(h^3). \tag{58}$$

Add these together

$$U_{j+1} + U_{j-1} - 2U_j = u''(x_j)h^2 + \mathcal{O}(h^3) \tag{59}$$

Hence

$$u''(x_j) \sim \frac{U_{j+1} + U_{j-1} - 2U_j}{h^2} + \mathcal{O}(h). \tag{60}$$

2 Schemes for updating time dependent problems

2.1 An explicit Euler method

Consider a first order ordinary differential equation which we can write in terms of operator $\mathbf{L}$,

$$\partial_t u + \mathbf{L}(u) = 0 \tag{61}$$

We approximate the time derivative to first order in time-step Δt,

$$\frac{u^{n+1} - u^n}{\Delta t} + \mathbf{L}(u^n) = 0. \tag{62}$$

Here u^n is the value of u at the n-th time-step. We want to find the value of u at the next or $n+1$ time-step. Notice that the operator is applied to u^n. When the operator is applied at the current time, the scheme is *explicit*. We now solve for u^{n+1};

$$u^{n+1} = (I + \Delta t\mathbf{L})u^n. \tag{63}$$

This is a simple and straightforward scheme. For the advection equation, it is unstable.

2.2 An implicit Euler method

We modify equation 62 so that the operator is applied to u^{n+1} not u^n.

$$\frac{u^{n+1} - u^n}{\Delta t} + \mathbf{L}(u^{n+1}) = 0. \tag{64}$$

We solve for u^{n+1}

$$(I + \Delta t\mathbf{L})u^{n+1} = u^n \tag{65}$$

$$u^{n+1} = (I + \Delta t\mathbf{L})^{-1}u^n. \tag{66}$$

Here I is the identity operator. Since we need the inverse of an operator or a matrix, this technique is more difficult to apply but it tends to be more stable.

Suppose instead of updating the field u each time-step, we only want to add a small quantity to the current field. We start again with equation 65 and subtract $(I + \Delta t\mathbf{L})u^n$ from both sides

$$\begin{aligned}(I + \Delta t\mathbf{L})u^{n+1} &= u^n \\ I + \Delta t\mathbf{L})(u^{n+1} - u^n) &= u^n - (I + \Delta t\mathbf{L})u^n = -\Delta t\mathbf{L}u^n \\ u^{n+1} - u^n &= (I + \Delta t\mathbf{L})^{-1}(-\Delta t\mathbf{L}u^n). \end{aligned} \tag{67}$$

2.3 The Crank-Nicolson method

We modify equation 62 so that the operator is applied to an average of u^{n+1} and u^n.

$$\frac{u^{n+1} - u^n}{\Delta t} + \frac{1}{2}\mathbf{L}(u^{n+1} + u^n) = 0. \tag{68}$$

We solve for u^{n+1}

$$\left(I + \frac{1}{2}\Delta t\mathbf{L}\right)u^{n+1} = \left(I - \frac{1}{2}\Delta t\mathbf{L}\right)u^n \tag{69}$$

$$u^{n+1} = \left(I + \frac{1}{2}\Delta t\mathbf{L}\right)^{-1}\left(I - \frac{1}{2}\Delta t\mathbf{L}\right)u^n. \tag{70}$$

For diffusive problems, this scheme is unconditionally stable.

2.4 An implicit Euler method in finite element methods

If the system is a finite element system, in place of the identity we have the mass operator.

$$M(u, v) = \int_\Omega dx\ uv \tag{71}$$

where u, v are test and trial functions. Equation 66 becomes

$$u^{n+1} = (M + \Delta t \mathbf{L})^{-1} u^n \tag{72}$$

and equation 67 becomes

$$u^{n+1} - u^n = (M + \Delta t \mathbf{L})^{-1} (-\Delta t \mathbf{L} u^n) \tag{73}$$

We define

$$M_* = M + \Delta t \mathbf{L} \tag{74}$$

and the difference $\delta u = u^{n+1} - u^n$. The difference in equation 73 can be written as

$$\delta u = M_*^{-1} (-\Delta t \mathbf{L} u^n). \tag{75}$$

2.5 IMEX which stands for semi-implicit Euler method or IMplicit Explicit methods

A semi-implicit method uses the $n+1$ timestep in some of the operators (which is like an implicit method), the n time-step in other operators (which is like an explicity method). The time derivative $\partial_t u$ is approximated as $\frac{u^{n+1} - u^n}{\Delta t}$ which is first order in time.

3 The finite element

Definition A finite element has

(i) A closed subset $\mathcal{K} \subset \mathbb{R}^n$ with nonempty interior and a piecewise smooth connected boundary.

(ii) A finite dimensional space $\mathcal{P}$ of functions of $\mathcal{K}$ called *shape functions*.

(iii) A basis $\mathcal{N} = \{\mathcal{N}_1, \mathcal{N}_\in ....\}$ for $\mathcal{P}'$ which is the set of *nodal variables*.

$\{\mathcal{K}, \mathcal{P}, \mathcal{N}\}$ is called a **finite element**.

This is Ciarlet's definition of a finite element (Ciarlet 1978).