INTRODUCTION TO FINITE ELEMENT METHODS

LONG CHEN

Finite element methods are based on the variational formulation of partial differential equations which only need to compute the gradient of a function. Although unknowns are still associated to nodes, the function composed by piece-wise polynomials on each element and thus the gradient can be computed element-wise. Finite element spaces can thus be constructed on general triangulations and this method is able to handle complex geometries and boundaries. Boundary condition is naturally build into the weak formulation or the function space. The variational approach also give solid mathematical foundation and make the error analysis more systematic. Generally speaking, finite element methods is the method of choice in all types of analysis for elliptic equations in complex domains.

1. GALERKIN METHODS

The finite element methods have been introduced as methods for approximate solution of variational problems. Let us consider the model problem: Poisson equation with homogenous Dirichlet boundary conditions

$$-\Delta u = f \ \text{ in } \Omega, \qquad u = 0 \ \text{ on } \partial\Omega. \tag{1}$$

Multiply a test function v, integrate over Ω, and use integration by parts to obtain the corresponding variational formulation: Find $u \in V = H_0^1(\Omega)$, such that

$$a(u,v) = \langle f, v\rangle, \quad \text{for all } \ v \in V. \tag{2}$$

where

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x, \quad \langle f, v\rangle = \int_\Omega f v \,\mathrm{d}x, \quad \text{for } f \in L_2(\Omega) \tag{3}$$

Clearly, in such case $a(u,u) = |u|_{1,\Omega}^2$ and the Poincaré inequality $\|u\|_{0,\Omega} \leq C|u|_{1,\Omega}$ for $H_0^1(\Omega)$ implies that $a(\cdot,\cdot)$ is an inner product on V, and thus the problem (2) has a unique solution by the Riesz representation theorem.

We now consider a class of methods, known as *Galerkin methods* which are used to approximate the solution to (2). Consider a finite dimensional subspace $V_h \subset V$, and let $V_h = \operatorname{span}\{\phi_1, \ldots, \phi_N\}$. For any function $v \in V_h$, there is a unique representation: $v = \sum_{i=1}^N v_i \phi_i$. We thus can define an isomorphism $V_h \cong \mathbb{R}^N$ by

$$v = \sum_{i=1}^N v_i \phi_i \longleftrightarrow \boldsymbol{v} = (v_1, \cdots, v_N)^\intercal, \tag{4}$$

and call $\boldsymbol{v}$ the coordinate vector of v relative to the basis $\{\phi_i\}_{i=1}^N$. Following the terminology in elasticity, we introduce the *stiffness matrix*

$$\boldsymbol{A} = (a_{ij})_{N\times N}, \ \text{ with } \quad a_{ij} = a(\phi_j, \phi_i),$$

Date: Created on Oct 2007. Latest update October 18, 2017.

and the *load vector* $\boldsymbol{f} = \{\langle f, \phi_k\rangle\}_{k=1}^N \in \mathbb{R}^N$. Then the coefficient vector can be obtained by solving the following linear algebraic system

$$\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f}.$$

It is straightforward to verify $\boldsymbol{A}$ is an SPD matrix and thus the solution $\boldsymbol{u}$ exists and unique.

The finite element methods is a special and most popular example of Galerkin methods by constructing a finite dimensional subspace V_h based on triangulations of the domain. The name comes from the fact that the domain is decomposed into finite number of elements.

2. Triangulations and Barycentric Coordinates

In this section, we discuss triangulations used in finite element methods. We would like to distinguish two structures related to a triangulation: one is the topology of a mesh determined by the combinatorial connectivity of vertices; another is the geometric shape which depends on both the connectivity and the location of vertices. Correspondingly there are two basic data structures used to represent a triangulation.

2.1. Geometric simplex and triangulation.

Let $\boldsymbol{x}_i = (x_{1,i}, \cdots, x_{n,i})^\intercal, i = 1, \cdots, n+1$ be $n+1$ points in $\mathbb{R}^n$. We say $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n+1}$ do not all lie in one hyper-plane if the n-vectors $\boldsymbol{x}_1\boldsymbol{x}_2, \cdots, \boldsymbol{x}_1\boldsymbol{x}_{n+1}$ are independent. This is equivalent to the matrix:

$$A = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,n+1} \\ x_{2,1} & x_{2,2} & \dots & x_{2,n+1} \\ \vdots & \vdots & & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,n+1} \\ 1 & 1 & \dots & 1 \end{pmatrix}$$

is non-singular. Given any point $\boldsymbol{x} = (x_1, \cdots, x_n) \in \mathbb{R}^n$, by solving the following linear system

$$A \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \lambda_{n+1} \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ 1 \end{pmatrix}, \tag{5}$$

we obtain unique $n+1$ real numbers $\lambda_i(\boldsymbol{x}), 1 \le i \le n+1$, such that for any $\boldsymbol{x} \in \mathbb{R}^n$

$$\boldsymbol{x} = \sum_{i=1}^{n+1} \lambda_i(\boldsymbol{x})\boldsymbol{x}_i, \quad \text{with} \quad \sum_{i=1}^{n+1} \lambda_i(\boldsymbol{x}) = 1. \tag{6}$$

The *convex hull* of the $d+1$ points $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{d+1}$ in $\mathbb{R}^n$

$$\tau := \{\boldsymbol{x} = \sum_{i=1}^{d+1} \lambda_i \boldsymbol{x}_i \,|\, 0 \le \lambda_i \le 1, i = 1 : d+1, \sum_{i=1}^{d+1} \lambda_i = 1\} \tag{7}$$

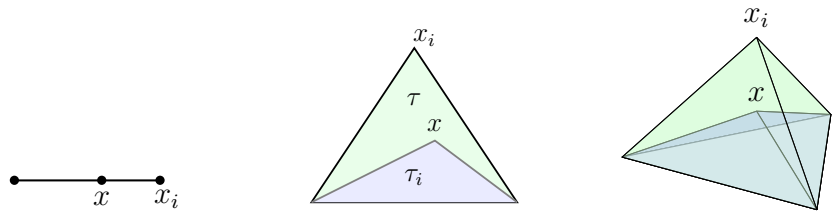
is defined as a *geometric d-simplex* generated (or spanned) by the vertices $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{d+1}$. For example, a triangle is a 2-simplex and a tetrahedron is a 3-simplex. For an integer $0 \le m \le d-1$, an m-dimensional face of τ is any m-simplex generated by $m+1$ vertices of τ. Zero dimensional faces are vertices and one-dimensional faces are called edges of τ. The $(d-1)$-face opposite to the vertex $\boldsymbol{x}_i$ will be denoted by F_i.

The numbers $\lambda_1(\boldsymbol{x}), \cdots, \lambda_{d+1}(\boldsymbol{x})$ are called *barycentric coordinates* of $\boldsymbol{x}$ with respect to the $d+1$ points $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{d+1}$. There is a simple geometric meaning of the barycentric

coordinates. Given a $\boldsymbol{x} \in \tau$, let $\tau_i(\boldsymbol{x})$ be the simplex with vertices $\boldsymbol{x}_i$ replaced by $\boldsymbol{x}$. Then, by the Cramer's rule for solving (5),

$$\lambda_i(\boldsymbol{x}) = \frac{|\tau_i(\boldsymbol{x})|}{|\tau|}, \tag{8}$$

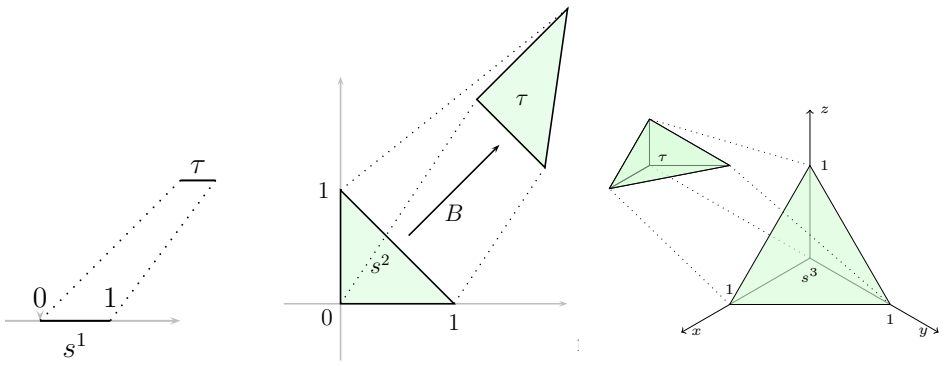
where $|\cdot|$ is the Lebesgue measure in $\mathbb{R}^d$, namely area in two dimensions and volume in three dimensions. Note that $\lambda_i(\boldsymbol{x})$ is an affine function of $\boldsymbol{x}$ and vanished on the face F_i.



(a) 1d simplex (b) 2d simplex (c) 3d simplex

FIGURE 1. Geometric explanation of barycentric coordinates.

It is convenient to have a standard simplex $s^n \subset \mathbb{R}^n$ spanned by the vertices $\mathbf{0}, \boldsymbol{e}_1, \cdots, \boldsymbol{e}_n$ where $\boldsymbol{e}_i = (0, \cdots, 1, \cdots, 0)$. Then any n-simplex $\tau \subset \mathbb{R}^n$ can be thought as an image of s^n through an affine map $B : s^n \to \tau$ with $B(\boldsymbol{e}_i) = \boldsymbol{x}_i$. See Figure 2. The simplex s^n is also often called *reference simplex*



(a) A standard simplex in $\mathbb{R}^1$ (b) A standard simplex in $\mathbb{R}^2$ (c) A standard simplex in $\mathbb{R}^3$

FIGURE 2. Reference simplexes in $\mathbb{R}^1, \mathbb{R}^2$ and $\mathbb{R}^3$

Let Ω be a polyhedral domain in $\mathbb{R}^d, d \geq 1$. A geometric triangulation (also called mesh or grid) $\mathcal{T}$ of Ω is a set of d-simplices such that

$$\cup_{\tau \in \mathcal{T}} \tau = \overline{\Omega}, \quad \text{and} \quad \overset{\circ}{\tau}_i \cap \overset{\circ}{\tau}_j = \varnothing, i \neq j.$$

Remark 2.1. In this course, we restrict ourself to simplicial triangulations. There are other type of meshes by partition the domain into quadrilateral (in 2-D), cubes, prisms (in 3-D), or polytopes in general.

There are two conditions that we shall impose on triangulations that are important in the finite element computation. The first requirement is a topological property. A triangulation $\mathcal{T}$ is called *conforming* or *compatible* if the intersection of any two simplexes τ and τ' in $\mathcal{T}$ is either empty or a common lower dimensional simplex (nodes in two dimensions, nodes and edges in three dimensions).

(a) A triangulation with a hanging node

(b) A conforming triangulation

FIGURE 3. Two triangulations. The left is non-conforming and the right is conforming.

The second important condition depends on the geometric structure. A set of triangulations $\mathscr{T}$ is called *shape regular* if there exists a constant c_0 such that

$$\max_{\tau\in\mathcal{T}} \frac{\operatorname{diam}(\tau)^d}{|\tau|} \leq c_0, \quad \text{for all } \mathcal{T} \in \mathscr{T}, \tag{9}$$

where $\operatorname{diam}(\tau)$ is the diameter of τ and $|\tau|$ is the measure of τ in $\mathbb{R}^d$. In two dimensions, it is equivalent to the minimal angle of each triangulation is bounded below uniformly in the shape regular class. We shall define $h_\tau = |\tau|^{1/n}$ for any $\tau \in \mathcal{T} \in \mathscr{T}$. By (9), $h_\tau \eqsim \operatorname{diam}(\tau)$ represents the size of an element $\tau \in \mathcal{T}$ for a shape regular triangulation $\mathcal{T} \in \mathscr{T}$.

In addition to (9), if

$$\frac{\max_{\tau\in\mathcal{T}} |\tau|}{\min_{\tau\in\mathcal{T}} |\tau|} \leq \rho, \quad \text{for all } \mathcal{T} \in \mathscr{T}, \tag{10}$$

$\mathscr{T}$ is called *quasi-uniform*. For quasi-uniform grids, define the mesh size of $\mathcal{T}$ as $h_\mathcal{T} := \max_{\tau\in\mathcal{T}} h_\tau$. It is used to measure the approximation rate. In FEM literature, we often write a triangulation as $\mathcal{T}_h$.

2.2. **Abstract simplex and simplicial complex.** To distinguish the topological structure with the geometric one, we now understand the points as abstract entities and introduce *abstract simplex* or *combinatorial simplex*. The set $\tau = \{v_1, \cdots, v_{d+1}\}$ of $d+1$ abstract points is called an abstract d-simplex. A face σ of a simplex τ is a simplex determined by a non-empty subset of τ. A k-face has $k+1$ points. A proper face is any face different from τ.

Let $\mathcal{N} = \{v_1, v_2, \cdots, v_N\}$ be a set of N abstract points. An *abstract simplicial complex* $\mathcal{T}$ is a set of simplices formed by finite subsets of $\mathcal{N}$ such that if $\tau \in \mathcal{T}$ is a simplex, then any face of τ is also a simplex in $\mathcal{T}$. By the definition, a two dimensional combinatorial complex $\mathcal{T}$ contains not only triangles but also edges and vertices of these triangles. A geometric triangulation defined before is only a set of d-simplex not its faces. By including all its face, we shall get a simplicial complex.

A subset $\mathcal{M} \subset \mathcal{T}$ is a subcomplex of $\mathcal{T}$ if $\mathcal{M}$ is a simplicial complex itself. Important classes of subcomplex includes the *star* or *ring* of a simplex. That is for a simplex $\sigma \in \mathcal{T}$

$$\operatorname{star}(\sigma) = \{\tau \in \mathcal{T}, \sigma \subset \tau\}.$$

If two, or more, simplices of $\mathcal{T}$ share a common face, they are called *adjacent* or *neighbors*. The boundary of $\mathcal{T}$ is formed by any proper face that belongs to only one simplex, and its faces.

By associating the set of abstract points with geometric points in $\mathbb{R}^n, n \geq d$, we obtain a geometric shape consisting of piecewise flat simplices. This is called a geometric realization of an abstract simplicial complex or, using the terminology of geometry, the embedding of $\mathcal{T}$ into $\mathbb{R}^n$. The embedding is uniquely determined by the identification of abstract and geometric vertices.

We would like to emphasis that they are two different structures of a triangulation $\mathcal{T}$: one is the topology of a mesh which is determined by the combinatorial connectivity of vertices; another is the geometric shape which depends on the location of the vertices. For example, a planar triangulation is a two dimensional abstract simplicial complex which can be embedded into $\mathbb{R}^2$ and thus called 2-D triangulation. A 2-D simplicial complex could also be embedding into $\mathbb{R}^3$ and result a triangulation of a surface. For these two different embedding, they have the same combinatory structure as an abstract simplicial complex but different geometric structure by representing a flat domain in $\mathbb{R}^2$ or a surface in $\mathbb{R}^3$.

3. Linear Finite Element Spaces

In this section we introduce the simplest linear finite element space of $H^1(\Omega)$ and use scaling argument to estimate the interpolation error.

3.1. Linear finite element space and the nodal interpolation.

Given a shape regular triangulation $\mathcal{T}_h$ of Ω, we set

$$V_h := \{v \,|\, v \in C(\overline{\Omega}), \text{ and } v|_\tau \in \mathcal{P}_1, \forall \tau \in \mathcal{T}_h\},$$

where $\mathcal{P}_1(\tau)$ denotes the space of polynomials of degree 1 (linear) on $\tau \in \mathcal{T}_h$. Whenever we need to deal with boundary conditions, we further define $V_{h,0} = V_h \cap H_0^1(\Omega)$. We note here that the global continuity is also necessary in the definition of V_h in the sense that if $u \in H^1(\Omega)$, and u is piecewise smooth, then u should be continuous.

We use N to denote the dimension of finite element spaces. For V_h, N is the number of vertices of the triangulation $\mathcal{T}_h$ and for $V_{h,0}$, N is the number of interior vertices. For linear finite element spaces, we have the so called *a nodal basis functions* $\{\phi_i, i = 1, \cdots N\}$ such that ϕ_i is piecewise linear (with respect to the triangulation) and $\phi_i(x_j) = \delta_{i,j}$ for all vertices x_j of $\mathcal{T}_h$. Therefore for any $v_h \in V_h$, we have the representation

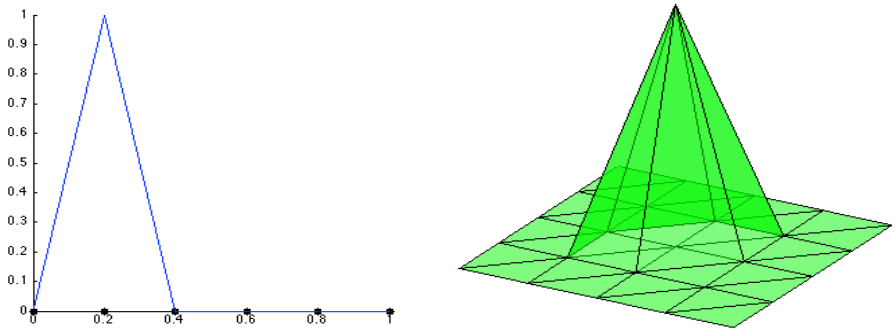
$$v_h(x) = \sum_{i=1}^{N} v_h(x_i)\phi_i(x).$$

Due to the shape of the nodal basis function, it is also called hat function. See Figure 4 for an illustration in 1-D and 2-D. Note that $\phi_i|_\tau$ is the corresponding barycentrical coordinates.

The nodal interpolation operator $I_h : C(\overline{\Omega}) \to V_h$ is defined as

$$(I_h u)(x) = \sum_{i=1}^{N} u(x_i)\phi_i(x),$$

and denoted by the short notation $u_I := I_h u$.



(a) 1d basis function (b) 2d basis function

FIGURE 4. Nodal basis functions in 1d and 2d.

3.2. **Scaling argument and inverse inequality.** Let $\hat{\tau} = s^n$ be the standard n-simplex which is also called reference simplex. Define an affine map $F : \hat{\tau} \mapsto \tau$, and the function $\hat{v}_h(\hat{x}) = v_h(F(\hat{x})), \forall \hat{x} \in \hat{\tau}$. The affine map F consists of translation, rotation and scaling. Essentially it is like a scaling of variables $x = h\hat{x}$. The following important relation between the norms on the reference simplex and physical simplex can be easily proved by changing of variable. The shape regularity of a simplex will be needed to bound the 2-norm of Jacobi matrix and its determinant in terms of h.

Lemma 3.1. *When τ is shape regular, we have*

$$\|\hat{D}^\alpha \hat{v}\|_{0,p,\hat{\tau}} \eqsim h_\tau^{\mathrm{sob}_n(k,p)} \|D^\alpha v\|_{0,p,\tau}, \quad \textit{for all} \ \ |\alpha| = k, \tag{11}$$

where the Sobolev number is $\mathrm{sob}_n(k,p) = k - \frac{n}{p}$.

Proof. Let $J = \left(\frac{\partial x}{\partial \hat{x}}\right)$ be the Jacobi matrix of the map F. Then $\hat{\nabla}\hat{v} = J\nabla v$ and consequently $|\hat{D}^\alpha \hat{v}| \eqsim h^k |D^\alpha v|$, cf. Exercise 3. Use the change of variable $\mathrm{d}x = |J|\,\mathrm{d}\hat{x} \eqsim h^n\,\mathrm{d}\hat{x}$, we have

$$\int_{\hat{\tau}} |\hat{D}^\alpha \hat{v}|^p \,\mathrm{d}\hat{x} \eqsim \int_\tau h^{kp} |D^\alpha v| h^{-n} \,\mathrm{d}x.$$

Then the results follows. □

For two Banach spaces B_0, B_1, the continuous embedding $B_1 \hookrightarrow B_0$ implies that

$$\|u\|_{B_0} \lesssim \|u\|_{B_1}, \quad \text{for all} \ \ u \in B_1.$$

The inequality in the reserve way $\|u\|_{B_1} \lesssim \|u\|_{B_0}$ may not true. Now considering finite element spaces $V_h \subset B_i, i = 0, 1$ endowed with two norms. For a fixed h, the dimension of V_h is finite. Since all the norms of finite dimensional spaces are equivalent, there exists constant C_h such that

$$\|u_h\|_{B_1} \leq C_h \|u_h\|_{B_0}, \quad \text{for all} \ \ u_h \in V_h. \tag{12}$$

The constant C_h in (12) depends on the size and shape of the domain. If we consider the restriction on one element τ and transfer to the reference element $\hat{\tau}$, the constant for the norm equivalence will not depend on h, i.e.,

$$\|\hat{u}_h\|_{B_1,\hat{\tau}} \lesssim \|\hat{u}_h\|_{B_0,\hat{\tau}}.$$

Using the map F, we can then determine the constant in terms of the mesh size h. This is called *scaling argument*.

As an example, we obtain the following typical inverse inequalities

$$|u_h|_{1,\tau} \lesssim h^{-1}\|u_h\|_\tau, \quad \text{and} \tag{13}$$

$$\|u_h\|_{0,p,\tau} \lesssim h^{n(1/p-1/q)}\|u_h\|_{0,q,\tau}, \quad 1 \le q \le p \le \infty. \tag{14}$$

Recall that we have the following refined embedding theorem

$$\|v\|_{0,p,\Omega} \le C(n,\Omega)\, p^{1-1/n}\|v\|_{1,n,\Omega}, \quad \text{for all } \ 1 \le p < \infty. \tag{15}$$

Using the inverse inequality and the above embedding result, for $u_h \in V_h$, we have

$$\|v_h\|_\infty \lesssim h^{-n/p}\|v_h\|_{0,p} \lesssim h^{-n/p}p^{1-1/n}\|v_h\|_{1,n}.$$

Now choosing $p = |\log h|$ and noting $h^{-n/|\log h|} \le C$, we get the following discrete embedding result:

$$\|v_h\|_\infty \lesssim |\log h|^{1-1/n}\|v_h\|_{1,n}, \quad \text{for all } \ v_h \in V_h. \tag{16}$$

In particular, when $n = 2$, we can almost control the maximum norm of a finite element function by its H^1 norm. Although the term $|\log h|$ is unbounded as $h \to 0$, it increases very slowly and appears as a constant for practical h.

3.3. **Error estimate of nodal interpolation.** We use the scaling argument to estimate the interpolation error $|u - u_I|_{1,\Omega}$ and refer to Exercise 4 for a proof using multipoint Taylor series.

Theorem 3.2. *For $u \in H^2(\Omega), \Omega \subset \mathbb{R}^n, n = 1, 2, 3$, and V_h the linear finite element space based on quasi-uniform triangulations $\mathcal{T}_h$, we have*

$$|u - u_I|_{1,\Omega} \lesssim h|u|_{2,\Omega}.$$

Proof. First of all, by the Sobolev embedding theorem, $H^2 \hookrightarrow C(\bar{\Omega})$ for $n \le 3$. Thus the nodal interpolation u_I is well defined.

Since $|\hat{u}|_{1,\hat{\tau}} \le \|\hat{u}\|_{2,\hat{\tau}}$, and

$$|\hat{u}_I|_{1,\hat{\tau}} \lesssim \|\hat{u}_I\|_{0,\infty,\hat{\tau}} \le \|\hat{u}\|_{0,\infty,\hat{\tau}} \lesssim \|\hat{u}\|_{2,\hat{\tau}},$$

we get the estimate in the reference simplex: for $\hat{u} \in H^2(\hat{\tau})$

$$|\hat{u} - \hat{u}_I|_{1,\hat{\tau}} \le |\hat{u}|_{1,\hat{\tau}} + |\hat{u}_I|_{1,\hat{\tau}} \lesssim \|\hat{u}\|_{2,\hat{\tau}}. \tag{17}$$

The nodal interpolation will preserve linear polynomials i.e. $\hat{p}_I = \hat{p}$ for $\hat{p} \in \mathcal{P}_1(\hat{\tau})$, then

$$|\hat{u} - \hat{u}_I|_{1,\hat{\tau}} = |(\hat{u} + \hat{p}) - (\hat{u} + \hat{p})_I|_{1,\hat{\tau}} \lesssim \|\hat{u} + \hat{p}\|_{2,\hat{\tau}}, \ \forall\, \hat{p} \in \mathcal{P}_1(\hat{\tau}),$$

and thus by the Bramble-Hilbert lemma

$$|\hat{u} - \hat{u}_I|_{1,\hat{\tau}} \lesssim \inf_{\hat{p}\in\mathcal{P}_1(\hat{\tau})} \|\hat{u} + \hat{p}\|_{2,\tau} \lesssim |\hat{u}|_{2,\hat{\tau}}. \tag{18}$$

We now use the scaling argument to transfer the inequality back to the simplex τ. First

$$|\hat{u}|_{2,\hat{\tau}} \lesssim h_\tau^{2-\frac{n}{2}}|u|_{2,\tau}.$$

To scale the left hand side, we need a property of the interpolation operator

$$\widehat{u - u_I} = \hat{u} - \hat{u}_I, \tag{19}$$

namely the interpolation is affine invariant, which can be verified easily by definition. Then by the scaling argument

$$h_\tau^{1-\frac{n}{2}}|u-u_I|_{1,\tau} \lesssim |\widehat{u-u_I}|_{1,\hat{\tau}} = |\hat{u}-\hat{u}_I|_{1,\hat{\tau}}.$$

Combing all the arguments above leads to the interpolation error estimate on a quasiuniform mesh. For $u \in H^2(\Omega)$,

$$|u-u_I|_{1,\Omega}^2 = \sum_{\tau\in\mathcal{T}_h} |u-u_I|_{1,\tau}^2 \lesssim \sum_{\tau\in\mathcal{T}_h} h_\tau^2 |u|_{2,\tau}^2 \eqsim h^2|u|_{2,\Omega}^2.$$

□

4. Finite Element Methods and Error Estimate

Finite element methods for solving Poisson equation is a special case of Galerkin method by choosing the subspace $V_h \subset V$ based on a triangulation $\mathcal{T}_h$ of the underlying domain. As an example, let us consider the linear finite element space V_h. The finite element approximation will be: find $u_h \in V_h$ such that

$$a(u_h, v_h) = \langle f, v_h\rangle, \quad \text{for all } \ v_h \in V_h. \tag{20}$$

Again the existence and uniqueness follows from the Riesz representation theorem since $f \in V' \subset V_h'$ is also a continuous linear functional on V_h and by Poincaré inequality $a(\cdot,\cdot)$ defines an inner production on $H_0^1(\Omega)$.

4.1. H^1 error estimate.

We first derive an important orthogonality result for projections. Let u and u_h be the solution of continuous and discrete equations respectively i.e.

$$\begin{aligned} a(u,v) &= \langle f, v\rangle \qquad \forall v \in H_0^1(\Omega),\\ a(u_h,v) &= \langle f, v\rangle \qquad \forall v \in V_h. \end{aligned}$$

By subtracting these two equations, we then get an important orthogonality

$$a(u-u_h, v_h) = 0 \qquad \forall v_h \in V_h, \tag{21}$$

which implies the following optimality of the finite element approximation

$$\|\nabla(u-u_h)\| = \inf_{v_h\in V_h} \|\nabla(u-v_h)\|. \tag{22}$$

Theorem 4.1. *Let u and u_h be the solution of continuous and discrete equations respectively. When $u \in H^2(\Omega)\cap H_0^1(\Omega)$, we have the following optimal order estimate:*

$$\|\nabla(u-u_h)\| \lesssim h\|u\|_2. \tag{23}$$

Furthermore when H^2-regularity result holds, we have

$$\|\nabla(u-u_h)\| \lesssim h\|f\|. \tag{24}$$

Proof. When $u \in H^2(\Omega)\cap H_0^1(\Omega)$, the nodal interpolation operator is well defined by the embedding theorem. By (22), we then have

$$\|\nabla(u-u_h)\| \le \|\nabla(u-u_I)\| \lesssim h\|u\|_2 \lesssim h\|f\|.$$

Here in the second $\lesssim$, we have used the error estimate of interpolation operator, and in the third one, we have useed the regularity result. □

4.2. L^2 **error estimate.** Now we estimate $\|u - u_h\|$. The main technical is the combination of the duality argument and the regularity result. It is known as Aubin-Nitsche duality argument or simply "Nitsche's trick".

Theorem 4.2. *Let u and u_h be the solution of continuous and discrete equations respectively. Suppose the H^2 regularity result holds, we then have the following optimal order approximation in L^2 norm*

$$\|u - u_h\| \lesssim h^2 \|u\|_2. \tag{25}$$

Proof. By the H^2 regularity result, there exists $w \in H^2(\Omega) \cap H_0^1(\Omega)$ such that

$$a(w, v) = (u - u_h, v), \quad \text{for all } v \in H_0^1(\Omega), \tag{26}$$

and $\|w\|_2 \leq C\|u - u_h\|$. Choosing $v = u - u_h$ in (26), we get

$$\begin{aligned}
\|u - u_h\|^2 &= a(w, u - u_h) \\
&= a(w - w_I, u - u_h) \\
&\leq \|\nabla(w - w_I)\| \|\nabla(u - u_h)\| && \text{(orthogonality)} \\
&\lesssim h\|w\|_2 \|\nabla(u - u_h)\| \\
&\lesssim h\|u - u_h\| \|\nabla(u - u_h)\| && \text{(regularity)}.
\end{aligned}$$

Canceling one $\|u - u_h\|$, we get

$$\|u - u_h\| \leq Ch\|\nabla(u - u_h)\| \lesssim h^2 \|u\|_2.$$

□

For the estimate in H^1 norm, when u is smooth enough, we can obtain the optimal first order estimate. But for L^2 norm, the duality argument requires H^2 elliptic regularity, which in turn requires that the polygonal domain be convex. In fact, for a non-convex polygonal domain, it will usually not be true that $\|u - u_h\| = \mathcal{O}(h^2)$ even if the solution u is smooth.

EXERCISE

1. A multi-index α is an k-tuple of non-negative integers $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_k)$. The length of α is defined by $|\alpha| = \sum_{i=1}^k \alpha_i$, and $\alpha! = \alpha_1! \ldots \alpha_n!$. For a given vector $\boldsymbol{x} = (x_1, x_2, \cdots, x_k)$, we define $\boldsymbol{x}^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_k^{\alpha_k}$. Finally let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_{n+1})$ denote the vector of barycentric coordinates.

For an $n+1$ multi-index α and an n-simplex τ, one has

$$\int_\tau \boldsymbol{\lambda}^\alpha(\boldsymbol{x}) d\boldsymbol{x} = \frac{\alpha! n!}{(|\alpha| + n)!} |\tau|. \tag{27}$$

We shall prove the identity (27) through the following sub-problems:

(1) $n = 1$ and $\tau = [0, 1]$. Prove that

$$\int_0^1 x^{\alpha_1}(1-x)^{\alpha_2} dx = \frac{\alpha_1! \alpha_2!}{(\alpha_1 + \alpha_2 + 1)!}$$

(2) $n = 2$ and $\tau = s^2$. Prove that

$$\int_\tau x^{\alpha_1} y^{\alpha_2} (1-x-y)^{\alpha_3} dx = \frac{\alpha_1! \alpha_2! \alpha_3!}{(\alpha_1 + \alpha_2 + \alpha_3 + 2)!}$$

(3) Prove the identity (27) for $\tau = s^n$ using induction on n.

(4) Prove the identity (27) for general simplex τ by using the transformation from the standard simplex s^n.

2. In this exercise, we give explicit formula of the stiffness matrix. Let τ be a triangle with vertices $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ and let $\lambda_1, \lambda_2, \lambda_3$ be corresponding barycentric coordinates.

(1) Let $\boldsymbol{n}_i$ be the outward normal vector of the edge e_i and d_i be the distance from $\boldsymbol{x}_i$ to e_i. Prove that

$$\nabla \lambda_i = -\frac{1}{d_i} \boldsymbol{n}_i.$$

(2) Let θ_i be the angle associated to the vertex $\boldsymbol{x}_i$. Prove that

$$\int_\tau \nabla \lambda_i \cdot \nabla \lambda_j \, \mathrm{d}x = -\frac{1}{2} \cot \theta_k,$$

where (i, j, k) is any permutation of $(1, 2, 3)$.

(3) Let $c_i = \cot \theta_i, i = 1$ to 3. If we define the local stiffness matrix $\boldsymbol{A}_\tau$ as 3×3 matrix formed by $\int_\tau \nabla \lambda_i \cdot \nabla \lambda_j \, dx, i, j = 1, 2, 3$. Show that

$$\boldsymbol{A}_\tau = \frac{1}{2} \begin{bmatrix} c_2 + c_3 & -c_3 & -c_2 \\ -c_3 & c_3 + c_1 & -c_1 \\ -c_2 & -c_1 & c_1 + c_2 \end{bmatrix}.$$

(4) Let e be an interior edge in the triangulation $\mathcal{T}$ with nodes x_i and x_j, and shared by two triangles τ_1 and τ_2. Denoted the angle in τ oppositing to e by θ_e^τ. Then prove that the entry $a_{ij} = \int_\Omega \nabla \phi_i \cdot \nabla \phi_j \, dx$ is

$$a_{ij} = -\frac{1}{2}(\cot \theta_e^{\tau_1} + \cot \theta_e^{\tau_2}).$$

Consequently $a_{ij} \leq 0$ if and only if $\theta_e^{\tau_1} + \theta_e^{\tau_2} \leq \pi$. By the way, if a 2-D triangulation satisfying $\theta_e^{\tau_1} + \theta_e^{\tau_2} \leq \pi$, it is called a Delaunay triangulation.

3. In this exercise, we compute the singular values of the affine map from the reference triangle $\hat{\tau}$ spanned by $\hat{a}_1 = (1,0), \hat{a}_2 = (0,1)$ and $\hat{a}_3 = (0,0)$ to a triangle τ with three vertices $a_i, i = 1,2,3$. One of such affine map is to match the local indices of three vertices, i.e., $F(\hat{a}_i) = a_i, i = 1,2,3$:

$$F(\hat{\boldsymbol{x}}) = B^{\intercal}(\hat{\boldsymbol{x}}) + c,$$

where

$$B = \begin{bmatrix} x_1 - x_3 & y_1 - y_3 \\ x_2 - x_3 & y_2 - y_3 \end{bmatrix}, \text{ and } c = (x_3, y_3)^{\intercal}.$$

(1) Estimate the $\sigma_{\max}(B)$ and $\sigma_{\min}(B)$ in terms of edge lengthes and angles of the triangle τ.
(2) Establish inequalities between $|\nabla v|$ and $|\hat{\nabla}\hat{v}|$ where $\hat{v}(\hat{x}) := v(F(\hat{x}))$.

4. In this exercise, we give an elementary proof of the interpolation error estimate. Let τ be a simplex with vertices $\boldsymbol{x}_i, i = 1, \ldots, n+1$ and $\{\lambda_i, i = 1, \ldots, n+1\}$ be the corresponding barycentric coordinates.

(1) Show that

$$u_I(\boldsymbol{x}) = \sum_{i=1}^{n+1} u(\boldsymbol{x}_i)\lambda_i(\boldsymbol{x}), \quad \sum_{i=1}^{n+1} \lambda_i(\boldsymbol{x}) = 1, \quad \text{and} \quad \sum_{i=1}^{n+1} (\boldsymbol{x} - \boldsymbol{x}_i)\lambda_i(\boldsymbol{x}) = 0.$$

(2) Let us introduce the auxiliary functions

$$g_i(t, \boldsymbol{x}) = u(\boldsymbol{x}_i + t(\boldsymbol{x} - \boldsymbol{x}_i)).$$

For $u \in C^2(\bar{\tau})$, prove the following error equations

$$(u_I - u)(\boldsymbol{x}) = \sum_{i=1}^{n+1} \lambda_i(\boldsymbol{x}) \int_0^1 t g_i''(t, \boldsymbol{x}) dt,$$

$$\nabla(u_I - u)(\boldsymbol{x}) = \sum_{i=1}^{n+1} \nabla\lambda_i \int_0^1 t g_i''(t, \boldsymbol{x}) dt.$$

Hint: Multiply the following Taylor series by λ_i and sum over i

$$g_i(0, \boldsymbol{x}) = g_i(1, \boldsymbol{x}) - g_i'(1, \boldsymbol{x}) + \int_0^1 t g_i''(t, \boldsymbol{x}) dt.$$

(3) Let $u \in W^{2,p}$ with $p > n/2$. Using the error formulate in (2) to prove

$$\|u - u_I\|_{0,p} \leq C_1 h^2 |u|_{2,p}, \quad |u - u_I|_{1,p} \leq C_2 h |u|_{2,p}.$$

Try to obtain a sharp constants in the above inequalities.

4. We can define the quadratic element by using piecewise quadratic polynomials.

(1) Inside one triangle, write out a basis of the quadratic element.
(2) Derive the error estimate on $|u - u_h|_1$, $\|u - u_h\|$, and $\|u - u_h\|_{-1}$.