

An overview of the finite element method

In this chapter, we give an overview of the finite element method for the discretization of partial differential equations. We will mainly consider the linear finite element discretization of the Poisson equations in one, two and three dimensions. We will walk through the main issues related to the finite element method, including the basic setup, theory, basic data structures and some solution methods.

The presentation in this chapter is meant to be concise and informative, rather than rigorous.

4.1 A model problem and variational formulation

4.1.1 The heat equation

We consider a solid material that occupies a region $\Omega \subset \mathbb{R}^3$ with the boundary $\partial\Omega$. Denote the temperature at point $x \in \mathbb{R}^3$ at the time instant t by $u(x, t)$. Suppose r is the heat received per unit volume by radiation. For any domain $D \subset \Omega$, the heat energy Q contained in the material is

$$(4.1) \quad Q = \int_D \rho \kappa_s u(x, t) dx,$$

where ρ is the density of the material and κ_s a physical characteristic of the material called the specific heat capacity. By Fourier's law, the flow rate of heat energy through a surface is proportional to the negative temperature gradient across the surface, i.e. $-\sigma \nabla u$, where σ is the thermal conductivity. Then the heat transfer in whole D will be governed by

$$(4.2) \quad \frac{\partial Q}{\partial t} = \int_{\partial D} \sigma \mathbf{n} \cdot \nabla u ds + \int_D r dx.$$

Suppose the material is isotropic, that means the physical parameters ρ, κ_s, σ are constants. Together with (4.1) and (4.2), it holds

$$\int_D \rho \kappa_s \frac{\partial u}{\partial t} dx = \int_{\partial D} \sigma \mathbf{n} \cdot \nabla u ds + \int_D r dx.$$

Using integration by part, we have

$$\int_D \rho \kappa_s \frac{\partial u}{\partial t} dx = \int_D \sigma \Delta u ds + \int_D r dx.$$

Because of the arbitrariness of D , the last equality equals to

$$\rho\kappa_s \frac{\partial u}{\partial t} = \sigma \Delta u + r \quad \text{in } \Omega \times (0, +\infty).$$

Now let $\kappa = \frac{\sigma}{\rho\kappa_s}$ and $f = \frac{r}{\rho\kappa_s}$. Therefore we get the heat equation as follows.

$$\frac{\partial u}{\partial t} - \kappa \Delta u = f \quad \text{in } \Omega \times (0, +\infty).$$

4.1.2 The Poisson equation

One of the most important and frequently encountered equations in many mathematical models of physical phenomena is the Poisson equation. Just as an example, the solution of this equation gives the electrostatic potential for a given charge distribution. It also frequently appears in structural mechanics, theoretical physics, and many other areas of science and engineering. It is named after the French mathematician Siméon-Denis Poisson. The Poisson equation is

$$(4.3) \quad -\Delta u = f, \quad x \in \Omega.$$

Here Δ is the Laplacian operator given by

$$\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \dots + \frac{\partial^2}{\partial x_d^2}$$

and Ω is a d -dimensional domain (e.g., a rod in 1d, a plate in 2d and a volume in 3d). The unknown function is the electrostatic potential u and the given data is the charge distribution f . If the charge distribution vanishes, this equation becomes Laplace's equation and the solution to the Laplace equation is called harmonic function.

Denote with $\partial\Omega$ the boundary of Ω , and with $\mathbf{n} = (n_1, \dots, n_d)^T$ the unit normal vector to $\partial\Omega$ pointing outside of Ω . For the Poisson equation, the following types of boundary conditions are often used. These are:

- Dirichlet (or first type) boundary condition:

$$(4.4) \quad u|_{\partial\Omega} = g_D$$

- Neumann (or second type) boundary condition:

$$(4.5) \quad \frac{\partial u}{\partial \mathbf{n}} = \nabla u \cdot \mathbf{n}|_{\partial\Omega} = g_N$$

- Mixed boundary condition:

$$(4.6) \quad u|_{\Gamma_D} = g_D, \quad \text{and} \quad \nabla u \cdot \mathbf{n}|_{\Gamma_N} = g_N$$

where $\Gamma_D \cup \Gamma_N = \partial\Omega$.

- Robin (or third type) boundary condition:

$$(4.7) \quad (\alpha u + \beta \nabla u \cdot \mathbf{n})|_{\partial\Omega} = g_R.$$

Dirichlet and Neumann boundary conditions are two special cases of the mixed boundary condition by taking $\Gamma_D = \partial\Omega$ or $\Gamma_N = \partial\Omega$, respectively. Mixed boundary condition itself is a special example of Robin boundary condition by taking the coefficient $\alpha = \chi_{\Gamma_D}$ and $\beta = \chi_{\Gamma_N}$, where χ is the characteristic function defined as usual. When the boundary data is zero, we call it homogenous.

For convenience of exposition, we will consider the following Poisson equation with homogeneous mixed boundary condition:

$$(4.8) \quad \begin{cases} -\Delta u = f, & x \in \Omega \\ u = 0, & x \in \Gamma_D \\ \frac{\partial u}{\partial \mathbf{n}} = 0, & x \in \Gamma_N. \end{cases}$$

Remark 5. For the Poisson equation with Neumann boundary condition

$$(4.9) \quad -\Delta u = f \text{ in } \Omega, \quad \frac{\partial u}{\partial \mathbf{n}} = g \text{ on } \partial\Omega,$$

there is a compatible condition for f and g :

$$(4.10) \quad \int_{\Omega} f \, dx = - \int_{\Omega} \Delta u \, dx = \int_{\partial\Omega} \frac{\partial u}{\partial \mathbf{n}} \, dS = \int_{\partial\Omega} g \, dS.$$

The solution is unique up to a constant. Namely if u is a solution to (4.9), so is $u + c$.

The equation (4.3) together with any of the boundary conditions given in (4.4)-(4.7), is called well-posed, if the solution exists and is unique, and moreover, depends continuously on the given data f . In other words, the differential problem above is well posed if its solution is unique and “small” perturbations in f lead to “small” perturbations in the solution u .

The problem of solving the Poisson equation, together with the boundary conditions is called a second order *boundary value problem*. Second order is to indicate that the highest order of the derivatives of u which appears in the equation is 2.

Example 3. If the spatial dimension is 1, i.e. $\Omega = (a, b)$ is an interval, we may set as boundary conditions $u(a) = 0, u'(b) = 0$. This corresponds to a Dirichlet condition on a part of the boundary ($x = a$) and Neumann condition on another part of the boundary ($x = b$). The boundary value problem then is: Find u , such that

$$(4.11) \quad -u'' = f, \quad x \in (a, b), \quad u(a) = 0, \quad u'(b) = 0.$$

The above considerations are not that rigorous, and the problem of existence and uniqueness of solution of a boundary value problem could be quite subtle. We need to seek the solution in the right space and the right topology, i.e. the “=” holds in what sense. It turns out that the point-wise topology is sometimes too strong.

Example 4. If the spatial dimension is $d = 1, 2, 3, \dots$, i.e. $\Omega = (0, 1)^d$ is a domain, we may set as Dirichlet boundary condition $u|_{\partial\Omega} = 0$. The boundary value problem then is: Find u , such that

$$(4.12) \quad \begin{aligned} -\Delta u &= f, & \text{in } \Omega, \\ u|_{\partial\Omega} &= 0, & \text{on } \partial\Omega. \end{aligned}$$

To give a brief idea of finite element method, we will consider the linear finite element method for this model problem in the rest of this chapter.

4.2 Variational formulation and Galerkin method

4.2.1 Sobolev spaces

Throughout this book, we will often use Sobolev spaces. Here we only briefly introduce Sobolev spaces $H^m(\Omega)$ ($0 \leq m \leq 2$). $L^2(\Omega) = H^0(\Omega)$ is the space of functions which are square integrable. $H^1(\Omega)$ consists of functions in $L^2(\Omega)$ whose all first order derivatives are in L^2 and $H^2(\Omega)$ consists of functions in $H^1(\Omega)$ whose all second order derivatives are in $L^2(\Omega)$

Thus, for a given $v \in H^1(\Omega)$, we have that

$$\int_{\Omega} v^2 < \infty, \quad \text{and} \quad \int_{\Omega} |\nabla v|^2 = \int_{\Omega} \sum_{k=1}^d \left| \frac{\partial v}{\partial x_k} \right|^2 < \infty,$$

and for $v \in H^2(\Omega)$, we have

$$\int_{\Omega} v^2 < \infty, \quad \int_{\Omega} |\nabla v|^2 < \infty, \quad \text{and} \quad \int_{\Omega} |\nabla^2 v|^2 = \sum_{k=1}^d \sum_{j=1}^d \int_{\Omega} \left| \frac{\partial^2 v}{\partial x_k \partial x_j} \right|^2 < \infty.$$

We can then measure how “large” an element in each of these spaces is, by introducing an analogue to the length (norm) of a vector in Euclidean space. Setting

$$\|v\|_{0,\Omega}^2 = \int_{\Omega} v^2,$$

the norm $\|\cdot\|_{1,\Omega}$ in $H^1(\Omega)$ is defined as follows:

$$\|v\|_{1,\Omega}^2 = \|v\|_{0,\Omega}^2 + |v|_{1,\Omega}^2, \quad \text{where} \quad |v|_{1,\Omega}^2 = \int_{\Omega} |\nabla v|^2.$$

The quantity $|v|_{1,\Omega}$ is called H^1 -seminorm. Similarly in $H^2(\Omega)$ we have the following norm and semi-norm:

$$\|v\|_{2,\Omega}^2 = \|v\|_{1,\Omega}^2 + |v|_{2,\Omega}^2, \quad \text{where} \quad |v|_{2,\Omega}^2 = \int_{\Omega} \sum_{k=1}^d \sum_{j=1}^d \left| \frac{\partial^2 v}{\partial x_k \partial x_j} \right|^2.$$

$H^1(\Omega)$ is the space of functions, which are square integrable and have square integrable gradient. Later we also need another functions space, with square integrable second derivatives, since for some Ω and f , the solution u not only has square integrable gradient, but also square integrable second order derivatives. In such case, if the function u , its gradient ∇u , and its Hessian are all square integrable, then we say that u belongs to the space $H^2(\Omega)$. The superscript indicates that all derivatives up-to and including second order are square integrable.

4.2.2 Variational formulations

The finite element method for the approximation of (4.11) is based on an equivalent variational formulation of (4.11)

Lemma 11. *Assume that u is continuous in $(0, 1)$, then the following statements are equivalent*

(1) $u(x) = 0$.

(2) $\int_0^1 u(x)v(x)dx = 0$ for any smooth (compactly supported) function v in $(0, 1)$.

Proof. It is obvious that (2) holds if $u(x) = 0$. Assume (2), if there exists $x_0 \in (0, 1)$ that $u(x_0) \neq 0$, say $u(x_0) \geq 0$. Since $u(x)$ is continuous, there exists $\delta > 0$ that $u(x) > 0$ for $x \in (x_0 - \delta, x_0 + \delta)$. Let $v(x) > 0$ be a smooth function on $(x_0 - \delta, x_0 + \delta)$. Then

$$\int_0^1 u(x)v(x)dx = \int_{x_0-\delta}^{x_0+\delta} u(x)v(x)dx > 0,$$

which is in contradiction with (2). Thus, $u(x) = 0$ which completes the proof. \square

This simple fact asserts that the point-wise identity in the first statement is equivalent to an average identity in the second statement.

Let us now use (4.11) to illustrate how to derive a variational formulation. We first introduce the following functional space:

$$(4.13) \quad V = \{v \in H^1(\Omega) : v(a) = 0\}.$$

Multiplying any function $v \in V$ on both hand sides of (4.11) and integrating by parts, we get

$$(4.14) \quad \int_a^b f v dx = - \int_0^1 u'' v dx = \int_0^1 u' v' dx.$$

Denote

$$a(u, v) = \int_0^1 u' v' dx.$$

We then have the variational formulation for (4.11) as follows: Find $u \in V$ such that

$$(4.15) \quad a(u, v) = (f, v) \quad \forall v \in V.$$

This means that u is a solution of (4.15) if it is a solution of (4.11). Conversely, by reversing the order of above derivation and using the variational principle stated in Lemma 11, we can see that u is a solution of (4.11) if it is a solution of (4.15) (and it is sufficiently smooth). For this reason, we say that (4.11) is equivalent to (4.15).

Consider the variational formulation for the two dimensional problem (4.12). Similarly, the variational formulation is also in the form of (4.15) with

$$(4.16) \quad a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v$$

and

$$(4.17) \quad V = \{v \in H^1(\Omega) : v|_{\partial\Omega} = 0\}.$$

The variational (weak) form can also be obtained by minimizing a quadratic functional (potential energy of the charged body Ω). Again we consider (4.12). We introduce the potential energy associated with the charge distribution f , as follows

$$E(v) = \frac{1}{2} \int_{\Omega} |\nabla v|^2 - \int_{\Omega} f v.$$

We claim that the function (the potential), u_{\min} which minimizes the energy $E(v)$ over $v \in H_0^1(\Omega)$ is also a solution to the variational form (4.15)-(4.17). Indeed, it is simple to show that the minimizer, u_{\min} , of $E(v)$ is unique (this is a quadratic functional over $H_0^1(\Omega)$). Moreover, this minimizer should zero out the first variation of $E(v)$, that is:

$$\lim_{\delta \rightarrow 0} \frac{E(u_{\min} + \delta w) - E(u_{\min})}{\delta} = 0, \quad \forall w \in H_0^1(\Omega).$$

Computing the limit gives exactly equation (4.15)-(4.17) for u_{\min} . Thus, the solution u of the Poisson equation and also the minimizer of $E(v)$ satisfy the same relation (4.15)-(4.17) for all functions $w \in H_0^1(\Omega)$, and they must be the same.

Another question that arises naturally is whether the solution to the weak form (4.15)-(4.17) will also satisfy the classical Poisson equation (4.12) for every point $x \in \Omega$. What we can see immediately, is that a function u needs not have second order derivatives in order to satisfy (4.15)-(4.17). Both right side and left side of this equation are finite and make sense if u has finite energy, namely $E(u) < \infty$. In other words, the Poisson equation is better to be understood in certain average/integral sense and not point-wise; and only first derivatives not second ones are required to be square integrable.

The PDE theory tells us that for convex Ω , square integrable f , and Dirichlet boundary conditions the solution of (4.15)-(4.17) also has square integrable second order derivatives, and will be a solution to (4.12). How many derivatives (or how *regular*) the solution to (4.15)-(4.17) has could be a difficult question. Mathematical results on this are known as *regularity* results in the PDE theory.

For the Poisson equation quite a bit is known about the regularity of the solution and how it depends on the domain Ω , the data f and the type of boundary conditions. For example, for non-convex domain in 2 spatial dimensions (such as L -shaped domain) it is known that the solution to (4.15)-(4.17) needs not have square integrable second derivatives, and in fact it does not have such regularity, even if f is smooth.

The equation (4.15)-(4.17) is known as the *variational form* or *weak form*, since the solution may have less derivatives (less regularity), than the required one by the *strong form* (4.12). In what follows by solution to Poisson equation, we mean the solution to the variational form (4.15)-(4.17). In some very idealized cases, for example Ω being a ball and for simple f , the solution to (4.15)-(4.17) can be found explicitly. However, in most applications this is not the case and quantitative information about the solution is obtained using numerical methods. Among the numerical methods known for the solution of such equations, the Finite Element Method (FEM) seems to be the method of choice in most applications. This is due to the fact that this method correctly represents the underlying physical principles (such as minimization of a potential energy), and also is applicable to domains Ω with complicated geometry and various types of boundary conditions.

4.2.3 Regularity of the solution

Consider the domain $\Omega \subset \mathbb{R}^2$ which is defined in a polar coordinate as $\Omega = \{(r, \theta) : 0 < r < 1 \text{ and } 0 < \theta < \frac{\pi}{\beta}\}$. Obviously if $\beta \geq 1$ then Ω is convex, while if $0 < \beta < 1$ then Ω violates the condition of H^2 -regularity result. Set $v = r^\beta \sin(\beta\theta)$ as the imaginary part of the analytic function z^β i.e. $v = \text{Im}(z^\beta)$. According to the properties of analytic function, we know $\Delta v = 0$. With this fact, it is easy to verify that $u = (1 - r^2)v$ is the solution of the Poisson equation with $f = 4(1 + \beta)v \in L^2(\Omega)$.

Now we check the regularity of u . The only possible singularity is at the origin. When r is near 0, the second derivative $D^\alpha u \sim r^{\beta-2}$ for any $|\alpha| = 2$. Considering the integral

$$\int_{\Omega} |D^\alpha u|^2 dx dy \lesssim \int_0^1 |D^\alpha u|^2 r dr = \int_0^1 r^{2(\beta-2)+1} dr$$

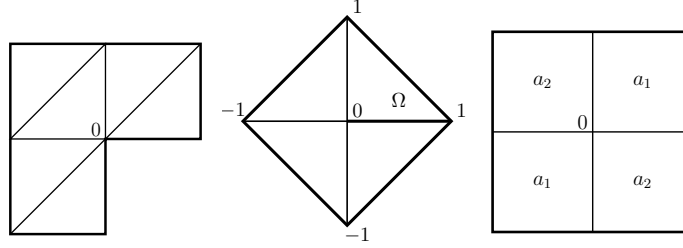


Fig. 4.1. Geometric explanation of barycentric coordinates

Therefore $u \in H^2(\Omega)$ if and only if $2(\beta - 2) + 1 > -1$, i.e., $\beta > 1$. Namely the domain Ω is convex. When β is fixed, by the same calculation, we see $u \in H^s(\Omega)$ for any $s < 1 + \beta$.

For general elliptic problem

$$(4.18) \quad -\operatorname{div}(A\nabla u) = f \text{ in } \Omega,$$

the lack of regularity could come from the discontinuity of the coefficients of A . See the example designed by Kellogg [?] with discontinuous diffusion coefficient.

Consider the partial differential equation (4.18) with $\Omega = (-1, 1)^2$ and the coefficient matrix A is piecewise constant: in the first and third quadrants, $A = a_1 I$; in the second and fourth quadrants, $A = a_2 I$. For $f = 0$, the exact solution in polar coordinates has been chosen to be $u(r, \theta) = r^\gamma \mu(\theta)$, where

$$\mu(\theta) = \begin{cases} \cos\left(\left(\frac{\pi}{2} - \sigma\right)\gamma\right) \cos\left(\left(\theta - \frac{\pi}{2} + \rho\right)\gamma\right) & \text{if } 0 \leq \theta \leq \frac{\pi}{2}, \\ \cos(\rho\gamma) \cos\left(\left(\theta - \pi + \sigma\right)\gamma\right) & \text{if } \frac{\pi}{2} \leq \theta \leq \pi, \\ \cos(\sigma\gamma) \cos\left(\left(\theta - \pi - \rho\right)\gamma\right) & \text{if } \pi \leq \theta \leq \frac{3\pi}{2}, \\ \cos\left(\left(\frac{\pi}{2} - \rho\right)\gamma\right) \cos\left(\left(\theta - \frac{3\pi}{2} - \sigma\right)\gamma\right) & \text{if } \frac{3\pi}{2} \leq \theta \leq 2\pi, \end{cases}$$

and the constants

$$\gamma = 0.1, \quad \rho = \pi/4, \quad \sigma = -14.9225565104455152,$$

and

$$a_1 = 161.4476387975881, \quad a_2 = 1.$$

For this example, we see $u \in H^{1+\gamma}(\Omega)$.

4.2.4 Galerkin methods

The finite element methods have been introduced as methods for approximate solutions of variational problems. Again, let us consider the model problem, which we have already discussed: approximating the solution to Laplace equation with Dirichlet boundary conditions, written in variational form, on a Lipschitz domain Ω . The corresponding model variational problem is: Find $u \in V$, such that

$$(4.19) \quad a(u, v) = f(v), \quad \forall v \in V.$$

Here V is a Hilbert space, $a(\cdot, \cdot) : V \times V \mapsto \mathbb{R}$ is a continuous bilinear form, $f(\cdot)$ is a continuous linear form. The aforementioned example (Laplace equation) corresponds to $V = H_0^1(\Omega)$, and

$$(4.20) \quad a(u, v) = \int_{\Omega} (\nabla u, \nabla v), \quad f(v) = \int_{\Omega} f v, \quad f \in L^2(\Omega)$$

Clearly, in such case $a(u, u) = |u|_{1, \Omega}^2$.

Lemma 12. *Poincaré inequality:*

$$\|v\|_{1, \Omega} \lesssim |v|_{1, \Omega}, \quad \forall v \in H_0^1(\Omega)$$

Proof. We first assume that the domain Ω contains the origin point $(0, 0)$ for 2d or $(0, 0, 0)$ for 3d. Note that $\frac{\partial(x_i v^2)}{\partial x_i} = v^2 \frac{\partial x_i}{\partial x_i} + 2x_i v \frac{\partial v}{\partial x_i}$. Take integration over Ω , we have

$$\begin{aligned} \|v\|_0^2 &= -2 \int_{\Omega} x_i v \frac{\partial v}{\partial x_i} + \int_{\Omega} \frac{\partial(x_i v^2)}{\partial x_i} \\ &= -2 \int_{\Omega} x_i v \frac{\partial v}{\partial x_i} \quad (v \in H_0^1(\Omega)) \\ &\leq 2(\max_{x \in \Omega} |x_i|) \|v\|_0 |v|_1. \end{aligned}$$

Thus,

$$\|v\|_0 \leq 2 \text{diam}(\Omega) |v|_1 \leq C(\Omega) |v|_1.$$

If the domain do not contain the original point, then we choose a point $(x_{0,1}, x_{0,2}) \in \Omega$ for 2d or $(x_{0,1}, x_{0,2}, x_{0,3}) \in \Omega$ for 3d. And note that $\frac{\partial((x_i - x_{0,i})v^2)}{\partial x_i} = v^2 \frac{\partial(x_i - x_{0,i})}{\partial x_i} + 2(x_i - x_{0,i})v \frac{\partial v}{\partial x_i} = v^2 + 2(x_i - x_{0,i})v \frac{\partial v}{\partial x_i}$. By similar process, we can get the inclusion. \square

The Poincaré inequality implies that $a(\cdot, \cdot)$ is an inner product on V , and thus the problem (4.19) has a unique solution.

Later, in addition to the above model problem, we will also consider approximations of problems with solutions in $H(\text{curl}; \Omega)$ and $H(\text{div}; \Omega)$. Thus the general form of $a(\cdot, \cdot)$ will be

$$(4.21) \quad a(u, v) = \int_{\Omega} \gamma_0 u v + \int_{\Omega} \gamma_1 D u D v,$$

where $D = \nabla, \text{curl}$ or div . Here γ_0 is a non-negative function and γ_1 is a positive function.

We now consider a class of methods, known as *Galerkin methods* which are used to approximate the solution to (4.19). Consider a finite dimensional subspace V_h of V , $V_h \subset V$, and let $V_h = \text{span}\{\varphi_1, \dots, \varphi_{n_h}\}$. With this setting, we have that $\dim V_h = n_h$. An approximate solution is then defined as follows: Find $u_h \in V_h$, such that

$$(4.22) \quad a(u_h, v_h) = f(v_h), \quad \text{for all } v_h \in V_h.$$

Using the basis in V_h , we may write

$$u_h = \sum_{k=1}^{n_h} \alpha_k \varphi_k.$$

Using the fact that $a(\cdot, \cdot)$ is bilinear and $f(\cdot)$ is linear, we obtain the following linear algebraic system for the unknown coefficients $\{\alpha_k\}_{k=1}^{n_h}$:

$$(4.23) \quad \sum_{k=1}^{n_h} a(\varphi_k, \varphi_j) \alpha_k = f(\varphi_j) \quad j = 1, \dots, n_h.$$

The right hand side of these equations (4.23) is known as *load vector* $F = \{f(\varphi_k)\}_{k=1}^{n_h} \in \mathbb{R}^{n_h}$, and the matrix defining the linear system $\{a(\varphi_j, \varphi_k)\}_{j,k=1}^{n_h}$ is known as *stiffness matrix*.

In the case when $a(\cdot, \cdot)$ defines an inner product, the following estimate is immediate consequence of the *projection theorem* in Hilbert spaces:

Theorem 24. *Let V be a Hilbert space, $a(\cdot, \cdot)$ be a continuous, symmetric, bilinear form defining an inner product on V . Let $f(\cdot)$ be a continuous linear form. If u is solution to the problem (4.19) and u_h is a solution to (4.22), then the following equality holds:*

$$(4.24) \quad \|u - u_h\|_a = \inf_{v \in V_h} \|u - v\|_a,$$

where $\|u - u_h\|_a^2 = a(u - u_h, u - u_h)$.

Proof. First, one sees that

$$a(u - u_h, v) = 0, \quad \forall v \in V_h.$$

Thus, for any $v \in V_h$ we have

$$\begin{aligned} \|u - u_h\|_a^2 &= a(u - u_h, u - u_h) \\ &= a(u - u_h, u - v) \\ &\leq \|u - u_h\|_a \|u - v\|_a. \end{aligned}$$

Taking the infimum on both sides of this equation then gives.

$$\|u - u_h\|_a \leq \inf_{v \in V_h} \|u - v\|_a.$$

The proof is complete, since

$$\|u - u_h\|_a \geq \inf_{v \in V_h} \|u - v\|_a,$$

by the definition of infimum. \square

By looking at the above theorem, we see that the Galerkin approximation u_h to the solution of (4.19) is quite natural and reasonable. Indeed this theorem states that: solving the equations (4.22) gives the best approximation to the unknown solution u with elements from V_h in the norm $\|\cdot\|_a$.

However, the above considerations are too general to be used in practice. Moreover, there are several important questions, which come immediately to mind, and we list some of them below.

- How to choose the finite dimensional space V_h in which an approximate solution is sought? Finite element spaces, which we introduce later provide a systematic way of choosing family of spaces V_h consisting of piece-wise polynomial functions.
- Is the problem (4.23) well posed and what can we say about the approximability, namely, is it true that for a suitable norm $\|\cdot\|$, we have that $\|u - u_h\| \rightarrow 0$, when $n_h \rightarrow \infty$. Such question is also addressed in this chapter, after introducing the finite element spaces.
- How to solve the resulting linear system of equations (4.23) at optimal computational cost, with number of arithmetic operations proportional to the dimension of V_h .
- How to use a current approximation u_h and obtain a better one, by increasing the dimension of V_h , but keep the computational cost as low as possible. This is the focus of the adaptive FE techniques.

Before we discuss error estimates in the finite element method, we need to introduce the relevant finite element spaces. We will focus here on the linear finite elements, just to fix ideas. These spaces are constructed by first triangulating Ω , that is covering with simplexes (intervals in 1d, triangles in 2d, and tetrahedrons in 3d). Then V_h is defined as the space of piece-wise linear continuous functions, where *piece-wise linear* means that the functions are linear on each simplex. The triangulations that we use are not arbitrary, and they possess certain properties, which we discuss next.

4.3 Finite element spaces

4.3.1 Triangulations and barycentric coordinates

In this section, we shall discuss the triangulations used in finite element methods. We would like to distinguish two structures related to the triangulation: 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. Correspondingly there are two basic data structures used to represent a triangulation.

Let $\mathbf{x}_i = (x_{1,i}, \dots, x_{d,i})^t, i = 1, \dots, d+1$ be $d+1$ points in \mathbb{R}^d which do not all lie in one hyper-plane. The *convex hull* of the $d+1$ points $\mathbf{x}_1, \dots, \mathbf{x}_{d+1}$

$$(4.25) \quad \tau := \left\{ \mathbf{x} = \sum_{i=1}^{d+1} \lambda_i \mathbf{x}_i \mid 0 \leq \lambda_i \leq 1, i = 1 : d+1, \sum_{i=1}^{d+1} \lambda_i = 1 \right\}$$

is defined as a *geometric d -simplex* generated (or spanned) by the vertices $\mathbf{x}_1, \dots, \mathbf{x}_{d+1}$. For example, a triangle is a 2-simplex and a tetrahedron is a 3-simplex. For an integer $0 \leq m \leq d-1$, an m -dimensional face of τ is any m -simplex generated by $m+1$ of the 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 \mathbf{x}_i will be denoted by F_i .

The numbers $\lambda_1(\mathbf{x}), \dots, \lambda_{d+1}(\mathbf{x})$ are called *barycentric coordinates* of \mathbf{x} with respect to the $d+1$ points $\mathbf{x}_1, \dots, \mathbf{x}_{d+1}$. There is a simple geometric meaning of the barycentric coordinates. Given a $\mathbf{x} \in \tau$, let $\tau_i(\mathbf{x})$ be the simplex with vertices \mathbf{x}_i replaced by \mathbf{x} . Then it can be easily shown that

$$(4.26) \quad \lambda_i(\mathbf{x}) = |\tau_i(\mathbf{x})|/|\tau|,$$

where $|\cdot|$ is the Lebesgue measure in \mathbb{R}^d , namely area in two dimensions and volume in three dimensions. Note that $\lambda_i(\mathbf{x})$ is affine function of \mathbf{x} and vanishes on the face F_i . We list the four basic properties of barycentric coordinate below:

1. $0 \leq \lambda_i(x) \leq 1$;
2. $\sum_{i=1}^{d+1} \lambda_i = 1$;
3. $\lambda_i \in P_1(\tau)$;
4. $\lambda_i(\mathbf{x}_j) = \delta_{ij}$.

4.3.2 Shape-regular and quai-uniform triangulations

Given a bounded tetrahedral domain $\Omega \subset \mathbb{R}^d$. A geometric triangulation (also called mesh or grid) \mathcal{T} of Ω is a set of d -simplices such that

$$\cup \tau = \overline{\Omega}, \quad \text{and} \quad \overset{\circ}{\tau}_i \cap \overset{\circ}{\tau}_j = \emptyset.$$

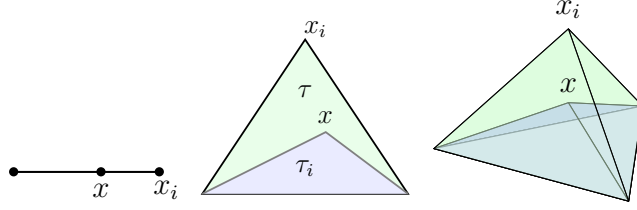


Fig. 4.2. Geometric explanation of barycentric coordinate

Denote

$$h_\tau = \text{diam}(\tau), \quad h = \max_{\tau \in \mathcal{T}_h} h_\tau; \quad \underline{h} = \min_{\tau \in \mathcal{T}_h} h_\tau,$$

The mesh domain $\Omega_h \stackrel{\text{def}}{=} \cup\{\tau : \tau \in \mathcal{T}_h\}$ is either equal to Ω (if Ω is a tetrahedral) or close to Ω in the sense that

$$\max_{x \in \tau \cap \partial\Omega_h} \text{dist}(x, \partial\Omega) \leq \sigma_0 h_\tau^2, \quad \forall \tau \in \mathcal{T}_h, \tau \cap \partial\Omega_h \neq \emptyset;$$

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).

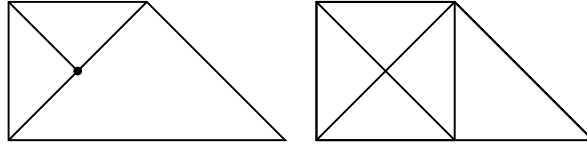


Fig. 4.3. Two triangulations. The left is non-conforming and the right is conforming.

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

$$(4.27) \quad \max_{\tau \in \mathcal{T}} \frac{\text{diam}(\tau)^d}{|\tau|} \leq c_0, \quad \forall \mathcal{T} \in \mathcal{T},$$

where $\text{diam}(\tau)$ is the diameter of τ and $|\tau|$ is the measure of τ in \mathbb{R}^d . This assumption can also be represented as

$$(4.28) \quad \sup_{h \in \mathbb{N}} \max_{\tau \in \mathcal{T}_h} \frac{h_\tau}{\rho_\tau} \leq \sigma_1$$

where ρ_τ denotes the radius of the ball inscribed in τ . 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 $\tau \in \mathcal{T} \in \mathcal{T}$. By (4.27), $h_\tau \approx \text{diam}(\tau)$ represents the size of an element $\tau \in \mathcal{T}$ for a shape regular triangulation $\mathcal{T} \in \mathcal{T}$.

In addition to (4.27), if

$$(4.29) \quad \frac{\max_{\tau \in \mathcal{T}} |\tau|}{\min_{\tau \in \mathcal{T}} |\tau|} \leq \rho, \quad \forall \mathcal{T} \in \mathcal{T},$$

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

The assumption (23.10) is a local assumption, as is meant by above definition, for $d = 2$ for example, it assures that each triangle will not degenerate into a segment in the limiting case. A triangulation satisfying this assumption is often called to be *shape regular*.

On the other hand, the assumption (4.29) is a global assumption, which says that the smallest mesh size is not too small compared with the largest mesh size of the same triangulation. By the definition, in a quasiuniform triangulation, all the elements are about the same size asymptotically.

Remark 6. In this course, unless otherwise noted, we restrict ourself to quasi-uniform simplicial triangulation. There are other type of meshes by partition the domain into quadrilateral (in 2-D), cubes (in 3-D), or other type of elements.

4.3.3 Piecewise linear finite element spaces

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

$$V_h := \{v \mid v \in C(\bar{\Omega}), \text{ and } v|_{\tau} \in P_1(\tau), \forall \tau \in \mathcal{T}_h\},$$

where $P_1(\tau)$ 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 has a square interable gradient, that is $u \in H^1(\Omega)$, and u is piecewise smooth, then u is continuous.

We always use n_h to denote the dimension of finite element spaces. For V_h , n_h is the number of vertices of the triangulation \mathcal{T}_h and for $V_{h,0}$, n_h is the number of interior vertices. For linear finite element spaces, we have the so called *a standard nodal basis functions* $\{\varphi_i, i = 1, \dots, n_h\}$ such that φ_i is piecewise linear (with respect to the triangulation) and $\varphi_i(x_j) = \delta_{i,j}$. Note that $\varphi_i|_{\tau}$ is the corresponding barycentric coordinates of x_i . See Figure 4.4 for an illustration in 2-D.

Therefore for any $v_h \in V_h$, we have the representation

$$v_h(x) = \sum_{i=1}^{n_h} v_h(x_i) \varphi_i(x).$$

Let us see how our construction looks like in one spatial dimension. Associated with the partition $\mathcal{T}_h = \{0 = x_0 < x_1 < \dots < x_{n_h} < x_{n_h+1} = 1\}$, we define a linear finite element space

$$V_{h,0} = \{v : v \text{ is continuous and piecewise linear w. r. t. } \mathcal{T}_h, v(0) = v(1) = 0\}.$$

A plot of a typical element of $V_{h,0}$ is shown in Fig. 4.5.

It is easily calculated (as we already mentioned), that the dimension of V_h is equal to the number of internal vertices, and the nodal basis functions spanning $V_{h,0}$ (for $i = 1, 2, \dots, n_h$) are (see also Fig. 4.4):

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

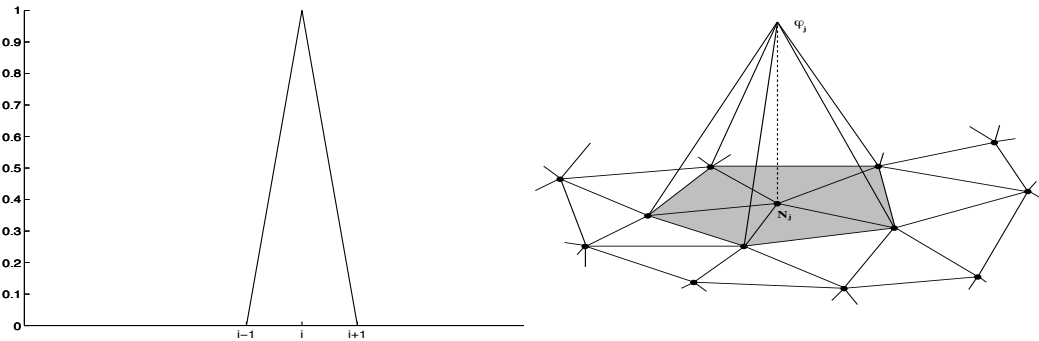


Fig. 4.4. Nodal basis functions in 1d and 2d

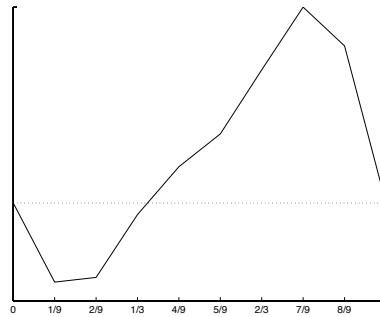


Fig. 4.5. Plot of a typical element from V_h .

4.3.4 Relationship with finite difference method

There exists some relation between the linear finite element and the finite difference method.

Example 5. Let us first consider a special example in one dimension. A direct calculation shows that, for the problem (4.11),

$$\int_0^1 \phi'_{j-1} \phi'_j = \int_0^1 \phi'_j \phi'_{j+1} = -\frac{1}{h}, \quad 1 < j < N$$

$$\int_0^1 (\phi'_j)^2 = \frac{2}{h}, \quad 1 \leq j \leq N.$$

Therefore, the finite element equation is reduced to

$$(4.30) \quad \frac{-\mu_{j-1} + 2\mu_j - \mu_{j+1}}{h} = f_j, \quad 1 \leq j \leq N,$$

where $f_j = \int_0^1 f \phi_j$.

Example 6. In two dimensions, we consider a unit square and uniform triangulations. Let $u_{i,j} = u(x_i, y_j)$. A direct calculation shows that the finite element equation is reduced to

$$(4.31) \quad 4u_{ij} - (u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}) = b_{i,j}.$$

For these two cases, the linear finite element method gives the same equations as the finite difference method.

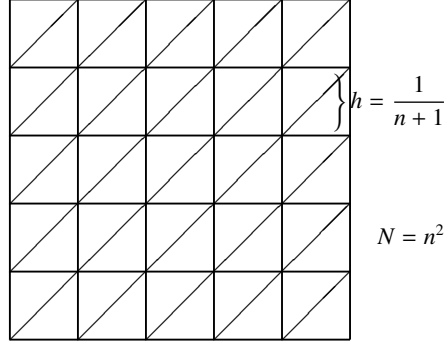


Fig. 4.6. A uniform grid

4.3.5 Nodal value interpolation

For a given smooth function u (continuous at least) we define its nodal interpolant $u_I \in V_h$, by

$$u_I(x_i) = \sum_{i=1}^{n_h} u(x_i)\varphi_i(x).$$

In another word, for a continuous u , the nodal interpolant is the piecewise linear function that takes the same values as u at the vertices of the triangulation.

Next we give an estimate on the *interpolation error* $u - u_I$. It turns out, that even though we do not know u , and hence u_I , we can still derive such an estimate. Recall the definition of the spaces $H^1(\Omega)$ and $H^2(\Omega)$, and the corresponding norms and semi-norms in these spaces. We mention here, that even though not obvious, the functions from $H^1(\Omega)$ in one dimension $\Omega = (0, 1)$, have well defined interpolant. For 2d and 3d this is no longer true, but it can be shown that the functions in $H^2(\Omega)$ have well defined interpolant. Such statements follow from the celebrated Sobolev embedding theorem, one of the important results in the theory of PDE, see [?] for more details.

We now state a general interpolation result, which holds in 1, 2, and 3 spatial dimensions.

Theorem 25. *Let Ω be a bounded Lipschitz domain in \mathbb{R}^d ($d = 1, 2, 3$). For any function $u \in H^2(\Omega)$, the following estimate holds:*

$$(4.32) \quad \|u - u_I\|_{0,\Omega} + h|u - u_I|_{1,\Omega} \leq Ch^2|u|_{2,\Omega}$$

where h is the maximum diameter of the elements in the triangulation, and C is an absolute constant, independent of u and h .

Proof. Let us first prove the estimate in one dimension and then give the proof for $1 \leq d \leq 3$.

One dimensional case: $d = 1$. Observe first that $e = (u - u_I)$ vanishes at the end points of each interval and e' is continuous, because e'' is square integrable. By the Rolle's theorem there exists $\xi_i \in (x_i, x_{i+1})$ such that $e'(\xi_i) = 0$. By the Fundamental Theorem of Calculus for $x \in (x_i, x_{i+1})$, we have that

$$e'(x) = \int_{\xi_i}^x e''(t) dt$$

Since u_I is linear on $[x_i, x_{i+1}]$ we have that $e''(t) = u''(t)$, and hence

$$[e'(x)]^2 = \left[\int_{\xi_i}^x u''(t) dt \right]^2.$$

Applying the Schwarz inequality to the right side then gives,

$$\begin{aligned} [e'(x)]^2 &\leq \left| \int_{\xi_i}^x 1^2 dt \right| \left| \int_{\xi_i}^x [u''(t)]^2 dt \right| \\ &\leq |\xi_i - x| \int_{\xi_i}^x [u''(t)]^2 dt. \end{aligned}$$

Integrating from x_i to x_{i+1} , and observing that

$$\int_{x_i}^{x_{i+1}} |\xi_i - x| dx = \frac{1}{2} [(\xi_i - x_i)^2 + (x_{i+1} - \xi_i)^2] \leq (x_i - x_{i+1})^2,$$

then gives

$$\int_{x_i}^{x_{i+1}} [e'(x)]^2 dx \leq (x_{i+1} - x_i)^2 \int_{x_i}^{x_{i+1}} [u''(t)]^2 dt.$$

Finally, summing up on $(0, 1)$ then leads to:

$$\begin{aligned} \|e'\|_{0,\Omega} &= \int_0^1 [e'(x)]^2 dx = \sum_{i=1}^{n_h-1} \int_{x_i}^{x_{i+1}} [e'(x)]^2 dx \\ &\leq \sum_{i=1}^{n_h-1} (x_{i+1} - x_i)^2 \int_{x_i}^{x_{i+1}} [u''(t)]^2 dt \\ &\leq \max_i (x_{i+1} - x_i)^2 \int_0^1 [u''(t)]^2 dt. \end{aligned}$$

Since $e(x_i) = 0$, for any $x \in (x_i, x_{i+1})$,

$$|e(x)| = \left| \int_{x_i}^x e'(t) dt \right| \leq h |e'|_{1,\tau} \leq h^2 |u|_{2,\tau}.$$

Summing up on $(0, 1)$ then leads to:

$$\|e\|_{0,\Omega} \leq h^2 |u|_{2,\Omega}.$$

This completes the proof of the estimate in one dimension.

General case. We now prove the result in general cases.

Let $x = (x^1, \dots, x^d)$ and $a_i = (a_i^1, \dots, a_i^d)$. Introducing the auxiliary functions

$$g_i(t) = u(a_i(t)), \text{ with } a_i(t) = a_i + t(x - a_i),$$

we have

$$g_i'(t) = (\nabla u)(a_i(t)) \cdot (x - a_i) = \sum_{l=1}^d (\partial_l u)(a_i(t))(x^l - a_i^l)$$

and

$$(4.33) \quad g_i''(t) = \sum_{k,l=1}^d (\partial_{kl}^2 u)(a_i(t))(x^k - a_i^k)(x^l - a_i^l).$$

By Taylor expansion

$$g_i(0) = g_i(1) - g_i'(1) + \int_0^1 t g_i''(t) dt.$$

Namely

$$u(a_i) = u(x) - (\nabla u)(x) \cdot (x - a_i) + \int_0^1 t g_i''(t) dt.$$

Note that

$$u_I(x) = \sum_{i=1}^{d+1} u(a_i) \lambda_i(x),$$

and

$$(4.34) \quad \sum_{i=1}^{d+1} \lambda_i(x) = 1, \text{ and } \sum_{i=1}^{d+1} (x - a_i) \lambda_i(x) = 0.$$

It follows that

$$(4.35) \quad (u_I - u)(x) = \sum_{i=1}^{d+1} \lambda_i(x) \int_0^1 t g_i''(t) dt.$$

Using the following change of variables

$$y = a_i(t) := a_i + t(x - a_i) : \tau \mapsto \tau_i^t \subset \tau \text{ with } dy = t^d dx,$$

we have

$$\begin{aligned} \|(\partial_{kl}^2 u)(a_i(t))\|_{0, \tau_i^t}^2 &= \int_{\tau} (\partial_{kl}^2 u)(y))^2 dx \\ &= \int_{\tau_i^t} (\partial_{kl}^2 u)(y))^2 \left| \frac{dx}{dy} \right| dy \\ &= t^{-d} \int_{\tau_i^t} (\partial_{kl}^2 u)(y))^2 dy \\ &\leq t^{-d} \|\partial_{kl}^2 u\|_{0, \tau}^2. \end{aligned}$$

Thus, by (4.33) and the fact that $|x^l - a_i^l| \leq h$, we obtain

$$\|g_i''(t)\|_{0,\tau} \leq h^2 \sum_{k,l=1}^d \|(\partial_{kl}^2 u)(a_i(t))\|_{0,\tau_i} \leq h^2 t^{-d/2} \|\partial_{kl}^2 u\|_{0,\tau}.$$

Now taking the $L^2(\tau)$ norm on both hand of sides of (4.35), we get

$$\begin{aligned} \|u_I - u\|_{0,\tau} &\leq h^2 \sum_{i=1}^{d+1} \max_{x \in \tau} |\lambda_i(x)| \int_0^1 t \|g_i''(t)\|_{0,\tau} dt \\ &\leq (d+1) \int_0^1 t^{1-d/2} dt h^2 \sum_{k,l=1}^d \|\partial_{kl}^2 u\|_{0,\tau} \\ &\leq \frac{2(d+1)}{4-d} h^2 \sum_{k,l=1}^d \|\partial_{kl}^2 u\|_{0,\tau} \\ &\leq \frac{4d(d+1)}{4-d} h^2 |u|_{2,\tau}. \end{aligned}$$

Now we prove the H^1 error estimate. By (4.34) and (4.35),

$$\sum_{i=1}^{d+1} \lambda_i(x) \nabla \int_0^1 t g_i''(t) dt = \sum_{i=1}^{d+1} \lambda_i(x) \nabla \left(u(a_i) - u(x) + (\nabla u)(x) \cdot (x - a_i) \right) = 0,$$

we get

$$(4.36) \quad (\nabla(u_I - u))(x) = \sum_{i=1}^{d+1} (\nabla \lambda_i)(x) \int_0^1 t g_i''(t) dt.$$

Note that $\nabla \lambda_i = -\frac{1}{d_i} \mathbf{n}_i$, where \mathbf{n}_i is the normal vector and d_i the perpendicular height of the edge opposite to the vertex a_i . Then the estimate for $|\nabla(u_I - u)|_{0,\tau}$ follows by a similar argument and the following obvious estimate

$$|\nabla \lambda_i| \lesssim \frac{1}{h}.$$

□

Remark 7. In one spatial dimension, the fact that u'' is square integrable implies that u' is continuous. This follows from the Sobolev embedding theorem, which we will state later, and the above proof works, i.e. we can apply Rolle's theorem. As we mentioned above, when the spatial dimension is 2 or 3, such relation is no longer true, that is, the Hessian of u being square integrable does not necessarily imply that the gradient of u is pointwise bounded, (or continuous).

4.4 Average interpolation for discontinuous functions

4.4.1 Sobolev embeddings

Let us discuss some special cases about Sobolev embedding

1. $H^2(\Omega) \hookrightarrow C(\overline{\Omega})$ if $1 \leq d \leq 3$. When $d = 4$, this relation does not hold any longer. For example, let $u(x) = \ln(\ln r)$ with $r^2 = \sum_{i=1}^4 x_i^2$. A direct computation gives $\partial_{x_i} u = x_i r^{-2} (\ln r)^{-1}$. Thus

$$\partial_{x_i x_j}^2 u = \delta_{ij} r^{-2} (\ln r)^{-1} - 2x_i x_j r^{-4} (\ln r)^{-1} - x_i x_j r^{-4} (\ln r)^{-2}$$

Since $x_i x_j r^{-2} \leq 1$, we have $\partial_{x_i x_j}^2 u \sim r^{-2} (\ln r)^{-1}$. Then we can claim that $u \in H^2(B(0, \frac{1}{2}))$ since

$$\int_{B(0, \frac{1}{2})} (\partial_{x_i x_j}^2 u)^2 dx \sim \int_0^{\frac{1}{2}} r^{-1} (\ln r)^{-2} dr = (\ln 2)^{-1} < +\infty.$$

On the other hand, it is straightforward that u is not continuous.

2. $W^{s,p}(\Omega) \hookrightarrow C(\Omega)$ if $sp > d$. If $sp = d$, the relation is not always true. When $s = 2$, $p = 1$ and $d = 2$, $W^{2,1}(\Omega) \hookrightarrow C(\Omega)$ is true; but $H^1(\Omega) \not\hookrightarrow C(\Omega)$.

Average interpolation

Denote the average interpolation operator A_τ by

$$A_\tau(v) = \frac{1}{|\tau|} \int_\tau v.$$

In this section, we will analyze the error $\|v - A_\tau(v)\|_{0,\tau}$.

Theorem 26. *Given any element $\tau \in T$, we have*

$$\|v\|_{0,\partial\tau} \lesssim h^{-1/2} \|v\|_{0,\tau} + h^{1/2} |v|_{1,\tau}$$

Proof. Let x_0 be the barycenter of τ . It is easy to see that

$$(x - x_0) \cdot \mathbf{n} \geq c_0 h_\tau, \quad \forall x \in \partial\tau.$$

Using Green formula and Cauchy-Swarchz inequality, we get

$$\begin{aligned} c_0 h \int_{\partial\tau} v^2 ds &\leq \int_{\partial\tau} v^2 (x - x_0) \cdot \mathbf{n} ds \\ &= \int_\tau \operatorname{div}((x - x_0)v^2) dx \\ &= \int_\tau (dv^2 + 2v\nabla v \cdot (x - x_0)) dx \\ &\lesssim \int_\tau v^2 dx + h_\tau \int_\tau |v\nabla v| dx \\ &\lesssim \int_\tau v^2 dx + h_\tau^2 \int_\tau |\nabla v|^2 dx. \end{aligned}$$

□

Lemma 13.

$$\|v - A_\tau(v)\|_{0,\tau} \lesssim h |v|_{1,\tau}.$$

Proof. Given any $x, y \in \tau$, we have

$$v(x) - v(y) = \int_0^1 (\nabla v)(y + t(x - y)) dt \cdot (x - y)$$

Let $z = y + t(x - y)$, then $dz = (1 - t)dy$, $x - z = (1 - t)(x - y)$, $|\tau_i| = (1 - t)^d |\tau|$. Taking the average with respect to y gives that

$$v(x) - A_\tau(v) = \frac{1}{|\tau|} \int_0^1 dt \int_\tau (\nabla v)(y + t(x - y)) \cdot (x - y) \leq \frac{h}{|\tau|} \int_0^1 dt \int_{\tau_i} dz (1 - t)^{1-d} |(\nabla v)(z)|$$

Take the L^2 -norm with respect to x variable on both hand side, we have

$$\|v(\cdot) - A_\tau(v)\|_{0,\tau} \leq \frac{h}{|\tau|} \int_0^1 (1 - t)^{1-d} dt \int_{\tau_i} \|(\nabla v)(y + t(\cdot - y))\|_{0,\tau} \lesssim h|v|_{1,\tau}.$$

□

Scott-Zhang interpolation

According to the embedding theorem, the nodal interpolation no longer makes sense in higher dimensions. However, we can still claim the error estimate above for $d \geq 4$. In this case, we need to introduce the famous Scott-Zhang interpolation.

For any element τ , $\lambda_1, \dots, \lambda_{d+1} \in P_1(\tau)$ be the barycentric coordinates. They form a basis of $P_1(\tau)$. There exists a dual basis $\{\psi_i\}_{i=1}^{d+1} \in P_1(\tau)$ such that

$$(\psi_i, \lambda_j) = \delta_{ij}$$

It is straightforward that

$$p(x) = \sum_{i=1}^{d+1} (\psi_i, p)_{0,\tau} \lambda_i(x), \quad \forall p \in P_1(\tau)$$

Given any nodal point a_i , pick an element τ_i with a vertex a_i , let $\psi_{a_i}^\tau$ be the dual basis function in $P_1(\tau)$ associated with the node a_i . Denote the Scott-Zhang interpolation $\Pi_h : V \rightarrow V_h$ by

$$\Pi_h v = \sum_{i=1}^N (\psi_{a_i}^{\tau_i}, v)_{0,\tau_i} \phi_i(x)$$

where ϕ_i are the nodal bases function.

Lemma 14. *It holds that*

1. $\Pi_h v_h = v_h, \quad \forall v_h \in V_h,$
2. $\|\Pi_h v\|_{0,\tau} \lesssim \|v\|_{0,\omega_\tau}, \quad \forall v \in L^2(\Omega),$
3. $\|v - \Pi_h v\|_{0,\tau} \lesssim h|v|_{1,\omega_\tau}$
4. $|\Pi_h v|_{1,\tau} \lesssim |v|_{1,\omega_\tau}, \quad \forall v \in H^1(\Omega),$
5. $\|v - \Pi_h v\|_{0,\tau} \lesssim h^2|v|_{2,\omega_\tau},$
6. $|v - \Pi_h v|_{1,\tau} \lesssim h|v|_{2,\omega_\tau}.$

Here ω_τ is the union of the element τ and its neighbors and N_{ω_τ} is the number of elements in ω_τ .

Proof.

1. It suffices to verify that $v_h(a_i) = (I_h v_h)(a_i)$ for all vertices a_i .
2. By definition, we have

$$\begin{aligned} \|\psi_{a_i}^{\tau_i}\|_{0,\tau_i} &= \sup_{c_j, j \neq i} \frac{(\psi_{a_i}^{\tau_i}, \lambda_i + \sum_{j \neq i} c_j \lambda_j)_{0,\tau_i}}{\|\lambda_i + \sum_{j \neq i} c_j \lambda_j\|} = \|\lambda_i\|_{0,\tau_i}^{-1}, \\ \|I_h v\|_{0,\tau} &= \left\| \sum_{i=1}^N (\psi_{a_i}^{\tau_i}, v)_{0,\tau_i} \phi_i(x) \right\|_{0,\tau} \\ &= \left\| \sum_{i=1}^3 (\psi_{a_i}^{\tau_i}, v)_{0,\tau_i} \phi_i(x) \right\|_{0,\tau} \\ &\lesssim \sum_{i=1}^3 \|\psi_{a_i}^{\tau_i}\|_{0,\tau_i} \|\phi_i\|_{0,\tau} \|v\|_{0,\tau_i} \\ &\lesssim \sum_{i=1}^3 \|v\|_{0,\omega_\tau} \end{aligned}$$

3. By quasiuniformity, there exists a constant $\gamma > 0$ such that

$$\omega_\tau \subset K(\tau) \quad \forall \tau \in \mathcal{T}_h$$

where $K(\tau) = x_\tau + \gamma(\tau - x_\tau)$ (with x_τ being the barycenter of τ) is the simplex that is similar to τ (with similarity constant being γ) and has the same barycenter x_τ .

By Sobolev extension theorem, it suffices to prove that

$$\|v - I_h v\|_{0,\tau} \lesssim h \|v\|_{1,K(\tau)} \quad \forall v \in H^1(\mathbb{R}^d).$$

Now given $\tau \in \mathcal{T}_h$, then

$$\|v - I_h v\|_{0,\tau} \lesssim h^{d/2} \|\hat{v} - \hat{I}_h v\|_{0,\hat{\tau}}.$$

Define $F(\hat{v}) = \|\hat{v} - \hat{I}_h v\|_{0,\hat{\tau}}$. Obviously $F(c) = 0$, $\forall c \in \mathbb{R}$. We conclude from Bramble-Hilbert Lemma that

$$F(\hat{v}) \lesssim |\hat{v}|_{1,K(\hat{\tau})}.$$

Therefore

$$\|v - I_h v\|_{0,\tau} \lesssim h^{d/2} |\hat{v}|_{1,K(\hat{\tau})} \lesssim h \|v\|_{1,K(\tau)}$$

The desired result then follows easily.

4. By the stability of I_h in L^2 sense, we have

$$\begin{aligned} |I_h v|_{1,\tau} &= |I_h(v - P_0 v)|_{1,\tau} \\ &\lesssim h^{-1} \|I_h(v - P_0 v)\|_{0,\tau} \\ &\lesssim h^{-1} (\|I_h(v - P_0 v) - (v - P_0 v)\|_{0,\tau} + \|(v - P_0 v)\|_{0,\tau}) \\ &\lesssim h^{-1} h (|v - P_0 v|_{1,\omega_\tau} + |v|_{1,\tau}) \\ &\lesssim |v|_{1,\omega_\tau}. \end{aligned}$$

5. Use similar techniques in 3.
6. Use similar techniques in 3.

□

There exists the following error estimate for the Scott-Zhang interpolation

$$|v - \Pi_h v|_{1,\tau} \leq Ch|v|_{2,\tau}.$$

Since $\Pi_h v \in V_h$ and $|u - u_h|_{1,\Omega} = \inf_{v_h \in V_h} |u - v_h|_{1,\Omega}$, we have

$$|u - u_h|_{1,\Omega} \leq |u - \Pi_h u|_{1,\Omega} \lesssim h|u|_{2,\Omega}$$

provided that $u \in H^2(\Omega)$.

4.5 Norm equivalence theorem

Lemma 15. Any l.s.c. (lower-semi-continuous) semi-norm $F(\cdot)$ on a Banach space V is bounded, namely

$$F(v) \lesssim \|v\|_V \quad \forall v \in V.$$

Proof. Let

$$V_k = \{v \in V : F(v) \leq k\}.$$

We claim that:

1. V_k is closed subset of V . In fact, if $v_n \in V_k$ and $v_n \rightarrow v$,

$$F(v) \leq \liminf_{n \rightarrow +\infty} F(v_n) \leq k < +\infty.$$

2. $V = \bigcup_{k=1}^{+\infty} V_k$

By Bair category Theorem, there exists $k, x_0 \in V, r > 0$ such that $x_0 + \frac{r}{2} \frac{v}{\|v\|} \in B(x_0, r) \subset V_k$, then

$$\begin{aligned} F\left(x_0 + \frac{r}{2} \frac{v}{\|v\|}\right) &\leq k, \\ F\left(\frac{r}{2} \frac{v}{\|v\|}\right) &\leq F(x_0) + k, \\ F(v) &\leq (F(x_0) + k) \frac{2}{r} \|v\|, \end{aligned}$$

which completes the proof. □

This result is sometimes known as Gelfand Lemma and its proof is similar to that of *uniform boundedness principle* in Banach space theory (c.f. [?]). The idea is to consider the decomposition $V = \bigcup_{k=1}^{\infty} \{v \in V : F(v) \leq k\|v\|_V\}$ and apply the Baire category theorem. The details of the proof are left to the readers.

Theorem 27. Assume that V is a Banach space, normed with $\|\cdot\|_V$, satisfying

$$(4.37) \quad \|v\|_V \lesssim F(v) + T(v), \quad \forall v \in V,$$

for some l.s.c seminorm $F(\cdot)$ on V and some functional T which is compact in the sense that any bounded and infinite set in V contains an infinite sequence $\{v_n\}$ so that $T(v_n - v_m) \rightarrow 0$ as $n, m \rightarrow \infty$. Then the following are true:

(i) For any other l.s.c. seminorm G over V , as long as $\ker(F) \cap \ker(G) = \{0\}$, then

$$\|v\|_V \approx F(v) + G(v) \quad \forall v \in V.$$

(ii) Let $V/\ker(F)$ be the ordinary quotient space and $\|\cdot\|_{V/\ker(F)}$ the corresponding quotient norm, then

$$\|v\|_{V/\ker(F)} \approx F(v), \quad \forall v \in V.$$

(iii) For any other l.s.c. seminorm B over V , as long as $\ker(F) \subset \ker(B)$, then

$$B(v) \lesssim F(v), \quad \forall v \in V.$$

Proof. An application of Lemma 15 gives that

$$F(v) + G(v) \lesssim \|v\|_V, \quad \forall v \in V.$$

To see the other direction of (i), we use a contradiction argument, namely we assume if what we want to show were not true, there would exist $\{v_n\} \subset V$ such that

$$(4.38) \quad \|v_n\|_V = 1, \quad \text{and} \quad F(v_n) + G(v_n) \rightarrow 0 \quad (\text{as } n \rightarrow \infty).$$

Since $\{v_n\}$ is bounded, by the compactness of the functional T , we may assume that

$$(4.39) \quad T(v_n - v_m) \rightarrow 0, \quad \text{as } n, m \rightarrow \infty.$$

It follows from the hypothesis (4.37), (4.38) and (4.39) that

$$\begin{aligned} \|v_n - v_m\|_V &\lesssim F(v_n - v_m) + T(v_n - v_m) \\ &\leq F(v_n) + F(v_m) + T(v_n - v_m) \rightarrow 0, \quad \text{as } n, m \rightarrow \infty. \end{aligned}$$

This means that $\{v_n\}$ is a Cauchy sequence on V . But V is a Banach space, hence there exists $v \in V$ so that

$$\lim_{n \rightarrow \infty} \|v_n - v\|_V = 0.$$

Since both F and G are l.s.c. seminorms, we conclude that

$$F(v) + G(v) \leq \limsup_{n \rightarrow \infty} F(v_n) + \limsup_{n \rightarrow \infty} G(v_n) = 0.$$

Hence $F(v) = G(v) = 0$, i.e., $v \in \ker(F) \cap \ker(G)$. By hypothesis, $v = 0$, but this contradicts to (4.38) which implies $\|v\|_V = 1$. This complete the proof of (i).

To prove (ii), we should first point out that $\ker(F)$ is obviously a subspace of V . Furthermore we note that $\ker(F)$ is finite dimensional, in fact, by the hypothesis (4.37), we can easily see that the unit ball in $\ker(F)$ is compact.

We need to show that

$$(4.40) \quad \inf_{\phi \in \ker(F)} \|v + \phi\|_V \lesssim F(v), \quad \forall v \in V.$$

as the other direction of the above inequality follows easily from Lemma 15. As mentioned above $m \stackrel{\text{def}}{=} \dim(\ker(F)) < \infty$, hence we can choose m functionals $\{f_k : k = 1, 2, \dots, m\}$ over $\ker(F)$ that forms a basis

for the dual space $(\ker(F))^*$. By Hahn-Banach theorem, we may assume that f_k 's are all defined on the whole of V . Set

$$G(v) = \sum_{k=1}^m |f_k(v)|.$$

Using the fact that $\{f_k\}$ forms a basis of $(\ker(F))^*$, for any $v \in V$, by solving a linear system with a nonsingular Gram matrix, we can find a $\phi_v \in \ker(F)$ such that

$$f_k(\phi_v) = -f_k(v), \quad k = 1, 2, \dots, m.$$

Namely

$$G(v + \phi_v) = 0.$$

Since G is obviously a l.s.c. seminorm on V and also $\ker(F) \cap \ker(G) = \{0\}$, we conclude from (i) that

$$\|v + \phi_v\|_V \lesssim F(v + \phi_v) + G(v + \phi_v) = F(v + \phi_v) = F(v),$$

which implies (4.40) and completes the proof of (ii).

We are now in a position to prove (iii). By Lemma 15, for any $v \in V$ and $\phi \in \ker(F)$

$$B(v + \phi) \lesssim \|v + \phi\|_V.$$

By hypothesis that $\ker(F) \subset \ker(B)$, we have $\phi \in \ker(B)$, thus $B(v) = B(v + \phi)$, therefore, using (ii), we get that

$$B(v) \lesssim \inf_{\phi \in \ker(F)} \|v + \phi\|_V \lesssim F(v),$$

completing the proof. \square

Given $m \geq 0$ and $p \geq 1$, we then take $V = W^{m+1,p}(\Omega)$, and

$$F(v) = |v|_{m+1,p}, \quad T(v) = \|v\|_{m,p},$$

we can see that (4.37) is trivially satisfied. Obviously, F is a l.s.c. seminorm by definition. As $W^{m+1,p}(\Omega)$ is compactly imbedded into $W^{m,p}(\Omega)$, T is a compact functional. Also it is straightforward to check that $\ker F = \mathcal{P}_m(\Omega)$. Choosing

$$G(v) = \|v\|_0,$$

it is derived by Theorem 27 that

$$\|v\|_{m+1,\Omega} \lesssim |v|_{m+1,\Omega} + \|v\|_{0,\Omega}.$$

We can also apply Theorem 27 to deduce the following (generalized) well-known results:

Theorem 28. 1. If G is a l.s.c. seminorm on V such that for $\phi \in \mathcal{P}_m(\Omega)$, $G(\phi) = 0$ iff $\phi = 0$, then

$$(4.41) \quad \|v\|_{m+1,p} \asymp G(v) + |v|_{m+1,p}, \quad \forall v \in W^{m+1,p}(\Omega),$$

2.

$$(4.42) \quad \inf_{\phi \in \mathcal{P}_m(\Omega)} \|v + \phi\|_{m+1,p} \asymp |v|_{m+1,p}, \quad \forall v \in W^{m+1,p}(\Omega),$$

3. [**Bramble-Hilbert Lemma**] If B is a l.s.c. seminorm on V such that for all $\phi \in \mathcal{P}_m(\Omega)$, $B(\phi) = 0$, then

$$(4.43) \quad B(v) \lesssim |v|_{m+1,p}, \quad \forall v \in W^{m+1,p}(\Omega),$$

(4.41) is often called *Sobolev norm equivalence theorem* in which G usually takes the form of $G(v) = \sum_{i=1}^m |f_i(v)|$, for some bounded linear functionals f_i 's. (4.43) is often called the Bramble-Hilbert Lemma.

A trivial consequence of the above theorem is

Corollary 5. *Assume that Γ is a measurable subset of $\partial\Omega$ with positive measure. Then*

$$\|v\|_{1,p} \lesssim |v|_{1,p} + \left| \int_{\Gamma} v ds \right| \quad \forall v \in W^{1,p}(\Omega).$$

The above results can obviously be extended to fractional order Sobolev space.

Theorem 29. *Let $\sigma \in (0, 1]$.*

1. *If G is a l.s.c. seminorm on V such that for $\phi \in \mathcal{P}_m(\Omega)$, $G(\phi) = 0$ iff $\phi = 0$, then*

$$(4.44) \quad \|v\|_{m+\sigma,p} \equiv G(v) + |v|_{m+\sigma,p} \quad \forall v \in W^{m+\sigma,p}(\Omega),$$

2.

$$(4.45) \quad \inf_{\phi \in \mathcal{P}_m(\Omega)} \|v + \phi\|_{m+\sigma,p} \equiv |v|_{m+\sigma,p}, \quad \forall v \in W^{m+\sigma,p}(\Omega),$$

3. *If B is a l.s.c. seminorm on V such that for all $\phi \in \mathcal{P}_m(\Omega)$, $B(\phi) = 0$, then*

$$(4.46) \quad B(v) \lesssim |v|_{m+\sigma,p}, \quad \forall v \in W^{m+\sigma,p},$$

As we know, the trace of the function in $H^\sigma(\Omega)$ is well-defined if $\sigma > \frac{1}{2}$, hence we can take $G(v) = \left| \int_{\Gamma} v dx \right|$ in (4.41) and have a special case of (4.41) as follows:

Theorem 30 (Poincaré Inequality). *Assume that $\Gamma \subset \partial\Omega$ is such that $\text{meas}(\Gamma) > 0$ and $\frac{1}{2} < \sigma \leq 1$, then*

$$\|v\|_{\sigma} \lesssim \left| \int_{\Gamma} v dx \right| + |v|_{\sigma}, \quad \forall v \in H^\sigma(\Omega).$$

Consequently

$$\|v\|_{\sigma} \lesssim |v|_{\sigma}, \quad \forall v \in H^\sigma(\Omega).$$

Lemma 16.

$$\|u\|_{0,\partial\Omega} \lesssim \epsilon^{-1} \|u\| + \epsilon \|u\|_1,$$

for any $u \in H^1(\Omega)$ and $\epsilon \in (0, 1)$.

Proof. It can be proved that (see Grisvard) that there exists a function $\rho \in [C^1(\bar{\Omega})]^2$ such that

$$\rho(x) \cdot \mathbf{n}(x) \geq 1, \quad \forall x \in \partial\Omega,$$

where $\mathbf{n}(x)$ is the outer normal direction of $\partial\Omega$ at x , we have

$$\begin{aligned} \int_{\partial\Omega} u^2 dx &\leq \int_{\partial\Omega} u^2 \rho \cdot \mathbf{n} dx \\ &= \int_{\Omega} \text{div} \rho u^2 dx + \int_{\Omega} 2u\rho \cdot \nabla u dx \\ &\lesssim \int_{\Omega} u^2 dx + \int_{\Omega} |u| |\nabla u| dx \\ &\lesssim (1 + \epsilon^{-2}) \int_{\Omega} u^2 dx + \epsilon^2 \int_{\Omega} |\nabla u|^2 dx, \end{aligned}$$

as desired. \square

4.5.1 Error estimates by scaling argument

A powerful general technique for obtaining finite element error estimate is a scaling argument. The main idea is to map a given finite element by an affine mapping to a fixed element of unit size (called reference element) and then apply the Bramble-Hilbert Lemma on the reference element and then map it back. Details follow.

It is convenient to have a standard simplex $s^n \subset \mathbb{R}^n$ spanned by the vertices $0, e_1, \dots, e_n$ where $e_i = (0, \dots, 1, \dots, 0)$. Then any n -simplex $\tau \subset \mathbb{R}^n$ can be thought as an image of s^n through an affine map $B : s^n \rightarrow \tau$ with $B(e_i) = x_i$. See Figure 4.7 (a).

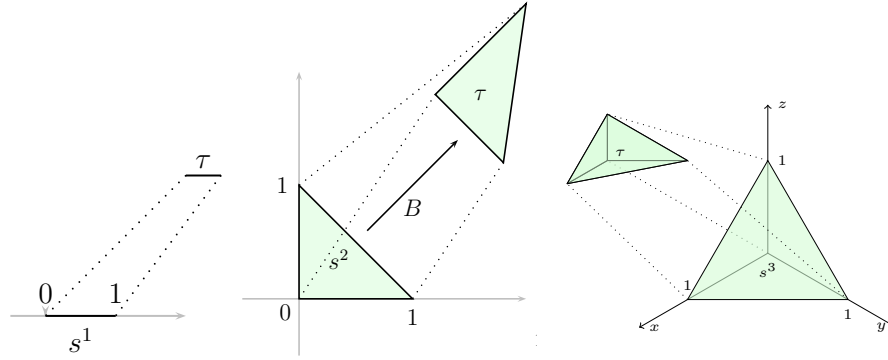


Fig. 4.7. Reference simplexes in $\mathbb{R}^1, \mathbb{R}^2$ and \mathbb{R}^3

For simplicity, we illustrate the idea in the case of two spatial dimension. With τ we denote a fixed (simplex) element from a given triangulation and with $\hat{\tau}$ we will by convention denote the reference element, as shown on fig. 4.7. For $\tau \in \mathcal{T}_h$ with vertices $\{a_1, a_2, a_3\}$ we assume that the angle between $(a_i - a_3)$ ($i = 1, 2$) is the largest angle of τ . For convenience we denote $\underline{b}_i = (a_i - a_3)$, $i = 1, 2$ to be the directonal vectors along sides of τ . The lengths of these vectors are denoted with $h_i = |\underline{b}_i|$ and the unit vectors along these sides are $e_i = (a_i - a_3)/h_i$, $i = 1, 2$.

With each fixed element $\tau \in \mathcal{T}_h$ we associate a diffeomorphic affine mapping $F : \hat{\tau} \rightarrow \tau$:

$$F(\hat{x}) = b + B\hat{x}$$

where $B = (\underline{b}_1, \underline{b}_2)$.

We define

$$\hat{v}(\hat{x}) = v(F(\hat{x})).$$

By chain rule, we have the following basic identities

$$(4.47) \quad \partial_{\hat{x}_i} \hat{v} = \nabla v \cdot \underline{b}_i = h_i \partial_{e_i} v$$

and

$$(4.48) \quad \partial_{\hat{x}_i} \partial_{\hat{x}_j} \hat{v} = h_i h_j \partial_{e_i} \partial_{e_j} v, \quad \partial_{\hat{x}_i} \partial_{\hat{x}_j} \partial_{\hat{x}_k} \hat{v} = h_i h_j h_k \partial_{e_i} \partial_{e_j} \partial_{e_k} v.$$

Generally, we have the lemma as follows:

Lemma 17. Let τ and $\hat{\tau} \subset \mathbb{R}^d$ be affine equivalent, i.e., there exists a bijective affine mapping

$$F : \hat{\tau} \rightarrow \tau, \quad F\hat{x} = B\hat{x} + b$$

with a nonsingular matrix B . If $v \in W^{m,p}(\tau)$, then $\hat{v} = v(F(\hat{x})) \in W^{m,p}(\hat{\tau})$, and there exists a constant $C = C(m, \hat{\tau})$ such that

$$|\hat{v}|_{m,p,\hat{\tau}} \leq C \|B\|^m |\det B|^{-1/2} |v|_{m,p,\tau}$$

$$|v|_{m,p,\tau} \leq C \|B^{-1}\|^m |\det B|^{1/2} |\hat{v}|_{m,p,\hat{\tau}}$$

Proof. Consider the derivative of order m as a multi-linear form. For $y_k = (y_{1,k}, y_{2,k}, \dots, y_{d,k})^T \in \mathbb{R}^d, k = 1, \dots, m$, define

$$(4.49) \quad D^m v(x)(y_1, \dots, y_m) = \sum_{1 \leq i_1, \dots, i_m \leq d} y_{i_1,1} \cdots y_{i_m,m} \partial_{i_1} \cdots \partial_{i_m} v(x).$$

from the chain rule, we have

$$(4.50) \quad \hat{D}^m \hat{v}(\hat{x})(\hat{y}_1, \dots, \hat{y}_m) = \sum_{1 \leq i_1, \dots, i_m \leq d} \hat{y}_{i_1,1} \cdots \hat{y}_{i_m,m} \hat{\partial}_{i_1} \cdots \hat{\partial}_{i_m} \hat{v}(\hat{x})$$

$$(4.51) \quad = \sum_{1 \leq i_1, \dots, i_m \leq d} \sum_{1 \leq j_1, \dots, j_m \leq d} \hat{y}_{i_1,1} \cdots \hat{y}_{i_m,m} b_{j_1 i_1} \cdots b_{j_m i_m} \partial_{j_1} \cdots \partial_{j_m} v(x)$$

$$(4.52) \quad = \sum_{1 \leq i_1, \dots, i_m \leq d} \sum_{1 \leq j_1, \dots, j_m \leq d} b_{j_1 i_1} \hat{y}_{i_1,1} \cdots \hat{y}_{i_m,m} b_{j_m i_m} \partial_{j_1} \cdots \partial_{j_m} v(x)$$

$$(4.53) \quad = D^m v(x)(B\hat{y}_1, \dots, B\hat{y}_m)$$

Thus

$$\|\hat{D}^m \hat{v}\|_{\mathcal{L}^m} \leq \|B\|^m \|D^m v\|_{\mathcal{L}^m}$$

where

$$\|D^m v\|_{\mathcal{L}^m} = \sup\{|D^m v(x)(y_1, \dots, y_m)| : |y_k| \leq 1, 1 \leq k \leq m\}.$$

We apply this estimate to the partial derivatives

$$\partial_{i_1} \cdots \partial_{i_m} v = D^m v(e_{i_1}, \dots, e_{i_m})$$

to get

$$(4.54) \quad \sum_{|\alpha|=m} |\hat{\partial}^\alpha \hat{v}|^p \leq d^m \max_{|\alpha|=m} |\hat{\partial}^\alpha \hat{v}|^p \leq d^m \|\hat{D}^m \hat{v}\|_{\mathcal{L}^m}^p \leq d^m \|B\|^{2m} \|D^m v\|_{\mathcal{L}^m}^p$$

$$(4.55) \quad \leq d^{2m} \|B\|^{2m} \sum_{|\alpha|=m} |\partial^\alpha v|^p$$

Finally we integrate, taking account of the transformation formula for multiple integrals

$$\int_{\hat{\tau}} \sum_{|\alpha|=m} |\hat{\partial}^\alpha \hat{v}|^p d\hat{x} \leq d^{2m} \|B\|^{2m} \int_{\tau} \sum_{|\alpha|=m} |\partial^\alpha v|^p |\det B^{-1}| dx.$$

This completes the proof of the first inequality. The other inequality is proved in a similar fashion. \square

Lemma 18. *Let τ and $\hat{\tau}$ be affine equivalent with*

$$F : \hat{x} \rightarrow B\hat{x} + b$$

being an invertible affine mapping. Then the upper bounds

$$(4.56) \quad \|B\| \leq \frac{h}{\hat{\rho}}, \quad \|B^{-1}\| \leq \frac{\hat{h}}{\rho}, \quad \left(\frac{\hat{h}}{\rho}\right)^d \leq |\det B| \leq \left(\frac{h}{\hat{\rho}}\right)^d$$

hold, where $h = \text{diam}(\tau)$, $\hat{h} = \text{diam}(\hat{\tau})$, ρ and $\hat{\rho}$ are the maximum diameter of the ball contained in τ and $\hat{\tau}$, respectively.

Proof. We may write

$$\|B\| = \frac{1}{\hat{\rho}} \sup_{|\xi|=\hat{\rho}} |B\xi|.$$

Given $\xi \in R^d$ so that $|\xi| = \hat{\rho}$, there exist $\hat{y}, \hat{z} \in \hat{\tau}$ such that $\hat{y} - \hat{z} = \xi$, $B\xi = F(\hat{y}) - F(\hat{z})$ with $F(\hat{y}), F(\hat{z}) \in \tau$. We deduce $|B\xi| \leq h$. This proves the first inequality in (4.56). The second inequality can be proved similarly. The last two inequalities are consequences of the identity $|\det B| = \frac{|\tau|}{|\hat{\tau}|}$. \square

Lemma 19 (Inverse inequality). *For any $\tau \in \mathcal{T}_h$, $u \in \mathcal{P}_m(\tau)$*

$$|u|_{1,\tau} \lesssim h_\tau^{-1} \|u\|_{0,\tau},$$

Proof. Map τ to a fixed unit size reference element $\hat{\tau}$ and then utilize the fact that any two norms on a fixed finite dimensional space are equivalent. Then by combining Lemma 17 and (4.56), we get the inequality.

$$|v|_{1,\tau} \lesssim h_\tau^{-1+\frac{d}{2}} |\hat{v}|_{1,\hat{\tau}} \lesssim h_\tau^{-1+\frac{d}{2}} \|\hat{v}\|_{0,\hat{\tau}} \lesssim h_\tau^{-1+\frac{d}{2}-\frac{d}{2}} \|v\|_{0,\tau} \lesssim h_\tau^{-1} \|v\|_{0,\tau}.$$

\square

We first present of L^p estimate which are independent of any regularity assumptions on the grids.

Lemma 20. *There exists an “absolute” constant C_1 such that the following estimate is true for any $\tau \in \mathcal{T}_h$:*

$$(4.57) \quad \|u - \mathcal{I}_h u\|_{L^p(\tau)} \leq C_1 \sum_{i,j=1}^n h_i h_j \|\partial_{e_i} \partial_{e_j} u\|_{L^p(\tau)} \quad \forall v \in W^{2,p}(\tau), 1 \leq p \leq \infty.$$

Proof. We prove (4.57) in the usual way. First we change the variables and after that we apply the Bramble-Hilbert lemma. We obtain:

$$(4.58) \quad \|u - \mathcal{I}_h u\|_{0,\tau}^2 = 2|\tau| \|\hat{u} - \hat{\mathcal{I}}_h \hat{u}\|_{0,\hat{\tau}}^2 \leq 2C_0 |\tau| \|\hat{u}\|_{2,\hat{\tau}}^2,$$

where $C_0 > 0$ is the best possible constant such that for every $\hat{v} \in H^2(\hat{\tau})$ with $\hat{v}(a_i) = 0$, $i = 1, 2, 3$ the following inequality holds:

$$|\hat{v}|_{0,\hat{\tau}} \leq C_0 |\hat{v}|_{2,\hat{\tau}}.$$

By (4.48)

$$(4.59) \quad 2|\tau| \|\partial_{\hat{x}_i} \partial_{\hat{x}_j} \hat{u}\|_{0,\hat{\tau}}^2 = h_i^2 h_j^2 \|\partial_{e_i} \partial_{e_j} u\|_{0,\tau}^2.$$

This completes the proof of (4.57). \square

Next lemma deals with local H^1 estimate. We assume that the triangulation \mathcal{T}_h satisfies the maximum angle condition, namely there exists a constant $\theta_0 > 0$, such that $\theta < \pi - \theta_0$ for every interior angle θ of any element in \mathcal{T}_h .

Lemma 21. Assume that $\theta \leq \pi - \theta_0$. Then the following inequality holds

$$\|u - \mathcal{I}_h u\|_{1,\tau} \leq C_2(\theta_0) \left(\sum_{i=1}^2 h_i \|\partial_{e_i}^2 u\|_{0,\tau} + (h_1 + h_2) \|\partial_{e_1} \partial_{e_2} u\|_{0,\tau} \right).$$

Proof. To prove the statement of the lemma we will use the similar technique, and the following obvious identity:

$$(4.60) \quad \partial_{\hat{x}_i}(\hat{\mathcal{I}}_h \hat{u}) = \hat{u}(\hat{\alpha}_i) - \hat{u}(\hat{\alpha}_3) = \int_{\hat{\alpha}_3}^{\hat{\alpha}_i} \partial_{\hat{x}_i} \hat{u}.$$

Assuming that for the maximal angle θ we have $\frac{\pi}{3} \leq \theta \leq \pi - \theta_0$ we have the following inequality

$$(4.61) \quad \|\nabla(u - \mathcal{I}_h u)\|_{0,\tau} \lesssim \|\partial_{e_1}(u - \mathcal{I}_h u)\|_{0,\tau} + \|\partial_{e_2}(u - \mathcal{I}_h u)\|_{0,\tau}.$$

After change of variables we have, by (4.48)

$$(4.62) \quad \|\partial_{e_1}(u - \mathcal{I}_h u)\|_{0,\tau}^2 = 2h_1^{-2} |\tau| \|\partial_{\hat{x}_1}(\hat{u} - \hat{\mathcal{I}}_h \hat{u})\|_{0,\hat{\tau}}^2.$$

Using the identity (4.67) and Bramble-Hilbert lemma we get the following estimate:

$$(4.63) \quad \|\partial_{\hat{x}_1}(\hat{u} - \hat{\mathcal{I}}_h \hat{u})\|_{0,\hat{\tau}} = \|\partial_{\hat{x}_1} \hat{u} - \int_{\hat{\alpha}_3}^{\hat{\alpha}_1} \partial_{\hat{x}_1} \hat{u}\|_{0,\hat{\tau}} \leq C |\partial_{\hat{x}_1} \hat{u}|_{1,\hat{\tau}}.$$

Now, going back to (4.62) we find that

$$(4.64) \quad \|\partial_{e_1}(u - \mathcal{I}_h u)\|_{0,\tau}^2 \leq Ch_1^{-2} |\tau| \left(\|\partial_{\hat{x}_1}^2 \hat{u}\|_{0,\hat{\tau}}^2 + \|\partial_{\hat{x}_2} \partial_{\hat{x}_1} \hat{u}\|_{0,\hat{\tau}}^2 \right).$$

Finally by (4.59) we get

$$(4.65) \quad \|\partial_{\hat{x}_1}(\hat{u} - \hat{\mathcal{I}}_h \hat{u})\|_{0,\hat{\tau}}^2 \leq C(h_1^2 \|\partial_{e_1}^2 u\|_{0,\tau}^2 + h_2^2 \|\partial_{e_2} \partial_{e_1} u\|_{0,\tau}^2)$$

Applying the same technique for the second component $\|\partial_{e_2}(u - \mathcal{I}_h u)\|_{0,\tau}$ we obtain the desired result. \square

Quadratic elements

We now discuss the extension of the result to higher order elements.

Lemma 22. Assume that $\theta \leq \pi - \theta_0$. Then the following inequality holds

$$(4.66) \quad \|u - \mathcal{I}_h u\|_{1,\tau} \leq C_2(\theta_0) \left(\sum_{i=1}^2 h_i^2 \|\partial_{e_i}^3 u\|_{0,\tau} + (h_1^2 + h_2^2) (\|\partial_{e_1}^2 \partial_{e_2} u\|_{0,\tau} + \|\partial_{e_1} \partial_{e_2}^2 u\|_{0,\tau}) \right).$$

Proof. For quadratic elements, we use the following identity (which can be verified by direct calculation):

$$(4.67) \quad \partial_{\hat{x}_1}(\hat{\mathcal{I}}_h \hat{u}) = (3 - 4\hat{x}_1) \int_0^{1/2} \partial_{\hat{x}_1} \hat{u} \, d\hat{x}_1 + (4\hat{x}_1 - 1) \int_{1/2}^1 \partial_{\hat{x}_1} \hat{u} \, d\hat{x}_1.$$

which, by Bramble-Hilbert lemma, implies

$$(4.68) \quad \|\partial_{\hat{x}_1}(\hat{u} - \hat{\mathcal{I}}_h \hat{u})\|_{0,\hat{\tau}} \lesssim |\partial_{\hat{x}_1} \hat{u}|_{2,\hat{\tau}}.$$

The rest of the proof is similar. \square

Theorem 31. Assume $d \leq 3$ and $\{\mathcal{T}_h : h \in \mathfrak{N}\}$ is quasiuniform, then, for any $2p > d$,

$$\|(I - I_h)v\|_{0,p} + h|(I - I_h)v|_{1,p} \lesssim h^2|v|_{2,p}, \quad \forall v \in W^{2,p}(\Omega) \cap H_0^1(\Omega).$$

Proof. A proof based on Bramble-Hilber Lemma is to be enclosed. \square

The assumption that $2p \geq d$ in above theorem is to guarantee $H^2(\Omega) \hookrightarrow C(\bar{\Omega})$ which is necessary to control the interpolation operator. The following result however relaxes this restriction in some special circumstances.

Remark 8. A similar procedure proves that

$$\inf_{q \in Q_m} \|v + q\|_{m+1,p,\Omega} \approx \sum_{i=1}^d \|\partial_{x_i}^{m+1} v\|_{0,\Omega}.$$

Thus, for Q_1 element,

$$\|(I - I_h)v\|_{0,p} + h|(I - I_h)v|_{1,p} \lesssim h^2 \sum_{i=1}^d \|\partial_{x_i}^{m+1} v\|_{0,\Omega}, \quad \forall v \in W^{2,p}(\Omega) \cap H_0^1(\Omega).$$

Lemma 23. Assume $\{\mathcal{T}_h : h \in \mathfrak{N}\}$ is regular and $l \geq 1$ is given. Then we have

$$\|(I - I_h)v\|_{0,p,\tau} \lesssim h_\tau \|\nabla v\|_{0,p,\tau}, \quad \forall v \in \mathcal{P}_l(\tau), \quad 1 \leq p \leq \infty, \quad \tau \in \mathcal{T}_h.$$

The estimate provided in above lemma does not in general hold for $v \in W^{1,p}(\tau)$ when the imbedding $W^{1,p}(\Omega) \hookrightarrow C(\bar{\Omega})$ does not hold (e.g. $2p \leq d$).

Proof. Let $\hat{\tau}$ be the standard reference element, for any $\tau \in \mathcal{T}_h$, we have an affine diffeomorphism:

$$F_\tau : \hat{\tau} \mapsto \tau.$$

For any function $v \in L^2(\tau)$, we adopt the following standard notation:

$$\hat{v}(\hat{x}) = v(F(\hat{x})), \quad \hat{x} \in \hat{\tau}.$$

We can show that (cf.[?])

$$(4.69) \quad \|(I - I_h)v\|_{0,p,\tau} \lesssim h_\tau^{\frac{d}{p}} \|(\hat{I} - \hat{I}_h)\hat{v}\|_{0,p,\hat{\tau}}.$$

Since $\mathcal{P}_l(\hat{\tau})$ is a fixed finite dimensional space, we have

$$\|(\hat{I} - \hat{I}_h)\hat{v}\|_{0,p,\hat{\tau}} \lesssim \|\hat{v}\|_{1,p,\hat{\tau}}.$$

Replacing \hat{v} by $\hat{v} + \hat{q}$ for $\hat{q} \in \mathcal{P}_0(\hat{\tau})$, we get from (4.42) that

$$\begin{aligned} \|(\hat{I} - \hat{I}_h)\hat{v}\|_{0,p,\hat{\tau}} &\lesssim \inf_{\hat{q} \in \mathcal{P}_0(\hat{\tau})} \|\hat{v} + \hat{q}\|_{1,p,\hat{\tau}} \\ &\lesssim \|\nabla \hat{v}\|_{0,p,\hat{\tau}} \lesssim h_\tau^{1-\frac{d}{p}} \|\nabla v\|_{0,p,\tau}. \end{aligned}$$

The desired result then follows by combining the above estimate with (4.69). \square

Theorem 32. For $0 \leq \beta \leq 1$, $\frac{d}{2} - 1 < \delta \leq 1$, we have

$$\|(I - I_h)v\|_{1-\beta} \lesssim h^{\beta+\delta} |v|_{1+\delta}, \quad \forall v \in H^{1+\delta}(\Omega) \cap H_0^1(\Omega).$$

Proof. We shall give the proof for $d \geq 2$ and leave the proof for $d = 1$ to the interested readers.

By the interpolation property of Sobolev space, it suffices to show that

$$(4.70) \quad \|(I - I_h)v\| \lesssim h^{1+\delta} |v|_{1+\delta}, \quad \forall v \in H^{1+\delta}(\Omega) \cap H_0^1(\Omega),$$

and

$$(4.71) \quad \|(I - I_h)v\|_1 \lesssim h^\delta |v|_{1+\delta}, \quad \forall v \in H^{1+\delta}(\Omega) \cap H_0^1(\Omega).$$

Proofs of (4.70) and (4.71) are similar. Hence we will only show (4.70). For this purpose, it suffices to show that

$$(4.72) \quad \|(I - I_h)v\|_{0,2,\tau} \lesssim h^{1+\delta} |\nabla v|_{\delta,\tau},$$

since by the integral representation of the fractional norm, we have

$$\sum_{\tau \in \mathcal{T}_h} |\nabla v|_{\delta,\tau}^2 \leq |\nabla v|_\delta^2.$$

The idea to show (4.72) is again to use the Bramble–Hilbert scaling technique. As before let $\hat{\tau}$ be the reference element, by changing variable, we have

$$\|(I - I_h)v\|_{0,2,\tau} \lesssim h^{\frac{d}{2}} \|(\hat{I} - \hat{I}_h)\hat{v}\|_{0,\hat{\tau}}.$$

Since $H^{1+\delta}(\hat{\tau}) \hookrightarrow C(\hat{\tau})$ and \hat{I}_h is invariant on linears, we get

$$\|(\hat{I} - \hat{I}_h)\hat{v}\|_{0,\hat{\tau}} \lesssim \inf_{\hat{q} \in \mathcal{P}_1(\hat{\tau})} \|\hat{v} + \hat{q}\|_{1+\delta,\hat{\tau}}.$$

In virtue of (4.42) in Theorem (28), we have

$$\inf_{\hat{q} \in \mathcal{P}_1(\hat{\tau})} \|\hat{v} + \hat{q}\|_{1+\delta,\hat{\tau}} \lesssim |\nabla \hat{v}|_{\delta,\hat{\tau}}.$$

Mapping $\hat{\tau}$ back to τ , it is elementary to check that

$$|\nabla \hat{v}|_{\delta,\hat{\tau}} \lesssim h^{1+\delta-\frac{d}{2}} |\nabla v|_{\delta,\tau}.$$

The desired result then follows. \square

4.5.2 Error estimates for finite element solutions

Basic energy norm estimate

Combining the interpolation estimate given in Theorem 25 and Theorem 24 we have the following basic *a priori* estimate in finite element method:

Theorem 33. *Let Ω be a bounded Lipschitz domain in \mathbb{R}^d ($d = 1, 2, 3$), V be a Hilbert space, $a(\cdot, \cdot)$ be a continuous, symmetric, bilinear form defining an inner product on V , and $f(\cdot)$ be a continuous linear form. If $u \in H^2(\Omega)$ is solution to the problem (4.19) and u_h is a solution to (4.22), then the following equality holds:*

$$|u - u_h|_a \leq Ch|u|_{2,\Omega}.$$

This shows that if we make the triangulation finer and finer (taking $h \rightarrow 0$), the finite element solution u_h will approach the solution of the variational problem, and with a rate proportional to h , provided that $u \in H^2(\Omega)$. Later we will show that the finite element solution approaches the solution to the variational problem even if u is not an H^2 function, although the rate of convergence can be slower, unless adaptive finite element techniques are used.

Duality argument and L^2 estimate

Now we turn to the estimate of error $\|u - u_h\|_{0,\Omega}$. It is known from Theorem 25 that for the interpolation operator, the error in $L^2(\Omega)$ norm is, with respect to h , one order higher than that in $H^1(\Omega)$ norm. Is it still true for the finite element solution u_h ? Next we introduce the technique of Aubin and Nistche, which is also called duality argument, to answer this question.

To analyze the error in $L^2(\Omega)$ norm, we introduce the dual problem of (4.19). For any $g \in L^2(\Omega)$, let $\phi_g \in H_0^1(\Omega)$ be the solution of the following variational problem,

$$(4.73) \quad a(\phi_g, v) = (g, v), \quad \forall v \in H_0^1(\Omega).$$

By (4.73), (4.19), (4.22) and

$$\|u - u_h\|_{0,\Omega} = \sup_{0 \neq g \in L^2(\Omega)} \frac{(g, u - u_h)}{\|g\|_{0,\Omega}},$$

it is derived that

$$\begin{aligned} (g, u - u_h) &= a(\phi_g, u - u_h) = \inf_{v_h \in V_h} a(\phi_g - v_h, u - u_h) \\ &\leq \inf_{v_h \in V_h} \|\phi_g - v_h\|_{1,\Omega} \|u - u_h\|_{1,\Omega}, \end{aligned}$$

and

$$(4.74) \quad \|u - u_h\|_{0,\Omega} \leq \|u - u_h\|_{1,\Omega} \sup_{0 \neq g \in L^2(\Omega)} \frac{1}{\|g\|_{0,\Omega}} \inf_{v_h \in V_h} \|\phi_g - v_h\|_{1,\Omega}.$$

Theorem 34. *Let Ω be a bounded Lipschitz domain in \mathbb{R}^d ($d = 1, 2, 3$), V be a Hilbert space, $a(\cdot, \cdot)$ be a continuous, symmetric, bilinear form defining an inner product on V , and $f(\cdot)$ be a continuous linear form. If $u \in H^2(\Omega)$ is solution to the problem (4.19) and u_h is a solution to (4.22), then the following equality holds:*

$$\|u - u_h\|_{0,\Omega} \leq Ch^2|u|_{2,\Omega}.$$

Since Ω is convex, we have

$$(4.75) \quad \|\phi_g\|_{2,\Omega} \lesssim \|g\|_{0,\Omega}.$$

By Theorem 25,

$$(4.76) \quad \inf_{v_h \in V_h} \|\phi_g - v_h\|_{1,\Omega} \leq \|\phi_g - \phi_{g,I}\|_{1,\Omega} \lesssim h|\phi_g|_{2,\Omega},$$

where $\phi_{g,I}$ is the interpolation of ϕ_g . By (4.74), (4.75), (4.76) and Theorem 33, it is derived that

$$\|u - u_h\|_{0,\Omega} \lesssim h\|u - u_h\|_{1,\Omega} \lesssim h^2|u|_{2,\Omega}.$$

Remark 9. For Poisson problem, we also have the estimate

$$\|u - u_h\|_{0,\infty,\Omega} \lesssim h^2 |\log h| |u|_{2,\infty,\Omega},$$

provided that $u \in W^{2,\infty}(\Omega)$.

4.6 Quasi-uniform and locally adaptive grids

In this section, we will give a heuristic argument of adaptive methods and report several numerical experiments to demonstrate the convergence behavior of this method.

4.6.1 Smooth function on a unit square and on an L-shaped domain

When the function is smooth, namely $u \in H^2(\Omega)$, the linear finite element method on quasi-uniform grids provides best asymptotic approximations as shown in Theorem 33. This optimal convergence rate is confirmed by our numerical experiments on a square domain and an L-shaped domain. We would like to point out, for domains like the L-shaped domain, the solution is rarely as smooth as the artificially produced solution.

4.6.2 Nonsmooth solution of an L-shaped domain

As mentioned before, the solution of the Poisson equation can become less regular on a concave domain even the data f on the right hand side is smooth. When this happens, as demonstrated in Table 4.6.4 for an L-shape domain, the convergence rate of the finite element approximation is degraded. From this computation, we observe that:

$$\|u - u_h\|_1 \leq Ch^{2/3} \approx N^{-1/3}.$$

This estimate can be theoretically justified very easily.

Because of the more singular behavior of the solution at the corner, the mesh needs to be appropriately refined around the corner in order to achieve more desirable accuracy. Adaptive method means that you should put more triangles near the singular point to obtain a good approximation. Thus, the resulting triangulation is not quasi-uniform any more. But the question is what the optimal grid is and how to find such a grid? Here we will first give a numerical example to illustrate the importance of the adaptive finite element method. From Table 4.6.4, for properly locally refined grid, we observe the following convergent rate:

$$\|u - u_N\|_1 \leq CN^{-1/2}.$$

Now, we give a heuristic argument to demonstrate why this is possible. According to the Bramble-Hilbert lemma and the scaling technique, see [?] for more details, we have the following estimate.

Lemma 24. *In two dimensions ($d = 2$), we have the following estimate:*

$$(4.77) \quad |u - u_I|_{1,\tau} \lesssim |\nabla^2 u|_{0,1,\tau} \equiv \int_{\tau} |\nabla^2 u(x)|$$

Let \mathcal{T} be a triangulation of the domain $\Omega \subset \mathbb{R}^2$ with N triangles, and let u_I or $u_{\mathcal{T}}$ be the nodal interpolation or the finite element approximation of u . Using the embedding: $W^{2,1}(\Omega) \subset H^1(\Omega) \cap C(\bar{\Omega})$, we know the nodal interpolation u_I is well defined and

$$|u - u_I|_{1,\tau} \leq C \|u\|_{2,1,\tau}, \quad \forall \tau \in \mathcal{T}.$$

Since the nodal interpolation preserves linear polynomials, we get

$$|u - u_I|_{1,\tau} \leq C |u|_{2,1,\tau}, \quad \forall \tau \in \mathcal{T},$$

and thus

$$|u - u_I|_{1,\Omega}^2 \lesssim \sum_{\tau} |u|_{2,1,\tau}^2.$$

We now try to minimize $\sum_{\tau} |u|_{2,1,\tau}^2$ by changing the underlying grids. By Cauchy-Schwarz inequality,

$$|u|_{2,1,\Omega} = \sum_{\tau} |u|_{2,1,\tau} \leq \left(\sum_{\tau} 1 \right)^{1/2} \left(\sum_{\tau} |u|_{2,1,\tau}^2 \right)^{1/2} = N^{1/2} \left(\sum_{\tau} |u|_{2,1,\tau}^2 \right)^{1/2}.$$

Thus, we have a lower bound: $\left(\sum_{\tau} |u|_{2,1,\tau}^2 \right)^{1/2} \geq N^{-1} |u|_{2,1,\Omega}$. The equality holds if and only if the following equi-distribution principle holds:

$$|u|_{2,1,\tau} = \text{constant} = N^{-1} |u|_{2,1,\Omega}.$$

Based on the above arguments, we obtain that if \mathcal{T} is a triangulation of N shape-regular elements satisfying:

$$(4.78) \quad |u|_{2,1,\tau} \leq \kappa_{\tau,N} |u|_{2,1,\Omega} \quad \text{and} \quad \sum_{\tau \in \mathcal{T}} \kappa_{\tau,N}^2 \leq c_0 N^{-1},$$

then

$$(4.79) \quad |u - u_I|_{1,\Omega} \lesssim N^{-1/2} |u|_{2,1,\Omega}.$$

In view of (4.78), equidistribution is indeed a sufficient condition for optimal convergent rate, but by no means this has to be a necessary condition. Can we really find such a grid that $|u|_{2,1,\tau}$ is a constant? This equidistribution principle can be violated but asymptotically optimal error estimate can still be maintained. For example, $\kappa_{\tau,N}^2$ could be big for some triangles provided the sum is bounded by $c_0 N^{-1}$; see (4.78).

As it turns out, rigorously speaking, we need a slightly stronger assumption on u (namely smoother than $W^{2,1}(\Omega)$), for example, $u \in W^{2,p}(\Omega)$ ¹ for some $p > 1$. This assumption is true for most practical domains. More precisely, for any $p > 1$, any N , we have a constructive algorithm [2] to find adaptively a shape-regular triangulation \mathcal{T} with $O(N)$ elements such that

$$|u|_{2,1,\tau} \leq c_0 N^{-1} |u|_{2,p,\Omega}.$$

As a result, since $|u - u_{\mathcal{T}}|_{1,\Omega} \lesssim |u - u_I|_{1,\Omega}$, we have the following error estimate

$$(4.80) \quad |u - u_{\mathcal{T}}|_{1,\Omega} \lesssim N^{-1/2} |u|_{2,p,\Omega}.$$

which is asymptotically best possible for an isotropic triangulation with $O(N)$ elements.

¹ it actually suffices if $M(\nabla^2 u) \in L^1(\Omega)$, where $M(f)$ is the Hardy-Littlewood maximal function of f

4.6.3 A posteriori error estimate

In practice, with the solution u unknown, finding the optimal grid becomes even more difficult. Adaptive methods use computational quantities to give an estimate of the error on each element, and tend to refine the mesh around those singularity points of the solution.

Recent works have shown that the estimate (4.80) can be practically realized [4, 32, 11, 39] by using appropriate a posteriori error estimates.

Theorem 35. *Let u_h be the finite element solution. It holds that*

$$\|\nabla(u - u_h)\|_0 \lesssim h\|f - \Delta u_h\|_0 + \left(\sum_e h \| [\nabla u_h] \cdot n \|_e^2 \right)^{1/2}.$$

Proof. For any $v \in V$,

$$\begin{aligned} (\nabla(u - u_h), \nabla v) &= (\nabla(u - u_h), \nabla(v - \Pi_h v)) = \sum_{\tau} (\nabla(u - u_h), \nabla(v - \Pi_h v))_{0,\tau} \\ &= \sum_{\tau} (f + \Delta u_h, v - \Pi_h v)_{0,\tau} + \sum_{\tau} \int_{\partial\tau} (v - \Pi_h v) \nabla(u - u_h) \cdot n \\ (4.81) \quad &= \sum_{\tau} (h(f + \Delta u_h), h^{-1}(v - \Pi_h v))_{0,\tau} + \sum_e \int_e (v - \Pi_h v) [\nabla(u - u_h)] \cdot n \\ &= \sum_{\tau} (h(f + \Delta u_h), h^{-1}(v - \Pi_h v))_{0,\tau} - \sum_e \int_e (v - \Pi_h v) [\nabla u_h] \cdot n \end{aligned}$$

Thanks to Cauchy-Schwarz inequality,

$$\begin{aligned} \left| \sum_{\tau} (h(f + \Delta u_h), h^{-1}(v - \Pi_h v))_{0,\tau} \right| &\leq \sum_{\tau} h \|f + \Delta u_h\|_{0,\tau} h^{-1} \|v - \Pi_h v\|_{0,\tau} \\ (4.82) \quad &\lesssim \left(\sum_{\tau} h^2 \|f + \Delta u_h\|_{0,\tau}^2 \right)^{1/2} \left(\sum_{\tau} \|\nabla v\|_{0,\omega_{\tau}}^2 \right)^{1/2} \\ &\lesssim h \|f + \Delta u_h\|_0 \|\nabla v\|_0. \end{aligned}$$

By the trace inequality in Theorem 26,

$$\begin{aligned} \left| \int_e (v - \Pi_h v) [\nabla u_h] \cdot n \right| &\leq \|h^{1/2} [\nabla u_h] \cdot n\|_{0,e} \|h^{-1/2} (v - \Pi_h v)\|_{0,e} \\ (4.83) \quad &\lesssim \|h^{1/2} [\nabla u_h] \cdot n\|_{0,e} (h^{-1} \|v - \Pi_h v\|_{0,\tau} + |v - \Pi_h v|_{1,\tau}) \\ &\lesssim \|h^{1/2} [\nabla u_h] \cdot n\|_{0,e} \|\nabla v\|_{0,\omega_{\tau}} \end{aligned}$$

Let $v = u - u_h$. Substituting (4.82) and (4.83) to (4.81) leads to

$$\|\nabla(u - u_h)\|_0 \lesssim h\|f + \Delta u_h\|_0 + \left(\sum_e h \| [\nabla u_h] \cdot n \|_{0,e}^2 \right)^{1/2}.$$

□

N	$ u - u_h _1$
8	3.616416e-01
21	2.290197e-01
65	1.451221e-01
225	9.262711e-02
833	5.898908e-02
3201	3.744845e-02
12545	2.371329e-02
49665	1.498890e-02

Table 4.1. Uniformly refined triangulation

N	$ u - u_h _1$
8	3.616416e-01
20	1.639911e-01
51	1.015016e-01
70	8.355636e-02
92	7.762996e-02
215	4.655665e-02
711	2.424167e-02
1273	1.768547e-02

Table 4.2. Locally refined triangulation

4.6.4 Some numerical examples

Fig 4.9 shows that adaptive grids use much less grid points to achieve the same error.

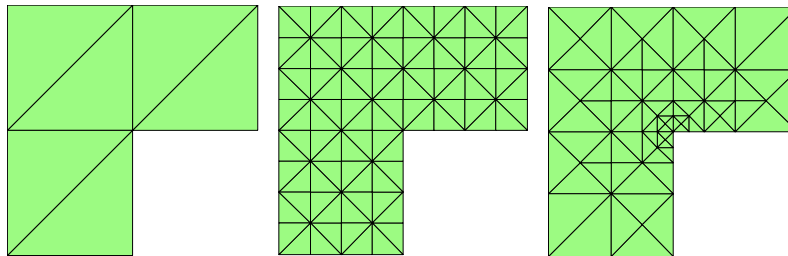


Fig. 4.8. Examples of uniformly refined mesh and adaptive refined mesh

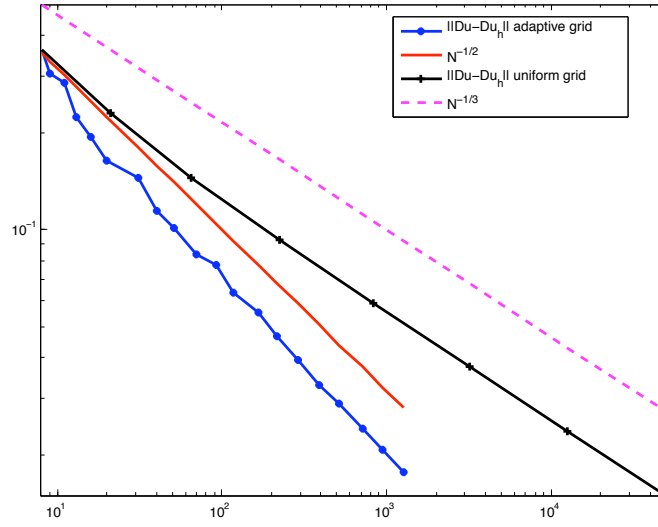


Fig. 4.9. Convergent rate of error

4.7 Numerical quadrature

To implement a finite element method, we need to compute various integrals on each element. In this section, we will present several numerical quadrature rules for elements in 1, 2 and 3 dimensions. We will also give the corresponding error estimates for these quadrature rules. The proof of these estimates will be left as exercises in future chapters when proper techniques become available.

The following tables are borrowed from [46].

4.7.1 $d = 1$

By changing of variable,

$$\int_a^b f(x) dx = \frac{b-a}{2} \int_{-1}^1 f\left(\frac{b-a}{2}x + \frac{b+a}{2}\right) dx.$$

We shall only consider the approximation of the integral in $[-1, 1]$. We choose interval $[-1, 1]$ (not $[0, 1]$) for the symmetric of points and weights. We shall use the following numerical quadrature

$$\int_{-1}^1 f(x) dx \approx \sum_{i=1}^n (f(-a_i) + f(a_i)) \omega_i.$$

The Gauss points and weights are listed in Figure 4.10.

$\pm a$		H
0	$n = 1$	2.000 000 000 000 000
$1/\sqrt{3}$	$n = 2$	1.000 000 000 000 000
$\sqrt{0.6}$	$n = 3$	5/9
0.000 000 000 000 000		8/9
0.861 136 311 594 953	$n = 4$	0.347 854 845 137 454
0.339 981 043 584 856		0.652 145 154 862 546
0.906 179 845 938 664	$n = 5$	0.236 926 885 056 189
0.538 469 310 105 683		0.478 628 670 499 366
0.000 000 000 000 000		0.568 888 888 888 889
0.932 469 514 203 152	$n = 6$	0.171 324 492 379 170
0.661 209 386 466 265		0.360 761 573 048 139
0.238 619 186 083 197		0.467 913 934 572 691
0.949 107 912 342 759	$n = 7$	0.129 484 966 168 870
0.741 531 185 599 394		0.279 705 391 489 277
0.405 845 151 377 397		0.381 830 050 505 119
0.000 000 000 000 000		0.417 959 183 673 469
0.960 289 856 497 536	$n = 8$	0.101 228 536 290 376
0.796 666 477 413 627		0.222 381 034 453 374
0.525 532 409 916 329		0.313 706 645 877 887
0.183 434 642 495 650		0.362 683 783 378 362
0.968 160 239 507 626	$n = 9$	0.081 274 388 361 574
0.836 031 107 326 636		0.180 648 160 694 857
0.613 371 432 700 590		0.260 610 696 402 935
0.324 253 423 403 809		0.312 347 077 040 003
0.000 000 000 000 000		0.330 239 355 001 260
0.973 906 528 517 172	$n = 10$	0.066 671 344 308 688
0.865 063 366 688 985		0.149 451 349 150 581
0.679 409 568 299 024		0.219 086 362 515 982
0.433 395 394 129 247		0.269 266 719 309 996
0.148 874 338 981 631		0.295 524 224 714 753

Fig. 4.10. Quadrature in 1-D

4.7.2 $d = 2$

In 2d, we shall use barycentric coordinate to present the formula of numerical quadrature. Let $(x_k, y_k), k = 1, 2, 3$ be coordinates of three vertices of τ , a point $p \in \tau$ is uniquely determined by its coordinate $p : (\lambda_1, \lambda_2, \lambda_3)$ by

$$p = (\lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3, \lambda_1 y_1 + \lambda_2 y_2 + \lambda_3 y_3).$$

We shall approximate

$$\int_{\tau} f(x) dx \approx \sum_{i=1}^n f(p_i) \omega_i |\tau|.$$

There list three ways to approximate $\int_{\tau} f(x) dx$:

1. $\frac{|\tau|}{3} \sum_{i=1}^3 f(a_i)$;
2. $\frac{|\tau|}{3} \sum_{1 \leq i \leq j \leq 3} f(a_{ij})$;
3. $|\tau| f(a_{123})$.

Consider the accuracy of the first case. Since

$$\int_{\tau} f(x) dx - \frac{|\tau|}{3} \sum_{i=1}^3 f(a_i) = 2|\tau| \int_{\hat{\tau}} \hat{f}(\hat{x}) d\hat{x} - \frac{|\tau|}{3} \sum_{i=1}^3 f(\hat{a}_i).$$

Denote

$$B(\hat{f}) = 2 \int_{\hat{\tau}} \hat{f}(\hat{x}) d\hat{x} - \frac{1}{3} \sum_{i=1}^3 f(\hat{a}_i).$$

If $\hat{f} \in P_1(\hat{\tau})$, $B(\hat{f}) = 0$, Bramble-Hilbert lemma gives

$$B(\hat{f}) \lesssim |\hat{f}|_{2,\infty,\hat{\tau}} \lesssim h^2 |f|_{2,\infty,\tau}.$$

Thus,

$$\int_{\tau} f(x) dx - \frac{|\tau|}{3} \sum_{i=1}^3 f(a_i) = O(h^2) |\mathcal{Q}| |f|_{2,\infty,\mathcal{Q}}.$$

Consider the accuracy of the second case. Denote

$$B(\hat{f}) = 2 \int_{\hat{\tau}} \hat{f}(\hat{x}) d\hat{x} - \frac{1}{3} \sum_{1 \leq i \leq j \leq 3} f(\hat{a}_{ij}).$$

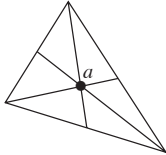
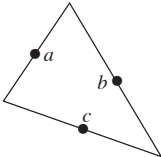
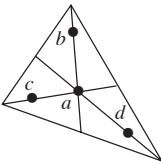
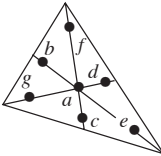
A simple computation gives that $B(\hat{f}) = 0, \forall \hat{f} \in P_2(\hat{\tau})$. By Bramble-Hilbert lemma,

$$B(\hat{f}) \lesssim |\hat{f}|_{3,\infty,\hat{\tau}} \lesssim h^3 |f|_{3,\infty,\tau}.$$

Thus,

$$\int_{\tau} f(x) dx - \frac{|\tau|}{3} \sum_{1 \leq i \leq j \leq 3} f(a_{ij}) = O(h^3) |\mathcal{Q}| |f|_{3,\infty,\mathcal{Q}}.$$

The barycentric coordinate of p_i and corresponding weight is listed in Figure 4.11.

Order	Figure	Error	Points	Triangular coordinates	Weights
Linear		$R = O(h^2)$	a	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	1
Quadratic		$R = O(h^3)$	a b c	$\frac{1}{2}, \frac{1}{2}, 0$ $0, \frac{1}{2}, \frac{1}{2}$ $\frac{1}{2}, 0, \frac{1}{2}$	$\frac{1}{3}$ $\frac{1}{3}$ $\frac{1}{3}$
Cubic		$R = O(h^4)$	a b c d	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$ $0.6, 0.2, 0.2$ $0.2, 0.6, 0.2$ $0.2, 0.2, 0.6$	$-\frac{27}{48}$ $\frac{25}{48}$
Quintic		$R = O(h^6)$	a b c d e f g	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$ $\alpha_1, \beta_1, \beta_1$ $\beta_1, \alpha_1, \beta_1$ $\beta_1, \beta_1, \alpha_1$ $\alpha_2, \beta_2, \beta_2$ $\beta_2, \alpha_2, \beta_2$ $\beta_2, \beta_2, \alpha_2$	0.225 000 000 0 0.132 394 152 7 0.125 939 180 5

with
 $\alpha_1 = 0.059\ 715\ 871\ 7$
 $\beta_1 = 0.470\ 142\ 064\ 1$
 $\alpha_2 = 0.797\ 426\ 985\ 3$
 $\beta_2 = 0.101\ 286\ 507\ 3$

Fig. 4.11. Quadrature in 2-D

4.7.3 $d = 3$

Similarly we shall approximate

$$\int_{\tau} f(x, y, z) dx dy dz \approx \sum_{i=1}^n f(p_i) \omega_i |\tau|.$$

The barycentric coordinate of p_i and corresponding weight is listed in Figure 4.12.

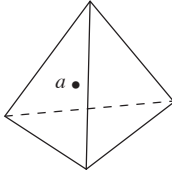
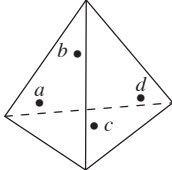
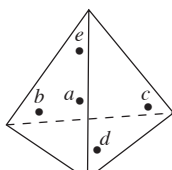
No.	Order	Figure	Error	Points	Tetrahedral coordinates	Weights
1	Linear		$R = O(h^2)$	a	$\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$	1
2	Quadratic		$R = O(h^3)$	a b c d	$\left. \begin{array}{l} \alpha, \beta, \beta, \beta \\ \beta, \alpha, \beta, \beta \\ \beta, \beta, \alpha, \beta \\ \beta, \beta, \beta, \alpha \end{array} \right\}$ $\alpha = 0.58541020$ $\beta = 0.13819660$	$\frac{1}{4}$
3	Cubic		$R = O(h^4)$	a b c d e	$\left. \begin{array}{l} \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4} \\ \frac{1}{2}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6} \\ \frac{1}{6}, \frac{1}{2}, \frac{1}{6}, \frac{1}{6} \\ \frac{1}{6}, \frac{1}{6}, \frac{1}{2}, \frac{1}{6} \\ \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{2} \end{array} \right\}$	$-\frac{4}{5}$ $\frac{9}{20}$

Fig. 4.12. Quadrature in 3-D

4.8 Stiffness matrix and properties

The solution of a finite difference discretization is reduced to the solution of a linear algebraic system such as (3.7). To solve this type of equations efficiently, it is important to understand the properties of the linear system, especially the properties related to the stiffness matrix A . Next, we will use the stiffness matrices derived in the previous sections as examples in order to study their properties in the following lemma.

Lemma 25. *The stiffness matrix A has the following properties:*

1. A is sparse, i.e., A has $O(N)$ nonzero entries.
2. A is symmetric, positive, and definite.
3. The condition number of A , defined by the ratio of the extreme eigenvalues of A ,

$$\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$$

satisfies

$$\kappa(A) = O(h^{-2}).$$

Furthermore,

$$\lambda_{\min}(A) = O(h^2), \lambda_{\max}(A) = O(1).$$

Proof. The first two conclusions in Lemma 25 are obvious. Let us discuss the eigenvalues of A . Let λ_k be an eigenvalue of A and $\xi_k \in \mathbb{R}^N \setminus \{0\}$ a corresponding eigenvector such that

$$A\xi^k = \lambda_k \xi^k.$$

Namely,

$$-\xi_{j-1}^k + 2\xi_j^k - \xi_{j+1}^k = \lambda_k \xi_j^k, \quad 1 \leq j \leq N,$$

with $\xi_0^k = \xi_{N+1}^k = 0$. It is not difficult to solve the above finite difference equation to obtain the following closed-form solution:

$$(4.84) \quad \lambda_k = 4 \sin^2 \frac{k\pi}{2(N+1)}, \quad \xi_j^k = \sin \frac{kj\pi}{N+1} \quad (1 \leq j \leq N).$$

Indeed, the relationship $A\xi^k = \lambda_k \xi^k$ can be verified by the following elementary trigonometric identities:

$$-\sin \frac{k(j-1)\pi}{N+1} + 2 \sin \frac{kj\pi}{N+1} - \sin \frac{k(j+1)\pi}{N+1} = 4 \sin^2 \frac{k\pi}{2(N+1)} \sin \frac{kj\pi}{N+1}.$$

We note that

$$\xi_j^k = \phi_k(x_j), \quad h^{-2}\lambda_k \rightarrow \mu_k \text{ as } h \rightarrow 0$$

where $\mu_k = (k\pi)^2$ and $\phi_k(x) = \sin(k\pi x)$ are the eigenpairs of the continuous problem (23.7), namely

$$-\phi_k''(x) = \mu_k \phi_k(x).$$

For $d = 2$, it is also easy to derive a closed-form solution of the eigenpairs of A given by

$$A = \text{tridiag}(-I, B, -I), \quad B = \text{tridiag}(-1, 4, -1)$$

(see also (23.3)) as follows:

$$(4.85) \quad \lambda_{kl}(A) = 4 \left(\sin^2 \frac{k\pi}{2(N+1)} + \sin^2 \frac{l\pi}{2(N+1)} \right)$$

and

$$(4.86) \quad \phi_{ij}^{kl} = \sin \frac{ki\pi}{N+1} \sin \frac{lj\pi}{N+1}, \quad 1 \leq i \leq N, \quad 1 \leq j \leq N.$$

These explicit expressions of eigenpairs clearly lead to Conclusion 3 in Lemma 25. \square

Remark 10. This result is actually valid for a large class of finite element discretizations of second-order elliptic boundary value problems in general domains.

Remark 11. For $d = 1$, the solution of (3.7) is easy to obtain by many different methods such as the simple Gaussian-elimination, which needs only $\mathcal{O}(N)$ operations. For $d = 2$, using the fact that eigenpairs of A are explicitly given as (4.85) and (4.86), the solution of (23.3) can also be obtained with $\mathcal{O}(N \log N)$ operations by what is known as the Fast Fourier Transform (FFT) method. However, for more general cases, these traditional methods cannot obtain the solution with $\mathcal{O}(N)$ or $\mathcal{O}(N \log N)$ operations. They usually require $\mathcal{O}(N^2)$ or $\mathcal{O}(N^3)$ operations. One main conclusion that we will draw and support in the rest of the notes is that a system such as (23.3) in more general situations can also be solved with $\mathcal{O}(N)$ or $\mathcal{O}(N \log N)$ operations by using more advanced algorithms such as multigrid methods.

4.9 Linear algebraic system and its solution

Let $\{\phi_j : 1 \leq j \leq N_h\}$ be the nodal basis functions. The the stiffness matrix $A = (a_{ij})$, $a_{ij} = a(\phi_i, \phi_j)$, has the following properties:

- A is sparse, has $O(N_h)$ nonzeros (since $a_{ij} = 0$ if x_i and x_j are not neighbors).
- A is SPD.
- The condition number of A satisfies $\kappa(A) = O(h^{-2})$.

In fact, A is obviously symmetric. Now, if $u_h = \sum_{j=1}^{N_h} \mu_j \phi_j \in R^{N_h}$, then for $\mu = (\mu_1, \dots, \mu_{N_h})$

$$\begin{aligned} \mu^t A \mu &= \sum_{i,j=1}^{N_h} a_{ij} \mu_i \mu_j = \sum_{i,j=1}^{N_h} a(\phi_i, \phi_j) \mu_i \mu_j \\ &= a(u_h, u_h) = \int_{\Omega} |\nabla u_h|^2 \geq 0. \end{aligned}$$

As a consequence, the stiffness matrix A for discretization on uniform mesh is block tri-diagonal $A = \text{diag}(-I, B, -I)$ with $B = \text{diag}(-1, 4, -1)$. By Lemma 25 the eigenvalues of A are as follows

$$\lambda_{ij}(A) = 4\left(\sin^2 \frac{i\pi}{2(N+1)} + \sin^2 \frac{j\pi}{2(N+1)}\right).$$

As a consequence, we have

$$\kappa(A) = \frac{\sin^2 \frac{N\pi}{2(N+1)} + \sin^2 \frac{N\pi}{2(N+1)}}{\sin^2 \frac{\pi}{2(N+1)} + \sin^2 \frac{\pi}{2(N+1)}},$$

which implies

$$\kappa(A) \approx N^2 \approx h^{-2}.$$