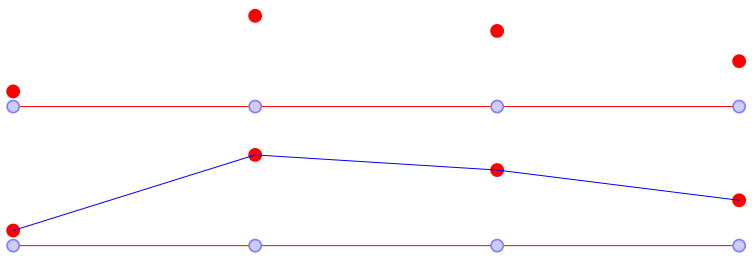
1

A brief introduction of finite element methods

1.1 Motivation and plans

Finite difference method can only be used on a rectangular domain or a domain which can be transformed to a rectangle such as a disk via polar coordinates. For complicated problems in real applications, finite element method is the most successful approach due to its rich theories and flexibility with geometries. In this notes, we mainly focus on rectangular domains, on which finite difference method is the most convenient choice. On the other hand, the finite element method on a rectangular domain can be implemented as a finite difference method when integrals are replaced with quadrature. For instance, there is no essential difference between grid point values and piece-wise linear polynomials represented by its grid point values. Even on unstructured triangular meshes, a linear polynomial on a triangle can be represented by its point values on three vertices of the triangle, which is often called *nodal representation*.



The conventional approach of constructing a finite difference method that we have seen in Chapter **??** includes two crucial steps: first, develop a consistent discretization or approximation to the differential operator and the boundary conditions then try to establish the stability, i.e., try to show $\|A^{-1}\| \leq C$ if the matrix-vector form of the scheme is $A\mathbf{u} = \mathbf{f}$. While the

first step is a relative easy task, the second step seems to be fine with the second order centered difference because we have eigenvalues and eigenvectors, which is however nearly impossible to find for higher order accurate schemes and more general problems such as variable coefficient problems. In general there are quite a few drawbacks and challenges in such a traditional approach. The key issues include:

- It is not elegant or convenient to design a high order accurate scheme to use Taylor expansion (for instance, do you actually enjoy solving Problem **??**). It also becomes harder to deal with boundary conditions in high order schemes.

- Stability is hard to establish in general: estimating the inverse of a matrix is always hard. If using only linear algebra, singular values and eigenvalues are impossible to estimate for more general schemes (think about a high order accurate scheme for $-(a(x)u')' = f$).

Remark 1.1. *9-point discrete Laplacian is successful, but only for Laplacian operator on uniform meshes with Dirichlet boundary conditions.*

Moreover, there are practical concerns:

- Loss of accuracy on non-uniform meshes: if the local truncation error is obtained by Taylor expansion on uniform grids, the proof of order of accuracy breaks down regardless of whether the actual scheme is still as accurate as on uniform grids or not.

- Loss of symmetry in the matrix A: the matrix in general is not symmetric and hard to symmetrize (think about $-\nabla(a(x,y)\nabla u) = f$ with Neumann boundary conditions).

Remark 1.2. *One of the main reasons why a symmetric A is much better is for purely Neumann boundary conditions. The exact solution is not unique for purely Neumann b.c.. So A in the numerical scheme $A\mathbf{u} = \mathbf{f}$ is not invertible thus the linear system $A\mathbf{u} = \mathbf{f}$ may not have a solution (unless $\mathbf{f}$ happens to lie in the column space of A, which is usually not true). So as we have seen in Section **??**, one would have to instead consider $A\mathbf{u} = \bar{\mathbf{f}}$ where $\bar{\mathbf{f}}$ is the projection of $\mathbf{f}$ onto the column space of A. The left null vector $\mathbf{v}$ (i.e.. $\mathbf{v}^T A = \mathbf{0}$) is needed for such a projection. If A is symmetric, then the left null vector is also the right null vector, which is usually $\begin{bmatrix}1 & 1 & \cdots & 1\end{bmatrix}^T$ since A approximates a differential operator. If A is not symmetric, then one would have to solve an eigenvalue problem $A^T\mathbf{v} = 0 * \mathbf{v}$ which is even more expensive (at least 2-3 times more expensive) than solving $A\mathbf{u} = \bar{\mathbf{f}}$. The difficulty of using a non-symmetric matrix for purely Neumann boundary conditions will also be explained in Section 1.9.*

Remark 1.3. *For solving $A\mathbf{u} = \bar{\mathbf{f}}$ as above, it is mathematically equivalent to solve the least square solution by solving $A^T A\mathbf{u} = A^T\mathbf{f}$, which is however a lot harder to solve numerically, because the condition number of $A^T A$ will be nearly the square of the condition number of A.*

All these concerns and difficulties can be solved by using finite element method! What is even better is that finite element method on rectangular meshes (or regular triangular meshes) looks like exactly a finite difference method. In this chapter, we will first see how a finite element method is defined then implement it as a finite difference method.

Caution to readers: this is a very brief introduction to the finite element method because

- We will give up certain math rigor such as complete definition of distribution and Sobolev spaces, proof of existence and uniqueness of variational formulation and important estimates. Instead they will be given and stated as facts.
- We focus mainly on rectangular domains and rectangular meshes.

Despite of these simplifications in mind, you will still learn and understand the key ingredients of the finite element method.

1.2 Preliminaries

1.2.1 Weak derivatives and Sobolev spaces

Let $C_0^\infty(\mathbb{R})$ be the set of all infinitely differentiable functions which are nonzero only on a finite interval.

If a function $f(x)$ is differentible, then after integration by parts, for any smooth function $v(x) \in C_0^\infty(\mathbb{R})$, we have

$$\int_{-\infty}^{+\infty} f(x)v'(x)dx = -\int_{-\infty}^{+\infty} f'(x)v(x)dx. \tag{1.1}$$

The function $f(x) = |x|$ is not differentiable but we can define its *weak or generalized derivative* as the step function $g(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases}$ in the sense of (1.1):

Definition 1.1. *A function $g(x)$ is defined to be the weak or generalized derivative of $f(x)$ if it satisfies*

$$\int_{-\infty}^{+\infty} f(x)v'(x)dx = -\int_{-\infty}^{+\infty} g(x)v(x)dx, \quad \forall v(x) \in C_0^\infty(\mathbb{R}).$$

Example 1.1. *It is straightforward to verify that the step function is the weak derivative of the absolute value function.*

If a function is differentible, then its weak derivative is simply the derivative. **From now on, in this Chapter, derivatives are understood as generalized derivatives.**

Next we need to define a few spaces:

- $$L^2([0,1]) = \left\{ f(x) : \int_0^1 f(x)^2 dx < \infty \right\}.$$

  For a general domain Ω, $L^2(\Omega)$ is similarly defined. Here integral is the Lebesgue integral if you know what it means. The $L^2(\Omega)$-norm will be denoted as

  $$\|f\|_{0,\Omega} = \|f\|_{L^2(\Omega)} = \left(\int_\Omega f(x)^2 dx \right)^{\frac{1}{2}}.$$

  When there is no confusion, we will drop Ω in the subscript, e.g., $\|f\|_0$ simply denotes the L^2-norm.

- $$H^1([0,1]) := \left\{ f(x), f'(x) \in L^2 : \int_0^1 [f(x)^2 + f'(x)^2] dx < \infty \right\}.$$

  The $H^1(\Omega)$-norm will be denoted as

  $$\|f\|_{1,\Omega} = \|f\|_{H^1(\Omega)} = \left(\int_\Omega [f(x)^2 + f'(x)^2] dx \right)^{\frac{1}{2}}.$$

  We also define a seminorm:

  $$|f|_{1,\Omega} = |f|_{H^1(\Omega)} = \left(\int_\Omega f'(x)^2 dx \right)^{\frac{1}{2}}.$$

  When there is no confusion, we will drop Ω in the subscript, e.g., $\|f\|_1$ simply denotes the H^1-norm.

  Fact: in one dimension, $H^1([0,1]) \subset C([0,1])$.

- $H_0^1([0,1])$ is the subset $H^1([0,1])$ with the property of vanishing at the boundary.

- H^2 space is similarly defined:

  $$H^2([0,1]) = \left\{ f(x), f'(x), f''(x) \in L^2 : \int_0^1 [f(x)^2 + f'(x)^2 + f''(x)^2] dx < \infty \right\}.$$

  Norm and semi-norm are

  $$\|f\|_{2,\Omega} = \|f\|_{H^2(\Omega)} = \left(\int_\Omega [f(x)^2 + f'(x)^2 + f''(x)^2] dx \right)^{\frac{1}{2}},$$

  $$|f|_{2,\Omega} = |f|_{H^2(\Omega)} = \left(\int_\Omega f''(x)^2 dx \right)^{\frac{1}{2}}.$$

- H^3 space and its norm are also similarly defined: just add $f'''(x)$.

More about continuity:

- The most general statement is from general Sobolev inequalities [1], which imply for a bounded open set $\Omega \subset \mathbb{R}^n$ with a C^1 boundary:

$$k > \frac{n}{2}, f(\mathbf{x}) \in H^k(\Omega) \Longrightarrow f(\mathbf{x}) \in C(\bar{\Omega}).$$

- The special case for one dimension: $f(x) \in H^1(-1,1) \Longrightarrow f(x) \in C[-1,1]$.
- Two dimensions: $f(x,y) \in H^2(\Omega) \Longrightarrow f(x,y) \in C(\bar{\Omega})$.
- Three dimensions: $f(x,y,z) \in H^2(\Omega) \Longrightarrow f(x,y,z) \in C(\bar{\Omega})$.
- In two dimensions, H^1 is not enough for continuity: consider Ω as a disk centered at the origin with radius $R = \frac{1}{2}$, then the following function cannot be made continous or even bounded by changing any point values:

$$f(x,y) = \left(-\log(x^2+y^2)\right)^{\alpha} \in H^1(\Omega), f(x,y) \notin C(\Omega),$$

where $\alpha \in (0, \frac{1}{2})$ is a constant. Let $r = \sqrt{x^2+y^2}$, we first have

$$\iint_{\Omega} |f|^2 dxdy = \int_{r=0}^{\frac{1}{2}} \int_{\theta=0}^{2\pi} [-\log r^2]^{2\alpha} r dr d\theta \leq C$$

because $[-\log r^2]^{2\alpha} r$ is bounded and continuous on $r \in [0, \frac{1}{2}]$. Then

$$|\nabla f| = \alpha \left(-\log r^2\right)^{\alpha-1} \frac{1}{r} = C(-\log r)^{\alpha-1} r^{-1}.$$

$$\iint_{\Omega} |\nabla f|^2 dxdy = \int_{r=0}^{\frac{1}{2}} \int_{\theta=0}^{2\pi} C(-\log r)^{2\alpha-2} r^{-2} r dr d\theta$$

$$= C \int_{r=0}^{\frac{1}{2}} (-\log r)^{2\alpha-2} r^{-1} dr (t = -\log r) = -C \int_{t=-\log\frac{1}{2}}^{+\infty} t^{2\alpha-2} dt < +\infty.$$

1.2.2 Interpolation and quadrature

Finite element methods are built upon basic tools including interpolation and quadrature (numerical integration).

- Lagrange interpolation is a convenient polynomial approximation to a function through its point values: given $k+1$ point values of $f(x)$ at $k+1$ grid points x_i $(i = 1, 2, \cdots, k+1)$, there is a unique polynomial $p(x)$ of degree k to satisfy $p(x_i) = f(x_i)$ $(i = 1, 2, \cdots, k+1)$.

The linear Lagrange interpolation at x_i, x_{i+1} for a function $f(x)$ is given by

$$\frac{x - x_{i+1}}{x_i - x_{i+1}} f_i + \frac{x - x_i}{x_{i+1} - x_i} f_{i+1}.$$

The quadratic Lagrange interpolation at x_{i-1}, x_i, x_{i+1} for a function $f(x)$ is given by

$$\frac{(x - x_i)(x - x_{i+1})}{(x_{i-1} - x_i)(x_{i-1} - x_{i+1})} f_{i-1} + \frac{(x - x_{i-1})(x - x_{i+1})}{(x_i - x_{i-1})(x_i - x_{i+1})} f_i + \frac{(x - x_i)(x - x_{i-1})}{(x_{i+1} - x_i)(x_{i+1} - x_{i-1})} f_{i+1}.$$

- Quadrature means numerical integration, which is to approximate integrals on computer.

$$\text{Trapezoidal rule} : \int_{-1}^{1} f(x)dx \approx f(-1) + f(1)$$
$$\text{Simpson's rule} : \int_{-1}^{1} f(x)dx \approx \frac{1}{3} f(-1) + \frac{4}{3} f(0) + \frac{1}{3} f(1)$$

Trapezoidal rule is exact if $f(x)$ is a linear polynomial. Simpson's rule is also 3-point Gauss-Lobatto rule or 3-point Newton-Cotes rule, which is exact if $f(x)$ is a cubic polynomial.

Consider an uniform mesh with grids $0 = x_0 < x_1 < \cdots < x_N < x_{N+1} = 1$ with spacing $h = \frac{1}{N+1}$ for the interval $[0, 1]$, which consists of $N + 1$ intervals $I_k = [x_{k-1}, x_k]$ $(k = 1, \cdots, N + 1)$. Then for each interval we can use a linear polynomial to approximate $f(x)$ if given $f_i = f(x_i)$. Let $\Pi_1 f(x)$ denote such a piecewise linear polynomial function.

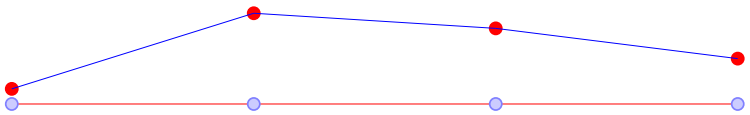


Figure 1.1: Four grid points and three intervals. For each interval, a linear polynomial is interpolated.

Next consider an uniform mesh with grids $0 = x_0 < x_1 < \cdots < x_N < x_{N+1} = 1$ with spacing $h = \frac{1}{N+1}$ for the interval $[0, 1]$. And this time we assume $N = 2n-1$ is odd. Then there are n small intervals $I_k = [x_{2k-2}, x_{2k}]$ $(k = 1, \cdots, n)$, on which we can define a piecewise quadratic interpolation polynomial, denoted by $\Pi_2 f(x)$.

Here are the facts that we will use without any proof first: for a smooth enought function $f(x)$, the interpolation error and quadrature error are given as

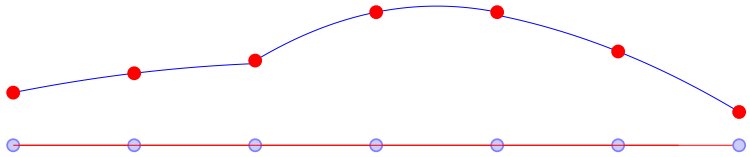


Figure 1.2: Seven grid points and three intervals. For each interval, a quadratic polynomial is interpolated.

- L^2 and H^1 errors of piecewise linear interpolation:

$$\|f - \Pi_1 f\|_0 \leq Ch^2|f|_2, \quad \|f - \Pi_1 f\|_1 \leq Ch|f|_2. \tag{1.2}$$

- L^2 and H^1 errors of piecewise quadratic interpolation:

$$\|f - \Pi_2 f\|_0 \leq Ch^3|f|_3, \quad \|f - \Pi_2 f\|_1 \leq Ch^2|f|_3. \tag{1.3}$$

- Quadrature error of trapezoidal rule for each small interval in Figure 1.1 :

$$\left| \int_0^1 f(x)dx - \sum_{k=1}^{N+1} \frac{1}{2} h[f(x_{k-1}) + f(x_k)] \right| \leq Ch^2|f|_2.$$

- Quadrature error of Simpson's rule for each small interval in Figure 1.2:

$$\left| \int_0^1 f(x)dx - \sum_{k=1}^{n} h[\frac{1}{3}f(x_{2k-1}) + \frac{4}{3}f(x_{2k}) + \frac{1}{3}f(x_{2k+1})] \right| \leq Ch^4|f|_4.$$

Notice that the estimate above only needs the minimal assumption on the function, e.g., for (1.2) we only need to assume $f(x) \in H^2(\Omega)$ (the second order derivative exists in the weak sense). The same order can be obtained by Taylor expansion, but obviously we need the derivatives to exist in the classical sense. All these estimates above can be easily derived from the Bramble-Hilbert Lemma in Section 1.7.2. On the other hand, you can simply assume these estimates are true for now.

1.3 1D BVP: homogeneous Dirichlet b.c.

1.3.1 Variational formulation

Given a function $f(x) \in L^2(0,1)$, consider solving

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

with boundary conditions

$$u(0) = 0, \quad u(1) = 0.$$

Multiplying a test function $v \in H_0^1(0,1)$, after integration by parts, we get

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

which can be denoted as

$$(u', v') = (f, v),$$

if we define

$$(f, g) := \int_0^1 f(x)g(x)dx.$$

It can be shown that the solution to the PDE is equivalent to the solution to the following variational formulation

$$\text{seek} \quad u \in H_0^1(0,1), \quad \text{satisfying} \quad (u', v') = (f, v), \forall v \in H_0^1(0,1). \tag{1.4}$$

Theorem 1.1. *Assume* $f(x) \in C([0,1])$ *and* $u(x) \in C^2([0,1])$ *satisfies* (1.4), *then* $-u''(x) = f(x)$.

Proof. After integration by parts in (1.4), we get

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

If $u''(x) + f(x) \neq 0$, then due to continuity, $u''(x) + f(x)$ is either positive or negative on an interval $[x_0, x_1] \subset [0,1]$. Without loss of generality, assume $u''(x) + f(x) > 0$ on $[x_0, x_1] \subset [0,1]$. Consider a test function

$$v(x) = \begin{cases} 0, & x < x_0 \\ (x - x_0)^2(x - x_1)^2, & x \in [x_0, x_1] \\ 0, & x > x_1 \end{cases},$$

and we have

$$\int_0^1 [u''(x) + f(x)]v(x)dx = \int_{x_0}^{x_1} [u''(x) + f(x)]v(x)dx > 0,$$

which is a contradiction. □

Why the variational formulation implies the PDE is one big step that we choose to skip. If this is your first time to learn finite element method, it is the best to accept this fact without spending time pursuing why. But if this is your fifth or even tenth time to read this chapter, it might be a good time to start to learn why it should be true, in a different book! Of course the solution of (1.4) is also the solution to the PDE only when it is a solution with a second order derivative at least in the weak sense. It can be shown that the solution of (1.4) has weak second order derivative, which is called *elliptic regularity theorem*.

Consider the 1D variable coefficient problem

$$-(a(x)u')' = f, \quad x \in (0,1), \quad u(0) = u(1) = 0, \tag{1.5}$$

where $a(x) > 0$ is a smooth coefficient, with boundary conditions.

We can introduce a new notation called bilinear form

$$\mathcal{A}(u,v) := \int_0^1 au'v'dx,$$

then the equivalent variational formulation is

$$\text{seek} \quad u \in H_0^1(0,1), \quad \text{satisfying} \quad \mathcal{A}(u,v) = (f,v), \forall v \in H_0^1(0,1). \tag{1.6}$$

1.3.2 The abstract finite element method

Given a mesh with n intervals I_j ($j = 1, \cdots, n$), let V^h denote the continuous piecewise polynomial of degree k approximation to the space $H^1(0,1)$:

$$V^h := \{v_h(x) \in C(0,1) : v_h(x) \text{ is polynomial of degree } k \text{ on each interval } I_j\}.$$

We will only consider $k = 1$ or $k = 2$, i.e., linear or quadratic polynomial approximation. In general, these small intervals I_j do not have to be uniform. But for convenience and also for the sake of constructing a finite difference scheme on a uniform mesh, let us assume they have an uniform interval size. Then Figure 1.1 and Figure 1.2 are illustrations of elements in V^h.

The space V_0^h is similarly defined as an approximation to $H_0^1(0,1)$:

$$V_0^h := \{v_h(x) \in C(0,1) : v_h(0) = v_h(1) = 0, v_h(x) \in P^k(I_j), \forall j\}.$$

A continous piecewise polynomial can have a weak derivative as defined by Definition 1.1, which is the piecewise derivative inside each interval, just like that the weak derivative of $f(x) = |x|$ is the step function. Thus we have the following fact:

$$V^h \subset H^1(0,1), \quad V_0^h \subset H_0^1(0,1).$$

Given V_0^h, the abstract finite element method for (1.6) is defined as

$$\text{seek} \quad u_h \in V_0^h, \quad \text{satisfying} \quad \mathcal{A}(u_h, v_h) = (f, v_h), \forall v_h \in V_0^h. \tag{1.7}$$

We call (1.7) the abstract finite element method because it can never be exactly implemented. For example, the right hand side integral (f, v_h) can never be computed exactly, unless $f(x)$ is a very simple function.

1.3.3 The abstract implementation

Assume we know how to compute all integrals in (1.7), e.g., if the coefficient $a(x) \equiv 1$ and $f(x)$ is a polynomial in (1.6), then all integrands are polynomials. Then let us think about how the scheme (1.7) should be implemented, e.g., in the scheme (1.7) what does arbitrariness of the test function v_h mean?

First of all, once the mesh is fixed and polynomial degree is fixed, the piecewise polynomial space V_0^h is a finite dimensional vector space. Assume it is N-dimensional with basis functions $\{\phi_i(x) : i = 1, \cdots, N\}$.

Second, in the scheme (1.7), both the left hand side and the right hand side are linear operators with respect to the test function v_h. Therefore, $A(u_h, v_h) = (f, v_h)$ for arbitrary test function v_h in an N-dimensional vector space V_0^h is equivalent to $A(u_h, v_h) = (f, v_h)$ for v_h being all basis functions $\phi_i(x)$. Namely, (1.7) is equivalent to

$$\mathcal{A}(u_h, \phi_i) = (f, \phi_i), \quad i = 1, \cdots, N.$$

Third, $u_h \in V_0^h$ implies that u_h is a linear combination of the basis functions:

$$u_h(x) = \sum_{j=1}^{N} u_j \phi_j(x).$$

Next, plugging in $u_h(x) = \sum_{j=1}^{N} u_j \phi_j(x)$ and using the linearity of the bilinear form $\mathcal{A}$, we get that

$$\sum_{j=1}^{N} u_j \mathcal{A}(\phi_j, \phi_i) = (f, \phi_i), \quad i = 1, \cdots, N,$$

which is a system of N linear equations.

The last step is to solve a linear system $S\mathbf{u} = \mathbf{f}$ where the *stiffness* matrix S has entries $S_{ij} = \mathcal{A}(\phi_j, \phi_i)$, and

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} (f, \phi_1) \\ (f, \phi_2) \\ \vdots \\ (f, \phi_N) \end{pmatrix}.$$

1.3.4 The simple practical implementation on uniform meshes

To implement the scheme (1.7), one needs to address the issue of how to compute integrals. One convenient choice is to use quadrature. Let us use trapezoidal rule for P^1 method (and Simpson's rule for P^2 method). Let $\mathcal{A}_h(\cdot, \cdot)$ and $\langle f, v_h \rangle_h$ denote the quadrature approximation to $\mathcal{A}(\cdot, \cdot)$ and (f, v_h) respectively. Then we get a new scheme

$$\text{seek} \quad u_h \in V_0^h, \quad \text{satisfying} \quad \mathcal{A}_h(u_h, v_h) = \langle f, v_h \rangle_h, \forall v_h \in V_0^h. \tag{1.8}$$

Recall that for both P^1 mesh Figure 1.1 and P^2 mesh 1.2, there are N interior grid points. Let $\phi_i(x)$ $(i = 1, \cdots, N)$ denote the basis functions in V_0^h satisfying

$$\phi_i(x_j) = \delta_{ij} = \begin{cases} 1 & i = j, \\ 0 & i \neq j, \end{cases}, \forall j = 1, \cdots, N.$$

This kind of basis is often called *Lagrangian basis* or *nodal basis*. For instance, $\phi_i(x)$ for P^1 method is given as

$$\phi_i(x) = \begin{cases} \frac{1}{h}(x - x_{i-1}), & x \in [x_{i-1}, x_i], \\ \frac{1}{h}(x_{i+1} - x), & x \in [x_i, x_{i+1}], \\ 0, & \text{otherwise}, \end{cases}$$

and its weak derivative is

$$\phi_i'(x) = \begin{cases} \frac{1}{h}, & x \in [x_{i-1}, x_i], \\ -\frac{1}{h}, & x \in [x_i, x_{i+1}], \\ 0, & \text{otherwise}. \end{cases}$$

For Lagrangian basis $\phi_i(x)$, if we set $u_j = u_h(x_j)$, then

$$u_h(x) = \sum_{i=1}^{N} u_j \phi_j(x),$$

thus the numerical solution u_h can also be denoted as a vector of point values

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}.$$

Plugging $\sum_{j=1}^{N} u_j \phi_j(x, y)$ into the bilinear form, we get

$$\mathcal{A}_h(u_h, v_h) = \sum_{j=1}^{N} u_j \mathcal{A}_h(\phi_i(x), v_h).$$

Since it suffices to ask $\mathcal{A}_h(u_h, v_h)$ to hold for $v_h = \phi_i$ for all i, the scheme (1.8) is equivalent to

$$\text{seek } \mathbf{u} \in \mathbb{R}^N, \quad \text{satisfying } \sum_{j=1}^{N} \mathcal{A}_h(\phi_j(x), \phi_i(x)) u_j = \langle f, \phi_i(x) \rangle_h, \forall i = 1, \cdots, N. \tag{1.9}$$

The right hand side can be explicitly written as

$$\langle f, \phi_i(x)\rangle_h = \sum_{k=0}^{N} \frac{1}{2} h[f(x_k)\phi_i(x_k) + f(x_{k+1})\phi_i(x_{k+1})] = f_i h.$$

So the matrix vector form of (1.9) is $S\mathbf{u} = h\mathbf{f}$ where the stiffness matrix S has its (i, j)-th entry as

$$S_{ij} = \mathcal{A}_h(\phi_j(x), \phi_i(x)).$$

Consider the simplest Laplacian case $a(x) \equiv 1$, then

$$S_{ij} = \mathcal{A}_h(\phi_j(x), \phi_i(x)) = \langle \phi_j'(x), \phi_i'(x)\rangle_h = \begin{cases} \frac{2}{h} & i = j \\ -\frac{1}{h} & i = j \pm 1 \\ 0 & \text{otherwise.} \end{cases}.$$

In other words, for solving $-u'' = f, u(0) = u(1) = 0$, the matrix vector form of the P^1 finite element method with trapezoidal quadrature is precisely the second order centered difference:

$$\frac{1}{h}\begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{N-1} \\ u_N \end{bmatrix} = h \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_{N-1} \\ f_N \end{bmatrix}.$$

For the variable coefficient problem $-(au')' = f, u(0) = u(1) = 0$, similarly we can derive the matrix vector form for the scheme (1.9) with piecewise linear basis:

$$\frac{1}{h}\frac{1}{2}\begin{pmatrix} a_0 + 2a_1 + a_2 & -a_1 - a_2 & & & \\ -a_1 - a_2 & a_1 + 2a_2 + a_3 & -a_2 - a_3 & & \\ & & & & \\ & & \ddots & \ddots & \ddots \end{pmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \\ \vdots \end{bmatrix} = h \begin{bmatrix} f_1 \\ f_2 \\ \\ \vdots \end{bmatrix} \tag{1.10}$$

Recall that the traditional finite difference scheme (**??**) in Chapter **??** is given as

$$\frac{1}{\Delta x^2}\begin{pmatrix} a_{\frac{1}{2}} + a_{\frac{3}{2}} & -a_{\frac{3}{2}} & & \\ -a_{\frac{3}{2}} & a_{\frac{3}{2}} + a_{\frac{5}{2}} & -a_{\frac{5}{2}} & \\ & \ddots & \ddots & \ddots \end{pmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \end{bmatrix},$$

and the matrix can be easily written as $B = \frac{1}{\Delta x^2} D^T A D$, where A be a diagonal matrix with diagonal entries $a_{\frac{1}{2}}, \cdots, a_{n+\frac{1}{2}}$ and

$$D = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ 0 & \ddots & \ddots & \\ & & -1 & 1 \\ & & & -1 \end{bmatrix}_{(n+1)\times n}.$$

Notice that the two schemes (**??**) and (1.10) would be exactly the same if we use an approximation $a_{j+\frac{1}{2}} \approx \frac{a_j + a_{j+1}}{2}$ for the mid point values of $a(x)$ in (**??**). For smooth $a(x)$, the approximation $a_{j+\frac{1}{2}} \approx \frac{a_j + a_{j+1}}{2}$ is second order accurate by Taylor expansion. Because of this, the stiffness matrix S in the finite element method (1.10) can be easily written as

$$S = \frac{1}{2}\frac{1}{h} D^T A D$$

where A is a diagonal matrix with diagonal entries $a_0+a_1, a_1+a_2, a_2+a_3, \cdots$.

Problem 1.1. *Are there any alternatives to compute or approximate integrals in* (1.7) *if we do not use quadrature?*

Problem 1.2. *For the variable coefficient problem* $-(au')' = f, u(0) = u(1) = 0$, *derive the equivalent matrix vector form* (1.10) *for the scheme* (1.9) *with piecewise linear basis.*

Problem 1.3. *Derive the basis functions* $\phi_i(x)$ *for the* P^2 *method and find the explicit matrix vector form of the scheme* (1.9) *for the* $-u'' = f, u(0) = u(1) = 0$.

Problem 1.4. *Implement both schemes* (**??**) *and* (1.10), *and compare their errors for a problem with a smooth solution* u *for a smooth coefficient* $a(x)$.

Problem 1.5. *For a rectangular domain* Ω, *consider a 2D variable coefficient problem*

$$-\nabla \cdot (a(\mathbf{x})\nabla u(\mathbf{x})) = f(\mathbf{x}), \quad \mathbf{x} \in \Omega$$

with homogeneous Dirichlet boundary condition, where $a(\mathbf{x}) > 0$ *is a scalar coefficient. Consider a uniform rectangular mesh and using* Q^1 *finite element method with trapezoidal quadrature for both* x *and* y *variables. The finite element method is to seek* $u_h \in V_0^h$ *satisfying*

$$\mathcal{A}_h(u_h, v_h) = \langle f, v_h \rangle.$$

Using notation in Chapter 2, the scheme can be written as

$$\left[\frac{1}{\Delta x^2}(D_x^T \otimes I_y)A_1(D_x \otimes I_y) + \frac{1}{\Delta y^2}(I_x \otimes D_y^T)A_2(I_x \otimes D_y)\right] vec(U) = vec(F),$$

where A_1 and A_2 are two diagonal matrices defined as follows.

Let a_1 be a 2D array of size $Ny\times(Nx+1)$ satisfying $a_1(j,i) = \frac{1}{2}a(x_i, y_j)+\frac{1}{2}a(x_{i-1}, y_j)$ and a_2 be a 2D array of size $(Ny+1)\times Nx$ satisfying $a_2(j,i) = \frac{1}{2}a(x_i, y_j) + \frac{1}{2}a(x_i, y_{j-1})$. Then A_1 and A_2 can be easily generated in MATLAB as sparse diagonal matrices:

```
A1=sparse(diag(a1(:)));
A2=sparse(diag(a2(:)));
```

Implement this scheme and test the accuracy for a smooth solution.

1.4 Basic properties of the bilinear form

1.4.1 Coercivity

We consider the bilinear form $\mathcal{A}(u,v) = \int_0^1 au'v'dx$ with the smooth coefficient $a(x)$ satisfying $0 \le \min\limits_{x\in[0,1]} a(x) \le a(x) \le \max\limits_{x\in[0,1]} a(x) < +\infty$ for any x. The first useful concept is called *coercivity* of the bilinear form:

$$\forall v \in H_0^1(\Omega), \quad \mathcal{A}(v,v) = \int_0^1 a(x)v'v'dx \ge \min_x a(x)|v|_1^2 \ge C\|v\|_1^2,$$

where $C > 0$ is a constant.

To establish the coercivity, we have used the fact that the H^1-seminorm $|\cdot|_{1,\Omega}$ and the H^1-norm $\|\cdot\|_{1,\Omega}$ are equivalent in $H_0^1(\Omega)$, i.e., there is a constant $C > 0$ depending only on Ω s.t. for any

$$\forall v \in H_0^1(\Omega), \quad C\|v\|_{1,\Omega}^2 \le |v|_{1,\Omega}^2 \le \|v\|_{1,\Omega}^2. \tag{1.11}$$

The second inequality in (1.11) is trivial. The first inequalilty in (1.11) simply says that the function value can be controlled by the derivatives, which is in general not true. For example, if $v(x) \equiv 1$ on $\Omega = [0,1]$, then $\|v\|_0 = 1$ and $|v|_1 = \|v'(x)\|_0 = 0$ thus the first inequalilty cannot hold for $v \notin H_0^1(\Omega)$.

For $\Omega = (0,1)$, here are some quick arguments to see why it is even possible to control function values by derivatives for $v(x) \in H_0^1(0,1)$. If $v(0) = 0$ and $v'(x)$ exists everywhere in the classical sense, then by the Mean Value Theorem we have $\frac{v(x)-v(0)}{x-0} = v'(y)$ for some $y \in (0,x)$, thus $v(x) = xv'(y)$ and $|v(x)| \le |v'(y)|$ for any $x \in [0,1]$. You can simply assume (1.11) is true for now, and read *Poincaré inequalilty* in the Appendix for a rigorous statement.

Remark 1.4. *The estimates in* (1.11) *hold even for a function $v(x)$ which vanishes only along a part or a very small part of the boundary of Ω.*

1.4.2 Continuity

The *continuity* of the bilinear form is simple implication of Cauchy Schwartz inequality:

$$\forall u, v \in H^1(\Omega), \mathcal{A}(u,v) = \int_0^1 au'v'dx \leq \max_x a(x) \int_0^1 u'v'dx$$

$$\leq \max_x a(x) \sqrt{\int_0^1 [u']^2 dx} \sqrt{\int_0^1 [v']^2 dx} \leq C\|u\|_1 \|v\|_1.$$

1.4.3 Coercivity is stability

Recall that the abstract finite element method can be casted as a linear system:

$$\sum_{j=1}^{N} u_j \mathcal{A}(\phi_j, \phi_i) = (f, \phi_i), \quad i = 1, \cdots, N.$$

Whenever a scheme is given as a linear system for an elliptic equation, it must be addressed whether the linear system has a solution. In other words, we need to show the solvability, i.e., the invertability of the *stiffness* matrix S with entries $S_{ij} = \mathcal{A}(\phi_j(x), \phi_i)$. In Chapter **??**, we computed eigenvalues of K matrix for constant coefficient problems so that we can show the nonsingularity of the matrix K.

Obviously, for a variable coefficient problem, e.g., the scheme (1.10), the eigenvalues of the stiffness matrix should be estimated rather than computed because it is nearly impossible to compute. Such an eigenvalue estimate can be done by the coercivity. In this section we only focus on the bilinear form $\mathcal{A}$ and we will discuss $\mathcal{A}_h$ later. For any $v_h \in V_0^h$, we have

$$\mathcal{A}(v_h, v_h) = \mathcal{A}(\sum_{i=1}^{n} v_i\phi_i, \sum_{j=1}^{n} v_j\phi_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathcal{A}(\phi_j, \phi_i) v_j v_i = \mathbf{v}^T S \mathbf{v}.$$

The coercivity says that

$$\mathcal{A}(v,v) \geq C_1\|v\|_1^2 \geq C_1\|v\|_0^2, \quad \forall v(x) \in H_0^1(\Omega),$$

where the constant C_1 only depends on the domain Ω. Thus

$$v_h(x) \in V_0^h \subset H_0^1(\Omega) \Rightarrow \mathcal{A}(v_h, v_h) \geq C_1\|v_h\|_0^2.$$

Notice that $\|v_h\|_0$ and $\|\mathbf{v}\|$ are both norms of the same finite dimensional vector space V_0^h, thus they are equivalent:

$$C_2\|\mathbf{v}\|^2 \leq \|v_h\|_0^2 \leq C_3\|\mathbf{v}\|^2,$$

where the constants C_2, C_3 depends on the dimension N of the vector space V_0^h.

Thus coercivity gives us

$$\mathbf{v}^T S\mathbf{v} = \mathcal{A}(v_h, v_h) \geq C_1\|v_h\|_0^2 \geq C_1C_2\|\mathbf{v}\|^2 \Rightarrow \frac{\mathbf{v}^T S\mathbf{v}}{\mathbf{v}^T\mathbf{v}} \geq C_1C_2.$$

Recall that $\mathcal{A}(u,v) = \int_0^1 a(x)u'(x)v'(x)dx$, thus $\mathcal{A}(u,v) = \mathcal{A}(v,u)$ implies that S is real symmetric. By the Courant-Fisher-Weyl min-max principle (see Appendix **??**), $\frac{\mathbf{v}^T S\mathbf{v}}{\mathbf{v}^T\mathbf{v}} \geq C_1C_2$ implies that the smallest eigenvalue of S is greater than or equal to $C_1C_2 > 0$. Therefore S is invertible.

In Section 1.6.4, we will further show that we can get a useful estimate for $\|S^{-1}\|$ from the coercivity.

1.5 Error estimates of the abstract finite element method

We have just seen that the finite element method with linear basis and quadrature recovers the K matrix for approximating second order derivative. Even though a finite element method with quadrature becomes a finite difference scheme, it is much more useful to understand the same scheme from the finite element perspective. In Chapter **??**, we had to find eigenvalues of K to discuss the stability thus convergence rate for the FD scheme $\frac{1}{h^2}K\mathbf{u} = \mathbf{f}$. Obviously, eigenvalues of a matrix are difficult to find in general, e,g., the matrix (1.10).

We first prove the error estimates for (1.7), i.e., the finite element method without quadrature. We will focus on the analysis for P^1 finite element method for the one-dimensional problem with homogeneous Dirichlet boundary conditions to understand the key components in the finite element method anaylysis, but keep in mind that all discussions in this section apply to much more general cases such as solving variable coefficient elliptic equations by P^k polynomial finite element method with Neumann boundary conditions on unstructured meshes for a curved domain in multiple dimensions.

1.5.1 H^1-norm estimate: stability and consistency imply convergence

The continuous piecewise polynomial has weak derivatives, thus we have $V_0^h \subset H_0^1(\Omega)$, and such a finite element method is called *conforming*. Here is one simple example of nonconforming finite element space for the Poisson equation: the discontinuous piecewise polynomial space is not a subspace of the $H_0^1(\Omega)$ function space.

Theorem 1.2 (Galerkin Orthogonality)**.** *Let u be the solution to* (1.6)*. The solution u_h to the conforming finite element method* (1.7) *satisfies:*

$$\mathcal{A}(u - u_h, w_h) = 0, \forall w_h \in V_0^h.$$

Proof. The exact solution u satisfies

$$\mathcal{A}(u, w) = (f, w), \forall w \in H_0^1(\Omega)$$

thus

$$\mathcal{A}(u, w_h) = (f, w_h), \forall w_h \in V_0^h \subset H_0^1(\Omega).$$

The numerical solution u_h satisfies

$$\mathcal{A}(u_h, w_h) = (f, w_h), \forall w_h \in V_0^h.$$

Subtracting these two equations, we get Galerkin Orthogonality, which is a straightforward implication from the choice of approximation space $V_0^h \subset H_0^1(\Omega)$. □

Galerkin Orthogonality simply says that the true error $u - u_h$ is somehow "orthogonal" to any test function w_h in V_0^h through the bilinear form $\mathcal{A}$.

Remark 1.5. *Galerkin Orthogonality is the analog of the consistency or truncation error in Chapter* **??**. *Since* $\mathcal{A}(u_h, w_h) = (f, w_h)$,

$$\mathcal{A}(u - u_h, w_h) = \mathcal{A}(u, w_h) - \mathcal{A}(u_h, w_h) = \mathcal{A}(u, w_h) - (f, w_h).$$

So Galerkin Orthogonality is the same as

$$\mathcal{A}(u, w_h) - (f, w_h) = 0, \forall w_h \in V_0^h,$$

which is nothing but replacing u_h *by* u *in the numerical scheme. On the other hand, it seems that the "truncation error" is zero here, which is due to the direct approximation of the variational form. Notice that the "truncation error"* $A(u, w_h) - (f, w_h)$ *is zero only when the test functions are in the approximated space* V_0^h.

Theorem 1.3 (Céa's Lemma)**.** *Let* u *be the solution to* (1.6). *The solution* u_h *to the conforming finite element method* (1.7) *satisfies:*

$$\|u - u_h\|_1 \leq C \inf_{w_h \in V_0^h} \|u - w_h\|_1.$$

Proof. First of all, we have $u_h \in V_0^h \subset H_0^1(\Omega)$ thus $u - u_h \in H_0^1(\Omega)$. The coercivity implies

$$C\|u - u_h\|_1^2 \leq \mathcal{A}(u - u_h, u - u_h).$$

Next, we have

$$\mathcal{A}(u - u_h, u - u_h) = \mathcal{A}(u - u_h, w_h - u_h) + \mathcal{A}(u - u_h, u - w_h).$$

Galerkin orthogonality implies $\mathcal{A}(u-u_h, w_h-u_h)=0$. So we get

$$\mathcal{A}(u-u_h, u-u_h)=\mathcal{A}(u-u_h, u-w_h)\le C\|u-u_h\|_1\|u-w_h\|_1,$$

where continuity is used.

Finally, we have

$$C\|u-u_h\|_1^2\le \mathcal{A}(u-u_h, u-u_h)=\mathcal{A}(u-u_h, u-w_h)\le C\|u-u_h\|_1\|u-w_h\|_1,$$

thus

$$\|u-u_h\|_1^2\le C\|u-u_h\|_1\|u-w_h\|_1.$$

So we have $\|u-u_h\|_1\le C\|u-w_h\|_1$ for any $w_h\in V_0^h$.

□

Céa's Lemma says the finite element solution error is controlled by the best piecewise polynomial approximation error, which we do not know. On the other hand, we do know polynomial interpolation error estimates (1.2) and (1.3). Assuming $u\in H^2(\Omega)$ or $u\in H^3(\Omega)$ (i.e., assuming u is smooth enough), we easily obtain error estimate in H^1-norm:

$$\|u-u_h\|_1\le C\inf_{w_h\in V_0^h}\|u-w_h\|_1\le C\|u-\Pi_k u\|_1=\begin{cases}Ch|u|_2, & k=1\\ Ch^2|u|_3, & k=2\end{cases}. \quad (1.12)$$

Céa's Lemma gives (1.12), which is the *convergence* of the finite element method. On the other hand, Céa's Lemma is implied by both Galerkin Orthogonality (consistency) and Coercivity (stability). So we get the same conclusion as in Chapter **??** and Chapter **??** for linear schemes solving linear PDEs:

$$\text{consistency}+\text{stability}\longrightarrow\text{convergence}.$$

Recall that $u_h\in V_0^h\subset H_0^1(\Omega)$ and $u\in H_0^1(\Omega)$, thus the error function $u-u_h$ is an element in the $H_0^1(\Omega)$ space, in which H^1-norm is equivalent to the H^1-seminorm. So the H^1-norm estimate above simply implies that P^1 finite element method generates a numerical solution satisfying that

$$\sqrt{\int_0^1|u'(x)-u_h'(x)^2|dx}=\mathcal{O}(h).$$

For function values, we will get one order higher, explained in the next subsection.

1.5.2 L^2-norm estimate: elliptic regularity and duality arguments

Notice that H^1 estimate cannot explain why P^1 method gives a second order accurate scheme. Recall that H^1-norm, i.e., $\|u-u_h\|_1$, measures the error

both in the function value and the first order derivative. The L^2-norm, i.e., $\|u-u_h\|_0$ measures the error in the function value. For example, we already know that the P^1 finite element method on a uniform mesh gives exactly the standard centered finite difference, which is second order accurate for the function value.

The L^2 estimate will be one order higher than H^1 estimate:

Theorem 1.4 (Aubin-Nitsche Lemma)**.** *Let u be the solution to* (1.6)*. The solution u_h to the conforming finite element method* (1.7) *satisfies:*

$$\|u - u_h\|_0 \leq Ch\|u - u_h\|_1,$$

where h is the mesh size.

For proving the Aubin-Nitsche Lemma, we need a basic fact about the Poisson equation, which is called *elliptic regularity*: the solution to (1.6) satisfies $\|u\|_2 \leq C\|f\|_0$, which simply says that the second order derivative of u and lower order ones are controlled by function value of $f(x)$.

Even though we only seek $u(x) \in H_0^1(\Omega)$ in (1.6), the elliptic regularity theorem guarantees that $f(x) \in L^2(\Omega) \Rightarrow u(x) \in H^2(\Omega)$. In particular, if $f(x)$ is infinitely differentiable, then so is $u(x)$. The elliptic regularity can be proven under certain assumptions for the domain Ω.

We also need a *dual* problem to help us here. A dual problem of (1.6) is to find $w \in H_0^1(\Omega)$ satisfying

$$\mathcal{A}(w, v) = (u - u_h, v), \quad \forall v \in H_0^1(\Omega). \tag{1.13}$$

The equivalent PDE form of the dual problem above is

$$-w''(x) = u(x) - u_h(x),$$

if the original PDE we want to solve is $-u''(x) = f(x)$.

The elliptic regularity theorem on the dual problem gives the following

$$\|w\|_2 \leq \|u - u_h\|_0.$$

For the dual problem, its finite element solution for finding $w_h \in V_0^h$ satisfying

$$\mathcal{A}(v_h, w_h) = (u - u_h, v_h), \quad \forall v_h \in V_0^h,$$

where we have used the symmetry of the bilinear form $\mathcal{A}(w, v) = \mathcal{A}(v, w)$. By Céa's Lemma and H^1-estimate applied to the finite element solution w_h, we have

$$\|w - w_h\|_1 \leq Ch|w|_2 \leq Ch\|w\|_2.$$

where we have used the interpolation error estimate.

So with elliptic regularity for the dual problem $\|w\|_2 \leq C\|u - u_h\|_0$, we get

$$\|w - w_h\|_1 \leq Ch\|w\|_2 \leq Ch\|u - u_h\|_0.$$

Proof. Let w_h be the finite element solution for the dual problem. Then Galerkin orthogonality implies $\mathcal{A}(u-u_h, w_h) = 0$, thus

$$\mathcal{A}(u-u_h, w) = \mathcal{A}(u-u_h, w-w_h) + \mathcal{A}(u-u_h, w_h) = \mathcal{A}(u-u_h, w-w_h).$$

Continuity implies

$$\mathcal{A}(u-u_h, w-w_h) \leq C\|u-u_h\|_1\|w-w_h\|_1.$$

Recall that w is the solution to the dual problem thus plugging in $v = u-u_h \in H_0^1(\Omega)$ in (1.13), we get

$$\mathcal{A}(u-u_h, w) = \mathcal{A}(w, u-u_h) = (u-u_h, u-u_h) = \|u-u_h\|_0^2$$

Finally, putting everything together

$$\|u-u_h\|_0^2 = \mathcal{A}(u-u_h, w) \leq C\|u-u_h\|_1\|w-w_h\|_1 \leq C\|u-u_h\|_1 h\|u-u_h\|_0,$$

which gives

$$\|u-u_h\|_0 \leq Ch\|u-u_h\|_1$$

□

With the H^1-norm estimate (1.12), the Aubin-Nitsche Lemma gives us the L^2-norm error estimates:

$$\|u-u_h\|_0 \leq Ch\|u-u_h\|_1 = \begin{cases} Ch^2|u|_2, & k=1 \\ Ch^3|u|_3, & k=2 \end{cases}. \tag{1.14}$$

This is consistent with what we already know: P^1 finite element method gives a second order accurate scheme for function values. The P^2 finite element method gives a third order accurate scheme for function values, which is consistent with the interpolation error order (1.3). However, if we implement P^2 finite element method as a finite difference scheme, we can actually get a fourth order accurate finite difference scheme, which is called *superconvergence*. It will be explained in the rest of the chapter.

Remark 1.6. *In estimates like* (1.12) *and* (1.14)*, it is already assumed that* u *should be smooth enough such that either* $u \in H^2(\Omega)$ *or* $u \in H^3(\Omega)$*. The elliptic regularity theorem implies that* $f(x) \in L^2(\Omega) \Rightarrow u \in H^2(\Omega)$ *and* $f(x) \in H^1(\Omega) \Rightarrow u \in H^3(\Omega)$*.*

1.5.3 Summarization and comparison

Now let us just focus on the P^1 finite element method and think about how the second order accuracy is proven differently from the one we did in Chapter **??**. In Chapter **??**, we computed the eigenvalues of the K matrix for proving stability $\|A^{-1}\| \leq C$ in a matrix-vector form of the scheme

$A\mathbf{u} = \mathbf{f}$. On the one hand, it only requires simpler knowledge of linear algebra. On the other hand, it is highly restrictive because we cannot even compute eigenvalues for a one-dimensional variable coefficient problem.

The discussion in this section obviously applies to the variable coefficient problem, but we need so many much more advanced tools such as Solobev spaces and elliptic regularity. Recall how exactly we can prove the second order error in P^1 finite element method for function values:

1. The homogeneous Dirichlet boundary condition is built into the function space $H_0^1(\Omega)$, which in return gives the *Poincaré inequalilty*:

$$\int_0^1 |v'(x)|^2 dx \geq C \left[\int_0^1 |v'(x)|^2 dx + \int_0^1 |v(x)|^2 dx \right], \quad \forall v(x) \in H_0^1(\Omega).$$

2. The *Poincaié inequalilty* gives the *coercivity*

$$\mathcal{A}(v_h, v_h) \geq C \|v_h\|_0^2,$$

   which is the *stability*.

3. From the fact that it is conforming $V_0^h \subset H_0^1(\Omega)$, *Galerkin orthogonality* is easily obtained:

$$\mathcal{A}(u - u_h, v_h) = 0, \quad \forall v_h \in V_0^h.$$

   Galerkin orthogonality is the *consistency*.

4. With *Galerkin orthogonality* and *coercivity*, we get Céa's Lemma, which says the finite element solution error is controlled by the best piecewise polynomial approximation error:

$$\|u - u_h\|_1 \leq C \inf_{w_h \in V_0^h} \|u - w_h\|_1.$$

   This is step is nothing but saything that consistency and stability imply convergence.

5. We know the interpolation error using P^k polynomials, so the H^1-estimate is simply by Céa's Lemma:

$$\|u - u_h\|_1 \leq C \inf_{v_h \in V_0^h} \|u - v_h\|_1 \leq C \|u - \Pi_k u\|_1 \leq \begin{cases} Ch|u|_2, & k = 1 \\ Ch^2|u|_3, & k = 2 \end{cases}.$$

6. Finally, with the elliptic regularity on a dual problem and almost everything above, we get the Aubin-Nitsche Lemma

$$\|u - u_h\|_0 \leq Ch \|u - u_h\|_1 \leq \begin{cases} Ch^2|u|_2, & k = 1 \\ Ch^3|u|_3, & k = 2 \end{cases}.$$

1.6 V^h-ellipticity: properties of the bilinear form with quadrature

Since in practice quadrature is used to implement the finite element method, we also need to know whether coercivity and continuity hold for $\mathcal{A}_h$. Usually the discrete continuity can be easily derived from Cauchy-Schwartz inequality. The discrete coercivity is called V^h-ellipticity.

We only consider the continuous piecewise linear space V_0^h as an example in this section. Let x_i $(i = 0, 1, \cdots, N+1)$ be an uniform mesh for the whole interval $[0, 1]$, where $x_0 = 0$ and $x_{N+1} = 1$ are boundary points. The grid spacing is $h = \frac{1}{N+1}$.

1.6.1 Equivalent norms of the piecewise linear polynomial space

Everything in this subsection can be derived by abstract arguments. But instead we use some explicit elementary tools to derive what we need for coercivity.

For any $v_h \in V_0^h$, let $v_i = v_h(x_i)$ and $\mathbf{v} = \begin{bmatrix} v_1 & \cdots & v_N \end{bmatrix}^T$. So $\|v_h\|_0$ and $\|\mathbf{v}\|$ are both norms of the same finite dimensional vector space V_0^h, thus they are equivalent:

$$C_2\|\mathbf{v}\|^2 \le \|v_h\|_0^2 \le C_3\|\mathbf{v}\|^2,$$

where the constants C_2, C_3 depends on the dimension N of the vector space V_0^h.

It is useful to figure out the exact dependence of of these constants on the dimension N or the mesh size h. For the one-dimensional problem continuous piecewise linear polynomial space V_0^h on a uniform mesh with mesh size h, we have

$$\|v_h\|_0^2 = \sum_{j=0}^{N} \int_{x_j}^{x_{j+1}} |v_h(x)|^2 dx = \sum_{j=0}^{N} \int_0^h \left[\frac{v_{j+1} - v_j}{h} x + v_j\right]^2 dx$$

$$= h \sum_{j=0}^{N} \left(\frac{1}{3} v_{j+1}^2 + \frac{5}{6} v_j^2 - \frac{1}{6} v_{j+1} v_j\right).$$

Recall that $v_h(x) \in V_0^h \Rightarrow v_0 = v_{N+1} = 0$. With two simple inequalilties

$$-\frac{1}{2} v_{j+1}^2 - \frac{1}{2} v_j^2 \le -v_{j+1} v_j \le \frac{1}{2} v_{j+1}^2 + \frac{1}{2} v_j^2,$$

we can derive

$$h\|\mathbf{v}\|^2 \le \|v_h\|_0^2 \le \frac{4}{3} h \|\mathbf{v}\|^2. \tag{1.15}$$

Let us consider $v_h'(x)$, which is only piecewise constant. Recall that $v_h(x) \in V_0^h$ is weakly differentiable thus $v_h'(x_j)$ is double valued unless $j = 0, N+1$. Let $v_h'(x_j)^-$ and $v_h'(x_j)^+$ denote two values obtained by taking derivatives in the intervals $[x_{j-1}, x_j]$ and $[x_j, x_{j+1}]$ respectively. For convenience, we will also abuse the notation by denoting

$$v_h'(x_j) := \frac{v_h'(x_j)^- + v_h'(x_j)^+}{2}, \quad [v_h'(x_j)]^2 := \frac{[v_h'(x_j)^-]^2 + [v_h'(x_j)^+]^2}{2}.$$

Recall that $v_0 = v_{N+1} = 0$. Let $\mathbf{v}'$ denote the following vector

$$\mathbf{v}' = \begin{pmatrix} v_h'(x_0) \\ v_h'(x_1) \\ v_h'(x_2) \\ \vdots \\ v_h'(x_{N+1}) \end{pmatrix} = \frac{1}{h} \begin{pmatrix} v_1 - v_0 \\ \frac{v_1 - v_0 + v_2 - v_1}{2} \\ \frac{v_2 - v_1 + v_3 - v_2}{2} \\ \vdots \\ v_{N+1} - v_N \end{pmatrix} = \frac{1}{h} \begin{pmatrix} v_1 \\ \frac{v_2}{2} \\ \frac{v_3 - v_1}{2} \\ \vdots \\ -v_N \end{pmatrix},$$

Remark 1.7. *Here for $j = 1, \cdots, N$, we have $[v_h'(x_j)]^2 := \frac{[v_h'(x_j)^-]^2 + [v_h'(x_j)^+]^2}{2} = \frac{v_{j+1} - v_{j-1}}{2h}$, which of course can be regarded as the centered finite difference approximation to the first order derivative at x_j.*

Remark 1.8. *From these happy coincidences with the second order centered difference, we should see that the piecewise linear space V_0^h is the better way to understand or derive the centered difference.*

Let $\bar{V}^h$ denote the vector space of piecewise constant on the intervals I_j. Then $v_h'(x)$ corresponds to an element in $\bar{V}^h$, and obviously $\|\mathbf{v}'\|$ and $\|v_h'\|_0$ can be regarded as two norms for measuring this element in the finite dimensional vector space $\bar{V}^h$, thus they should be equivalent. However, to derive the coercivity of $\mathcal{A}_h(v_h, v_h)$, we need to be careful with the dependence of constants on the dimension N or mesh size h. Similar to (1.15), we can derive

$$\frac{1}{2} h \|\mathbf{v}'\|^2 \leq \|v_h'\|_0^2 \leq 2h \|\mathbf{v}'\|^2. \tag{1.16}$$

Problem 1.6. *Derive* (1.16)*, where $\|v_h'\|_0$ is the L^2-norm for the function $v_h'(x)$. Hint: let $\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_{N+1} \end{pmatrix}$ denote the constants that $v_h'(x)$ corresponds to. Notice that the boundary condition $v_0 = v_{N+1} = 0$ implies that $\sum_{j=1}^{N+1} c_j = 0$. Then $\|v_h'\|_0^2 = h \sum_{j=1}^{N+1} c_j^2$. Derive what $\|\mathbf{v}'\|^2$ should be in terms of c_j.*

1.6.2 Coercivity

If using trapezoidal rule for P^1 finite element method in each cell $I_j = [x_j, x_{j+1}]$ in Figure 1.1, then for any $v_h(x) \in V_0^h$ we have

$$\begin{aligned}
\mathcal{A}_h(v_h, v_h) &= \sum_{j=0}^{N} \frac{h}{2} \left(a(x_j)[v_h'(x_j)^+]^2 + a(x_{j+1})[v_h'(x_{j+1})^-]^2 \right) \\
&\geq \min_j a(x_j) \sum_{j=0}^{N} \frac{h}{2} \left([v_h'(x_j)^+]^2 + [v_h'(x_{j+1})^-]^2 \right) \\
&= \min_j a(x_j) \left(\frac{h}{2}[v_h'(x_0)]^2 + h\sum_{j=1}^{N}[v_h'(x_j)]^2 + \frac{h}{2}[v_h'(x_{N+1})]^2 \right) \\
&\geq \min_j a(x_j) \frac{h}{2} \|\mathbf{v}'\|^2 \geq \min_j a(x_j) \frac{1}{4} \|v_h'(x)\|_0^2 \\
&= \min_j a(x_j) \frac{1}{4} |v_h(x)|_1^2 \geq C \min_x a(x) \frac{1}{4} \|v_h(x)\|_1^2,
\end{aligned}$$

where we have used (1.16) and (1.11) in the last two lines, and the constant C is independent of h or N.

1.6.3 Continuity

The continuity for $\mathcal{A}_h$ is straightforward: for any $w_h, v_h \in V_0^h$, we have

$$\begin{aligned}
&\mathcal{A}_h(w_h, v_h) = \sum_{j=0}^{N} \frac{h}{2} \left(a(x_j)[w_h'(x_j)^+][v_h'(x_j)^+] + a(x_{j+1})[w_h'(x_{j+1})^-][v_h'(x_{j+1})^-] \right) \\
\leq &\max_j a(x_j) \frac{h}{2} \sum_{j=0}^{N} \left(\left| [w_h'(x_j)^+][v_h'(x_j)^+] \right| + \left| [w_h'(x_{j+1})^-][v_h'(x_{j+1})^-] \right| \right) \\
\leq &\max_j a(x_j) \frac{h}{2} \sqrt{\sum_{j=0}^{N} \left([w_h'(x_j)^+]^2 + [w_h'(x_{j+1})^-]^2 \right)} \sqrt{\sum_{j=0}^{N} \left([v_h'(x_j)^+]^2 + [v_h'(x_{j+1})^-]^2 \right)},
\end{aligned}$$

where we have used the Cauchy Schwartz inequality for vectors

$$\sum_i a_i b_i \leq \sqrt{\sum_i a_i^2} \sqrt{\sum_i b_i^2}.$$

Recall that we have defined $[v_h'(x_j)]^2 := \frac{[v_h'(x_j)^-]^2 + [v_h'(x_j)^+]^2}{2}$, thus

$$[v_h'(x_j)^-]^2 + [v_h'(x_j)^+]^2 = 2[v_h'(x_j)]^2$$

and

$$\sqrt{\sum_{j=0}^{N} \left([v_h'(x_j)^+]^2 + [v_h'(x_{j+1})^-]^2 \right)} \leq \sqrt{\sum_{j=0}^{N} 2[v_h'(x_j)]^2} = \sqrt{2} \|\mathbf{v}'\|.$$

With (1.16), we get the continuity

$$\mathcal{A}_h(w_h, v_h) \leq \max_x a(x) \frac{h}{2} \sqrt{2}\|\mathbf{w}'\| \sqrt{2}\|\mathbf{v}'\| \leq 2 \max_x a(x) ||w_h'||_0 ||v_h'||_0$$

$$\leq 2 \max_x a(x) ||w_h||_1 ||v_h||_1.$$

1.6.4 Coercivity implies stability of the finite difference scheme

Recall that in Section 1.4.3 we have shown the nonsingularity of the *stiffness* matrix for the abstract finite element method without any quadrature.

Now we are ready to discuss how the V^h-ellipticity or the discrete coercivity can imply nonsingularity of the *stiffness* matrix S with entries $S_{ij} = \mathcal{A}_h(\phi_j, \phi_i)$ for the finite element method with quadrature. In particular, for P^1 finite element method with trapezoidal rule solving a variable coefficient problem, from (1.10) we know the stiffness matrix can be written as

$$S = \frac{1}{h}\frac{1}{2}\begin{pmatrix} a_0 + 2a_1 + a_2 & -a_1 - a_2 & & \\ -a_1 - a_2 & a_1 + 2a_2 + a_3 & -a_2 - a_3 & \\ & \ddots & \ddots & \ddots \end{pmatrix}$$

Since $v_h(x) = \sum_{j=1}^N v_j \phi_j(x)$, we have

$$\mathcal{A}_h(v_h, v_h) = \mathcal{A}_h(\sum_{j=1}^N v_j \phi_j(x), \sum_{i=1}^N v_i \phi_i(x)) = \sum_{i=1}^N \sum_{j=1}^N \mathcal{A}_h(\phi_j(x), \phi_i(x)) v_i v_j = \mathbf{v}^T S \mathbf{v}.$$

With the coercivity in Section 1.6.2 and (1.15), we have

$$\mathcal{A}_h(v_h, v_h) \geq C\|v_h\|_1^2 \geq C\|v_h\|_0^2 \geq Ch\|\mathbf{v}\|^2$$

So we have $\mathbf{v}^T S \mathbf{v} \geq Ch\|\mathbf{v}\|^2$ for any $\mathbf{v} \in \mathbb{R}^N$, which implies S is positive definite. The symmetry of S is implied by $\mathcal{A}_h(w_h, v_h) = \mathcal{A}_h(v_h, w_h)$. So S is invertible. By the Courant-Fisher-Weyl min-max principle (see Appendix **??**), $\frac{\mathbf{v}^T S \mathbf{v}}{\mathbf{v}^T \mathbf{v}} \geq Ch$ implies that the smallest eigenvalue of S is greater than or equal to $Ch > 0$. Therefore S is invertible.

The matrix-vector form of (1.10) is $S\mathbf{u} = h\mathbf{f}$, thus $\mathbf{u} = hS^{-1}\mathbf{f}$. Since we have shown S is real symmetric positive definite, thus singular values are also eigenvalues for both S and S^{-1}. So $\|S^{-1}\|$ is the simply the reciprocal of the smallest eigenvalue of S. Therefore we get $\|hS^{-1}\| = h\|S^{-1}\| \leq C$, which is precisely the stability in the sense of traditional finite difference method in Chapter **??**.

Problem 1.7. *Recall that the traditional finite difference scheme* (**??**) *in Chapter* **??** *is given as*

$$\frac{1}{\Delta x^2}\begin{pmatrix} a_{\frac{1}{2}}+a_{\frac{3}{2}} & -a_{\frac{3}{2}} & & \\ -a_{\frac{3}{2}} & a_{\frac{3}{2}}+a_{\frac{5}{2}} & -a_{\frac{5}{2}} & \\ & \ddots & \ddots & \ddots \end{pmatrix}\begin{bmatrix} u_1 \\ u_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \end{bmatrix}.$$

Apply the discussion in this section to prove the stability of this scheme. The consistency or the truncation error of the scheme (**??**) *is straightforward to derive. Once we have the stability, we have its convergence following Chapter* **??**. *Hint: it becomes trivial if we can have an equivalent scheme in the following form*

$$\frac{1}{h}\frac{1}{2}\begin{pmatrix} b_0+2b_1+b_2 & -b_1-b_2 & & \\ -b_1-b_2 & b_1+2b_2+b_3 & -b_2-b_3 & \\ & \ddots & \ddots & \ddots \end{pmatrix}\begin{bmatrix} u_1 \\ u_2 \\ \vdots \end{bmatrix} = h\begin{bmatrix} f_1 \\ f_2 \\ \vdots \end{bmatrix}.$$

So how do we define b_i so that they are equivalent?

1.7 Error estimates of the finite element method with quadrature

In order to derive the error estimates of the finite element method with quadrature (1.8), we need to show all the lemmas and theorems in Section 1.5 also hold when $\mathcal{A}(\cdot,\cdot)$ is replaced by $\mathcal{A}_h(\cdot,\cdot)$. If this is your first time to read this chapter, you can assume that this is true and skip this section.

1.7.1 First Strang Lemma

The First Strang Lemma is the Céa's Lemma for the scheme (1.8).

Theorem 1.5. *[First Strang Lemma]*

$$\|u-u_h\|_1 \le C \inf_{v_h\in V_0^h}\left\{\|u-v_h\|_1 + \sup_{w_h\in V_0^h}\frac{|\mathcal{A}(v_h,w_h)-\mathcal{A}_h(v_h,w_h)|}{\|w_h\|_1}\right\}$$
$$+C\sup_{w_h\in V_0^h}\frac{|\langle f,w_h\rangle_h-(f,w_h)|}{\|w_h\|_1}.$$

Remark 1.9. *Compared to Céa's Lemma, the extra terms in the First Strang Lemma is nothing but quadrature error terms.*

Proof. First, we can rewrite the bilinear form

$$\mathcal{A}_h(u_h-v_h,u_h-v_h) = \mathcal{A}_h(u_h,u_h-v_h)-\mathcal{A}_h(v_h,u_h-v_h)+\mathcal{A}(u-v_h,u_h-v_h)-\mathcal{A}(u-v_h,u_h-v_h)$$

$= \mathcal{A}_h(u_h, u_h - v_h) - \mathcal{A}_h(v_h, u_h - v_h) + \mathcal{A}(u - v_h, u_h - v_h) + \mathcal{A}(v_h, u_h - v_h) + \mathcal{A}(u, u_h - v_h).$

By coercivity of $\mathcal{A}_h$, and the facts $\mathcal{A}_h(u_h, u_h - v_h) = \langle f, u_h - v_h \rangle_h$ and $A(u, u_h - v_h) = (f, u_h - v_h)$, we get

$$C\|u_h - v_h\|_1^2 \leq \mathcal{A}_h(u_h - v_h, u_h - v_h) = \mathcal{A}(u - v_h, u_h - v_h) + \mathcal{A}(v_h, u_h - v_h) - \mathcal{A}_h(v_h, u_h - v_h)$$
$$+ \langle f, u_h - v_h \rangle_h - (f, u_h - v_h), \quad \forall v_h \in V_0^h.$$

With $\mathcal{A}(u - v_h, u_h - v_h) \leq C_2 \|u - v_h\|_1 \|u_h - v_h\|_1$, we have

$$C\|u_h - v_h\|_1 \leq C_2 \|u - v_h\|_1 + \frac{\mathcal{A}(v_h, u_h - v_h) - \mathcal{A}_h(v_h, u_h - v_h)}{\|u_h - v_h\|_1} + \frac{\langle f, u_h - v_h \rangle_h - (f, u_h - v_h)}{\|u_h - v_h\|_1}$$

thus

$$\|u_h - v_h\|_1 \leq C\|u - v_h\|_1 + C \sup_{w_h \in V_0^h} \frac{|\mathcal{A}(v_h, w_h) - \mathcal{A}_h(v_h, w_h)|}{\|w_h\|_1} + C \sup_{w_h \in V_0^h} \frac{|\langle f, w_h \rangle_h - (f, w_h)|}{\|w_h\|_1}.$$

The proof is done after using the triangle inequality:

$$\|u - u_h\|_1 \leq \|u - v_h\|_1 + \|u_h - v_h\|_1$$

□

1.7.2 Quadrature estimate: Bramble Hilbert Lemma

The first Strang Lemma means that the Céa's Lemma holds up to the quadrature error, which can be estimated by the Bramble Hilbert Lemma:

Theorem 1.6 (Bramble Hilbert Lemma)**.** *For some integer $k \geq 0$, let $\mathcal{L}$ be a continuous linear form on the space $H^{k+1}(0,1)$ with the property that $\forall p(x) \in P^k(\Omega)$ (all polynomials of degree k), $\mathcal{L}[p(x)] = 0$. Then*

$$|\mathcal{L}(f)| \leq C\|\mathcal{L}\|_{k+1}^* |f|_{k+1},$$

where $\|\cdot\|_{k+1}^*$ is the operator norm and $|f|_{k+1} = \sqrt{\int_\Omega |f^{(k+1)}(x)|^2 dx}$ is the H^{k+1}-seminorm.

Remark 1.10. *The notation in the Bramble Hilbert Lemma are abstract but one typical example of such a linear operator is the interpolation error operator. For instance, given point values f_i of some function $f(x)$ on a uniform mesh, we can do a piecewise linear polynomial interpolation as in Section 1.2.2. The interpolation error is a linear operator w.r.t. $f(x)$, and the interpolation error is always zero if $f(x)$ is a linear polynomial. Then the Bramble Hilbert Lemma implies that this piecewise linear interpolation error is controlled by $|f|_2$, which contains the second order derivative (in the weak sense). On the other hand, we can also get similar conclusion that the piecewise linear interpolation error is dominated by or related to $f''(x)$ through Taylor expansion. So if you prefer, you can think of the Bramble Hilbert Lemma as the better alternative as opposed to performing Taylor expansion.*

Remark 1.11. *The power of the abstraction in the Bramble Hilbert Lemma lies in the fact that we easily extend the interpolation and quadrature error estimates in Section 1.2.2 to unstructured meshes on any shape of domain. Recall that the H^1-norm error estimate is built upon the the interpolation error estimate. This is why the arguments for deriving error estimates in this chapter also apply to any general setup such as problems in multiple dimensions.*

Consider the quadrature error operator, which is linear and also zero for polynomials of certain degree. For instance, if considering the trapezoidal rule for each interval in Figure 1.1, then

$$\int_0^1 f(x)dx - \sum_{i=0}^{N} \frac{1}{2}h[f(x_i)+f(x_{i+1})] = \sum_{i=0}^{N} \left(\int_{x_i}^{x_{i+1}} f(x)dx - \frac{1}{2}h[f(x_i)+f(x_{i+1})] \right).$$

Consider a mapping from the small cell $[x_i, x_{i+1}]$ to the reference cell $[0,1]$ by

$$x = h\hat{x} + x_i, \quad \hat{f}(\hat{x}) = f(h\hat{x} + x_i).$$

Let

$$E_i(f) = \int_{x_i}^{x_{i+1}} f(x)dx - \frac{1}{2}h[f(x_i)+f(x_{i+1})]$$

be the quadrature error on a small inverval, and

$$\hat{E}(\hat{f}) = \int_0^1 \hat{f}(\hat{x})d\hat{x} - \frac{1}{2}[\hat{f}(0)+\hat{f}(1)]$$

be the quadrature error on a reference interval $[0,1]$. Then $\hat{E}$ is the linear operator $\mathcal{L}$ in the Bramble Hilbert Lemma on $\Omega = [0,1]$ and we have

$$|E_i(f)| = h|\hat{E}(\hat{f})| \leq hC|\hat{f}|_2 = hC\sqrt{\int_0^1 \hat{[}f''(\hat{x})]^2 d\hat{x}} = h^{2.5}C\sqrt{\int_{x_i}^{x_{i+1}} [f''(x)]^2 dx}.$$

With Cauchy Schwartz inequality for vectors $\sum_i a_i b_i \leq \sqrt{\sum_i a_i^2}\sqrt{\sum_i b_i^2}$, we get the total quadrature error as

$$\sum_{i=0}^{N} |E_i(f)| \leq Ch^2 \sum_{i=0}^{N} \sqrt{h}\sqrt{\int_{x_i}^{x_{i+1}} [f''(x)]^2 dx} \leq Ch^2 \sqrt{\sum_{i=0}^{N} h}\sqrt{\sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} [f''(x)]^2 dx}$$

$$= Ch^2|f|_2.$$

So we have just proven that

$$\left| \int_0^1 f(x)dx - \sum_{i=0}^{N} \frac{1}{2}h[f(x_i)+f(x_{i+1})] \right| \leq Ch^2|f|_2. \tag{1.17}$$

1.7.3 Error estimates

We only demonstrate the main idea why the error estimates for the abstract finite element method can still hold after quadrature error is involved. We focus on the simplest example. Consider the scheme (1.8) for $a(x) \equiv 1$, i.e., the scheme $\frac{1}{h^2}K\mathbf{u} = \mathbf{f}$. The integrand in the bilinear $\mathcal{A}(u_h, v_h)$ is simply piecewise constant because u_h and v_h are piecewise linear. Thus we have $\mathcal{A}(u_h, v_h) = \mathcal{A}_h(u_h, v_h)$ and the first Strang Lemma reduces to

$$\|u - u_h\|_1 \leq C \inf_{v_h \in V_0^h} \|u - v_h\|_1 + C \sup_{w_h \in V_0^h} \frac{|\langle f, w_h\rangle_h - (f, w_h)|}{\|w_h\|_1}.$$

For a piecewise polynomial w_h, its second order derivative only exists on each interval, thus we abuse the notation by letting $|w_h|_2$ denote (this is usually called Broken Sobolev space, i.e., the Sobolev space on each small interval)

$$|w_h|_2^2 = \sum_i \int_{x_i}^{x_{i+1}} [w_h''(x)]^2 dx.$$

With this modification of seminorm (you can verify that (1.17) still holds if replacing f by w_h), by (1.17), we have

$$|\langle f, w_h\rangle_h - (f, w_h)| \leq Ch^2 |fw_h|_2.$$

Notice that in each interval $(fw_h)'' = (f'w_h + fw_h')' = f''w_h + 2f'w_h'$ because $w''(x) \equiv 0$ within each interval. Thus with Cauchy Schwartz inequality, we have

$$\begin{aligned} |\langle f, w_h\rangle_h - (f, w_h)| &\leq Ch^2 |fw_h|_2 = Ch^2 |f''w_h + 2f'w_h'|_0 \\ &\leq Ch^2 (|f''|_0 |w_h|_0 + 2|f'||w_h'|) \leq Ch^2 \|f\|_2 \|w\|_1. \end{aligned}$$

Therefore we obtain the H^1 estimate as

$$\|u - u_h\|_1 \leq Ch|u|_2 + Ch^2 \|f\|_2.$$

Similarly, the Aubin-Nitsche Lemma also holds up to quadrature error.

The conclusion is very simple: the orders in the estimates in (1.12) and (1.14) still hold in the estimates for the scheme with quadrature (1.8).

1.8 Generalization: general domain in two dimensions

We will have a quick glance at how everything can be easily extended to a general setup. Consider solving a two-dimensional Poisson equation:

$$-\nabla \cdot (A(x, y)\nabla u(x, y)) = f(x, y), \quad (x, y) \in \Omega$$

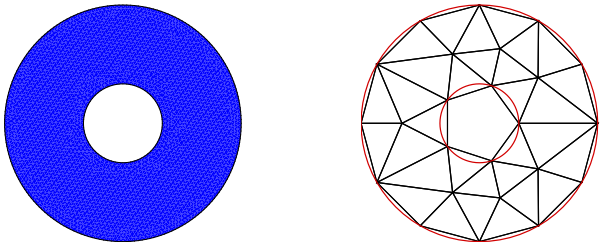


Figure 1.3: Left: the domain Ω. Right: the approximated domain Ω_h via a triangular mesh.

with homogeneous Dirichlet boundary conditions $u(x, y) = 0$ along the domain boundary Ω for a bounded region Ω, where $A(x, y)$ is a 2×2 matrix coefficient.

We just mention some key ingredients in the generalization to see how an easy extension is possible in the first place:

1. Multiplying the test function and integration by parts, we get the equivalent variational formulation for the PDE:

$$\text{seek } u \in H_0^1(\Omega), \quad \iint_\Omega \nabla v^T A \nabla u dxdy = \iint_\Omega f v dxdy, \quad \forall v \in H_0^1(\Omega),$$

   which can be denoted as $\mathcal{A}(u, v) = (f, v)$.

2. Construct an unstructured triangular mesh, which gives an approximated domain Ω_h as shown in Figure 1.3. Notice that the approximated boundary $\partial\Omega_h$ is a piecewise segment approximation to the curved boundary $\partial\Omega$, which induces a second order geometric error thus any finite element method defined on this Ω_h can be at most second order accurate even if using very high order polynomial basis. On the other hand, we can easily fix this issue by using curved triangle along the boundary, but quadrature on curved triangles are more expensive. For simplicity, we just consider the mesh shown in Figure 1.3.

3. We define V_0^h as the continuous piecewise linear polynomial space on the mesh shown in Figure 1.3, with the property of vanishing on $\partial\Omega_h$. An abstract finite element method is natually given as

$$\text{seek } u_h \in V_0^h, \quad \iint_\Omega \nabla v_h^T A \nabla u_h dxdy = \iint_\Omega f v_h dxdy, \quad \forall v_h \in V_0^h,$$

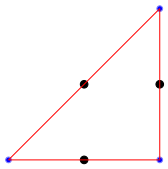
   which can be denoted as $\mathcal{A}(u_h, v_h) = (f, v_h)$.

4. Now let us consider what kind of coefficient $A(x, y)$ can ensure coercivity. For instance, if we assume that A is real symmetric and its smallest eigenvalue has a uniform positive lower bound, i.e., $\lambda(A) \geq C > 0$ for any (x, y), then by the Courant-Fisher-Weyl Min-Max principle,

$$\frac{\nabla v^T A \nabla v}{\nabla v^T \nabla v} \geq C \Rightarrow \mathcal{A}(v, v) \geq C|v|_1^2 \geq C\|v\|_1^2, \quad \forall v \in H_0^1(\Omega).$$

where we have used the fact that H^1-seminorm and H^1-norm are equivalent in $H_0^1(\Omega)$.

5. Assume V_0^h is N-dimensional. We can define *Lagrangian* basis (also called *nodal basis*) functions $\phi_i(x, y)$ on Ω_h just like the one-dimensional case. For instance, a linear polynomial is completely determined by its point values at three vertices on the triangle, and a quadratic polynomial is completely determined by its point values at three vertices and three edge centers on the triangle.



6. Plugging in $u_h(x, y) = \sum_{j=1}^N u_j \phi_j(x)$ we get a linear system

$$\sum_{j=1}^N u_j \mathcal{A}(\phi_j, \phi_i) = (f, \phi_i), \quad i = 1, \cdots, N,$$

and the *stiffness* matrix S has entries $S_{ij} = \mathcal{A}(\phi_j, \phi_i)$.

7. It can be shown that weak partial derivatives of any $v_h \in V_0^h$ exist thus it is conforming: $V_0^h \subset H_0^1(\Omega)$. So the proof of Galerkin Orthogonality holds. Coercivity and Galerkin Orthogonality imply Céa's Lemma. Once we have Céa's Lemma, the H^1-norm error is controlled by the interpolation error, which can be given via Bramble-Hilbert Lemma. Similary, the Aubin-Nitsche Lemma also holds.

8. The quadrature using only three vertices is exact for linear polynomials on a triangle. The quadrature using three vertices and three edge centers is exact for quadratic polynomials on a triangle. With a suitable quadrature, the finite element method can be represented as

$$\text{seek } u_h \in V_0^h, \quad \mathcal{A}_h(u_h, v_h) = \langle f, v_h \rangle_h, \quad \forall v_h \in V_0^h.$$

9. Finally, if you are curious whether this is still a finite difference scheme if using a structured triangular mesh, e.g., one rectangle is splitted into

two triangles in a rectangular mesh, then the answer is yes! With no surprises, the P^1 finite element method will give the same 5-point discrete Laplacian scheme as in Chapter **??**. The P^2 polynomial finite element method for $-\Delta u = f$ gives a fourth order accurate (superconvergence) finite difference scheme with the following stencil:

$$\begin{array}{ccccccccccc} & & & & & & 1 & & & & \\ & & -1 & & & & -4 & & & & \\ \text{edge center} & -1 & 4 & -1 & \text{vertex} & 1 & 12 & -4 & 1 & . \\ & & -1 & & & & -4 & & & & \\ & & & & & & 1 & & & & \end{array}$$

1.9 Generalization: purely Neumann b.c.

Consider a one-dimensional problem

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

$$u'(0) = \sigma_0, u'(1) = \sigma_1.$$

Recall that $f(x)$ must be compatible with the boundary conditions:

$$\int_0^1 f(x)dx = -a_1\sigma_1 + a_0\sigma_0, \tag{1.18}$$

which is obtained by integrating the PDE.

1.9.1 Quotient space $H^1(\Omega)/P^0(\Omega)$

Recall that this boundary value problem does not have a unique solution: if $u(x)$ is a solution, then so is $u(x) + c$ for any constant c. This non-uniqueness issue must be addressed properly. To this end, it is natural to consider a quotient space in which two functions differing by only a constant are regarded as the same function.

Let $P^0(\Omega)$ be the linear space of all polynomials of degree zero, i.e., all constants. We first introduce an *equivalent* class by

$$\dot{v}(x) := \{w(x) = v(x) + c, \quad c \in P^0(\Omega)\}.$$

In other words, if two functions $v(x)$ and $w(x)$ are different only by a constant c, then we regard them to be in the same equivalent class, which is a set. Any element $w(x)$ in an equivalent class $\dot{v}(x)$ is called a representation of the equivalent class $\dot{v}(x)$. For instance, in this section, $v(x)$ means a representation of the equivalent class $\dot{v}(x)$ which $v(x)$ belongs to.

The quotient space $H^1(\Omega)/P^0(\Omega)$ is defined as

$$H^1(\Omega)/P^0(\Omega) = \{\dot{v}(x) : v(x) \in H^1(\Omega)\}.$$

The norm the quotient space $H^1(\Omega)/P^0(\Omega)$ is defined as

$$\|\dot{v}\|_1 := \inf_{w\in\dot{v}} \|w\|_1,$$

where $\|w\|_1$ is the H^1-norm of the representation $w(x)$. This definition can be explicitly written as

$$\|\dot{v}\|_1 := \inf_{c\in P^0(\Omega)} \|v(x)+c\|_1 = \min_{c\in\mathbb{R}} \sqrt{\int_\Omega |v(x)+c|^2 dx + \int_\Omega \left|\frac{d}{dx}(v(x)+c)\right|^2 dx}.$$

So we get

$$\|\dot{v}\|_1^2 := \min_{c\in\mathbb{R}} \int_\Omega |v(x)+c|^2 dx + \int_\Omega |v'(x)|^2 dx,$$

which is nothing but a minimization with respect to c. Also, it is a simple quadratic function of the number c, so the minimizer is the average of $v(x)$, $c = \frac{1}{|\Omega|}\int_\Omega v(x)dx$. For the domain $\Omega = (0,1)$, let $\bar{v} = \int_0^1 v(x)dx$ be the average of the function $v(x)$ over Ω. Then the quotient space $H^1(\Omega)/P^0(\Omega)$ can be equivalently written as

$$\|\dot{v}\|_1^2 = \int_\Omega |v(x)-\bar{v}|^2 dx + \int_\Omega |v'(x)|^2 dx.$$

This quotient space norm is also equivalent to the seminorm $|v|_1$:

$$C\|\dot{v}\|_1 \le |v|_1 \le \|\dot{v}\|_1, \quad C>0.$$

The first inequalilty is true because of the following *Poincaré inequalilty* (see Appendix for a generic statement):

$$\int_\Omega |v(x)-\bar{v}|^2 dx \le C\int_\Omega |v'(x)|^2 dx.$$

1.9.2 Variational formulation and coercivity

Multiplying a test function and intergration by parts, we can get a variational form:

$$\int_0^1 a(x)u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx + a_1\sigma_1 v(1) - a_0\sigma_0 v(0).$$

Obviously, both side stay the same if we replace $u(x)$ by $u(x)+c$ for any constant c. Now if we replace $v(x)$ by $v(x)+c$, the left hand side stays the same, and the right hand side also stays the same because of the compatibility condition (1.18).

So the equivalent variational formulation is to seek $\dot{u} \in H^1(\Omega)/P^0(\Omega)$ such that

$$\int_\Omega a(x)u'(x)v'(x)dx = \int_\Omega f(x)v(x)dx + a_1\sigma_1 v(1) - a_0\sigma_0 v(0), \quad \forall \dot{v} \in H^1(\Omega)/P^0(\Omega).$$

It can be denoted by the same short hand notation as:

$$\mathcal{A}(u, v) = (f, v) + a_1\sigma_1 v(1) - a_0\sigma_0 v(0), \quad \forall \dot{v} \in H^1(\Omega)/P^0(\Omega).$$

The Cauchy-Schwartz inequalilty implies the continuity of the bilinear form. Since quotient space norm is also equivalent to the H^1 seminorm, we also have the coercivity:

$$\mathcal{A}(v, v) \geq C\|\dot{v}\|_1, \quad \forall v \in H^1(\Omega)/P^0(\Omega).$$

1.9.3 The finite element method

On a mesh with intervals I_j, we define the space V^h as an approximation to $H^1(0, 1)$:

$$V^h = \{v_h(x) \in C(0, 1) : v_h(x) \in P^k(I_j), \forall j\}.$$

We can also define a quotient space V^h/P^0 similarly:

$$V^h/P^0 = \{\dot{v}_h(x) : v_h(x) \in V^h\}.$$

The finite element method is to seek $\dot{u}_h(x) \in V^h/P^0$ such that

$$\mathcal{A}(u_h, v_h) = (f, v_h) + a_1\sigma_1 v_h(1) - a_0\sigma_0 v_h(0), \quad \forall \dot{v}_h \in V^h/P^0.$$

Notice that we use representations u_h and v_h in the bilinear form $\mathcal{A}(u_h, v_h)$, instead of their equivalent classes $\dot{u}_h$ and $\dot{v}_h$. All the previous arguments for error estimates can be established similarly, and the only difference is that the underlying function space is the quotient space $H^1(\Omega)/P^0(\Omega)$, even though we just plug in functions into the variational form as before.

1.9.4 Coercivity implies the stiffness matrix null space

For simplicity, we assume homogeneous Neumann boundary condition $\sigma_0 = \sigma_1 = 0$, and constant coefficient $a(x) = 1$. Then for the P^1 basis finite element method, the bilinear form with trapezoidal quadrature $\mathcal{A}_h(u_h, v_h)$ is the same as $\mathcal{A}(u_h, v_h)$.

Recall our uniform grid points are

$$0 = x_0 < x_1 < \cdots < x_N < x_{N+1} = 1.$$

Let $\phi_i(x), i = 0, 1, \cdots, N + 1$ be the Lagrangian basis or *nodal basis* of V^h. Then the stiffness matrix $S \in \mathbb{R}^{(N+2)\times(N+2)}$ has entries $S_{ij} = \mathcal{A}_h(\phi_j, \phi_i) = \mathcal{A}(\phi_j, \phi_i)$. Here we have abused notation by allowing indices i, j to take value 0.

With similar notation as before, e.g., $\mathbf{v} \in \mathbb{R}^{N+2}$ denoting a vector of point values $v_h(x_j)$, we have

$$\mathbf{v}^T S\mathbf{v} = \mathcal{A}(v_h, v_h) \geq C\|\dot{v}_h\|_1 \geq 0,$$

thus S is still real symmetric and positive semi-definite.

Since the boundary value problem does not have a unique solution, the stiffness matrix S must have a nontrival null space. As a matter of fact, the constant one vector $\mathbf{1}$ is in its null space. We first have

$$\forall \mathbf{v}, \quad \mathbf{v}^T S\mathbf{1} = \mathcal{A}(1, v_h) = 0 \Rightarrow \mathbf{v} \perp S\mathbf{1}, \quad \forall \mathbf{v} \Rightarrow S\mathbf{1} = \mathbf{0}.$$

Next, we want to show that the coercivity implies that the null space of S is one-dimensional:

$$S\mathbf{v} = \mathbf{0} \Leftrightarrow \mathbf{v}^T S\mathbf{v} = 0 \Leftrightarrow \mathcal{A}(v_h, v_h) = 0 \Rightarrow \|\dot{v}_h\|_1 = 0,$$

where the last step is due to the coercivity. Thus

$$S\mathbf{v} = \mathbf{0} \Rightarrow \|\dot{v}_h\|_1 = 0 \Leftrightarrow \dot{v}_h(x) = \dot{0} \Leftrightarrow v_h(x) \equiv c \Leftrightarrow \mathbf{v} = c\mathbf{1},$$

because a function in the quotient space has zero norm if and only if it is $\dot{0}$, which is the property of a norm.

1.9.5 The finite difference form

For simplicity, just consider the constant coefficient case $a(x) = 1$, for piecewise linear basis with trapezoidal quadrature on the uniform grid, the finite element method can be equivalently written as

$$\begin{aligned}
&\frac{1}{h}(u_1 - u_0) = \frac{h}{2} f_0 + a_0\sigma_0 \\
&\frac{1}{h}(-u_{j-1} + 2u_j - u_{j+1}) = h f_j, \quad j = 1, \cdots, N \\
&\frac{1}{h}(u_{N+1} - u_N) = \frac{h}{2} f_{N+1} + a_1\sigma_1
\end{aligned}$$

which is exactly the same as the traditional finite difference scheme in Section **??**.

Now the finite element theory can give error estimates like (1.12) and (1.14). On the other hand, it is straightforward to check the truncation error at x_0 or x_{N+1} is only first order, even though the Neumann boundary condition was approximated by a second order centered difference in Section **??**. It is quite difficult to show that this scheme is second order accurate following arguments in Chapter **??**, epsecially for a variable coefficient problem in multiple dimensions. But we know this scheme is indeed second order accurate in the sense of (1.14), which demonstrates the superiority of the finite element method compared to traditional finite difference method.

For $a(x) = 1$, the stiffness matrix S with entries $\mathcal{A}_h(\phi_j, \phi_i)$ can be written out as

$$S = \frac{1}{h}\begin{pmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{pmatrix}_{(N+2)\times(N+2)}.$$

In general, we have

$$S = \frac{1}{h}\frac{1}{2}\begin{pmatrix} a_0+a_1 & -a_0-a_1 & & & & \\ -a_0-a_1 & a_0+2a_1+a_2 & -a_1-a_2 & & & \\ & -a_1-a_2 & a_1+2a_2+a_3 & -a_2-a_3 & & \\ & & & & & \\ & & & \ddots & \ddots & \ddots \end{pmatrix}_{(N+2)\times(N+2)}.$$

1.9.6 How to solve the singular linear system

Now consider the finite element method with quadrature for a variable coefficient problem:

$$\mathcal{A}_h(u_h, v_h) = \langle f, v_h\rangle_h + a_1\sigma_1 v_h(1) - a_0\sigma_0 v_h(0), \quad \forall \dot{v}_h \in V^h/P^0.$$

For piecewise linear finite element method, if using a Lagrangian basis or *nodal basis* $\phi_j(x)$, by plugging in $v_h = \phi_i(x)$, we get a singular linear system

$$S\begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_N \\ u_{N+1} \end{pmatrix} = h\begin{pmatrix} \frac{1}{2}f_0 \\ f_1 \\ \vdots \\ f_N \\ \frac{1}{2}f_{N+1} \end{pmatrix} + \begin{pmatrix} a_0\sigma_0 \\ 0 \\ \vdots \\ 0 \\ a_1\sigma_1 \end{pmatrix}.$$

Since the null space of S is $Span\{\mathbf{1}\}$, the least square solution of this singular linear system corresponds to the facts that the finite element method is defined on a quotient space V^h/P^0, and we look for a solution in a quotient space $H^(\Omega)/P^0(\Omega)$ in the original variational formulation. We do not elaborate more about it here, but at least all discussions about quotient spaces and the null spaces are consistent.

Let $S\mathbf{u} = \tilde{f}$ denote this singular linear system. The first question we should ask is whether it has a solution since S is not invertible. A linear system $S\mathbf{u} = \tilde{f}$ if and only if $\tilde{f}$ is in the column space of S, denoted by $Col(S)$. To this end, we need the orthogonal complement of $Col(S)$, denoted by $Col(S)^\perp$:

$$Col(S)^\perp = \{\mathbf{y} \in \mathbb{R}^{N+2} : \mathbf{y}^T S = \mathbf{0}.\}$$

Since S is symmetric, $\mathbf{y}^T S = \mathbf{0} \Leftrightarrow S\mathbf{y} = \mathbf{0}$, thus $Col(S)^\perp$ is the null space of S. In particular, we know that $\mathbf{1}$ is the basis of $Col(S)^\perp$. We have

$$\tilde{f} \in Col(S) \Leftrightarrow \tilde{f} \perp Col(S)^\perp \Leftrightarrow \tilde{f} \perp \mathbf{1}$$

$$\Leftrightarrow \frac{1}{2}hf_0 + h\sum_{j=1}^{N} f_j + \frac{1}{2}hf_{N+1} + a_0\sigma_0 + a_1\sigma_1 = 0$$

which is nothing but a discrete compatibility condition.

For a function $f(x)$ satisfying the compatibility condition, its point values may not necessarily satisfy the discrete compatibility condition. We can simply project $\tilde{f}$ to the column space of S. Let $\bar{f}$ be the projection vector, then $S\mathbf{u} = \bar{f}$ is ensured to have a solution, and we can use iterative solvers in Chapter **??** such as conjugate gradient method or its preconditioned version directly on $S\mathbf{u} = \bar{f}$ to find the least square solution to $S\mathbf{u} = \tilde{f}$. Since we know what $Col(S)^\perp$ is, the projection $\bar{f}$ is quite easy to find. We summerize it as follows:

1. The projection $\bar{f}$ is computed as

$$\bar{f} = \tilde{f} - \frac{\langle \mathbf{1}, \tilde{f} \rangle}{\|\mathbf{1}\|^2}\mathbf{1}.$$

   It is easy to verify $\langle \mathbf{1}, \bar{f} \rangle = 0$.

2. Solve $S\mathbf{u} = \bar{f}$ by direct or iterative solvers. See Chapter **??**.

Remark 1.12. *Iterative solvers like conjugate gradient may not work well directly on $S\mathbf{u} = \tilde{f}$ epsecially if the discrete compatibility error is large.*

Remark 1.13. *To find the least square solution to $S\mathbf{u} = \tilde{f}$, it is mathematically equivalent to solve the normal equation $S^T S\mathbf{u} = S^T\tilde{f}$ which is ensured to have a solution for any $\tilde{f}$. However, $S^T S\mathbf{u} = S^T\tilde{f}$ is much harder to solve. For example, if S is invertible, then the condition number of $S^T S$ is about the square of the condition number of S.*

Remark 1.14. *If S is not symmetric, in order to find the projection $\bar{f}$, we need to compute the left null vector $\mathbf{y}$ first: solving $S^T\mathbf{y} = \mathbf{0} = 0 * \mathbf{y}$ is an eigenvector problem, which is much more expensive than solving a linear system of the same size. For instance, for a nonsingular system $A\mathbf{x} = b$, iterative solvers are based on minimizing a function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - b^T\mathbf{x}$. For $A\mathbf{x} = 0$, if minimizing $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x}$, we simply get $\mathbf{x} = \mathbf{0}$, which is a solution that we do not want at all. For getting the nonzero solution to $A\mathbf{x} = 0$, roughtly speaking, we would have to minimize $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x}$ over the sphere $\{\mathbf{x} : \|\mathbf{x}\| = 1\}$.*

Remark 1.15. *For all remarks above, it is highly desired to have a symmetric S. The symmetry of the matrix S with entries $S_{ij} = \mathcal{A}_h(\phi_j, \phi_i) = \mathcal{A}_h(\phi_i, \phi_j)$ holds trivially even for a two-dimensional or three-dimensional problem $-\nabla \cdot (A\nabla u) = f$ with a real symmetric matrix coefficient A. This is one of the key advantages of using finite element method for purely Neumann boundary conditions. It is in general quite difficult to construct a real symmetrix matrix for variable coefficient problems with Neumann boundary in multiple dimensions by traditional finite difference method.*

1.10 Generalization: nonhomogeneou Dirichlet b.c.

Consider solving

$$-u''(x) = f(x), \quad x \in (0,1),$$
$$u(0) = \sigma_0, u(1) = \sigma_1.$$

The standard approach is to assume that there exists a smooth enough function $g(x)$ satisfying the same boundary condition. Then the function $\tilde{u} = u - g$ satisfies

$$-\tilde{u}''(x) = f(x) + g''(x), \quad x \in (0,1),$$
$$\tilde{u}(0) = \tilde{u}(1) = 0.$$

Obviously everything in Section 1.3 can be easily applied to construct and anaylze a finite element method for $\tilde{u} \in H_0^1(\Omega)$, provided that we know what $g(x)$ is, which is easy to construct in one-dimension but not necessarily in multiple dimensions.

However, we only need to know existence of the smooth function $g(x)$ and an actual implementation can be made irrevelant to what exactly $g(x)$ should be. The same order from the L^2-norm error estimate (1.14) can still hold.

By multiplying a test function $v \in H_0^1(\Omega)$ and intergration by parts, we get the equivalent variational form for seeking $\tilde{u} \in H_0^1(\Omega)$ satisfying

$$\int_0^1 \tilde{u}'(x)v'(x)dx = \int_0^1 f(x)v(x)dx - \int_0^1 g'(x)v'(x)dx, \quad \forall v \in H_0^1(\Omega),$$

which can be denoted as

$$\mathcal{A}(\tilde{u}, v) = (f, v) - \mathcal{A}(g, v), \quad \forall v \in H_0^1(\Omega).$$

1.10.1 A scheme in theory

An abstract finite element method that we should never implement is to find $\tilde{u}_h \in V_0^h$ satisfying

$$\mathcal{A}(\tilde{u}_h, v_h) = (f, v_h) - \mathcal{A}(g, v_h), \quad \forall v_h \in V_0^h.$$

Assume $g(x)$ is a nice function so that we can still derive the error estimates (1.12) and (1.14). For example, if $g''(x)$ exists, then after integration by parts for test function $v_h(x) \in V_0^h$, the abstract finite element is equivalent to seeking $\tilde{u}_h \in V_0^h$ satisfying

$$\mathcal{A}(\tilde{u}_h, v_h) = (f + g'', v_h), \quad \forall v_h \in V_0^h.$$

If we treat $f - g''$ as the right hand side function, then the error estimates (1.12) and (1.14) can still hold for $\tilde{u}_h - \tilde{u}$.

The numerical solution that we want is

$$u_h := \tilde{u}_h + g(x).$$

Be careful that we no longer have $u_h \in V^h$. By moving $\mathcal{A}(g, v_h)$ to the left hand side, we get

$$\mathcal{A}(u_h, v_h) = (f, v_h), \quad \forall v_h \in V_0^h.$$

Also $u_h - u$ satisfies the error estimates (1.12) and (1.14).

Next, assume we use quadrature, so we have

$$\mathcal{A}_h(\tilde{u}_h, v_h) = \langle f, v_h \rangle_h - \mathcal{A}_h(g, v_h), \quad \forall v_h \in V_0^h, \tag{1.19}$$

or equivalently

$$\mathcal{A}_h(u_h, v_h) = \langle f, v_h \rangle_h, \quad \forall v_h \in V_0^h. \tag{1.20}$$

Assume the estimates (1.12) and (1.14) still hold after using quadrature for the scheme (1.19).

1.10.2 A scheme for implementation

We consider the piecewise linear Lagrangian interpolation polynomial for $g(x)$ at grid points x_i, denoted by $g_h(x) = \Pi_1 g(x) \in V^h$. For *nodal basis* $\{\phi_j(x)\}_{j=0}^{N+1}$ of V^h, we simply have

$$g_h(x) = \sum_{j=0}^{N+1} g_j \phi_j(x) \in V^h,$$

where $g_0 = \sigma_0, g_{N+1} = \sigma_1, x_0 = 0, x_{N+1} = 1$. Then we consider a new scheme seeking $\tilde{u}_h \in V_0^h$ satisfying

$$\mathcal{A}_h(\tilde{u}_h, v_h) = \langle f, v_h \rangle_h - \mathcal{A}_h(g_h, v_h), \quad \forall v_h \in V_0^h. \tag{1.21}$$

The difference between the scheme (1.21) and the scheme (1.19) is where using $g(x)$ or its polynomial interpolation $g_h(x)$.

Let $u_h(x) = \tilde{u}_h(x) + g_h(x) \in V^h$, then we can rewrite (1.21) equivalently as

$$\mathcal{A}_h(u_h, v_h) = \langle f, v_h \rangle_h, \quad \forall v_h \in V_0^h. \tag{1.22}$$

This time, since $u_h(x) = \tilde{u}_h(x) + g_h(x) \in V^h$, we have

$$u_h(x) = \sum_{j=0}^{N+1} u_j \phi_j(x),$$

where $u_i = u_h(x_i)$.

Here I need to emphasize that V^h is $(N+2)$-dimensional with basis $\{\phi_j(x)\}_{j=0}^{N+1}$, whereras the test function space V_0^h is only N-dimensional with basis $\{\phi_j(x)\}_{j=1}^{N}$.

Obviously, plugging this representation into (1.22) and test function space basis $\phi_i(x)$ for $i = 1, \cdots, N$, we get a linear system

$$\sum_{j=0}^{N+1} \mathcal{A}_h(\phi_j, \phi_i) u_j = h f_j, \quad \forall i = 1, \cdots, N.$$

Of course the linear system should have only N unknowns because of Dirichlet boundary $u_0 = \sigma_0$ and $u_{N+1} = \sigma_1$. The scheme is precisely

$$\frac{1}{h}(-u_{j-1} + 2u_j - u_{j+1}) = h f_j, \quad j = 1, \cdots, N, \tag{1.23}$$

where $u_0 = \sigma_0, u_{N+1} = \sigma_1$.

Remark 1.16. *Notice that the scheme* (1.22) *is equivalent to the following scheme seeking* $u_h(x) \in V_0^h$ *satisfying*

$$A_h(u_h, v_h) = \langle f, v_h \rangle_h - A_h(\sigma_h, v_h), \quad \forall v_h \in V_0^h, \tag{1.24}$$

where $\sigma_h \in V^h$ *is the Lagrangian interpolation of the trivial nonsmooth extension function:*

$$\sigma_h(x_0) = \sigma_0, \sigma_h(x_{N+1}) = \sigma_1, \quad \sigma_h(x_i) = 0, i = 1, \cdots, N.$$

Remark 1.17. *The scheme* (1.22) *or* (1.24) *has nothing to do with what* $g(x)$ *is. On the other hand, with the existence of smooth* $g(x)$*, the error estimates can be easily established via the analysis of the scheme* (1.19)*. To establish error estimates for* (1.22) *or* (1.24)*, notice that their only difference from* (1.19) *is the following*

$$\mathcal{A}_h(g, v_h) - \mathcal{A}_h(g_h, v_h),$$

which can be analyzed through the interpolation error estimates on $\|g - g_h\|_1$ *and* $\|g - g_h\|_0$*. For instance, for convergence in* H^1 *norm, similar to the First Strang Lemma Theorem 1.5, we will have to deal with*

$$\sup_{w_h \in V_0^h} \frac{|\mathcal{A}_h(g, w_h) - \mathcal{A}_h(g_h, w_h)|}{\|w_h\|_1},$$

which can be easily done by discrete continuity of the bilinear form:

$$\frac{|\mathcal{A}_h(g, w_h) - \mathcal{A}_h(g_h, w_h)|}{\|w_h\|_1} \leq C \|g - g_h\|_1.$$

The scheme (1.23) is exactly the same as taking the scheme for purely Neumann boundary at interior grid points $j = 1, \cdots, N$ in Section 1.9. This is not a coincidence at all. This fact remains true even for high order polynomial basis with variable coefficients, which means that we have a neat treatment of boundary condition in finite element method. In particular, for a variable coefficient problem, by taking the scheme at interior grid points $j = 1, \cdots, N$ in Section 1.9, we obtain the P^1 finite element method with trapezoidal quadrature for the nonhomogeneous Dirichlet boundary as

$$\frac{-(a_{j-1} + a_j)u_{j-1} + (a_{j-1} + 2a_j + a_{j+1})u_j - (a_j + a_{j+1})u_{j+1}}{2h} = hf_j, \quad j = 1, \cdots, N,$$

where $u_0 = \sigma_0, u_{N+1} = \sigma_1$.

1.10.3 A scheme in theory for 2D general domain Ω

Consider solving a two-dimensional Poisson equation with nonhomogeneous Dirichlet boundary condition for a bounded region Ω:

$$-\nabla \cdot (A(x, y)\nabla u(x, y)) = f(x, y), \quad (x, y) \in \Omega,$$

$$u(x, y) = \sigma(x, y), \quad (x, y) \in \partial\Omega.$$

where $A(x, y)$ is a 2×2 matrix coefficient.

Assume there exists a smooth extension function $g(x, y)$ satisfying that $g|_{\partial\Omega}(x, y) = \sigma(x, y)$, then $\tilde{u} = u - g \in H_0^1(\Omega)$ satisfying

$$\mathcal{A}(\tilde{u}, v) = (f, v) - \mathcal{A}(g, v), \quad \forall v \in H_0^1(\Omega)$$

where the bilinear form is $\mathcal{A}(u, v) = \iint_\Omega \nabla v^T A \nabla u dx dy$.

Given a triangulation of the domain Ω_h as shown in (1.3), assume either Ω is polygonal or we use curved triangles, so that $\partial\Omega_h = \partial\Omega$. Define a continuous piecewise polynomial space $V_0^h \subset H_0^1(\Omega)$, then an abstract finite element method that can be easily analyzed is to find $\tilde{u}_h \in V_0^h$ satisfying

$$\mathcal{A}(\tilde{u}_h, v_h) = (f, v_h) - \mathcal{A}(g, v_h), \quad \forall v_h \in V_0^h.$$

The scheme with quadrature is written as

$$\mathcal{A}_h(\tilde{u}_h, v_h) = \langle f, v_h \rangle_h - \mathcal{A}_h(g, v_h), \quad \forall v_h \in V_0^h.$$

or equivalently

$$u_h = \tilde{u}_h + g, \quad \mathcal{A}_h(u_h, v_h) = \langle f, v_h \rangle_h, \quad \forall v_h \in V_0^h. \tag{1.25}$$

1.10.4 A scheme for implementation for 2D general domain Ω

The error estimates of (1.25) can be easily established. For the ease of implementation, we define $g_h(x)$ as the Lagrangian interpolation of $g(x)$ over nodal points in the mesh, which will be explained below.

Then we implement a different scheme

$$\tilde{u}_h \in V_0^h, \quad \mathcal{A}_h(\tilde{u}_h, v_h) = \langle f, v_h \rangle_h - \mathcal{A}_h(g_h, v_h), \quad \forall v_h \in V_0^h.$$

or equivalently

$$u_h = \tilde{u}_h + g_h \in V^h, \quad \mathcal{A}_h(u_h, v_h) = \langle f, v_h \rangle_h, \quad \forall v_h \in V_0^h. \tag{1.26}$$

For convenience, let $\mathbf{x}$ denote (x, y). Now we need to make some assumptions which are quite practical at least for P^1 and P^2:

I. V_0^h is N-dimensional and V^h is $(N+n)$-dimensional.

II. V^h has a Lagrangian basis (*nodal basis*) $\{\phi_j(\mathbf{x})\}_{j=1}^{N+n}$ satisfying

$$\phi_j(\mathbf{x}_i) = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases},$$

for the points $\mathbf{x}_i : i = 1, \cdots, N+n$.

III. V_0^h has a Lagrangian basis $\{\phi_j(\mathbf{x})\}_{j=1}^{N}$ satisfying

$$\phi_j(\mathbf{x}_i) = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases},$$

for the points $\mathbf{x}_i, i = 1, \cdots, N$. For example, $\mathbf{x}_i$ are three vertices of all triangles for a continuous piecewise linear polynomial on a triangular mesh. For a continuous piecewise quadratic polynomial on a triangular mesh, $\mathbf{x}_i$ are three vertices and three edge centers of all triangles.

IV. The quadrature points used in $\mathcal{A}_h(\cdot, \cdot)$ is a subset of $\{\mathbf{x}_i, i = 1, \cdots, N\}$. For instance, the quadrature using three vertices with equal weight is exact for integrating a linear polynomial on a triangle thus second order accurate by Bramble-Hilbert Lemma, and the quadrature using only three edge centers with equal weight is exact for integrating a quadratic polynomial on a triangle thus third order accurate by Bramble-Hilbert Lemma.

So the points $\{\mathbf{x}_i\}_{i=1}^{N}$ are interior points inside the domain Ω and the points $\{\mathbf{x}_i\}_{i=N+1}^{N+n}$ are boundary points, along the boundary $\partial\Omega_h = \Omega$ (not true in general but we assumed it).

Let $u_j = u_h(\mathbf{x}_j)$ and $\sigma_j = \sigma(\mathbf{x}_j)$, then

$$u_h(\mathbf{x}) = \sum_{j=1}^{N+n} u_j\phi_j(\mathbf{x}) = \sum_{j=1}^{N} u_j\phi_j(\mathbf{x}) + \sum_{j=N+1}^{N+n} \sigma_j\phi_j(\mathbf{x}).$$

Under these assumptions, the scheme (1.26) is exactly the same as

$$\mathcal{A}_h(\sum_{j=1}^{N+n} u_j\phi_j, v_h) = \langle f, v_h\rangle_h, \quad \forall v_h \in V_0^h. \tag{1.27}$$

or equivalently

$$\sum_{j=1}^{N} \quad \mathcal{A}_h(\phi_j, \phi_i)u_j = \langle f, \phi_i\rangle_h - \sum_{j=N+1}^{N+n} \mathcal{A}_h(\phi_j, \phi_i)\sigma_j, \quad i = 1, \cdots, N.$$

If you ever wonder what the simplest boundary treatment for a high order accurate scheme should be, (1.26) gives a perfect answer.

To establish the convergence in H^1-norm and L^2-norm for the scheme (1.26) or (1.27), we first can have the error estimates for (1.25), then analyze the only difference between (1.26) and (1.25):

$$\mathcal{A}_h(g, v_h) - \mathcal{A}_h(g_h, v_h),$$

which is related to the interpolation error estimates on $\|g - g_h\|_1$ and $\|g - g_h\|_0$.

1.10.5 The error in the 2-norm over grid point values

Obviously, the implementation in previous subsection has absolutely nothing to do with what $g(x)$ is. As a matter of fact, the implementation (1.23) is our classical finite difference scheme. But there is still one catch that I have not mentioned, for implementing the finite element method as a finite difference scheme.

To be specific, in (1.23), we can only get point values of $u_h(x)$ at x_j, even though in practice we are quite happy with that already. On the other hand, if we have $g(x)$ and we solve (1.19), then we get $u_h(x) = \tilde{u}_h(x) + g(x)$ for any $x \in (0, 1)$.

In terms of the error estimates, the L^2-norm (1.14) measures the error for all x in the interval $(0, 1)$. In the scheme (1.23), since we only have $u_h(x_j)$, the errors can be measured only at these grid points. For P^1 finite element method, it is straightforward to show that (1.14) for $u_h(x)$ implies the scheme (1.23) is second order accurate in the 2-norm:

$$\|\mathbf{e}\|_2 = \sqrt{h\sum_{j=1}^{N} e_j^2} = \sqrt{h\sum_{j=1}^{N} |u_j - u(x_j)|^2}.$$

The 2-norm above is an approximation to L^2-norm error $\|e_h\|_0$ by the trapezoidal quadrature for the error $e_h = u_h - u$:

$$\|e_h\|_0 = \sqrt{\int_0^1 |e_h(x)|^2 dx},$$

where $e_h(0) = e_h(1) = 0$ because $u_h(x)$ satisfies the boundary condition.

Remark 1.18. *For P^k basis finite element with $k \geq 2$, the error order for function values at (k+1)-point Gauss-Lobatto quadrature points are $(k+2)$-th order in the 2-norm. This one order higher phenomenon is called superconvergence of function values. We can use finite element method with quadratic polynomial to get a fourth order accurate finite difference scheme! Of course it can no longer be derived from L^2-norm estimate* (1.14)*, which is only third order accurate for P^2.*

1.11 Generalization: a general elliptic operator

Next, we consider an elliptic equation in the following form

$$-(a(x)u'(x))' + b(x)u'(x) + c(x)u(x) = f(x), \quad x \in (0,1), \qquad u(0) = u(1) = 0,$$

where $a(x) \geq \min_x a(x) > 0$ and $c(x) \geq 0$.

The variational form is still $\mathcal{A}(u,v) = (f,v)$ where

$$\mathcal{A}(u,v) = \int_0^1 a(x)u'(x)v'(x) + b(x)u'(x)v(x) + c(x)u(x)v(x)dx.$$

First of all, unless $b(x) = 0$, we lose the symmetry of the bilinear form, and $\mathcal{A}(u,v) = \mathcal{A}(v,u)$ is not true in general. Thus the stiffness matrix will no longer be symmetric. But other than this, almost everything above can be extended, under suitable assumptions.

For simplicity, we will just focus on how to establish the coercivity. Since $c(x) \geq 0$, we have

$$\mathcal{A}(v,v) \geq \int_0^1 a(x)|v'(x)|^2 dx + \int_0^1 b(x)v'(x)v(x)dx.$$

For the second order derivative term, recall that by *Poincaré inequality* we have

$$\int_0^1 a(x)|v'(x)|^2 dx \geq \min_x a(x)|v|_1^2 \geq C \min_x a(x)\|v\|_1^2, \quad \forall v \in H_0^1(\Omega),$$

where the constant C depends only on Ω.

For the first order derivative, after integration by parts, we get

$$\int_0^1 b(x)v'(x)v(x)dx = \int_0^1 b(x)\frac{d}{dx}\frac{v^2(x)}{2}dx = -\int_0^1 b'(x)\frac{v^2(x)}{2}dx, \quad \forall v \in H_0^1(\Omega). \tag{1.28}$$

In two dimensions, for a first order derivative term like $\mathbf{b} \cdot \nabla u$, after integration by parts, we have

$$\iint_\Omega (\mathbf{b} \cdot \nabla v) v d\mathbf{x} = -\iint_\Omega \frac{v^2}{2} (\nabla \cdot \mathbf{b}) d\mathbf{x}, \quad \forall v \in H_0^1(\Omega). \tag{1.29}$$

So we can get the *coercivity* $\mathcal{A}(v, v) \geq C\|v\|_1^2$ under the following assumptions:

1. If $b'(x) \equiv 0$, then the term in (1.28) is gone. In two dimensions, if $\nabla \cdot \mathbf{b} \equiv 0$, i.e., $\mathbf{b}$ is *incompressible*, then the term in (1.29) is gone in two dimensions.

2. If $b'(x) \leq 0$ in one dimension or $\nabla \cdot \mathbf{b} \leq 0$ in two dimensions, then we have
$$\mathcal{A}(v, v) \geq \int_0^1 a(x)|v'(x)|^2 dx \geq C\|v\|_1^2.$$

3. If $b'(x) \geq 0$, then we have to assume $\max_x b'(x) < 2C \min_x a(x)$ where C is the constant in the *Poincaré inequality*, thus
$$\mathcal{A}(v, v) \geq \min_x a(x) C\|v\|_1^2 - \max_x b'(x) \frac{1}{2}\|v\|_0^2 \geq (C \min_x a(x) - \frac{1}{2} \max_x b'(x))\|v\|_1^0.$$

Remark 1.19. *For the case $b'(x) \geq 0$, obviously we need the diffusion term $-(au')'$ to be strong enough to dominate the convection term bu'. However, if the diffusion coefficient is very small compared to $b'(x)$, then the coercivity will be lost, thus all arguments in finite element theory based on coercivity will also break down. In pratice, this reflects on the difficulties of using finite element theory to construct a scheme for convection dominated problems, e.g., $\max_x b'(x) >> \max_x a(x)$ or $a(x)$ is nearly zero.*

1.12 Generalization: higher order accuracy via P^2

We only discuss the constant coefficient case. If you are interested, you can find the variable coefficient case in [3].

1.12.1 Dirichlet b.c.

Let V^h and V_0^h denote the corresponding spaces of continuous piecewise quadratic polynomial was shown in Figure 1.2. The difference between V^h and V_0^h is that elements in V_0^h are always zero on the boundary.

The scheme (1.8) with piecewise quadratic basis and Simpson's quadrature (3-point Gauss-Lobatto quadrature) has a matrix form $S\mathbf{u} = M\mathbf{f}$ where

S is stiffness matrix and M is the *lumped mass matrix*. On a uniform mesh, it can be written as $H\mathbf{u} = \mathbf{f}$ with

$$H = M^{-1}S = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & & & & \\ -2 & \frac{7}{2} & -2 & \frac{1}{4} & & & & \\ & -1 & 2 & -1 & & & & \\ & \frac{1}{4} & -2 & \frac{7}{2} & -2 & \frac{1}{4} & & \\ & & & -1 & 2 & -1 & & \\ & & & & \ddots & \ddots & & \\ & & & \frac{1}{4} & -2 & \frac{7}{2} & -2 & \\ & & & & & -1 & 2 & \end{pmatrix}$$

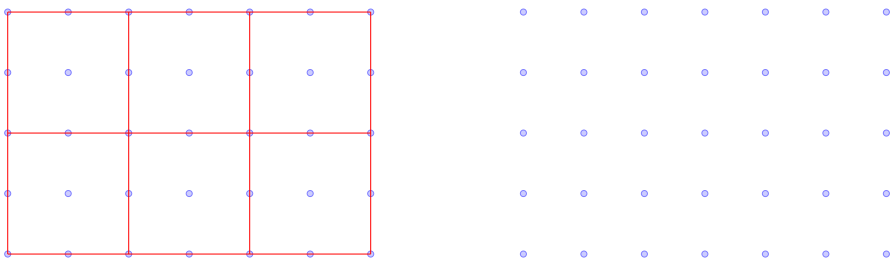
For the nonhomogeneou Dirichlet boundary value problem, the scheme (1.25) can be written as

$$\frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2} = f_i, \quad \text{if } x_i \text{ is a mid point,} \tag{1.30a}$$

$$\frac{u_{i-2} - 8u_{i-1} + 14u_i - 8u_{i+1} + u_{i+2}}{4h^2} = f_i, \quad \text{if } x_i \text{ is a cell end,} \tag{1.30b}$$

where $u_0 = \sigma_0$ and $u_{N+1} = \sigma_1$

Remark 1.20. *The finite difference scheme* (1.30) *can only be defined on a grid with odd number of grid points, because it is obtained by taking end points and midpoints in a finite element mesh as shown in Figure 1.2. This is the only drawback when using* (1.30) *as a finite difference scheme.*



(a) The quadrature points (b) Finite difference grid

Figure 1.4: An illustration of Q^2 element and the 3×3 Gauss-Lobatto quadrature.

Now consider solving $-\Delta u = f$ on a rectangular domain with homogeneous Dirichlet boundary condition. We can use (1.8) with piecewise Q^2 (tensor product of quadratic polynomial) basis and Simpon's rule or 3×3 Gauss-Lobatto quadrature rule on an uniform rectangular mesh as shown in Figure 1.4. With the same array notation in Chapter **??** The equivalent matrix vector form of the scheme (1.8) is

$$(H \otimes I + I \otimes H)vec(U) = vec(F), \tag{1.31}$$

or equivalently

$$(S \otimes M + M \otimes S)vec(U) = (M \otimes M)vec(F).$$

Remark 1.21. *The linear system in* (1.31) *can be easily solved by first computing eigenvalue decomposition of H then the eigenvector method as in the Chapter* **??**. *The eigenvalue decomposition of H can be computed in MATLAB, which is affordable since H is a small matrix compared to $H \otimes I + I \otimes H$.*

Remark 1.22. *The stiffness matrix S is always symmetric and the lumped mass matrix M is diagonal. The matrix H or $H \otimes I + I \otimes H$ is not symmetric, but $S \otimes M + M \otimes S$ is real symmetric. If a symmetric linear system is preferred, then the original symmetric form can be used.*

1.12.2 Neumann b.c.

For one-dimensional homogeneous Neumann boundary, the scheme can be written as

$$\begin{aligned}
\frac{7u_0 - 8u_1 + u_2}{2h^2} &= f_0, \\
\frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2} &= f_i, \quad \text{if } x_i \text{ is a mid point} \\
\frac{u_{i-2} - 8u_{i-1} + 14u_i - 8u_{i+1} + u_{i+2}}{4h^2} &= f_i, \quad \text{if } x_i \text{ is a cell end but not a boundary point,} \\
\frac{u_{N-1} - 8u_N + 7u_{N+1}}{2h^2} &= f_{N+1}.
\end{aligned}$$

1.12.3 The fourth order accuracy as a finite difference scheme

The fourth order accuracy of (1.31) is proved in [3].

The standard finite element error estimate for schemes in this section is third order in L^2-norm. But it can be proven that (1.30) is actually fourth order accurate in the 2-norm over grid points.

First of all, we can check that the finite difference approximation to the second order derivative in (1.30) is only second order accurate, even for the one in (1.30b). Second, if we use this second order approximation to solve a second order PDE such as $-u''(x) = f$, we get a fourth order accurate scheme! As a matter of fact, it can be rigorously proven that this scheme is fourth order accurate for commonly used linear second order PDEs [3, 2] for

- Elliptic equation $-\Delta u = f$.
- Parabolic equation $u_t = \Delta u$.
- Wave equation $u_{tt} = \Delta u$.

- Schrödinger equation $\mathbb{i}u_t = \Delta u$.
- Variable coefficient version of the equations above.

All error estimates in this notes are *a priori* error estimates, which means that the order holds if the exact solution $u(x)$ is smooth enough. For instance, the fourth order accuracy of (1.31) can be proven only if assuming $u \in H^4(\Omega)$. In practice, we often use high order accurate schemes for nonsmooth solutions, for which high order *a priori* error estimates can no longer hold. So a natural question is whether it still makes sense to use a high order accurate scheme like (1.30) on uniform meshes, which is nonethesless often used in applications. In Figure 1.5, there is a comparison of between the second order finite difference (1.23) and the fourth order finite difference(1.30) for solving the following generalized Allen-Cahn equation

$$\phi_t + u\phi_x + v\phi_y = \mu\Delta\phi - \frac{F'(\phi)}{\varepsilon}, \quad (x, y) \in \Omega, \tag{1.32}$$

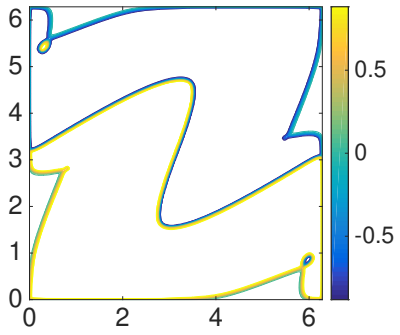
where u, v are given incompressible velocity field, and $F'(\cdot)$ is some fixed energy potential term. With the first order accurate implicit explicit (IMEX) time discretization, it becomes

$$\frac{\phi^{n+1} - \phi^n}{\Delta t} + u^{n+1}\phi_x^{n+1} + v^{n+1}\phi_y^{n+1} = \mu\Delta\phi^{n+1} - \frac{F'(\phi^n)}{\varepsilon}. \tag{1.33}$$

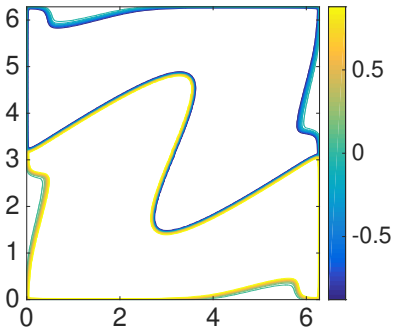
For the differential operators in (1.33), we can used two finite difference schemes derived from P^1 and P^2 finite element method with quadrature. For the second order derivative, they are (1.23) and (1.30). In Figure 1.5, we can see that the solution has a sharp interface, which gives large gradient thus smoothness or regularity of $\phi(x, y)$ is lost, yet the fourth order spatial discretization is still superior because the second order spatial discretization gives a wrong solution on the same coarse 239×239 grid. Higher order time accuracy here does not help the second order spatial discretization on the same coarse 239×239 grid. This is somehow intuitive since usually time evolution is a lot smoother thus spatial error is dominant in these problems.

1.13 Superconvergence

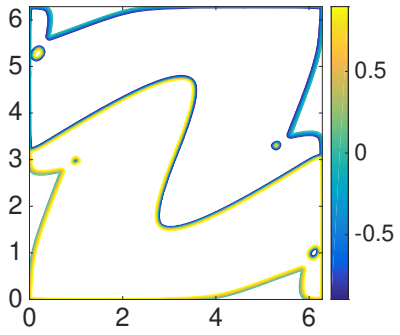
For the scheme (1.30), the error order at quadrature points (two cell ends and the middle point) is one order higher than L^2 error, which is computed for all x in the domain. Such a phenomenon that error at certain points is smaller is called *superconvergence*. On the other hand, it is straightforward to verify that the local truncation error of (1.30a) and (1.30b) is only second order. Recall that the local truncation error is not the true error. The phenomenon that local truncation error at particular locations has lower order than the true error order is called *supraconvergence*. The full proof



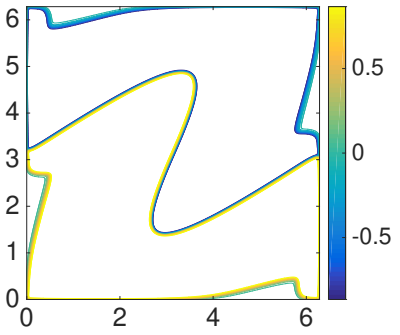
(a) Second order scheme with first order IMEX on a 239×239 grid



(b) Fourth order scheme with first order IMEX on a 239×239 grid



(c) Second order scheme with third order IMEX BDF on a 239×239 grid



(d) Reference Solution

Figure 1.5: Allen-Cahn with log energy at $T = 1.8$.

of why the scheme (1.30) is fourth order accurate in 2-norm over all grid points is quite complicated, see [3, 2]. In this section, we will only see some quick reasons why superconvergence is even possible.

1.13.1 The delta function

Our heuristic understanding of the delta function is $\delta(x) = \begin{cases} +\infty, & x = 0 \\ 0, & x \neq 0 \end{cases}$, which is however not a conventional function at all. One rigorous understanding of it is to define it as a *functional*, mapping any continuous function with compact support $f(x)$ linearly to a number $f(0)$. It is often denoted by an integral, i.e., the definition of the symbol $\delta(x)$ is defined to satisfy

$$\int_{-\infty}^{+\infty} f(x)\delta(x)dx = f(0), \quad \forall f(x) \in C_0(\mathbb{R}).$$

Recall that the function $f(x) = |x|$ is not differentiable but we can define its *weak or generalized derivative* as the step function $g(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases}$. Now let us compute the weak derivative of the step function by integration by parts:

$$\forall v(x) \in C_0^\infty(\mathbb{R}), \quad \int_{-\infty}^{+\infty} g(x)v'(x)dx = \int_0^{+\infty} v'(x)dx - \int_{-\infty}^{0} v'(x)dx = -2v(0).$$

With the definition of $\delta(x)$ above, we have

$$\forall v(x) \in C_0^\infty(\mathbb{R}) \subset C_0(\mathbb{R}), \quad \int_{-\infty}^{+\infty} v(x)\delta(x)dx = v(0),$$

thus

$$\forall v(x) \in C_0^\infty(\mathbb{R}), \quad \int_{-\infty}^{+\infty} g(x)v'(x)dx = -\int_{-\infty}^{+\infty} v(x)\frac{1}{2}\delta(x)dx.$$

Therefore, we have obtained $\frac{d^2}{dx^2}|x| = 2\delta(x)$, in the weak derivative sense. The symbol $\delta_a(x) := \delta(x - a)$ satisfies

$$\int_{-\infty}^{+\infty} f(x)\delta_a(x)dx = \int_{-\infty}^{+\infty} f(x)\delta(x - a)dx = f(a), \quad \forall f(x) \in C_0(\mathbb{R}).$$

Thus we also have $\frac{d^2}{dx^2}\frac{1}{2}|x - a| = \delta(x - a)$.

1.13.2 The one-dimensional Green's function

For the boundary value problem $-u''(x) = f(x), \quad x \in (0, 1), \quad u(0) = u(1) = 0$, its Green's function $G_a(x)$ is defined to satisfy

$$-\frac{d^2}{dx^2}G_a(x) = \delta_a(x), \quad G_a(0) = G_a(1) = 0,$$

where $a \in (0,1)$ is a fixed number.

Following the discussion in the previous subsection, it is straightforward to verify that

$$G_a(x) = \begin{cases} \frac{1}{2}(1-a)x, & x \le a \\ -\frac{1}{2}ax + \frac{1}{2}a, & x > a \end{cases},$$

thus

$$\frac{d}{dx}G_a(x) = \begin{cases} \frac{1}{2}(1-a), & x \le a \\ -\frac{1}{2}a, & x > a \end{cases}, \quad \frac{d^2}{dx^2}G_a(x) = \delta_a(x).$$

Notice that $G_a(x)$ is a continous piecewise linear function, but this is true only for one-dimensional problem.

1.13.3 Superconvergence at knots in one dimension

For the one-dimensional problem $-u''(x) = f(x), \quad x \in (0,1), \quad u(0) = u(1) = 0$, assume we have a mesh of intervals I_j, on which we define continuous piecewise polynomial spaces V^h and V_0^h.

The abstract finite element method is to seek $u_h \in V_0^h$ satisfying

$$(u_h', v_h') = (f, v_h), \quad \forall v_h \in V_0^h. \tag{1.34}$$

Recall that the solution u_h has Galerkin Orthogonality:

$$(u' - u_h', v_h') = 0, \quad \forall v_h \in V_0^h,$$

Let $e(x) = u(x) - u_h(x) \in H_0^1([0,1]) \subset C_0([0,1])$, then Galerkin Orthogonality can be written as

$$(e', v_h') = 0, \quad \forall v_h \in V_0^h.$$

Let x_i be the cell end of some interval I_j and we call x_i a knot. Then we consider the Green's function at $a = x_i$, e.g, $G_{x_i}(x)$, which is a piecewise linear polynomial defined on the same mesh, thus $G_{x_i}(x) \in V_0^h$. Now let us take a special test function $v_h = G_{x_i}(x)$ in the Galerkin Orthogonality:

$$(e', G_{x_i}(x)') = 0 \Rightarrow \int_0^1 e'(x) G_{x_i}(x)' dx = 0 \Rightarrow -\int_0^1 e(x) \frac{d^2}{dx^2} G_{x_i}(x) dx = 0$$

$$\Rightarrow -\int_0^1 e(x)\delta_{x_i}(x)dx = 0 \Rightarrow -\int_{-\infty}^{+\infty} e(x)\delta_{x_i}(x)dx = 0 \Rightarrow e(x_i) = 0,$$

where we have extended $e(x)$ to the whole real line by zero extension.

This means that the error at knots x_i are zero! Notice that this is the property to the abstract scheme (1.34) for any P^k basis, which we however

do not implement. For instance, for P^1 basis on a uniform mesh, the scheme (1.34) is the same as

$$\frac{1}{h}\begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{N-1} \\ u_N \end{bmatrix} = \begin{bmatrix} (f,\phi_1) \\ (f,\phi_2) \\ (f,\phi_3) \\ \vdots \\ (f,\phi_{N-1}) \\ (f,\phi_N) \end{bmatrix}. \tag{1.35}$$

But usually we implement it by approximating the integral (f,ϕ_i) by the trapezoidal rule, which is second order accurate. If we do compute the integrals (f,ϕ_i) exactly, then the scheme (1.35) has zero error.

For the P^2 basis on uniform mesh, the Simpson's rule is exact for the left hand integral (u_h', v_h'), thus (1.34) can be written as

$$\frac{2h}{3}\frac{-u_{i-1}+2u_i-u_{i+1}}{h^2} = (f,\phi_i), \quad \text{if } x_i \text{ is a mid point,} \tag{1.36a}$$

$$\frac{h}{3}\frac{u_{i-2}-8u_{i-1}+14u_i-8u_{i+1}+u_{i+2}}{4h} = (f,\phi_i), \quad \text{if } x_i \text{ is a cell end.} \tag{1.36b}$$

The error of the scheme (1.36) is zero at the cell end x_i (knots). Of course, in the scheme (1.30), we use Simpson's rule for approximating the integrals (f,ϕ_i), which is fourth order accurate. So at least now intuitively it is not a surprise that the scheme (1.30) should be fourth order accurate at the knots. For the fourth order accurate at the midpoint, we need some more arguments, which will not be explained in this notes.

Remark 1.23. *In general, by the standard superconvergence theory of P^k ($k \geq 2$) finite element method* (1.34) *(even for a variable coefficient problem in multiple dimensions), function values of $u_h(x)$ are $(k+2)$-th order accurate at Gauss-Labotto points for each small interval in 2-norm, as opposed to $(k+1)$-th order in the L^2-norm error estimate, and derivatives of $u_h(x)$ are $(k+2)$-th order accurate at Gauss points, as opposed to k-th order in the H^1-norm error estimate.*

1.14 Comparison with traditional finite difference method

1.14.1 Advantages of the finite element method

Troughout this chapter, we have seen many things that cannot happen or cannot be explained in the traditional finite difference method. Even on

uniform meshes for a rectangular domain, the finite element method is still superior from any perspective, because it gives us a finite difference with all desired properties. We summerize some comparsions in Table 1.1.

1.14.2 Limitations of the finite element method

In general, the finite element method is quite successful, for solving an elliptic equation $-\Delta u = f$ or some other types of equations including parabolic equations $u_t = \Delta u$, wave equations $u_{tt} = \Delta u$, Schrödinger equation $\mathbb{i}u_t = \Delta u$, etc. These equations all contain the Laplacian operator $-\Delta u$, for which a *coercive* bilinear form $\mathcal{A}(u, v) = (u', v')$ can be defined. Another different example is the biharmonic equation $u''''(x) = f$, for which we can also define a similar variational formulation with coercivity, thus the finite element method for this kind of fourth order PDE is also quite successful.

The foundation of the success for the finite element method, when missing, is also source of the limitations of the finite element method in applications. It could be quite or extremely difficult to use finite element method for equations lack of coercive operators. One simple example of such equations is the simple convection $u_t + u_x = 0$ which will be discussed in Chapter **??**, or its nonlinear version *nonlinear conservation laws* $u_t + f(u)_x = 0$ which will be discussed in Chapter **??**. Another example is the *Hamilton-Jacobi* equation $u_t + f(u_x) = 0$, e.g., $u_t + |\nabla u| = 0$, which is also closely related to *nonlinear conservation laws*. A formal application of the finite element method to these equations, with certain modifications to achieve stability or even convergence, is always possible, but many provable properties in this chapter will be no longer true.

Table 1.1: Comparison of traditional FD and finite element method for solving $-\nabla \cdot (A\nabla u) = f$ on Ω.

	traditional FD	FEM
Equation	approximates PDE	approximates variational form
Boundary condition	direct approximation	absorbed in V_0^h and variational form
Curved domain	a mapping to rectangular grid	Ω is easily approximated by Ω_h
Rectangular Ω	a rectangular grid	becomes finite difference
S matrix	nonsymmetric in general	always symmetric
Consistency	Taylor expansion	Galerkin Orthogonality
Stability	singular values	coercivity
Convergence	in 2-norm	H^1 and L^2 estimates
General tools	Calculus & Linear Algebra	functional analysis, PDE theory, etc
Error order	truncation error order	interpolation error order
Higher order schemes	large stencil, inducing difficulty near boundary	no difficulty at the boundary
Variable coefficient	difficult to construct higher order schemes	easy to to construct higher order schemes
Superconvergence	N/A	P^2 gives a 4th order FD
General implementation	form a matrix directly	computing some $S_{ij} = \mathcal{A}(\phi_j, \phi_i)$ to get S
Rectangular Ω	just solve a linear system	implement it as a FD scheme
Purely Neumann b.c.	left null vector is expensive to compute	left null vector is always $\mathbf{1}$

Bibliography

[1] Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1998.

[2] Hao Li, Danielö Appelo, and Xiangxiong Zhang. Accuracy of Spectral Element Method for Wave, Parabolic, and Schrödinger Equations. *SIAM Journal on Numerical Analysis*, 60(1):339–363, 2022.

[3] Hao Li and Xiangxiong Zhang. Superconvergence of high order finite difference schemes based on variational formulation for elliptic equations. *Journal of Scientific Computing*, 82(2):36, 2020.