Chapter 2
Introduction to Finite Element Methods

The finite element method (FEM) is arguably one of the most robust and popular numerical methods used for solving various partial differential equations (PDEs). Due to the diligent work of many researchers over the past several decades, the fundamental theory and implementation of FEM have been well established as evidenced by many excellent books published in this area (e.g., [4, 20, 21, 39, 51, 54, 65, 78, 158, 163, 243]).

In this chapter, we provide a brief introduction to the basic FEM theory and programming techniques in order to prepare readers for extending these skills to solve metamaterial Maxwell's equations in later chapters.

The outline of this chapter is as follows: In Sect. 2.1, we introduce some basic concepts about constructing two-dimensional (2-D) and three-dimensional (3-D) Lagrange finite elements. Then in Sect. 2.2, we provide a succinct introduction to Sobolev spaces. After that, we present some classic finite element results such as the interpolation error estimates for Lagrange finite elements in Sect. 2.3. To prepare readers for more complicated analysis and algorithmic implementation in later chapters, we then provide a brief introduction to some basic finite element error analysis tools for elliptic type problems in Sect. 2.4. Finally, in Sect. 2.5, we introduce some standard coding techniques for implementing Lagrange finite elements for solving the second order elliptic problems.

2.1 Introduction to Finite Elements

Suppose that we want to numerically solve a given PDE on a fixed domain Ω. To use the finite element method, basically we need to proceed the following steps:

1. Rewrite a given PDE into an equivalent weak formulation.
2. Subdivide the physical domain Ω into smaller simple geometrical subdomains (or elements). Often we use tetrahedra, hexahedra or prisms for a 3-D domain, and triangles or quadrilaterals in a 2-D domain.

J. Li and Y. Huang, *Time-Domain Finite Element Methods for Maxwell's Equations in Metamaterials*, Springer Series in Computational Mathematics 43,
DOI 10.1007/978-3-642-33789-5_2, © Springer-Verlag Berlin Heidelberg 2013

3. Design a proper finite element, which is often denoted as a triple (K, P_K, Σ_K) according to [78]. Here K is a geometric element, P_K is a space of functions on K, and Σ_K is the so-called *degrees of freedom* of the finite element. For efficiency and simplicity reasons, P_K is often formed by polynomials. The degrees of freedom are often formed by values (or derivatives) of a function at the element vertices, or some integral forms of a function on the element edges and/or on the element.
4. Construct a finite element solution formed by basis functions of P_K to approximate the infinite dimensional solution in the weak formulation. Doing this leads to a system of discretized linear (or nonlinear) equations.
5. Solve the system of discretized equations and postprocess the obtained solution to get the numerical solution for the original given PDE.

In this section, we focus on the third step. The rest steps will be elaborated in later sections.

First, let us introduce some common notation for polynomial spaces used throughout the book. Let P_k be the space of polynomials of maximum total degree k in d variables $x_1, \cdots, x_d$, and $\tilde{P}_k$ be the space of polynomials of total degree exactly k in d variables $x_1, \cdots, x_d$. Hence a polynomial $p \in P_k$ if and only if it can be written as

$$p(\mathbf{x}) = \sum_{\alpha_1+\cdots+\alpha_d \le k} c_{\alpha_1,\cdots,\alpha_d} x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$$

at any point $\mathbf{x} = (x_1, \cdots, x_d)$, and a polynomial $\tilde{p} \in \tilde{P}_k$ if and only if it can be written as

$$\tilde{p}(\mathbf{x}) = \sum_{\alpha_1+\cdots+\alpha_d = k} c_{\alpha_1,\cdots,\alpha_d} x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$$

for proper coefficients $c_{\alpha_1,\cdots,\alpha_d}$. Here all α_i are assumed to be non-negative integers.

It is easy to see that in $\mathscr{R}^d$, the dimensions of the spaces P_k and $\tilde{P}_k$ are

$$dim(P_k) = \binom{k+d}{k} = \frac{(k+d)\cdots(k+1)}{d!} \tag{2.1}$$

and

$$dim(\tilde{P}_k) = dim(P_k) - dim(P_{k-1}), \tag{2.2}$$

respectively.

On a d-dimensional rectangle, we need a tensor-product polynomial space $Q_{l_1,\cdots,l_d}$, which is formed by polynomials of maximum degree l_k in x_k, where $1 \le k \le d$, i.e., a polynomial $q \in Q_{l_1,\cdots,l_d}$ if and only if it can be written as

$$q(\mathbf{x}) = \sum_{0 \le \alpha_1 \le l_1, \cdots, 0 \le \alpha_d \le l_d} c_{\alpha_1, \cdots, \alpha_d} x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$$

for some coefficients $c_{\alpha_1, \cdots, \alpha_d}$.

The dimension of $Q_{l_1, \cdots, l_d}$ is easy to calculate as

$$dim(Q_{l_1, \cdots, l_d}) = (l_1 + 1)(l_2 + 1) \cdots (l_d + 1).$$

Before we move forward, let us introduce the *unisolvent* concept used in finite element.

Definition 2.1. A finite element (K, P_K, Σ_K) is *unisolvent* if the set of degrees of freedom Σ_K uniquely defines a function in P_K.

Below are some examples of unisolvent finite elements.

Example 2.1. Consider a triangle K with vertices $(x_i, y_i), i = 1, 2, 3,$ ordered counterclockwisely. It is known that the area A of this triangle can be calculated as

$$A = \frac{1}{2} \begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix}.$$

Now we can define the so-called *barycentric coordinates* $\lambda_i(x, y)$ of the triangle, which is often denoted as

$$\lambda_i(x, y) = \frac{1}{2A}(\alpha_i + \beta_i x + \gamma_i y), \quad i = 1, 2, 3, \tag{2.3}$$

where constants α_i, β_i and γ_i are

$$\alpha_i = x_j y_k - x_k y_j, \;\; \beta_i = y_j - y_k, \;\; \gamma_i = -(x_j - x_k),$$

where $i \neq j \neq k$, and i, j and k permute naturally.

It is not difficult to see that: for any $1 \le i, j \le 3$, we have

$$\lambda_i(x_j, y_j) = \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases}$$

from which we see that any function $u \in P_1$ on K can be uniquely represented by

$$u(x, y) = \sum_{i=1}^{3} u(x_i, y_i) \lambda_i(x, y) \quad \forall\ (x, y) \in K.$$

Using the triple notation, we can denote this unisolvent element (often called as P_1 element) as:

$$
\begin{aligned}
K &= \{\text{The triangle with vertices } (x_i, y_i), i = 1, 2, 3, \} \\
P_K &= \text{Polynomials of degree 1 in variables } x \text{ and } y, \\
\Sigma_K &= \{\text{Function values at the vertices: } u(x_i, y_i), i = 1, 2, 3.\}
\end{aligned}
$$

Using the barycentric coordinates λ_i, we can define other finite elements. Below is the so-called P_2 element:

Example 2.2.

$$
\begin{aligned}
K = \{&\text{A triangle with vertices } a_i(x_i, y_i), i = 1, 2, 3, \\
&\qquad \text{and edge midpoints } a_{ij},\ 1 \le i < j \le 3\}, \\
P_K = {}&\text{Polynomials of degree 2 in variables } x \text{ and } y. \\
\Sigma_K = \{&\text{Function values at vertices and midpoints: } u(a_i), u(a_{ij}), i, j = 1, 2, 3.\}
\end{aligned}
$$

We can prove that P_2 element is unisolvent.

Lemma 2.1. *Any function $u \in P_2$ on K is uniquely determined by its values at all vertices and edge midpoints, i.e., by $u(a_i)$, $1 \le i \le 3$, and $u(a_{ij})$, $1 \le i < j \le 3$.*

Proof. Note that the total number of degrees of freedom in Σ_K is equal to 6, which is the same as $dim(P_2)$. Hence we only need to show that if $u(a_i) = u(a_{ij}) = 0$, then $u \equiv 0$. Note that the restriction of u to edge a_2a_3 is a quadratic function in one variable and vanishes at three distinct points (i.e., at a_2, a_3 and a_{23}), hence $u(x, y)$ must be zero on this edge. This implies that u should contain a factor $\lambda_1(x, y)$.

By the same argument, u must be zero on edge a_1a_2, which implies that u should contain a factor $\lambda_3(x, y)$. Similarly, because u is zero on edge a_1a_3, u should also contain a factor $\lambda_2(x, y)$. Therefore, we can write

$$u(x, y) = c\lambda_1(x, y)\lambda_2(x, y)\lambda_3(x, y),$$

which becomes a third-order polynomial unless $c = 0$. Hence, $u \equiv 0$, which concludes the proof. □

Actually, any function u in P_2 element can be explicitly represented as [78, p. 47]:

$$u(x, y) = \sum_{i=1}^{3} u(a_i)\lambda_i(x, y)(2\lambda_i(x, y) - 1) + \sum_{1 \le i < j \le 3} 4u(a_{ij})\lambda_i(x, y)\lambda_j(x, y).$$

Similarly, we can construct a P_1 element on a tetrahedron.

Example 2.3.

$$\begin{aligned} \textit{Denote}\quad K &= \{\text{A tetrahedron with four vertices } V_i(x_i, y_i, z_i), i = 1,2,3,4\}, \\ P_K &= \text{Polynomials of degree 1 in three variables}, \\ \Sigma_K &= \{\text{Function values at the vertices: } u(V_i), i = 1,2,3,4\}. \end{aligned}$$

On element K, any function u of P_K can be written as

$$u(x,y,z) = a + bx + cy + dz, \tag{2.4}$$

where the coefficients a, b, c, and d can be determined by enforcing (2.4) equal to the given values $u(V_i)$ at the four vertices of the tetrahedron. Introducing the short notation $u_i = u(V_i)$, we have

$$\begin{aligned} a + bx_1 + cy_1 + dz_1 &= u_1, \\ a + bx_2 + cy_2 + dz_2 &= u_2, \\ a + bx_3 + cy_3 + dz_3 &= u_3, \\ a + bx_4 + cy_4 + dz_4 &= u_4, \end{aligned}$$

solving which we obtain

$$a = \frac{1}{6V} \begin{vmatrix} u_1 & u_2 & u_3 & u_4 \\ x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \\ z_1 & z_2 & z_3 & z_4 \end{vmatrix} = \frac{1}{6V}(a_1u_1 + a_2u_2 + a_3u_3 + a_4u_4),$$

$$b = \frac{1}{6V} \begin{vmatrix} 1 & 1 & 1 & 1 \\ u_1 & u_2 & u_3 & u_4 \\ y_1 & y_2 & y_3 & y_4 \\ z_1 & z_2 & z_3 & z_4 \end{vmatrix} = \frac{1}{6V}(b_1u_1 + b_2u_2 + b_3u_3 + b_4u_4),$$

$$c = \frac{1}{6V} \begin{vmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \\ u_1 & u_2 & u_3 & u_4 \\ z_1 & z_2 & z_3 & z_4 \end{vmatrix} = \frac{1}{6V}(c_1u_1 + c_2u_2 + c_3u_3 + c_4u_4),$$

$$d = \frac{1}{6V} \begin{vmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \\ u_1 & u_2 & u_3 & u_4 \end{vmatrix} = \frac{1}{6V}(d_1u_1 + d_2u_2 + d_3u_3 + d_4u_4),$$

where the coefficients a, b, c and d can be determined from the expansion of determinants, and V denotes the volume of the element, which can be expressed as

$$V = \frac{1}{6}\begin{vmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \\ z_1 & z_2 & z_3 & z_4 \end{vmatrix}.$$

Substituting a, b, c and d back into (2.4) and collecting like terms u_j, we have

$$u(x, y, z) = \sum_{j=1}^{4} u(x_j, y_j, z_j) N_j(x, y, z), \tag{2.5}$$

where functions N_j are given by

$$N_j(x, y, z) = \frac{1}{6V}(a_j + b_j x + c_j y + d_j z). \tag{2.6}$$

We like to remark that $N_j(x, y, z)$ are often called the *shape functions* of the finite element, and they have the property

$$N_j(x_i, y_i, z_i) = \delta_{ij} = \begin{cases} 1 & i = j, \\ 0 & i \neq j. \end{cases}$$

By a similar technique, we can construct finite elements on d-rectangles. First, we give an example on a rectangular element.

Example 2.4. Consider a rectangle $K = [x_c - h_x, x_c + h_x] \times [y_c - h_y, y_c + h_y]$, whose four vertices are oriented counterclockwisely, starting with the bottom-left vertex $V_1 = (x_c - h_x, y_c - h_y)$. Then we can denote the Q_1 rectangular element by the triple:

$$\begin{aligned} K &= \{\text{The rectangle with four vertices } V_i, i = 1, 2, 3, 4\}, \\ P_K &= \text{Polynomial } Q_{1,1} \text{ on } K, \\ \Sigma_K &= \{\text{Function values at the vertices: } u(V_i), i = 1, 2, 3, 4\}. \end{aligned}$$

It is easy to check that any function u of $Q_{1,1}$ on K can be uniquely represented as follows:

$$u(x, y) = \sum_{j=1}^{4} u(x_j, y_j) N_j(x, y),$$

where the shape functions N_j are given by

$$N_1(x, y) = \frac{1}{A}(x_c + h_x - x)(y_c + h_y - y),$$

$$\begin{aligned}
N_2(x,y) &= \frac{1}{A}(x - x_c + h_x)(y_c + h_y - y),\\
N_3(x,y) &= \frac{1}{A}(x - x_c + h_x)(y - y_c + h_y),\\
N_4(x,y) &= \frac{1}{A}(x - x_c + h_x)(y - y_c + h_y),
\end{aligned}$$

here $A = 4h_x h_y$ denotes the area of the rectangle.

Now we give an example on a cubic element.

Example 2.5. Consider a cube

$$K = [x_c - h_x, x_c + h_x] \times [y_c - h_y, y_c + h_y] \times [z_c - h_z, z_c + h_z],$$

whose eight vertices are oriented counterclockwisely (four on the bottom face, and four on the top face), starting with the front-bottom-left vertex $V_1 = (x_c - h_x, y_c - h_y, z_c - h_z)$. We can define a unisolvent Q_1 cubic element by the triple:

$$\begin{aligned}
K &= \{\text{The cube with 8 vertices } V_i, i = 1, 2, \cdots, 8\},\\
P_K &= \text{Polynomial } Q_{1,1,1} \text{ on } K,\\
\Sigma_K &= \{\text{Function values at the vertices: } u(V_i), i = 1, 2, \cdots, 8\}.
\end{aligned}$$

It is not difficult to check that any function u of $Q_{1,1,1}$ on K can be uniquely represented as follows:

$$u(x,y,z) = \sum_{j=1}^{8} u(x_j, y_j, z_j) N_j(x,y,z),$$

where the shape functions N_j are given by

$$\begin{aligned}
N_1(x,y,z) &= \frac{1}{V}(x_c + h_x - x)(y_c + h_y - y)(z_c + h_z - z),\\
N_2(x,y,z) &= \frac{1}{V}(x - x_c + h_x)(y_c + h_y - y)(z_c + h_z - z),\\
N_3(x,y,z) &= \frac{1}{V}(x - x_c + h_x)(y - y_c + h_y)(z_c + h_z - z),\\
N_4(x,y,z) &= \frac{1}{V}(x_c + h_x - x)(y - y_c + h_y)(z_c + h_z - z),
\end{aligned}$$

and the other four N_j have the same form as above except that the last terms are changed to $z - (z_c - h_z)$. Here $V = 8h_x h_y h_z$ denotes the volume of the cube.

2.2 Functional Analysis and Sobolev Spaces

2.2.1 Basic Functional Analysis

In our later analysis of Maxwell's equations, we shall appeal to many basic theorems from functional analysis. In this section, we simply summarize some definitions and theorems to be used in later sections. Readers interested in details can consult specialized books such as [53].

Let X be a normed linear space with norm $||\cdot||_X$.

Definition 2.2. A sequence $\{u_k\}_{k=1}^{\infty} \subset X$ *converges* to $u \in X$, denoted as $u_k \to u$, if $\lim_{k\to\infty} ||u_k - u||_X = 0$. If for any $\epsilon > 0$, there exists $N > 0$ such that

$$||u_k - u_m||_X < \epsilon \quad \text{for any } k, m \geq N,$$

then the sequence $\{u_k\}_{k=1}^{\infty} \subset X$ is called a *Cauchy sequence*.

Definition 2.3. The space X is called *complete* if each Cauchy sequence in X converges; namely, whenever $\{u_k\}_{k=1}^{\infty}$ is a Cauchy sequence, there exists $u \in X$ such that $u_k \to u$.

Definition 2.4. A complete normed linear space is called a *Banach space*.

Definition 2.5. A collection $\mathscr{M}$ of subsets of $\mathscr{R}^d$ is called *σ-algebra* if

(i) $\emptyset, \mathscr{R}^d \in \mathscr{M}$;
(ii) $A \in \mathscr{M}$ implies that $\mathscr{R}^d \setminus A \in \mathscr{M}$;
(iii) $\{A_k\}_{k=1}^{\infty} \subset \mathscr{M}$ implies that $\cup_{k=1}^{\infty} A_k, \cap_{k=1}^{\infty} A_k \in \mathscr{M}$.

It can be proved that there exists a σ-algebra $\mathscr{M}$ of subsets of $\mathscr{R}^d$ and a mapping $|\cdot| : \mathscr{M} \to [0, +\infty]$ with the following properties:

(i) Every open subset of $\mathscr{R}^d$ and every closed subset of $\mathscr{R}^d$ belong to $\mathscr{M}$.
(ii) If A is a ball of $\mathscr{R}^d$, then $|A|$ equals the d-dimensional volume of A.

The sets in $\mathscr{M}$ are often called *Lebesgue measurable* sets and $|\cdot|$ is *Lebesgue measure*. Hence Lebesgue measure provides a way of describing the volume of subsets of $\mathscr{R}^d$.

Definition 2.6. A function $f : \mathscr{R}^d \to \mathscr{R}$ is called a *measurable function* if $f^{-1}(S) \in \mathscr{M}$ for every open subset $S \subset \mathscr{R}$.

Definition 2.7. Let Ω be an open subset of $\mathscr{R}^d$, and $1 \leq p \leq \infty$. For a measurable function $f : \Omega \to \mathscr{R}$, we define the norm

$$||f||_{L^p(\Omega)} = (\int_{\Omega} |f|^p dx)^{1/p} \quad \text{if } 1 \leq p < \infty$$

and

$$||f||_{L^\infty(\Omega)} = \inf_{S\subset\Omega,|S|=0} \sup_{\Omega\setminus S} |f(x)| = \text{ess sup}_\Omega |f(x)| \quad \text{if } p = \infty.$$

Furthermore, $L^p(\Omega)$ is defined to be the linear space of all measurable functions $f : \Omega \to \mathscr{R}$ for which $||f||_{L^p(\Omega)} < \infty$.

It is known that $L^p(\Omega), 1 \leq p \leq \infty$, is a Banach space.

Definition 2.8. Let X be a real linear space. A mapping $(\cdot,\cdot) : X \times X \to \mathscr{R}$ is called an *inner product* if

(i) $(u, v) = (v, u)$ for any $u, v \in X$;
(ii) Each of the maps $u \to (u, v)$ and $v \to (u, v)$ is linear on X;
(iii) $(u, u) \geq 0$ for any $u \in X$;
(iv) $(u, u) = 0$ if any only if $u = 0$.

Furthermore, if $(\cdot,\cdot)$ is an inner product, the associated norm is

$$||u||_X = (u, u)^{1/2} \quad \text{for any} \quad u \in X.$$

Moreover, if X is complete with respect to the norm $||\cdot||_X$, then X is called a *Hilbert space*.

A simple example of a Hilbert space is $L^2(\Omega)$, which has the scalar inner product

$$(u, v) = \int_\Omega u(x)v(x)dx.$$

Two elementary estimates for Hilbert spaces are often used in numerical analysis. One is the Cauchy-Schwarz inequality

$$|(u, v)| \leq ||u||_X ||v||_X \quad \forall\, u, v \in X. \tag{2.7}$$

The other one is the arithmetic-geometric mean inequality: for any $u, v \in X$ and $\delta > 0$, we have

$$|(u, v)| \leq \frac{\delta}{2}||u||_X^2 + \frac{1}{2\delta}||v||_X^2. \tag{2.8}$$

2.2.2 Sobolev Spaces

Sobolev spaces are named after the Russian mathematician Sergei Sobolev, who introduced this concept around 1950. A Sobolev space is a space of functions equipped with a norm which can be used to measure both the size and regularity of a function. Hence, Sobolev spaces play a very important role in analyzing partial differential equations. In this section, we just present some important properties of Sobolev spaces related to our later usage. For more comprehensive discussions, readers can consult specialized books on Sobolev spaces (e.g. [2]).

Let $C_0^\infty(\Omega)$ (or $D(\Omega)$) denote the space of infinitely differentiable functions with compact support in Ω. Furthermore, let $\alpha = (\alpha_1, \cdots, \alpha_d)$ be an d-tuple of nonnegative integers and denote its length by $|\alpha| = \sum_{i=1}^d \alpha_i$.

Before we introduce the weak derivative concept, let us provide some motivation. Suppose we are given a function $u \in C^1(\Omega)$. If $\phi \in C_0^\infty(\Omega)$, using integration by parts we have

$$\int_\Omega u\phi_{x_i} dx = -\int_\Omega u_{x_i}\phi dx \quad i = 1, 2, \cdots, d. \tag{2.9}$$

Here no boundary integral term exists, since ϕ has compact support in Ω and vanishes on $\partial\Omega$. Similarly, if $u \in C^k(\Omega)$ for some integer $k > 1$, we can obtain

$$\int_\Omega u D^\alpha \phi dx = (-1)^{|\alpha|} \int_\Omega D^\alpha u \phi dx \quad \forall\, \phi \in C_0^\infty(\Omega). \tag{2.10}$$

Now (2.10) has a problem if u is not C^k, since $D^\alpha u$ on the right hand side has no obvious meaning. We resolve this difficulty by introducing the concept of weak derivative. Let us denote the set of locally integrable functions

$$L^1_{loc}(\Omega) = \{v : \ v \in L^1(K), \ \text{ for all compact set } K \subset \Omega\}.$$

Definition 2.9. Let $u, v \in L^1_{loc}(\Omega)$. We say v is the α-th *weak partial derivative* of u, denoted as $v = D^\alpha u$, provided that

$$\int_\Omega u D^\alpha \phi dx = (-1)^{|\alpha|} \int_\Omega v\phi dx \tag{2.11}$$

holds true for all $\phi \in C_0^\infty(\Omega)$

It is easy to see that if a weak derivative exists, then it is uniquely defined up to a set of measure zero. Furthermore, for any $u \in C^{|\alpha|}(\Omega)$, the weak derivative $D^\alpha u$ exists and equals the classic derivative.

With the above preparations, we can define the Sobolev space $W^{k,p}(\Omega)$.

Definition 2.10. The Sobolev space $W^{k,p}(\Omega)$ consists of all locally summable functions $u : \Omega \to R$, i.e., $u \in L^1_{loc}(\Omega)$, such that for each multi-index α with $|\alpha| \leq k$, $D^\alpha u$ exists in the weak sense and belongs to $L^p(\Omega)$. Moreover, for any $u \in W^{k,p}(\Omega)$, its norm is defined to be

$$||u||_{W^{k,p}(\Omega)} = \begin{cases} (\sum_{|\alpha|\leq k} \int_\Omega |D^\alpha u|^p dx)^{1/p} & \text{if } 1 \leq p < \infty, \\ \sum_{|\alpha|\leq k} \text{ess sup}_\Omega |D^\alpha u| & \text{if } p = \infty. \end{cases}$$

It is known that $W^{k,p}(\Omega)$ is a Banach space (see e.g., [51, p. 28]).

We like to remark that a function belonging to a Sobolev space can be discontinuous and/or unbounded. A simple example is $u(x) = |x|^{-\alpha}$, which is unbounded at $x = 0$ for $\alpha > 0$. However, $u(x)$ belongs to some Sobolev space.

Lemma 2.2. *Let* $u(x) = |x|^{-\alpha}$ $(x \neq 0)$ *defined in the open unit ball* $\Omega = B(0,1)$ *in* $\mathscr{R}^d$. *Then* $u \in W^{1,p}(\Omega)$ *if and only if* $\alpha < \frac{d-p}{p}$.

Proof. Note that u is smooth away from 0, and

$$u_{x_i}(x) = -\alpha x_i |x|^{-(\alpha+2)} \quad (x \neq 0).$$

For any $\phi \in C_0^\infty(\Omega)$ and fixed $\epsilon > 0$, we have

$$\int_{\Omega \backslash B(0,\epsilon)} u\phi_{x_i} dx = \int_{\partial B(0,\epsilon)} u\phi n_i ds - \int_{\Omega \backslash B(0,\epsilon)} u_{x_i}\phi dx,$$

where n_i denotes the unit inward normal on $\partial B(0,\epsilon)$.

If $\alpha + 1 < d$, then $|Du(x)| = \frac{\alpha}{|x|^{\alpha+1}} \in L^1(\Omega)$, in which case,

$$|\int_{\partial B(0,\epsilon)} u\phi n_i ds| \leq ||\phi||_{L^\infty} \int_{\partial B(0,\epsilon)} \epsilon^{-\alpha} ds \leq C\epsilon^{d-1-\alpha} \rightarrow 0 \text{ as } \epsilon \rightarrow 0.$$

Hence for any $0 \leq \alpha < d - 1$, we have

$$\int_{\Omega} u\phi_{x_i} dx = -\int_{\Omega} u_{x_i}\phi dx \quad \forall\, \phi \in C_0^\infty(\Omega).$$

Similarly, it is easy to see that

$$|Du(x)| = \frac{\alpha}{|x|^{\alpha+1}} \in L^p(\Omega) \text{ if and only if } (\alpha + 1)p < d,$$

which concludes the proof. □

Definition 2.11. Given a subset $S \subset X$, the *closure* of S in X (usually denoted as $\overline{S}$) is the set of all limits of convergent subsequence of S using the X norm. Furthermore, if $\overline{S} = X$, we say that the subset S is *dense* in X.

Definition 2.12. $W_0^{k,p}(\Omega)$ is denoted as the *closure* of $C_0^\infty(\Omega)$ in $W^{k,p}(\Omega)$. When $p = 2$, we usually write

$$H^k(\Omega) = W^{k,2}(\Omega) \quad (\text{or } H_0^k(\Omega) = W_0^{k,2}(\Omega)),$$

since $H^k(\Omega)$ is a Hilbert space.

Definition 2.13. A function $f : \Omega \to \mathscr{R}$ is called *Lipschitz continuous* if

$$|f(x) - f(y)| \leq L|x - y|$$

for some constant $L > 0$ and all $x, y \in \Omega$.

Definition 2.14. Let Ω be an open and bounded domain of $\mathscr{R}^d (d \geq 2)$ with boundary $\partial\Omega$. We say that Ω is *Lipschitz* (or $\partial\Omega$ is a Lipschitz boundary), if there exists a finite open cover $U_1, \cdots, U_m$ of $\partial\Omega$ such that for $j = 1, \cdots, m$:

(i) $\partial\Omega \cap U_j$ is the graph of a Lipschitz function $\phi_j : \mathscr{R}^{d-1} \to \mathscr{R}$, and
(ii) $\Omega \cap U_j$ is on one side of this graph.

Namely, for any $x = (\tilde{x}, x_d) \in U_j$, where $\tilde{x} = (x_1, \cdots, x_{d-1})$, there exists a Lipschitz function ϕ_j such that $x_d = \phi_j(\tilde{x})$, $\Omega \cap U_j = \{x : x_d > \phi_j(\tilde{x})\}$ and $\partial\Omega \cap U_j = \{x : x_d = \phi_j(\tilde{x})\}$.

Definition 2.15. A normed space U is said to be *embedded* in another normed space V, denoted as $U \hookrightarrow V$, if

(i) U is a linear subspace of V;
(ii) The injection of U into V is continuous, i.e., there exists a constant $C > 0$ such that $||u||_V \leq C||u||_U \quad \forall\, u \in U$.

As we mentioned earlier, Sobolev spaces provide a way of quantifying the degree of smoothness of functions. The following theorem summaries some classic results [2].

Theorem 2.1. *(Sobolev embedding theorem) Suppose that Ω is an open set of R^d with a Lipschitz continuous boundary, and $1 \leq p < \infty$. Then the following embedding results hold true:*

(i) If $0 \leq kp < d$, then $W^{k,p}(\Omega) \hookrightarrow L^{p_*}(\Omega)$ for $p_* = \frac{dp}{d-kp}$.
(ii) If $kp = d$, then $W^{k,p}(\Omega) \hookrightarrow L^q(\Omega)$ for any q such that $p \leq q < \infty$.
(iii) If $kp > d$, then $W^{k,p}(\Omega) \hookrightarrow C^0(\overline{\Omega})$.

When we deal with time-dependent problems, we need some Sobolev spaces involving time. Let X denote a real Banach space with norm $|| \cdot ||_X$.

Definition 2.16. The space $L^p(0, T; X)$ consists of all measurable functions $u : [0, T] \to X$ with endowed norm

$$||u||_{L^p(0,T;X)} = (\int_0^T ||u(t)||_X^p dt)^{1/p} < \infty, \quad \text{if } 1 \leq p < \infty,$$

and

$$||u||_{L^\infty(0,T;X)} = \text{ess sup}_{0 \leq t \leq T} ||u(t)||_X < \infty, \quad \text{if } p = \infty.$$

Definition 2.17. The space $C(0,T;X)$ consists of all continuous functions $u : [0,T] \to X$ with

$$||u||_{C(0,T;X)} = \max_{0\le t\le T} ||u(t)||_X < \infty.$$

Definition 2.18. The Sobolev space $W^{k,p}(0,T;X)$ comprises all functions $u \in L^p(0,T;X)$ such that $\frac{d^\alpha u}{dt^\alpha}, 1 \le \alpha \le k$, exists in the weak sense and belongs to $L^p(0,T;\Omega)$. Furthermore,

$$||u||_{W^{k,p}(0,T;X)} = \begin{cases} (\int_0^T (||u(t)||_X^p + \sum_{1\le\alpha\le k} ||\frac{d^\alpha u}{dt^\alpha}||_X^p)dt)^{1/p} & \text{if } 1 \le p < \infty, \\ \text{ess sup}_{0\le t\le T}(||u(t)||_X + \sum_{1\le\alpha\le k} ||\frac{d^\alpha u}{dt^\alpha}||_X) & \text{if } p = \infty. \end{cases}$$

When $p = 2$, we denote

$$H^k(0,T;X) = W^{k,2}(0,T;X).$$

For many applications to be discussed later, we need a space of vector functions with a square-integrable divergence. Such a space is often denoted as $H(\text{div};\Omega)$, which is defined by

$$H(\text{div};\Omega) = \{\mathbf{v} \in (L^2(\Omega))^d : \nabla\cdot\mathbf{v} \in L^2(\Omega)\}, \tag{2.12}$$

with norm $||\mathbf{v}||_{H(\text{div};\Omega)} = (||\mathbf{v}||^2_{(L^2(\Omega))^d} + ||\nabla\cdot\mathbf{v}||^2_{L^2(\Omega)})^{1/2}$, where $\nabla\cdot$ is the divergence operator defined as

$$\nabla\cdot\mathbf{v} = \sum_{i=1}^d \frac{\partial v_i}{\partial x_i}, \quad \text{for all } \mathbf{v} \in (C_0^\infty(\Omega))^d,\ d = 2,3.$$

It is easy to prove that $H(\text{div};\Omega)$ is a Hilbert space with the inner product

$$(\mathbf{u},\mathbf{v})_{div} = \int_\Omega (\mathbf{u}\cdot\mathbf{v} + (\nabla\cdot\mathbf{u})(\nabla\cdot\mathbf{v}))d\mathbf{x}.$$

Similar to space $W_0^{k,p}(\Omega)$, we denote $H_0(\text{div};\Omega)$ as the closure of $(C_0^\infty(\Omega))^d$ with respect to the norm $||\cdot||_{H(\text{div};\Omega)}$.

Later, when we deal with Maxwell's equations, we need a space of vector functions with a square-integrable curl. In standard notation, we define this space by

$$H(\text{curl};\Omega) = \{\mathbf{u} \in (L^2(\Omega))^d : \nabla\times\mathbf{u} \in (L^2(\Omega))^d\}, \tag{2.13}$$

with norm $||\mathbf{u}||_{H(\text{curl};\Omega)} = (||\mathbf{u}||^2_{(L^2(\Omega))^d} + ||\nabla\times\mathbf{u}||^2_{(L^2(\Omega))^d})^{1/2}$, where $\nabla\times$ is the curl operator defined as

$$\nabla\times\mathbf{u} = (\frac{\partial u_3}{\partial x_2} - \frac{\partial u_2}{\partial x_3}, \frac{\partial u_1}{\partial x_3} - \frac{\partial u_3}{\partial x_1}, \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2})',$$

for any 3-D vector $\mathbf{u} = (u_1, u_2, u_3)' \in (C_0^\infty(\Omega))^d,\ d = 3$. For a 2-D vector, the curl operator becomes as

$$\nabla \times \mathbf{u} = \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2}, \quad \text{for all } \mathbf{u} = (u_1, u_2)' \in (C_0^\infty(\Omega))^2.$$

It is easy to prove that $H(\text{curl}; \Omega)$ is a Hilbert space with the inner product

$$(\mathbf{u}, \mathbf{v})_{curl} = \int_\Omega (\mathbf{u} \cdot \mathbf{v} + (\nabla \times \mathbf{u}) \cdot (\nabla \times \mathbf{v})) d\mathbf{x}.$$

Furthermore, $H_0(\text{curl}; \Omega)$ is denoted as the closure of $(C_0^\infty(\Omega))^d$ with respect to the norm $||\cdot||_{H(\text{curl};\Omega)}$.

Similarly, with higher regularity, we can define the space

$$H^\alpha(\text{curl}; \Omega) = \{\mathbf{v} \in (H^\alpha(\Omega))^d : \nabla \times \mathbf{v} \in (H^\alpha(\Omega))^d\}, \tag{2.14}$$

for any $\alpha > 0$ with norm $||\mathbf{v}||_{\alpha,\text{curl}} = (||\mathbf{v}||^2_{(H^\alpha(\Omega))^d} + ||\nabla \times \mathbf{v}||^2_{(H^\alpha(\Omega))^d})^{1/2}$. When $\alpha = 0$, $H^\alpha(\text{curl}; \Omega)$ reduces to $H(\text{curl}; \Omega)$.

2.3 Classic Finite Element Theory

To use the finite element method to solve a PDE, we have to build up a finite dimensional space of functions on the physical domain Ω. To do this, we first need to generate a finite element mesh covering the domain Ω, i.e., we need to construct a set T_h of non-overlapping elements K_i satisfying the following conditions:

(i) $\overline{\Omega} = \bigcup_{K_i \in T_h} \overline{K_i}$;
(ii) Each $K \in T_h$ is a Lipschitz domain, and has a positive measurement;
(iii) For any distinct elements K_1 and K_2 in T_h, we have $K_1 \cap K_2 = \emptyset$. In other words, any two neighboring elements have to meet at the common vertices, match exactly at a common edge or a common face.

A mesh satisfying conditions (i)–(iii) is often called *conforming mesh*. Note that in hp finite element method [97, 98, 255], condition (iii) is often violated, in which case we have the so-called *non-conforming mesh*.

2.3.1 Conforming and Non-conforming Finite Elements

After creating a mesh for Ω, we can use the element-wise defined finite elements to construct a global finite element space on Ω by lumping together all the element degrees of freedom. The key here is how to define the element degrees of freedom to guarantee the needed global smoothness for the finite element solution. For this, we need to classify finite elements into two classes: *conforming* or *non-conforming*.

Definition 2.19. Let V be a space of functions. The finite element (K, P_K, Σ_K) is said to be V *conforming* if its corresponding global finite element space is a subspace of V. Otherwise, the finite element is said to be V *nonconforming*.

It is well-known that for a finite element space to be $H^1(\Omega)$ conforming, the global finite element function has to be continuous as stated in the next lemma (cf. [78, Theorem 2.1.1] and [217, Lemma 5.3]).

Lemma 2.3. *Let K_1 and K_2 be two non-overlapping Lipschitz domains having a common interface Λ such that $\overline{K_1} \cap \overline{K_2} = \Lambda$. Assume that $u_1 \in H^1(K_1)$ and $u_2 \in H^1(K_2)$, and $u \in L^2(K_1 \cup K_2 \cup \Lambda)$ be defined by*

$$u = \begin{cases} u_1 & \text{on } K_1, \\ u_2 & \text{on } K_2. \end{cases}$$

Then $u_1 = u_2$ on Λ implies that $u \in H^1(K_1 \cup K_2 \cup \Lambda)$.

Proof. Suppose that we have a function $u \in L^2(K_1 \cup K_2 \cup \Lambda)$ defined by $u|_{K_i} = u_i, i = 1, 2$, and $u_1 = u_2$ on Λ. To prove that $u \in H^1(K_1 \cup K_2 \cup \Lambda)$, for any function $\boldsymbol{\phi} \in (C_0^\infty(K_1 \cup K_2 \cup \Lambda))^d$, using integration by parts, we have

$$\int_{K_1 \cup K_2 \cup \Lambda} u \frac{\partial \boldsymbol{\phi}}{\partial x_i} d\mathbf{x} = \int_{K_1} u \frac{\partial \boldsymbol{\phi}}{\partial x_i} d\mathbf{x} + \int_{K_2} u \frac{\partial \boldsymbol{\phi}}{\partial x_i} d\mathbf{x}$$
$$= -\int_{K_1} \frac{\partial (u|_{K_1})}{\partial x_i} \boldsymbol{\phi} d\mathbf{x} - \int_{K_2} \frac{\partial (u|_{K_2})}{\partial x_i} \boldsymbol{\phi} d\mathbf{x} + \int_\Lambda (u_1 \boldsymbol{\phi} \cdot \mathbf{n}_{i,1} + u_2 \boldsymbol{\phi} \cdot \mathbf{n}_{i,2}) ds,$$

where $\mathbf{n}_{i,j}$ denotes the unit outward normal to ∂K_j, $j = 1, 2$, respectively.

Denote $v_i = \frac{\partial (u|_{K_l})}{\partial x_i}, i = 1, \cdots, d,$ on $K_l, l = 1, 2$. Using the assumption that $u_1 = u_2$ on Λ, we see that the boundary integral term vanishes. Hence, we have

$$\int_{K_1 \cup K_2 \cup \Lambda} u \frac{\partial \boldsymbol{\phi}}{\partial x_i} d\mathbf{x} = -\int_{K_1 \cup K_2 \cup \Lambda} \mathbf{v} \cdot \boldsymbol{\phi} d\mathbf{x},$$

which shows that $u \in H^1(K_1 \cup K_2 \cup \Lambda)$ by the definition of weak derivative. □

Examples 2.1–2.5 given in Sect. 2.1 are H^1 conforming elements. Many popular nonconforming elements are illustrated in the nice paper by Carstensen and Hu [63]. Below we present two examples of non-conforming elements. The first one is the so-called rotated Q1 element.

Example 2.6. Consider a rectangle $K = [x_c - h_x, x_c + h_x] \times [y_c - h_y, y_c + h_y]$, whose four vertices are oriented counterclockwise, starting with the bottom-left vertex $V_1(x_c - h_x, y_c - h_y)$. The four mid-edge points starting from the bottom edge are denoted as $M_i, i = 1, 2, 3, 4$. The Q_1 rectangular element is defined by the following triple:

$$\begin{aligned} K &= \{\text{The rectangle with four vertices } V_i, i = 1,2,3,4\}, \\ P_K &= \text{Polynomial with basis } 1, x, y, x^2, y^2 \text{ on } K, \\ \Sigma_K &= \{\text{Function values at the mid-edge points: } u(M_i), i = 1,2,3,4\}. \end{aligned}$$

The unisolvence of this finite element (K, P_K, Σ_K) can be proved as follows. Assume that an arbitrary function u of P_K on K is represented as

$$u(x,y) = a_1 + a_2(x - x_c) + a_3(y - y_c) + a_4[(x - x_c)^2 - (y - y_c)^2], \quad (2.15)$$

where the unknown coefficients a_1, a_2, a_3 and a_4 can be determined by the interpolation conditions

$$u(M_i) = u_i, i = 1,2,3,4. \quad (2.16)$$

Imposing (2.16) at the four mid-edge points, we have

$$\begin{aligned} a_1 - h_y a_3 - h_y^2 a_4 &= u_1, \\ a_1 + h_y a_3 - h_y^2 a_4 &= u_3, \\ a_1 + h_x a_2 + h_x^2 a_4 &= u_2, \\ a_1 - h_x a_2 + h_x^2 a_4 &= u_4, \end{aligned}$$

which gives the following unique solution

$$\begin{aligned} a_1 &= [h_x^2(u_1 + u_3) + h_y^2(u_2 + u_4)]/2(h_x^2 + h_y^2), \\ a_2 &= (u_2 - u_4)/2h_x, \\ a_3 &= (u_3 - u_1)/2h_y, \\ a_4 &= [(u_2 + u_4) - (u_1 + u_3)]/2(h_x^2 + h_y^2). \end{aligned}$$

Another popular non-conforming element is Wilson's rectangular element.

Example 2.7. Consider the same rectangle $K = [x_c - h_x, x_c + h_x] \times [y_c - h_y, y_c + h_y]$ as Example 2.6. The Wilson element can be defined by the following triple:

$$\begin{aligned} K &= \{\text{The rectangle with four vertices } V_i, i = 1,2,3,4\}, \\ P_K &= \text{Polynomial } P_2 \text{ on } K, \\ \Sigma_K &= \{\text{Function values at the vertices: } u(V_i), i = 1,2,3,4, \text{ and mean values of} \\ &\qquad \text{the second derivatives of } u \text{ over } K: \frac{1}{|K|}\int_K \frac{\partial^2 u}{\partial x^2} d\mathbf{x}, \frac{1}{|K|}\int_K \frac{\partial^2 u}{\partial y^2} d\mathbf{x}\}. \end{aligned}$$

The unisolvence of this finite element (K, P_K, Σ_K) can be proved as follows. For a function u of P_2 on K, we can represent it as

$$u(x, y) = a_1+a_2(x-x_c)+a_3(y-y_c)+a_4(x-x_c)(y-y_c)+a_5(x-x_c)^2+a_6(y-y_c)^2, \tag{2.17}$$

where the unknown coefficients a_i can be determined by the given degrees of freedom Σ_K, which lead to the system of linear equations:

$$\begin{aligned}
&a_1 - h_x a_2 - h_y a_3 + h_x h_y a_4 + h_x^2 a_5 + h_y^2 a_6 = u_1,\\
&a_1 + h_x a_2 - h_y a_3 - h_x h_y a_4 + h_x^2 a_5 + h_y^2 a_6 = u_2,\\
&a_1 + h_x a_2 + h_y a_3 + h_x h_y a_4 + h_x^2 a_5 + h_y^2 a_6 = u_3,\\
&a_1 - h_x a_2 + h_y a_3 - h_x h_y a_4 + h_x^2 a_5 + h_y^2 a_6 = u_4,\\
&\int_K \frac{\partial^2 u}{\partial x^2} d\mathbf{x} = 2|K| a_5,\\
&\int_K \frac{\partial^2 u}{\partial y^2} d\mathbf{x} = 2|K| a_6.
\end{aligned}$$

Solving the above system, we can obtain the following unique solution

$$\begin{aligned}
a_5 &= \frac{1}{2|K|}\int_K \frac{\partial^2 u}{\partial x^2} d\mathbf{x}, \quad a_6 = \frac{1}{2|K|}\int_K \frac{\partial^2 u}{\partial y^2} d\mathbf{x},\\
a_1 &= \frac{u_1+u_2+u_3+u_4}{4} - \frac{h_x^2}{2|K|}\int_K \frac{\partial^2 u}{\partial x^2} d\mathbf{x} - \frac{h_y^2}{2|K|}\int_K \frac{\partial^2 u}{\partial y^2} d\mathbf{x},\\
a_2 &= (u_2 - u_1 + u_3 - u_4)/4h_x,\\
a_3 &= (u_3 + u_4 - u_1 - u_2)/4h_y,\\
a_4 &= [(u_3 - u_4) - (u_2 - u_1)]/4h_x h_y.
\end{aligned}$$

2.3.2 Basic Interpolation Error Estimates

Given a finite element (K, P_K, Σ_K), we can define a local Lagrange interpolant:

$$\Pi_K^k v = \sum_{i=1}^{n} v(a_i)\phi_i,$$

where a_i are the vertices of K, $v(a_i)$ are the degrees of freedom, and ϕ_i are the shape functions. Examples include the first-order interpolant

$$\Pi_K^1 v(x,y) = \sum_{i=1}^{3} v(a_i)\lambda_i(x,y),$$

and the second-order interpolant

$$\Pi_K^2 v(x,y) = \sum_{i=1}^{3} v(a_i)\lambda_i(2\lambda_i - 1) + \sum_{1\le i<j\le 3} 4v(a_{ij})\lambda_i\lambda_j,$$

mentioned in Sect. 2.1. Examples $\Pi_K^1 v$ and $\Pi_K^2 v$ are Lagrange interpolations, which satisfy the property $\Pi_K^{1,2} v(a_i) = v(a_i), i = 1,2,3$. More complicated interpolants such as Example 2.7 involve other degrees of freedom.

By piecing together the local interpolants, we can define a corresponding global interpolant Π_h^k as follows:

$$(\Pi_h^k v)|_K = \Pi_K^k(v|_K) \quad \forall\, K \in T_h.$$

We assume further that each element K of T_h can be obtained as an affine mapping of a reference element $\hat{K}$, i.e.,

$$K = F_K(\hat{K}), \quad F_K(\hat{x}) = B_K\hat{x} + b_K, \tag{2.18}$$

where B_K is a $d \times d$ non-singular matrix. The rest of this section is concerned about the estimate of interpolation error $v - \Pi_h^k v$.

Lemma 2.4. *Denote $\hat{v}(\hat{x}) = v(F_K(\hat{x}))$. If $v \in W^{m,p}(K), m \ge 0, p \in [1,\infty]$, then $\hat{v} \in W^{m,p}(\hat{K})$. Moreover, there exists a constant $C = C(m,p,d)$ such that*

$$|v|_{m,p,K} \le C||B_K^{-1}||^m |det(B_K)|^{1/p} |\hat{v}|_{m,p,\hat{K}}, \quad \forall\, \hat{v} \in W^{m,p}(\hat{K}), \tag{2.19}$$

and

$$|\hat{v}|_{m,p,\hat{K}} \le C||B_K||^m |det(B_K)|^{-1/p} |v|_{m,p,K}, \quad \forall\, v \in W^{m,p}(K), \tag{2.20}$$

where $||\cdot||$ denotes the matrix norm associated to the Euclidean norm in R^d.

Proof. For simplicity we just show the proof of (2.19) for $p = 2$. A complete proof can be found in the classic book by Ciarlet [78, Theorem 3.1.2]. Since $C^\infty(K)$ is dense in $H^m(K)$, it is sufficient to prove (2.19) for a smooth function v.

Using the chain rule and the mapping F_K, we have

$$|v|^2_{m,2,K} = \sum_{|\alpha|=m} ||D^\alpha v||^2_{0,K} \le C||B_K^{-1}||^{2m} \sum_{|\beta|=m} \int_K |\hat{D}^\beta \hat{v}|^2 d\mathbf{x}$$
$$\le C||B_K^{-1}||^{2m} \sum_{|\beta|=m} ||\hat{D}^\beta \hat{v}||^2_{0,\hat{K}} \cdot (\det(B_K)), \tag{2.21}$$

which completes the proof of (2.19) for $p = 2$. □

To obtain an explicit bound on $||B_K||$, $||B_K^{-1}||$ and $\det(B_K)$, we introduce some notation. For an element K, we denote

$$h_K = \text{diameter of the smallest sphere (or circle) containing } K,$$

and

$$\rho_K = \text{diameter of the largest sphere (or circle) inscribed in } K.$$

Similarly, $h_{\hat{K}}$ and $\rho_{\hat{K}}$ are used for the reference element $\hat{K}$.

Noting that $\det(B_K) = \mathrm{vol}(K)/\mathrm{vol}(\hat{K})$, we easily have

$$C_1\rho_K^d \le |\det(B_K)| \le C_2 h_K^d, \tag{2.22}$$

where the positive constants C_1 and C_2 are independent of h_K and ρ_K.

Lemma 2.5. *The following estimates hold*

$$||B_K^{-1}|| \le \frac{h_{\hat{K}}}{\rho_K}, \quad ||B_K|| \le \frac{h_K}{\rho_{\hat{K}}}.$$

Proof. By definition, we have

$$||B_K^{-1}|| = \frac{1}{\rho_K} \sup_{|\xi|=\rho_K} |B_K^{-1}\xi|. \tag{2.23}$$

For any ξ satisfying $|\xi| = \rho_K$, we can always find two points $x, y \in K$ such that $x - y = \xi$. Note that $|B_K^{-1}\xi| = |B_K^{-1}(x-y)| = |\hat{x} - \hat{y}| \le h_{\hat{K}}$, substituting which into (2.23) completes the proof of the first inequality. The other one can be proved in a similar way. □

Lemma 2.6 ([78, Theorem 3.1.1]). *There exists a constant $C(\hat{K})$ such that*

$$\inf_{\hat{p}\in P_k(\hat{K})} ||\hat{v} + \hat{p}||_{k+1,p,\hat{K}} \le C(\hat{K})|\hat{v}|_{k+1,p,\hat{K}} \quad \forall\, \hat{v} \in W^{k+1,p}(\hat{K}).$$

With the above preparations, we can prove the following interpolation error estimates.

Theorem 2.2 ([78, Theorem 3.1.5]). *Let $(\hat{K}, \hat{P}_K, \hat{\Sigma}_K)$ be a finite element. If*

$$
\begin{aligned}
&W^{k+1,p}(\hat{K}) \hookrightarrow C^s(\hat{K}),\\
&W^{k+1,p}(\hat{K}) \hookrightarrow W^{m,q}(\hat{K}),\\
&P_k(\hat{K}) \subset \hat{P}_K \subset W^{m,q}(\hat{K}),
\end{aligned}
$$

hold true for integers $m, k \geq 0$, and some numbers $p, q \in [1, \infty]$, where s denotes the greatest order of derivatives occurring in the degrees of freedom set $\hat{\Sigma}_K$, then there exists a constant C, depending on $\hat{K}$, $\hat{P}_K$ and $\hat{\Sigma}_K$, such that

$$|v - \Pi_K^k v|_{m,q,K} \leq C|det(B_K)|^{1/q-1/p} \frac{h_K^{k+1}}{\rho_K^m} |v|_{k+1,p,K} \quad \forall\ v \in W^{k+1,p}(K).$$

Proof. Noting that the shape functions $\hat{\phi}_i$ in $\hat{K}$ are given by $\hat{\phi}_i = \phi_i \circ F_K$, we obtain

$$\widehat{\Pi_K^k v} = \Pi_h^k v \circ F_K = \sum_i v(a_i)(\phi_i \circ F_K) = \sum_i v(F_K(\hat{a}_i))\hat{\phi}_i = \Pi_{\hat{K}}^k \hat{v},$$

from which and Lemmas 2.4–2.6, we have

$$
\begin{aligned}
|v - \Pi_K^k v|_{m,q,K} &\leq C||B_K^{-1}||^m |det(B_K)|^{1/q} |\hat{v} - \widehat{\Pi_K^k v}|_{m,q,\hat{K}} \leq \frac{C}{\rho_K^m} |det(B_K)|^{1/q} |\hat{v} - \Pi_{\hat{K}}^k \hat{v}|_{m,q,\hat{K}}\\
&\leq \frac{C}{\rho_K^m} |det(B_K)|^{1/q} ||I - \Pi_{\hat{K}}^k||_{\mathscr{L}(W^{k+1,p};W^{m,q})} \inf_{\hat{p} \in P_k} ||\hat{v} + \hat{p}||_{k+1,p,\hat{K}}\\
&\leq \frac{C}{\rho_K^m} |det(B_K)|^{1/q} C|\hat{v}|_{k+1,p,\hat{K}}\\
&\leq \frac{C}{\rho_K^m} |det(B_K)|^{1/q} ||B_K||^{k+1} |det(B_K)|^{-1/p} |v|_{k+1,p,K}\\
&\leq \frac{C}{\rho_K^m} h_K^{k+1} |det(B_K)|^{1/q-1/p} |v|_{k+1,p,K},
\end{aligned}
$$

which completes the proof. □

The parameter ρ_K can be eliminated if the finite element mesh is regular.

Definition 2.20. A triangulation T_h of Ω is called *regular* when $h = \max_{K \in T_h} h_K$ approaches zero, if there exists a constant $\sigma \geq 1$, independent of h, such that

$$\frac{h_K}{\rho_K} \leq \sigma \quad \text{for any } K \in T_h.$$

Under the regularity assumption, from Theorem 2.2 and (2.22), we can obtain the following interpolation error estimate over a physical domain Ω.

Theorem 2.3. *In addition to the assumptions of Theorem 2.2, if the mesh of Ω is regular, then*

$$|v - \Pi_h^k v|_{m,q,\Omega} \leq C h^{d(1/q-1/p)+k+1-m} |v|_{k+1,p,\Omega} \quad \forall\, v \in W^{k+1,p}(\Omega).$$

2.4 Finite Element Analysis for Elliptic Problems

In this section, we will present some basic finite element error analysis techniques developed for elliptic problems. For simplicity, we limit our discussion to conforming finite elements. More general discussions can be found in classic books such as [51, 78, 243].

2.4.1 Abstract Convergence Theory

Consider an abstract variational problem: Find $u \in V$ such that

$$A(u,v) = F(v) \quad \forall\, v \in V, \tag{2.24}$$

where V denotes a real Hilbert space and $F \in V'$. Here V' denotes the dual space of V. The following famous Lax-Milgram lemma justifies the existence and uniqueness of the solution for this problem.

Theorem 2.4 (Lax-Milgram lemma [78, p. 8]). *Let V be a real Hilbert space with norm $||\cdot||_V$, $A(\cdot,\cdot)$ be a bilinear form from $V \times V$ to R, and $F(\cdot)$ be a linear continuous functional from V to R. Furthermore, suppose that $A(\cdot,\cdot)$ is bounded:*

$$\exists \beta > 0 \text{ such that } |A(w,v)| \leq \beta ||w||_V ||v||_V \quad \text{for all } w, v \in V,$$

and coercive:

$$\exists \alpha > 0 \text{ such that } |A(v,v)| \geq \alpha ||v||_V^2 \quad \text{for all } v \in V.$$

Then, there exists a unique solution $u \in V$ to (2.24) and

$$||u||_V \leq \frac{1}{\alpha} ||F||_{V'}.$$

Assume that a family of finite dimensional subspaces V_h is constructed to approximate the infinite dimensional space V, i.e.,

$$\inf_{v_h \in V_h} ||v - v_h||_V \to 0 \text{ as } h \to 0, \quad \text{for all } v \in V.$$

The standard Galerkin FEM for solving (2.24) is: Find $u_h \in V_h$ such that

$$A(u_h, v_h) = F(v_h) \quad \forall\, v_h \in V_h. \tag{2.25}$$

From (2.24) and (2.25), we obtain the Galerkin orthogonality relation:

$$A(u - u_h, v_h) = 0 \quad \forall\, v_h \in V_h. \tag{2.26}$$

Combining (2.26) with the coercivity and continuity of $A(\cdot,\cdot)$, we have

$$\begin{aligned}\alpha ||u - u_h||_V^2 &\le A(u - u_h, u - u_h) = A(u - u_h, u - v_h)\\ &\le \beta ||u - u_h||_V ||u - v_h||_V,\end{aligned}$$

which leads to the following stability and convergence result.

Theorem 2.5 (Céa lemma). *Under the assumption of Theorem 2.4, there exists a unique solution u_h to (2.25) and*

$$||u_h||_V \le \frac{1}{\alpha} ||F||_{V'}.$$

Furthermore, if u denotes the solution to (2.24), then

$$||u - u_h||_V \le \frac{\beta}{\alpha} \inf_{v_h \in V_h} ||u - v_h||_V, \tag{2.27}$$

i.e., u_h converges to u as $h \to 0$.

2.4.2 Error Estimate for an Elliptic Problem

Here we demonstrate how the abstract convergence theory presented in last section can be used for a specific problem. Without loss of generality, let us consider the following elliptic boundary value problem

$$-\triangle u + u = f \quad \text{in } \Omega, \tag{2.28}$$

$$u = 0 \quad \text{on } \partial\Omega. \tag{2.29}$$

Multiplying (2.28) by a test function $v \in H_0^1(\Omega)$ and using the Green's formula

$$-\int_\Omega \triangle u v dx = -\int_{\partial\Omega} \frac{\partial u}{\partial n} v ds + \int_\Omega \sum_{i=1}^{d} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i}, \quad \forall\, u \in H^2(\Omega), v \in H^1(\Omega), \tag{2.30}$$

where $\frac{\partial}{\partial n} = \sum_{i=1}^{d} n_i \frac{\partial}{\partial x_i}$ is the normal derivative operator, we can obtain an equivalent variational problem: Find $u \in H_0^1(\Omega)$ such that

$$A(u,v) \equiv (\nabla u, \nabla v) + (u,v) = (f,v), \quad \forall\, v \in H_0^1(\Omega). \tag{2.31}$$

Application of the Cauchy-Schwarz inequality shows that

$$|A(u,v)| = |(\nabla u, \nabla v) + (u,v)| \le ||\nabla u||_0 ||\nabla v||_0 + ||u||_0 ||v||_0 \le 2||u||_1 ||v||_1,$$

which means that $A(\cdot,\cdot)$ is continuous on $H_0^1(\Omega) \times H_0^1(\Omega)$.

On the other hand, we easily obtain

$$A(v,v) = ||\nabla v||_0^2 + ||v||_0^2 = ||v||_1^2,$$

which proves the coercivity of $A(\cdot,\cdot)$. Therefore, by the Lax-Milgram lemma, the variational problem (2.31) has a unique solution $u \in H_0^1(\Omega)$.

To solve (2.31) by the finite element method, we construct a finite dimensional subspace V_h of $H_0^1(\Omega)$:

$$V_h = \{v_h \in H_0^1(\Omega);\ \forall\ K \in T_h, v_h|_K \in P_k(K)\}, \tag{2.32}$$

where T_h is a regular family of triangulations of Ω.

The finite element approximation $u_h \in V_h$ of (2.31) is: Find $u_h \in V_h$ such that

$$(\nabla u_h, \nabla v_h) + (u_h, v_h) = (f, v_h), \quad \forall\, v_h \in V_h.$$

For the elliptic problem (2.28) and (2.29), the following optimal error estimates hold true in both H^1 and L^2 norms.

Theorem 2.6. *Let Ω be a polygonal domain of $R^d, d = 2,3$, with Lipschitz boundary, and T_h be a regular family of triangulations of Ω. Let V_h be defined in (2.32). If the exact solution $u \in H^s(\Omega) \cap H_0^1(\Omega), s \ge 2$, the error estimate holds*

$$||u - u_h||_1 \le Ch^l ||u||_{l+1}, \quad l = \min(k, s-1), k \ge 1.$$

Suppose, furthermore, for each $e \in L^2(\Omega)$, the solution w of the adjoint problem (see (2.35) below) of (2.28) belongs to $H^2(\Omega)$ and satisfies

$$||w||_2 \le C||e||_0, \quad \forall\, e \in L^2(\Omega). \tag{2.33}$$

Then we have

$$||u - u_h||_0 \le Ch^{l+1} ||u||_{l+1}, \quad l = \min(k, s-1), k \ge 1.$$

Proof. By the Céa lemma, for any $u \in H^s(\Omega) \cup H_0^1(\Omega), s \geq 2$, we have

$$||u - u_h||_1 \leq 2 \inf_{v_h \in V_h} ||u - v_h||_1 \leq 2||u - \Pi_h^k u||_1 \leq Ch^l ||u||_{l+1},$$

where $l = \min(k, s-1)$, and Π_h^k is the finite element interpolation operator defined in Sect. 2.3.

To derive error estimates in the L^2-norm, we need to use the so-called Aubin-Nitsche technique (also called duality argument). Denote error $e = u - u_h$, and let w be the solution of the adjoint problem of (2.28):

$$-\triangle w + w = e \text{ in } \Omega, \quad w = 0 \text{ on } \partial\Omega, \tag{2.34}$$

whose variational formulation is: Find $w \in H_0^1(\Omega)$ such that

$$A(v, w) = (e, v) \quad \forall\ v \in H_0^1(\Omega). \tag{2.35}$$

Hence, by the Galerkin orthogonality relation, we have

$$\begin{aligned} ||u - u_h||_0^2 &= (e, u - u_h) = A(u - u_h, w) = A(u - u_h, w - \Pi_h^k w) \\ &\leq 2||u - u_h||_1 ||w - \Pi_h^k w||_1 \leq Ch^l ||u||_{l+1} \cdot h||w||_2, \end{aligned} \tag{2.36}$$

which, along with the bound (2.33), leads to

$$||u - u_h||_0 \leq Ch^{l+1} ||u||_{l+1},$$

which is optimal in the L^2-norm. This concludes the proof. □

Using more sophisticated weighted-norm technique, error estimates in the L^∞ norm can be proved [78, p. 165].

Theorem 2.7. *Under the assumption of* $u \in H_0^1(\Omega) \cap W^{k+1,\infty}(\Omega), k \geq 1$, *we have*

$$||u - u_h||_{\infty,\Omega} + h||\nabla(u - u_h)||_{\infty,\Omega} \leq Ch^2 |\ln h| |u|_{2,\infty,\Omega}, \ \textit{for } k = 1,$$

and

$$||u - u_h||_{\infty,\Omega} + h||\nabla(u - u_h)||_{\infty,\Omega} \leq Ch^{k+1} |u|_{k+1,\infty,\Omega}, \ \textit{for } k \geq 2.$$

2.5 Finite Element Programming for Elliptic Problems

In this section, we introduce the basic procedures for programming a finite element method used for solving the second-order elliptic boundary value problems. We want to remark that finite element programming is quite a sophisticated task and can be written in a stand-alone book. Readers can find more advanced and sophisticated programming algorithms in books devoted to this subject (e.g., [21, 62, 97, 107, 112, 158, 174, 240]). For example, [21] presents the package PLTMG, which solves elliptic problems using adaptive FEM in MATLAB; [174] introduces Diffpack, a sophisticated toolbox for solving PDEs in C++; [252] elaborates the adaptive finite element software ALBERTA written in ANSI-C; [107] introduces an open-source MATLAB package IFISS, which can be used to solve convection-diffusion, Stokes, and Navier-Stokes equations. Demkowicz [97] presents a self-contained hp-finite element package for solving elliptic problems and time-harmonic Maxwell's equations.

In the rest of this section, we detail the important steps for programming a finite element method in MATLAB to solve the elliptic problem:

$$-\triangle u + u = f \quad \text{in } \Omega = (0,1)^2, \tag{2.37}$$

$$u = 0 \quad \text{on } \partial\Omega, \tag{2.38}$$

by using Q_1 element. The material of this section is modified from our previous book [187, Chap. 7].

2.5.1 *The Basic Steps*

2.5.1.1 FEM Mesh Generation

As we mentioned earlier, we need to subdivide the physical domain Ω into small elements. For simplicity, here we generate a uniform rectangular mesh T_h for Ω. The mesh structure can be clearly described by four variables and four arrays. More specifically, we use nx and ny for the total number of elements in the x- and y-direction, respectively; ne and np for the total number of elements and the total number of nodes in T_h, respectively. We use the 1-D array $x(1:np)$ and $y(1:np)$ to represent the nodal x and y coordinates, respectively. A 2-D array conn(1:ne,1:4) is used to store the connectivity matrix for the mesh, i.e., $conn(i, 1:4)$ specifies the four node numbers of element i. A 2-D array $gbc(1:np, 1:2)$ is used to store the indicators for Dirichlet nodes in $gbc(1:np, 1)$, and the corresponding boundary values in $gbc(1:np, 2)$.

The rectangular mesh T_h can be generated using the MATLAB code *getQ1mesh.m*, which is shown below.

```
% Generate Q1 mesh on [xlow, xhigh]x[ylow, yhigh]
% nx,ny: number of elements in each direction
% x,y: 1-D array for nodal coordinates
% conn(1:ne,1:4): connectivity matrix
% ne, np: total numbers of elements, nodes generated
% gbc(1:np, 1:2): store Dirichlet node labels and values
```

```
function[x,y,conn,ne,np,gbc] ...
      = getQ1mesh(xlow,xhigh,ylow,yhigh,nx,ny)

ne = nx*ny;
np = (nx+1)*(ny+1);

% create nodal coordinates
dx=(xhigh - xlow)/nx;  dy=(yhigh - ylow)/ny;
for i = 1:(nx+1)
   for j=1:(ny+1)
      x((ny+1)*(i-1)+j) = dx*(i-1);
      y((ny+1)*(i-1)+j) = dy*(j-1);
   end
end

% form the connectivity matrix:
%  start from low-left corner countclockwisely.
for j=1:nx
   for i=1:ny
      ele = (j-1)*ny + i;
      conn(ele,1) = ele + (j-1);
      conn(ele,2) = conn(ele,1) + ny + 1;
      conn(ele,3) = conn(ele,2) + 1;
      conn(ele,4) = conn(ele,1) + 1;
   end
end

% pick out Dirichlet BC nodes
for i=1:np
    if (abs(x(i) - xlow) < 0.1*dx ...
          | abs(x(i) - xhigh) < 0.1*dx)
        gbc(i,1)=1;    % find one BC node
    elseif (abs(y(i) - ylow) < 0.1*dy ...
          | abs(y(i) - yhigh) < 0.1*dy)
        gbc(i,1)=1;    % find one BC node
    end
end
```

A simple 4 × 4 rectangular mesh generated with this code is shown in Fig. 2.1, where the nodal numbers and element numbers are provided. Note that the nodes of each element are oriented counterclockwisely. For example, the connectivity matrices for elements 1 and 2 are given by:

$$conn(1, 1:4) = 1, 6, 7, 2, \quad conn(2, 1:4) = 2, 7, 8, 3.$$

2.5.1.2 Forming FEM Equations

The finite element method for solving (2.37) and (2.38) is: Find $u_h \in \mathbf{v}_h \subset H_0^1(\Omega)$ such that

$$(\nabla u_h, \nabla \phi_h) + (u_h, \phi_h) = (f, \phi_h) \quad \forall\ \phi_h \in \mathbf{v}_h, \tag{2.39}$$

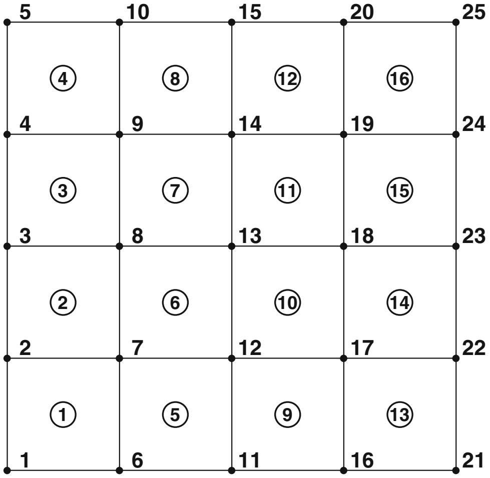

Fig. 2.1 An exemplary Q_1 element: node and element labelings

where the Q_1 finite element space is defined as

$$\mathbf{V}_h = \{v \in H_0^1(\Omega);\ \forall\ K \in T_h, v|_K \in Q_1(K) \text{ and } v|_{K\cap\partial\Omega} = 0\}.$$

On each rectangular element K, the exact solution u is approximated by

$$u_h^K(x,y) = \sum_{j=1}^{4} u_j^K \psi_j^K(x,y), \tag{2.40}$$

which leads to a 4×4 element coefficient matrix of (2.39) with entries

$$A_{ij} \equiv \int_K \nabla\psi_j^K \cdot \nabla\psi_i^K dxdy + \int_K \psi_j^K \psi_i^K dxdy,\ \ i,j = 1,\cdots,4. \tag{2.41}$$

2.5.1.3 Calculation of Element Matrices

The calculation of A_{ij} is often carried out on a reference rectangle $\hat{K}$ with vertices

$$(\xi_1,\eta_1) = (-1,-1),\ (\xi_2,\eta_2) = (1,-1),\ (\xi_3,\eta_3) = (1,1),\ (\xi_4,\eta_4) = (-1,1),$$

whose corresponding shape functions

$$\hat{\psi}_i(\xi,\eta) = \frac{1}{4}(1+\xi_i\xi)(1+\eta_i\eta), \quad i = 1,\cdots,4. \tag{2.42}$$

The mapping between an arbitrary rectangular element with vertices $(x_i, y_i), 1 \leq i \leq 4$, and the reference element is given by

$$x = \sum_{j=1}^{4} x_j \hat{\psi}_j(\xi, \eta), \quad y = \sum_{j=1}^{4} y_j \hat{\psi}_j(\xi, \eta). \tag{2.43}$$

The basis function $\psi_j(x, y)$ on a general rectangle is defined as

$$\psi_j(x, y) = \hat{\psi}_j(\xi(x, y), \eta(x, y)),$$

from which we have

$$\frac{\partial \hat{\psi}_j}{\partial \xi} = \frac{\partial \psi_j}{\partial x}\frac{\partial x}{\partial \xi} + \frac{\partial \psi_j}{\partial y}\frac{\partial y}{\partial \xi}, \tag{2.44}$$

$$\frac{\partial \hat{\psi}_j}{\partial \eta} = \frac{\partial \psi_j}{\partial x}\frac{\partial x}{\partial \eta} + \frac{\partial \psi_j}{\partial y}\frac{\partial y}{\partial \eta}, \tag{2.45}$$

i.e.,

$$J^T \cdot \nabla \psi_j = \begin{bmatrix} \frac{\partial \hat{\psi}_j}{\partial \xi} \\ \frac{\partial \hat{\psi}_j}{\partial \eta} \end{bmatrix}, \quad j = 1, 2, 3,$$

where we denote the gradient

$$\nabla \psi_j = \begin{bmatrix} \frac{\partial \psi_j}{\partial x} \\ \frac{\partial \psi_j}{\partial y} \end{bmatrix}$$

and the Jacobi matrix J of the mapping as

$$J \equiv \begin{bmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \xi} & \frac{\partial y}{\partial \eta} \end{bmatrix}.$$

In the above, J^T denotes the transpose of J.

From (2.44) and (2.45), we see that

$$\nabla \psi_j = (J^T)^{-1} \begin{bmatrix} \frac{\partial \hat{\psi}_j}{\partial \xi} \\ \frac{\partial \hat{\psi}_j}{\partial \eta} \end{bmatrix} = \frac{1}{det(J)} \begin{bmatrix} \frac{\partial y}{\partial \eta} & -\frac{\partial y}{\partial \xi} \\ -\frac{\partial x}{\partial \eta} & \frac{\partial x}{\partial \xi} \end{bmatrix} \begin{bmatrix} \frac{\partial \hat{\psi}_j}{\partial \xi} \\ \frac{\partial \hat{\psi}_j}{\partial \eta} \end{bmatrix} \tag{2.46}$$

$$= \frac{1}{det(J)} \begin{bmatrix} (\sum_{j=1}^4 y_j \frac{\partial \hat{\psi}_j}{\partial \eta}) \frac{\partial \hat{\psi}_j}{\partial \xi} - (\sum_{j=1}^4 y_j \frac{\partial \hat{\psi}_j}{\partial \xi}) \frac{\partial \hat{\psi}_j}{\partial \eta} \\ -(\sum_{j=1}^4 x_j \frac{\partial \hat{\psi}_j}{\partial \eta}) \frac{\partial \hat{\psi}_j}{\partial \xi} + (\sum_{j=1}^4 x_j \frac{\partial \hat{\psi}_j}{\partial \xi}) \frac{\partial \hat{\psi}_j}{\partial \eta} \end{bmatrix}. \tag{2.47}$$

Since it is too complicated to evaluate A_{ij} analytically, below we use Gaussian quadrature over the reference rectangle to approximate A_{ij}. For example,

$$\int_K G(x,y)dxdy = \int_{\hat{K}} G(\hat{x},\hat{y})|J|d\xi d\eta$$
$$\approx \sum_{m=1}^{N}(\sum_{n=1}^{N} G(\xi_m,\eta_n)|J|\omega_n)\omega_m,$$

where ξ_m and η_m, and ω_m and ω_n are the quadrature points and weights in the ξ and η directions, respectively. In our implementation, we use the second-order Gaussian quadrature rule, in which case

$$\xi_1 = -\frac{1}{\sqrt{3}}, \quad \xi_2 = \frac{1}{\sqrt{3}}, \quad \omega_1 = \omega_2 = 1.$$

The right-hand side vector (f,ϕ_i) is approximated as

$$(f,\phi_i) \approx (\frac{1}{4}\sum_{j=1}^{4} f_j,\phi_i), \quad i = 1,2,3,4.$$

Below is our MATLAB code *elemA.m*, which is used to calculate the element matrix and right-hand side vector.

```
function [ke,rhse] = elemA(conn,x,y,gauss,rhs,e);

% Q1 elementary stiffness matrix
ke = zeros(4,4);
rhse=zeros(4,1);
one = ones(1,4);
psiJ = [-1, +1, +1, -1]; etaJ = [-1, -1, +1, +1];

% coordinates of each element 'e'
for j=1:4
   je = conn(e,j);   % get the node label
   xe(j) = x(je); ye(j) = y(je); % get the coordinates
end

for i=1:2  % loop over gauss points in eta
  for j=1:2   % loop over gauss points in psi
     eta = gauss(i);  psi = gauss(j);
  % construct shape functions: starting at low-left corner
     NJ=0.25*(one + psi*psiJ).*(one + eta*etaJ);
  % derivatives of shape functions in reference coordinate
     NJpsi = 0.25*psiJ.*(one + eta*etaJ);    % 1x4 array
     NJeta = 0.25*etaJ.*(one + psi*psiJ);    % 1x4 array
     % derivatives of x and y wrt psi and eta
     xpsi = NJpsi*xe';  ypsi = NJpsi*ye';
     xeta = NJeta*xe';  yeta = NJeta*ye';
     Jinv = [yeta, -xeta; -ypsi, xpsi];      % 2x2 array
     jcob = xpsi*yeta - xeta*ypsi;
   % derivatives of shape functions in original coordinate
     NJdpsieta = [NJpsi; NJeta];             % 2x4 array
     NJdxy = Jinv*NJdpsieta;                 % 2x4 array
     % assemble element stiffness matrix ke: 4x4 array
     ke = ke + (NJdxy(1,:))'*(NJdxy(1,:))/jcob ...
             + (NJdxy(2,:))'*(NJdxy(2,:))/jcob ...
             + NJ(1,:)'*NJ(1,:)*jcob;
     rhse = rhse + rhs*NJ'*jcob;
   end
end
```

2.5.1.4 Assembly of Global Matrix

To obtain the global coefficient matrix, we need to assembly the contributions from each element coefficient matrix, i.e., we need to loop through all elements in the mesh, find the corresponding global nodal label for each local node, and put them in the right locations of the global coefficient matrix. A pseudo code is listed below:

```
for n=1:NE     % loop through all elements
  % computer element coefficient matrix
   for i=1:4    % loop through all nodes
      i1 = conn(n,i)
      for j=1:4
         j1=conn(n,j)
         Ag(i1,j1)=Ag(i1,j1)+Aloc(i,j)
      End
   End
End
```

After the assembly process, we need to impose the given Dirichlet boundary conditions. Suppose that after assembly we obtain a global linear system

$$A\mathbf{u} = \mathbf{b}. \tag{2.48}$$

Assume that we have to impose a Dirichlet boundary condition $u = u(x_k, y_k)$ at the k-th global node. A simple way to impose this boundary condition is as follows: First, replace each entry b_i of $\mathbf{b}$ by $b_i - A_{ik}u_k$; Then reset all entries in the k-th row and k-th column of A to 0, and the diagonal entry A_{kk} to 1; Finally, replace the k-th entry b_k of $\mathbf{b}$ by u_k.

This algorithm can be implemented by a pseudo code listed below:

```
for k=1:NG     % loop through all global nodes
    % identify all Dirichlet nodes
    If (gbc(k,1) = 1) then
       for i=1:NG
          b(i) = b(i) - Ag(i,k)*gbc(k,2)
          Ag(i,k) = 0
          Ag(k,i) = 0
       End
       Ag(k,k) = 1
       b(k) = gbc(k,2)
    End
End
```

2.5.2 A MATLAB Code for Q_1 Element

A driver function *ellip_Q1.m* developed for implementing the Q_1 element for solving the elliptic equation (2.37) is shown below:

```
%-----------------------------------------------------
% 2D Q1 FEM for solving
%     -Lap*u + u = f(x,y)
% on a rectangular domain (xlow,xhigh)x(ylow,yhigh)
% with Dirichlet BC condition: u=g  on boundary
%-----------------------------------------------------

clear all;

% readers can change the parameters to
% reset their own rectangualr domain and the mesh size
xlow = 0.0; xhigh = 1.0;
ylow = 0.0; yhigh = 1.0;
nx=20; ny=20;

% Gaussian quadrature points
gauss = [-1/sqrt(3), 1/sqrt(3)];

% generate a Q1 mesh
[x,y,conn,ne,np,gbc] = ...
    getQ1mesh(xlow,xhigh,ylow,yhigh,nx,ny);

% specify an exact solution and use it for Dirichlet BC
for i=1:np
```

```
    uex(i)=sin(pi*x(i)).*cos(pi*y(i));
    if(gbc(i,1)==1)         % find a Dirichlet node
       gbc(i,2) = uex(i);  % assign the BC value
    end
end

% initialize the coefficient matrix and the rhs vector
Ag = zeros(np); bg=zeros(np,1);

nloc = 4;   % number of nodes per element
for ie = 1:ne    % loop over all elements
   rhs= (feval(@SRC,x(conn(ie,1)),y(conn(ie,1)))...
       + feval(@SRC,x(conn(ie,2)),y(conn(ie,2)))...
       + feval(@SRC,x(conn(ie,3)),y(conn(ie,3)))...
       + feval(@SRC,x(conn(ie,4)),y(conn(ie,4))))/nloc;

   [Aloc,rhse] = elemA(conn,x,y,gauss,rhs,ie);
   % assemble local matrices into the global matrix
   for i=1:nloc;
      irow = conn(ie,i);   % global row index
      bg(irow)=bg(irow) + rhse(i);
      for j=1:nloc;
         icol = conn(ie,j);  %global column index
         Ag(irow, icol) = Ag(irow, icol) + Aloc(i,j);
      end;
   end;
end;

% impose the Dirichlet BC
for m=1:np
 if(gbc(m,1)==1)
   for i=1:np
    bg(i) = bg(i) - Ag(i,m) * gbc(m,2);
    Ag(i,m) = 0; Ag(m,i) = 0;
   end
   Ag(m,m) = 1.0; bg(m) = gbc(m,2);
  end
end

%solve the equation
ufem = Ag\bg;

% display the max pointwise error
disp('Max error='), max(ufem-uex'),
```

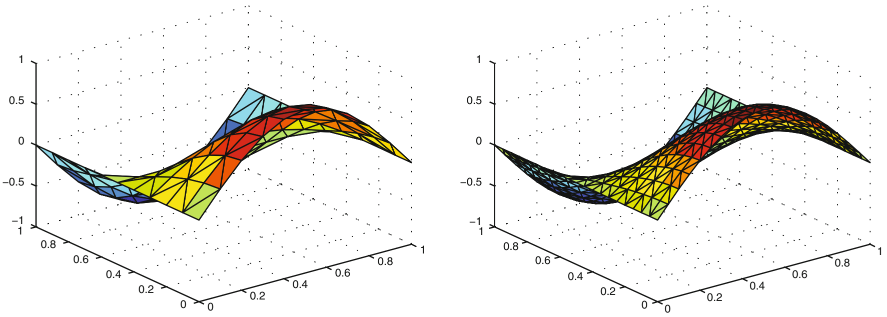

Fig. 2.2 Numerical solutions obtained on $nx = ny = 10$ (*left*), and $nx = ny = 20$ (*right*) grids

```
% plot the FEM solution
tri = delaunay(x,y);
trisurf(tri,x,y,ufem);
```

In this code, we solve the Eq. (2.37) with non-homogenous Dirichlet boundary condition $u = g$ with the exact solution

$$u = \sin \pi x \cos \pi y,$$

which leads to $f = (2\pi^2 + 1)u$.

The problem can be solved with several uniformly refined grids by just changing nx and ny in driver function *ellip_Q1.m*. For example, maximum pointwise errors obtained with $nx = ny = 10, 20, 40$ grids are $0.0084, 0.0021, 5.2583e{-}004$, respectively, which clearly shows the $O(h^2)$ convergence rate in the L^∞-norm. Exemplary numerical solutions obtained with $nx = ny = 10$ and 20 are shown in Fig. 2.2.