Selected Topics in Finite Element Methods

Zhiming Chen

Haijun Wu

Institute of Computational Mathematics, Chinese Academy of Sciences, Beijing, 100080, China.

E-mail address: zmchen@lsec.cc.ac.cn

Department of Mathematics, Nanjing University, Nanjing, Jiangsu, 210093, China.

E-mail address: hjw@nju.edu.cn

Contents

CHAPTER 0

A brief introduction to finite element methods

1. Two-point boundary value problem and weak formulation

Consider the two-point boundary value problem: Given a constant $a \geq 0$ and a function $f(x)$, find $u(x)$ such that

$$\begin{aligned} -u'' + au &= f(x), \quad 0 < x < 1, \\ u(0) = 0, &\quad u'(1) = 0. \end{aligned} \tag{0.1}$$

If u is the solution to (0.1) and $v(x)$ is any (sufficiently regular) function such that $v(0) = 0$, then integration by parts yields

$$\begin{aligned} &\int_0^1 -u''v \,\mathrm{d}x + \int_0^1 auv \,\mathrm{d}x \\ &\quad = -u'(1)v(1) + u'(0)v(0) + \int_0^1 u'(x)v'(x)\,\mathrm{d}x + \int_0^1 auv\,\mathrm{d}x \\ &\quad = \int_0^1 fv\,\mathrm{d}x. \end{aligned}$$

Let us introduce the bilinear form

$$A(u,v) = \int_0^1 (u'v' + auv)\,\mathrm{d}x,$$

and define

$$V = \left\{ v \in L^2([0,1]) : \; A(v,v) < \infty \text{ and } v(0) = 0 \right\}.$$

Then we can say that the solution u to (0.1) is characterized by

$$u \in V \qquad \text{such that} \qquad A(u,v) = \int_0^1 f(x)v(x)\,\mathrm{d}x \quad \forall v \in V. \tag{0.2}$$

which is called the *variational* or *weak* formulation of (0.1).

We remark that the boundary condition $u(0) = 0$ is called *essential* as it appears in the variational formulation explicitly, i.e., in the definition of V. This type of boundary condition is also called “Dirichlet” boundary condition. The boundary condition $u'(0) = 0$ is called *natural* as it is incorporated implicitly. This type of boundary condition is often referred to by the name “Neumann”.

THEOREM 1.1. *Suppose $f \in C^0([0,1])$ and $u \in C^2([0,1])$ satisfies (0.2). Then u solves (0.1).*

PROOF. Let $v \in V \bigcap C^1([0,1])$. Then integration by parts gives

$$\int_0^1 fv\,\mathrm{d}x = A(u,v) = \int_0^1 -u''v\,\mathrm{d}x + \int_0^1 auv\,\mathrm{d}x + u'(1)v(1). \tag{0.3}$$

Thus, $\int_0^1 (f + u'' - au)v \, dx = 0$ for all $v \in V \bigcap C^1([0,1])$ such that $v(1) = 0$. Let $w = f + u'' - au \in C^0([0,1])$. If $w \not\equiv 0$, then $w(x)$ is of one sign in some interval $[b,c] \subset [0,1]$, with $b < c$. Choose $v(x) = (x-b)^2(x-c)^2$ in $[b,c]$ and $v \equiv 0$ outside $[b,c]$. But then $\int_0^1 wv \, dx \neq 0$ which is a contradiction. Thus $-u'' + au = f$. Now apply (0.3) with $v(x) = x$ to find $u'(1) = 0$. So u solves (0.1). □

2. Piecewise polynomial spaces – the finite element method

2.1. Meshes. Let $\mathcal{M}_h$ be a partition of $[0,1]$:

$$0 = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = 1.$$

The points $\{x_i\}$ are called *nodes*. Let $h_i = x_i - x_{i-1}$ be the length of the i-th subinterval $[x_{i-1}, x_i]$. Define $h = \max_{1 \le i \le n} h_i$.

2.2. Finite element spaces. We shall approximate the solution $u(x)$ by using the continuous piecewise linear functions over $\mathcal{M}_h$. Introduce the linear space of functions

$$\begin{aligned} V_h = \big\{ v \in C^0([0,1]) : v(0) = 0, \\ v|_{[x_{i-1},x_i]} \text{ is a linear polynomial, } i = 1, \cdots, n \big\}. \end{aligned} \tag{0.4}$$

It is clear that $V_h \subset V$.

2.3. The finite element method. The finite element discretization of (0.2) reads as:

$$\text{Find } u_h \in V_h \quad \text{such that} \quad A(u_h, v_h) = \int_0^1 f(x) v_h(x) \, dx \quad \forall v_h \in V_h. \tag{0.5}$$

2.4. A nodal basis. For $i = 1, \cdots, n$, define $\phi_i \in V_h$ by the requirement that $\phi_i(x_j) = \delta_{ij}$ = the Kronecker delta, as shown in Fig. 1:

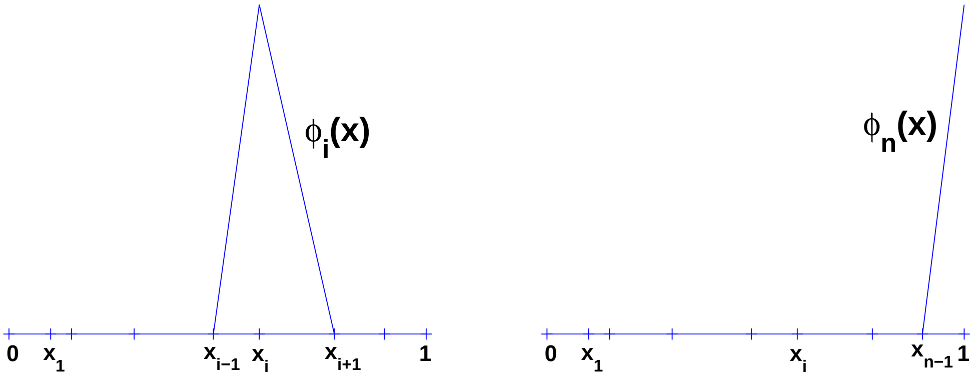

Figure 1. piecewise linear basis function ϕ_i.

$$\phi_i = \begin{cases} \frac{x-x_{i-1}}{h_i}, & x_{i-1} \le x \le x_i, \\ \frac{x_{i+1}-x}{h_{i+1}}, & x_i < x \le x_{i+1}, \\ 0, & x < x_{i-1} \text{ or } x > x_{i+1}, \end{cases} \qquad 1 \le i \le n-1.$$

$$\phi_n = \begin{cases} \frac{x-x_{n-1}}{h_n}, & x_{n-1} \le x \le 1, \\ 0, & x < x_{n-1}. \end{cases}$$

For any $v_h \in V_h$, let v_i be the value of v_h at the node x_i, i.e.,

$$v_i = v_h(x_i), \quad i = 1, 2, \cdots, n,$$

then

$$v_h = v_1\phi_1(x) + v_2\phi_2(x) + \cdots + v_n\phi_n(x).$$

2.5. The finite element equations. Let

$$u_h = u_1\phi_1 + u_2\phi_2 + \cdots + u_n\phi_n, \quad u_1, \cdots, u_n \in \mathbb{R},$$

where $u_i = u_h(x_i)$.

Let $v_h = \phi_i$, $i = 1, \cdots, n$ in (0.5), then we obtain an algebraic linear system in unknowns $u_1, u_2, \cdots, u_n$:

$$A(\phi_1, \phi_i)u_1 + A(\phi_2, \phi_i)u_2 + \cdots + A(\phi_n, \phi_i)u_n = \int_0^1 f(x)\phi_i \,\mathrm{d}x, \qquad i = 1, \cdots, n. \tag{0.6}$$

Denote by

$$k_{ij} = A(\phi_j, \phi_i) = \int_0^1 \phi_j'\phi_i' + a\phi_j\phi_i \,\mathrm{d}x, \quad f_i = \int_0^1 f(x)\phi_i \,\mathrm{d}x,$$

and

$$K = \big(k_{ij}\big)_{n\times n}, \quad F = \big(f_i\big)_{n\times 1}, \quad U = \big(u_i\big)_{n\times 1},$$

then (0.6) can be rewritten as:

$$KU = F \tag{0.7}$$

Here K is called the *stiffness* matrix.

It is clear that $A(\phi_j, \phi_i) = 0$ if x_i and x_j are not adjacent to each other. Therefore K is *sparse*.

We recall that the Simpson quadrature rule

$$\int_c^d \phi(x)\,\mathrm{d}x \simeq \frac{d-c}{6}\left[\phi(c) + 4\phi(\frac{c+d}{2}) + \phi(d)\right]$$

is accurate for polynomials of degree ≤ 3. To compute $A(\phi_j, \phi_i)$, we first calculate the following integrals over the subinterval $[x_{i-1}, x_i]$:

$$\begin{aligned}
&\int_{x_{i-1}}^{x_i} \phi_i'\phi_i' \,\mathrm{d}x = \int_{x_{i-1}}^{x_i} \frac{1}{h_i^2} \,\mathrm{d}x = \frac{1}{h_i},\\
&\int_{x_{i-1}}^{x_i} \phi_{i-1}'\phi_{i-1}' \,\mathrm{d}x = \frac{1}{h_i},\\
&\int_{x_{i-1}}^{x_i} \phi_i'\phi_{i-1}' \,\mathrm{d}x = \int_{x_{i-1}}^{x_i} -\frac{1}{h_i^2} \,\mathrm{d}x = -\frac{1}{h_i},\\
&\int_{x_{i-1}}^{x_i} \phi_i\phi_i \,\mathrm{d}x = \int_{x_{i-1}}^{x_i} \Big(\frac{x - x_{i-1}}{h_i}\Big)^2 \,\mathrm{d}x = \frac{h_i}{6}(1 + \frac{4}{4} + 0) = \frac{h_i}{3},\\
&\int_{x_{i-1}}^{x_i} \phi_{i-1}\phi_{i-1} \,\mathrm{d}x = \int_{x_{i-1}}^{x_i} \Big(\frac{x_i - x}{h_i}\Big)^2 \,\mathrm{d}x = \frac{h_i}{3},\\
&\int_{x_{i-1}}^{x_i} \phi_i\phi_{i-1} \,\mathrm{d}x = \int_{x_{i-1}}^{x_i} \Big(\frac{x - x_{i-1}}{h_i}\Big)\Big(\frac{x_i - x}{h_i}\Big) \,\mathrm{d}x = \frac{h_i}{6}(0 + \frac{4}{4} + 0) = \frac{h_i}{6}. \qquad (0.8)
\end{aligned}$$

Hence

$$\int_0^1 \phi_i'\phi_i' \,\mathrm{d}x = \begin{cases} \int_{x_{i-1}}^{x_i} \phi_i'\phi_i' \,\mathrm{d}x + \int_{x_i}^{x_{i+1}} \phi_i'\phi_i' \,\mathrm{d}x = \dfrac{1}{h_i} + \dfrac{1}{h_{i+1}} \\ \qquad\qquad\qquad\qquad\qquad i = 1, \cdots, n-1, \\ \dfrac{1}{h_n}, \quad i = n, \end{cases}$$

$$\int_0^1 \phi_i'\phi_{i-1}' \,\mathrm{d}x = \int_{x_{i-1}}^{x_i} \phi_i'\phi_{i-1}' \,\mathrm{d}x = -\frac{1}{h_i}, \quad i = 2, \cdots, n,$$

$$\int_0^1 \phi_i\phi_i \,\mathrm{d}x = \begin{cases} \dfrac{h_i + h_{i+1}}{3}, \quad i = 1, \cdots, n-1, \\ \dfrac{h_n}{3}, \quad i = n, \end{cases}$$

$$\int_0^1 \phi_i\phi_{i-1} \,\mathrm{d}x = \frac{h_i}{6}, \quad i = 2, \cdots, n.$$

Therefore

$$A(\phi_i, \phi_i) = \int_0^1 \phi_i'\phi_i' \,\mathrm{d}x + a\int_0^1 \phi_i\phi_i \,\mathrm{d}x = \begin{cases} \dfrac{1}{h_i} + \dfrac{1}{h_{i+1}} + \dfrac{a}{3}(h_i + h_{i+1}), \\ \qquad\qquad i = 1, \cdots, n-1, \\ \dfrac{1}{h_n} + \dfrac{a}{3}h_n, \quad i = n, \end{cases}$$

$$A(\phi_i, \phi_{i-1}) = A(\phi_{i-1}, \phi_i) = -\frac{1}{h_i} + \frac{a}{6}h_i, \quad i = 2, \cdots, n.$$

Combining the above equations and (0.6) yields

$$\begin{cases} \big[\frac{a(h_1+h_2)}{3} + \frac{1}{h_1} + \frac{1}{h_2}\big]u_1 + \big(\frac{ah_2}{6} - \frac{1}{h_2}\big)u_2 = f_1, \\ \big(\frac{ah_i}{6} - \frac{1}{h_i}\big)u_{i-1} + \big[\frac{a(h_i+h_{i+1})}{3} + \frac{1}{h_i} + \frac{1}{h_{i+1}}\big]u_i + \big(\frac{ah_{i+1}}{6} - \frac{1}{h_{i+1}}\big)u_{i+1} = f_i \\ \qquad\qquad\qquad\qquad\qquad i = 2, \cdots, n-1, \\ \big(\frac{ah_n}{6} - \frac{1}{h_n}\big)u_{n-1} + \big[\frac{ah_n}{3} + \frac{1}{h_n}\big]u_n = f_n. \end{cases}$$

2.6. The interpolant. Given $u \in C^0([0,1])$, the interpolant $u_I \in V_h$ of u is determined by

$$u_I = \sum_{i=1}^{n} u(x_i)\phi_i.$$

It clear that $u_I(x_i) = u(x_i)$, $i = 0, 1, \cdots, n$, and

$$u_I(x) = \frac{x_i - x}{h_i} u(x_{i-1}) + \frac{x - x_{i-1}}{h_i} u(x_i) \qquad \text{for } x \in [x_{i-1}, x_i].$$

Denote by $\tau_i = [x_{i-1}, x_i]$ and by $\|g\|_{L^2(\tau_i)} = \left(\int_{x_{i-1}}^{x_i} g^2 \,\mathrm{d}x\right)^{1/2}$.

THEOREM 2.1.

$$\|u - u_I\|_{L^2(\tau_i)} \leq \frac{1}{\pi} h_i \|u'\|_{L^2(\tau_i)}, \tag{0.9}$$

$$\|u - u_I\|_{L^2(\tau_i)} \leq \frac{1}{\pi^2} h_i^2 \|u''\|_{L^2(\tau_i)}, \tag{0.10}$$

$$\|u' - u_I'\|_{L^2(\tau_i)} \leq \frac{1}{\pi} h_i \|u''\|_{L^2(\tau_i)}. \tag{0.11}$$

PROOF. We only prove (0.9) and leave the others as an exercise. We first change (0.9) to the reference interval $[0,1]$. Let $\hat{x} = (x - x_{i-1})/h_i$ and let $\hat{e}(\hat{x}) = u(x) - u_I(x)$. Note that $\hat{e}(0) = \hat{e}(1) = 0$ and $k = u_I'$ is a constant. The inequality (0.9) is equivalent to

$$\|\hat{e}\|_{L^2([0,1])}^2 = \frac{1}{h_i} \|u - u_I\|_{L^2(\tau_i)}^2 \leq \frac{h_i}{\pi^2} \|u'\|_{L^2(\tau_i)}^2 = \frac{1}{\pi^2} \|\hat{e}' + kh_i\|_{L^2([0,1])}^2,$$

that is

$$\|\hat{e}\|_{L^2([0,1])}^2 \leq \frac{1}{\pi^2} \|\hat{e}'\|_{L^2([0,1])}^2 + \frac{1}{\pi^2} \|kh_i\|_{L^2([0,1])}^2. \tag{0.12}$$

Introduce the space $W = \left\{w \in L^2([0,1]) : \ w' \in L^2([0,1]) \text{ and } w(0) = w(1) = 0\right\}$. Let

$$\lambda_1 = \inf_{w \in W, w \neq 0} R[w] = \inf_{w \in W, w \neq 0} \frac{\|w'\|_{L^2([0,1])}^2}{\|w\|_{L^2([0,1])}^2}.$$

By variational calculus it is easy to see that $R[w]$ is the Rayleigh quotient of the following eigenvalue problem:

$$-w'' = \lambda w, w \in W.$$

Therefore $\lambda_1 = \pi^2$ is the smallest eigenvalue of the above problem, and hence (0.12) holds. This completes the proof of (0.9). □

2.7. A priori error estimate. Introduce the energy norm

$$|||v||| = A(v,v)^{1/2}.$$

From the Cauchy inequality,

$$A(u,v) \leq |||u||| \, |||v|||.$$

By taking $v = v_h \in V_h$ in (0.2) and subtracting it from (0.5), we have the following fundamental orthogonality

$$A(u - u_h, v_h) = 0 \qquad \forall v_h \in V_h. \tag{0.13}$$

Therefore

$$|\!|\!| u-u_h |\!|\!|^2 = A(u-u_h, u-u_h) = A(u-u_h, u-u_I) \le |\!|\!| u-u_h |\!|\!| \, |\!|\!| u-u_I |\!|\!|,$$

It follows from Theorem 2.1 that

$$\begin{aligned}
|\!|\!| u-u_h |\!|\!| \le |\!|\!| u-u_I |\!|\!| &= \left[\sum_{i=1}^{n}(\|u'-u_I'\|^2_{L^2(\tau_i)} + a\,\|u-u_I\|^2_{L^2(\tau_i)})\right]^{1/2} \\
&\le \left[\sum_{i=1}^{n}\left(\left(\frac{h}{\pi}\right)^2 \|u''\|^2_{L^2(\tau_i)} + a\left(\frac{h}{\pi}\right)^4 \|u''\|^2_{L^2(\tau_i)}\right)\right]^{1/2} \\
&= \frac{h}{\pi}\left[\left(1+a\left(\frac{h}{\pi}\right)^2\right)\int_0^1 (u'')^2 \,\mathrm{d}x\right]^{1/2}.
\end{aligned}$$

We have proved the following error estimate.

Theorem 2.2.

$$|\!|\!| u-u_h |\!|\!| \le \frac{h}{\pi}\left(1+a\left(\frac{h}{\pi}\right)^2\right)^{1/2} \|u''\|_{L^2([0,1])}\,.$$

Since the above estimate depends on the unknown solution u, it is called the *a priori* error estimate.

2.8. A posteriori error estimates. We will derive error estimates independent of the unknown solution u.

Let $e = u - u_h$. Then

$$\begin{aligned}
A(e,e) &= A(u-u_h, e-e_I) \\
&= \int_0^1 f\cdot(e-e_I)\,\mathrm{d}x - \int_0^1 u_h'(e-e_I)'\,\mathrm{d}x - \int_0^1 a u_h(e-e_I)\,\mathrm{d}x \\
&= \int_0^1 (f-au_h)(e-e_I)\,\mathrm{d}x - \sum_{i=1}^{n}\int_{x_{i-1}}^{x_i} u_h'(e-e_I)'\,\mathrm{d}x
\end{aligned}$$

Since u_h' is constant on each interval (x_{i-1}, x_i),

$$\begin{aligned}
A(e,e) &= \sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}(f-au_h)(e-e_I)\,\mathrm{d}x \\
&\le \sum_{i=1}^{n}\|f-au_h\|_{L^2(\tau_i)}\,\|e-e_I\|_{L^2(\tau_i)} \\
&\le \sum_{i=1}^{n}\frac{h_i}{\pi}\|f-au_h\|_{L^2(\tau_i)}\,\|e'\|_{L^2(\tau_i)}\,.
\end{aligned}$$

Here we have used Theorem 2.1 to derive the last inequality.

Define the local error estimator on the element $\tau_i = [x_{i-1}, x_i]$ as follows

$$\eta_i = \frac{1}{\pi}h_i\,\|f-au_h\|_{L^2(\tau_i)}\,. \tag{0.14}$$

Then

$$|\!|\!| e |\!|\!|^2 \le \left(\sum_{i=1}^{n}\eta_i^2\right)^{1/2}\|e'\| \le \left(\sum_{i=1}^{n}\eta_i^2\right)^{1/2}|\!|\!| e |\!|\!|.$$

That is, we have the following a posteriori error estimate.

Theorem 2.3 (Upper bound).

$$|||u - u_h||| \leq \left(\sum_{i=1}^{n} \eta_i^2\right)^{1/2}. \tag{0.15}$$

Now a question is if the above upper bound overestimates the true error. To answer this question we introduce the following theorem that gives a lower bound of the true error.

Theorem 2.4 (Lower bound). *Define* $|||\phi|||_{\tau_i} = \left(\int_{x_{i-1}}^{x_i}((\phi')^2 + a\phi^2)\,\mathrm{d}x\right)^{1/2}$. *Let* $(f - au_h)_i = \frac{1}{h_i}\int_{x_{i-1}}^{x_i}(f - au_h)\,\mathrm{d}x$ *and* $\mathrm{osc}_i = \frac{1}{\pi}h_i\,\|f - au_h - (f - au_h)_i\|_{L^2(\tau_i)}$. *Then*

$$\eta_i - \Big(1 + \frac{\sqrt{30}}{5}\Big)\mathrm{osc}_i \leq \frac{1}{\pi}\Big(12 + \frac{6ah_i^2}{5}\Big)^{\frac{1}{2}} |||u - u_h|||_{\tau_i}. \tag{0.16}$$

Proof. Suppose ψ is differentiable over each τ_i and continuous on $[0,1]$. It is clear that

$$A(e,\psi) = \int_0^1 (f - au_h)\psi\,\mathrm{d}x - \sum_{i=1}^{n}\int_{x_{i-1}}^{x_i} u_h'\psi'\,\mathrm{d}x \tag{0.17}$$

Define $\psi_i(x) = 4\phi_{i-1}(x)\phi_i(x)$ if $x \in \tau_i$ and $\psi_i(x) = 0$ otherwise. Choose $\psi = \alpha_i\psi_i$ such that

$$\int_0^1 (f - au_h)_i\psi\,\mathrm{d}x = h_i^2\,\|(f - au_h)_i\|_{L^2(\tau_i)}^2.$$

From (0.8),

$$\alpha_i = \frac{h_i^3(f - au_h)_i}{\int_0^1 \psi_i\,\mathrm{d}x} = \frac{3}{2}h_i^{\frac{1}{2}}h_i\,\|(f - au_h)_i\|_{L^2(\tau_i)}.$$

Therefore, by simple calculations,

$$h_i^{-1}\|\psi\|_{L^2(\tau_i)} = \frac{\sqrt{30}}{5}h_i\,\|(f - au_h)_i\|_{L^2(\tau_i)},\ \|\psi'\|_{L^2(\tau_i)} = 2\sqrt{3}h_i\,\|(f - au_h)_i\|_{L^2(\tau_i)}.$$

From (0.17),

$$A(e,\psi) = \int_0^1 (f-au_h)\psi\,\mathrm{d}x = \int_{x_{i-1}}^{x_i}(f-au_h-(f-au_h)_i)\psi\,\mathrm{d}x + h_i^2\,\|(f - au_h)_i\|_{L^2(\tau_i)}^2.$$

We have,

$$\begin{aligned} h_i^2\,\|(f - au_h)_i\|_{L^2(\tau_i)}^2 &\leq |||e|||_{\tau_i}|||\psi|||_{\tau_i} + \mathrm{osc}_i\pi h_i^{-1}\,\|\psi\|_{L^2(\tau_i)} \\ &= \left(\Big(12 + \frac{6ah_i^2}{5}\Big)^{\frac{1}{2}}|||e|||_{\tau_i} + \frac{\pi\sqrt{30}}{5}\mathrm{osc}_i\right)h_i\,\|(f - au_h)_i\|_{L^2(\tau_i)}, \end{aligned}$$

which implies

$$h_i\,\|(f - au_h)_i\|_{L^2(\tau_i)} \leq \Big(12 + \frac{6ah_i^2}{5}\Big)^{\frac{1}{2}}|||e|||_{\tau_i} + \frac{\pi\sqrt{30}}{5}\mathrm{osc}_i$$

Now the proof is completed by using $\eta_i \leq \frac{1}{\pi}h_i\,\|(f - au_h)_i\|_{L^2(\tau_i)} + \mathrm{osc}_i$. □

We remark that the term osc_i is of high order compared to η_i if f and a are smooth enough on τ_i.

EXAMPLE 2.5. We solve the following problem by the linear finite element method.

$$
\begin{aligned}
&-u'' + 10000u = 1, \quad 0 < x < 1,\\
&u(0) = u(1) = 0.
\end{aligned}
$$

The true solution (see Fig. 2) is

$$
u = \frac{1}{10000}\left(1 - \frac{e^{100x} + e^{100(1-x)}}{1 + e^{100}}\right).
$$

If we use the uniform mesh obtained by dividing the interval $[0,1]$ into 1051 subintervals of equal length, then the error $|||u - u_h||| = 2.7438 \times 10^{-5}$. On the other hand, if we use a non-uniform mesh as shown in Fig. 2 which also contains 1051 subintervals, then the error $|||u - u_h||| = 1.9939 \times 10^{-6}$ is smaller than that obtained by using the uniform mesh.

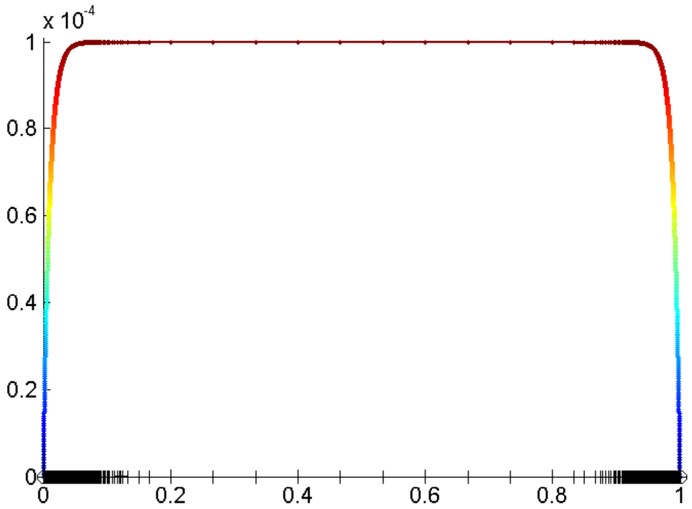

FIGURE 2. Example 2.5. The finite element solution and the mesh.

3. Exercises

EXERCISE 0.1. Prove (0.10) and (0.11).

EXERCISE 0.2. Let $u \in V$, show that the interpolant $u_I \in V_h$ is the best approximation of u in the norm $\left\|\frac{d}{dx}\cdot\right\|_{L^2([0,1])}$, that is,

$$
\|(u - u_I)'\|_{L^2([0,1])} = \inf_{v_h \in V_h} \|(u - v_h)'\|_{L^2([0,1])}\,.
$$

EXERCISE 0.3. Use Example 2.5 to verify numerically the a posteriori error estimates in Theorem 2.3 and 2.4.

CHAPTER 1

Variational Formulation of Elliptic Problems

In this chapter we shall introduce the variational formulation of the elliptic boundary value problem

$$Lu = f \qquad \text{in } \Omega, \qquad u = 0 \qquad \text{on } \partial\Omega, \tag{1.1}$$

where Ω is a bounded open subset of $\mathbb{R}^d$ $(d = 1, 2, 3)$ and $u : \Omega \to \mathbb{R}$ is the unknown function. Here $f : \Omega \to \mathbb{R}$ is a given function and L denotes the second-order partial differential operator of the form

$$Lu = -\sum_{i,j=1}^{d} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}\right) + \sum_{i=1}^{d} b_i(x)\frac{\partial u}{\partial x_i} + c(x)u \tag{1.2}$$

for given coefficients $a_{ij}, b_i, c,\ i, j = 1, 2, \cdots, d$.

We shall assume the partial differential operator L is *uniformly elliptic*, that is, there exists a constant $\theta > 0$ such that

$$\sum_{i,j=1}^{d} a_{ij}(x)\xi_i\xi_j \geqslant \theta|\xi|^2 \qquad \text{for a.e. } x \in \Omega \text{ and all } \xi \in \mathbb{R}^d.$$

1.1. Basic concepts of Sobolev space

Let Ω be an open subset in $\mathbb{R}^d$. We define $C_0^\infty(\Omega)$ to be the linear space of infinitely differentiable functions with compact support in Ω. Let $L^1_{\text{loc}}(\Omega)$ be the set of locally integrable functions:

$$L^1_{\text{loc}}(\Omega) = \left\{f : \ f \in L^1(K) \quad \forall \text{ compact set } K \subset \text{ interior } \Omega\right\},$$

We start with the definition of weak derivatives.

Definition 1.1. Assume $f \in L^1_{\text{loc}}(\Omega), 1 \leqslant i \leqslant d$, we say $g_i \in L^1_{\text{loc}}(\Omega)$ is the *weak partial derivative* of f with respect to x_i in Ω if

$$\int_\Omega f\frac{\partial \varphi}{\partial x_i}\,\mathrm{d}x = -\int_\Omega g_i\varphi\,\mathrm{d}x \qquad \forall\, \varphi \in C_0^\infty(\Omega).$$

We write

$$\partial_{x_i} f = \frac{\partial f}{\partial x_i} = g_i, \quad i = 1, 2, \cdots, d, \qquad \nabla f = (\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_d})^T.$$

Similarly, for a multi-index $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_d) \in \mathbb{N}^d$ with length $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_d$, $\partial^\alpha f \in L^1_{\text{loc}}(\Omega)$ is defined by

$$\int_\Omega \partial^\alpha f \varphi \, \mathrm{d}x = (-1)^{|\alpha|} \int_\Omega f \partial^\alpha \varphi \, \mathrm{d}x \qquad \forall \, \varphi \in C_0^\infty(\Omega),$$

where $\partial^\alpha = \partial_{x_1}^{\alpha_1} \partial_{x_2}^{\alpha_2} \cdots \partial_{x_d}^{\alpha_d}$.

EXAMPLE 1.2. *Let $d = 1, \Omega = (-1, 1)$, and $f(x) = 1 - |x|$. The weak derivative of f is*

$$g = \begin{cases} 1 & \textit{if } x \leqslant 0, \\ -1 & \textit{if } x > 0. \end{cases}$$

The weak derivative of g does not exist.

DEFINITION 1.3 (Sobolev space). For a non-negative integer k and a real $p \geqslant 1$, we define

$$W^{k,p}(\Omega) = \{u \in L^p(\Omega) : \partial^\alpha u \in L^p(\Omega) \quad \text{for all } |\alpha| \leqslant k\}.$$

The space is a Banach space with the norm

$$\|u\|_{W^{k,p}(\Omega)} = \begin{cases} \big(\sum\limits_{|\alpha| \leqslant k} \|\partial^\alpha u\|^p_{L^p(\Omega)} \big)^{1/p}, & 1 \leqslant p < +\infty; \\ \max\limits_{|\alpha| \leqslant k} \|\partial^\alpha u\|_{L^\infty(\Omega)}, & p = +\infty. \end{cases}$$

The closure of $C_0^\infty(\Omega)$ in $W^{k,p}(\Omega)$ is denoted by $W_0^{k,p}(\Omega)$. It is also a Banach space. When $p = 2$, we denote

$$H^k(\Omega) = W^{k,2}(\Omega), \qquad H_0^k(\Omega) = W_0^{k,2}(\Omega).$$

The space $H^k(\Omega)$ is a Hilbert space when equipped with the inner product

$$(u, v)_{k,\Omega} = \sum_{|\alpha| \leqslant k} \int_\Omega \partial^\alpha u \partial^\alpha v \, \mathrm{d}x.$$

EXAMPLE 1.4.

(1) *Let $\Omega = (0, 1)$ and consider the function $u = x^\alpha$. One easily verifies that $u \in L^2(\Omega)$ if $\alpha > -\frac{1}{2}$, $u \in H^1(\Omega)$ if $\alpha > \frac{1}{2}$, and $u \in H^k(\Omega)$ if $\alpha > k - \frac{1}{2}$.*

(2) *Let $\Omega = \{x \in \mathbb{R}^2 : |x| < 1/2\}$ and consider the function $f(x) = \log \big| \log |x| \big|$. Then $f \in W^{1,p}(\Omega)$ for $p \leqslant 2$ but $f \notin L^\infty(\Omega)$. This example shows that functions in $H^1(\Omega)$ are neither necessarily continuous nor bounded.*

Now we consider the regularization of functions in Sobolev space. Let ρ be a non-negative, real-valued function in $C_0^\infty(\mathbb{R}^d)$ with the property

$$\int_{\mathbb{R}^d} \rho(x)\,\mathrm{d}x = 1, \quad \operatorname{supp}(\rho) \subset \{x : |x| \leqslant 1\}\,. \tag{1.3}$$

An example of such a function is

$$\rho(x) = \begin{cases} Ce^{\frac{1}{|x|^2-1}} & \text{if } |x| < 1, \\ 0 & \text{if } |x| \geqslant 1, \end{cases} \tag{1.4}$$

where the constant C is so chosen that $\int_{\mathbb{R}^d} \rho(x)\,\mathrm{d}x = 1$. For $\epsilon > 0$, the function $\rho_\epsilon(x) = \epsilon^{-d}\rho(x/\epsilon)$ belongs to $C_0^\infty(\mathbb{R}^d)$ and $\operatorname{supp}(\rho_\epsilon) \subset \{x : |x| < \epsilon\}$. ρ_ϵ is called the *mollifier* and the convolution

$$u_\epsilon(x) = (\rho_\epsilon * u)(x) = \int_{\mathbb{R}^d} \rho_\epsilon(x-y)u(y)\,\mathrm{d}y \tag{1.5}$$

is called the *regularization* of u. Regularization has several important and useful properties that are summarized in the following lemma.

Lemma 1.1.

(i) *If $u \in L^1_{\text{loc}}(\mathbb{R}^d)$, then for every $\epsilon > 0$, $u_\epsilon \in C^\infty(\mathbb{R}^n)$ and $\partial^\alpha(\rho_\epsilon * u) = (\partial^\alpha \rho_\epsilon) * u$ for each multi-index α;*

(ii) *If $u \in C(\mathbb{R}^d)$, then u_ε converges uniformly to u on compact subsets of $\mathbb{R}^d$;*

(iii) *If $u \in L^p(\mathbb{R}^d), 1 \leqslant p < \infty$, then $u_\epsilon \in L^p(\mathbb{R}^d)$, $\|u_\epsilon\|_{L^p(\mathbb{R}^d)} \leqslant \|u\|_{L^p(\mathbb{R}^d)}$, and $\lim_{\epsilon\to 0} \|u_\epsilon - u\|_{L^p(\mathbb{R}^d)} = 0$.*

Proof. (i) follows directly from (1.5).

(ii) is obvious by observing that

$$\begin{aligned} |u_\varepsilon(x) - u(x)| &\leqslant \int_{\mathbb{R}^d} \rho_\epsilon(x-y)\,|u(x) - u(y)|\,\mathrm{d}y \\ &\leqslant (\max\rho)\varepsilon^{-d} \int_{|y-x|\leqslant\varepsilon} |u(x) - u(y)|\,\mathrm{d}y \end{aligned}$$

and that u is uniformly continuous on compact sets.

To show (iii), let $p' \in \mathbb{R}$ such that $1/p + 1/p' = 1$. Then by Hölder inequality

$$\begin{aligned}
&\|u_\epsilon\|_{L^p(\mathbb{R}^d)} \\
&\leqslant \left\{ \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} |u(y)| \rho_\epsilon(x-y)\,\mathrm{d}y \right)^p \mathrm{d}x \right\}^{1/p} \\
&\leqslant \left\{ \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} |u(y)|^p \rho_\epsilon(x-y)\,\mathrm{d}y \right) \cdot \left(\int_{\mathbb{R}^d} \rho_\epsilon(x-y)\,\mathrm{d}y \right)^{p/p'} \mathrm{d}x \right\}^{1/p} \\
&= \left\{ \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |u(y)|^p \rho_\epsilon(x-y)\,\mathrm{d}y\,\mathrm{d}x \right\}^{1/p} \\
&= \|u\|_{L^p(\mathbb{R}^d)}. \qquad (1.6)
\end{aligned}$$

For $u \in L^p(\mathbb{R}^d)$ and any $\delta > 0$, we choose a continuous function v with compact support such that $\|u - v\|_{L^p(\mathbb{R}^d)} \leqslant \delta/3$. From (ii), $\|v_\epsilon - v\|_{L^p(\mathbb{R}^d)} \leqslant \delta/3$ for ϵ sufficiently small. By the triangle inequality and (1.6),

$$\|u_\epsilon - u\|_{L^p(\mathbb{R}^d)} \leqslant \|u_\epsilon - v_\epsilon\|_{L^p(\mathbb{R}^d)} + \|v_\epsilon - v\|_{L^p(\mathbb{R}^d)} + \|u - v\|_{L^p(\mathbb{R}^d)} \leqslant \delta. \quad (1.7)$$

This completes the proof of (iii). □

In this book, a *domain* is referred to an open and connected set. The following lemma will be useful in proving the Poincaré-Friedrichs inequality.

LEMMA 1.2. *Let Ω be a domain, $u \in W^{1,p}(\Omega)$, $1 \leqslant p \leqslant \infty$, and $\nabla u = 0$ a.e. on Ω, then u is constant on Ω.*

PROOF. For any bounded subdomain K of Ω and $\epsilon > 0$, let K_ϵ be the ϵ-neighborhood of K, that is, K_ϵ is the union of all balls $B(x, \epsilon)$, $x \in K$. Let u be extended to be zero outside Ω and let $u_\epsilon = u * \rho_\epsilon$. If $K_\epsilon \subset \Omega$ for some $\epsilon > 0$, then $\nabla u_\epsilon = (\nabla u) * \rho_\epsilon = 0$ in K. Since u_ϵ is smooth, we deduce that u_ϵ is constant in K. On the other hand, by Lemma 1.1, $u_\epsilon \to u$ in $L^1(K)$. Thus u is constant in K. This completes the proof. □

THEOREM 1.5 (Properties of weak derivatives). *Assume $1 \leqslant p < +\infty$.*

(i) *(Product rule) If $f, g \in W^{1,p}(\Omega) \cap L^\infty(\Omega)$, then $fg \in W^{1,p}(\Omega)$ and*

$$\frac{\partial (fg)}{\partial x_i} = \frac{\partial f}{\partial x_i} g + f \frac{\partial g}{\partial x_i} \quad \text{a.e. in } \Omega, \quad i = 1, 2, \cdots, d;$$

(ii) *(Chain rule) If $f \in W^{1,p}(\Omega)$ and $F \in C^1(\mathbb{R}), F' \in L^\infty(\mathbb{R})$, then $F(f) \in W^{1,p}(\Omega)$ and*

$$\frac{\partial F(f)}{\partial x_i} = F'(f) \frac{\partial f}{\partial x_i} \qquad \text{a.e. in } \Omega, \quad i = 1, 2, \cdots, d;$$

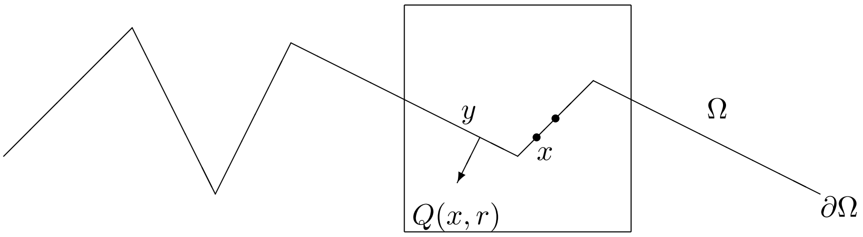

FIGURE 1. The domain with a Lipschitz boundary

(iii) *If $f \in W^{1,p}(\Omega)$ and F is piecewise smooth on $\mathbb{R}$ with $F' \in L^\infty(\mathbb{R})$, then $F(f) \in W^{1,p}(\Omega)$. Furthermore, if $\mathcal{L}$ is the set of all corner points of F, we have, for $i = 1, 2, \cdots, d$,*

$$\frac{\partial F(f)}{\partial x_i} = \begin{cases} F'(f)\frac{\partial f}{\partial x_i} & \textit{if} \quad f(x) \notin \mathcal{L}, \\ 0 & \textit{if} \quad f(x) \in \mathcal{L}. \end{cases}$$

In order to introduce further properties of Sobolev spaces, we introduce the following condition on the boundary of the domain.

DEFINITION 1.6 (Lipschitz domain). We say that a domain Ω has a *Lipschitz boundary* $\partial\Omega$ if for each point $x \in \partial\Omega$ there exist $r > 0$ and a Lipschitz mapping $\varphi : \mathbb{R}^{d-1} \to \mathbb{R}$ such that — upon rotating and relabeling the coordinate axes if necessary — we have

$$\Omega \cap Q(x, r) = \{y : \varphi(y_1, \cdots, y_{d-1}) < y_d\} \cap Q(x, r),$$

where $Q(x, r) = \{y : |y_i - x_i| < r, i = 1, 2, \cdots, d\}$. We call Ω a Lipschitz domain if it has a Lipschitz boundary.

THEOREM 1.7. *Let Ω be a Lipschitz domain in $\mathbb{R}^d$.*

(i) *Let $\mathcal{D}(\bar{\Omega})$ be the set of all functions $\varphi|_\Omega, \varphi \in C_0^\infty(\mathbb{R}^d)$. Then $\mathcal{D}(\bar{\Omega})$ is dense in $W^{k,p}(\Omega)$ for all integers $k \geqslant 0$ and real p with $1 \leqslant p < \infty$;*

(ii) *Let $u \in W^{k,p}(\Omega)$ and let $\widetilde{u}$ denote its extension by zero outside Ω. If $\widetilde{u} \in W^{k,p}(\mathbb{R}^d)$, for $k \geqslant 1, 1 \leqslant p < \infty$, then $u \in W_0^{k,p}(\Omega)$;*

(iii) *If in addition Ω is bounded and $k \geqslant 1, 1 \leqslant p \leqslant \infty$, there exists a continuous linear extension operator $\mathbb{E}$ from $W^{k,p}(\Omega)$ to $W^{k,p}(\mathbb{R}^d)$ such that $\mathbb{E}u = u$ in Ω.*

The following theorem plays an important role in the application of Sobolev spaces.

THEOREM 1.8 (Sobolev Imbedding Theorem). *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain and $1 \leqslant p \leqslant \infty$. Then*

(i) *If* $0 \leqslant k < d/p$, *the space* $W^{k,p}(\Omega)$ *is continuously imbedded in* $L^q(\Omega)$ *with* $q = dp/(d-kp)$ *and compactly imbedded in* $L^{q'}(\Omega)$ *for any* $1 \leqslant q' < q$;

(ii) *If* $k = d/p$, *the space* $W^{k,p}(\Omega)$ *is compactly imbedded in* $L^q(\Omega)$ *for any* $1 \leqslant q < \infty$;

(iii) *If* $0 \leqslant m < k - \frac{d}{p} < m+1$, *the space* $W^{k,p}(\Omega)$ *is continuously imbedded in* $C^{m,\alpha}(\bar{\Omega})$ *for* $\alpha = k - \frac{d}{p} - m$, *and compactly imbedded in* $C^{m,\beta}(\bar{\Omega})$ *for all* $0 \leqslant \beta < \alpha$.

Example 1.9. $H^1(\Omega)$ *is continuously imbedded in* $C^{0,1/2}(\bar{\Omega})$ *for* $d=1$, *in* $L^q(\Omega)$, $1 \leqslant q < \infty$, *for* $d = 2$, *and in* $L^6(\Omega)$ *for* $d = 3$.

Theorem 1.10 (Poincaré-Friedrichs Inequality). *Let* $\Omega \subset \mathbb{R}^d$ *be a bounded Lipschitz domain and* $1 \leqslant p \leqslant \infty$. *Then*

$$\|u\|_{L^p(\Omega)} \leqslant C_p \|\nabla u\|_{L^p(\Omega)} \quad \forall u \in W_0^{1,p}(\Omega),$$
$$\|u - u_\Omega\|_{L^p(\Omega)} \leqslant C_p \|\nabla u\|_{L^p(\Omega)} \quad \forall u \in W^{1,p}(\Omega),$$

where $u_\Omega = \frac{1}{|\Omega|}\int_\Omega u(x)\,\mathrm{d}x$.

Proof. We only give the proof of the first inequality. Assume it is false. Then there exists a sequence $\{u_n\} \subset W_0^{1,p}(\Omega)$ such that

$$\|u_n\|_{L^p(\Omega)} = 1, \quad \|\nabla u_n\|_{L^p(\Omega)} \leqslant \frac{1}{n}.$$

By the compactness imbedding theorem, there exists a subsequence (still denoted by) u_n and a function $u \in L^p(\Omega)$ such that $u_n \to u$ in $L^p(\Omega)$. By the completeness of $L^p(\Omega)$ we know that $\nabla u_n \to 0$ in $L^p(\Omega)^d$. Thus, by the definition of weak derivative, $\nabla u = 0$, which implies, by Lemma 1.2, that $u = 0$. This contradicts the fact that $\|u\|_{L^p(\Omega)} = 1$. □

Next we study the trace of functions in $W^{k,p}$ for which we first introduce the Sobolev spaces of non-integer order k. There are several definitions of fractional Sobolev spaces which unfortunately are not equivalent. Here we shall use the following one.

Definition 1.11 (Fractional Sobolev space). For two real numbers s, p with $p \geqslant 1$ and $s = k + \sigma$ where $\sigma \in (0,1)$. We define $W^{s,p}(\Omega)$ when $p < \infty$ as the set of all functions $u \in W^{k,p}(\Omega)$ such that

$$\int_\Omega\int_\Omega \frac{|\partial^\alpha u(x) - \partial^\alpha u(y)|^p}{|x-y|^{d+\sigma p}}\,\mathrm{d}x\,\mathrm{d}y < +\infty \quad \forall |\alpha| = k.$$

Likewise, when $p = \infty$, $W^{s,\infty}(\Omega)$ is the set of all functions $u \in W^{k,\infty}(\Omega)$ such that

$$\max_{|\alpha|=k} \operatorname*{ess\,sup}_{x,y\in\Omega, x\neq y} \frac{|\partial^\alpha u(x) - \partial^\alpha u(y)|}{|x-y|^\sigma} < \infty \quad \forall|\alpha| = k.$$

$W^{s,p}(\Omega)$ when $p < \infty$ is a Banach space with the norm

$$\|u\|_{W^{s,p}(\Omega)} = \left\{ \|u\|^p_{W^{k,p}(\Omega)} + \sum_{|\alpha|=k} \int_\Omega \int_\Omega \frac{|\partial^\alpha u(x) - \partial^\alpha u(y)|^p}{|x-y|^{d+\sigma p}} \,\mathrm{d}x\,\mathrm{d}y \right\}^{1/p}$$

with the obvious modification when $p = \infty$.

The closure of $C_0^\infty(\Omega)$ in $W^{s,p}(\Omega)$ is denoted by $W_0^{s,p}(\Omega)$. It is also a Banach space. When $p = 2$, we denote $H^s(\Omega) = W^{s,2}(\Omega)$ and $H_0^s(\Omega) = W_0^{s,2}(\Omega)$.

We remark that the statement of Sobolev Imbedding Theorem 1.8 is valid for fractional Sobolev spaces. The density result and the extension result in Theorem 1.7 are valid as well for fractional Sobolev spaces when $s > 0$.

Now we examine the boundary values of functions in $W^{s,p}(\Omega)$. The fractional Sobolev space $W^{s,p}(\Gamma)$ on the boundary Γ of Ω can be defined by using the atlas of the boundary Γ and using the definition of fractional Sobolev space in Definition 1.6 locally. As we are mostly interested in the case when $s < 1$ we make use of the following equivalent definition of Sobolev space on the boundary.

Definition 1.12 (Sobolev space on the boundary). Let Ω be a bounded Lipschitz domain in $\mathbb{R}^d$ with boundary Γ. Let s, p be two real numbers with $0 \leqslant s < 1$ and $1 \leqslant p < \infty$. We define $W^{s,p}(\Gamma)$ as the set of all functions $u \in L^p(\Omega)$ such that

$$\int_\Gamma \int_\Gamma \frac{|u(x) - u(y)|^p}{|x-y|^{d-1+sp}} \,\mathrm{d}s(x)\,\mathrm{d}s(y) < \infty.$$

$W^{s,p}(\Gamma)$ is a Banach space with the norm

$$\|u\|_{W^{s,p}(\Gamma)} = \left\{ \|u\|^p_{L^p(\Gamma)} + \int_\Gamma \int_\Gamma \frac{|u(x) - u(y)|^p}{|x-y|^{d-1+sp}} \,\mathrm{d}s(x)\,\mathrm{d}s(y) \right\}^{1/p}.$$

As usual, when $p = 2$, $H^s(\Gamma) = W^{s,2}(\Gamma)$.

We know that if u is continuous on $\bar{\Omega}$ then its restriction to the boundary $\partial\Omega$ is well-defined and continuous. If however, u is a function in some Sobolev space, the restriction $u|_{\partial\Omega}$ may not be defined in a pointwise sense. To interpret boundary values of Sobolev functions properly, we introduce the following trace theorem for Sobolev spaces.

THEOREM 1.13 (Trace Theorem). *Let Ω be a bounded Lipschitz domain with boundary Γ, $1 \leqslant p < \infty$, and $1/p < s \leqslant 1$.*

(i) *There exists a bounded linear mapping*

$$\gamma_0 : W^{s,p}(\Omega) \text{ onto } W^{s-1/p,p}(\Gamma)$$

such that $\gamma_0(u) = u$ on Γ for all $u \in W^{s,p}(\Omega) \cap C(\bar{\Omega})$;

(ii) *For all $v \in C^1(\bar{\Omega})$ and $u \in W^{1,p}(\Omega)$,*

$$\int_\Omega u \frac{\partial v}{\partial x_i} \,\mathrm{d}x = -\int_\Omega \frac{\partial u}{\partial x_i} v \,\mathrm{d}x + \int_\Gamma \gamma_0(u)\, v\, n_i \,\mathrm{d}s,$$

where n_i denotes the i-th component of the unit outward normal to Γ;

(iii) $W_0^{1,p}(\Omega) = \{u \in W^{1,p}(\Omega) : \gamma_0(u) = 0\}$.

(iv) *γ_0 has a continuous right inverse, that is, there exists a constant C such that, $\forall g \in W^{s-1/p,p}(\Gamma)$, there exists $u_g \in W^{s,p}(\Omega)$ satisfying*

$$\gamma_0(u_g) = g \quad \text{and} \quad \|u_g\|_{W^{s,p}(\Omega)} \leqslant C\, \|g\|_{W^{s-1/p,p}(\Gamma)}\,.$$

$\gamma_0(u)$ is called the trace of u on the boundary $\Gamma = \partial\Omega$. Noting that γ_0 is surjective and the property (iv) is a consequence of (i) and the open mapping theorem. The function u_g is said to be a *lifting* of g in $W^{s,p}(\Omega)$. In what follows, whenever no confusion can arise, we write u instead of $\gamma_0(u)$ on boundaries.

1.2. Variational formulation

We assume $f \in L^2(\Omega)$ and the coefficients in (1.2) satisfies $a_{ij}, b_i, c \in L^\infty(\Omega), i,\, j = 1, 2, \cdots, d$.

Assuming for the moment the solution u is a smooth function, we multiply $Lu = f$ in (1.1) by a smooth function $\varphi \in C_0^\infty(\Omega)$, and integrate over Ω, to find

$$\int_\Omega \left(\sum_{i,j=1}^{d} a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial \varphi}{\partial x_i} + \sum_{i=1}^{d} b_i \frac{\partial u}{\partial x_i} \varphi + cu\varphi \right) \mathrm{d}x = \int_\Omega f\varphi \,\mathrm{d}x, \tag{1.8}$$

where we have used the integration by parts formula in Theorem 1.13 in the first term on the left hand side. There are no boundary terms since $\varphi = 0$ on $\partial\Omega$. By the density argument we deduce that (1.8) is valid for any $\varphi \in H_0^1(\Omega)$, and the resulting equation makes sense if $u \in H_0^1(\Omega)$. We choose the space

$H_0^1(\Omega)$ to incorporate the boundary condition from (1.1) that "$u = 0$" on $\partial\Omega$. This motivates us to define the bilinear form $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ as follows

$$a(u,\varphi) = \int_\Omega \Big(\sum_{i,j=1}^{d} a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial \varphi}{\partial x_i} + \sum_{i=1}^{d} b_i \frac{\partial u}{\partial x_i} \varphi + cu\varphi \Big)\, \mathrm{d}x.$$

Definition 1.14. $u \in H_0^1(\Omega)$ ia called a *weak solution* of the boundary value problem (1.1) if

$$a(u,\varphi) = (f,\varphi) \qquad \forall\, \varphi \in H_0^1(\Omega),$$

where $(\cdot,\cdot)$ denotes the inner product on $L^2(\Omega)$.

More generally, we can consider the boundary value problem (1.1) for $f \in H^{-1}(\Omega)$, the dual space of $H_0^1(\Omega)$. For example, f is defined by

$$\langle f,\varphi\rangle = \int_\Omega \Big(f_0\varphi + \sum_{i=1}^{d} f_i \frac{\partial \varphi}{\partial x_i} \Big)\, \mathrm{d}x, \quad \forall\, \varphi \in H_0^1(\Omega),$$

where $f_i \in L^2(\Omega), i = 0, 1, \cdots, d$, and $\langle \cdot,\cdot\rangle$ denotes the duality pairing of $H^{-1}(\Omega)$ and $H_0^1(\Omega)$.

Definition 1.15. Suppose $f \in H^{-1}(\Omega)$. $u \in H_0^1(\Omega)$ is called a weak solution of (1.1) if

$$a(u,\varphi) = \langle f,\varphi\rangle \qquad \forall\, \varphi \in H_0^1(\Omega).$$

The inhomogeneous boundary-value problem

$$Lu = f \qquad \text{in } \Omega, \qquad u = g \qquad \text{on } \partial\Omega,$$

can be transformed to the homogeneous one if $g \in H^{1/2}(\Gamma)$ is the trace of some function $w \in H^1(\Omega)$. Then $\tilde{u} = u - w \in H_0^1(\Omega)$ is a weak solution of the problem

$$L\tilde{u} = \tilde{f} \qquad \text{in } \Omega, \qquad \tilde{u} = 0 \qquad \text{on } \partial\Omega,$$

where $\tilde{f} = f - Lw \in H^{-1}(\Omega)$.

Theorem 1.16 (Lax-Milgram Lemma). *Assume that V is a real Hilbert space, with norm $\|\cdot\|$ and inner product $(\cdot,\cdot)$. Assume that $a : V \times V \to \mathbb{R}$ is a bilinear form, for which there exist constants $\alpha, \beta > 0$ such that*

$$|a(u,v)| \leqslant \beta \|u\|\|v\| \qquad \forall\, u, v \in V, \tag{1.9}$$

and

$$a(v,v) \geqslant \alpha \|v\|^2 \qquad \forall\, v \in V. \tag{1.10}$$

Let $f : V \to \mathbb{R}$ be a bounded linear functional on V. Then there exists a unique element $u \in V$ such that

$$a(u, v) = \langle f, v \rangle \qquad \forall\, v \in V. \tag{1.11}$$

The bilinear form a is called V-elliptic (or V-coercive) if it satisfies (1.10). The Lax-Milgram lemma is a consequence of the following generalized Lax-Milgram lemma.

THEOREM 1.17 (Generalized Lax-Milgram Lemma). *Let U and V be real Hilbert spaces and let $a(\cdot,\cdot)$ denote a bounded bilinear form on $U \times V$. Consider the variational problem: Find $u \in U$ such that*

$$a(u, v) = \langle f, v \rangle \qquad \forall v \in V. \tag{1.12}$$

(1.12) *attains a unique solution $u \in U$ for any $f \in V'$ if and only if $a(u, v)$ satisfies the following two conditions:*

(i) *There exists a constant α such that*

$$\inf_{u\in U, \|u\|_U=1} \sup_{v\in V, \|v\|_V=1} |a(u, v)| \geqslant \alpha > 0;$$

(ii) *For every $v \in V, v \neq 0$,*

$$\sup_{u\in U} |a(u, v)| > 0.$$

(i) *is called the* inf-sup condition *of the bilinear form.*

PROOF. Denote by $(\cdot,\cdot)$ the inner product on V. From the Riesz representation theorem, there exist two bounded linear operators $J : U \to V$ and $K : V' \to V$ such that

$$(Ju, v) = a(u, v) \quad \forall u \in U, v \in V,$$
$$(Kf, v) = \langle f, v \rangle \quad \forall v \in V, f \in V'.$$

Then the problem (1.12) is equivalent to: Find $u \in U$ such that

$$Ju = Kf \tag{1.13}$$

Since

$$\sup_{v\in V, \|v\|_V=1} |a(u, v)| = \sup_{v\in V, \|v\|_V=1} |(Ju, v)| = \|Ju\|_V\,, \tag{1.14}$$

(i) is equivalent to: there exists $\alpha > 0$ such that

$$\inf_{u\in U, \|u\|_U=1} \|Ju\|_V \geqslant \alpha. \tag{1.15}$$

If (i) and (ii) hold, then from (1.15) J is injective and $R(J)$, the range of J, is closed. It follows from (ii) that for any $v \in V, v \neq 0$,

$$\sup_{u \in U} |a(u,v)| = \sup_{u \in U} |(Ju,v)| > 0$$

which implies that $R(J)^{\perp} = \{0\}$, that is, J is surjective. Therefore, J is invertible and hence (1.13) attains a unique solution in U for any $f \in V'$.

It remains to prove the necessity. If (1.12) attains a unique solution in U for any $f \in V'$, then J is invertible and the open mapping theorem implies that J^{-1} is continuous. There exist $\alpha > 0$ such that

$$\left\| J^{-1} v \right\|_U \leqslant \frac{1}{\alpha} \left\| v \right\|_V \text{ or } \left\| Ju \right\|_V \geqslant \alpha \left\| u \right\|_U .$$

From (1.14), (i) holds. (ii) is a consequence of $R(J) = V$. This completes the proof of the theorem. □

Remark 1.18. The Generalized Lax-Milgram Lemma is still valid if U and V are complex Hilbert spaces and $a(\cdot,\cdot)$ is a bounded sesquilinear form.

Corollary 1.19. *If L is uniformly elliptic, $b_i = 0$ for $i = 1, \cdots, d$, and $c(x) \geqslant 0$. Suppose $f \in H^{-1}(\Omega)$. Then the boundary value problem $Lu = f$ in Ω has a unique weak solution $u \in H_0^1(\Omega)$.*

Theorem 1.20 (Regularity). *Assume that $a_{ij} \in C^1(\bar{\Omega}), b_i, c \in L^\infty(\Omega), i, j = 1, \cdots, d$, and $f \in L^2(\Omega)$. Suppose that $u \in H_0^1(\Omega)$ is the weak solution of the problem $Lu = f$ in Ω. Assume that $\partial\Omega$ is smooth ($C^{1,1}$) or Ω is convex. Then $u \in H^2(\Omega)$ satisfies the estimate*

$$\|u\|_{H^2(\Omega)} \leqslant C(\|f\|_{L^2(\Omega)} + \|u\|_{L^2(\Omega)}).$$

Bibliographic notes. The standard reference on Sobolev spaces is Adams [**1**]. Here we mainly follow the development in Evans [**32**] and Girault and Ravairt [**34**]. Further results on regularity theory for elliptic equations can be found in Gilbarg and Trudinger [**33**] for smooth domains and in Grisvard [**35**], Dauge [**25**] for non-smooth domains. The generalized Lax-Milgram Theorem 1.17 is due to Nečas [**45**].

1.3. Exercises

Exercise 1.1. If Ω is an open subset in $\mathbb{R}^d$ and K is a compact subset of Ω, show that there exists a function $\varphi \in C_0^\infty(\mathbb{R}^d)$ such that $\operatorname{supp}(\varphi) \subset \Omega$ and $\varphi = 1$ in K.

Exercise 1.2 (Partition of unity). Let $\{O_i\}, i = 1, \cdots, k$, be a family of open sets in $\mathbb{R}^d$ that covers a compact set K. Then there exists a family of functions $\varphi_i \geqslant 0$, $\varphi_i \in C_0^\infty(O_i)$, and $\sum_{i=1}^k \varphi_i = 1$ in K. $\{\varphi_i\}$ is called a partition of unity subordinate to $\{O_i\}$.

Exercise 1.3. Let Ω be a bounded domain with Lipschitz boundary. For any $g \in H^{1/2}(\partial\Omega)$, let $u_g \in H^1(\Omega)$ be the weak solution of the following Dirichlet boundary value problem

$$-\Delta u = 0 \quad \text{in } \Omega, \qquad u = g \quad \text{on } \partial\Omega.$$

Show that there exist constants C_1 and C_2 such that

$$C_1 \|g\|_{H^{1/2}(\partial\Omega)} \leqslant \|u_g\|_{H^1(\Omega)} \leqslant C_2 \|g\|_{H^{1/2}(\partial\Omega)}.$$

Exercise 1.4. Let Ω be a bounded domain in $\mathbb{R}^d$. A function $u \in H^1(\Omega)$ is a weak solution of Neumann problem

$$-\triangle u = f \qquad \text{in } \Omega, \qquad \frac{\partial u}{\partial \nu} = 0 \qquad \text{on } \partial\Omega$$

if

$$\int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x = \int_\Omega f v \,\mathrm{d}x \qquad \forall\, v \in H^1(\Omega). \tag{1.16}$$

Suppose $f \in L^2(\Omega)$. Prove (1.16) has a weak solution if and only if $\int_\Omega f \,\mathrm{d}x = 0$.

Exercise 1.5. Let Ω be a bounded domain in $\mathbb{R}^d$. A function $u \in H_0^2(\Omega)$ is a weak solution of the homogeneous boundary value problem for the biharmonic equation

$$\triangle^2 u = f \qquad \text{in } \Omega, \qquad u = \frac{\partial u}{\partial \nu} = 0 \qquad \text{on } \partial\Omega$$

provided

$$\int_\Omega \triangle u \cdot \triangle v \,\mathrm{d}x = \int_\Omega f v \,\mathrm{d}x \qquad \forall\, v \in H_0^2(\Omega). \tag{1.17}$$

Suppose $f \in L^2(\Omega)$. Prove that there exists a unique weak solution of (1.17).

CHAPTER 2

Finite Element Methods for Elliptic Equations

2.1. Galerkin method for variational problems

Let V be a real Hilbert space with the norm $\|\cdot\|_V$ and inner product $(\cdot,\cdot)_V$. Assume that the bilinear form $a : V \times V \to \mathbb{R}$ satisfies (1.9) and (1.10), i.e. a is bounded and V-elliptic. Let $f : V \to \mathbb{R}$ be a bounded linear functional on V. We consider the variational problem to find $u \in V$ such that

$$a(u,v) = \langle f, v\rangle \qquad \forall\ v \in V, \tag{2.1}$$

where $\langle\cdot,\cdot\rangle$ denotes the duality pairing between V and V'.

Let V_h be a subspace of V which is finite dimensional, h stands for a discretization parameter. The Galerkin method of the variation problem is then to find $u_h \in V_h$ such that

$$a(u_h,v_h) = \langle f, v_h\rangle \qquad \forall\ v \in V_h. \tag{2.2}$$

Suppose that $\{\phi_1,\cdots,\phi_N\}$ is a basis for V_h. Then (2.2) is equivalent to

$$a(u_h,\phi_i) = \langle f,\phi_i\rangle, \quad i = 1,\cdots,N.$$

Writing u_h in the form

$$u_h = \sum_{j=1}^{N} z_j\phi_j, \tag{2.3}$$

we are led to the system of equations

$$\sum_{j=1}^{N} a(\phi_j,\phi_i)z_j = \langle f,\phi_i\rangle, \quad i = 1,\cdots,N,$$

which we can write in the matrix-vector form as

$$Az = b,$$

where $A_{ij} = a(\phi_j, \phi_i)$ and $b_i = \langle f, \phi_i \rangle$. Since a is V-elliptic, the matrix A is positive definite:

$$z^T Az = \sum_{i,j=1}^{N} z_i A_{ij} z_j = a\Big(\sum_{j=1}^{N} z_j \phi_j, \sum_{i=1}^{N} z_i \phi_i\Big) = a(u_h, u_h) \geqslant \alpha \|u_h\|_V^2,$$

and so $z^T Az > 0$, for any $z \neq 0$. The matrix A is called the *stiffness matrix*.

THEOREM 2.1 (Céa Lemma). *Suppose the bilinear form* $a(\cdot, \cdot)$ *satisfies* (1.9) *and* (1.10), *i.e., a is bounded and V-elliptic. Suppose u and* u_h *are the solutions of the variational problem* (2.1) *and its Galerkin approximation* (2.2), *respectively. Then*

$$\|u - u_h\|_V \leqslant \frac{\beta}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_V. \tag{2.4}$$

PROOF. Since $V_h \subset V$, by the definition of u and u_h,

$$a(u, v_h) = \langle f, v_h \rangle \qquad \forall\ v_h \in V_h,$$
$$a(u_h, v_h) = \langle f, v_h \rangle \qquad \forall\ v_h \in V_h.$$

It follows by subtraction we obtain the following *Galerkin orthogonality*

$$a(u - u_h, v_h) = 0 \qquad \forall\ v_h \in V_h, \tag{2.5}$$

which implies that $a(u - u_h, v_h - u_h) = 0$. Thus

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

After dividing by $\|u - u_h\|_V$, the assertion is established. □

According to Céa Lemma, the accuracy of a numerical solution depends essentially on choosing function spaces which are capable of approximating the solution u well. For polynomials, the order of approximation is determined by the smoothness of the solution. However, for boundary-value problems, the smoothness of the solution typically decreases as we approach the boundary. Thus, it may not be advantageous to insist on a high accuracy by forcing the degree of the polynomials to be high.

There are several methods related.

Rayleigh-Ritz method. When the bilinear form $a : V \times V \to \mathbb{R}$ is symmetric, then the variational problem (2.1) is equivalent to the minimization problem

$$\min_{v \in V} J(v), \quad J(v) := \frac{1}{2} a(v, v) - \langle f, v \rangle. \tag{2.6}$$

The Rayleigh-Ritz method is then to solve (2.6) by solving $u_h \in V_h$ as

$$\min_{v_h \in V_h} J(v_h).$$

Usually one finds u_h as in (2.3) by solving the equation $(\partial/\partial z_i)J(\sum_{j=1}^{N} z_j\phi_j) = 0$.

Galerkin method. The weak equation (2.2) is solved for problems where the bilinear form is not necessarily symmetric. If the weak equations arise from a variational problem with a positive quadratic form, then often the term *Ritz-Galerkin method* is used.

Petrov-Galerkin method. We seek $u_h \in V_h$ with

$$a(u_h, v) = \langle f, v\rangle \qquad \forall\ v \in S_h,$$

where V_h is termed the *trial space*, S_h is termed the *test space*, and the two spaces V_h and S_h need not be the same but have the same dimension.

Finite element method. The finite element method can be regarded as a special kind of Galerkin method that uses piecewise polynomials to construct discrete approximating function spaces.

2.2. The construction of finite element spaces

In practice, the spaces used in finite element methods over which we solve the variational problems are called finite element spaces. We partition the given domain Ω into (finitely many) subdomains, and consider functions which reduce to a polynomial on each subdomain (element). For planar problems, the elements can be triangles or quadrilaterals. For three-dimensional problems, we can use tetrahedrons, hexahedrons, etc. For simplicity, we restrict our discussion primarily to the piecewise polynomial approximations over triangular (2D) or tetrahedral (3D) elements.

2.2.1. The finite element.

Definition 2.2. A *finite element* is a triple $(K, \mathcal{P}, \mathcal{N})$ with the following properties:

(i) $K \subset \mathbb{R}^d$ is a closed set with piecewise smooth boundary (the *element*);

(ii) $\mathcal{P}$ is a finite-dimensional space of functions on K (the *space functions*);

(iii) $\mathcal{N} = \{N_1, \cdots, N_n\}$ is a basis for $\mathcal{P}'$ (the *nodal variables* or *degrees of freedom*).

DEFINITION 2.3. Let $(K, \mathcal{P}, \mathcal{N})$ be a finite element, and let $\{\psi_1, \psi_2, \cdots, \psi_n\}$ be the basis for $\mathcal{P}$ dual to $\mathcal{N}$, that is, $N_i(\psi_j) = \delta_{ij}$. It is called the *nodal basis* for $\mathcal{P}$.

It is clear that the following expansion holds for any $v \in \mathcal{P}$:

$$v(x) = \sum_{i=1}^{n} N_i(v)\psi_i(x).$$

Denote by P_k the set of polynomials of degree $\leqslant k$.

EXAMPLE 2.4 (The linear element). *Let K be a simplex in $\mathbb{R}^d$ with vertices A_i $(i = 1, \cdots, d+1), \mathcal{P} = P_1$, and $\mathcal{N} = \{N_1, \cdots, N_{d+1}\}$, where $N_i(v) = v(A_i)$ for any $v \in \mathcal{P}$. Then $(K, \mathcal{P}, \mathcal{N})$ is a finite element.*

The nodal basis $\{\lambda_1(x), \cdots, \lambda_{d+1}(x)\}$ of the linear element satisfies

$$\lambda_i(x) \text{ is linear and } \lambda_i(A_j) = \delta_{ij}, \quad i, j = 1, \cdots, d+1. \tag{2.7}$$

Given a simplex K in $\mathbb{R}^d$ it is often convenient to consider the associated *barycentric coordinates* defined as the ordered $(d+1)$-tuple $(\lambda_1(x), \lambda_2(x), \cdots, \lambda_{d+1}(x))$, where $\lambda_i(x)$ satisfies (2.7). Let α_i be the Cartesian coordinates of the vertex A_i. We have the following relationship between the Cartesian coordinates and the barycentric coordinates:

$$\sum_{i=1}^{d+1} \lambda_i(x) = 1 \text{ and } x = \sum_{i=1}^{d+1} \alpha_i \lambda_i(x). \tag{2.8}$$

The relationship is obvious since any two linear functions are equal if they coincide at the vertices $A_i, i = 1, \cdots, d+1$. Note that the barycenter of K has barycentric coordinates $(\frac{1}{d+1}, \cdots, \frac{1}{d+1})$.

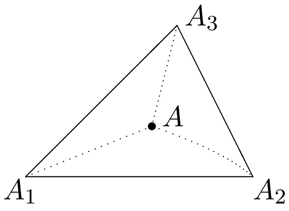

FIGURE 1. The triangle and the barycentric coordinates.

Next we consider a geometric interpretation of the barycentric coordinates in dimension 2. Let the coordinates of A_i be (a_i, b_i) and let

$$S = \frac{1}{2}\begin{vmatrix} a_1 & b_1 & 1 \\ a_2 & b_2 & 1 \\ a_3 & b_3 & 1 \end{vmatrix}$$

be the directional area of the triangle K. $S > 0$ if A_1, A_2, A_3 is ordered counter-clockwise, $S < 0$ otherwise. For any point $A(x_1, x_2)$ in the element K (see Figure 1), by connecting A with three vertices of K, we obtain three triangles. It is clear that $\lambda_i(A)$ is the ratio of areas

$$\lambda_1 = \frac{|\triangle AA_2A_3|}{|\triangle A_1A_2A_3|}, \qquad \lambda_2 = \frac{|\triangle A_1AA_3|}{|\triangle A_1A_2A_3|}, \qquad \lambda_3 = \frac{|\triangle A_1A_3A|}{|\triangle A_1A_2A_3|}.$$

That is

$$\lambda_1 = \frac{1}{2S}\begin{vmatrix} x_1 & x_2 & 1 \\ a_2 & b_2 & 1 \\ a_3 & b_3 & 1 \end{vmatrix}, \lambda_2 = \frac{1}{2S}\begin{vmatrix} a_1 & b_1 & 1 \\ x_1 & x_2 & 1 \\ a_3 & b_3 & 1 \end{vmatrix}, \lambda_3 = \frac{1}{2S}\begin{vmatrix} a_1 & b_1 & 1 \\ a_2 & b_2 & 1 \\ x_1 & x_2 & 1 \end{vmatrix}. (2.9)$$

EXAMPLE 2.5 (The Argyris element). *Let K be a triangle in $\mathbb{R}^2$, $\mathcal{P} = P_5$ of dimension 21, $\mathcal{N}$ = the 21 degrees of freedom shown in Figure 2. "•" denotes the evaluation at that point, the inner circle denotes the evaluation of the gradient at the center, the outer circle denotes the evaluation of three second derivatives at the center, and the arrows represent the evaluation of the normal derivatives at three midpoints.*

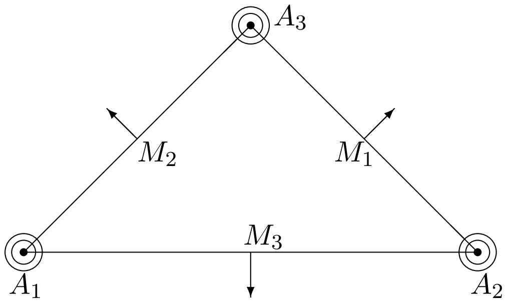

FIGURE 2. The degrees of freedom of the Argyris element

We claim that $\mathcal{N} = \{N_1, N_2, \cdots, N_{21}\}$ determines P_5. Suppose that for some $P \in P_5$, $N_i(P) = 0$, for $i = 1, 2, \cdots, 21$. All we need is to prove $P \equiv 0$. From (2.8), P is a fifth order polynomial in λ_1 and λ_2. Since the edge A_2A_3 is on the line $\lambda_1 = 0$, the restriction of P to A_2A_3 is a fifth order polynomial in λ_2. Moreover,

$$P(A_j) = \frac{\partial P}{\partial \lambda_2}(A_j) = \frac{\partial^2 P}{\partial \lambda_2^2}(A_j) = 0, \quad j = 2, 3.$$

Therefore

$$P\big|_{A_2A_3} = P(0, \lambda_2) = 0. \tag{2.10}$$

On the other hand,

$$\frac{\partial P}{\partial \lambda_1}(A_j) = \frac{\partial}{\partial \lambda_2}\frac{\partial P}{\partial \lambda_1}(A_j) = 0, \quad j = 2, 3.$$

Since $\nabla\lambda_1$ *is parallel to the unit outer normal to* A_2A_3*,* $\frac{\partial P}{\partial \lambda_1}(M_1) = 0$*. Notice that* $\frac{\partial P}{\partial \lambda_1}\big|_{A_2A_3}$ *is a forth order polynomial, we have* $\frac{\partial P}{\partial \lambda_1}|_{A_2A_3} \equiv 0$*, that is,*

$$\frac{\partial P}{\partial \lambda_1}(0, \lambda_2) = 0. \tag{2.11}$$

Combining (2.10) *and* (2.11)*, we have* $P = \lambda_1^2 P_1$*. Similarly,* $P = \lambda_1^2\lambda_2^2\lambda_3^2 Q = 0$*. Since* Q *is a polynomial,* $Q \equiv 0$*, and hence* $P \equiv 0$*.* □

Definition 2.6. Given a finite element $(K, \mathcal{P}, \mathcal{N})$, let the set $\{\psi_i : 1 \leqslant i \leqslant n\} \subset \mathcal{P}$ be the nodal basis of $\mathcal{P}$. If v is a function for which all $N_i \in \mathcal{N}$, $i = 1, \cdots, n$, are defined, then we define the *local interpolant* by

$$I_K v := \sum_{i=1}^{n} N_i(v)\psi_i.$$

It is easy to see that I_K is linear and $I_K u = u$ for $u \in \mathcal{P}$.

Example 2.7 (Lagrange interpolant of linear elements). *Let* $(K, \mathcal{P}, \mathcal{N})$ *be the linear finite element with nodal basis* $\{\psi_i\}$*. The Lagrange interpolant is defined as*

$$(I_K v)(x) := \sum_{i=1}^{d+1} v(x_i)\psi_i(x) \quad \forall v \in C(K).$$

We now piece together the elements.

Definition 2.8. A triangular (tetrahedral) mesh $\mathcal{M}_h$ is a partition of the domain Ω in $\mathbb{R}^d$ ($d = 2, 3$) into a finite collection of triangles (tetrahedrons) $\{K_i\}$ satisfying the following conditions:

(i) The intersection of the interior of any two elements is empty;
(ii) $\cup K_i = \bar{\Omega}$;
(iii) No vertex of any triangle (tetrahedron) lies in the interior of an edge (a face or an edge) of another triangle (tetrahedron).

In this book, a triangular or tetrahedral mesh is called a *triangulation*, or simply a *mesh*.

THEOREM 2.9. *Let Ω be a bounded domain in $\mathbb{R}^d$ and $\mathcal{M}_h = \{K_j\}_{j=1}^J$ be a partition of Ω, that is, $\cup K_i = \bar{\Omega}$ and $K_i \cap K_j = \emptyset$ if $i \neq j$. Assume that ∂K_i $(i = 1, \cdots, J)$ are Lipschitz. Let $k \geqslant 1$. Then a piecewise infinitely differentiable function $v : \bar{\Omega} \to \mathbb{R}$ over the partition $\mathcal{M}_h$ belongs to $H^k(\Omega)$ if and only if $v \in C^{k-1}(\bar{\Omega})$.*

PROOF. We only prove the case $k = 1$. For $k > 1$, the assertion follows from a consideration of the derivatives of order $k - 1$.

Let $v \in C(\bar{\Omega})$. For $i = 1, 2$, define

$$w_i(x) = \frac{\partial v}{\partial x_i} \quad \text{for} \quad x \in \Omega,$$

where on the edges we can take either of the two limiting values. Let $\varphi \in C_0^\infty(\Omega)$,

$$\begin{aligned}\int_\Omega \varphi w_i \, \mathrm{d}x &= \sum_{K \in \mathcal{M}_h} \int_K \varphi \frac{\partial v}{\partial x_i} \, \mathrm{d}x \\ &= \sum_{K \in \mathcal{M}_h} \left(-\int_K v \frac{\partial \varphi}{\partial x_i} \, \mathrm{d}x + \int_{\partial K} \varphi v \cdot n_i \, \mathrm{d}s \right) = -\int_\Omega v \frac{\partial \varphi}{\partial x_i} \, \mathrm{d}x,\end{aligned}$$

where $\mathbf{n}_K = (n_1, \cdots, n_d)^T$ is the unit outward normal to ∂K. This shows that w_i is the weak derivative of v and hence $v \in H^1(\Omega)$.

Conversely, let $v \in H^1(\Omega)$. Let x be on the interior of an edge e shared by two elements K_1, K_2. Then there exists a neighborhood B small enough such that $B \subset K_1 \cup K_2$. Denote by $v_i = v|_{K_i}$, $i = 1, 2$. Then, by Green formula, for all $\boldsymbol{\varphi} \in C_0^\infty(B)^d$,

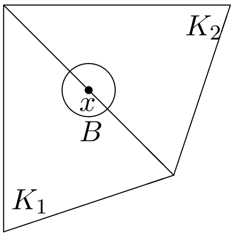

FIGURE 3. The neighborhood of a point x on the common side of two elements K_1 and K_2.

$$\int_{K_i} \nabla v \cdot \boldsymbol{\varphi} \, \mathrm{d}x = -\int_{K_i} v \nabla \cdot \boldsymbol{\varphi} \, \mathrm{d}x + \int_{\partial K_i} v_i (\boldsymbol{\varphi} \cdot \mathbf{n}_{K_i}) \, \mathrm{d}s, \quad i = 1, 2.$$

Since $v \in H^1(\Omega)$, we have

$$\int_\Omega \nabla v \cdot \boldsymbol{\varphi} \, \mathrm{d}x = -\int_\Omega v \nabla \cdot \boldsymbol{\varphi} \, \mathrm{d}x.$$

Thus

$$\int_e (v_1 - v_2)\phi \, \mathrm{d}s = 0 \qquad \forall \, \phi \in C_0^\infty(B).$$

This implies that $v_1 = v_2$ at x. Therefore, v is continuous across any inter-element edges. □

Example 2.10 (Conforming linear element). *Let $(K, \mathcal{P}, \mathcal{N})$ be the linear finite element defined in Example 2.4. Since any piecewise linear function is continuous as long as it is continuous at the vertices, we can introduce*

$$V_h = \{v : v|_K \in P_1, \ \ \forall K \in \mathcal{M}_h, \ v \textit{ is continuous at the vertices of the elements}\}.$$

By Theorem 2.9, $V_h \subset H^1(\Omega)$, V_h is a H^1-conforming *finite element space.*

Example 2.11 (Crouzeix-Raviart element). *Let K be a triangle in $\mathbb{R}^2$, $\mathcal{P} = P_1$ the set of linear polynomials, and $\mathcal{N} = \{N_1, N_2, N_3\}$, where $N_i(v) = v(M_i)$ and M_i, $i = 1, 2, 3$, are the midpoints of three edges. It is easy to see that $(K, \mathcal{P}, \mathcal{N})$ is a finite element. Define*

$$\hat{V}_h = \{v : v|_K \in P_1, \ \ \forall K \in \mathcal{M}_h, \ v \textit{ is continuous at the midpoints of the triangle edges}\}.$$

Then $\hat{V}_h \subset L^2(\Omega)$ but $\hat{V}_h \not\subset H^1(\Omega)$. $\hat{V}_h$ is an example of H^1-nonconforming *finite element spaces.*

Example 2.12. *Let $(K, \mathcal{P}, \mathcal{N})$ be the Argyris element and define*

$$\widetilde{V}_h = \{v : v|_K \in P_5, \ \ \forall \, K \in \mathcal{M}_h, \ v \textit{ and its partial derivatives up to second order are continuous at the vertices of the triangle elements, } v \textit{ has continuous normal derivatives at the midpoints of the triangle edges.}\}.$$

It can be shown that $\widetilde{V}_h \subset C^1(\bar{\Omega})$. Therefore $\widetilde{V}_h \subset H^2(\Omega)$ is a H^2-conforming finite element space.

2.3. Computational consideration

The computation of finite element methods can be divided into three steps:

1. Construction of a mesh by partitioning Ω;
2. Setting up the stiffness matrix;
3. Solution of the system of equations.

In this section we consider the computation of the stiffness matrix. The solution of the system of equations will be treated in Chapter 5. The construction of the mesh will be discussed in Chapter 4 and Chapter 10.

We will only consider conforming linear finite element approximations to elliptic equations of second order. In this case, the stiffness matrix can be assembled elementwise. For simplicity, we consider only the principal part

$$a(u,v) = \int_\Omega \sum_{k,l=1}^{d} a_{kl}(x)\frac{\partial u}{\partial x_l}\frac{\partial v}{\partial x_k}\,\mathrm{d}x = \int_\Omega \mathbf{a}(x)\nabla u\cdot\nabla v\,\mathrm{d}x,$$

where $\mathbf{a}(x) = \big(a_{kl}(x)\big)_{d\times d}$. Let $\{\phi_j\}_{j=1}^J$ be a nodal basis of the linear finite element space $V_h^0 = V_h \cap H_0^1(\Omega)$ so that $\phi_i(x_j) = \delta_{ij},\ i,j = 1,\cdots,J$, where $\{x_j\}_{j=1}^J$ is the set of interior nodes of the mesh $\mathcal{M}_h$. Then

$$A_{ij} = a(\phi_j,\phi_i) = \sum_{K\in\mathcal{M}_h}\int_K \mathbf{a}(x)\nabla\phi_j\cdot\nabla\phi_i\,\mathrm{d}x. \tag{2.12}$$

In forming the sum, we need only take account of those triangles which overlap the support of both ϕ_i and ϕ_j. Note that $A_{ij} = 0$ if the x_i and x_j are not adjacent. The stiffness matrix $A = (A_{ij})$ is sparse.

In practice, for every element $K \in \mathcal{M}_h$, we find the additive contribution from (2.12) to the stiffness matrix. Since on each element K, the nodal basis function reduces to one of the barycentric coordinate functions $\lambda_p, p = 1,2,\cdots d+1$. Thus we need only to evaluate the following $(d+1)\times(d+1)$ matrix

$$A_K:\ (A_K)_{pq} = \int_K \mathbf{a}(x)\nabla\lambda_q\cdot\nabla\lambda_p\,\mathrm{d}x. \tag{2.13}$$

Here A_K is called the *element stiffness matrix*. Denote by K_p the global index of the p-th vertex of the element K. Then $\phi_{K_p}|_K = \lambda_p$ and the global stiffness matrix may be assembled through the element stiffness matrices as

$$A_{ij} = \sum_{\substack{K,p,q\\ K_p=i,K_q=j}} (A_K)_{pq}\ . \tag{2.14}$$

For example, consider two adjacent nodal points x_i and x_j on a triangular mesh as shown in Figure 4. Suppose the local indices of the vertices of the elements K^{I}, K^{II}, ..., K^{VI} are labeled as in the same figure. Then

$$K_2^{\mathrm{I}} = K_1^{\mathrm{II}} = K_3^{\mathrm{III}} = K_2^{\mathrm{IV}} = K_1^{\mathrm{V}} = K_3^{\mathrm{VI}} = i, \quad K_3^{\mathrm{I}} = K_2^{\mathrm{VI}} = j.$$

and (2.14) implies that

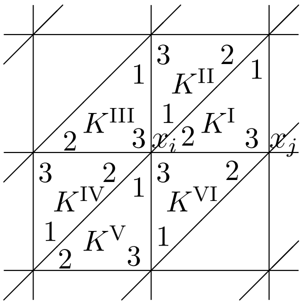

FIGURE 4. Global and local indices.

$$\begin{aligned} A_{ij} &= (A_{K^{\mathrm{I}}})_{23} + (A_{K^{\mathrm{VI}}})_{32}, \\ A_{ii} &= (A_{K^{\mathrm{I}}})_{22} + (A_{K^{\mathrm{II}}})_{11} + (A_{K^{\mathrm{III}}})_{33} + (A_{K^{\mathrm{IV}}})_{22} + (A_{K^{\mathrm{V}}})_{11} + (A_{K^{\mathrm{VI}}})_{33}. \end{aligned}$$

The computation of (2.13) can be simplified by using the following property of the conforming linear finite element space.

THEOREM 2.13. *The conforming linear finite element space V_h associated with a triangulation $\mathcal{M}_h$ of $\Omega \subset \mathbb{R}^d$ is an* affine family *in the sense that there exists a finite element $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ called the* reference finite element *with the following properties: For every $K \in \mathcal{M}_h$, there exists an affine mapping $F_K : \hat{K} \to K$ such that for every $v \in V_h$, its restriction to K has the form*

$$v(x) = \hat{v}(F_K^{-1}(x)) \quad \textit{with some} \quad \hat{v} \in \hat{\mathcal{P}}.$$

PROOF. It is obvious. We only need to set

$$\hat{K} = \left\{ \hat{x} = (\xi_1, \cdots, \xi_d) \in \mathbb{R}^d : \ \xi_1, \cdots, \xi_d \geqslant 0, \ 1 - \sum_{i=1}^{d} \xi_i \geqslant 0 \right\},$$

$\hat{\mathcal{P}} = P_1$, and $\hat{\mathcal{N}} = \{\hat{N}_i, i = 1, \cdots, d+1\}$, where $\hat{N}_i(p) = p(\hat{A}_i)$ for any $p \in \hat{\mathcal{P}}$, where $\{\hat{A}_i\}$ is the set of vertices of $\hat{K}$. □

To compute (2.13), we transform the element K into the reference element $\hat{K}$. Let

$$x = F_K\hat{x} = B_K\hat{x} + b_K$$

be the corresponding linear mapping. Then

$$(A_K)_{pq} = \frac{|K|}{|\hat{K}|}\int_{\hat{K}} \hat{\mathbf{a}}(\hat{x})(B_K^{-T}\hat{\nabla}\hat{\lambda}_q)\cdot(B_K^{-T}\hat{\nabla}\hat{\lambda}_p)\,\mathrm{d}\hat{x}. \tag{2.15}$$

Here $|K|$ is the measure of K and $|\hat{K}|$ is the measure of $\hat{K}$.

EXAMPLE 2.14. *Let K be a triangle element with vertices $A_i(a_i, b_i), i = 1, 2, 3$, and let F_K be the affine mapping defined by $\hat{x} = F_K^{-1}(x)$ (cf. (2.9)):*

$$\hat{x}_1 = \lambda_1(x) = \frac{1}{2|K|}\big((b_2 - b_3)x_1 + (a_3 - a_2)x_2 + a_2b_3 - a_3b_2\big),$$
$$\hat{x}_2 = \lambda_2(x) = \frac{1}{2|K|}\big((b_3 - b_1)x_1 + (a_1 - a_3)x_2 + a_3b_1 - a_1b_3\big).$$

It is clear that F_K^{-1} maps the points A_1, A_2, and A_3 in the x_1x_2 plane to the points $(1,0), (0,1)$, and $(0,0)$ in the $\hat{x}_1\hat{x}_2$ plane, respectively. Obviously,

$$B_K^{-T} = \frac{1}{2|K|}\begin{pmatrix} b_2 - b_3 & b_3 - b_1 \\ a_3 - a_2 & a_1 - a_3 \end{pmatrix}.$$

Noting that $\hat{\lambda}_i(\hat{x}) = \lambda_i(x)$ and $\lambda_1 + \lambda_2 + \lambda_3 = 1$, we have

$$\hat{\nabla}\hat{\lambda}_1 = \begin{pmatrix}1\\0\end{pmatrix}, \quad \hat{\nabla}\hat{\lambda}_2 = \begin{pmatrix}0\\1\end{pmatrix}, \quad \hat{\nabla}\hat{\lambda}_3 = \begin{pmatrix}-1\\-1\end{pmatrix}.$$

Then the element stiffness matrix A_K can be computed by using (2.15) *and the global stiffness matrix can be assembled by using* (2.14).

For partial differential equations with variable coefficients, the evaluation of the integrals (2.15) is usually accomplished by using a quadrature formula. Some examples of quadrature formulas on the two-dimensional reference element (see Figure. 5) are as follows. The quadrature formula

$$\int_{\hat{K}} \hat{\varphi}(\hat{x})\mathrm{d}\hat{\mathrm{x}} \sim |\hat{K}|\hat{\varphi}(\hat{a}_{123}), \tag{2.16}$$

is exact for polynomials of degree $\leqslant 1$, i.e.,

$$\int_{\hat{K}} \hat{\varphi}(\hat{x})\,\mathrm{d}\hat{x} = |\hat{K}|\hat{\varphi}(\hat{a}_{123}) \qquad \forall\,\hat{\varphi} \in P_1(\hat{K}).$$

Here $\hat{a}_{123}$ is the barycenter of $\hat{K}$.

The quadrature formula

$$\int_{\hat{K}} \hat{\varphi}(\hat{x}) \mathrm{d}\hat{\mathrm{x}} \sim \frac{|\hat{K}|}{3} \sum_{1 \leqslant i < j \leqslant 3} \hat{\varphi}(\hat{a}_{ij}), \tag{2.17}$$

is exact for polynomials of degree $\leqslant 2$. Here $\hat{a}_{12}, \hat{a}_{23}$, and $\hat{a}_{13}$ are the mid-edge points of $\hat{K}$.

The quadrature formula

$$\int_{\hat{K}} \hat{\varphi}(\hat{x}) \mathrm{d}\hat{\mathrm{x}} \sim \frac{|\hat{K}|}{60} \left(3 \sum_{i=1}^{3} \hat{\varphi}(\hat{a}_i) + 8 \sum_{1 \leqslant i < j \leqslant 3} \hat{\varphi}(\hat{a}_{ij}) + 27 \hat{\varphi}(\hat{a}_{123}) \right) \tag{2.18}$$

is exact for polynomials of degree $\leqslant 3$.

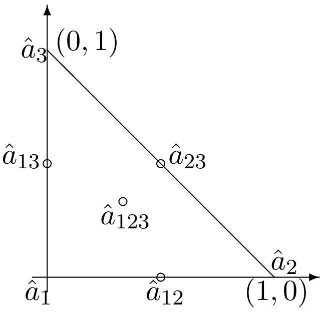

FIGURE 5. The reference element $\hat{K}$ for the quadrature formulas (2.16),(2.17), and (2.18).

Table 1 shows the sample points (ξ_i, η_i) and weights for *Gaussian quadrature* formula which is exact for polynomials of degree $\leqslant 5$

$$\int_{\hat{K}} \hat{\varphi}(\hat{x}) \mathrm{d}\hat{\mathrm{x}} \sim \sum_{i=1}^{7} w_i \hat{\varphi}(\xi_i, \eta_i). \tag{2.19}$$

Bibliographic notes. The material in this chapter is classical. We refer to the book of Ciarlet [**23**] for further information on the construction and computation of finite elements. The quadrature formulas (2.19) in Section 2.3 is taken from Braess [**11**].

2.4. Exercises

EXERCISE 2.1. Construct the nodal basis functions for the Crouzeix-Raviort element using barycentric coordinates.

i	ξ_i	η_i	w_i
1	$1/3$	$1/3$	$9/80$
2	$(6+\sqrt{15})/21$	$(6+\sqrt{15})/21$	$\dfrac{155+\sqrt{15}}{2400}$
3	$(9-2\sqrt{15})/21$	$(6+\sqrt{15})/21$	
4	$(6+\sqrt{15})/21$	$(9-2\sqrt{15})/21$	
5	$(6-\sqrt{15})/21$	$(6-\sqrt{15})/21$	$\dfrac{155-\sqrt{15}}{2400}$
6	$(9+2\sqrt{15})/21$	$(6-\sqrt{15})/21$	
7	$(6-\sqrt{15})/21$	$(9+2\sqrt{15})/21$	

TABLE 1. The sample points (ξ_i, η_i) and weights for the seven-point Gaussian quadrature rule over the reference element $\hat{K}$.

EXERCISE 2.2. Show that the finite element space based on the Argyris element is a subspace of C^1 and thus is indeed H^2-conforming.

EXERCISE 2.3. Show the quadrature scheme (2.17) is exact for polynomials of degree $\leqslant 2$.

EXERCISE 2.4. Let K be a triangle in $\mathbb{R}^2$. Compute the element mass matrix

$$M_K = \left(\int_K \lambda_i \lambda_j \,\mathrm{d}x \right)_{i,j=1}^{3}.$$

EXERCISE 2.5. Let K be a triangle in $\mathbb{R}^2$. Show that

$$\int_K \lambda_1^p \lambda_2^q \lambda_3^r \,\mathrm{d}x = 2|K| \frac{p!q!r!}{(p+q+r+2)!}, \quad p, q, r \geqslant 0, \text{integer}.$$

EXERCISE 2.6. Consider the Poisson equation $-\Delta u = 1$ on the unit square with homogeneous Dirichlet boundary condition. Compute the stiffness matrix of the linear finite element method on the standard triangulation of the unit square constructed by first dividing the unit square into n^2 subsquares of equal size and then connecting the southwest-to-northeast diagonal of each subsquare. Here n is an positive integer. Compare the stiffness matrix with the coefficient matrix of the five-point difference equations.

CHAPTER 3

Convergence Theory of Finite Element Methods

In this chapter we consider the convergence of the finite element method for solving elliptic equations. From Céa lemma in Theorem 2.1 we know that the error of the finite element solution is bounded by $\inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)}$. This quantity will be estimated by the scaling argument that we develop in the first section.

3.1. Interpolation theory in Sobolev spaces

We start from the following result which plays an important role in the error analysis of finite element methods. This result generalizes the second Poincaré-Friedrichs inequality in Theorem 1.10.

Theorem 3.1 (Deny-Lions). *Let Ω be a bounded Lipschitz domain. For any $k \geqslant 0$, there exists a constant $C(\Omega)$ such that*

$$\inf_{p \in P_k(\Omega)} \|v + p\|_{H^{k+1}(\Omega)} \leqslant C(\Omega)|v|_{H^{k+1}(\Omega)} \quad \forall v \in H^{k+1}(\Omega). \tag{3.1}$$

Here $P_k(\Omega)$ is the set of polynomials over Ω with degree $\leqslant k$.

Proof. Let $N = \dim P_k(\Omega)$ and let $f_i, 1 \leqslant i \leqslant N$, be a basis of the dual space of $P_k(\Omega)$. Using the Hahn-Banach extension theorem, there exist continuous linear forms over the space $H^{k+1}(\Omega)$, again denote by $f_i, 1 \leqslant i \leqslant N$, such that for any $p \in P_k(\Omega)$, $f_1(p) = f_2(p) = \cdots = f_N(p) = 0$ if and only if $p = 0$. We will show that there exists a constant $C(\Omega)$ such that

$$\|v\|_{H^{k+1}(\Omega)} \leqslant C(\Omega)\Big(|v|_{H^{k+1}(\Omega)} + \sum_{i=1}^{N} |f_i(v)|\Big) \quad \forall v \in H^{k+1}(\Omega). \tag{3.2}$$

(3.1) is a direct consequence of (3.2) because for any $v \in H^{k+1}(\Omega)$, there exists a $p \in P_k(\Omega)$ such that $f_i(p) = -f_i(v), 1 \leqslant i \leqslant N$.

If inequality (3.2) is false, there exists a sequence $\{v_n\}$ of functions $v_n \in H^{k+1}(\Omega)$ such that

$$\|v_n\|_{H^{k+1}(\Omega)} = 1, \quad |v_n|_{H^{k+1}(\Omega)} + \sum_{i=1}^{N} |f_i(v_n)| \leqslant \frac{1}{n}. \tag{3.3}$$

Since $\{v_n\}$ is bounded in $H^{k+1}(\Omega)$, by the compactness imbedding theorem, there exists a subsequence, again denoted by $\{v_n\}$, and a function $v \in H^k(\Omega)$ such that

$$\|v_n - v\|_{H^k(\Omega)} \to 0 \qquad \text{as} \qquad n \to \infty. \tag{3.4}$$

Since, by (3.3), $|v_n|_{H^{k+1}(\Omega)} \to 0$, and since the space $H^{k+1}(\Omega)$ is complete, we conclude from (3.4) , that the sequence $\{v_n\}$ converges in $H^{k+1}(\Omega)$. The limit v of this sequence satisfies

$$|v|_{H^{k+1}(\Omega)} + \sum_{i=1}^{N} |f_i(v)| = 0.$$

Thus, it follows from Lemma 1.2 that $v \in P_k(\Omega)$, and hence $v = 0$. But this contradicts the equality $\|v\|_{H^{k+1}(\Omega)} = 1$. □

A direct consequence of this theorem is the following lemma which is called Bramble-Hilbert Lemma in the literature.

LEMMA 3.1 (Bramble-Hilbert). *Let Ω be a bounded Lipschitz domain and let $k \geqslant 0$ be an integer. Denote by $X = H^{k+1}(\Omega)$. Let Y be a Banach space and let $f \in \mathcal{L}(X, Y)$ be a continuous linear operator from X to Y such that $f(p) = 0$ for any $p \in P_k(\Omega)$. Then there exists a constant $C(\Omega)$ such that*

$$\|f(v)\|_Y \leqslant C(\Omega)\|f\|_{\mathcal{L}(X,Y)}|v|_{H^{k+1}(\Omega)} \quad \forall\ v \in H^{k+1}(\Omega),$$

where $\|\cdot\|_{\mathcal{L}(X,Y)}$ is the operator norm.

The error analysis of the finite element method depends on the scaling argument which makes use of the relation of Sobolev norms under the affine transform.

LEMMA 3.2. *Let Ω and $\hat{\Omega} \subset \mathbb{R}^d$ be affine equivalent, i.e., there exists a bijective affine mapping*

$$F : \hat{\Omega} \to \Omega, \quad F\hat{x} = B\hat{x} + b$$

with a nonsingular matrix B. If $v \in H^m(\Omega)$, then $\hat{v} = v \circ F \in H^m(\hat{\Omega})$, and there exists a constant $C = C(m, d)$ such that

$$|\hat{v}|_{H^m(\hat{\Omega})} \leqslant C\|B\|^m |\det B|^{-1/2} |v|_{H^m(\Omega)},$$
$$|v|_{H^m(\Omega)} \leqslant C\|B^{-1}\|^m |\det B|^{1/2} |\hat{v}|_{H^m(\hat{\Omega})}.$$

Here $\|\cdot\|$ denotes the matrix norm associated with the Euclidean norm in $\mathbb{R}^d$.

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

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

From the chain rule, we have

$$\begin{aligned}
\hat{D}^m \hat{v}(\hat{x})(\hat{y}_1, \cdots, \hat{y}_m) &= \sum_{1 \leqslant i_1, \cdots, i_m \leqslant d} \hat{y}_{i_1 1} \cdots \hat{y}_{i_m m} \hat{\partial}_{i_1} \cdots \hat{\partial}_{i_m} \hat{v}(\hat{x}) \\
&= \sum_{1 \leqslant i_1, \cdots, i_m \leqslant d} \sum_{1 \leqslant j_1, \cdots, j_m \leqslant d} \hat{y}_{i_1 1} \cdots \hat{y}_{i_m m} b_{j_1 i_1} \cdots b_{j_m i_m} \partial_{j_1} \cdots \partial_{j_m} v(x) \\
&= \sum_{1 \leqslant j_1, \cdots, j_m \leqslant d} \sum_{1 \leqslant i_1, \cdots, i_m \leqslant d} b_{j_1 i_1} \hat{y}_{i_1 1} \cdots b_{j_m i_m} \hat{y}_{i_m m} \partial_{j_1} \cdots \partial_{j_m} v(x) \\
&= D^m v(x)(B\hat{y}_1, \cdots, B\hat{y}_m).
\end{aligned}$$

Thus

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

where

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

Apply this estimate to the partial derivatives $\partial_{i_1} \partial_{i_2} \cdots \partial_{i_m} v = D^m v(e_{i_1}, \cdots, e_{i_m})$ to get

$$\begin{aligned}
\sum_{|\alpha|=m} |\hat{\partial}^\alpha \hat{v}|^2 &\leqslant d^m \max_{|\alpha|=m} |\hat{\partial}^\alpha \hat{v}|^2 \leqslant d^m \|\hat{D}^m \hat{v}\|^2_{\mathcal{L}^m} \leqslant d^m \|B\|^{2m} \|D^m v\|^2_{\mathcal{L}^m} \\
&\leqslant d^{2m} \|B\|^{2m} \sum_{|\alpha|=m} |\partial^\alpha v|^2.
\end{aligned}$$

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

$$\int_{\hat{\Omega}} \sum_{|\alpha|=m} |\hat{\partial}^\alpha \hat{v}|^2 \mathrm{d}\hat{x} \leqslant d^{2m} \|B\|^{2m} \int_\Omega \sum_{|\alpha|=m} |\partial^\alpha v|^2 |\det B^{-1}| \mathrm{d}x.$$

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

To apply Lemma 3.2, it is desirable to estimate $\|B\|$ and $\|B^{-1}\|$ in terms of simple geometric quantities.

Lemma 3.3. *Let Ω and $\hat{\Omega}$ be affine equivalent with*

$$F:\ \hat{x}\in\mathbb{R}^d \mapsto B\hat{x}+b\in\mathbb{R}^d$$

being an invertible affine mapping. Then the upper bounds

$$\|B\|\leqslant\frac{h}{\hat{\rho}},\quad \|B^{-1}\|\leqslant\frac{\hat{h}}{\rho},\quad \Big(\frac{\rho}{\hat{h}}\Big)^d\leqslant|\det B|\leqslant\Big(\frac{h}{\hat{\rho}}\Big)^d \tag{3.5}$$

hold, where $h=\operatorname{diam}(\Omega),\hat{h}=\operatorname{diam}(\hat{\Omega}),\rho$ and $\hat{\rho}$ are the maximum diameter of the ball contained in Ω and $\hat{\Omega}$, respectively.

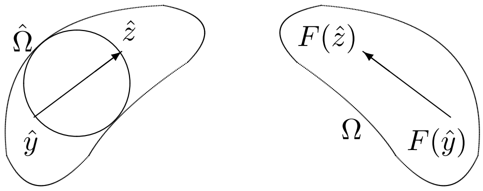

Figure 1. The affine mapping between Ω and $\hat{\Omega}$.

Proof. We may write

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

Given $\xi\in\mathbb{R}^d$ so that $|\xi|=\hat{\rho}$, there exist $\hat{y},\hat{z}\in\hat{\Omega}$ such that $\hat{y}-\hat{z}=\xi$ (see Figure 1). $B\xi=F(\hat{y})-F(\hat{z})$ with $F(\hat{y}),F(\hat{z})\in\Omega$. We deduce $|B\xi|\leqslant h$. This proves the first inequality in (3.5). The second inequality can be proved similarly. The last two inequalities are consequences of the identity $|\det B|=|\Omega|\,/|\hat{\Omega}|$. □

Theorem 3.2. *Suppose $m-d/2>l$. Let $(\hat{K},\hat{\mathcal{P}},\hat{\mathcal{N}})$ be a finite element satisfying*

(i) $\mathcal{P}_{m-1}\subset\hat{\mathcal{P}}\subset H^m(\hat{K})$*;*

(ii) $\hat{\mathcal{N}}\subset C^l(\hat{K})'$.

Then for $0\leqslant i\leqslant m$ and $\hat{v}\in H^m(\hat{K})$ we have

$$|\hat{v}-\hat{I}\hat{v}|_{H^i(\hat{K})}\leqslant C(m,d,\hat{K})|\hat{v}|_{H^m(\hat{K})},$$

where $\hat{I}$ is the local interpolation operator of the finite element defined in Definition 2.6.

Proof. We first prove $\hat{I}$ is bounded from $H^m(\hat{K})$ to $H^i(\hat{K})$. Let $\hat{\mathcal{N}} = \{\hat{N}_1, \cdots, \hat{N}_n\}$ and let $\{\hat{\phi}_1, \cdots, \hat{\phi}_n\}$ be the dual basis. Then

$$\begin{aligned}
\|\hat{I}\hat{u}\|_{H^i(\hat{K})} &= \left\| \sum_{j=1}^{n} \hat{N}_j(\hat{u})\hat{\phi}_j \right\|_{H^i(\hat{K})} \leqslant \sum_{j=1}^{n} |\hat{N}_j(\hat{u})| \|\hat{\phi}_j\|_{H^i(\hat{K})} \\
&\leqslant \sum_{j=1}^{n} \|\hat{N}_j\|_{C^l(\hat{K})'} \|\hat{\phi}_j\|_{H^m(\hat{K})} \|\hat{u}\|_{C^l(\hat{K})} \\
&\leqslant C \|\hat{u}\|_{C^l(\hat{K})} \leqslant C \|\hat{u}\|_{H^m(\hat{K})} .
\end{aligned}$$

Here we have used the Sobolev Imbedding Theorem 1.8 in the last inequality. Next by Theorem 3.1

$$\begin{aligned}
|\hat{v} - \hat{I}\hat{v}|_{H^i(\hat{K})} &= \inf_{\hat{p} \in P_{m-1}} |\hat{v} - \hat{p} - \hat{I}(\hat{v} - \hat{p})|_{H^i(\hat{K})} \\
&\leqslant C \inf_{\hat{p} \in P_{m-1}} \|\hat{v} - \hat{p}\|_{H^m(\hat{K})} \leqslant C |\hat{v}|_{H^m(\hat{K})} .
\end{aligned}$$

This completes the proof of the theorem. □

Definition 3.3. Let $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ be a finite element and $x = F(\hat{x}) = B\hat{x} + b$ be an affine map. Let $v = \hat{v} \circ F^{-1}$. The finite element $(K, \mathcal{P}, \mathcal{N})$ is *affine-interpolation equivalent* to $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ if

(i) $K = F(\hat{K})$;
(ii) $\mathcal{P} = \{p : \hat{p} \in \hat{\mathcal{P}}\}$;
(iii) $\widehat{Iv} = \hat{I}\hat{v}$.

Here $\hat{I}\hat{v}$ and Iv are the $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$-interpolant and the $(K, \mathcal{P}, \mathcal{N})$-interpolant, respectively.

Definition 3.4. A family of meshes $\{\mathcal{M}_h\}$ is called *regular* or *shape regular* provided there exists a number $\kappa > 0$ such that each $K \in \mathcal{M}_h$ contains a ball of diameter ρ_K with $\rho_K \geqslant h_K/\kappa$.

Theorem 3.5. *Let $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ satisfy the conditions of Theorem 3.2 and let $(K, \mathcal{P}, \mathcal{N})$ be affine-interpolation equivalent to $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$. Then for $0 \leqslant i \leqslant m$ and $v \in H^m(K)$ we have*

$$|v - Iv|_{H^i(K)} \leqslant C h_K^{m-i} |v|_{H^m(K)} ,$$

where C depends on $m, d, \hat{K}$, and h_K/ρ_K.

Proof. By Lemmas 3.2–3.3 and Theorem 3.2 we have

$$\begin{aligned}|v - Iv|_{H^i(K)} &\leqslant C\|B^{-1}\|^i |\det B|^{1/2} \left|\widehat{v - Iv}\right|_{H^i(\hat{K})} \\ &= C\|B^{-1}\|^i |\det B|^{1/2} \left|\hat{v} - \hat{I}\hat{v}\right|_{H^i(\hat{K})} \\ &\leqslant C\|B^{-1}\|^i |\det B|^{1/2} |\hat{v}|_{H^m(\hat{K})} \leqslant C\|B^{-1}\|^i \|B\|^m |v|_{H^m(K)} \\ &\leqslant C\Big(\frac{\hat{h}_{\hat{K}}}{\rho_K}\Big)^i \Big(\frac{h_K}{\hat{\rho}_{\hat{K}}}\Big)^m |v|_{H^m(K)} \leqslant C\frac{\hat{h}^i_{\hat{K}}}{\hat{\rho}^m_{\hat{K}}}\Big(\frac{h_K}{\rho_K}\Big)^i h_K^{m-i} |v|_{H^m(K)}.\end{aligned}$$

This completes the proof. □

As a consequence of the above theorem we have the following estimate.

Theorem 3.6. *Suppose $\{\mathcal{M}_h\}$ is a regular family of meshes of a polyhedral domain $\Omega \subset \mathbb{R}^d$. Let $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ be a reference finite element satisfying the conditions of Theorem 3.2 for some l and m. For all $K \in \mathcal{M}_h$, suppose $(K, \mathcal{P}_K, \mathcal{N}_K)$ is affine-interpolation equivalent to $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$. Then for $0 \leqslant i \leqslant m$, there exists a positive constant $C(\hat{K}, d, m, \kappa)$ such that*

$$\Big(\sum_{K\in\mathcal{M}_h} \|v - I_h v\|^2_{H^i(K)}\Big)^{1/2} \leqslant C h^{m-i} |v|_{H^m(\Omega)}, \ h = \max_{K\in\mathcal{M}_h} h_K, \ \forall\, v \in H^m(\Omega),$$

where $I_h v$ is the global interpolant *defined by $I_h v|_K = I_K v$ for all $K \in \mathcal{M}_h$.*

Now we consider the inverse estimates which are useful in the error analysis of finite element methods. We first introduce the quasi-uniform meshes.

Definition 3.7. A family of meshes $\{\mathcal{M}_h\}$ is called *quasi-uniform* if there exists a constant ν such that

$$h/h_K \leqslant \nu \qquad \forall\, K \in \mathcal{M}_h,$$

where $h = \max_{K\in\mathcal{M}_h} h_K$.

Theorem 3.8. *Let $\{\mathcal{M}_h\}$ be a shape-regular quasi-uniform family of triangulations of Ω and let X_h be a finite element space of piecewise polynomials of degree less than or equal to p. Then for $m \geqslant l \geqslant 0$, there exists a constant $C = C(p, \kappa, \nu, m)$ such that for any $v_h \in X_h$,*

$$\left(\sum_{K\in\mathcal{M}_h} |v_h|^2_{H^m(K)}\right)^{1/2} \leqslant C h^{l-m} \left(\sum_{K\in\mathcal{M}_h} |v_h|^2_{H^l(K)}\right)^{1/2}.$$

Proof. From Lemma 3.2, we have

$$|v|_{H^m(K)} \leqslant C\|B_K^{-1}\|^m|\mathrm{det}B_K|^{1/2}|\hat{v}|_{H^m(\hat{K})}.$$

Since

$$\begin{aligned}|\hat{v}|_{H^m(\hat{K})} &= \inf_{p\in P_{l-1}(\hat{K})} |\hat{v}+p|_{H^m(\hat{K})} \leqslant \inf_{p\in P_{l-1}(\hat{K})} \|\hat{v}+p\|_{H^m(\hat{K})} \\ &\leqslant C \inf_{p\in P_{l-1}(\hat{K})} \|\hat{v}+p\|_{H^l(\hat{K})} \leqslant C|\hat{v}|_{H^l(\hat{K})},\end{aligned}$$

it follows from Lemma 3.2 and Lemma 3.3 that

$$\begin{aligned}|v|_{H^m(K)} &\leqslant C\|B_K^{-1}\|^m\|B_K\|^l|v|_{H^l(K)} \\ &\leqslant C\left(\frac{\hat{h}_{\hat{K}}}{\rho_K}\right)^m\left(\frac{h_K}{\hat{\rho}_{\hat{K}}}\right)^l|v|_{H^l(K)} \leqslant Ch_K^{l-m}|v|_{H^l(K)}.\end{aligned} \tag{3.6}$$

This completes the proof. □

From (3.6) we have the following *local inverse estimates* on an element K:

$$|v|_{H^m(K)} \leqslant Ch_K^{l-m}|v|_{H^l(K)}, \quad \forall m \geqslant l \geqslant 0,\ v\in P_p(K), \tag{3.7}$$

where C depends only on p, m, and the shape regularity of the element K.

3.2. The energy error estimate

Let Ω be a polyhedral domain in $\mathbb{R}^d$ and $\{\mathcal{M}_h\}$ be a regular family of triangulations of the domain. Let V_h be the piecewise linear conforming finite element space over $\mathcal{M}_h$. Denote $V_h^0 = V_h \cap H_0^1(\Omega)$. Let $u \in H_0^1(\Omega)$ be the weak solution of the variational problem

$$a(u,v) = \langle f, v\rangle \qquad \forall\, v\in H_0^1(\Omega), \tag{3.8}$$

and $u_h \in V_h^0$ be the corresponding finite element solution

$$a(u_h, v_h) = \langle f, v_h\rangle \quad \forall\, v_h \in V_h^0. \tag{3.9}$$

We assume the bilinear form $a : H_0^1(\Omega)\times H_0^1(\Omega) \to \mathbb{R}$ is bounded and $H_0^1(\Omega)$-elliptic:

$$|a(u,v)| \leqslant \beta\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)}, \quad a(u,u) \geqslant \alpha\|u\|^2_{H^1(\Omega)}, \quad \forall u,v\in H_0^1(\Omega).$$

Then we know from Lax-Milgram Lemma that (3.8) and (3.9) have a unique solution u, u_h, respectively.

THEOREM 3.9. *If the solution $u \in H_0^1(\Omega)$ has the regularity $u \in H^2(\Omega)$, then there exists a constant C independent of h such that*

$$\|u - u_h\|_{H^1(\Omega)} \leqslant Ch|u|_{H^2(\Omega)}.$$

PROOF. By Céa lemma and the finite element interpolation estimate in Theorem 3.6,

$$\|u - u_h\|_{H^1(\Omega)} \leqslant C \inf_{v_h \in V_h^0} \|u - v_h\|_{H^1(\Omega)} \leqslant C\|u - I_h u\|_{H^1(\Omega)} \leqslant Ch|u|_{H^2(\Omega)}.$$

□

If the solution of the problem (3.8) does not in $H^2(\Omega)$, we still have the convergence of finite element methods.

THEOREM 3.10 (Convergence). *If the solution u only belongs to $H_0^1(\Omega)$, we still have*

$$\lim_{h \to 0} \|u - u_h\|_{H^1(\Omega)} = 0.$$

PROOF. We only need to prove

$$\lim_{h \to 0} \inf_{v_h \in V_h^0} \|u - v_h\|_{H^1(\Omega)} = 0.$$

For $u \in H_0^1(\Omega)$ and any $\epsilon > 0$, there exists a function $v_\epsilon \in C_0^\infty(\Omega)$ such that

$$\|u - v_\epsilon\|_{H^1(\Omega)} \leqslant \epsilon.$$

On the other hand, by the interpolation estimate in Theorem 3.6,

$$\|v_\epsilon - I_h v_\epsilon\|_{H^1(\Omega)} \leqslant Ch|v_\epsilon|_{H^2(\Omega)}.$$

Thus

$$\inf_{v_h \in V_h^0} \|u - v_h\|_{H^1(\Omega)} \leqslant \|u - I_h v_\epsilon\|_{H^1(\Omega)} \leqslant \epsilon + Ch|v_\epsilon|_{H^2(\Omega)}.$$

By letting $h \to 0$ we get

$$\overline{\lim}_{h \to 0} \inf_{v_h \in V_h^0} \|u - v_h\|_{H^1(\Omega)} \leqslant \epsilon \quad \text{for any} \quad \epsilon > 0.$$

This completes the proof. □

3.3. The L^2 error estimate

We assume the adjoint variational problem of (3.8) is regular in the following sense

(i) For any $g \in L^2(\Omega)$, the problem

$$a(v, \varphi_g) = (g, v) \qquad \forall\ v \in H_0^1(\Omega) \tag{3.10}$$

attains a unique solution $\varphi_g \in H^2(\Omega) \cap H_0^1(\Omega)$;

(ii) There exists a constant C such that

$$\|\varphi_g\|_{H^2(\Omega)} \leqslant C\|g\|_{L^2(\Omega)}.$$

THEOREM 3.11. *Assume the solution of problem* (3.8) $u \in H^2(\Omega)$ *and the adjoint problem is regular. Then there exists a constant C independent of h such that*

$$\|u - u_h\|_{L^2(\Omega)} \leqslant Ch^2|u|_{H^2(\Omega)}.$$

PROOF. For $g = u - u_h$, let φ_g be the solution of (3.10). Then

$$\begin{aligned}(u - u_h, g) = a(u - u_h, \varphi_g) &= a(u - u_h, \varphi_g - I_h\varphi_g)\\ &\leqslant \beta\|u - u_h\|_{H^1(\Omega)}\|\varphi_g - I_h\varphi_g\|_{H^1(\Omega)}\\ &\leqslant Ch^2|u|_{H^2(\Omega)}|\varphi_g|_{H^2(\Omega)}\\ &\leqslant Ch^2\|g\|_{L^2(\Omega)}|u|_{H^2(\Omega)}.\end{aligned}$$

This completes the proof. □

The argument used in the proof of Theorem 3.11 is termed Aubin-Nitsche trick in the literature.

Bibliographic notes. The results in this chapter are taken for Ciarlet [**23**] to which we refer for further developments in the finite element a priori error analysis. The Deny-Lions Theorem is from [**26**]. The Bramble-Hilbert Lemma is proved in [**13**].

3.4. Exercises

EXERCISE 3.1. Let $m \geqslant 0$ and let $1 \leqslant p \leqslant \infty$. Show that, under the conditions of Lemma 3.2, there exists a constant $C = C(m, p, d)$ such that

$$\begin{aligned}|\hat{v}|_{W^{m,p}(\hat{\Omega})} &\leqslant C\|B\|^m|\det B|^{-1/p}|v|_{W^{m,p}(\Omega)},\\ |v|_{W^{m,p}(\Omega)} &\leqslant C\|B^{-1}\|^m|\det B|^{1/p}|\hat{v}|_{W^{m,p}(\hat{\Omega})}.\end{aligned}$$

Exercise 3.2. Under the assumption of Theorem 3.8, show that

$$\|v_h\|_{L^p(\partial\Omega)} \leqslant Ch^{-1/p}\,\|v_h\|_{L^p(\Omega)} \qquad \forall\ v_h \in X_h.$$

Exercise 3.3. Let K be an element in $\mathbb{R}^d$ with diameter h_K. Prove the scaled trace inequality

$$\|v\|_{L^2(\partial K)} \leqslant C\left(h_K^{1/2}\|\nabla v\|_{L^2(K)} + h_K^{-1/2}\|v\|_{L^2(K)}\right) \quad \forall\ v \in H^1(K).$$

Exercise 3.4. Let $\{\mathcal{M}_h\}$ be a regular and quasi-uniform family of triangulations and V_h be the H^1-conforming linear finite element space. Let Q_h be the L^2-projection to $V_h \subset H^1(\Omega)$, i.e.,

$$(Q_h v, v_h) = (v, v_h) \quad \forall\ v_h \in V_h.$$

Show that $\|Q_h v\|_{H^1(\Omega)} \leqslant C\|v\|_{H^1(\Omega)}$ for all $v \in H^1(\Omega)$.

CHAPTER 4

Adaptive Finite Element Methods

The adaptive finite element method based on a posteriori error estimates provides a systematic way to refine or coarsen the meshes according to the local a posteriori error estimator on the elements. The purpose of this chapter is to describe the basic idea of the adaptive finite element method using the example of solving the Poisson equation.

4.1. An example with singularity

We know from Chapter 3 that if the solution of the elliptic problem $Lu = f$ has the regularity $u \in H^2(\Omega)$, then the linear finite element method has the optimal convergence order $O(h)$. For the domain with reentrant corners, however, the solution is no longer in $H^2(\Omega)$. So the classical finite element method fails to provide satisfactory result. The purpose of this chapter is to construct one way to attack this problem. But first we construct an example to illustrate the singular behavior around reentrant corners.

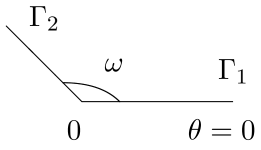

FIGURE 1. The sector S_ω.

We consider the harmonic functions in the sector $S_\omega = \{(r,\theta): \ 0 < r < \infty, \ 0 \leqslant \theta \leqslant \omega\}$, where $0 < \omega < 2\pi$. We look for the solution of the form $u = r^\alpha \mu(\theta)$ for the Laplace equation $-\triangle u = 0$ in S_ω with boundary condition $u = 0$ on $\Gamma_1 \cup \Gamma_2$, where

$$\Gamma_1 = \{(r,\theta): \ r > 0, \ \theta = 0\} \quad , \quad \Gamma_2 = \{(r,\theta): \ r > 0, \ \theta = \omega\}$$

Let $u = r^\alpha \mu(\theta)$. Since in polar coordinates

$$\Delta u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2},$$

we have

$$\triangle u = \alpha(\alpha-1)r^{\alpha-2}\mu(\theta) + \alpha r^{\alpha-2}\mu(\theta) + r^{\alpha-2}\mu''(\theta) = 0,$$

which implies

$$\mu''(\theta) + \alpha^2\mu(\theta) = 0.$$

Therefore $\mu(\theta) = A\sin\alpha\theta + B\cos\alpha\theta$. The boundary condition $\mu(0) = \mu(\omega) = 0$ yields that $\alpha = k\pi/\omega$ and $\mu(\theta) = A\sin(\frac{k\pi}{\omega}\theta), k = 1, 2, 3, \cdots$. Therefore, the boundary value problem $\triangle u = 0$ in $S_\omega, u = 0$ on $\Gamma_1 \cup \Gamma_2$ has a solution

$$u = r^\alpha \sin(\alpha\theta), \quad \alpha = \frac{\pi}{\omega}.$$

Lemma 4.1. *$u \notin H^2(S_\omega \cap B_R)$ for any $R > 0$ if $\pi < \omega < 2\pi$.*

Proof. By direct calculation

$$\begin{aligned}\int_\Omega \left|\frac{\partial^2 u}{\partial r^2}\right|^2 \mathrm{dx} &= \int_0^R \int_0^\omega \left|\alpha(\alpha-1)r^{\alpha-2}\sin(\alpha\theta)\right|^2 r\mathrm{d}r\mathrm{d}\theta \\ &= \alpha^2(\alpha-1)^2\int_0^\omega |\sin^2(\alpha\theta)|^2\mathrm{d}\theta \cdot \int_0^R r^{2\alpha-3}\mathrm{dr} \\ &= cr^{2(\alpha-1)}|_0^R.\end{aligned}$$

This completes the proof. □

Example 4.1. Let us consider the Laplace equation on the L-shaped domain Ω of Figure 2 with the Dirichlet boundary condition so chosen that the true solution is $u = r^{2/3}\sin(2\theta/3)$ in polar coordinates.

Let $\mathcal{M}_h$ be a uniform triangulation of Ω, and u_h be the solution of the linear finite element method over $\mathcal{M}_h$. Since $u \notin H^2(\Omega)$, the H^1 error estimate of u_h in Theorem 3.9 does not hold for this L-shaped domain problem. To find the convergence rate of the linear finite element approximation u_h, we solve the L-shaped domain problem using a sequence of uniform refined meshes $\mathcal{M}_j$ which is obtained by connecting the edge midpoints of $\mathcal{M}_{j-1}$ starting from the mesh $\mathcal{M}_0$ shown in Figure 2 (left). Figure 2 (right) plots the H^1 error $\|u - u_{hj}\|_{H^1(\Omega)}$ versus $2^j = h_0/h_j$ in log-log coordinates, where u_{hj} is the finite element approximation over $\mathcal{M}_j$ and h_j is the maximum diameter of triangles in $\mathcal{M}_j$. It shows that the following error estimate holds for the linear finite element approximation of the L-shaped problem over uniform triangulations:

$$\|u - u_h\|_{H^1(\Omega)} \leqslant Ch^{2/3}. \tag{4.1}$$

The implementation details of this example are given in Section 10.2.

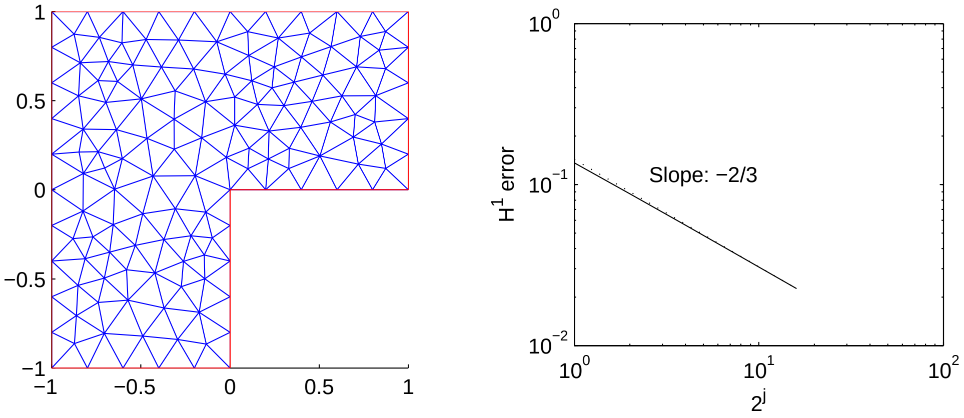

FIGURE 2. Example 4.1: the L-shaped domain and the initial mesh (left). The H^1 error versus 2^j in log-log coordinates and the dotted reference line with slope $-2/3$ (right).

4.2. A posteriori error analysis

Let $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ be a bounded polyhedral domain and $\mathcal{M}_h$ be a shape regular triangulation of Ω. The set of all interior sides of the mesh $\mathcal{M}_h$ is denoted as $\mathcal{B}_h$. Let V_h be the standard H^1-conforming linear finite element space, $V_h^0 = V_h \cap H_0^1(\Omega)$. For any $K \in \mathcal{M}_h$, let h_K be the diameter of K. For any $e \in \mathcal{B}_h$ with $e = K_1 \cap K_2$, let $\Omega_e = K_1 \cup K_2$ and let h_e be the diameter of e as before.

We consider the variational problem to find $u \in H_0^1(\Omega)$ such that

$$(a\nabla u, \nabla v) = (f, v) \qquad \forall\, v \in H_0^1(\Omega), \tag{4.2}$$

where a is assumed to be a piecewise constant function, $f \in L^2(\Omega)$, and Ω is not necessarily convex. Suppose that $a(x)$ is constant on each $K \in \mathcal{M}_h$.

Let $u_h \in V_h^0$ be the finite element solution of the discrete problem

$$(a\nabla u_h, \nabla v_h) = (f, v_h) \qquad \forall\, v_h \in V_h^0. \tag{4.3}$$

In this section, we first introduce the Clément interpolation operator for non-smooth functions, then introduce the a posteriori error estimates including the upper bound and lower bound.

4.2.1. The Clément interpolation operator.

The Clément interpolation operator to be introduced has a definition for any function in $L^1(\Omega)$, comparing to the Lagrange interpolation operator which is defined for continuous functions.

Let $\{x_j\}_{j=1}^{\bar{J}}$ be the set of nodes of the mesh $\mathcal{M}_h$, and $\{\phi_j\}_{j=1}^{\bar{J}}$ be the set of nodal basis functions. For any x_j, define $S_j = \operatorname{supp}(\phi_j)$, the star surround x_j. Since the triangulation $\mathcal{M}_h$ is regular, the number of elements in S_j is bounded by a constant depending only on the minimum angle of the mesh $\mathcal{M}_h$. Consequently, the macro-elements S_j can only assume a finite number of different configurations.

Denote by $\Lambda = \{\hat{S}\}$ the set of reference configurations. The number of reference configurations $\#\Lambda$ depends only on the minimum angle of $\mathcal{M}_h$. For any S_j, let $\hat{S}_j$ be the corresponding reference configuration in Λ and let F_j be a C^0-diffeomorphism from $\hat{S}_j$ to S_j such that $F_i|_{\hat{K}}$ is affine for any $\hat{K} \subset \hat{S}_j$. Define $\hat{R}_j : \ L^1(\hat{S}_j) \to P_1(\hat{S}_j)$ the L^2 projection operator by

$$\hat{R}_j\hat{\psi} \in P_1(\hat{S}_j) : \int_{\hat{S}_j} (\hat{R}_j\hat{\psi})\hat{v}_h \mathrm{d}x = \int_{\hat{S}_j} \hat{\psi}\hat{v}_h \mathrm{d}x \qquad \forall \hat{v}_h \in P_1(\hat{S}_j), \tag{4.4}$$

for any $\hat{\psi} \in L^1(\hat{S}_j)$. For any $\psi \in L^1(\Omega)$, denote by $\hat{\psi}_j = \psi \circ F_j$. Let $\{x_j\}_{j=1}^{J}$ be the set of interior nodes. The Clément interpolation operators Π_h and Π_h^0 are then defined by

$$\Pi_h : \ L^1(\Omega) \to V_h, \quad \Pi_h\psi = \sum_{j=1}^{\bar{J}} (\hat{R}_j\hat{\psi}_j)(F_j^{-1}(x_j))\phi_j,$$

$$\Pi_h^0 : \ L^1(\Omega) \to V_h^0, \quad \Pi_h^0\psi = \sum_{j=1}^{J} (\hat{R}_j\hat{\psi}_j)(F_j^{-1}(x_j))\phi_j.$$

Theorem 4.2. *There exists a constant C depending only on the minimum angle of $\mathcal{M}_h$ such that for any $\psi \in H_0^1(\Omega)$*

$$\|\psi - \Pi_h^0\psi\|_{L^2(K)} \leqslant Ch_K\|\nabla\psi\|_{L^2(\widetilde{K})} \qquad \forall K \in \mathcal{M}_h, \tag{4.5}$$

$$\|\psi - \Pi_h^0\psi\|_{L^2(e)} \leqslant Ch_e^{1/2}\|\nabla\psi\|_{L^2(\tilde{e})} \qquad \forall e \in \mathcal{B}_h, \tag{4.6}$$

$$\left\|\nabla\Pi_h^0\psi\right\|_{L^2(K)} \leqslant C\|\nabla\psi\|_{L^2(\widetilde{K})} \qquad \forall K \in \mathcal{M}_h, \tag{4.7}$$

where $\widetilde{K}$ *is the union of all elements in* $\mathcal{M}_h$ *having nonempty intersection with* K*, and* $\tilde{e} = \widetilde{K}_1 \cup \widetilde{K}_2$ *with* $e = K_1 \cap K_2$*.*

Proof. The proof is divided into three steps.

1°) (4.6) and (4.7) are direct consequences of (4.5). Let $\psi_K = \frac{1}{|K|}\int_K \psi \mathrm{d}x$ be the average of ψ on K, then it follows from the local inverse estimate in (3.7), Theorem 1.10, and (4.5) that

$$\begin{aligned}\|\nabla \Pi_h^0 \psi\|_{L^2(K)} &= \|\nabla(\Pi_h^0\psi - \psi_K)\|_{L^2(K)} \leqslant C h_K^{-1}\|\Pi_h^0\psi - \psi_K\|_{L^2(K)} \\ &\leqslant C h_K^{-1}\big(\|\Pi_h^0\psi - \psi\|_{L^2(K)} + \|\psi - \psi_K\|_{L^2(K)}\big) \\ &\leqslant C\|\nabla\psi\|_{L^2(\widetilde{K})}.\end{aligned}$$

On the other hand, by the scaled trace inequality in Exercise 3.3, for $e \subset \partial K$ for some $K \in \mathcal{M}_h$,

$$\begin{aligned}\|\psi - \Pi_h^0\psi\|_{L^2(e)} &\leqslant C\big(h_e^{-1/2}\|\psi - \Pi_h^0\psi\|_{L^2(K)} + h_e^{1/2}\|\nabla(\psi - \Pi_h^0\psi)\|_{L^2(K)}\big) \\ &\leqslant C h_e^{1/2}\|\nabla\psi\|_{L^2(\widetilde{K})} \leqslant C h_e^{1/2}\|\nabla\psi\|_{L^2(\tilde{e})}.\end{aligned}$$

2°) We have, from Theorem 3.1 and the inverse inequality, for any $\hat{\psi} \in H^1(\hat{S}_j)$,

$$\|\hat{\psi} - \hat{R}_j\hat{\psi}\|_{L^2(\hat{S}_j)} \leqslant \inf_{\hat{p}\in P_1(\hat{S}_j)} \|\hat{\psi} - \hat{p}\|_{L^2(\hat{S}_j)} \leqslant C\|\nabla\hat{\psi}\|_{L^2(\hat{S}_j)}, \tag{4.8}$$

$$\begin{aligned}\|\nabla\hat{R}_j\hat{\psi}\|_{L^2(\hat{S}_j)} &= \|\nabla\hat{R}_j(\hat{\psi} - \hat{\psi}_{\hat{S}_j})\|_{L^2(\hat{S}_j)} \leqslant C\big\|\hat{R}_j\big(\hat{\psi} - \hat{\psi}_{\hat{S}_j}\big)\big\|_{L^2(\hat{S}_j)} \\ &\leqslant C\big\|\hat{\psi} - \hat{\psi}_{\hat{S}_j}\big\|_{L^2(\hat{S}_j)} \leqslant C\|\nabla\hat{\psi}\|_{L^2(\hat{S}_j)}.\end{aligned} \tag{4.9}$$

Denote by h_j the diameter of S_j. Since $\sum_{j=1}^{J}\phi_j = 1$, we have

$$\begin{aligned}\|\psi - \Pi_h\psi\|_{L^2(K)} &= \Big\|\sum_{x_j\in K}\big(\psi - (\hat{R}_j\hat{\psi}_j)(F_j^{-1}(x_j))\big)\phi_j\Big\|_{L^2(K)} \\ &\leqslant C\sum_{x_j\in K}\big\|\psi - (\hat{R}_j\hat{\psi}_j)(F_j^{-1}(x_j))\big\|_{L^2(S_j)} \\ &\leqslant C\sum_{x_j\in K} h_j^{d/2}\|\hat{\psi}_j - (\hat{R}_j\hat{\psi}_j)(F_j^{-1}(x_j))\|_{L^2(\hat{S}_j)} \\ &\leqslant \sum_{x_j\in K} h_j^{d/2}\Big(\big\|\hat{\psi}_j - \hat{R}_j\hat{\psi}_j\big\|_{L^2(\hat{S}_j)} + \big\|\hat{R}_j\hat{\psi}_j - (\hat{R}_j\hat{\psi}_j)(F_j^{-1}(x_j))\big\|_{L^2(\hat{S}_j)}\Big) \\ &\leqslant \sum_{x_j\in K} h_j^{d/2}\Big(\big\|\hat{\psi}_j - \hat{R}_j\hat{\psi}_j\big\|_{L^2(\hat{S}_j)} + \big\|\nabla\hat{R}_j\hat{\psi}_j\big\|_{L^2(\hat{S}_j)}\Big) \\ &\leqslant \sum_{x_j\in K} h_j^{d/2}\|\nabla\hat{\psi}_j\|_{L^2(\hat{S}_j)} \leqslant C h_K\|\nabla\psi\|_{L^2(\widetilde{K})}.\end{aligned}$$

3°) To conclude the proof we must consider the case when $K \in \mathcal{M}_h$ has a node on the boundary $\partial\Omega$ because, otherwise, $\Pi_h^0\psi = \Pi_h\psi$ on K. Notice that if $x_j \in \partial\Omega$ then there exists a side $e_j \subset \partial\Omega$ including x_j as one of its vertices. Let $\hat{e}_j = F_j^{-1}(e_j)$. Since $\psi = 0$ on e_j for $\psi \in H_0^1(\Omega)$, we have

$$\begin{aligned}|(\hat{R}_j\hat{\psi}_j)(F_j^{-1}(x_j))| &\leqslant \left|(\hat{R}_j\hat{\psi}_j)(F_j^{-1}(x_j)) - \frac{1}{|\hat{e}_j|}\int_{\hat{e}_j}\hat{R}_j\hat{\psi}_j\right| + \frac{1}{|\hat{e}_j|}\left|\int_{\hat{e}_j}\hat{R}_j\hat{\psi}_j\right| \\ &\leqslant C\|\nabla\hat{R}_j\hat{\psi}_j\|_{L^2(\hat{S}_j)} + C\|\hat{\psi}_j - \hat{R}_j\hat{\psi}_j\|_{L^2(\hat{e}_j)} \\ &\leqslant C\|\nabla\hat{\psi}_j\|_{L^2(\hat{S}_j)} \leqslant Ch_j^{1-d/2}\|\nabla\psi\|_{L^2(S_j)}.\end{aligned}$$

Therefore, on boundary element $K \in \mathcal{M}_h$,

$$\begin{aligned}\|\Pi_h\psi - \Pi_h^0\psi\|_{L^2(K)} &\leqslant \sum_{x_j\in\partial\Omega\cap K}|(\hat{R}_j\hat{\psi}_j)(F_j^{-1}(x_j))|\,\|\phi_j\|_{L^2(K)} \\ &\leqslant C\sum_{x_j\in\partial\Omega\cap K}h_j^{1-d/2}\|\nabla\psi\|_{L^2(S_j)}h_j^{d/2} \\ &\leqslant C\sum_{x_j\in\partial\Omega\cap K}h_K\|\nabla\psi\|_{L^2(S_j)} \leqslant Ch_K\|\nabla\psi\|_{L^2(\widetilde{K})}.\end{aligned}$$

This completes the proof of the theorem. □

4.2.2. A posteriori error estimates.

For any $e \in \mathcal{B}_h$ with $e = K_1 \cap K_2$ we define the jump residual for u_h by

$$J_e = \left([\![a(x)\nabla u_h]\!]\cdot\nu\right)\big|_e := a(x)\nabla u_h|_{K_1}\cdot\nu_1 + a(x)\nabla u_h|_{K_2}\cdot\nu_2, \tag{4.10}$$

where ν_i is the unit outer normal of ∂K_i restricted to e. For convenience, define $J_e = 0$ for any side $e \subset \partial\Omega$. For any $K \in \mathcal{M}_h$, define the error indicator η_K by

$$\eta_K^2 := h_K^2\|f\|_{L^2(K)}^2 + h_K\sum_{e\subset\partial K}\|J_e\|_{L^2(e)}^2. \tag{4.11}$$

For any domain $G \subset \Omega$ let $|\!|\!|\cdot|\!|\!|_G = \|a^{1/2}\nabla\cdot\|_{L^2(G)}$. Note that $|\!|\!|\cdot|\!|\!|_\Omega$ is the energy norm in $H_0^1(\Omega)$.

Theorem 4.3 (*Upper bound*). *There exists a constant $C_1 > 0$ which depends only on the minimum angle of the mesh $\mathcal{M}_h$ and the minimum value of $a(x)$ such that*

$$|\!|\!|u - u_h|\!|\!|_\Omega \leqslant C_1\left(\sum_{K\in\mathcal{M}_h}\eta_K^2\right)^{1/2}.$$

Proof. Define $\mathcal{R} \in H^{-1}(\Omega)$ as the residual through

$$\langle \mathcal{R}, \varphi \rangle = (f, \varphi) - (a\nabla u_h, \nabla \varphi) = (a\nabla(u - u_h), \nabla \varphi), \qquad \forall\, \varphi \in H_0^1(\Omega).$$

By (4.3) we obtain the Galerkin orthogonality $\langle \mathcal{R}, v_h \rangle = 0$ for any $v_h \in V_h^0$. Thus

$$\begin{aligned}
(a\nabla(u - u_h), \nabla \varphi) &= \langle \mathcal{R}, \varphi - \Pi_h^0 \varphi \rangle \\
&= (f, \varphi - \Pi_h^0 \varphi) - (a\nabla u_h, \nabla(\varphi - \Pi_h^0 \varphi)) \\
&= (f, \varphi - \Pi_h^0 \varphi) - \sum_{K \in \mathcal{M}_h} \int_K a\nabla u_h \cdot \nabla(\varphi - \Pi_h^0 \varphi) \mathrm{d}x \\
&= (f, \varphi - \Pi_h^0 \varphi) - \sum_{K \in \mathcal{M}_h} \int_{\partial K} a\nabla u_h \cdot \nu (\varphi - \Pi_h^0 \varphi)\, \mathrm{d}s \\
&= \sum_{K \in \mathcal{M}_h} \int_K f(\varphi - \Pi_h^0 \varphi) \mathrm{d}x - \sum_{e \in \mathcal{B}_h} \int_e J_e\, (\varphi - \Pi_h^0 \varphi)\, \mathrm{d}s \\
&\leqslant C \left(\sum_{K \in \mathcal{M}_h} \| h_K f \|_{L^2(K)}^2 \right)^{1/2} \| \nabla \varphi \|_{L^2(\Omega)} \\
&\qquad + C \left(\sum_{e \in \mathcal{B}_h} \| h_e^{1/2} J_e \|_{L^2(e)}^2 \right)^{1/2} \| \nabla \varphi \|_{L^2(\Omega)} \\
&\leqslant C_1 \left(\sum_{K \in \mathcal{M}_h} \eta_K^2 \right)^{1/2} |||\varphi|||_\Omega.
\end{aligned}$$

The theorem follows by taking $\varphi = u - u_h \in H_0^1(\Omega)$. □

Theorem 4.4 (*Local lower bound*). *There exists a constant $C_2 > 0$ which depends only on the minimum angle of the mesh $\mathcal{M}_h$ and the maximum value of $a(x)$ such that for any $K \in \mathcal{M}_h$*

$$\eta_K^2 \leqslant C_2 |||u - u_h|||_{K^*}^2 + C_2 \sum_{K \subset K^*} h_K^2 \| f - f_K \|_{L^2(K)}^2,$$

where $f_K = \frac{1}{|K|} \int_K f \mathrm{d}x$ and K^* is the union of all elements sharing at least one common side with K.

Proof. From the proof of Theorem 4.3,

$$(a\nabla(u - u_h), \nabla \varphi) = \sum_{K \in \mathcal{M}_h} \int_K f \varphi \mathrm{d}x - \sum_{e \in \mathcal{B}_h} \int_e J_e \varphi\, \mathrm{d}s, \quad \forall \varphi \in H_0^1(\Omega). \tag{4.12}$$

The remains of the proof is divided into two steps.

1°) For any $K \in \mathcal{M}_h$, Let $\varphi_K = (d+1)^{d+1}\lambda_1 \cdots \lambda_{d+1}$ be the canonical bubble function in K, we choose the constant α_K such that $\varphi = \alpha_K \varphi_K$ satisfies

$$\int_K f_K \varphi \mathrm{d}x = h_K^2 \|f_K\|_{L^2(K)}^2.$$

It is clear that

$$|\alpha_K| = \frac{h_K^2 |f_K| |K|}{\int_K \varphi_K \mathrm{d}x} \leqslant C h_K^{1-\frac{d}{2}} \|h_K f_K\|_{L^2(K)}$$

and thus

$$h_K^{-1} \|\varphi\|_{L^2(K)}, \quad \|\nabla \varphi\|_{L^2(K)} \leqslant C |\alpha_K| \, h_K^{-1} \, |K|^{1/2} \leqslant C \|h_K f_K\|_{L^2(K)}.$$

Now

$$\|h_K f\|_{L^2(K)}^2 \leqslant 2\|h_K f_K\|_{L^2(K)}^2 + 2\|h_K(f - f_K)\|_{L^2(K)}^2$$

and it follows from (4.12) and $\varphi \in H_0^1(K)$ that

$$\begin{aligned}\|h_K f_K\|_{L^2(K)}^2 = \int_K f_K \varphi \mathrm{d}x = \int_K (f_K - f)\varphi \mathrm{d}x + \int_K a \nabla(u - u_h) \nabla \varphi \mathrm{d}x \\ \leqslant C \|h_K(f - f_K)\|_{L^2(K)} \|h_K^{-1} \varphi\|_{L^2(K)} + C |||u - u_h|||_K \|\nabla \varphi\|_{L^2(K)} \\ \leqslant C \|h_K f_K\|_{L^2(K)} \left(|||u - u_h|||_K^2 + \|h_K(f - f_K)\|_{L^2(K)}^2 \right)^{1/2}.\end{aligned}$$

Therefore,

$$\|h_K f\|_{L^2(K)}^2 \leqslant C \left(|||u - u_h|||_K^2 + \|h_K(f - f_K)\|_{L^2(K)}^2 \right).$$

2°) For any side $e \subset \partial K \cap \Omega$, let $\psi_e = d^d \lambda_1 \cdots \lambda_d$ be the bubble function, where $\lambda_1, \cdots, \lambda_d$ are the barycentric coordinate functions associate with the nodes of e. Denote by $\psi = \beta_e \psi_e$ the function satisfies

$$\int_e J_e \psi = h_K \|J_e\|_{L^2(e)}^2,$$

It is easy to check that

$$|\beta_e| \leqslant C h_K |J_e| \leqslant C h_K^{1-\frac{d}{2}} h_K^{1/2} \|J_e\|_{L^2(e)}$$

and thus

$$h_K^{-1} \|\psi\|_{L^2(\Omega_e)}, \quad \|\nabla \psi\|_{L^2(\Omega_e)} \leqslant C \, |\beta_e| \, h_K^{-1} \, |\Omega_e|^{1/2} \leqslant C h_K^{1/2} \|J_e\|_{L^2(e)}.$$

Now it follows from (4.12) and $\psi \in H_0^1(\Omega_e)$ that

$$h_K\|J_e\|_{L^2(e)}^2 = \int_e J_e\psi = \int_{\Omega_e} f\psi \mathrm{d}x - \int_{\Omega_e} a\nabla(u-u_h)\nabla\psi \mathrm{d}x$$
$$\leqslant Ch_K^{1/2}\|J_e\|_{L^2(e)} \left(\sum_{K\subset\Omega_e} \|h_K f\|_{L^2(K)}^2 + |||u-u_h|||_{\Omega_e}^2 \right)^{1/2} .$$

This completes the proof upon using the estimate for $\|h_K f\|_{L^2(K)}$. □

The lower bound in Theorem 4.4 implies that up to a high order quantity $\left(\sum_{K\subset K^*} h_K^2\|f-f_K\|_{L^2(K)}^2\right)^{1/2}$, the local energy error $|||u-u_h|||_{K^*}$ is bound from below by the error indicator η_K.

4.3. Adaptive algorithm

Based on the local error indicators, the usual adaptive algorithm solving the variational problem (4.5) may be described as loops of the form

$$\mathsf{Solve} \longrightarrow \mathsf{Estimate} \longrightarrow \mathsf{Mark} \longrightarrow \mathsf{Refine}. \tag{4.13}$$

The important convergence property, which guarantees the iterative loop (4.13) terminates in finite number of iterations starting from any given initial mesh, depends on the proper design of marking strategies. There are several marking strategies proposed in the literature. Here we give a brief review.

1. The error equidistribution strategy: Given $\theta > 1$ and a tolerance TOL, mark all elements K such that
$$\eta_K \geqslant \theta\frac{\mathrm{TOL}}{\sqrt{M}} ,$$
where M is the number of elements in $\mathcal{M}_h$.
2. The maximum strategy: Given $\theta \in (0,1)$, mark all elements K such that
$$\eta_K \geqslant \theta \max_{K'\in\mathcal{M}_h} \eta_{K'} .$$
3. The Dörfler Strategy. Given $\theta \in (0,1]$, mark elements in a subset $\hat{\mathcal{M}}_h$ of $\mathcal{M}_h$ such that
$$\eta_{\hat{\mathcal{M}}_h} \geqslant \theta\eta_{\mathcal{M}_h}. \tag{4.14}$$

Given a triangulation $\mathcal{M}_H$ and and a set of marked elements $\hat{\mathcal{M}}_H \subset \mathcal{M}_H$, the refinement of $\mathcal{M}_H$ usually consists of two steps: refining the marked elements and removing the hanging nodes. We make the following assumption on the first step:

$$\begin{aligned}&\textit{Any marked simplex is subdivided into several subsimplices such that}\\&\textit{the measure of each subsimplex} \leqslant \frac{1}{m}\times \textit{ the measure of its father simplex.}\end{aligned} \tag{4.15}$$

Here $m > 1$ is a fixed number. For example, in the case of one time bisection, $m = 2$. We remark that some unmarked simplices may be refined in the step of removing hanging nodes.

4.4. Convergence analysis

In this section we consider the convergence of the adaptive finite element algorithm based on the Dörfler strategy. We start with the following lemma.

LEMMA 4.2. *Let $\mathcal{M}_h$ be a refinement of $\mathcal{M}_H$ such that $V_H \subset V_h$. Then the following relation holds*

$$|||u - u_h|||_\Omega^2 = |||u - u_H|||_\Omega^2 - |||u_h - u_H|||_\Omega^2.$$

PROOF. The proof is straightforward by the Galerkin orthogonality since $u_h - u_H \in V_H^0$. □

Let

$$\tilde{\eta}_K^2 : = \tilde{h}_K^2 \|f\|_{L^2(K)}^2 + \tilde{h}_K \sum_{e\subset\partial K} \|J_e\|_{L^2(e)}^2, \qquad \text{where } \tilde{h}_K := |K|^{1/d}. \tag{4.16}$$

It is clear that there exist positive constants c_1 and c_2 such that

$$c_2\eta_K \leqslant \tilde{\eta}_K \leqslant c_1\eta_K. \tag{4.17}$$

The modified the error indicator $\tilde{\eta}_K$ enjoys the following reduction property.

LEMMA 4.3. *Let $\hat{\mathcal{M}}_H \subset \mathcal{M}_H$ be the set of elements marked for refinement and let $\mathcal{M}_h$ be a refinement of $\mathcal{M}_H$ satisfying the assumption* (4.15). *Then there exists a constant C_3 depending only the minimum angle of the meshes and the maximum value of $a(x)$ such that, for any $\delta > 0$,*

$$\tilde{\eta}_{\mathcal{M}_h}^2 \leqslant (1+\delta)\Big(\tilde{\eta}_{\mathcal{M}_H}^2 - \big(1 - \frac{1}{\sqrt[d]{m}}\big)\tilde{\eta}_{\hat{\mathcal{M}}_H}^2\Big) + \big(1+\frac{1}{\delta}\big)C_3|||u_h - u_H|||_\Omega^2.$$

Proof. From the Young inequality with parameter δ,

$$\begin{aligned}\tilde\eta^2_{\mathcal{M}_h} &= \sum_{K\in\mathcal{M}_h}\Big(\tilde h_K^2\|f\|^2_{L^2(K)} + \tilde h_K\sum_{e\subset\partial K\cap\Omega}\big\|\big([\![a\nabla(u_H+u_h-u_H)]\!]\cdot\nu\big)\big|_e\big\|^2_{L^2(e)}\Big)\\ &\leqslant \sum_{K\in\mathcal{M}_h}\Big(\tilde h_K^2\|f\|^2_{L^2(K)} + (1+\delta)\tilde h_K\sum_{e\subset\partial K\cap\Omega}\big\|\big([\![a\nabla u_H]\!]\cdot\nu\big)\big|_e\big\|^2_{L^2(e)}\Big)\\ &\quad + \big(1+\frac{1}{\delta}\big)\sum_{K\in\mathcal{M}_h}\tilde h_K\sum_{e\subset\partial K\cap\Omega}\big\|\big([\![a\nabla(u_h-u_H)]\!]\cdot\nu\big)\big|_e\big\|^2_{L^2(e)}\\ &:=\mathrm{I}+\mathrm{II}.\end{aligned}$$

Note that $\big([\![a\nabla u_H]\!]\cdot\nu\big)\big|_e = 0$ for any e in the interior of some element $K'\in\mathcal{M}_H$ and that $\tilde h_K = |K|^{1/d} \leqslant \frac{1}{\sqrt[d]{m}}\widetilde H_{K'}$ for any $K\subset K'\in\hat{\mathcal{M}}_H$. We have

$$\begin{aligned}\mathrm{I} \leqslant &(1+\delta)\sum_{K\subset K'\in\mathcal{M}_H\setminus\hat{\mathcal{M}}_H}\Big(\tilde h_K^2\|f\|^2_{L^2(K)} + \tilde h_K\sum_{e\subset\partial K\cap\Omega}\big\|\big([\![a\nabla u_H]\!]\cdot\nu\big)\big|_e\big\|^2_{L^2(e)}\Big)\\ &+(1+\delta)\sum_{K\subset K'\in\hat{\mathcal{M}}_H}\Big(\tilde h_K^2\|f\|^2_{L^2(K)} + \tilde h_K\sum_{e\subset\partial K\cap\Omega}\big\|\big([\![a\nabla u_H]\!]\cdot\nu\big)\big|_e\big\|^2_{L^2(e)}\Big)\\ \leqslant&(1+\delta)\sum_{K'\in\mathcal{M}_H\setminus\hat{\mathcal{M}}_H}\big(\widetilde H_{K'}^2\|f\|^2_{L^2(K')} + \widetilde H_{K'}\sum_{e'\subset\partial K'\cap\Omega}\big\|\big([\![a\nabla u_H]\!]\cdot\nu\big)\big|_{e'}\big\|^2_{L^2(e')}\big)\\ &+\frac{1+\delta}{\sqrt[d]{m}}\sum_{K'\in\hat{\mathcal{M}}_H}\big(\widetilde H_{K'}^2\|f\|^2_{L^2(K')} + \widetilde H_{K'}\sum_{e'\subset\partial K'\cap\Omega}\big\|\big([\![a\nabla u_H]\!]\cdot\nu\big)\big|_{e'}\big\|^2_{L^2(e')}\big)\\ =&(1+\delta)\tilde\eta^2_{\mathcal{M}_H\setminus\hat{\mathcal{M}}_H} + \frac{1+\delta}{\sqrt[d]{m}}\tilde\eta^2_{\hat{\mathcal{M}}_H} = (1+\delta)\Big(\tilde\eta^2_{\mathcal{M}_H} - \big(1-\frac{1}{\sqrt[d]{m}}\big)\tilde\eta^2_{\hat{\mathcal{M}}_H}\Big).\end{aligned}$$

Next we estimate II. For any $e\in\mathcal{B}_h$, denote by K_1 and K_2 the two elements having common side e. We have

$$\begin{aligned}\mathrm{II} \leqslant &C\big(1+\frac{1}{\delta}\big)\sum_{e\in\mathcal{B}_h}h_e\,\big\|\big([\![a\nabla(u_h-u_H)]\!]\cdot\nu\big)\big|_e\big\|^2_{L^2(e)}\\ =&C\big(1+\frac{1}{\delta}\big)\sum_{e\in\mathcal{B}_h}h_e\,\|a\nabla(u_h-u_H)|_{K_1}\cdot\nu_1 + a\nabla(u_h-u_H)|_{K_2}\cdot\nu_2\|^2_{L^2(e)}\\ \leqslant&C\big(1+\frac{1}{\delta}\big)\sum_{e\in\mathcal{B}_h}h_e\big(\,\|a\nabla(u_h-u_H)|_{K_1}\|^2_{L^2(e)} + \|a\nabla(u_h-u_H)|_{K_2}\|^2_{L^2(e)}\,\big)\\ \leqslant&C\big(1+\frac{1}{\delta}\big)\sum_{e\in\mathcal{B}_h}\|a\nabla(u_h-u_H)\|^2_{K_1\cup K_2} \leqslant \big(1+\frac{1}{\delta}\big)C_3|\!|\!|u_h-u_H|\!|\!|^2_{\Omega}.\end{aligned}$$

The proof follows by combining the above three estimates. □

THEOREM 4.5. *Let $\theta \in (0,1]$, and let $\{\mathcal{M}_k, u_k\}_{k\geqslant 0}$ be the sequence of meshes and discrete solutions produced by the adaptive finite element algorithm based on the Dörfler marking strategy and the assumption* (4.15)*. Suppose the family of meshes $\{\mathcal{M}_k\}$ is shape regular. Then there exist constants $\gamma > 0$, $C_0 > 0$, and $0 < \alpha < 1$, depending solely on the shape-regularity of $\{\mathcal{M}_k\}$, m, and the marking parameter θ, such that*

$$\Big(|||u-u_k|||_\Omega^2 + \gamma\eta_{\mathcal{M}_k}^2\Big)^{1/2} \leqslant C_0\alpha^k. \tag{4.18}$$

PROOF. We first show that there exist constants $\gamma_0 > 0$ and $0 < \alpha < 1$ such that

$$|||u-u_{k+1}|||_\Omega^2 + \gamma_0\tilde{\eta}_{\mathcal{M}_{k+1}}^2 \leqslant \alpha^2\Big(|||u-u_k|||_\Omega^2 + \gamma_0\tilde{\eta}_{\mathcal{M}_k}^2\Big). \tag{4.19}$$

For convenience, we use the notation

$$e_k := |||u-u_k|||_\Omega, \quad \tilde{\eta}_k := \tilde{\eta}_{\mathcal{M}_k}, \quad \lambda := 1 - \frac{1}{\sqrt[d]{m}}.$$

From Lemma 4.2, Lemma 4.3, and the Dörfler strategy, we know that

$$\tilde{\eta}_{k+1}^2 \leqslant (1+\delta)\big(1-\lambda\theta^2\big)\tilde{\eta}_k^2 + \big(1+\frac{1}{\delta}\big)C_3(e_k^2 - e_{k+1}^2). \tag{4.20}$$

Next by Theorem 4.3 and (4.17) we have

$$e_k^2 \leqslant \widetilde{C}_1\tilde{\eta}_k^2, \qquad \text{where } \widetilde{C}_1 = C_1/c_2. \tag{4.21}$$

Let $\beta = \big(1+\frac{1}{\delta}\big)C_3$. Then, it follows from (4.20) and (4.21) that, for $0 < \zeta < 1$,

$$\begin{aligned} e_{k+1}^2 + \frac{1}{\beta}\tilde{\eta}_{k+1}^2 \leqslant & e_k^2 + \frac{1}{\beta}(1+\delta)\big(1-\lambda\theta^2\big)\tilde{\eta}_k^2 \\ \leqslant & \zeta\, e_k^2 + \Big((1-\zeta)\widetilde{C}_1 + \frac{1}{\beta}(1+\delta)\big(1-\lambda\theta^2\big)\Big)\tilde{\eta}_k^2 \\ = & \zeta\,\Big(e_k^2 + \frac{1}{\beta}\Big(\beta\zeta^{-1}(1-\zeta)\widetilde{C}_1 + \zeta^{-1}(1+\delta)\big(1-\lambda\theta^2\big)\Big)\tilde{\eta}_k^2\Big). \end{aligned}$$

We choose $\delta > 0$ such that $(1+\delta)(1-\lambda\theta^2) < 1$ and choose ζ such that $\beta\zeta^{-1}(1-\zeta)\widetilde{C}_1 + \zeta^{-1}(1+\delta)\big(1-\lambda\theta^2\big) = 1$ which amounts to take

$$\zeta = \frac{(1+\delta)\big(1-\lambda\theta^2\big) + \beta\widetilde{C}_1}{1+\beta\widetilde{C}_1} < 1.$$

This implies (4.19) holds with

$$\gamma_0 = \frac{1}{\beta} = \frac{\delta}{(1+\delta)C_3} \quad \text{and} \quad \alpha^2 = \zeta = \frac{\delta(1+\delta)(1-\lambda\theta^2) + (1+\delta)\widetilde{C}_1C_3}{\delta + (1+\delta)\widetilde{C}_1C_3}.$$

To conclude the proof, we note that by (4.17) $\tilde{\eta}_k \geqslant c_2\eta_k$ and thus (4.18) is valid with

$$\gamma = \gamma_0 c_2^2 \quad \text{and} \quad C_0 = \left(|||u - u_0|||_\Omega^2 + \gamma_0 \tilde{\eta}_0^2 \right)^{1/2}.$$

This completes the proof of the theorem. □

In two dimensional case, extensive numerical experiments strongly suggest that the adaptive finite element method based on a posteriori error estimates described in this chapter enjoys the remarkable property that the meshes and the associated numerical complexity are quasi-optimal in the sense that the linear finite element discretization error is proportional to $N^{-1/2}$ in terms of the energy norm, where N is the number of elements of the underlying mesh. Theorem 4.5, however, does not provide any hint on this important property.

Example 4.6. Consider the L-shaped domain problem in Example 4.1 using the adaptive algorithm base on the maximum strategy. Figure 3 plots the mesh after 10 adaptive iterations (left) and plots the H^1 errors $\|u - u_k\|_{H^1(\Omega)}$ versus N_k in log-log coordinates (right), where u_k is the finite element approximation over $\mathcal{M}_k$, the mesh after k iterations, and N_k is the total number of degrees of freedom in $\mathcal{M}_k$. It shows that

$$\|u - u_k\|_{H^1(\Omega)} \approx O(N_k^{-1/2}), \tag{4.22}$$

is valid asymptotically as $k \to \infty$. We notice that the convergence rate is quasi-optimal. The implementation details of this example are given in Section 10.3.

Bibliographic notes. The study of adaptive finite element methods based on a posteriori error estimates is started in Babuška and Rheihnbold [**6**]. The upper bound in Theorem 4.3 is from Babuška and Miller [**5**] and the local lower bound in Theorem 4.4 is from Verfürth [**50**]. Further results on the a posteriori error estimates for stationary problems can be found in the book Verfürth [**51**]. The convergence of adaptive algorithms is first considered in Dörfler [**27**]. Section 4.4 is based on the work of Cascon et al [**18**] where the convergence of the adaptive finite element methods based on Dörfler strategy using the error indicator $\tilde{\eta}_K$ is proved. The Clément interpolation operator for non-smooth functions is introduced in [**24**]. Studies on the quasi-optimal convergence of adaptive finite element methods can be found in [**18**] and the extensive references therein.

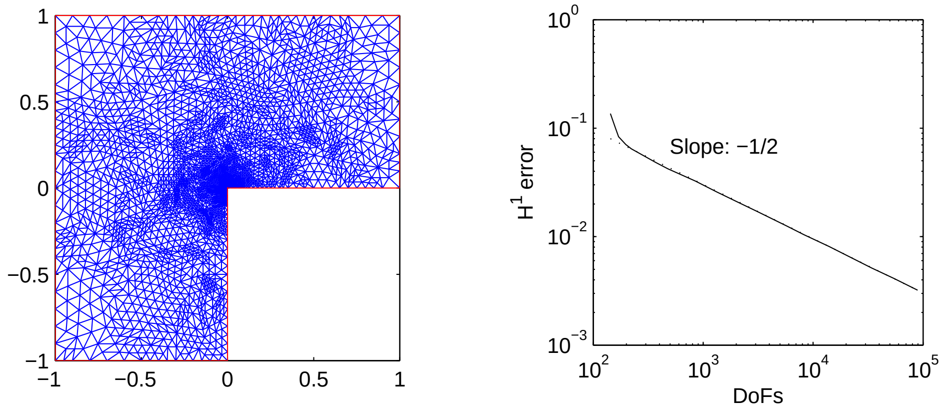

FIGURE 3. Example 4.6: the mesh after 10 adaptive iterations (left). The H^1 error versus the total number of degrees of freedom in log-log coordinates and the dotted reference line with slope $-1/2$ (right).

4.5. Exercises

EXERCISE 4.1. Find the general solution of the form $u = r^\alpha \mu(\theta)$ to the Laplace equation $-\triangle u = 0$ in the sector S_ω which satisfies the boundary conditions

(i) $\dfrac{\partial u}{\partial \nu} = 0$ on $\Gamma_1 \cup \Gamma_2$;

(ii) $u = 0$ on Γ_1, $\dfrac{\partial u}{\partial \nu} = 0$ on Γ_2.

EXERCISE 4.2. Show that there exists a constant C depending only on the minimum angle of $\mathcal{M}_h$ such that (4.8) and (4.9) hold.

EXERCISE 4.3. Let Ω be a bounded polyhedral domain in $\mathbb{R}^d$ $(d = 2, 3)$. Prove the following error estimate for the Clément interpolation operator

$$\|\varphi - \Pi_h \varphi\|_{H^k(K)} \leqslant C h_K^{2-k} |\varphi|_{H^2(\widetilde{K})} \quad \forall \varphi \in H^2(\Omega), \quad k = 0, 1.$$

EXERCISE 4.4. Let $\Omega \subset \mathbb{R}^2$ be a bounded polygon. For $f \in L^2(\Omega)$ and $g \in C(\partial\Omega)$, let $u \in H^1(\Omega)$ be the weak solution of $-\Delta u = f$ in Ω, $u = g$ on $\partial\Omega$. Let $u_h \in V_h$ be the conforming linear finite element approximation such that $u_h = I_h g$ on $\partial\Omega$. Derive an a posteriori error estimate for $\|\nabla(u - u_h)\|_{L^2(\Omega)}$.

EXERCISE 4.5. Let $\Omega = (0, 1)$. Derive a posteriori error estimate for the conforming linear finite element approximation to the two-point boundary value problem $-u'' = f$ in $\Omega, u(0) = \alpha, u'(1) = \beta$.

CHAPTER 5

Finite Element Multigrid Methods

The multigrid method provides an optimal complexity algorithm for solving discrete elliptic boundary value problems. The error bounds of the approximate solution obtained from the full multigrid algorithm are comparable to the theoretical error bounds of the the finite element solution, while the amount of computational work involved is proportion only to the number of unknowns in the discretized equations.

The multigrid method has two main features: smoothing on the current grid and error correction on the coarse grid. The smoothing step has the effect of damping the oscillatory part of the error. The smooth part of the error can then be corrected on the coarse grid.

5.1. The model problem

Let $\Omega \subset \mathbb{R}^d$ $(d = 1, 2, 3)$ be a convex polyhedral domain and

$$a(u, v) = \int_\Omega (\alpha \nabla u \cdot \nabla v + \beta uv)\,\mathrm{d}x \tag{5.1}$$

where α and β are smooth functions such that for some $\alpha_0, \alpha_1, \beta_1 \in \mathbb{R}^+$ we have $\alpha_0 \leqslant \alpha(x) \leqslant \alpha_1$ and $0 \leqslant \beta(x) \leqslant \beta_1$ for all $x \in \Omega$. We consider the Dirichlet problem: Find $u \in V = H_0^1(\Omega)$ such that

$$a(u, v) = (f, v) \qquad \forall\ v \in V, \tag{5.2}$$

where $f \in L^2(\Omega)$ and $(\cdot, \cdot)$ denotes the L^2 inner product.

Let $\mathcal{M}_k$ be a sequence of meshes of Ω obtained successively by standard uniform refinements. Let V_k be the H^1-conforming linear finite element space over $\mathcal{M}_k$ whose functions vanish on $\partial\Omega$. The discrete problem on V_k is then to find $u_k \in V_k$ such that

$$a(u_k, v_k) = (f, v_k) \qquad \forall\ v_k \in V_k. \tag{5.3}$$

We introduce the L^2 and H^1 projection operators

$$(Q_k\varphi, v_k) = (\varphi, v_k)\,, \quad a(P_k\psi, v_k) = a(\psi, v_k) \qquad \forall\ v_k \in V_k,$$

where $\varphi \in L^2(\Omega)$ and $\psi \in H_0^1(\Omega)$. Then by using the Aubin-Nitsche trick (cf. Section 3.3) we have

$$\|w - P_k w\|_{L^2(\Omega)} \leqslant Ch_k \|w\|_A \quad \forall w \in H_0^1(\Omega),$$

where $h_k = \max_{K \in \mathcal{M}_k} h_K$ and $\|\cdot\|_A = a(\cdot,\cdot)^{1/2}$. From $v - P_{k-1}v = (I - P_{k-1})(I - P_{k-1})v$, we then have the following approximation property

$$\|(I - P_{k-1})v\|_{L^2(\Omega)} \leqslant Ch_k \|(I - P_{k-1})v\|_A \qquad \forall\ v \in V_k. \tag{5.4}$$

5.2. Iterative methods

Let $A_k : V_k \to V_k$ be defined by

$$(A_k w_k, v_k) = a(w_k, v_k) \qquad \forall\ v_k \in V_k.$$

Then the finite element scheme (5.3) can be rewritten in the form

$$A_k u_k = f_k := Q_k f. \tag{5.5}$$

Let $\{\phi_k^i : \ i = 1, \cdots, n_k\}$ denote the nodal basis of V_k. Given any $v_k = \sum_{i=1}^{n_k} v_{k,i}\phi_k^i \in V_k$, define $\widetilde{v}_k, \widetilde{\widetilde{v}}_k \in \mathbb{R}^{n_k}$ as follows

$$(\widetilde{v}_k)_i = v_{k,i}, \quad (\widetilde{\widetilde{v}}_k)_i = (v_k, \phi_k^i), \quad i = 1, \cdots, n_k. \tag{5.6}$$

Let $\widetilde{A}_k = \left[a(\phi_k^j, \phi_k^i)\right]_{i,j=1}^{n_k}$ be the stiffness matrix. We have the following matrix representation of (5.5):

$$\widetilde{A}_k \widetilde{u}_k = \widetilde{\widetilde{f}}_k, \tag{5.7}$$

We want to consider the following linear iterative method for (5.7): Given $\widetilde{u}^{(0)} \in \mathbb{R}^{n_k}$

$$\widetilde{u}^{(n+1)} = \widetilde{u}^{(n)} + \widetilde{R}_k(\widetilde{\widetilde{f}}_k - \widetilde{A}_k \widetilde{u}^{(n)}), \quad n = 0, 1, 2, \cdots. \tag{5.8}$$

$\widetilde{R}_k$ is called the iterator of $\widetilde{A}_k$. Note that (5.8) converges if the spectral radius $\rho(I - \widetilde{R}_k \widetilde{A}_k) < 1$. If we define a linear operator $R_k : V_k \mapsto V_k$ as

$$R_k g = \sum_{i,j=1}^{n_k} (\widetilde{R}_k)_{ij}(g, \phi_k^j)\phi_k^i, \tag{5.9}$$

then $\widetilde{R_k g} = \widetilde{R}_k \widetilde{\widetilde{g}}$, so that the algorithm (5.8) for the matrix equation (5.7) is equivalent to the following linear iterative algorithm for the operator equation (5.5): Given $u^{(0)} \in V_k$

$$u^{(n+1)} = u^{(n)} + R_k(f_k - A_k u^{(n)}), \quad n = 0, 1, 2, \cdots.$$

Here we have used the fact that $\widetilde{\widetilde{A_k u^{(n)}}} = \widetilde{A}_k \widetilde{u}^{(n)}$. It is clear that the error propagation operator is $I - R_k A_k$.

Noting that $\widetilde{A}_k$ is symmetric and positive definite, we write $\widetilde{A}_k = \widetilde{D} - \widetilde{L} - \widetilde{L}^T$ with $\widetilde{D}$ and $-\widetilde{L}$ being the diagonal and the lower triangular part of $\widetilde{A}_k$ respectively. We recall the following choices of $\widetilde{R}_k$ that result in various different iterative methods:

$$\widetilde{R}_k = \begin{cases} \frac{\omega}{\rho(\widetilde{A}_k)} I & \text{Richardson;} \\ \omega \widetilde{D}^{-1} & \text{Damped Jacobi;} \\ (\widetilde{D} - \widetilde{L})^{-1} & \text{Gauss-Seidel;} \\ (\widetilde{D} - \widetilde{L})^{-T} \widetilde{D} (\widetilde{D} - \widetilde{L})^{-1} & \text{Symmetrized Gauss-Seidel.} \end{cases} \tag{5.10}$$

Lemma 5.1. *We have*

(i) *The Richardson method converges if and only if* $0 < \omega < 2$*;*
(ii) *The Damped Jacobi method converges if and only if* $0 < \omega < \frac{2}{\rho(\widetilde{D}^{-1}\widetilde{A}_k)}$*;*
(iii) *The Gauss-Seidel method and symmetrized Gauss-Seidel method always converge.*

Lemma 5.2. *The damped Jacobi iterative method for solving* (5.7) *is equivalent to the following iterative scheme in the space* V_k :

$$u_k^{(n+1)} = u_k^{(n)} + R_k(f_k - A_k u_k^{(n)}) \quad , \quad R_k = \omega \sum_{i=1}^{n_k} P_k^i A_k^{-1},$$

where P_k^i *is the projection operator to the subspace spanned by* $\{\phi_k^i\}$*:*

$$a(P_k^i w_k, \phi_k^i) = a(w_k, \phi_k^i) \qquad \forall\ w_k \in V_k. \tag{5.11}$$

Proof. From (5.11) we know that

$$P_k^i w_k = \frac{a(w_k, \phi_k^i)}{a(\phi_k^i, \phi_k^i)} \phi_k^i, \quad i = 1, 2, \cdots, n_k.$$

Recall that the iterator of the damped Jacobi iterative method is $\widetilde{R}_k = \omega \widetilde{D}^{-1} = \operatorname{diag}\big(\omega / a(\phi_k^1, \phi_k^1), \cdots, \omega / a(\phi_k^{n_k}, \phi_k^{n_k})\big)$. It follows from (5.9) that

$$R_k g = \omega \sum_{i=1}^{n_k} \frac{(g, \phi_k^i)}{a(\phi_k^i, \phi_k^i)} \phi_k^i = \omega \sum_{i=1}^{n_k} \frac{a(A_k^{-1} g, \phi_k^i)}{a(\phi_k^i, \phi_k^i)} \phi_k^i = \omega \sum_{i=1}^{n_k} P_k^i A_k^{-1} g \quad \forall g \in V_k,$$

which completes the proof. □

Lemma 5.3. *The standard Gauss-Seidel iterative method for solving* (5.7) *is equivalent to the following iterative scheme in the space* V_k :

$$u_k^{(n+1)} = u_k^{(n)} + R_k(f_k - A_k u_k^{(n)}), \quad R_k = (I - E_k)A_k^{-1},$$

where $E_k = (I - P_k^{n_k}) \cdots (I - P_k^1)$.

Lemma 5.4. *The symmetrized Gauss-Seidel iterative method for solving* (5.7) *is equivalent to the following iterative scheme in the space* V_k :

$$u_k^{(n+1)} = u_k^{(n)} + R_k(f_k - A_k u_k^{(n)}), \quad R_k = (I - E_k^* E_k)A_k^{-1},$$

where $E_k = (I - P_k^{n_k}) \cdots (I - P_k^1)$ *and* $E_k^* = (I - P_k^1) \cdots (I - P_k^{n_k})$ *is the conjugate operator of* E_k *with respect to* $a(\cdot, \cdot)$.

The proofs of Lemma 5.3 and Lemma 5.4 are left as Exercise 5.1.

It is well-known that the classical iterative methods listed in (5.10) are inefficient for solving (5.7) when n_k is large. But they have an important "smoothing property" that we discuss now. For example, Richardson iteration for (5.7) reads as

$$\widetilde{u}^{(n+1)} = \widetilde{u}^{(n)} + \frac{\omega}{\rho(\widetilde{A}_k)}(\widetilde{\widetilde{f}}_k - \widetilde{A}_k \widetilde{u}^{(n)}), \quad n = 0, 1, 2, \cdots .$$

Let $\widetilde{A}_k \widetilde{\phi}_i = \mu_i \widetilde{\phi}_i$ with $\mu_1 \leqslant \mu_2 \leqslant \cdots \leqslant \mu_{n_k}$, $(\widetilde{\phi}_i, \widetilde{\phi}_j) = \delta_{ij}$ and $\widetilde{u}_k - \widetilde{u}^0 = \sum_{i=1}^{n_k} \alpha_i \widetilde{\phi}_i$, then

$$\widetilde{u}_k - \widetilde{u}^{(n)} = \sum_i \alpha_i (1 - \omega \mu_i / \mu_{n_k})^n \widetilde{\phi}_i.$$

For a fixed $\omega \in (0, 2)$, it is clear that $(1 - \omega\mu_i/\mu_{n_k})^n$ converges to zero very fast as $n \to \infty$ if μ_i is close to μ_{n_k}. This means that the high frequency modes in the error get damped very quickly.

Let us illustrate the smoothing property of the Gauss-Seidel method by a simple numerical example. Consider the Poisson equation $-\Delta u = 1$ with homogeneous Dirichlet condition on the unit square which is discretized by the uniform triangulation. Figure 1 shows that high frequency errors are well annihilated by Gauss-Seidel iterations.

For the above model problem, Brandt applied the "local mode analysis" to show that: The damped Jacobi method achieve its optimal smoothing property when $\omega = 4/5$; the Gauss-Seidel method is a better smoother than the damped Jacobi method; the Gauss-Seidel method with red-black ordering is a better smoother than the one with lexicographic ordering. We also

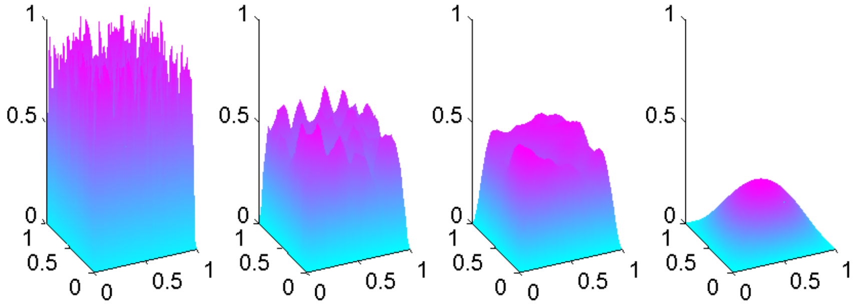

FIGURE 1. Error after $0, 3, 9$ and 200 Gauss-Seidel iterations, respectively, with 2113 unknowns

note that the red-black Gauss-Seidel and Jacobi method have better parallel features.

5.3. The multigrid V-cycle algorithm

The basic idea in a multigrid strategy is that smoothing on the current grid and error correction on a coarser grid. Let $R_k : V_k \to V_k$ be a linear smoother and R_k^t be the adjoint of R_k with respect to $(\cdot, \cdot)$. The multigrid V-cycle algorithm for solving (5.5) can be written as

$$u_k^{(n+1)} = u_k^{(n)} + \mathbb{B}_k(f_k - A_k u_k^{(n)}), \quad n = 0, 1, 2, \cdots, \tag{5.12}$$

where the iterator $\mathbb{B}_k$ is defined by the following algorithm.

ALGORITHM 5.1. (V-cycle iterator). For $k = 1$, define $\mathbb{B}_1 = A_1^{-1}$. Assume that $\mathbb{B}_{k-1} : V_{k-1} \mapsto V_{k-1}$ is defined. For $g \in V_k$, define the iterator $\mathbb{B}_k : V_k \mapsto V_k$ through the following steps.

(1) Pre-smoothing: For $y_0 = 0 \in V_k$ and $j = 1, \cdots, m$,

$$y_j = y_{j-1} + R_k(g - A_k y_{j-1}).$$

(2) Coarse grid correction: $y_{m+1} = y_m + \mathbb{B}_{k-1} Q_{k-1}(g - A_k y_m)$,

(3) Post-smoothing: For $j = m + 2, \cdots, 2m + 1$,

$$y_j = y_{j-1} + R_k^t(g - A_k y_{j-1}).$$

Define $\mathbb{B}_k g = y_{2m+1}$.

In the following, we assume that R_k is symmetric with respect to $(\cdot,\cdot)$ and positive semi-definite. Denote by $y = A_k^{-1}g$, then we have

$$y_{2m+1} - y = (I - R_kA_k)^m(I - \mathbb{B}_{k-1}Q_{k-1}A_k)(I - R_kA_k)^m(y_0 - y).$$

Thus

$$I - \mathbb{B}_kA_k = (I - R_kA_k)^m(I - \mathbb{B}_{k-1}Q_{k-1}A_k)(I - R_kA_k)^m$$

On the other hand, for any $v_k \in V_k, w_{k-1} \in V_{k-1}$, we have

$$\begin{aligned}(Q_{k-1}A_kv_k, w_{k-1}) &= (A_kv_k, w_{k-1}) = a(v_k, w_{k-1})\\ &= a(P_{k-1}v_k, w_{k-1}) = (A_{k-1}P_{k-1}v_k, w_{k-1})\end{aligned}$$

that is, $Q_{k-1}A_k = A_{k-1}P_{k-1}$. Therefore we have the following two-level recurrence relation.

Lemma 5.5. *Let $K_k = I - R_kA_k$. Then*

$$I - \mathbb{B}_kA_k = K_k^m((I - P_{k-1}) + (I - \mathbb{B}_{k-1}A_{k-1})P_{k-1})K_k^m \qquad \text{on } V_k.$$

The following lemma is left as an Exercise 5.2.

Lemma 5.6. *We have*

$$a(K_kv, w) = a(v, K_kw) \quad \text{and} \quad (\mathbb{B}_kv, w) = (v, \mathbb{B}_kw) \qquad \forall\ v, w \in V_k.$$

The following abstract estimate plays an important role in the analysis of multigrid method.

Theorem 5.1. *Assume that $R_k : V_k \to V_k$ is symmetric with respect to $(\cdot,\cdot)$, positive semi-definite, and satisfies*

$$a((I - R_kA_k)v, v) \geqslant 0 \qquad \forall\ v \in V_k. \tag{5.13}$$

Moreover

$$(R_k^{-1}v, v) \leqslant \alpha a(v, v) \qquad \forall\ v \in (I - P_{k-1})V_k. \tag{5.14}$$

Then we have

$$0 \leqslant a((I - \mathbb{B}_kA_k)v, v) \leqslant \delta a(v, v) \qquad \forall\ v \in V_k, \tag{5.15}$$

where $\delta = \alpha/(\alpha + 2m)$.

Proof. We prove by induction, (5.15) is trivial when $k = 1$ since $\mathbb{B}_1 = A_1^{-1}$. Let us now assume (5.15) is true for $k - 1$:

$$0 \leqslant a((I - \mathbb{B}_{k-1}A_{k-1})v, v) \leqslant \delta a(v, v) \qquad \forall\ v \in V_{k-1}. \tag{5.16}$$

Then it follows from Lemma 5.5 that for any $v \in V_k$,

$$\begin{aligned}
&a((I-\mathbb{B}_kA_k)v,v)\\
&\quad = a(K_k^m(I-P_{k-1})K_k^m v,v) + a(K_k^m(I-\mathbb{B}_{k-1}A_{k-1})P_{k-1}K_k^m v,v)\\
&\quad = a((I-P_{k-1})K_k^m v,K_k^m v) + a((I-\mathbb{B}_{k-1}A_{k-1})P_{k-1}K_k^m v,P_{k-1}K_k^m v)\\
&\quad \geqslant a((I-P_{k-1})K_k^m v,K_k^m v) = a((I-P_{k-1})K_k^m v,(I-P_{k-1})K_k^m v)\\
&\quad \geqslant 0.
\end{aligned}$$

For the upper bound, we have

$$\begin{aligned}
a((I-\mathbb{B}_kA_k)v,v) &\leqslant a((I-P_{k-1})K_k^m v,K_k^m v) + \delta a(P_{k-1}K_k^m v,P_{k-1}K_k^m v)\\
&= (1-\delta)a((I-P_{k-1})K_k^m v,K_k^m v) + \delta a(K_k^m v,K_k^m v).
\end{aligned}$$

Now

$$\begin{aligned}
&a((I-P_{k-1})K_k^m v,K_k^m v) = ((I-P_{k-1})K_k^m v,A_kK_k^m v)\\
&\quad = (R_k^{-1}(I-P_{k-1})K_k^m v,R_kA_kK_k^m v)\\
&\quad \leqslant (R_k^{-1}(I-P_{k-1})K_k^m v,(I-P_{k-1})K_k^m v)^{1/2}(R_kA_kK_k^m v,A_kK_k^m v)^{1/2}\\
&\quad \leqslant \sqrt{\alpha}a((I-P_{k-1})K_k^m v,K_k^m v)^{1/2}a((I-K_k)K_k^m v,K_k^m v)^{1/2}.
\end{aligned}$$

Thus

$$a((I-P_{k-1})K_k^m v,K_k^m v) \leqslant \alpha a((I-K_k)K_k^m v,K_k^m v).$$

Since $R_k : V_k \to V_k$ is symmetric and simi-definite, by Lemma 5.6 we know that K_k is symmetric with respect to $a(\cdot,\cdot)$ and $0 \leqslant a(K_kv,v) \leqslant a(v,v)$. Thus the eigenvalues of K_k belong to $[0,1]$. Hence

$$a((I-K_k)K_k^{2m}v,v) \leqslant a((I-K_k)K_k^i v,v) \qquad \forall\ 0 \leqslant i \leqslant 2m,$$

and consequently

$$a((I-K_k)K_k^{2m}v,v) \leqslant \frac{1}{2m}\sum_{i=0}^{2m-1} a((I-K_k)K_k^i v,v) = \frac{1}{2m}a((I-K_k^{2m})v,v).$$

This yields

$$\begin{aligned}
a((I-\mathbb{B}_kA_k)v,v) &\leqslant (1-\delta)\frac{\alpha}{2m}a(v,v) + \left(\delta - \frac{\alpha}{2m}(1-\delta)\right)a(K_k^m v,K_k^m v)\\
&= \frac{\alpha}{\alpha+2m}a(v,v).
\end{aligned}$$

This completes the proof of the theorem. □

Now we define the smoothers which satisfy the assumptions in Theorem 5.1. Let

$$V_k = \sum_{i=1}^{K} V_k^i$$

and denote by $P_k^i : V_k \to V_k^i$ the projection

$$a(P_k^i v, w) = a(v, w) \qquad \forall\ w \in V_k^i,\ \forall\ v \in V_k.$$

We introduce the following additive and multiplicative Schwarz operator

$$R_k^a = \sum_{i=1}^{K} P_k^i A_k^{-1}$$

and

$$R_k^m = (I - E_k^* E_k) A_k^{-1},$$

where $E_k = (I - P_k^K) \cdots (I - P_k^1)$ and $E_k^* = (I - P_k^1) \cdots (I - P_k^K)$ is the conjugate operator of E_k with respect to $a(\cdot, \cdot)$.

If $K = n_k$ and $V_k^i = \operatorname{span}\{\phi_k^i\}$, then, from Lemma 5.2 and Lemma 5.4, R_k^a and R_k^m are the iterators of the Jacobi method and the symmetrized Gauss-Seidel method, respectively.

THEOREM 5.2. *Assume there exist constants $\beta, \gamma > 0$ such that*

(i) $$\sum_{i=1}^{K}\sum_{j=1}^{K} a(v_k^i, w_k^j) \leqslant \beta \Big(\sum_{i=1}^{K} a(v_k^i, v_k^i)\Big)^{\frac{1}{2}} \Big(\sum_{j=1}^{K} a(w_k^j, w_k^j)\Big)^{\frac{1}{2}}, \qquad \forall v_k^i, w_k^i \in V_k^i;$$

(ii) $$\inf_{\substack{v=\sum_{i=1}^{K} v_k^i \\ v_k^i \in V_k^i}} \sum_{i=1}^{K} a(v_k^i, v_k^i) \leqslant \gamma a(v, v) \quad \forall\ v \in (I - P_{k-1}) V_k.$$

Then

1°) *For $\omega \leqslant 1/\beta$, $R_k = \omega R_k^a$ satisfies the assumptions in Theorem 5.1 with $\alpha = \gamma/\omega$.*

2°) *$R_k = R_k^m$ satisfies the assumptions in Theorem 5.1 with $\alpha = \beta^2 \gamma$.*

PROOF. 1°) For any $v \in V_k$,

$$\begin{aligned}a(R_k^a A_k v, v) = a\left(\sum_{i=1}^K P_k^i v, v\right) &\leqslant a(v,v)^{1/2} a\left(\sum_{i=1}^K P_k^i v, \sum_{i=1}^K P_k^i v\right)^{1/2}\\ &\leqslant \beta^{1/2} a(v,v)^{1/2}\left(\sum_{i=1}^K a(P_k^i v, v)\right)^{1/2}\\ &= \beta^{1/2} a(v,v)^{1/2} a(R_k^a A_k v, v)^{1/2}.\end{aligned}$$

Thus

$$a\big((I-\omega R_k^a A_k)v, v\big) \geqslant 0 \qquad \text{if} \qquad \omega \leqslant 1/\beta.$$

Next we prove (5.14) by showing that

$$((R_k^a)^{-1}v, v) = \inf_{\substack{v=\sum_{i=1}^K v_k^i\\ v_k^i \in V_k^i}} \sum_{i=1}^K a(v_k^i, v_k^i) \quad \text{for any } v \in (I-P_{k-1})V_k. \tag{5.17}$$

Denote by $\Theta = R_k^a$, and $v = \sum_{i=1}^K v_k^i$, $v_k^i \in V_k^i$. Then

$$\begin{aligned}(\Theta^{-1}v, v) &= \sum_{i=1}^K (\Theta^{-1}v, v_k^i) = \sum_{i=1}^K a(A_k^{-1}\Theta^{-1}v, v_k^i) = \sum_{i=1}^K a(P_k^i A_k^{-1}\Theta^{-1}v, v_k^i)\\ &\leqslant \left(\sum_{i=1}^K a(P_k^i A_k^{-1}\Theta^{-1}v, P_k^i A_k^{-1}\Theta^{-1}v)\right)^{1/2}\left(\sum_{i=1}^K a(v_k^i, v_k^i)\right)^{1/2}\\ &= \left(\sum_{i=1}^K (P_k^i A_k^{-1}\Theta^{-1}v, \Theta^{-1}v)\right)^{1/2}\left(\sum_{i=1}^K a(v_k^i, v_k^i)\right)^{1/2},\end{aligned} \tag{5.18}$$

that is

$$(\Theta^{-1}v, v) \leqslant (v, \Theta^{-1}v)^{1/2}\left(\sum_{i=1}^K a(v_k^i, v_k^i)\right)^{1/2}.$$

Thus

$$(\Theta^{-1}v, v) \leqslant \sum_{i=1}^K a(v_k^i, v_k^i) \qquad \forall\, v = \sum_{i=1}^K v_k^i.$$

To show the equality in (5.17) we only need to take $v_k^i = P_k^i A_k^{-1}\Theta^{-1}v$. This proves the assertion for R_k^a.

2°) Since $R_k^m = (I - E_k^* E_k)A_k^{-1}$, we have

$$a\big((I - R_k^m A_k)v, v\big) = a(E_k v, E_k v) \geqslant 0.$$

Note that (5.18) holds for any invertible operator on V_k. By letting $\Theta = R_k^m$ in (5.18) we have

$$\big((R_k^m)^{-1}v, v\big) \leqslant \big(R_k^a(R_k^m)^{-1}v, (R_k^m)^{-1}v\big)^{1/2} \left(\sum_{i=1}^{K} a(v_k^i, v_k^i)\right)^{1/2}.$$

It follows from (ii) that

$$\big((R_k^m)^{-1}v, v\big) \leqslant \gamma^{1/2}\big(R_k^a(R_k^m)^{-1}v, (R_k^m)^{-1}v\big)^{1/2} a(v,v)^{1/2}. \tag{5.19}$$

Now we show

$$\big(R_k^a v, v\big) \leqslant \beta^2\big(R_k^m v, v\big), \qquad \forall\ v \in V_k. \tag{5.20}$$

Denote by $y = A_k^{-1}v$, then

$$\big(R_k^m v, v\big) = \big((I - E^*E)A_k^{-1}v, v\big) = a\big((I - E^*E)y, y\big) = a(y,y) - a(E_k y, E_k y)$$

Let $E_k^0 = I$ and $E_k^i = (I - P_k^i)\cdots(I - P_k^1),\ i = 1, \cdots, K$. Then

$$E_k^i = (I - P_k^i)E_k^{i-1} \quad \text{and} \quad E_k^K = E_k.$$

Therefore

$$\begin{aligned} a(E_k^i y, E_k^i y) &= a\big((I - P_k^i)E_k^{i-1}y, E_k^{i-1}y\big) \\ &= a(E_k^{i-1}y, E_k^{i-1}y) - a(P_k^i E_k^{i-1}y, E_k^{i-1}y), \end{aligned}$$

which yields

$$a(E_k y, E_k y) = a(y,y) - \sum_{i=1}^{K} a(P_k^i E_k^{i-1}y, E_k^{i-1}y).$$

Consequently,

$$\big(R_k^m v, v\big) = \sum_{i=1}^{K} a(P_k^i E_k^{i-1}y, E_k^{i-1}y).$$

On the other hand,

$$\big(R_k^a v, v\big) = \sum_{i=1}^{K} \big(P_k^i A_k^{-1}v, v\big) = \sum_{i=1}^{K} a(P_k^i y, y) = \sum_{i=1}^{K} a(P_k^i y, P_k^i y).$$

We deduce from $E_k^j = E_k^{j-1} - P_k^j E_k^{j-1}$ that

$$E_k^i y = y - \sum_{j=1}^{i} P_k^j E_k^{j-1}y$$

and thus

$$P_k^i y = P_k^i E_k^i y + P_k^i \sum_{j=1}^{i} P_k^j E_k^{j-1}y = P_k^i \sum_{j=1}^{i} P_k^j E_k^{j-1}y.$$

Now

$$\begin{aligned}\big(R_k^a v, v\big) &= \sum_{i=1}^K a(P_k^i y, \sum_{j=1}^i P_k^j E_k^{j-1} y) = \sum_{i=1}^K \sum_{j=1}^i a(P_k^i y, P_k^j E_k^{j-1} y)\\ &\leqslant \beta \left(\sum_{i=1}^K a(P_k^i y, P_k^i y)\right)^{1/2} \left(\sum_{j=1}^K a(P_k^j E_k^{j-1} y, P_k^j E_k^{j-1} y)\right)^{1/2}\\ &= \beta\big(R_k^a v, v\big)^{1/2}\big(R_k^m v, v\big)^{1/2}.\end{aligned}$$

This proves (5.20). Finally, we deduce from (5.19) that

$$\big((R_k^m)^{-1} v, v\big) \leqslant \beta\gamma^{1/2}\big((R_k^m)^{-1} v, v\big)^{1/2} a(v, v).$$

This completes the proof of the theorem. □

5.4. The finite element multigrid V-cycle algorithm

Now we apply the abstract result in last section to solve the discrete elliptic problem (5.3). Let $K = n_k$ and $V_k^i = \text{span}\{\phi_k^i\}$, the subspace spanned by nodal basis function $\phi_k^i, i = 1, 2, \cdots, n_k$. Then the condition (i) in Theorem 5.2 is easily satisfied by the local property of finite element nodal basis functions. For any $v \in (I - P_{k-1})V_k$, it remains to find the decomposition $v = \sum_{i=1}^{n_k} v_k^i$, $v_k^i \in V_k^i$, so that (ii) of Theorem 5.2 is satisfied. To do so, we take the canonical decomposition $v = \sum_{i=1}^{n_k} v(x_k^i)\phi_k^i$ with $v_k^i = v(x_k^i)\phi_k^i \in V_k^i$. It is easy to see that

$$\sum_{i=1}^{n_k} \|v_k^i\|_{L^2(\Omega)}^2 \leqslant C \sum_{i=1}^{n_k} h_k^d v(x_k^i)^2 \leqslant C\|v\|_{L^2(\Omega)}^2$$

by the scaling argument. Thus by the inverse estimate and (5.4) we get

$$\sum_{i=1}^{n_k} a(v_k^i, v_k^i) \leqslant Ch_k^{-2} \sum_{i=1}^{n_k} \|v_k^i\|_{L^2(\Omega)}^2 \leqslant Ch_k^{-2}\|v\|_{L^2(\Omega)}^2 \leqslant Ca(v, v),$$

for any $v \in (I - P_{k-1})V_k$.

Theorem 5.3. *Let $\mathbb{B}_k$ be the standard multigrid V-cycle with symmetric Gauss-Seidel relaxation as the smoothing operator. Then the exists a constant C independent of $\mathcal{M}_k$ and $m \geqslant 1$ such that*

$$\|I - \mathbb{B}_k A_k\|_A \leqslant \frac{C}{C + m},$$

where

$$\|I - \mathbb{B}_k A_k\|_A = \sup_{0\neq v\in V_k} \frac{a((I-\mathbb{B}_k A_k)v, v)}{\|v\|_A}.$$

Example 5.4. Consider the Poisson equation $-\Delta u = 1$ with homogeneous Dirichlet condition on unit square discretized with uniform triangulations. We solve the problem by the V-cycle algorithm (5.12) with zero initial value, Gauss-Seidel smoother ($m = 2$), and stopping rule

$$\|\widetilde{\widetilde{f}}_k - \widetilde{A}_k\widetilde{u}_k^{(n)}\|_\infty / \|\widetilde{\widetilde{f}}_k - \widetilde{A}_k\widetilde{u}_k^{(0)}\|_\infty < 10^{-6}.$$

The initial mesh consists of 4 triangles. Table 1 shows the number of multigrid iterations after 1–10 uniform refinements by the "newest vertex bisection" algorithm. The final mesh consists of 4194304 triangles and 2095105 interior nodes. For an implementation of the V-cycle algorithm we refer to Section 10.4.

N	5	25	113	481	1985	8065	32513	130561	523265	2095105
l	3	6	6	7	7	7	7	7	7	7

Table 1. Number of multigrid iterations (l) versus number of degrees of freedom (N) for Example 5.4.

5.5. The full multigrid and work estimate

We shall now describe a more efficient multigrid technique, called the *full multigrid* cycle. Recall that u_k is the k^{th} level finite element solution. By the convergence theory of finite element methods in Chapter 3 we have the following error estimates:

$$\|u - u_k\|_A \leqslant c_1 h_k, \quad k \geqslant 1, \tag{5.21}$$

where $c_1 > 0$ is a constant independent of k. The full multigrid method (FMG) is based on the following two observations:

(1) $u_{k-1} \in V_{k-1} \subset V_k$ is closed to $u_k \in V_k$ and hence can be used as an initial guess for an iterative scheme for solving u_k;

(2) Each u_k can be solved within its truncation error by a multigrid iterative scheme.

Algorithm 5.2. (FMG).

For $k = 1$, $\hat{u}_1 = A_1^{-1} f_1$;

For $k \geqslant 2$, let $\hat{u}_k = \hat{u}_{k-1}$, and iterate $\hat{u}_k \leftarrow \hat{u}_k + \mathbb{B}_k(f_k - A_k\hat{u}_k)$ for l times.

Denote by $\tilde{h}_k = \max_{K\in\mathcal{M}_k} |K|^{1/d}$. It is clear that there exists a positive number $p > 1$ such that $\tilde{h}_k = \tilde{h}_{k-1}/p$ and that $\tilde{h}_k$ is equivalent to h_k, that is, there exist positive constants c_2 and c_3 depending only on the minimum angle of the meshes such that $c_2\tilde{h}_k \leqslant h_k \leqslant c_3\tilde{h}_k$. The following theorem says that the above full multigrid algorithm can produce results with errors comparable to the errors of the finite element solutions.

Theorem 5.5. *Assume that Theorem 5.1 holds and that $\delta^l < 1/p$. Then*

$$\|u_k - \hat{u}_k\|_A \leqslant \frac{c_3 p\delta^l}{c_2(1-p\delta^l)}\, c_1 h_k, \quad k \geqslant 1.$$

Proof. By Theorem 5.1 we have

$$\|u_k - \hat{u}_k\|_A \leqslant \delta^l \|u_k - \hat{u}_{k-1}\|_A \leqslant \delta^l(\|u_k - u_{k-1}\|_A + \|u_{k-1} - \hat{u}_{k-1}\|_A).$$

Noting that $\|u_1 - \hat{u}_1\|_A = 0$, we conclude that

$$\begin{aligned}\|u_k - \hat{u}_k\|_A &\leqslant \sum_{n=1}^{k-1}(\delta^l)^n \|u_{k-n+1} - u_{k-n}\|_A \leqslant \sum_{n=1}^{k-1}(\delta^l)^n \|u - u_{k-n}\|_A\\ &\leqslant c_1\sum_{n=1}^{k-1}(\delta^l)^n h_{k-n} \leqslant c_1c_3\sum_{n=1}^{k-1}(\delta^l)^n\tilde{h}_{k-n}\\ &\leqslant c_1c_3\tilde{h}_k\sum_{n=1}^{k-1}(p\delta^l)^n \leqslant \frac{c_1c_3}{c_2}\frac{p\delta^l}{1-p\delta^l}h_k.\end{aligned}$$

This completes the proof. □

We now turn our attention to the *work estimate.* It is clear that

$$n_k = \mathrm{dim}V_k \sim \frac{1}{h_k^d} \sim \frac{1}{\tilde{h}_k^d} \sim (p^d)^k. \tag{5.22}$$

Theorem 5.6. *The work involved in the FMG is $O(n_k)$.*

Proof. Let W_k denote the work in the k^{th} level V-cycle iteration. Together, the smoothing and correction steps yield

$$W_k \leqslant Cmn_k + W_{k-1}.$$

Hence

$$W_k \leqslant Cm(n_1 + n_2 + \cdots + n_k) \leqslant Cn_k.$$

Let $\hat{W}_k$ denote the work involved in obtaining $\hat{u}_k$ in the FMG. Then

$$\hat{W}_k \leqslant \hat{W}_{k-1} + lW_k \leqslant \hat{W}_{k-1} + Cn_k.$$

Thus we have

$$\hat{W}_k \leqslant C(n_1 + \cdots + n_k) \leqslant Cn_k.$$

This completes the proof. □

This theorem shows that the FMG has an optimal computational complexity $O(n_k)$ to compute the solution within truncation error. In contrast, the computational complexity of the k^{th} level V-cycle iteration is not optimal, because its number of operations required to compute the solution within truncation error is $O(n_k \log \frac{1}{h_k}) = O(n_k \log n_k)$.

5.6. The adaptive multigrid method

The distinct feature of applying multigrid methods on adaptively refined finite element meshes is that the number of nodes of $\mathcal{M}_k$ may not grow exponentially with respect to the number of mesh refinements k. In practice, local relaxation schemes are used in applying multigrid methods on adaptively refined finite element meshes.

Let $\widetilde{\mathcal{N}}_k$ be the set of nodes on which local Gauss-Seidel relaxation are carried out

$$\widetilde{\mathcal{N}}_k = \{z \in \widetilde{\mathcal{N}}_k : \ z \text{ is a new node or } z \in \mathcal{N}_{k-1} \text{ but } \phi_k^z \neq \phi_{k-1}^z\},$$

where ϕ_k^z is the nodal basis function at the node z in V_k. For convenience we denote $\widetilde{\mathcal{N}}_k = \{x_k^j : \ j = 1, \cdots, \tilde{n}_k\}$. The local Gauss-Seidel iterative operator is given by

$$R_k = (I - (I - P_k^{\tilde{n}_k}) \cdots (I - P_k^1))A_k^{-1}.$$

The following theorem is proved by Wu and Chen [**52**].

THEOREM 5.7. *Let the meshes $\mathcal{M}_k, 0 \leqslant k \leqslant J$, be obtained by the "newest vertex bisection" algorithm in Section 10.4. Let each element $K \in \mathcal{M}_k$ is obtained by refining some element $K' \in \mathcal{M}_{k-1}$ finite number of times so that $h_{K'} \leqslant Ch_K$. Then the standard multigrid V-cycle with local Gauss-Seidel relaxation satisfies*

$$\|I - \mathbb{B}_k A_k\|_A < \delta$$

for some constant $\delta < 1$ independent of k and $\mathcal{M}_k$.

Bibliographic notes. There is a rich literature on the mathematical theory of multigrid methods. We refer to Brandt [**14**], the book Bramble [**12**], and the review paper Xu [**53**] for further mathematical results. Our development in Section 5.3 follows Arnold et al [**3**]. The full multigrid method is introduced in Brandt [**15**]. The convergence of the adaptive multigrid finite element method is considered in Wu and Chen [**52**].

5.7. Exercises

Exercise 5.1. Prove Lemma 5.3 and Lemma 5.4.

Exercise 5.2. Prove Lemma 5.6.

Exercise 5.3. Let R_k be symmetric with respect to $\big(\cdot,\cdot\big)$ and let $K_k = I - R_k A_k$. Then R_k is semi-definite and satisfies

$$a\big(K_k v, v\big) \geqslant 0 \qquad \forall\ v \in V_k$$

is equivalent to

$$\|K_k\|_A \leqslant 1 \quad \text{and} \quad \|I - K_k\|_A \leqslant 1.$$

CHAPTER 6

Mixed Finite Element Methods

In this chapter we consider mixed finite element methods for solving partial differential equations that can be formulated in the variational saddle point form. We first introduce the abstract framework for the approximation of saddle point problems. Then we apply the general results to two examples, the Possion equation in the mixed form and the Stokes problem.

6.1. Abstract framework

Let X, M be two Hilbert spaces and assume

$$a : X \times X \to \mathbb{R}, \quad b : X \times M \to \mathbb{R}$$

are continuous bilinear forms. Let $f \in X'$ and $g \in M'$. We denote both the dual pairing of X with X' and that of M with M' by $\langle \cdot, \cdot \rangle$. We consider the following problem: Find $(u, \lambda) \in X \times M$ such that

$$\begin{aligned} a(u,v) + b(v,\lambda) &= \langle f, v\rangle \qquad \forall v \in X, \\ b(u,\mu) &= \langle g, \mu\rangle \qquad \forall \mu \in M. \end{aligned} \tag{6.1}$$

Define the Lagrange functional:

$$\mathcal{L}(u,\lambda) := \frac{1}{2}a(u,u) - \langle f, u\rangle + \big[b(u,\lambda) - \langle g, \lambda\rangle\big] \qquad \forall (u,\lambda) \in X \times M.$$

It is easy to see that every solution (u, λ) of problem (6.1) must satisfy the saddle point property

$$\mathcal{L}(u,\mu) \leqslant \mathcal{L}(u,\lambda) \leqslant \mathcal{L}(v,\lambda) \qquad \forall (v,\mu) \in X \times M.$$

It is often easier to handle the saddle point equation (6.1) if we reformulate it as operator equations. Introduce

$$A : X \to X' \quad : \quad \langle Au, v\rangle = a(u,v) \qquad \forall v \in X.$$

Similarly, we associate a mapping B and its adjoint mapping B' with the form b:

$$B : X \to M' \quad : \quad \langle Bu, \mu\rangle = b(u,\mu) \qquad \forall \mu \in M,$$

$$B' : M \to X' \quad : \quad \langle B'\lambda, v\rangle = b(v,\lambda) \qquad \forall v \in X.$$

Then (6.1) is equivalent to

$$\begin{aligned} Au + B'\lambda &= f \quad \text{in } X', \\ Bu &= g \quad \text{in } M'. \end{aligned} \tag{6.2}$$

Define

$$V = \ker(B) = \{v \in X : \; b(v,\mu) = 0 \qquad \forall \mu \in M\}. \tag{6.3}$$

Lemma 6.1. *The following assertions are equivalent:*

(i) *There exists a constant* $\beta > 0$ *such that*

$$\inf_{\mu\in M} \sup_{v\in X} \frac{b(v,\mu)}{\|v\|_X \|\mu\|_M} \geqslant \beta; \tag{6.4}$$

(ii) *The operator* $B : \; V^{\perp} \to M'$ *is an isomorphism, and*

$$\|Bv\|_{M'} \geqslant \beta \|v\|_X \quad \forall v \in V^{\perp}; \tag{6.5}$$

(iii) *The operator* $B' : \; M \to V^0 \subset X'$ *is an isomorphism, and*

$$\|B'\mu\|_{X'} \geqslant \beta \|\mu\|_M \quad \forall \mu \in M. \tag{6.6}$$

Here V^0 *is the polar set*

$$V^0 = \{l \in X' : \langle l, v\rangle = 0 \quad \forall v \in V\}.$$

Proof. By Riesz Representation Theorem, there exist canonical isometric isomorphisms

$$\pi_X : X' \to X, \qquad \pi_M : M' \to M$$

such that

$$\begin{aligned} (\pi_X l, v) &= \langle l, v\rangle \qquad \forall v \in X, \; \forall l \in X', \\ (\pi_M g, \mu) &= \langle g, \mu\rangle \qquad \forall \mu \in M, \; \forall \, g \in M'. \end{aligned}$$

It is easy to check that V^0 and $V^{\perp}$ is isomorphic under the mapping π_X. In fact, for any $l \in V^0$, $(\pi_X l, v) = \langle l, v\rangle = 0$ for any $v \in V$. This implies $\pi_X l \in V^{\perp}$. The inverse is also valid.

We prove now the equivalence of (i) and (iii). It is clear that (6.4) is equivalent to (6.6). So we only need to show that $B' : \; M \to V^0 \subset X'$ is an isomorphism. By (6.6) we know that $B' : \; M \to R(B')$ is an isomorphism. We now show $R(B') = V^0$. First we have $R(B')$ is closed and $R(B') \subset V^0$. In fact, for any $v \in V$ and $\mu \in M$, we know $\langle B'\mu, v\rangle = \langle Bv, \mu\rangle = 0$. That is $R(B') \subset V^0$. By isometry π_X we know that $\pi_X R(B')$ is a closed subspace of $\pi_X V^0 = V^{\perp}$. If $v \in \pi_X R(B')^{\perp}$,

$$(\pi_X B'\mu, v) = 0 \quad \forall \mu \in M \qquad \Leftrightarrow \qquad \langle Bv, \mu\rangle = 0 \quad \forall \mu \in M \qquad \Leftrightarrow \qquad v \in V.$$

Thus $V^{\perp} = \pi_X R(B')^{\perp}$. This proves the equivalence of (i) and (iii).

Next we prove the equivalence of (ii) and (iii). We consider the diagram

$$\begin{tikzcd}
V^{\perp} \arrow[r, "B"] \arrow[d, "\pi_X^{-1}"'] & M' \arrow[d, "\pi_M"] \\
X' \supset V^0 & M \arrow[l, "B'"]
\end{tikzcd}$$

For any $v \in V^{\perp}, \pi_X^{-1} v \in V^0$, thus there exists a $\mu \in M$ such that $B'\mu = \pi_X^{-1} v$ and $\|\mu\|_M \leqslant \beta^{-1} \|B'\mu\|_{X'} = \beta^{-1} \|\pi_X^{-1} v\|_{X'} = \beta^{-1} \|v\|_X$. Thus

$$\|Bv\|_{M'} = \sup_{\lambda \in M} \frac{\langle Bv, \lambda \rangle}{\|\lambda\|_M} \geqslant \frac{\langle B'\mu, v \rangle}{\|\mu\|_M} \geqslant \frac{\|v\|_X^2}{\beta^{-1}\|v\|_X} = \beta \|v\|_X.$$

Now we prove $B : V^{\perp} \to M'$ is an isomorphism. First $R(B)$ is a closed subspace in M', which implies $\pi_M R(B)$ is a closed subspace in M. If $\mu \in \pi_M R(B)^{\perp}$, then

$$(\mu, \pi_M Bv) = 0 \quad \forall v \in X \quad \Leftrightarrow \quad \langle B'\mu, v \rangle = 0 \quad \forall v \in X \quad \Leftrightarrow \quad B'\mu = 0.$$

Hence $\mu = 0$ since B' is an isomorphism. This shows (iii) implies (ii). Similarly, one can show (ii) implies (iii). This completes the proof. □

THEOREM 6.1. *Assume that*

(i) *The bilinear form a is V-elliptic, i.e.,*

$$a(v, v) \geqslant \alpha \|v\|_X^2 \qquad \forall v \in V, \quad \textit{for some } \alpha > 0;$$

(ii) *The bilinear form b satisfies the inf-sup condition* (6.4).

Then the saddle point problem (6.1) *has a unique solution* $(u, \lambda) \in X \times M$ *which satisfies*

$$\|u\|_X + \|\lambda\|_M \leqslant C(\|f\|_{X'} + \|g\|_{M'}).$$

PROOF. From Lemma 6.1, $B : V^{\perp} \to M'$ is an isomorphism, there exists an element $u_0 \in V^{\perp}$ such that $Bu_0 = g$ and

$$\|u_0\|_X \leqslant \beta^{-1} \|Bu_0\|_{M'} = \beta^{-1} \|g\|_{M'}.$$

Let $w = u - u_0$, then (6.2) is equivalent to

$$Aw + B'\lambda = f - Au_0\,, \qquad Bw = 0.$$

Since A is V-elliptic, by Lax-Milgram Lemma, there exists a unique $w \in V$ such that $Aw = f - Au_0$ in V' and

$$\alpha \|w\|_X \leqslant \|f\|_{X'} + C\|u_0\|_X \leqslant \|f\|_{X'} + \beta^{-1} C \|g\|_{M'}.$$

Finally, since $f - Au_0 - Aw \in V^0$, by Lemma 6.1, there exists a unique $\lambda \in M$ such that $B'\lambda = f - Au_0 - Aw$ and

$$\beta \|\lambda\|_M \leqslant \|B'\lambda\|_{X'} = \|f - Au_0 - Aw\|_{X'} \leqslant \|f\|_{X'} + C\|u_0 + w\|_X.$$

Thus $(u, \lambda) = (u_0 + w, \lambda) \in X \times M$ is the solution of (6.1) and satisfies

$$\begin{aligned} \|u\|_X &\leqslant \beta^{-1}\|g\|_{M'} + \alpha^{-1}\|f\|_{X'} + \alpha^{-1}\beta^{-1}C\|g\|_{M'} \\ &= \alpha^{-1}\|f\|_{X'} + (1 + \alpha^{-1}C)\beta^{-1}\|g\|_{M'}, \\ \|\lambda\|_M &\leqslant \beta^{-1}(1 + \alpha^{-1}C)\|f\|_{X'} + (1 + \alpha^{-1}C)C\beta^{-2}\|g\|_{M'}. \end{aligned}$$

This completes the proof. □

Now we choose finite dimensional subspaces $X_h \subset X$, $M_h \subset M$ and solve the discrete approximation problem: Find $(u_h, \lambda_h) \in X_h \times M_h$ such that

$$\begin{aligned} a(u_h, v_h) + b(v_h, \lambda_h) &= \langle f, v_h \rangle \qquad \forall v_h \in X_h, \\ b(u_h, \mu_h) &= \langle g, \mu_h \rangle \qquad \forall \mu_h \in M_h. \end{aligned} \tag{6.7}$$

Define

$$B_h : X_h \to M_h' \quad : \quad \langle B_h u_h, \mu_h \rangle = b(u_h, \mu_h) \qquad \forall \mu_h \in M_h,$$

and

$$V_h = \ker(B_h) = \{v_h \in X_h : \ b(v_h, \mu_h) = 0 \qquad \forall \mu_h \in M_h\}.$$

Theorem 6.2. *Assume that there exist positive constants α_h and β_h such that*

(i) *The bilinear form a is V_h-elliptic, i.e.,*

$$a(v_h, v_h) \geqslant \alpha_h \|v_h\|_X^2 \qquad \forall v_h \in V_h; \tag{6.8}$$

(ii) *The bilinear form b satisfies the inf-sup condition:*

$$\inf_{\mu_h \in M_h} \sup_{v_h \in X_h} \frac{b(v_h, \mu_h)}{\|v_h\|_X \|\mu_h\|_M} \geqslant \beta_h. \tag{6.9}$$

Then the discrete problem (6.7) *has a unique solution $(u_h, \lambda_h) \in X_h \times M_h$ which satisfies*

$$\|u - u_h\|_X + \|\lambda - \lambda_h\|_M \leqslant C \left(\inf_{v_h \in X_h} \|u - v_h\|_X + \inf_{\mu_h \in M_h} \|\lambda - \mu_h\|_M \right).$$

Here the constant C depends on α_h, β_h.

Proof. Introduce the set

$$Z_h(g) = \{w_h \in X_h : b(w_h, \mu_h) = \langle g, \mu_h \rangle \ \ \forall \mu_h \in M_h\}. \tag{6.10}$$

Clearly, $Z_h(g)$ is non-empty because B_h is surjective. Let v_h be arbitrary in X_h. Since B_h verifies (6.9), the reciprocal of Lemma 6.1 (ii) implies the existence of $r_h \in X_h$ such that

$$b(r_h, \mu_h) = b(u - v_h, \mu_h) \;\; \forall \mu_h \in M_h, \quad \beta_h \left\| r_h \right\|_X \leqslant C \left\| u - v_h \right\|_X . \qquad (6.11)$$

It is clear that $r_h + v_h \in Z_h(g)$. Let $w_h = r_h + v_h$, $y_h = u_h - w_h$. Then $y_h \in V_h$, which implies

$$\begin{aligned} \alpha_h \left\| y_h \right\|_X^2 \leqslant a(u_h - w_h, y_h) &= a(u_h - u, y_h) + a(u - w_h, y_h) \\ &= b(y_h, \lambda - \lambda_h) + a(u - w_h, y_h) \\ &= b(y_h, \lambda - \mu_h) + a(u - w_h, y_h) \\ &\leqslant C \left\| y_h \right\|_X \left\| \lambda - \mu_h \right\|_M + C \left\| u - w_h \right\|_X \left\| y_h \right\|_X . \end{aligned}$$

Therefore

$$\left\| y_h \right\|_X \leqslant C \left\| \lambda - \mu_h \right\|_M + C \left\| u - w_h \right\|_X .$$

It follows from the triangle inequality and (6.11) that

$$\left\| u - u_h \right\|_X = \left\| u - v_h - r_h - y_h \right\|_X \leqslant C \left\| \lambda - \mu_h \right\|_M + C \left\| u - v_h \right\|_X .$$

We now estimate $\lambda - \lambda_h$. Since $b(v_h, \lambda - \lambda_h) = a(u_h - u, v_h)$ for all $v_h \in X_h$, we have for any $\mu_h \in M_h$,

$$b(v_h, \mu_h - \lambda_h) = a(u_h - u, v_h) + b(v_h, \mu_h - \lambda) \qquad \forall v_h \in X_h.$$

Combining (6.9) then implies

$$\left\| \mu_h - \lambda_h \right\|_M \leqslant C \left\| u - u_h \right\|_X + C \left\| \lambda - \mu_h \right\|_M \leqslant C \left\| u - v_h \right\|_X + C \left\| \lambda - \mu_h \right\|_M ,$$

which completes the proof of the theorem. □

In the practical application, the verification of the inf-sup condition (6.9) can be done through the following lemma due to Fortin.

Lemma 6.2. *Suppose that the bilinear form $b : X \times M \to \mathbb{R}$ satisfies the inf-sup condition. In addition, suppose that for the subspaces X_h, M_h, there exists a bounded linear projection $\pi_h : X \to X_h$ such that*

$$b(v - \pi_h v, \mu_h) = 0 \qquad \forall \mu_h \in M_h.$$

Then if $\|\pi_h\| \leqslant C$ for some constant independent of h, the finite element spaces X_h, M_h satisfy the inf-sup condition in (6.9) with β_h independent of h.

Proof. By the assumption,

$$\beta\|\mu_h\|_M \leqslant \sup_{v\in X}\frac{b(v,\mu_h)}{\|v\|_X} = \sup_{v\in X}\frac{b(\pi_h v,\mu_h)}{\|v\|_X} \leqslant C\sup_{v\in X}\frac{b(\pi_h v,\mu_h)}{\|\pi_h v\|_X}$$
$$\leqslant C\sup_{v_h\in X_h}\frac{b(v_h,\mu_h)}{\|v_h\|_X}.$$

This proves the lemma. □

6.2. The Poisson equation as a mixed problem

Let Ω be a bounded polygonal domain in $\mathbb{R}^2$. We consider the Dirichlet problem of the Poisson equation

$$-\triangle u = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega.$$

Let $\boldsymbol{\sigma} = \nabla u$, then we have $-\operatorname{div}\boldsymbol{\sigma} = f$. Introduce the Sobolev space

$$H(\operatorname{div};\Omega) = \{\boldsymbol{\tau}\in L^2(\Omega)^2 : \operatorname{div}\boldsymbol{\tau}\in L^2(\Omega)\}$$

with norm

$$\|\boldsymbol{\tau}\|_{H(\operatorname{div},\Omega)} = \big(\,\|\boldsymbol{\tau}\|^2_{L^2(\Omega)} + \|\operatorname{div}\boldsymbol{\tau}\|^2_{L^2(\Omega)}\,\big)^{1/2}.$$

Then the original problem can be put in the mixed form: Find $(\boldsymbol{\sigma},u)\in H(\operatorname{div};\Omega)\times L^2(\Omega)$ such that

$$\begin{aligned}(\boldsymbol{\sigma},\boldsymbol{\tau}) + (\operatorname{div}\boldsymbol{\tau},u) &= 0 && \forall\boldsymbol{\tau}\in H(\operatorname{div};\Omega),\\ (\operatorname{div}\boldsymbol{\sigma},v) &= -(f,v) && \forall v\in L^2(\Omega).\end{aligned}\tag{6.12}$$

Set $X = H(\operatorname{div};\Omega), M = L^2(\Omega)$. Let $a : X\times X\to\mathbb{R}$, $b : X\times M\to\mathbb{R}$ be the bilinear forms

$$a(\boldsymbol{\sigma},\boldsymbol{\tau}) = (\boldsymbol{\sigma},\boldsymbol{\tau}), \qquad b(\boldsymbol{\tau},v) = (\operatorname{div}\boldsymbol{\tau},v).$$

Clearly the forms a, b are continuous. From (6.3), $V = \{\boldsymbol{\tau}\in X : \operatorname{div}\boldsymbol{\tau} = 0\}$. Therefore, for any $\boldsymbol{\tau}\in V$, we have

$$a(\boldsymbol{\tau},\boldsymbol{\tau}) = \|\boldsymbol{\tau}\|^2_{L^2(\Omega)} = \|\boldsymbol{\tau}\|^2_X.$$

Thus a is elliptic in the kernel V. It remains to check the inf-sup condition for b. For any $v\in L^2(\Omega)$, let $w\in H^1_0(\Omega)$ be the weak solution of $-\Delta w = v$ in Ω. Then $\boldsymbol{\tau} = -\nabla w\in H(\operatorname{div};\Omega)$ satisfies

$$\|\boldsymbol{\tau}\|_X = \|-\Delta w\|_{L^2(\Omega)} + \|\boldsymbol{\tau}\|_{L^2(\Omega)} \leqslant C\|v\|_{L^2(\Omega)}.$$

Thus

$$\sup_{\boldsymbol{\tau}'\in X}\frac{b(\boldsymbol{\tau}',v)}{\|\boldsymbol{\tau}'\|_X} \geqslant \frac{(\operatorname{div}\boldsymbol{\tau},v)}{\|\boldsymbol{\tau}\|_X} \geqslant \frac{\|v\|^2_{L^2(\Omega)}}{C\|v\|_{L^2(\Omega)}} = \beta\|v\|_{L^2(\Omega)}.\tag{6.13}$$

This shows that (6.12) has a unique solution $(\boldsymbol{\sigma}, u) \in X \times M$. It is easy to show that $u \in H_0^1(\Omega)$ by the first equation in (6.12).

We now consider the finite element approximation of (6.12). Let $\mathcal{M}_h$ be a shape regular mesh over Ω. We introduce the Raviart-Thomas element which is conforming in $H(\mathrm{div};\Omega)$. The first hint to proceed is the following result whose proof is similar to Theorem 2.9 and is omitted..

LEMMA 6.3. *Let Ω be a bounded domain. Then a piecewise infinitely differentiable function $\mathbf{v} : \bar{\Omega} \to \mathbb{R}$ belongs to $H(\mathrm{div}\,;\Omega)$ if and only if $\mathbf{v} \cdot \mathbf{n}$ is continuous across any inter-element side.*

DEFINITION 6.3. The lowest order Raviart-Thomas element is a triple $(K, \mathcal{P}, \mathcal{N})$ with the following properties:

(i) $K \subset \mathbb{R}^2$ is a triangle with three edges e_1, e_2, e_3;

(ii) $\mathcal{P} = \{\mathbf{p} \in P_1(K)^2 : \mathbf{p} = \mathbf{a}_K + c_K x, \mathbf{a}_K \in \mathbb{R}^2, c_K \in \mathbb{R}\}$;

(iii) $\mathcal{N} = \{N_i : i = 1, 2, 3\}$ is a basis of $\mathcal{P}'$,

$$N_i(\mathbf{p}) = \frac{1}{|e_i|} \int_{e_i} \mathbf{p} \cdot \mathbf{n}_i \, \mathrm{d}s \quad \forall \mathbf{p} \in \mathcal{P}.$$

Here $\mathbf{n}_i$ is the unit outer normal vector to e_i.

Notice that if $N_i(\mathbf{p}) = 0\,(i = 1, 2, 3)$ for some $\mathbf{p} \in \mathcal{P}$, then

$$\mathrm{div}\,\mathbf{p} = \frac{1}{|K|} \int_K \mathrm{div}\,\mathbf{p} \; \mathrm{d}s = \frac{1}{|K|} \int_{\partial K} \mathbf{p} \cdot \mathbf{n} \, \mathrm{d}s = 0.$$

This implies $c_K = 0$ and consequently $\mathbf{a}_K = 0$. This shows that $\mathcal{N}$ is a basis of $\mathcal{P}'$, the dual space of $\mathcal{P}$.

Next let K be a triangle with vertices $A_i, 1 \leqslant i \leqslant 3$. Let $F_K : \hat{K} \to K$ be the affine transform from the reference element $\hat{K}$ to K

$$x = F_K(\hat{x}) = B_K \hat{x} + \mathbf{b}_K, \quad \hat{x} \in \hat{K}.$$

Notice that the unit outward normal vectors $\mathbf{n}, \hat{\mathbf{n}}$ satisfy

$$\mathbf{n} \circ F_K = (B_K^{-1})^T \hat{\mathbf{n}} / |(B_K^{-1})^T \hat{\mathbf{n}}|.$$

For any scaler function φ defined on K, we associate

$$\hat{\varphi} = \varphi \circ F_K, \qquad \text{that is,} \qquad \hat{\varphi}(\hat{x}) = \varphi(B_K \hat{x} + \mathbf{b}_K).$$

For any vector valued function $\boldsymbol{\sigma}$ defined on K, we associate

$$\hat{\boldsymbol{\sigma}} = B_K^{-1} \boldsymbol{\sigma} \circ F_K, \qquad \text{that is,} \qquad \hat{\boldsymbol{\sigma}}(\hat{x}) = B_K^{-1} \boldsymbol{\sigma}(B_K \hat{x} + \mathbf{b}_K).$$

LEMMA 6.4. *We have*

(i) $\boldsymbol{\sigma} \in \mathcal{P}(K) \Leftrightarrow \hat{\boldsymbol{\sigma}} \in \mathcal{P}(\hat{K})$;
(ii) $N_i(\boldsymbol{\sigma}) = 0 \Leftrightarrow \hat{N}_i(\hat{\boldsymbol{\sigma}}) = 0 \quad \forall \boldsymbol{\sigma} \in \mathcal{P}(K), \quad i = 1, 2, 3$;
(iii) $N_i(\boldsymbol{\sigma}) = 0 \Rightarrow \boldsymbol{\sigma} \cdot \mathbf{n}_i = 0$ on $e_i \quad \forall \boldsymbol{\sigma} \in \mathcal{P}(K), \quad i = 1, 2, 3.$

Proof. (i) If $\boldsymbol{\sigma} = \mathbf{a}_K + c_K x \in \mathcal{P}(K)$, then

$$\hat{\boldsymbol{\sigma}} = B_K^{-1}(\mathbf{a}_K + c_K(B_K\hat{x} + \mathbf{b}_K)) = B_K^{-1}(\mathbf{a}_K + c_K\mathbf{b}_K) + c_K\hat{x} \in \mathcal{P}(K).$$

The proof of the reverse is the same.

(ii) For any $\boldsymbol{\sigma} \in \mathcal{P}(K)$, we have

$$\begin{aligned}\hat{N}_i(\hat{\boldsymbol{\sigma}}) &= \frac{1}{|\hat{e}_i|}\int_{\hat{e}_i} \hat{\boldsymbol{\sigma}} \cdot \hat{\mathbf{n}}_i \mathrm{d}\hat{\mathrm{s}} = \frac{1}{|\hat{e}_i|} \cdot \frac{|\hat{e}_i|}{|e_i|}\int_{e_i} B_K^{-1}\boldsymbol{\sigma} \cdot B_K^T\mathbf{n}_i \cdot |(B_K^{-1})^T\hat{\mathbf{n}}_i|\,\mathrm{d}s \\ &= \frac{1}{|e_i|}\int_{e_i} \boldsymbol{\sigma} \cdot \mathbf{n}_i\,\mathrm{d}s \cdot |(B_K^{-1})^T\hat{\mathbf{n}}_i| = |(B_K^{-1})^T\hat{\mathbf{n}}_i| \cdot N_i(\boldsymbol{\sigma}). \qquad (6.14)\end{aligned}$$

(iii) It is a direct consequence of

$$\hat{\boldsymbol{\sigma}} \cdot \hat{\mathbf{n}}_i = |(B_K^{-1})^T\hat{\mathbf{n}}_i| \ \ \boldsymbol{\sigma} \cdot \mathbf{n}_i.$$

and the corresponding result in the reference element. □

Lemma 6.5. *There exists an operator* $\pi_K : H^1(K)^2 \to \mathcal{P}(K)$ *and a constant* C *such that*

$$\int_{e_i} (\pi_K\boldsymbol{\sigma} - \boldsymbol{\sigma}) \cdot \mathbf{n}_i\,\mathrm{d}s = 0, \quad i = 1, 2, 3, \qquad (6.15)$$

and

$$\|\boldsymbol{\sigma} - \pi_K\boldsymbol{\sigma}\|_{H(\mathrm{div}\,,K)} \leqslant C\frac{h_K^2}{\rho_K}\big(|\boldsymbol{\sigma}|_{H^1(K)} + |\mathrm{div}\,\boldsymbol{\sigma}|_{H^1(K)}\big). \qquad (6.16)$$

Proof. We notice that (6.15) uniquely defines the interpolation operator π_K and $N_i(\boldsymbol{\sigma}) = N_i(\pi_K\boldsymbol{\sigma})$. By (6.14) we know that

$$\hat{N}_i(\widehat{\pi_K\boldsymbol{\sigma}}) = |(B_K^{-1})^T\hat{\mathbf{n}}_i|N_i(\pi_K\boldsymbol{\sigma}) = |(B_K^{-1})^T\hat{\mathbf{n}}_i|N_i(\boldsymbol{\sigma}) = \hat{N}_i(\hat{\boldsymbol{\sigma}}).$$

Thus we have

$$\hat{\pi}_K\hat{\boldsymbol{\sigma}} = \widehat{\pi_K\boldsymbol{\sigma}}.$$

This implies

$$\begin{aligned}\|\boldsymbol{\sigma} - \pi_K\boldsymbol{\sigma}\|_{L^2(K)} &= \frac{|K|^{1/2}}{|\hat{K}|^{1/2}}\|B_K(\hat{\boldsymbol{\sigma}} - \hat{\pi}_K\hat{\boldsymbol{\sigma}})\|_{L^2(\hat{K})} \\ &\leqslant \|B_K\|\frac{|K|^{1/2}}{|\hat{K}|^{1/2}}\|\hat{\boldsymbol{\sigma}} - \hat{\pi}_K\hat{\boldsymbol{\sigma}}\|_{L^2(\hat{K})}.\end{aligned}$$

By Theorem 3.1 and the definition of $\hat{\pi}_K$,

$$\begin{aligned}\|\hat{\boldsymbol{\sigma}} - \hat{\pi}_K\hat{\boldsymbol{\sigma}}\|_{L^2(\hat{K})} &= \inf_{\mathbf{c}\in P_0^2} \|\hat{\boldsymbol{\sigma}} - \mathbf{c} - \hat{\pi}_K(\hat{\boldsymbol{\sigma}} - \mathbf{c})\|_{L^2(\hat{K})} \\ &\leqslant C \inf_{\mathbf{c}\in P_0^2} \big(\|\hat{\boldsymbol{\sigma}} - \mathbf{c}\|_{L^2(\hat{K})} + \|\hat{\boldsymbol{\sigma}} - \mathbf{c}\|_{H^1(\hat{K})}\big) \\ &\leqslant C|\hat{\boldsymbol{\sigma}}|_{H^1(\hat{K})}.\end{aligned}$$

Thus, by Lemma 3.2 and Lemma 3.3,

$$\begin{aligned}\|\boldsymbol{\sigma} - \pi_K\boldsymbol{\sigma}\|_{L^2(K)} &\leqslant C\|B_K\|\frac{|K|^{1/2}}{|\hat{K}|^{1/2}}|\hat{\boldsymbol{\sigma}}|_{H^1(\hat{K})} \leqslant C\|B_K\|^2\|B_K^{-1}\|\ |\boldsymbol{\sigma}|_{H^1(K)} \\ &\leqslant C\frac{h_K^2}{\rho_K}|\boldsymbol{\sigma}|_{H^1(K)}.\end{aligned}$$

On the other hand, from (6.15),

$$\int_K \operatorname{div}\boldsymbol{\sigma}\mathrm{d}x = \int_K \operatorname{div}\pi_K\boldsymbol{\sigma}\mathrm{d}x.$$

Thus

$$\|\operatorname{div}(\boldsymbol{\sigma} - \pi_K\boldsymbol{\sigma})\|_{L^2(K)} \leqslant \inf_{c\in P_0} \|\operatorname{div}\boldsymbol{\sigma} - c\|_{L^2(K)} \leqslant Ch_K|\operatorname{div}\boldsymbol{\sigma}|_{H^1(K)}.$$

This completes the proof. □

We define the finite element spaces

$$\begin{aligned}X_h : &= \{\boldsymbol{\tau} \in H(\mathrm{div};\Omega) :\ \boldsymbol{\tau}|_K \in \mathcal{P}(K)\ \ \forall K \in \mathcal{M}_h\}, \\ M_h : &= \{v \in L^2(\Omega) :\ v|_K \in P_0(K)\ \ \forall K \in \mathcal{M}_h\}.\end{aligned}$$

By Lemma 6.3 and 6.4(iii) we know X_h is well-defined. The discrete problem to approximate (6.12) is: Find $(\boldsymbol{\sigma}_h, u_h) \in X_h \times M_h$ such that

$$\begin{aligned}(\boldsymbol{\sigma}_h, \boldsymbol{\tau}_h) + (\operatorname{div}\boldsymbol{\tau}_h, u_h) &= 0 \qquad\qquad \forall \boldsymbol{\tau}_h \in X_h, \\ (\operatorname{div}\boldsymbol{\sigma}_h, v_h) &= -(f, v_h) \qquad \forall v_h \in M_h.\end{aligned} \tag{6.17}$$

Lemma 6.6. *For any $p \in L^2(\Omega)$, there exists a function $\boldsymbol{\tau} \in H^1(\Omega)^2$ such that* $\operatorname{div}\boldsymbol{\tau} = p$ *and* $\|\boldsymbol{\tau}\|_{H^1(\Omega)} \leqslant C\|p\|_{L^2(\Omega)}$.

Proof. We extend p to be zero outside the domain Ω and denote the extension by $\tilde{p}$. Let B_R be a circle of radius R that includes $\bar{\Omega}$. Let w be the solution of the problem

$$-\Delta w = \tilde{p} \ \text{ in } B_R, \quad w = 0 \ \text{ on } \partial B_R.$$

By the regularity theorem for elliptic equations in Theorem 1.20 we know that $w \in H^2(B_R)$ and $\|w\|_{H^2(\Omega)} \leqslant \|w\|_{H^2(B_R)} \leqslant C\|\tilde{p}\|_{L^2(B_R)} = C\|p\|_{L^2(\Omega)}$. This shows the lemma by setting $\boldsymbol{\tau} = -\nabla w$. □

THEOREM 6.4. *Assume that $u \in H^2(\Omega)$ and $f \in H^1(\Omega)$. Let $\mathcal{M}_h$ be a shape regular mesh of Ω. Then the problem (6.17) has a unique solution and there exists a constant C such that*

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{H(\mathrm{div}\,;\Omega)} + \|u - u_h\|_{L^2(\Omega)} \leqslant Ch(\|u\|_{H^2(\Omega)} + \|f\|_{H^1(\Omega)}).$$

PROOF. The proof is divided into two steps.

1°) Notice that

$$\begin{aligned} V_h &= \{\boldsymbol{\tau}_h \in X_h : \ (\mathrm{div}\,\boldsymbol{\tau}_h, v_h) = 0 \quad \forall v_h \in M_h\} \\ &= \{\boldsymbol{\tau}_h \in X_h : \ \mathrm{div}\,\boldsymbol{\tau}_h = 0 \quad \text{on} \quad K \in \mathcal{M}_h\}. \end{aligned}$$

Thus $a(\cdot,\cdot)$ is V_h-elliptic with the constant $\alpha_h = 1$.

2°) We show

$$\inf_{v_h \in M_h} \sup_{\boldsymbol{\tau}_h \in X_h} \frac{(\mathrm{div}\,\boldsymbol{\tau}_h, v_h)}{\|\boldsymbol{\tau}_h\|_{H(\mathrm{div}\,;\Omega)}\|v_h\|_{L^2(\Omega)}} \geqslant \beta > 0. \tag{6.18}$$

By Lemma 6.6, for any $v_h \in M_h$, there exists a function $\boldsymbol{\tau} \in H^1(\Omega)^2$ such that

$$\mathrm{div}\,\boldsymbol{\tau} = v_h \text{ in } \Omega \quad \text{and} \quad \|\boldsymbol{\tau}\|_{H^1(\Omega)} \leqslant C\|v_h\|_{L^2(\Omega)}.$$

We define the interpolation operator $\pi_h : \ H^1(\Omega)^2 \to X_h$ by using the local operator in Lemma 6.5 to get

$$\int_\Omega \mathrm{div}\,(\pi_h \boldsymbol{\tau})\,\mathrm{d}x = \int_\Omega \mathrm{div}\,\boldsymbol{\tau}\,\mathrm{d}x.$$

Moreover, we have

$$\begin{aligned} \|\pi_h\boldsymbol{\tau}\|_{L^2(\Omega)} &\leqslant \|\boldsymbol{\tau}\|_{L^2(\Omega)} + \|\boldsymbol{\tau} - \pi_h\boldsymbol{\tau}\|_{L^2(\Omega)} \\ &\leqslant \|\boldsymbol{\tau}\|_{L^2(\Omega)} + Ch\|\boldsymbol{\tau}\|_{H^1(\Omega)} \leqslant C\|v_h\|_{L^2(\Omega)}, \\ \|\mathrm{div}\,\pi_h\boldsymbol{\tau}\|_{L^2(\Omega)} &\leqslant \|\mathrm{div}\,\boldsymbol{\tau}\|_{L^2(\Omega)} = \|v_h\|_{L^2(\Omega)}. \end{aligned}$$

Now for any $v_h \in L^2(\Omega)$,

$$\sup_{\boldsymbol{\tau}_h \in X_h} \frac{(\mathrm{div}\,\boldsymbol{\tau}_h, v_h)}{\|\boldsymbol{\tau}_h\|_{H(\mathrm{div}\,;\Omega)}} \geqslant \frac{(\mathrm{div}\,\pi_h\boldsymbol{\tau}, v_h)}{\|\pi_h\boldsymbol{\tau}\|_{H(\mathrm{div}\,;\Omega)}} \geqslant \frac{(\mathrm{div}\,\boldsymbol{\tau}, v_h)}{C\|v_h\|_{L^2(\Omega)}} \geqslant C\|v_h\|_{L^2(\Omega)}.$$

This shows (6.18) and thus completes the proof by using the abstract result Theorem 6.2 and Lemma 6.5. □

6.3. The Stokes problem

Let Ω be a bounded Lipschitz domain in $\mathbb{R}^d (d = 2, 3)$. We consider the Stokes problem

$$\begin{aligned} -\nu\triangle\mathbf{u} + \nabla p &= \mathbf{f} \qquad \text{in} \quad \Omega, \\ \operatorname{div}\mathbf{u} &= 0 \qquad \text{in} \quad \Omega, \\ \mathbf{u} &= 0 \qquad \text{on} \quad \partial\Omega. \end{aligned} \tag{6.19}$$

We set $X = H_0^1(\Omega)^d, M = L_0^2(\Omega)$, and

$$a(u, v) = \nu(\nabla u, \nabla v), \qquad b(\mathbf{v}, q) = -(q, \operatorname{div}\mathbf{v}).$$

Then (6.19) is equivalent to the variational formulation: Find a pair $(\mathbf{u}, p) \in X \times M$ such that

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) &= \langle \mathbf{f}, \mathbf{v} \rangle \quad \forall \mathbf{v} \in X, \\ b(\mathbf{u}, q) &= 0 \quad \forall q \in M. \end{aligned} \tag{6.20}$$

To prove the well-posedness of the above variational problem, we need the following theorem whose proof is outside the scope of this book and is omitted.

THEOREM 6.5 (Nečas). *Let Ω be a bounded Lipschitz domain. There exists a constant $C > 0$, depending only on Ω, such that*

$$\begin{aligned} &\|p\|_{L^2(\Omega)} \leqslant C\big(\|p\|_{H^{-1}(\Omega)} + \|\nabla p\|_{H^{-1}(\Omega)} \big) \quad \forall p \in L^2(\Omega), \\ &\|p\|_{L^2(\Omega)} \leqslant C \, \|\nabla p\|_{H^{-1}(\Omega)} \quad \forall p \in L_0^2(\Omega), \end{aligned}$$

where $H^{-1}(\Omega)$ is the dual space of $H_0^1(\Omega)$ and $\nabla p \in H^{-1}(\Omega)^d$ is defined by

$$\langle \nabla p, \mathbf{v} \rangle = -(p, \operatorname{div}\mathbf{v}) \quad \forall \mathbf{v} \in H_0^1(\Omega)^d. \tag{6.21}$$

By (6.3), $V = \big\{ \mathbf{v} \in H_0^1(\Omega)^d : \operatorname{div}\mathbf{v} = 0 \big\}$. Then the polar set of V is

$$V^0 = \Big\{ \mathbf{l} \in H^{-1}(\Omega)^d : \langle \mathbf{l}, \mathbf{v} \rangle = 0 \quad \forall \mathbf{v} \in V \Big\}.$$

LEMMA 6.7. *Let Ω be a bounded Lipschitz domain. Then*

(i) *The operator ∇ is an isomorphism of $L_0^2(\Omega)$ to V^0;*

(ii) *The operator* div *is an isomorphism of $V^\perp$ to $L_0^2(\Omega)$.*

PROOF. (i) First we know that ∇ is an isomorphism of $L_0^2(\Omega)$ to $R(\nabla) \subset H^{-1}(\Omega)^d$. By Theorem 6.5, $R(\nabla)$ is closed. Since $\langle \nabla p, \mathbf{v} \rangle = -\langle p, \operatorname{div}\mathbf{v} \rangle = 0$ for any $p \in L_0^2(\Omega)$ and $\mathbf{v} \in V$, we know $R(\nabla) \subset V^0$. Let π be the canonical mapping between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$, then $\pi R(\nabla)$ is a closed subspace of

$\pi V^0 = V^\perp$. If there exists a function $\mathbf{v} \in V^\perp, \mathbf{v} \neq 0$ such that $\mathbf{v} \in \pi R(\nabla)^\perp$, that is,

$$0 = (\pi \nabla p, \mathbf{v}) = \langle \nabla p, \mathbf{v} \rangle = -(p, \operatorname{div} \mathbf{v}) \quad \forall p \in L_0^2(\Omega),$$

then $\operatorname{div} \mathbf{v} = 0$, that is, $\mathbf{v} \in V$, a contradiction! Thus $\pi R(\nabla) = V^\perp$ and hence $R(\nabla) = V^0$.

(ii) This is a direct consequence of (i) since div is the dual operator of ∇ and $\pi V^0 = V^\perp$. This completes the proof. □

It is easy to see by Poincaré inequality that $a(\cdot, \cdot)$ is X-elliptic and thus also V-elliptic. The continuous inf-sup condition follows from Lemma 6.7 and thus (6.20) has a unique solution.

In the rest of this section, suppose Ω is a bounded polygonal domain in $\mathbb{R}^2$. Next we consider the finite element approximation of (6.20). Let $\mathcal{M}_h$ be a shape regular mesh. We will approximate the velocity by the "mini" finite element which we now introduce. On each element K we approximate the velocity by a polynomial of the form

$$\mathcal{P}(K) = (P_1 \oplus \{\lambda_1 \lambda_2 \lambda_3\})^2$$

and the pressure by a polynomial of $P_1(K)$. We define

$$\begin{aligned} \tilde{X}_h =& \{v \in C(\bar{\Omega})^2 : v|_K \in \mathcal{P}(K) \quad \forall K \in \mathcal{M}_h\}, \quad X_h = \tilde{X}_h \cap H_0^1(\Omega), \\ \tilde{M}_h =& \{q \in C(\bar{\Omega}) : \ q|_K \in P_1(K) \quad \forall K \in \mathcal{M}_h\}, \quad M_h = \tilde{M}_h \cap L_0^2(\Omega). \end{aligned}$$

The degrees of freedom are the simplest ones, namely the values of the velocity at the vertices and the center of K, the values of the pressure at the vertices of K. The discrete problem is: Find a pair $(u_h, p_h) \in X_h \times M_h$ such that

$$\begin{aligned} a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) = \langle \mathbf{f}, \mathbf{v}_h \rangle \qquad \forall \mathbf{v}_h \in X_h, \\ b(\mathbf{u}_h, q_h) = 0 \qquad \forall q_h \in M_h. \end{aligned}$$

Lemma 6.8. *There exists a constant $\beta > 0$ independent of h such that*

$$\inf_{q_h \in M_h} \sup_{\mathbf{v}_h \in X_h} \frac{(q_h, \operatorname{div} \mathbf{v}_h)}{\|\mathbf{v}_h\|_{H^1(\Omega)}} \geqslant \beta > 0.$$

Proof. We define an operator π_h satisfying the condition in Lemma 6.2. For any $\mathbf{v} \in H_0^1(\Omega)^2$, we want to construct a function $\pi_h \mathbf{v} \in X_h$ that satisfies

$$\int_\Omega \pi_h \mathbf{v} \cdot \nabla \mu_h \, \mathrm{d}x = \int_\Omega \mathbf{v} \cdot \nabla \mu_h \, \mathrm{d}x \qquad \forall \mu_h \in M_h.$$

Since $\nabla\mu_h \in P_0(K)^2$ on each K, this equality induces us to define $\pi_h\mathbf{v}$ in X_h such that

$$(\pi_h\mathbf{v})(a) = (r_h\mathbf{v})(a) \quad \forall \text{ node } a \text{ of } \mathcal{M}_h,$$

and

$$\int_K \pi_h\mathbf{v}\,\mathrm{d}x = \int_K \mathbf{v}\,\mathrm{d}x \qquad \forall K \in \mathcal{M}_h,$$

where r_h is the Clément interpolation operator in Section 4.2. Clearly $\pi_h : H_0^1(\Omega)^2 \to X_h$ is well-defined. Moreover

$$\int_\Omega \operatorname{div}(\pi_h\mathbf{v} - \mathbf{v})\mu_h \mathrm{d}x = 0 \qquad \forall \mu_h \in M_h.$$

On each element $K \in \mathcal{M}_h$, we have

$$\pi_h\mathbf{v}|_K = r_h\mathbf{v}|_K + \boldsymbol{\beta}_K\lambda_1\lambda_2\lambda_3,$$

where

$$\boldsymbol{\beta}_K = \int_K (\mathbf{v} - r_h\mathbf{v})\mathrm{d}x \Big/ \int_K \lambda_1\lambda_2\lambda_3\mathrm{d}x.$$

By the scaling argument and Theorem 4.2

$$|\boldsymbol{\beta}_K| \leqslant Ch_K^{-1}\|\mathbf{v} - r_h\mathbf{v}\|_{L^2(K)} \leqslant C\|\mathbf{v}\|_{H^1(\tilde{K})},$$

where $\tilde{K}$ is the union of all elements having non-empty intersection with K. Hence

$$\|\pi_h\mathbf{v}\|_{H^1(\Omega)} \leqslant C\|\mathbf{v}\|_{H^1(\Omega)}.$$

This proves the lemma by Lemma 6.2. □

The following theorem is a direct consequence of Theorem 6.2.

Theorem 6.6. *Let the solution* $(\mathbf{u}, p)$ *of the Stokes problem satisfy*

$$\mathbf{u} \in \left(H^2(\Omega) \cap H_0^1(\Omega)\right)^2, \qquad p \in H^1(\Omega) \cap L_0^2(\Omega).$$

Then

$$\|\mathbf{u} - \mathbf{u}_h\|_{H^1(\Omega)} + \|p - p_h\|_{L^2(\Omega)} \leqslant Ch(|\mathbf{u}|_{H^2(\Omega)} + |p|_{H^1(\Omega)}).$$

Bibliographic notes. There is a rich literature on mixed finite element methods. We refer to the monographs Girault and Raviart [**34**], Brezzi and Fortin [**16**] for further studies. Theorem 6.5 is due to Nečas [**46**]. Duvaut and Lions [**28**] gives a proof of Theorem 6.5 when the boundary of the domain is smooth.

6.4. Exercises

EXERCISE 6.1. Formulate the mixed formulation of the Neumann Problem

$$-\triangle u = f \quad \text{in} \quad \Omega, \qquad \frac{\partial u}{\partial \mathbf{n}} = g \quad \text{on} \quad \partial\Omega,$$

and prove the unique existence of the solution to the corresponding saddle point problem.

EXERCISE 6.2. For the Stokes problem, let

$$X_h = \{\mathbf{v} \in C(\bar{\Omega})^2 : \ \mathbf{v}|_K \in P_1(K)^2 \quad \forall K \in \mathcal{M}_h, \ \mathbf{v}|_{\partial\Omega} = 0\},$$
$$M_h = \{q \in L_0^2(\Omega) : \ q|_K \in P_0(K) \quad \forall K \in \mathcal{M}_h\}.$$

Does the inf-sup condition

$$\inf_{q \in M_h} \sup_{\mathbf{v} \in X_h} \frac{(q, \operatorname{div} \mathbf{v})}{\|q\|_{L^2(\Omega)} \|\mathbf{v}\|_{H^1(\Omega)}} \geqslant \beta > 0$$

hold?

EXERCISE 6.3. Construct the local nodal basis functions for the lowest order Raviart-Thomas finite element.

CHAPTER 7

Finite Element Methods for Parabolic Problems

In this chapter we consider finite element methods for solving the initial boundary value problem of the following parabolic equation:

$$\begin{aligned} \frac{\partial u}{\partial t} - \sum_{i,j=1}^{d} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}\right) + c(x)u = f \quad &\text{in } \Omega \times (0,T), \\ u = 0 \quad &\text{on } \Gamma \times (0,T), \\ u(\cdot,0) = u_0(\cdot) \quad &\text{in } \Omega, \end{aligned} \tag{7.1}$$

where Ω is a bounded domain in $\mathbb{R}^d$ with boundary Γ, $T > 0$, $u = u(x,t)$, and a_{ij}, c are bounded functions on Ω, $a_{ij} = a_{ji}$, and there exists a constant $\alpha_0 > 0$ such that

$$\sum_{i,j=1}^{d} a_{ij}(x)\xi_i\xi_j \geqslant \alpha_0|\xi|^2, \quad c(x) \geqslant 0 \quad \text{for a.e. } x \in \Omega \text{ and all } \xi \in \mathbb{R}^d. \tag{7.2}$$

7.1. The weak solutions of parabolic equations

We start with introducing the function spaces involving time.

Definition 7.1. Let X be a real Banach space with norm $\|\cdot\|$.

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

$$\|u\|_{L^p(0,T;X)} := \left(\int_0^T \|u(t)\|^p \,\mathrm{d}t\right)^{1/p} < \infty$$

for $1 \leqslant p < \infty$, and

$$\|u\|_{L^\infty(0,T;X)} := \sup_{0\leqslant t\leqslant T} \|u(t)\| < \infty.$$

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

$$\|u\|_{C([0,T];X)} := \max_{0\leqslant t\leqslant T} \|u(t)\| < \infty.$$

Definition 7.2. Let $u \in L^1(0,T;X)$. We say $v \in L^1(0,T;X)$ is the weak derivative of u, written $u' = v$, provided

$$\int_0^T \phi'(t)u(t)\,\mathrm{d}t = -\int_0^T \phi(t)v(t)\,\mathrm{d}t \qquad \forall \phi \in C_0^\infty(0,T).$$

The Sobolev space $W^{1,p}(0,T;X)$ is then defined as the set of all functions $u \in L^p(0,T;X)$ such that u' exists in the weak sense and belongs to $L^p(0,T;X)$. When $p = 2$, we write $H^1(0,T;X) = W^{1,2}(0,T;X)$.

Theorem 7.3. *Let $u \in W^{1,p}(0,T;X)$ for some $1 \leqslant p \leqslant \infty$. Then*

(i) *After possibly being redefined on a set of measure zero $u \in C([0,T];X)$;*
(ii) *$u(t) = u(s) + \int_s^t u'(\tau)\mathrm{d}\tau$ for all $0 \leqslant t \leqslant T$;*
(iii) *We have the estimate*

$$\max_{0\leqslant t\leqslant T} \|u(t)\| \leqslant C\|u\|_{W^{1,p}(0,T;X)},$$

the constant C depending only on T.

The proof of this theorem is left as an exercise.

Theorem 7.4. *Suppose $u \in L^2(0,T;H_0^1(\Omega))$ with $u' \in L^2(0,T;H^{-1}(\Omega))$. Then after possibly redefined on a set of measure zero, $u \in C([0,T];L^2(\Omega))$ and the mapping $t \mapsto \|u(t)\|^2_{L^2(\Omega)}$ is absolutely continuous with*

$$\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)} = 2\langle u'(t), u(t)\rangle \qquad \text{for a.e. } 0 \leqslant t \leqslant T. \tag{7.3}$$

Furthermore, we have the estimate

$$\max_{0\leqslant t\leqslant T} \|u(t)\|_{L^2(\Omega)} \leqslant C\left(\|u\|_{L^2(0,T;H_0^1(\Omega))} + \|u'\|_{L^2(0,T;H^{-1}(\Omega))}\right), \tag{7.4}$$

for constant C dependening only on T.

Proof. 1°) We extend u to be zero on $(-\infty,0)$ and (T,∞) and define the regularization $u_\epsilon = \rho_\epsilon * u$, where ρ_ϵ is the usual mollifier on $\mathbb{R}^1$. It is clear that $u'_\epsilon = \rho_\epsilon * u'$ on $(\epsilon, T-\epsilon)$. Then for $\epsilon, \delta > 0$,

$$\frac{d}{dt}\| u_\epsilon(t) - u_\delta(t) \|^2_{L^2(\Omega)} = 2(u'_\epsilon(t) - u'_\delta(t), u_\epsilon(t) - u_\delta(t)).$$

Thus, for all $0 \leqslant s,t \leqslant T$,

$$\begin{aligned}\| u_\epsilon(t) - u_\delta(t) \|^2_{L^2(\Omega)} =& \| u_\epsilon(s) - u_\delta(s) \|^2_{L^2(\Omega)} \\ &+ 2\int_s^t \langle u'_\epsilon(\tau) - u'_\delta(\tau), u_\epsilon(\tau) - u_\delta(\tau)\rangle \mathrm{d}\tau.\end{aligned} \tag{7.5}$$

Fix any point $s \in (0,T)$ for which $u_\epsilon(s) \to u(s)$ in $L^2(\Omega)$. Hence

$$\begin{aligned}&\limsup_{\epsilon,\delta\to 0}\sup_{0\leqslant t\leqslant T} \| u^\epsilon(t) - u^\delta(t) \|_{L^2(\Omega)}\\ &\leqslant \lim_{\epsilon,\delta\to 0}\int_0^T \left(\| u'_\epsilon(\tau) - u'_\delta(t) \|^2_{H^{-1}(\Omega)} + \| u_\epsilon(\tau) - u_\delta(t) \|^2_{H^1(\Omega)}\right) \mathrm{d}\tau = 0.\end{aligned}$$

This implies there exists a $v \in C([0,T]; L^2(\Omega))$ such that $u_\epsilon \to v$ in $C([0,T]; L^2(\Omega))$. Since $u_\epsilon \to u$ for a.e. t, we conclude $v = u$ a.e. in $(0,T)$.

2°) Similar to (7.5) we have

$$\| u_\epsilon(t) \|^2_{L^2(\Omega)} = \| u_\epsilon(s) \|^2_{L^2(\Omega)} + 2\int_s^t \langle u'_\epsilon(\tau), u_\epsilon(\tau)\rangle \mathrm{d}\tau.$$

By letting $\epsilon \to 0$ we obtain,

$$\| u(t) \|^2_{L^2(\Omega)} = \| u(s) \|^2_{L^2(\Omega)} + 2\int_s^t \langle u'(\tau), u(\tau)\rangle \mathrm{d}\tau \qquad \forall s,t \in [0,T]. \tag{7.6}$$

This proves $\|u(t)\|_{L^2(\Omega)}$ is absolutely continuous and the equality (7.3). (7.4) is a direct consequence of (7.6). □

To define the weak solution of the problem (7.1) we introduce the the bilinear form $a(u,v) : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$

$$a(u,v) = \int_\Omega \left(\sum_{i,j=1}^d a_{ij}(x)\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i} + c(x)uv\right) \mathrm{d}x.$$

It is clear by the assumption (7.2) that a is bounded and V-elliptic, that is, there exist constants $\alpha_0, \beta > 0$ such that

$$|a(u,v)| \leqslant \beta \,\|u\|_{H^1(\Omega)}\,\|v\|_{H^1(\Omega)} \qquad \forall\ u,v \in H^1_0(\Omega), \tag{7.7}$$

and

$$a(v,v) \geqslant \alpha \,\|v\|^2_{H^1(\Omega)} \qquad \forall\ v \in H^1_0(\Omega). \tag{7.8}$$

We have the following definition of weak solutions for parabolic problems.

Definition 7.5. Given $f \in L^2(0,T; H^{-1}(\Omega))$ and $u_0 \in L^2(\Omega)$. We call a function

$$u \in L^2(0,T; H^1_0(\Omega)), \quad \partial_t u \in L^2(0,T; H^{-1}(\Omega)),$$

is a solution of the parabolic initial boundary value problem (7.1) if

(i) $\langle \partial_t u, v\rangle + a(u,v) = \langle f, v\rangle \quad \forall v \in H^1_0(\Omega)$ and a.e. $0 \leqslant t \leqslant T$;

(ii) $u(x,0) = u_0(x)$ in Ω.

By Theorem 7.4 we see $u \in C([0,T];L^2(\Omega))$, and thus the equality in (ii) makes sense.

THEOREM 7.6. *There exists a unique weak solution to the problem* (7.1). *Moreover, the following stability estimate holds*

$$\| \partial_t u \|_{L^2(0,T;H^{-1}(\Omega))} + \| u \|_{L^2(0,T;H^1(\Omega))} \leqslant C\| u_0 \|_{L^2(\Omega)} + C\| f \|_{L^2(0,T;H^{-1}(\Omega))}.$$

PROOF. Let $N \geqslant 1$ be an integer and $\tau = T/N$ be the time step. Denote $t^n = n\tau$, $n = 1, \cdots, N$. For $n = 1, \cdots, N$, we consider the elliptic problem to find $U^n \in H_0^1(\Omega)$ such that

$$(\bar\partial U^n, v) + a(U^n, v) = \langle \bar f^n, v\rangle \quad \forall v \in H_0^1(\Omega), \tag{7.9}$$

where $\bar\partial U^n = (U^n - U^{n-1})/\tau$, $\bar f^n = \frac{1}{\tau}\int_{t^{n-1}}^{t^n} f(x,t)\,\mathrm{d}t$, and $U^0 = u_0$. By (7.2) and the Lax-Milgram Lemma, (7.9) has a unique weak solution $U^n \in H_0^1(\Omega)$. Notice that

$$(\bar\partial U^n, U^n) = \frac{1}{2}\| U^n \|_{L^2(\Omega)}^2 + \frac{1}{2}\| U^{n-1} \|_{L^2(\Omega)}^2 + \frac{1}{2}\| U^n - U^{n-1} \|_{L^2(\Omega)}^2,$$

by taking $v = U^n$ in (7.9), we can obtain the energy estimate

$$\begin{aligned}&\max_{1\leqslant n\leqslant N} \| U^n \|_{L^2(\Omega)}^2 + \sum_{n=1}^N \| U^n - U^{n-1} \|_{L^2(\Omega)}^2 + \sum_{n=1}^N \tau\alpha_0\| U^n \|_{H^1(\Omega)}^2 \\ &\leqslant C\| u_0 \|_{L^2(\Omega)}^2 + C\|f\|_{L^2(0,T;H^{-1}(\Omega))}^2.\end{aligned} \tag{7.10}$$

Now we define the U_τ and $\bar U_\tau$ through the following relations

$$U_\tau(t) = l(t)U^n + (1 - l(t))U^{n-1}, \quad \bar U_\tau(t) = U^n \quad \forall t \in (t^{n-1}, t^n), \quad \forall n \geqslant 1,$$

where $l(t) = (t - t^{n-1})/\tau$. Then by (7.10)

$$\begin{aligned}&\| U_\tau \|_{L^2(0,T;H^1(\Omega))}^2 + \| \bar U_\tau \|_{L^2(0,T;H^1(\Omega))}^2 \\ &\leqslant C\| u_0 \|_{L^2(\Omega)}^2 + C\|f\|_{L^2(0,T;H^{-1}(\Omega))}^2.\end{aligned} \tag{7.11}$$

Moreover, it follows from the equation (7.9) than

$$\| \partial_t U_\tau \|_{L^2(0,T;H^{-1}(\Omega))}^2 \leqslant C\| u_0 \|_{L^2(\Omega)}^2 + C\|f\|_{L^2(0,T;H^{-1}(\Omega))}^2. \tag{7.12}$$

Therefore we know that there exist functions

$$u \in L^2(0,T;H_0^1(\Omega)) \cap H^1(0,T;H^{-1}(\Omega)), \quad \bar u \in L^2(0,T;H_0^1(\Omega))$$

such that

$$\begin{aligned}&U_\tau \to u \quad \text{weakly in } L^2(0,T;H_0^1(\Omega)) \cap H^1(0,T;H^{-1}(\Omega)),\\ &\bar U_\tau \to \bar u \quad \text{weakly in } L^2(0,T;H_0^1(\Omega)).\end{aligned}$$

But by (7.10)

$$\|U_\tau - \bar{U}_\tau\|^2_{L^2(0,T;L^2(\Omega))} \leqslant \sum_{n=1}^{N} \tau \|U^n - U^{n-1}\|^2_{L^2(\Omega)} \leqslant C\tau,$$

which by taking $\tau \to 0$ implies $u = \bar{u}$ a.e. in $\Omega \times (0,T)$.

Now by (7.9) we have

$$(\partial_t U_\tau, v) + a(\bar{U}_\tau, v) = \langle \bar{f}_\tau, v\rangle \quad \forall v \in H^1_0(\Omega) \ \text{ a.e. in } (0,T),$$

where $\bar{f}_\tau = \bar{f}^n$ for $t \in (t^{n-1}, t^n)$. We then have, for any $\phi \in C^\infty_0(0,T)$,

$$\int_0^T \left[(\partial_t U_\tau, v) + a(\bar{U}_\tau, v)\right] \phi \,\mathrm{d}t = \int_0^T \langle \bar{f}_\tau, v\rangle \phi \,\mathrm{d}t \quad \forall v \in H^1_0(\Omega).$$

Let $\tau \to 0$ in above equality, we obtain

$$\int_0^T \left[(\partial_t u, v) + a(u, v)\right] \phi \,\mathrm{d}t = \int_0^T \langle f, v\rangle \phi \,\mathrm{d}t \quad \forall v \in H^1_0(\Omega), \phi \in C^\infty_0(0,T),$$

which implies

$$\langle \partial_t u, v\rangle + a(u, v) = \langle f, v\rangle \quad \forall v \in H^1_0(\Omega) \ \text{ a.e. in } (0,T).$$

This proves the existence of weak solution. The stability estimate of the theorem follows from (7.11)-(7.12) by letting $\tau \to 0$. The uniqueness is a direct consequence of the stability estimate. □

7.2. The semidiscrete approximation

The problem (7.1) will be discretized and analyzed in two steps. In the first step we shall approximate $u(x,t)$ by means of a function $u_h(x,t)$ which, for each fixed t, belongs to a finite element space V^0_h. This function will be a solution of an h-dependent finite system of ordinary differential equations in time and is referred to as a *spatially discrete*, or *semidiscrete*, *solution*. We then proceed to discretize this system in the time to produce a fully discrete time stepping scheme for the approximate solution of (7.1). In this section we consider the spatially semidiscrete approximation and the a priori L^2 and H^1 error estimates.

Suppose that $\Omega \subset \mathbb{R}^d$ is a bounded polyhedral domain and $\mathcal{M}_h$ is a shape regular partition of Ω. Let $V^0_h \subset H^1_0(\Omega)$ be the standard piecewise linear conforming finite element space. We may then pose the approximate problem to find $u_h = u_h(x,t)$, belong to V^0_h for each t, such that

$$(u_{h,t}, v_h) + a(u_h, v_h) = (f, v_h) \quad \forall v_h \in V^0_h, \quad t > 0, \quad u_h(\cdot, 0) = u_{h0}, \quad (7.13)$$

where u_{h0} is some approximation of u_0 in V^0_h.

In terms of the nodal basis $\{\phi_j\}_{j=1}^J$ for V_h^0, our semidiscrete problem may be stated: Find the coefficients $z_j(t)$ in $u_h(x,t) = \sum_{j=1}^J z_j(t)\phi_j(x)$ such that

$$\sum_{j=1}^J z_j'(t)(\phi_j,\phi_i) + \sum_{j=1}^J z_j(t)a(\phi_j,\phi_i) = (f,\phi_i), \quad i = 1,\cdots,J,$$

and, with z_{j0} the components of u_{h0}, $z_j(0) = z_{j0}$ for $j = 1,\cdots,J$. In the matrix notation this may be expressed as

$$Mz'(t) + Az(t) = b(t), \quad t > 0, \text{ with } z(0) = z_0, \tag{7.14}$$

where $M = (m_{ij})$ is the mass matrix with elements $m_{ij} = (\phi_j,\phi_i)$, $A = (a_{ij})$ the stiffness matrix with $a_{ij} = a(\phi_j,\phi_i)$, $b = (b_i)$ the vector with entries $b_i = (f,\phi_i)$, $z(t)$ the vector of unknowns $z_j(t)$, and $z_0 = (z_{j0})$. Since M is positive definite and invertible, the system of ordinary differential equations (7.14) has a unique solution for $t > 0$.

Next we estimate the error between u_h and u. To do so we first prove a stability result for the semidiscrete problem. Throughout this chapter, C will denote a positive generic constant independent of h, t, and can have different values in different places.

Theorem 7.7. *Let $r(t) \in L^2(\Omega)$, and $\theta_h(t) = \theta_h(\cdot,t) \in V_h^0$ satisfies*

$$(\theta_{h,t}, v_h) + a(\theta_h, v_h) = (r, v_h) \quad \forall v_h \in V_h^0, \quad t > 0. \tag{7.15}$$

Then

$$\|\theta_h(t)\|_{L^2(\Omega)} \leqslant \|\theta_h(0)\|_{L^2(\Omega)} + \int_0^t \|r(s)\|_{L^2(\Omega)} \,\mathrm{d}s, \tag{7.16}$$

$$\|\theta_h(t)\|_{H^1(\Omega)} \leqslant C\,\|\theta_h(0)\|_{H^1(\Omega)} + C\Big(\int_0^t \|r(s)\|_{L^2(\Omega)}^2 \,\mathrm{d}s\Big)^{1/2}. \tag{7.17}$$

Proof. We choose $v_h = \theta_h(t)$ in (7.15) and conclude

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\,\|\theta_h(t)\|_{L^2(\Omega)}^2 + a(\theta_h(t),\theta_h(t)) = (r(t),\theta_h(t)).$$

From (7.8) and the Cauchy inequality,

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\,\|\theta_h(t)\|_{L^2(\Omega)}^2 \leqslant \|r(t)\|_{L^2(\Omega)}\,\|\theta_h(t)\|_{L^2(\Omega)}\,.$$

Since $\frac{\mathrm{d}}{\mathrm{d}t} \|\theta_h(t)\|_{L^2(\Omega)}$ might not be differentiable when $\theta_h = 0$, we add ε^2 to obtain

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} \|\theta_h(t)\|^2_{L^2(\Omega)} &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\big(\|\theta_h(t)\|^2_{L^2(\Omega)} + \varepsilon^2\big)\\ &= \big(\|\theta_h(t)\|^2_{L^2(\Omega)} + \varepsilon^2\big)^{1/2} \frac{\mathrm{d}}{\mathrm{d}t}\big(\|\theta_h(t)\|^2_{L^2(\Omega)} + \varepsilon^2\big)^{1/2}\\ &\leqslant \|r(t)\|_{L^2(\Omega)} \|\theta_h(t)\|_{L^2(\Omega)},\end{aligned}$$

and hence

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(\|\theta_h(t)\|^2_{L^2(\Omega)} + \varepsilon^2\big)^{1/2} \leqslant \|r(t)\|_{L^2(\Omega)}.$$

After integration and letting $\varepsilon \to 0$ we conclude that (7.16) holds.

In order to prove (7.17), we use again (7.15), now with $v_h = \theta_{h,t}$, we obtain

$$\|\theta_{h,t}\|^2_{L^2(\Omega)} + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} a(\theta_h, \theta_h) = (r(t), \theta_{h,t}) \leqslant \frac{1}{4} \|r(t)\|^2_{L^2(\Omega)} + \|\theta_{h,t}\|^2_{L^2(\Omega)}.$$

Therefore

$$\frac{\mathrm{d}}{\mathrm{d}t} a(\theta_h(t), \theta_h(t)) \leqslant \frac{1}{2} \|r(t)\|^2_{L^2(\Omega)},$$

and hence by integration

$$a(\theta_h(t), \theta_h(t)) \leqslant a(\theta_h(0), \theta_h(0)) + \frac{1}{2}\int_0^t \|r(s)\|^2_{L^2(\Omega)} \,\mathrm{d}s.$$

Now (7.17) follows from (7.7) and (7.8). □

For the purpose of error estimates between u_h and u, we introduce the so called *elliptic* or *Ritz projection* R_h onto V_h^0:

$$a(R_h\varphi, v_h) = a(\varphi, v_h) \quad \forall v_h \in V_h^0, \qquad \text{for any } \varphi \in H_0^1(\Omega). \tag{7.18}$$

As an immediate consequence of Theorem 3.9 and 3.11 we have the following error estimate:

$$\begin{aligned}\|R_h\varphi - \varphi\|_{L^2(\Omega)} + h \|R_h\varphi - \varphi\|_{H^1(\Omega)} &\leqslant Ch^2 |\varphi|_{H^2(\Omega)}\\ &\forall \varphi \in H^2(\Omega) \cap H_0^1(\Omega),\end{aligned} \tag{7.19}$$

$$\|R_h\varphi - \varphi\|_{L^2(\Omega)} \leqslant Ch \|\varphi\|_{H^1(\Omega)} \quad \forall \varphi \in H_0^1(\Omega), \tag{7.20}$$

where h is the maximum diameter of the elements in $\mathcal{M}_h$.

Now we are ready to prove the following estimates in $L^2(\Omega)$ and $H^1(\Omega)$ for the error between the solutions of the semidiscrete problem and the continuous problem.

THEOREM 7.8. *Let u and u_h be the solutions of* (7.1) *and* (7.13), *respectively. Suppose that*

$$\|u_{h0} - u_0\|_{L^2(\Omega)} + h\,\|u_{h0} - u_0\|_{H^1(\Omega)} \leqslant Ch^2\,|u_0|_{H^2(\Omega)}\,. \tag{7.21}$$

Then for $t \geqslant 0$,

$$\|u_h(t) - u(t)\|_{L^2(\Omega)} \leqslant Ch^2\left(|u_0|_{H^2(\Omega)} + \int_0^t |u_t(s)|_{H^2(\Omega)}\,\mathrm{d}s\right), \tag{7.22}$$

$$\begin{aligned}\|u_h(t) - u(t)\|_{H^1(\Omega)} \leqslant Ch\Bigg(& |u_0|_{H^2(\Omega)} + |u(t)|_{H^2(\Omega)} \\ &+ \left(\int_0^t \|u_t(s)\|^2_{H^1(\Omega)}\,\mathrm{d}s\right)^{1/2}\Bigg).\end{aligned} \tag{7.23}$$

PROOF. We split the error into two parts:

$$u_h - u = \theta_h + \rho, \qquad \text{where } \theta_h = u_h - R_h u \in V_h^0, \quad \rho = R_h u - u. \tag{7.24}$$

The second term is easily bounded by (7.19),

$$\|\rho(t)\|_{L^2(\Omega)} + h\,\|\rho(t)\|_{H^1(\Omega)} \leqslant Ch^2\,|u(t)|_{H^2(\Omega)}\,. \tag{7.25}$$

In order to estimate θ_h, we note that by our definitions

$$\begin{aligned}(\theta_{h,t}, v_h) + a(\theta_h, v_h) =&(u_{h,t}, v_h) + a(u_h, v_h) - (R_h u_t, v_h) - a(R_h u, v_h) \\ =&(f, v_h) - a(u, v_h) - (R_h u_t, v_h) \\ =&(u_t - R_h u_t, v_h),\end{aligned} \tag{7.26}$$

or

$$(\theta_{h,t}(t), v_h) + a(\theta_h(t), v_h) = (\rho_t(t), v_h) \quad \forall v_h \in V_h^0, \quad t > 0. \tag{7.27}$$

From Theorem 7.7 with $r(t) = -\rho_t(t)$,

$$\|\theta_h(t)\|_{L^2(\Omega)} \leqslant \|\theta_h(0)\|_{L^2(\Omega)} + \int_0^t \|\rho_t(s)\|_{L^2(\Omega)}\,\mathrm{d}s.$$

Now (7.22) follows from (7.21), (7.25),

$$\|\rho_t(s)\|_{L^2(\Omega)} = \|u_t(s) - R_h u_t(s)\|_{L^2(\Omega)} \leqslant Ch^2\,|u_t(s)|_{H^2(\Omega)}\,, \tag{7.28}$$

$$|u(t)|_{H^2(\Omega)} = \left|u_0 + \int_0^t u_t(s)\,\mathrm{d}s\right|_{H^2(\Omega)} \leqslant |u_0|_{H^2(\Omega)} + \int_0^t |u_t(s)|_{H^2(\Omega)}\,\mathrm{d}s,$$

and

$$\theta_h(0) = u_{h0} - R_h u_0 = u_{h0} - u_0 + u_0 - R_h u_0.$$

Similarly, from Theorem 7.7 with $r(t) = -\rho_t(t)$,

$$\|\theta_h(t)\|_{H^1(\Omega)} \leqslant C\,\|\theta_h(0)\|_{H^1(\Omega)} + C\Big(\int_0^t \|\rho_t(s)\|^2_{L^2(\Omega)}\,\mathrm{d}s\Big)^{1/2}. \qquad (7.29)$$

Then the estimate (7.23) follows from (7.21), (7.25), and

$$\|\rho_t(s)\|_{L^2(\Omega)} = \|u_t(s) - R_h u_t(s)\|_{L^2(\Omega)} \leqslant Ch\,\|u_t(s)\|_{H^1(\Omega)}\,. \qquad (7.30)$$

This completes the proof of the theorem. □

If we choose $u_{h0} = I_h u_0$ or $u_{h0} = R_h u_0$ or $u_{h0} = Q_h u_0$ then (7.21) holds. Here I_h is the standard finite element interpolation operator onto V_h^0, and Q_h is the L^2 projection onto V_h^0. We also remark that if we take $u_{h0} = R_h u_0$ then $\theta_h(0) = 0$ and hence (7.29) and (7.28) imply that $\|\theta_h(t)\|_{H^1(\Omega)} = \|u_h(t) - R_h u(t)\|_{H^1(\Omega)} = O(h^2)$, that is, the finite element solution $u_h(t)$ is superconvergent to the elliptic projection $R_h u(t)$ of the exact solution in $H^1(\Omega)$ norm for any fixed $t \geqslant 0$.

7.3. The fully discrete approximation

In this section we turn to the analysis of some simple time discretization schemes. We begin by the *backward Euler-Galerkin method*. Let τ be the time step and U^n the approximation in V_h^0 of $u(t)$ at $t = t_n = n\tau$, this method is defined by replacing the time derivative in (7.13) by a backward difference quotient, or if $\bar\partial U^n = (U^n - U^{n-1})/\tau$,

$$(\bar\partial U^n, v_h) + a(U^n, v_h) = (f(t_n), v_h) \quad \forall v_h \in V_h^0, \quad n \geqslant 1, \quad U^0 = u_{h0}, \ (7.31)$$

For given U^{n-1} this defines U^n implicitly from the equation

$$(U^n, v_h) + \tau a(U^n, v_h) = (U^{n-1} + \tau f(t_n), v_h) \quad \forall v_h \in V_h^0.$$

With the notation as in the semidiscrete situation, this may be written

$$(M + \tau A)z^n = Mz^{n-1} + \tau b(t_n), \quad n \geqslant 1, \text{ with } z(0) = z_0, \qquad (7.32)$$

where $M + \tau A$ is positive definite.

Following the argument in the semidiscrete case, we first prove a stability estimate for the backward Euler fully discrete problem (7.31).

Theorem 7.9. *Let $r^n \in L^2(\Omega)$ and $\theta^n \in V_h^0$ satisfy*

$$(\bar\partial \theta^n, v_h) + a(\theta^n, v_h) = (r^n, v_h) \quad \forall v_h \in V_h^0, \quad n \geqslant 1. \qquad (7.33)$$

Then

$$\|\theta^n\|_{L^2(\Omega)} \leqslant \left\|\theta^0\right\|_{L^2(\Omega)} + \sum_{j=1}^{n} \tau \left\|r^j\right\|_{L^2(\Omega)}, \tag{7.34}$$

$$\|\theta^n\|_{H^1(\Omega)} \leqslant C \left\|\theta^0\right\|_{H^1(\Omega)} + C \left(\sum_{j=1}^{n} \tau \left\|r^j\right\|^2_{L^2(\Omega)} \right)^{1/2} . \tag{7.35}$$

Proof. Choosing $v_h = \theta^n$, we have $(\bar{\partial}\theta^n, \theta^n) \leqslant (r^n, \theta^n)$, or

$$\|\theta^n\|^2_{L^2(\Omega)} \leqslant \left\|\theta^{n-1}\right\|_{L^2(\Omega)} \|\theta^n\|_{L^2(\Omega)} + \tau \|r^n\|_{L^2(\Omega)} \|\theta^n\|_{L^2(\Omega)},$$

so that

$$\|\theta^n\|_{L^2(\Omega)} \leqslant \left\|\theta^{n-1}\right\|_{L^2(\Omega)} + \tau \|r^n\|_{L^2(\Omega)},$$

and, by repeated application, we have that (7.34) holds.

To prove (7.35), we choose instead $v_h = \bar{\partial}\theta^n$ to obtain

$$(\bar{\partial}\theta^n, \bar{\partial}\theta^n) + a(\theta^n, \bar{\partial}\theta^n) = (r^n, \bar{\partial}\theta^n) \leqslant \frac{1}{4} \|r^n\|^2_{L^2(\Omega)} + \left\|\bar{\partial}\theta^n\right\|^2_{L^2(\Omega)},$$

or

$$\begin{aligned} a(\theta^n, \theta^n) \leqslant & a(\theta^n, \theta^{n-1}) + \frac{\tau}{4} \|r^n\|^2_{L^2(\Omega)}, \\ \leqslant & \frac{1}{2} a(\theta^n, \theta^n) + \frac{1}{2} a(\theta^{n-1}, \theta^{n-1}) + \frac{\tau}{4} \|r^n\|^2_{L^2(\Omega)}, \end{aligned}$$

so that

$$a(\theta^n, \theta^n) \leqslant a(\theta^{n-1}, \theta^{n-1}) + \frac{\tau}{2} \|r^n\|^2_{L^2(\Omega)},$$

which together with (7.7) and (7.8) implies that (7.35) holds. □

We need the following Taylor formula with integral remainder in the subsequent analysis:

$$\varphi(t) = \sum_{k=0}^{m} \frac{\varphi^{(k)}(a)}{k!} (t-a)^k + \frac{1}{m!} \int_a^t \varphi^{(m+1)}(s)(t-s)^m \, \mathrm{d}s. \tag{7.36}$$

Theorem 7.10. *Suppose that*

$$\|u_{h0} - u_0\|_{L^2(\Omega)} + h \|u_{h0} - u_0\|_{H^1(\Omega)} \leqslant Ch^2 |u_0|_{H^2(\Omega)} .$$

With U^n and u the solutions of (7.31) *and* (7.1)*, respectively, we have*

$$\begin{aligned}\|U^n - u(t_n)\|_{L^2(\Omega)} \leqslant & Ch^2 \left(|u_0|_{H^2(\Omega)} + \int_0^{t_n} |u_t(s)|_{H^2(\Omega)} \, \mathrm{d}s \right) \\ & + C\tau \int_0^{t_n} \|u_{tt}(s)\|_{L^2(\Omega)} \, \mathrm{d}s, \qquad (7.37)\\ \|U^n - u(t_n)\|_{H^1(\Omega)} \leqslant & Ch \Bigg(|u_0|_{H^2(\Omega)} + |u(t_n)|_{H^2(\Omega)} \\ & + \left(\int_0^{t_n} \|u_t(s)\|^2_{H^1(\Omega)} \, \mathrm{d}s \right)^{1/2} \Bigg) \\ & + C\tau \left(\int_0^{t_n} \|u_{tt}(s)\|^2_{L^2(\Omega)} \, \mathrm{d}s \right)^{1/2}, \qquad n \geqslant 0. \qquad (7.38)\end{aligned}$$

Proof. In analogy with (7.24) we write

$$U^n - u(t_n) = (U^n - R_h u(t_n)) + (R_h u(t_n) - u(t_n)) = \theta^n + \rho^n,$$

and ρ^n is bounded as claimed in (7.25). This time, a calculation corresponding to (7.26) yields

$$(\bar{\partial}\theta^n, v_h) + a(\theta^n, v_h) = -(\omega^n, v_h) \quad \forall v_h \in V_h^0, \quad n \geqslant 1, \qquad (7.39)$$

where

$$\begin{aligned}\omega^n &= R_h \bar{\partial} u(t_n) - u_t(t_n) \\ &= (R_h - I)\bar{\partial} u(t_n) + (\bar{\partial} u(t_n) - u_t(t_n)) = \omega_1^n + \omega_2^n. \qquad (7.40)\end{aligned}$$

We write

$$\begin{aligned}\tau\omega_1^j &= (R_h - I) \int_{t_{j-1}}^{t_j} u_t \, \mathrm{d}s = \int_{t_{j-1}}^{t_j} (R_h - I) u_t(s) \, \mathrm{d}s, \\ \tau\omega_2^j &= u(t_j) - u(t_{j-1}) - \tau u_t(t_j) = - \int_{t_{j-1}}^{t_j} (s - t_{j-1}) u_{tt}(s) \, \mathrm{d}s,\end{aligned}$$

and obtain

$$\begin{aligned}\tau \| \, \omega_1^j \, \|_{L^2(\Omega)} &\leqslant Ch^2 \int_{t_{j-1}}^{t_j} |u_t(s)|_{H^2(\Omega)} \, \mathrm{d}s, \\ \tau \| \, \omega_1^j \, \|_{L^2(\Omega)} &\leqslant Ch \int_{t_{j-1}}^{t_j} \|u_t(s)\|_{H^1(\Omega)} \, \mathrm{d}s, \\ \tau \| \, \omega_2^j \, \|_{L^2(\Omega)} &\leqslant C\tau \int_{t_{j-1}}^{t_j} \|u_{tt}(s)\|_{L^2(\Omega)} \, \mathrm{d}s.\end{aligned}$$

Together our estimates and using Theorem 7.9 with $r^n = -\omega^n$ complete the proof of the theorem. □

We now turn to the *Crank-Nicolson Galerkin method*. Here the semidiscrete equation is discretized in a symmetric fashion around the point $t_{n-\frac{1}{2}} = (n-\frac{1}{2})\tau$, which will produce a second order accurate method in time. More precisely, we define U^n in V_h^0 recursively for $n \geqslant 1$ by

$$(\bar{\partial}U^n, v_h)+a\left(\frac{U^n+U^{n-1}}{2}, v_h\right) = (f(t_{n-\frac{1}{2}}), v_h) \quad \forall v_h \in V_h^0, \quad n \geqslant 1, \tag{7.41}$$

with $U^0 = u_{h0}$. Here the equation for U^n may be written in the matrix form as

$$(M+\frac{1}{2}\tau A)z^n = (M-\frac{1}{2}\tau A)z^{n-1} + \tau b(t_{n-\frac{1}{2}}),$$

with a positive definite matrix $M + \frac{1}{2}\tau A$. We have the following stability result for (7.41).

Theorem 7.11. *Let $r^n \in L^2(\Omega)$ and $\theta^n \in V_h^0$ satisfy*

$$(\bar{\partial}\theta^n, v_h) + a\left(\frac{\theta^n+\theta^{n-1}}{2}, v_h\right) = (r^n, v_h) \quad \forall v_h \in V_h^0, \quad n \geqslant 1. \tag{7.42}$$

Then (7.34) *and* (7.35) *hold, that is,*

$$\|\theta^n\|_{L^2(\Omega)} \leqslant \left\|\theta^0\right\|_{L^2(\Omega)} + \sum_{j=1}^{n} \tau \left\|r^j\right\|_{L^2(\Omega)}, \tag{7.43}$$

$$\|\theta^n\|_{H^1(\Omega)} \leqslant C\left\|\theta^0\right\|_{H^1(\Omega)} + C\left(\sum_{j=1}^{n} \tau \left\|r^j\right\|_{L^2(\Omega)}^2\right)^{1/2}. \tag{7.44}$$

Proof. (7.43) and (7.44) can be proved by choosing $v_h = (\theta^n + \theta^{n-1})/2$ and $v_h = \bar{\partial}\theta^n$ in (7.42), respectively, we omit the details. □

Now the error estimate reads as follows.

Theorem 7.12. *Suppose that*

$$\|u_{h0} - u_0\|_{L^2(\Omega)} + h\|u_{h0} - u_0\|_{H^1(\Omega)} \leqslant Ch^2 |u_0|_{H^2(\Omega)}.$$

Let U^n and u be the solutions of (7.41) *and* (7.1), *respectively. Then we have for* $n \geqslant 0$,

$$\begin{aligned}
&\|U^n - u(t_n)\|_{L^2(\Omega)}\\
&\quad\leqslant Ch^2 \left(|u_0|_{H^2(\Omega)} + \int_0^{t_n} |u_t(s)|_{H^2(\Omega)}\,\mathrm{d}s \right)\\
&\qquad + C\tau^2 \int_0^{t_n} \Big(\|u_{ttt}(s)\|_{L^2(\Omega)} + \|u_{tt}(s)\|_{H^2(\Omega)} \Big)\,\mathrm{d}s, \qquad (7.45)\\
&\|U^n - u(t_n)\|_{H^1(\Omega)}\\
&\quad\leqslant Ch \left(|u_0|_{H^2(\Omega)} + |u(t_n)|_{H^2(\Omega)} + \left(\int_0^{t_n} \|u_t(s)\|^2_{H^1(\Omega)}\,\mathrm{d}s \right)^{1/2} \right)\\
&\qquad + C\tau^2 \left(\int_0^{t_n} (\|u_{ttt}(s)\|^2_{L^2(\Omega)} + \|u_{tt}(s)\|^2_{H^2(\Omega)})\,\mathrm{d}s \right)^{1/2}. \qquad (7.46)
\end{aligned}$$

Proof. Since ρ^n is bounded as above, we only need to consider θ^n. We have

$$(\bar{\partial}\theta^n, v_h) + a\left(\frac{\theta^n + \theta^{n-1}}{2}, v_h \right) = -(\omega^n, v_h) \quad \forall v_h \in V_h^0, \quad n \geqslant 1,$$

where now

$$\omega^n = \omega_1^n + \omega_2^n + \omega_3^n, \quad \omega_1^n = (R_h - I)\bar{\partial}u(t_n), \quad \omega_2^n = \bar{\partial}u(t_n) - u_t(t_{n-\frac{1}{2}}),$$

and $\omega_3^n \in V_h^0$ such that

$$(\omega_3^n, v_h) = a\left(\frac{u(t_n) + u(t_{n-1})}{2} - u(t_{n-\frac{1}{2}}), v_h \right) \quad \forall v_h \in V_h^0.$$

Since θ^0 and ω_1^j are estimated as before, to apply Theorem 7.11, it remains to bound the terms in ω_2^j and ω_3^j. This can be done by using Taylor formula (7.36). We omit the details. □

7.4. A posteriori error analysis

In this section we consider the a posteriori error estimates for the finite element method for solving linear parabolic problems which are the basis of the time and space adaptive algorithm in next section.

Let Ω be a polyhedron domain in $\mathbb{R}^d$ $(d = 1, 2, 3)$, $\Gamma = \partial\Omega$ and $T > 0$, we consider the following linear parabolic equation:

$$\begin{aligned}
u_t - \operatorname{div}(a(x)\nabla u) = f \quad &\text{in } \Omega \times (0, T),\\
u = 0 \text{ on } \Gamma \times (0, T), \quad &u(x, 0) = u_0(x) \text{ in } \Omega,
\end{aligned} \qquad (7.47)$$

where $f \in L^2(0,T;L^2(\Omega))$ and $u_0 \in L^2(\Omega)$, and the coefficient $a(x)$ is assumed to be piecewise constant and positive. The weak formulation of (7.47) reads as follows: Find $u \in L^2(0,T;H_0^1(\Omega)) \cap H^1(0,T;H^{-1}(\Omega))$ such that $u(\cdot,0) = u_0(\cdot)$, and for a.e. $t \in (0,T)$ the following relation holds

$$\langle u_t, \varphi\rangle + (a\nabla u, \nabla\varphi) = \langle f, \varphi\rangle \quad \forall \varphi \in H_0^1(\Omega). \tag{7.48}$$

We consider the backward Euler fully discrete approximation with variable time steps for (7.48). Let τ_n be the n-th time-step size and set

$$t^n := \sum_{i=1}^{n} \tau_i, \quad \varphi^n(\cdot) = \varphi(\cdot, t^n)$$

for any function φ continuous in $(t^{n-1}, t^n]$. Let N be the total number of steps, that is $t^N \geqslant T$. At each time step n, $n = 1, 2, \cdots, N$, we denote by $\mathcal{M}^n$ a uniformly regular partition of Ω into simplexes which is obtained from $\mathcal{M}^{n-1}$ by using refinement/coarsening procedures. Let $V_0^n \subset H_0^1(\Omega)$ indicate the usual space of conforming linear finite elements over $\mathcal{M}^n$. Let $U_h^0 = Q_0 u_0$, where $Q_0 : L^2(\Omega) \to V_0^0$ is the L^2 projection operator into the linear finite element space V_0^0 over the initial mesh $\mathcal{M}^0$. Then the fully discrete finite element approximation at the n-th time step reads as follows: Given $U_h^{n-1} \in V^{n-1}$, then $\mathcal{M}^{n-1}$ and τ_{n-1} are modified as described below to give rise to $\mathcal{M}^n$ and τ_n and thereafter $U_h^n \in V_0^n$ computed according to the following prescription:

$$\left\langle \bar\partial^n U_h^n, v\right\rangle + (a\nabla U_h^n, \nabla v) = \left\langle \bar f^n, v\right\rangle \quad \forall v \in V_0^n. \tag{7.49}$$

Here $\bar\partial^n U_h^n = (U_h^n - U_h^{n-1})/\tau_n$ is the backward difference quotient, and

$$\bar f^n = \frac{1}{\tau_n}\int_{t^{n-1}}^{t^n} f(x,t)\,\mathrm{d}t.$$

Denote $\mathcal{B}^n$ the collection of interior inter-element sides e of $\mathcal{M}^n$ in Ω. h_K stands for the diameter of $K \in \mathcal{M}^n$ and h_e stands for the size of $e \in \mathcal{B}^n$. We define the interior residual

$$R^n := \bar f^n - \bar\partial^n U_h^n,$$

along with the jump residual across $e \in \mathcal{B}^n$

$$J_e^n := [\![a\nabla U_h^n]\!]_e \cdot \nu_e = (a\nabla U_h^n|_{K_1} - a\nabla U_h^n|_{K_2}) \cdot \nu_e, \qquad e = \partial K_1 \cap \partial K_2,$$

using the convention that the unit normal vector ν_e to e points from K_2 to K_1. We observe that the integration by parts implies

$$(a\nabla U_h^n, \nabla\varphi) = -\sum_{e\in\mathcal{B}^n}\int_e J_e^n \varphi ds \qquad \forall\varphi\in H_0^1(\Omega). \tag{7.50}$$

Introduce the energy norm $|||\varphi|||_\Omega = (a\nabla\varphi, \nabla\varphi)^{1/2}$. We have the following upper bound estimate.

THEOREM 7.13. *For any integer $1 \leqslant m \leqslant N$, there exists a constant $C > 0$ depending only the minimum angle of meshes $\mathcal{M}^n, n = 1, 2, \cdots, m$, and the coefficient $a(x)$ such that the following a posteriori error estimate holds*

$$\begin{aligned}
&\frac{1}{2}\|u^m - U_h^m\|_{L^2(\Omega)}^2 + \sum_{n=1}^m \int_{t^{n-1}}^{t^n} |||u - U_h^n|||_\Omega^2 dt \\
&\leqslant \left\|u_0 - U_h^0\right\|_{L^2(\Omega)}^2 + \sum_{n=1}^m \tau_n \eta_{\text{time}}^n + C\sum_{n=1}^m \tau_n\eta_{\text{space}}^n \\
&\quad + 2\Big(\sum_{n=1}^m \int_{t^{n-1}}^{t^n} \left\|f - \bar{f}^n\right\|_{L^2(\Omega)} \mathrm{d}t\Big)^2,
\end{aligned} \tag{7.51}$$

where the time error indicator η_{time}^n and space error indicator η_{space}^n are given by

$$\eta_{\text{time}}^n = \frac{1}{3}|||U_h^n - U_h^{n-1}|||_\Omega^2, \qquad \eta_{\text{space}}^n = \sum_{e\in\mathcal{B}^n}\eta_e^n$$

with the local error indicator η_e^n defined as

$$\eta_e^n = \frac{1}{2}\sum_{K\in\Omega_e} h_K^2\|\ R^n\ \|_{L^2(K)}^2 + h_e\|\ J_e^n\ \|_{L^2(e)}^2.$$

Here Ω_e is the collection of two elements sharing the common side $e \in \mathcal{B}^n$.

PROOF. From (7.49) we know that, for any $\varphi \in H_0^1(\Omega)$ and $v \in V_0^n$,

$$\begin{aligned}
&\langle\bar{\partial}^n U^n, \varphi\rangle + (a\nabla U_h^n, \nabla\varphi) \\
&= -\langle R^n, \varphi - v\rangle + (a\nabla U_h^n, \nabla(\varphi - v)) + \langle\bar{f}^n, \varphi\rangle.
\end{aligned} \tag{7.52}$$

For any $t \in (t^{n-1}, t^n]$, we denote by

$$U_h(t) = l(t)U_h^n + (1 - l(t))U_h^{n-1}, \quad l(t) = (t - t^{n-1})/\tau_n.$$

Then from (7.48) and (7.52) we have, for a.e. $t \in (t^{n-1}, t^n]$, and for any $\varphi \in H_0^1(\Omega), v \in V_0^n$,

$$\begin{aligned}&\Big\langle \frac{\partial(u-U_h)}{\partial t}, \varphi \Big\rangle + \langle a\nabla(u-U_h^n), \nabla\varphi\rangle \\ &= \langle R^n, \varphi - v\rangle - (a\nabla U_h^n, \nabla(\varphi - v)) + \langle f - \bar{f}^n, \varphi\rangle.\end{aligned}$$

Now we resort to the Clément interpolation operator $r_n : H_0^1(\Omega) \to V_0^n$ defined in Subsection 4.2.1, which satisfies the following local approximation properties by Theorem 4.2, for any $\varphi \in H_0^1(\Omega)$,

$$\|\,\varphi - r_n\varphi\,\|_{L^2(K)} + h_K\|\,\nabla(\varphi - r_n\varphi)\,\|_{L^2(K)} \leqslant C^* h_K \|\,\nabla\varphi\,\|_{L^2(\tilde{K})}, \quad (7.53)$$

$$\|\,\varphi - r_n\varphi\,\|_{L^2(e)} \leqslant C^* h_e^{1/2} \|\,\nabla\varphi\,\|_{L^2(\tilde{e})}, \quad (7.54)$$

where $\tilde{A}$ is the union of all elements in $\mathcal{M}^n$ surrounding the sets $A = K \in \mathcal{M}^n$ or $A = e \in \mathcal{B}^n$. The constant C^* depends only on the minimum angle of mesh $\mathcal{M}^n$. Based on this interpolation operator, by taking $\varphi = u - U_h \in H_0^1(\Omega)$, $v = r_n(u - U_h) \in V_0^n$, and using (7.50) and the identity

$$(a\nabla(u-U_h^n), \nabla(u-U_h)) = \frac{1}{2}|\!|\!|u-U_h^n|\!|\!|_\Omega^2 + \frac{1}{2}|\!|\!|u-U_h|\!|\!|_\Omega^2 - \frac{1}{2}|\!|\!|U_h-U_h^n|\!|\!|_\Omega^2,$$

we deduce that

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\|\,u-U_h\,\|_{L^2(\Omega)}^2 + \frac{1}{2}|\!|\!|u-U_h^n|\!|\!|_\Omega^2 + \frac{1}{2}|\!|\!|u-U_h|\!|\!|_\Omega^2 \\ &= \frac{1}{2}|\!|\!|U_h-U_h^n|\!|\!|_\Omega^2 + \langle R^n, (u-U_h) - r_n(u-U_h)\rangle \\ &\quad + \sum_{e\in\mathcal{B}^n}\int_e J_e^n[(u-U_h) - r_n(u-U_h)]ds + \langle f - \bar{f}^n, u - U_h\rangle. \qquad (7.55)\end{aligned}$$

For any $t^* \in (t^{m-1}, t^m]$, by integrating (7.55) in time from 0 to t^*, using (7.53)-(7.54) and exploiting the standard argument in finite element a posteriori analysis, we have

$$\begin{aligned}&\frac{1}{2}\|\,(u-U_h)(t^*)\,\|_{L^2(\Omega)}^2 + \frac{1}{2}\sum_{n=1}^m \int_{t^{n-1}}^{t^n\wedge t^*} \Big(|\!|\!|u-U_h^n|\!|\!|_\Omega^2 + |\!|\!|u-U_h|\!|\!|_\Omega^2\Big)dt \\ &\leqslant \frac{1}{2}\|\,u_0 - U_h^0\,\|_{L^2(\Omega)}^2 + \frac{1}{2}\sum_{n=1}^m\int_{t^{n-1}}^{t^n} |\!|\!|U_h-U_h^n|\!|\!|_\Omega^2 dt \\ &\quad + C\sum_{n=1}^m\int_{t^{n-1}}^{t^n}(\eta_{\text{space}}^n)^{1/2}|\!|\!|u-U_h|\!|\!|_\Omega dt + \frac{1}{4}\max_{0\leqslant t\leqslant t^*}\|\,u-U_h\,\|_{L^2(\Omega)}^2 \\ &\quad + \Big(\sum_{n=1}^m\int_{t^{n-1}}^{t^n}\|\,f-\bar{f}^n\,\|_{L^2(\Omega)}dt\Big)^2,\end{aligned}$$

where $t^n \wedge t^* = \min(t^n, t^*)$. This implies the desired estimate (7.51) by using the fact that

$$\begin{aligned}\int_{t^{n-1}}^{t^n} \| U_h - U_h^n \|_\Omega^2 dt &= \int_{t^{n-1}}^{t^n} (1 - l(t))^2 \| U_h^n - U_h^{n-1} \|_\Omega^2 dt \\ &= \frac{1}{3}\tau_n \| U_h^n - U_h^{n-1} \|_\Omega^2.\end{aligned}$$

This completes the proof. □

In our a posteriori error estimate at the n-th time step, the time discretization error is controlled by $\| U_h^n - U_h^{n-1} \|_\Omega$ and $\displaystyle\int_{t^{n-1}}^{t^n} \| f - \bar{f}^n \|_{L^2(\Omega)} \,\mathrm{d}t$, which can only be reduced by reducing the time-step sizes τ_n. On the other hand, the time-step size τ_n essentially controls the semi-discretization error: the error between the exact solution u and the solution U^n of the following problem

$$\left\langle \frac{U^n - U^{n-1}}{\tau_n}, \varphi \right\rangle + (a\nabla U^n, \nabla\varphi) = \langle \bar{f}^n, \varphi \rangle \quad \forall \varphi \in H_0^1(\Omega). \tag{7.56}$$

Thus $\| U_h^n - U_h^{n-1} \|_\Omega$ is not a good error indicator for time discretization unless the space discretization error is sufficiently resolved. In the adaptive method for time-dependent problems, we must do space mesh and time-step size adaptation simultaneously. Ignoring either one of them may not provide correct error control of approximation to the problem.

Our objective next is to prove the following estimate for the local error which ensures over-refinement will not occur for the refinement strategy based on our space error indicator. First we note that for given $U_h^{n-1} \in V_0^{n-1}$, let $U_*^n \in H_0^1(\Omega)$ be the solution of the following continuous problem

$$\left\langle \frac{U_*^n - U_h^{n-1}}{\tau_n}, \varphi \right\rangle + (a\nabla U_*^n, \nabla\varphi) = \langle \bar{f}^n, \varphi \rangle \quad \forall \varphi \in H_0^1(\Omega), \tag{7.57}$$

Then the space error indicator $\eta_{\rm space}^n$ controls only the error between U_h^n and U_*^n, not between U_h^n and U^n (or the exact solution u).

For any $K \in \mathcal{M}^n$ and $\varphi \in L^2(\Omega)$, we define $P_K\varphi = \frac{1}{|K|}\int_K \varphi dx$, the average of φ over K. For any $n = 1, 2, \cdots$, we also need the notation

$$\hat{C}_n = \max_{K\in\mathcal{M}^n} (h_K^2/\tau_n) + 1, \qquad h_K = \mathrm{diam}\,(K). \tag{7.58}$$

THEOREM 7.14. *Let $U_*^n \in H_0^1(\Omega)$ be the solution of the auxiliary problem (7.57). Then there exist constants $C_2, C_3 > 0$ depending only on the minimum* *angle of $\mathcal{M}^n$ and the coefficient $a(x)$ such that for any $e \in \mathcal{B}^n$, the following estimate holds*

$$\begin{aligned}\eta_e^n \leqslant& C_2\hat{C}_n \sum_{K\in\Omega_e} \Big(\frac{1}{\tau_n}\| U_*^n - U_h^n \|_{L^2(K)}^2 + |\!|\!| U_*^n - U_h^n |\!|\!|_K^2\Big)\\ &+ C_3 \sum_{K\in\Omega_e} h_K^2\| R^n - P_K R^n \|_{L^2(K)}^2. \qquad (7.59)\end{aligned}$$

Proof. The proof extends the idea to prove the local lower bound for elliptic equations in Theorem 4.4. For any $K \in \mathcal{M}^n$, let $\psi_K = (d+1)^{d+1}\lambda_1 \cdots \lambda_{d+1}$ be the bubble function, where $\lambda_1, \cdots, \lambda_{d+1}$ are the barycentric coordinate functions. By the standard scaling argument, we have the following inf-sup relation that holds for some constant β depending only on the minimum angle of $K \in \mathcal{M}^n$

$$\inf_{v_h\in P_1(K)} \sup_{\varphi_h\in P_1(K)} \frac{\displaystyle\int_K v_h\varphi_h\psi_K dx}{\| \varphi_h \|_{L^2(K)}\| v_h \|_{L^2(K)}} \geqslant \beta > 0,$$

Thus there exists a function $\varphi^n \in P_1(K)$ with $\| \varphi^n \|_{L^2(K)} = 1$ such that

$$\begin{aligned}&\beta\| P_K R^n \|_{L^2(K)}\\ &\leqslant \int_K (P_K R^n)\psi_K\varphi^n dx\\ &= \int_K (P_K R^n - R^n)\psi_K\varphi^n dx + \int_K \Big(\bar{f}^n - \frac{U_h^n - U_h^{n-1}}{\tau_n}\Big)\psi_K\varphi^n dx\\ &= \int_K (P_K R^n - R^n)\psi_K\varphi^n dx + \int_K \frac{U_*^n - U_h^n}{\tau_n}\ \psi_K\varphi^n dx + (a\nabla U_*^n, \nabla(\psi_K\varphi^n))_K,\end{aligned}$$

where we have used (7.57) in the last identity. Since $U_h^n \in P_1(K)$ and $\psi_K = 0$ on ∂K, simple integration by parts implies that $(a\nabla U_h^n, \nabla(\psi_K\varphi^n))_K = 0$. Thus, we have

$$\begin{aligned}\| P_K R^n \|_{L^2(K)} \leqslant& C\| R^n - P_K R^n \|_{L^2(K)} + C\tau_n^{-1}\| U_*^n - U_h^n \|_{L^2(K)}\\ &+ C|\!|\!| U_*^n - U_h^n |\!|\!|_K |\!|\!| \psi_K\varphi^n |\!|\!|_K.\end{aligned}$$

Since $|\!|\!| \psi_K\varphi^n |\!|\!|_K \leqslant Ch_K^{-1}$ by inverse estimate, we conclude by the definition of $\hat{C}_n$ in (7.58) that

$$\begin{aligned}\| P_K R^n \|_{L^2(K)} \leqslant& C\| R^n - P_K R^n \|_{L^2(K)}\\ &+ C\hat{C}_n^{1/2} h_K^{-1}\Big(\frac{1}{\tau_n}\| U_*^n - U_h^n \|_{L^2(K)}^2 + |\!|\!| U_*^n - U_h^n |\!|\!|_K^2\Big)^{1/2}.\end{aligned}$$

Therefore, we have

$$\begin{aligned} h_K^2\| R^n \|_{L^2(K)}^2 \leqslant & Ch_K^2\| R^n - P_K R^n \|_{L^2(K)}^2 \\ & + C\hat{C}_n\Big(\frac{1}{\tau_n}\| U_*^n - U_h^n \|_{L^2(K)}^2 + \|\!| U_*^n - U_h^n |\!\|_K^2\Big). \end{aligned}$$

For any $e \in \mathcal{B}^n$, let $\psi_e = d^d\lambda_1 \cdots \lambda_d$ be the bubble function, where $\lambda_1, \cdots, \lambda_d$ are the barycentric coordinate functions associated with the nodes of e. Denote by $\psi^n = J_e^n\psi_e \in H_0^1(\Omega)$. Then since J_e^n is constant on $e \in \mathcal{B}^n$, we get, after integration by parts, that

$$\begin{aligned} \| J_e^n \|_{L^2(e)}^2 \leqslant & C\int_e J_e^n\psi^n dx = -C\sum_{K\in\Omega_e}\int_K a(x)\nabla U_h^n\nabla\psi^n dx \\ = & C\sum_{K\in\Omega_e}\int_K a(x)\nabla(U_*^n - U_h^n)\nabla\psi^n dx \\ & - C\sum_{K\in\Omega_e}\int_K\Big(R^n - \frac{U_*^n - U_h^n}{\tau^n}\Big)\psi^n dx, \end{aligned}$$

where we have used the definition of U_*^n in (7.57). Moreover, it is easy to see that

$$\| \nabla\psi^n \|_{L^2(K)} \leqslant Ch_e^{-1/2}\| J_e^n \|_{L^2(e)}, \| \psi^n \|_{L^2(K)} \leqslant Ch_e^{1/2}\| J_e^n \|_{L^2(e)}, \ \forall K \in \Omega_e.$$

Thus

$$\begin{aligned} h_e\| J_e^n \|_{L^2(e)}^2 \leqslant & C\sum_{K\in\Omega_e} h_K^2\| R^n \|_{L^2(K)}^2 \\ & + C\hat{C}^n\sum_{K\in\Omega_e}\Big(\|\!| U_*^n - U_h^n |\!\|_K^2 + \frac{1}{\tau^n}\|U_*^n - U_h^n\|_{L^2(K)}\Big). \end{aligned}$$

This completes the proof. □

7.5. The adaptive algorithm

We start by considering the algorithm for time-step size control. The adjustment of the time-step size is based on the error equi-distribution strategy: the time discretization error should be equally distributed to each time interval $(t^{n-1}, t^n), n = 1, \cdots, N$. Let $\texttt{TOL}_{\texttt{time}}$ be the total tolerance allowed for the part of a posteriori error estimate in (7.51) related to the time discretization, that is,

$$\sum_{n=1}^N \tau_n\eta_{\rm time}^n + 2\Big(\sum_{n=1}^N\int_{t^{n-1}}^{t^n}\| f - \bar{f}^n \|_{L^2(\Omega)}dt\Big)^2 \leqslant \texttt{TOL}_{\texttt{time}}. \qquad (7.60)$$

A natural way to achieve (7.60) is to adjust the time-step size τ_n such that the following relations are satisfied

$$\eta_{\rm time}^n \leqslant \frac{\mathtt{TOL_{time}}}{2T}, \qquad \frac{1}{\tau_n}\int_{t^{n-1}}^{t^n} \| f-\bar{f}^n \|_{L^2(\Omega)} dt \leqslant \frac{1}{2T}\sqrt{\mathtt{TOL_{time}}}. \tag{7.61}$$

Let $\mathtt{TOL_{space}}$ be the tolerance allowed for the part of a posteriori error estimate in (7.51) related to the spatially semidiscrete approximation. Then the usual stopping criterion for the mesh adaptation is to satisfy the following relation at each time step n

$$\eta_{\rm space}^n \leqslant \frac{\mathtt{TOL_{space}}}{T}. \tag{7.62}$$

This stopping rule is appropriate for mesh refinements but not for mesh coarsening. We will use the coarsening error indicator based on the following theorem.

Define the weighted norm of $H^1(\Omega)$ with parameter $\tau_n > 0$

$$\| \varphi \|_{\tau_n,\Omega} = \Big(\frac{1}{\tau_n}\| \varphi \|_{L^2(\Omega)}^2 + |\!|\!|\varphi|\!|\!|_\Omega^2\Big)^{1/2} \quad \forall \varphi \in H^1(\Omega) \tag{7.63}$$

THEOREM 7.15. *Given $U_h^{n-1} \in V^{n-1}$ and $\tau_n > 0$. Let $\mathcal{M}_H^n$ be a coarsening of the mesh $\mathcal{M}^n$. Let $U_H^n \in V_0^{n,H}$, $U_h^n \in V_0^n$ be the solutions of the discrete problem (7.49) over meshes $\mathcal{M}_H^n$ and $\mathcal{M}^n$, respectively. Then the following error estimate is valid*

$$\| U_*^n - U_H^n \|_{\tau_n,\Omega}^2 \leqslant \| U_*^n - U_h^n \|_{\tau_n,\Omega}^2 + \| U_h^n - I_H^n U_h^n \|_{\tau_n,\Omega}^2,$$

where $I_H^n : C(\bar{\Omega}) \to V_0^{n,H}$ is the standard linear finite element interpolation operator.

PROOF. By definition, $U_H^n \in V_0^{n,H}$ and $U_h^n \in V_0^n$ satisfy the following relations

$$\Big(\frac{U_H^n - U_h^{n-1}}{\tau_n}, v\Big) + (a\nabla U_H^n, \nabla v) = \langle \bar{f}^n, v\rangle \quad \forall v \in V_0^{n,H}, \tag{7.64}$$

$$\Big(\frac{U_h^n - U_h^{n-1}}{\tau_n}, v\Big) + (a\nabla U_h^n, \nabla v) = \langle \bar{f}^n, v\rangle \quad \forall v \in V_0^{n}, \tag{7.65}$$

Since $\mathcal{M}_H^n$ is a coarsening of $\mathcal{M}^n$, we have $V_0^{n,H} \subset V_0^n$. Thus $U_H^n - U_h^n \in V_0^n$. Now the equation (7.65) together with (7.57) implies the following Galerkin orthogonal identity:

$$\Big(\frac{U_*^n - U_h^n}{\tau_n}, U_H^n - U_h^n\Big) + (a\nabla(U_*^n - U_h^n), \nabla(U_H^n - U_h^n)) = 0.$$

Hence

$$\| U_*^n - U_H^n \|_{\tau_n,\Omega}^2 = \| U_*^n - U_h^n \|_{\tau_n,\Omega}^2 + \| U_H^n - U_h^n \|_{\tau_n,\Omega}^2. \tag{7.66}$$

Next, by subtracting (7.64) from (7.65) and taking $v = U_H^n - I_H^n U_h^n \in V_0^{n,H}$, we obtain the following Galerkin orthogonal relation

$$\Big\langle \frac{U_H^n - U_h^n}{\tau_n}, U_H^n - I_H^n U_h^n \Big\rangle + (a\nabla(U_H^n - U_h^n), \nabla(U_H^n - I_H^n U_h^n)) = 0,$$

which implies

$$\begin{aligned} \| U_H^n - U_h^n \|_{\tau_n,\Omega}^2 &= \| U_h^n - I_H^n U_h^n \|_{\tau_n,\Omega}^2 - \| U_H^n - I_H^n U_h^n \|_{\tau_n,\Omega}^2 \\ &\leqslant \| U_h^n - I_H^n U_h^n \|_{\tau_n,\Omega}^2. \end{aligned}$$

This completes the proof by using (7.66). □

Theorem 7.15 suggests us to introduce the following coarsening error indicator

$$\eta_{\text{coarse}}^n = \frac{1}{\tau_n} \| I_H^n U_h^n - U_h^n \|_{L^2(\Omega)}^2 + |\!|\!| I_H^n U_h^n - U_h^n |\!|\!|_{\Omega}^2. \tag{7.67}$$

The nice feature of this indicator lies in that it does not depend on U_H^n, the solution of the coarsened problem. This property allows us to do coarsening only once, without checking whether the coarsened solution U_H^n satisfies some stopping criterion such as (7.62). Combining above ideas together, we arrive at the following adaptive algorithm for one single time step.

Algorithm 7.1. (Time and space adaptive algorithm) Given tolerances $\texttt{TOL}_{\texttt{time}}, \texttt{TOL}_{\texttt{space}}$ and $\texttt{TOL}_{\texttt{coarse}}$, parameters $\delta_1 \in (0,1), \delta_2 > 1$ and $\theta_{\text{time}} \in (0,1)$. Given U_h^{n-1} from the previous time-step at time t^{n-1} with the mesh $\mathcal{M}^{n-1}$ and the time-step size τ_{n-1}.

```
1. $\mathcal{M}^n := \mathcal{M}^{n-1}$, $\tau_n := \tau_{n-1}$, $t^n := t^{n-1} + \tau_n$
   solve the discrete problem for $U_h^n$ on $\mathcal{M}^n$ using data $U_h^{n-1}$
   compute error estimates on $\mathcal{M}^n$
2. while (7.61) is not satisfied do
       $\tau_n := \delta_1 \tau_n$, $t^n := t^{n-1} + \tau_n$
       solve the discrete problem for $U_h^n$ on $\mathcal{M}^n$ using data $U_h^{n-1}$
       compute error estimates on $\mathcal{M}^n$
   end while
3. while $\eta_{\text{space}}^n > \texttt{TOL}_{\texttt{space}}/T$ do
       refine mesh $\mathcal{M}^n$ producing a modified $\mathcal{M}^n$
       solve the discrete problem for $U_h^n$ on $\mathcal{M}^n$ using data $U_h^{n-1}$
```

```
        compute error estimates on $\mathcal{M}^n$
        while (7.61) is not satisfied do
            $\tau_n := \delta_1\tau_n$, $t^n := t^{n-1} + \tau_n$
            solve the discrete problem for $U_h^n$ on $\mathcal{M}^n$ using data $U_h^{n-1}$
            compute error estimates on $\mathcal{M}^n$
        end while
    end while
4.  coarsen $\mathcal{M}^n$ producing a modified mesh $\mathcal{M}^n$ according to $\eta^n_{\text{coarse}} \leqslant \frac{\texttt{TOL}_{\texttt{coarse}}}{T}$
    solve the discrete problem for $U_h^n$ on $\mathcal{M}^n$ using data $U_h^{n-1}$
5.  if
        $\eta^n_{\text{time}} \leqslant \theta_{\text{time}} \frac{\texttt{TOL}_{\texttt{time}}}{2T}, \ \frac{1}{\tau_n}\int_{t^{n-1}}^{t^n} \| f - \bar{f}^n \|_{L^2(\Omega)} dt \leqslant \frac{1}{2T}\sqrt{\theta_{\text{time}}\texttt{TOL}_{\texttt{time}}},$
    then
        $\tau_n := \delta_2\tau_n$
    end if
```

A good choice of the parameters in above algorithm for the backward Euler scheme in time is to take $\delta_1 = 0.5, \delta_2 = 2$, and $\theta_{\text{time}} = 0.5$. The goal of the first three steps in above algorithm is to reduce the time-step size and refine the mesh so that the time and space error indicators become smaller than the respective tolerances. We achieve this goal by first reducing the time-step size to have the time error estimate below the tolerance while keeping the mesh unchanged. In Step 5, when the time error indicator is much smaller than the tolerance, the step size is enlarged (coarsened) by a factor $\delta_2 > 1$. In this case, the actual time step is not re-calculated, only the initial time-step size for the next time step is changed.

We have the following theorem which guarantees the reliability of the above algorithm in terms of error control.

Theorem 7.16. *For $n \geqslant 1$, assume that Algorithm 7.1 terminates and generates the final mesh $\mathcal{M}_H^n$, time-step size τ_n, and the the corresponding discrete solution U_H^n. Here the mesh $\mathcal{M}_H^n$ is coarsened from the mesh $\mathcal{M}^n$ produced by the first three steps. Then for any integer $1 \leqslant m \leqslant N$, there exists a constant C depending only on the minimum angles of $\mathcal{M}^n, n = 1, 2, \cdots, m$,* *and the coefficient $a(x)$ such that the following estimate holds*

$$\begin{aligned}&\frac{1}{2}\| u^m - U_H^m \|^2_{L^2(\Omega)} + \sum_{n=1}^m \int_{t^{n-1}}^{t^n} \|\!| u - U_H^n |\!\|^2_\Omega \, dt \\ &\leqslant \| u_0 - U_h^0 \|^2_{L^2(\Omega)} + \frac{t^m}{T}\mathtt{TOL}_{\mathtt{time}} \\ &\quad + C\,\frac{t^m}{T}\mathtt{TOL}_{\mathtt{space}} + C\,\hat{C}_H^m\,\frac{t^m}{T}\mathtt{TOL}_{\mathtt{coarse}},\end{aligned}\tag{7.68}$$

where $\hat{C}_H^m = \max\{h_K^2/\tau_n : K \in \mathcal{M}_H^n, n = 1, 2, \cdots, m\}$.

PROOF. Let U_h^n be the solution of the discrete problem (7.49) over the mesh $\mathcal{M}^n$ and with the time-step size τ_n. Then upon the termination of Algorithm 7.1 we have that

$$\begin{aligned}&\eta^n_{\text{time}} \leqslant \frac{\mathtt{TOL}_{\mathtt{time}}}{2T}, \quad \frac{1}{\tau_n}\int_{t^{n-1}}^{t^n} \| f - \bar{f}^n \|_{L^2(\Omega)}\, dt \leqslant \frac{1}{2T}\sqrt{\mathtt{TOL}_{\mathtt{time}}}, \\ &\eta^n_{\text{space}} \leqslant \frac{\mathtt{TOL}_{\mathtt{space}}}{T}.\end{aligned}$$

From (7.49) we know that, for any $\varphi \in H_0^1(\Omega)$,

$$\begin{aligned}&\Big\langle \frac{U_H^n - U_h^{n-1}}{\tau_n}, \varphi \Big\rangle + (a\nabla U_H^n, \nabla\varphi) \\ &= \Big\langle \frac{U_H^n - U_h^n}{\tau_n}, \varphi \Big\rangle + (a\nabla(U_H^n - U_h^n), \nabla\varphi) \\ &\quad - \langle R^n, \varphi\rangle + (a\nabla U_h^n, \nabla\varphi) + \langle \bar{f}^n, \varphi\rangle.\end{aligned}\tag{7.69}$$

Since $\mathcal{M}_H^n$ is a coarsening of $\mathcal{M}^n$, by the Galerkin orthogonal relation as in Theorem 3.1, we have

$$\Big\langle \frac{U_H^n - U_h^n}{\tau_n}, v_H \Big\rangle + (a\nabla(U_H^n - U_h^n), \nabla v_H) = 0 \quad \forall v_H \in V_0^{n,H}.$$

On the other hand, since U_h^n is the discrete solution over mesh $\mathcal{M}^n$, we have

$$-\langle R^n, v\rangle + (a\nabla U_h^n, \nabla v) = 0 \quad \forall v \in V_0^n.$$

Thus from (7.48) and (7.69) we deduce that, for a.e. $t \in (t^{n-1}, t^n]$ and for any $\varphi \in H_0^1(\Omega)$, $v_H \in V_0^{n,H}$, $v \in V_0^n$,

$$\begin{aligned}&\Big\langle \frac{\partial(u - U_H)}{\partial t}, \varphi \Big\rangle + (a\nabla(u - U_H^n), \nabla\varphi) \\ &= \langle R^n, \varphi - v\rangle - (a\nabla U_h^n, \nabla(\varphi - v)) + \langle f - \bar{f}^n, \varphi\rangle \\ &\quad - \Big\langle \frac{U_H^n - U_h^n}{\tau_n}, \varphi - v_H \Big\rangle - (a\nabla(U_H^n - U_h^n), \nabla(\varphi - v_H)),\end{aligned}$$

where for any $t \in (t^{n-1}, t^n]$, $U_H(t) = l(t)U_H^n + (1 - l(t))U_h^{n-1}$ with $l(t) = (t - t^{n-1})/\tau_n$. By taking $v_H = \Pi_H^n \varphi \in V_0^{n,H}$, the Clément interpolant of $\varphi \in H_0^1(\Omega)$ in $V_0^{n,H}$, we get, after using the estimate (7.53) for the Clément interpolation operator, that

$$\begin{aligned}
&\Big|\Big\langle \frac{U_H^n - U_h^n}{\tau_n}, \varphi - \Pi_H^n\varphi \Big\rangle + (a\nabla(U_H^n - U_h^n), \nabla(\varphi - \Pi_H^n\varphi))\Big| \\
&\leqslant C\Big(\sum_{K\in\mathcal{M}_H^n} h_K^2\tau_n^{-2}\| U_H^n - U_h^n \|_{L^2(K)}^2 + \|\!| U_H^n - U_h^n \|\!|_\Omega^2\Big)^{1/2} \|\!|\varphi\|\!|_\Omega \\
&\leqslant C(\hat{C}_H^m)^{1/2}\| U_H^n - U_h^n \|_{\tau_n,\Omega}\|\!|\varphi\|\!|_\Omega.
\end{aligned}$$

Again, since $\mathcal{M}_H^n$ is a coarsening of $\mathcal{M}^n$, from the proof of Theorem 3.1 and the Step 4 in Algorithm 7.1 we know that

$$\| U_H^n - U_h^n \|_{\tau_n,\Omega} \leqslant \| I_H^n U_h^n - U_h^n \|_{\tau_n,\Omega} \leqslant (\eta_{\text{coarse}}^n)^{1/2} \leqslant \sqrt{\frac{\mathtt{TOL_{coarse}}}{T}},$$

which yields

$$\begin{aligned}
&\Big|\Big\langle \frac{U_H^n - U_h^n}{\tau_n}, \varphi - \Pi_H^n\varphi \Big\rangle + (a\nabla(U_H^n - U_h^n), \nabla(\varphi - \Pi_H^n\varphi))\Big| \\
&\leqslant C\sqrt{\frac{\hat{C}_H^m \mathtt{TOL_{coarse}}}{T}}\|\!|\varphi\|\!|_\Omega.
\end{aligned}$$

The rest of the proof is similar to that of in Theorem 2.1. Here we omit the details. □

In practical computations, it is natural to choose the coarsening tolerance $\mathtt{TOL_{coarse}}$ much smaller than the space tolerance $\mathtt{TOL_{space}}$. However, the additional factor $\hat{C}_H^m$ in the estimate (7.68) suggests that the ratio between the coarsening tolerance and the time tolerance should also be small.

Bibliographic notes. The Sobolev space involving time in Section 7.1 follows Evans [**32**]. A comprehensive account on the mathematical theory of parabolic equations can be found in Ladyzhenskaya et al [**40**]. Sections 7.2 and 7.3 follow the development in Thomée [**49**] where further results on finite element methods for parabolic problems can be found. There are several approaches for deriving a posteriori error estimates for parabolic problems, e.g., the duality argument by Eriksson and Johnson [**30, 31**] and the energy argument by Picasso [**47**]. The analysis in Sections 7.4 is from Chen and Jia [**20**] and improves the results of [**47**]. The space and time adaptive algorithms based on a posteriori error estimates and their implementation are considered in Schmidt and Siebert [**48**]. Section 7.5 is taken from [**20**] to which we refer for the discussion on the termination of the adaptive algorithm 7.1 in finite number of steps.

7.6. Exercises

EXERCISE 7.1. Prove Theorem 7.3.

EXERCISE 7.2. Under the assumptions of Theorem 7.7, prove that, for $t > 0$,

$$\|\theta_h(t)\|_{L^2(\Omega)} \leqslant e^{-\alpha t} \|\theta_h(0)\|_{L^2(\Omega)} + \int_0^t e^{-\alpha(t-s)} \|r(s)\|_{L^2(\Omega)} \, \mathrm{d}s,$$

where α is the constant in (7.8).

EXERCISE 7.3. Let u and u_h be the solutions of (7.1) and (7.13), respectively. Suppose that $\|u_{h0} - u_0\|_{L^2(\Omega)} \leqslant Ch^2 |u_0|_{H^2(\Omega)}$. Then for $t \geqslant 0$,

$$\int_0^t \|u_h(s) - R_h u(s)\|_{H^1(\Omega)}^2 \, \mathrm{d}s \leqslant Ch^4 \Big(|u_0|_{H^2(\Omega)}^2 + \int_0^t |u_t(s)|_{H^2(\Omega)}^2 \, \mathrm{d}s \Big).$$

EXERCISE 7.4. Complete the proofs of Theorem 7.11 and Theorem 7.12.

CHAPTER 8

Finite Element Methods for Maxwell Equations

The Maxwell equations comprise four first-order partial differential equations linking the fundamental electromagnetic quantities, the electric field $\mathbf{E}$, the magnetic induction $\mathbf{B}$, the magnetic field $\mathbf{H}$, the electric flux density $\mathbf{D}$, the electric current density $\mathbf{J}$, and the space charge density ρ:

$$\begin{aligned}\nabla\times\mathbf{H}&=\mathbf{J}+\partial_t\mathbf{D},\qquad \mathrm{div}\mathbf{D}=\rho,\\ \nabla\times\mathbf{E}&=-\partial_t\mathbf{B},\qquad \mathrm{div}\mathbf{B}=0.\end{aligned}$$

They are usually supplemented by the following linear constitutive laws

$$\mathbf{D}=\varepsilon\mathbf{E},\quad \mathbf{B}=\mu\mathbf{H},$$

where ε is the dielectric permittivity and μ the magnetic permeability. In the wave form, we have

$$\begin{aligned}&\varepsilon\partial_{tt}^2\mathbf{E}+\nabla\times\left(\mu^{-1}\nabla\times\mathbf{E}\right)=-\partial_t\mathbf{J},\qquad \mathrm{div}(\varepsilon\mathbf{E})=\rho,\\ &\mu\partial_{tt}^2\mathbf{H}+\nabla\times\left(\varepsilon^{-1}\nabla\times\mathbf{H}\right)=\nabla\times\left(\varepsilon^{-1}\mathbf{J}\right),\qquad \mathrm{div}(\mu\mathbf{H})=0.\end{aligned}$$

Usually, time harmonic solutions are considered, that is,

$$\mathbf{E}(x,t)=\Re(\hat{\mathbf{E}}(x)e^{-\mathbf{i}\omega t}),\mathbf{H}(x,t)=\Re(\hat{\mathbf{H}}(x)e^{-\mathbf{i}\omega t}),\mathbf{J}(x,t)=\Re(\hat{\mathbf{J}}(x)e^{-\mathbf{i}\omega t}),$$

where $\omega>0$ is the angular frequency, then

$$\begin{aligned}&\nabla\times(\mu^{-1}\nabla\times\hat{\mathbf{E}})-\varepsilon\omega^2\hat{\mathbf{E}}=\mathbf{i}\omega\hat{\mathbf{J}},\qquad \mathrm{div}(\varepsilon\hat{\mathbf{E}})=\rho,\\ &\nabla\times(\varepsilon^{-1}\nabla\times\hat{\mathbf{H}})-\mu\omega^2\hat{\mathbf{H}}=\nabla\times(\varepsilon^{-1}\hat{\mathbf{J}}),\qquad \mathrm{div}(\mu\hat{\mathbf{H}})=0.\end{aligned}$$

In this chapter we consider adaptive edge element methods for solving the time-harmonic Maxwell equations. We will first introduce the function space $H(\mathrm{curl};\Omega)$ and its conforming finite element discretization, the lowest order Nédélec edge element method. Then we will derive the a priori and a posteriori error estimate for the edge element method.

8.1. The function space $H(\mathrm{curl};\Omega)$

Let Ω be a bounded domain in $\mathbb{R}^3$ with a Lipschitz boundary Γ. we define

$$H(\mathrm{curl};\Omega) = \{\mathbf{v} \in L^2(\Omega)^3 : \ \nabla \times \mathbf{v} \in L^2(\Omega)^3\}$$

with the norm

$$\|\mathbf{v}\|_{H(\mathrm{curl};\Omega)} = \left(\|\mathbf{v}\|^2_{L^2(\Omega)} + \|\nabla \times \mathbf{v}\|^2_{L^2(\Omega)}\right)^{1/2}.$$

We define $H_0(\mathrm{curl};\Omega)$ to be the closure of $C_0^\infty(\Omega)^3$ in $H(\mathrm{curl};\Omega)$. $H(\mathrm{curl};\Omega)$ and $H_0(\mathrm{curl};\Omega)$ are Hilbert spaces.

Lemma 8.1. *Let Ω be a bounded Lipschitz domain. Let $\mathbf{v} \in H(\mathrm{curl};\mathbb{R}^3)$ vanish outside Ω. Then $\mathbf{v} \in H_0(\mathrm{curl};\Omega)$.*

Proof. Suppose for the moment that the domain Ω is strictly star-shaped with respect to one of its points. Without loss of generality, we may take the point as the origin. Then

$$\theta\bar{\Omega} \subset \Omega \quad \forall \theta \in [0,1) \quad \text{and} \quad \bar{\Omega} \subset \theta\Omega \quad \forall \theta > 1.$$

Define, for $\theta \in (0,1)$,

$$\mathbf{v}_\theta(x) = \mathbf{v}(x/\theta) \quad \forall x \in \mathbb{R}^3.$$

It is obvious that $\mathbf{v}_\theta$ has a compact support in Ω for $\theta \in (0,1)$ and

$$\lim_{\theta\to 1} \mathbf{v}_\theta = \mathbf{v} \ \text{ in } H(\mathrm{curl}, \mathbb{R}^3).$$

For any $\epsilon > 0$, let $\rho_\epsilon(x) = \epsilon^{-d}\rho(x/\epsilon) \in C_0^\infty(\mathbb{R}^3)$ be the mollifier function where $\rho(x)$ is defined in (1.4). Recall that $\rho_\epsilon = 0$ for $|x| > \epsilon$. Hence, for $\epsilon > 0$ sufficiently small, $\rho_\epsilon * \mathbf{v}_\theta$ is in $C_0^\infty(\Omega)^3$ and from Lemma 1.1,

$$\lim_{\epsilon\to 0}\lim_{\theta\to 1}(\rho_\epsilon * \mathbf{v}_\theta) = \mathbf{v} \ \text{ in } H(\mathrm{curl}, \Omega).$$

In the general case, Ω can be covered by a finite family of open sets

$$\Omega \subset \cup_{1\leqslant i\leqslant q} O_i$$

such that each $\Omega_i = \Omega \cap O_i$ is Lipschitz, bounded and strictly star-shaped. Let $\{\chi_i\}_{1\leqslant i\leqslant q}$ be a partition of unity subordinate to the family $\{O_i\}_{1\leqslant i\leqslant q}$, that is,

$$\chi_i \in C_0^\infty(O_i), \quad 0 \leqslant \chi_i \leqslant 1, \quad \text{and} \quad \sum_{i=1}^q \chi_i = 1 \ \text{ in } \Omega.$$

Then $\mathbf{v} = \sum_{i=1}^q \chi_i \mathbf{v}$ in $\mathbb{R}^3$. Clearly $\chi_i\mathbf{v} \in H(\mathrm{curl};\Omega)$ with support in Ω_i. Therefore, we can finish the proof by using the result for the strictly star-shaped domain in the first part of the proof. □

Theorem 8.1. *Let $\mathcal{D}(\bar{\Omega})$ be the set of all functions $\phi|_\Omega$ with $\phi \in C_0^\infty(\mathbb{R}^3)$. Then $\mathcal{D}(\bar{\Omega})^3$ is dense in $H(\text{curl};\Omega)$.*

Proof. Let $\mathbf{l}$ belong to $H(\text{curl};\Omega)'$, the dual space of $H(\text{curl};\Omega)$. As $H(\text{curl};\Omega)$ is a Hilbert space, by Riesz representation theorem, we can associate with $\mathbf{l}$ a function $\mathbf{u}$ in $H(\text{curl};\Omega)$ such that

$$\langle \mathbf{l}, \mathbf{v}\rangle = (\mathbf{u}, \mathbf{v}) + (\mathbf{w}, \nabla \times \mathbf{v}) \qquad \forall \mathbf{v} \in H(\text{curl};\Omega),$$

where

$$\mathbf{w} = \nabla \times \mathbf{u}.$$

Now assume that $\mathbf{l}$ vanishes on $\mathcal{D}(\bar{\Omega})^3$ and let $\tilde{\mathbf{u}}, \tilde{\mathbf{w}}$ be respectively the extension of $\mathbf{u}$, $\mathbf{w}$ by zero outside Ω. Then we have

$$\int_{\mathbb{R}^3} \{\tilde{\mathbf{u}} \cdot \mathbf{v} + \tilde{\mathbf{w}} \cdot \nabla \times \mathbf{v}\}\, \mathrm{d}x = 0 \qquad \forall \mathbf{v} \in C_0^\infty(\mathbb{R}^3)^3.$$

This implies that

$$\tilde{\mathbf{u}} = -\nabla \times \tilde{\mathbf{w}}.$$

Therefore $\tilde{\mathbf{w}} \in H(\text{curl};\mathbb{R}^3)$, since $\tilde{\mathbf{u}} \in L^2(\mathbb{R}^3)^3$. Now by Lemma 8.1 we have $\mathbf{w} \in H_0(\text{curl};\Omega)$. As $C_0^\infty(\Omega)^3$ is dense in $H_0(\text{curl};\Omega)$, let $\mathbf{w}_\epsilon$ be a sequence of functions in $C_0^\infty(\Omega)^3$ that tends to $\mathbf{w}$ in $H(\text{curl};\Omega)$ as $\epsilon \to 0$, then

$$\langle \mathbf{l}, \mathbf{v}\rangle = \lim_{\epsilon\to 0} \{(-\nabla \times \mathbf{w}_\epsilon, \mathbf{v}) + (\mathbf{w}_\epsilon, \nabla \times \mathbf{v})\} = 0 \qquad \forall \mathbf{v} \in H(\text{curl};\Omega).$$

Therefore, $\mathbf{l}$ vanishes on $\mathcal{D}(\bar{\Omega})^3$ implies that $\mathbf{l}$ also vanishes on $H(\text{curl};\Omega)$. This completes the proof. □

The following theorem about the tangential trace of the functions in $H(\text{curl};\Omega)$ is a direct consequence of the above theorem.

Theorem 8.2. *The mapping $\gamma_\tau :\ \mathbf{v} \to \mathbf{v} \times \mathbf{n}|_\Gamma$ defined on $\mathcal{D}(\bar{\Omega})^3$ can be extended by continuity to a linear and continuous mapping from $H(\text{curl};\Omega)$ to $H^{-1/2}(\Gamma)^3$. Moreover, the following Green formula holds*

$$\langle \mathbf{v}\times\mathbf{n},\, \mathbf{w}\rangle_\Gamma = \int_\Omega \mathbf{v}\cdot\nabla\times\mathbf{w}\,\mathrm{d}x - \int_\Omega \nabla\times\mathbf{v}\cdot\mathbf{w}\,\mathrm{d}x \quad \forall\, \mathbf{w} \in H^1(\Omega)^3,\ \mathbf{v} \in H(\text{curl};\Omega).$$

We remark that the trace operator γ_τ is not a surjective mapping. The following characterization of the $H_0(\text{curl};\Omega)$ follows from the definition of the space $H_0(\text{curl};\Omega)$ and Lemma 8.1.

Lemma 8.2. *We have*

$$H_0(\text{curl};\Omega) = \{\mathbf{v} \in H(\text{curl},\Omega) : \mathbf{v} \times \mathbf{n} = 0 \text{ on } \Gamma\}.$$

The Helmholtz decomposition plays an important role in the analysis and computation of electromagnetic fields. We start with the following generalization of the classical Stokes theorem.

THEOREM 8.3. *Let Ω be a simply connected Lipschitz domain. Then $\mathbf{u} \in L^2(\Omega)^3$ and $\nabla \times \mathbf{u} = 0$ if and only if there exists a function $\varphi \in H^1(\Omega)/\mathbb{R}$ such that $\mathbf{u} = \nabla\varphi$.*

PROOF. We first prove that $\mathbf{u} = \nabla\varphi$ for some $\varphi = L^2_{\text{loc}}(\Omega)$. Since $\nabla\varphi \in L^2(\Omega)^3$, we then easily have $\varphi \in L^2(\Omega)$. In fact, using the argument in Lemma 8.1 we may assume Ω is strictly star-shaped. Then we can introduce φ_θ as in Lemma 8.1. Since $\varphi \in L^2_{loc}(\Omega)$ and $\nabla\varphi \in L^2(\Omega)^3$, we know that $\nabla\varphi_\theta \to \nabla\varphi$ and $\int_D \varphi_\theta \to \int_D \varphi$ as $\theta \to 1$, where D is some compact subset of Ω. Now since $\varphi \in L^2_{loc}(\Omega)$ we know $\varphi_\theta \in L^2(\Omega)$. By using Poincaré inequality we know that φ_θ is a Cauchy sequence in $L^2(\Omega)$ as $\theta \to 1$. Thus there exists a $\varphi_1 \in L^2(\Omega)$ such that $\varphi_\theta \to \varphi_1$ in $L^2(\Omega)$. This implies $\varphi = \varphi_1 \in L^2(\Omega)$.

To show $\mathbf{u} = \nabla\varphi$ for some $\varphi = L^2_{\text{loc}}(\Omega)$, first we find a sequence of simply connected Lipschitz domain $\{\Omega_m\}_{m\geqslant 1}$ such that

$$\bar{\Omega}_m \subset \Omega, \quad \Omega_m \subset \Omega_{m+1}, \quad \Omega = \cup_{m\geqslant 1}\Omega_m.$$

In Ω_m we can smooth $\mathbf{u}$ so that its curl is zero and so we can apply the classical Stokes theorem for C^1 functions. For any $\epsilon > 0$, let $\rho_\epsilon(x) = \epsilon^{-d}\rho(x/\epsilon) \in C_0^\infty(\mathbb{R}^3)$ be the mollifier function where $\rho(x)$ is defined in (1.4). Recall that $\rho_\epsilon = 0$ for $|x| > \epsilon$. Denote by $\tilde{\mathbf{u}}$ the zero extension of $\mathbf{u}$ outside Ω. Then $\rho_\epsilon * \tilde{\mathbf{u}} \in C_0^\infty(\mathbb{R}^3)^3$, and

$$\rho_\epsilon * \tilde{\mathbf{u}} \to \tilde{\mathbf{u}} \text{ in } L^2(\mathbb{R}^3)^3, \qquad \nabla \times (\rho_\epsilon * \tilde{\mathbf{u}}) = \rho_\epsilon * \nabla \times \tilde{\mathbf{u}}.$$

For sufficiently small ϵ, we have $\cup_{x\in\Omega_m} B(x;\epsilon) \subset \Omega$. Thus

$$\nabla \times (\rho_\epsilon * \tilde{\mathbf{u}}) = 0 \text{ in } \Omega_m.$$

Now from the classical Stokes theorem, there is a smooth function $\varphi_\epsilon \in H^1(\Omega_m)/\mathbb{R}$ such that

$$\rho_\epsilon * \tilde{\mathbf{u}} = \nabla\varphi_\epsilon \text{ in } \Omega_m.$$

Let $\epsilon \to 0$, we know that there is a function $\varphi_m \in H^1(\Omega_m)$ such that $\varphi_\epsilon \to \varphi_m$ in $H^1(\Omega_m)/\mathbb{R}$, and

$$\mathbf{u} = \nabla\varphi_m \text{ in } \Omega_m.$$

But $\nabla\varphi_m = \nabla\varphi_{m+1}$ in Ω_m. Thus φ_m, φ_{m+1} differ by only a constant which we can choose as zero. Therefore

$$\varphi_{m+1} = \varphi_m \text{ in } \Omega_m \ \forall m \geqslant 1.$$

This defines a function $\varphi \in L^2_{\mathrm{loc}}(\Omega)$ such that $\mathbf{u} = \nabla\varphi$. □

Our next goal is to show that a vector field whose divergence vanishes must be a curl filed. We assume $\partial\Omega$ has $p+1$ connected parts $\Gamma_i, 0 \leqslant i \leqslant p$, and Γ_0 is the exterior boundary. We denote Ω_i the domain encompassed by Γ_i, $1 \leqslant i \leqslant p$ (see Figure 1).

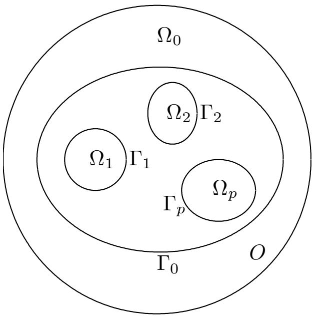

FIGURE 1. The domain Ω and the ball O.

THEOREM 8.4. *A vector field* $\mathbf{v} \in L^2(\Omega)^3$ *satisfies*

$$\operatorname{div}\mathbf{v} = 0 \quad \textit{in } \Omega, \qquad \langle \mathbf{v}\cdot\mathbf{n}, 1\rangle_{\Gamma_i} = 0, \quad 0 \leqslant i \leqslant p, \tag{8.1}$$

if and only if there is a vector potential $\mathbf{w} \in H^1(\Omega))^3$ *such that*

$$\mathbf{v} = \nabla \times \mathbf{w}. \tag{8.2}$$

Moreover, $\mathbf{w}$ *may be chosen such that* $\operatorname{div}\mathbf{w} = 0$ *and the following estimate holds*

$$\|\mathbf{w}\|_{H^1(\Omega)} \leqslant C\|\mathbf{v}\|_{L^2(\Omega)}. \tag{8.3}$$

PROOF. 1°) Let $\mathbf{w} \in H^1(\Omega)^3$ and $\mathbf{v} = \nabla \times \mathbf{w}$. Obviously $\operatorname{div}\mathbf{v} = 0$. For $0 \leqslant i \leqslant p$, let $\chi_i \in C_0^\infty(\mathbb{R}^3)$ be the cut-off function such that $0 \leqslant \chi_i \leqslant 1$, $\chi_i = \delta_{ij}$ in the neighborhood of Γ_j. Define $\mathbf{v}_i = \nabla \times (\chi_i \mathbf{w})$. Then

$$\langle \mathbf{v}\cdot\mathbf{n}, 1\rangle_{\Gamma_i} = \langle \mathbf{v}_i\cdot\mathbf{n}, 1\rangle_{\Gamma} = \int_\Omega \operatorname{div}\mathbf{v}_i dx = 0, \qquad 0 \leqslant i \leqslant p.$$

This shows (8.1).

2°) Now let us assume (8.1) holds. First we extend $\mathbf{v}$ to be a function in $\mathbb{R}^3$ so that its divergence is zero. Let O be a ball containing Ω (see Figure 1). For $0 \leqslant i \leqslant p$, denote by $\theta_i \in H^1(\Omega_i)$ the solution of the following problem

$$-\Delta\theta_0 = 0 \text{ in } \Omega_0 = O \setminus \big(\bar{\Omega} \cup_{i=1}^p \Omega_i\big),\ \partial_{\mathbf{n}}\theta_0 = \mathbf{v}\cdot\mathbf{n} \text{ on } \Gamma_0,\ \partial_{\mathbf{n}}\theta_0 = 0 \text{ on } \partial O,$$
$$-\Delta\theta_i = 0 \text{ in } \Omega_i,\ \partial_{\mathbf{n}}\theta_i = \mathbf{v}\cdot\mathbf{n} \text{ on } \Gamma_i,\ 0 \leqslant i \leqslant p.$$

Define

$$\tilde{\mathbf{v}} = \begin{cases} \mathbf{v} & \text{in } \ \Omega, \\ \nabla\theta_i & \text{in } \ \Omega_i, \ \ 1 \leqslant i \leqslant p, \\ 0 & \text{in } \ \mathbb{R}^3 \backslash \bar{O}. \end{cases}$$

Then $\tilde{\mathbf{v}} \in L^2(\mathbb{R}^3)$ and $\operatorname{div} \tilde{\mathbf{v}} = 0$. Let $\hat{\mathbf{v}} = (\hat{v}_1, \hat{v}_2, \hat{v}_3)^T$ be the Fourier transform of $\tilde{\mathbf{v}}$

$$\hat{\mathbf{v}}(\xi) = \int_{\mathbb{R}^3} e^{-2\pi\mathbf{i}(x,\xi)}\tilde{\mathbf{v}}(x)\,\mathrm{d}x, \quad (x,\xi) = \sum_{i=1}^3 x_i\xi_i.$$

By taking the Fourier transform of (8.1) we obtain

$$\xi_1\hat{v}_1 + \xi_2\hat{v}_2 + \xi_3\hat{v}_3 = 0. \tag{8.4}$$

Notice that if (8.2) is satisfied we need

$$\hat{\mathbf{v}} = 2\pi\mathbf{i}(\xi_2\hat{w}_3 - \xi_3\hat{w}_2, \xi_3\hat{w}_1 - \xi_1\hat{w}_3, \xi_1\hat{w}_2 - \xi_2\hat{w}_1)^T. \tag{8.5}$$

If $\operatorname{div}\mathbf{w} = 0$ we need

$$\xi_1\hat{w}_1 + \xi_2\hat{w}_2 + \xi_3\hat{w}_3 = 0. \tag{8.6}$$

Solving $\hat{\mathbf{w}}$ from (8.4)-(8.6) we get

$$\hat{\mathbf{w}} = \frac{1}{2\pi\mathbf{i}|\xi|^2}(\xi_3\hat{v}_2 - \xi_2\hat{v}_3, \xi_1\hat{v}_3 - \xi_3\hat{v}_1, \xi_2\hat{v}_1 - \xi_1\hat{v}_2)^T.$$

We will define $\mathbf{w}$ as the inverse Fourier transform of the above function. Obviously $\nabla\mathbf{w} \in L^2(\Omega)^{3\times3}$ because

$$|\xi_j\hat{\mathbf{w}}_k| \leqslant \frac{1}{2\pi}(|\hat{v}_1| + |\hat{v}_2| + |\hat{v}_3|). \tag{8.7}$$

Now we show $\mathbf{w} \in L^2(\Omega)^3$. Denote by $\omega \in C_0^\infty(\mathbb{R}^3)$ the function which is 1 in the neighborhood of the origin. Then

$$\hat{\mathbf{w}}(\xi) = \omega(\xi)\hat{\mathbf{w}}(\xi) + (1-\omega(\xi))\hat{\mathbf{w}}(\xi).$$

From (8.5) we know that $\hat{v}_j(0) = 0$. Since $\hat{v}_j(\xi)$ is holomorphic, we know that, in the neighborhood of the origin,

$$\hat{v}_j(\xi) = \sum_{k=1}^{3} \frac{\partial \hat{v}_j}{\partial \xi_k}\xi_k + O(|\xi|^2).$$

Thus $\hat{\mathbf{w}}$ is bounded in the neighborhood of the origin. Now $\omega\hat{\mathbf{w}}$ has the compact support, its inverse Fourier transform is holomorphic and its restriction to Ω belongs to $L^2(\Omega)^3$. On the other hand, $(1-\omega)\hat{\mathbf{w}}$ is zero in the neighborhood of the origin. Hence $(1-\omega)\hat{\mathbf{w}} \in L^2(\mathbb{R}^3)^3$ and its inverse Fourier transform in $L^2(\mathbb{R}^3)^3$. This proves the inverse Fourier transform of $\hat{\mathbf{w}}$ is in $L^2(\Omega)^3$.

Clearly $\mathbf{w}$ can be chosen up to an arbitrary constant. Thus (8.3) follows from (8.7), the Parserval identity, and Poincaré inequality. □

The following Helmholtz decomposition theorem is now a direct consequence of Theorems 8.3 and 8.4.

THEOREM 8.5. *Any vector field* $\mathbf{v} \in L^2(\Omega)^3$ *has the following orthogonal decomposition*

$$\mathbf{v} = \nabla q + \nabla \times \mathbf{w},$$

where $q \in H^1(\Omega)/\mathbb{R}$ *is the unique solution of the following problem*

$$(\nabla q, \nabla\varphi) = (\mathbf{v}, \nabla\varphi) \qquad \forall \varphi \in H^1(\Omega),$$

and $\mathbf{w} \in H^1(\Omega)^3$ *satisfies* $\mathrm{div}\,\mathbf{w} = 0$ *in* Ω, $\nabla \times \mathbf{w} \cdot \mathbf{n} = 0$ *on* Γ.

We conclude this section by proving the embedding theorem for function spaces $\mathbf{X}_N(\Omega)$ and $\mathbf{X}_T(\Omega)$ which will be used in our subsequent analysis

$$\mathbf{X}_N(\Omega) = \{\mathbf{v} \in L^2(\Omega)^3 : \ \nabla \times \mathbf{v} \in L^2(\Omega)^3, \ \mathrm{div}\mathbf{v} \in L^2(\Omega), \ \mathbf{v} \times \mathbf{n} = 0 \text{ on } \Gamma\},$$
$$\mathbf{X}_T(\Omega) = \{\mathbf{v} \in L^2(\Omega)^3 : \ \nabla \times \mathbf{v} \in L^2(\Omega)^3, \ \mathrm{div}\mathbf{v} \in L^2(\Omega), \ \mathbf{v} \cdot \mathbf{n} = 0 \text{ on } \Gamma\}.$$

THEOREM 8.6. *If* Ω *is a* C^1 *or convex domain,* $\mathbf{X}_N(\Omega), \mathbf{X}_T(\Omega)$ *are continuously embedded into* $H^1(\Omega)^3$.

PROOF. Without loss of generality, we may assume Ω is also simply connected and has connected boundary. For, otherwise, Ω is the union of finite number of domains Ω_k having above properties. We can introduce the partition of unity χ_k subordinate to Ω_k and apply the result for each $\chi_k\mathbf{v}$.

1°) Let $\mathbf{v} \in \mathbf{X}_T(\Omega)$. By Theorem 8.4, for $\nabla \times \mathbf{v}$, we have vector potential $\mathbf{w} \in H^1(\Omega)^3$ such that

$$\nabla \times \mathbf{w} = \nabla \times \mathbf{v}, \quad \mathrm{div}\,\mathbf{w} = 0 \ \text{ in } \Omega.$$

Moreover, $\|\mathbf{w}\|_{H^1(\Omega)} \leqslant C\|\nabla \times \mathbf{v}\|_{L^2(\Omega)}$. Thus $\nabla \times (\mathbf{v} - \mathbf{w}) = 0$ and by Theorem 8.3, $\mathbf{v} - \mathbf{w} = \nabla\varphi$ for some function $\varphi \in H^1(\Omega)$. Moreover, φ satisfies

$$-\Delta\varphi = -\mathrm{div}\,\mathbf{v} \ \text{ in } \Omega, \quad \partial_{\mathbf{n}}\varphi = -\mathbf{w}\cdot\mathbf{n} \ \text{ on } \Gamma. \tag{8.8}$$

Since $\mathbf{w} \in H^1(\Omega)^3$, $\mathbf{w}\cdot\mathbf{n} \in H^{1/2}(\Gamma)$. Now if Ω is a C^1 domain, then the regularity theory for elliptic equations implies $\varphi \in H^2(\Omega)$. Thus $\mathbf{v} = \mathbf{w} + \nabla\varphi$ belongs to $H^1(\Omega)^3$. Moreover,

$$\begin{aligned} |\varphi|_{H^2(\Omega)} \leqslant & C\left(\| \,\mathrm{div}\mathbf{v}\,\|_{L^2(\Omega)} + \|\,\mathbf{w}\cdot\mathbf{n}\,\|_{H^{1/2}(\Gamma)}\right) \\ \leqslant & C\left(\| \,\mathrm{div}\mathbf{v}\,\|_{L^2(\Omega)} + \|\,\nabla\times\mathbf{v}\,\|_{L^2(\Omega)}\right). \end{aligned}$$

Thus

$$|\mathbf{v}|_{H^1(\Omega)} \leqslant C\left(\| \,\mathrm{div}\mathbf{v}\,\|_{L^2(\Omega)} + \|\,\nabla\times\mathbf{v}\,\|_{L^2(\Omega)}\right).$$

2°) Let $\mathbf{v} \in \mathbf{X}_N(\Omega)$. Let O be a ball in $\mathbb{R}^3$ that includes Ω. Denote by $\tilde{\mathbf{v}}$ the zero extension to O of $\mathbf{v}$. By Theorem 8.4, there exists a function $\mathbf{w} \in H^1(O)^3$ such that

$$\nabla\times\mathbf{w} = \nabla\times\tilde{\mathbf{v}}, \quad \mathrm{div}\,\mathbf{w} = 0 \ \text{ in } O.$$

Now $\mathbf{w}$ is curl free in $O\backslash\bar{\Omega}$ and by Theorem 8.3, $\mathbf{w} = \nabla\psi$ for some $\psi \in H^2(O\backslash\bar{\Omega})$. On the other hand, $\nabla\times(\tilde{\mathbf{v}} - \mathbf{w}) = 0$ and O is simply connected, again by Theorem 8.3, $\mathbf{v} - \mathbf{w} = \nabla\varphi$ for some function $\varphi \in H^1(O)$. Clearly φ satisfies

$$-\Delta\varphi = -\mathrm{div}\mathbf{v} \ \text{ in } \Omega, \quad \varphi = -\psi \ \text{ on } \Gamma. \tag{8.9}$$

If Ω is a C^1 domain, then $\varphi \in H^2(\Omega)$ and consequently $\mathbf{v} = \mathbf{w} + \nabla\varphi \in H^1(\Omega)^3$. Moreover,

$$|\mathbf{v}|_{H^1(\Omega)} \leqslant C\left(\| \,\mathrm{div}\mathbf{v}\,\|_{L^2(\Omega)} + \|\,\nabla\times\mathbf{v}\,\|_{L^2(\Omega)}\right).$$

3°) If Ω is a convex domain, the regularity theory for elliptic equation ensures that the solution φ of (8.9) is in $H^2(\Omega)$ and thus $\mathbf{X}_N(\Omega)$ is continuously embedded into $H^1(\Omega)^3$ by using the same argument in 2°). For the case of $\mathbf{X}_T(\Omega)$, the regularity of the elliptic equation with Neumann condition in (8.8) is unknown. A different approach is used to prove the embedding theorem. We refer the reader to the monograph [**34**]. □

8.2. The curl conforming finite element approximation

In this section we only consider the lowest order Nédélec finite element space.

Definition 8.7. The lowest order Nédélec finite element is a triple $(K, \mathcal{P}, \mathcal{N})$ with the following properties

(i) $K \subset \mathbb{R}^3$ is a tetrahedron;

(ii) $\mathcal{P} = \{\mathbf{u} = \mathbf{a}_K + \mathbf{b}_K \times x \quad \forall \mathbf{a}_K, \mathbf{b}_K \in \mathbb{R}^3\}$;

(iii) $\mathcal{N} = \{M_e : M_e(\mathbf{u}) = \int_e (\mathbf{u} \cdot \mathbf{t})\, \mathrm{d}l \quad \forall$ edge e of K, $\forall \mathbf{u} \in \mathcal{P}\}$. $M_e(\mathbf{u})$ is called the moment of $\mathbf{u}$ on the edge e.

Note that if $\mathbf{u} \in \mathcal{P}_1(K)^3$ then $\nabla \times \mathbf{u}$ is a constant vector, say $\nabla \times \mathbf{u} = 2\mathbf{b}_K$, which implies $\nabla \times (\mathbf{u} - \mathbf{b}_K \times x) = 0$ in K. We get $\mathbf{u} = \nabla\varphi + \mathbf{b}_K \times x$ for some $\varphi \in \mathcal{P}_2(K)$. When $\mathbf{b}_K = 0$, $\mathbf{u}$ should approximate the function in $L^2(K)^3$, the minimum requirement is $\varphi \in \mathcal{P}_1(K)$, that is, $\nabla\varphi = \mathbf{a}_K$ for some constant vector $\mathbf{a}_K$ in $\mathbb{R}^3$. This motivates the shape functions in $\mathcal{P}$.

Lemma 8.3. *The nodal basis of the lowest order Nédélec element is* $\{\lambda_i \nabla \lambda_j - \lambda_j \nabla \lambda_i, 1 \leqslant i < j \leqslant 4\}$. *Here* λ_j, $j = 1, 2, 3, 4$, *are barycentric coordinate functions of the element* K.

Proof. Let K be the tetrahedron with four vertices A_1, A_2, A_3, A_4 corresponding to $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, respectively. We first notice that the normal to the face F_{123} with vertices A_1, A_2, A_3 is parallel to $\nabla\lambda_4$. In fact, for any edge e of F_{123} with tangential vector $\mathbf{t}_e$,

$$\nabla\lambda_4 \cdot \mathbf{t}_e = \frac{\partial \lambda_4}{\partial \mathbf{t}_e} = 0.$$

Similarly, we have the normal to the face F_{234} is parallel to $\nabla\lambda_1$.

Now we show that the basis function corresponding to the edge e_{14} is $\mathbf{u}_{14} = \lambda_1 \nabla\lambda_4 - \lambda_4 \nabla\lambda_1$. In fact, since $\lambda_1 = 0$ on the face F_{234} and $\lambda_4 = 0$ on the face F_{123}, we have

$$\int_e \mathbf{u}_{14} \cdot \mathbf{t}_e \, \mathrm{d}l = 0 \quad \forall\ e \neq e_{14}.$$

It remains to prove

$$\int_{e_{14}} \mathbf{u}_{14} \cdot \mathbf{t}_{14} \, \mathrm{d}l = 1.$$

Noting that $\lambda_1 + \lambda_4 = 1$ on e_{14}, we have

$$\mathbf{u}_{14} = \nabla\lambda_4 - \lambda_4 \nabla(\lambda_1 + \lambda_4) = \nabla\lambda_4 + \lambda_4 \nabla(\lambda_2 + \lambda_3).$$

Therefore

$$\int_{e_{14}} \mathbf{u}_{14} \cdot \mathbf{t}_{14}\, \mathrm{d}l = \int_{e_{14}} \nabla\lambda_4 \cdot \mathbf{t}_{14}\, \mathrm{d}l = \int_{e_{14}} \frac{\partial \lambda_4}{\partial \mathbf{t}_{14}}\, \mathrm{d}l = 1.$$

This completes the proof. □

Let K be a tetrahedron with vertices A_i, $1 \leqslant i \leqslant 4$, and let $F_K : \hat{K} \to K$ be the affine transform from the reference element $\hat{K}$ to K

$$x = F_K(\hat{x}) = B_K\hat{x} + \mathbf{b}_K, \quad \hat{x} \in \hat{K}, \quad B_K \text{ is invertible},$$

so that $F_K(\hat{A}_i) = A_i, 1 \leqslant i \leqslant 4$. Notice that the normal and tangential vectors $\mathbf{n}, \hat{\mathbf{n}}$ and $\mathbf{t}, \hat{\mathbf{t}}$ to the faces satisfy

$$\mathbf{n} \circ F_K = (B_K^{-1})^T\hat{\mathbf{n}}/|(B_K^{-1})^T\hat{\mathbf{n}}|, \qquad \mathbf{t} \circ F_K = B_K\hat{\mathbf{t}}/|B_K\mathbf{t}|.$$

For any scaler function φ defined on K, we associate

$$\hat{\varphi} = \varphi \circ F_K, \qquad \text{that is,} \qquad \hat{\varphi}(\hat{x}) = \varphi(B_K\hat{x} + \mathbf{b}_K).$$

For any vector valued function $\mathbf{u}$ defined on K, we associate

$$\hat{\mathbf{u}} = B_K^T\mathbf{u} \circ F_K, \qquad \text{that is,} \qquad \hat{\mathbf{u}}(\hat{x}) = B_K^T\mathbf{u}(B_K\hat{x} + \mathbf{b}_K). \tag{8.10}$$

Denote by $\mathbf{u} = (u_1, u_2, u_3), \hat{\mathbf{u}} = (\hat{u}_1, \hat{u}_2, \hat{u}_3)$. We introduce

$$\mathcal{C} = \left(\frac{\partial u_i}{\partial x_j} - \frac{\partial u_j}{\partial x_i}\right)_{i,j=1}^3 \quad \text{and} \quad \hat{\mathcal{C}} = \left(\frac{\partial \hat{u}_i}{\partial \hat{x}_j} - \frac{\partial \hat{u}_j}{\partial \hat{x}_i}\right)_{i,j=1}^3.$$

Then we have

$$\mathcal{C} \circ F_K = (B_K^{-1})^T\hat{\mathcal{C}}B_K^{-1}. \tag{8.11}$$

In fact,

$$\frac{\partial \hat{u}_i}{\partial \hat{x}_j} = \frac{\partial}{\partial \hat{x}_j}\Big(\sum_k b_{ki}(u_k \circ F_K)\Big) = \sum_{k,l} b_{ki}\frac{\partial u_k}{\partial x_l}b_{lj}$$

and

$$\frac{\partial \hat{u}_i}{\partial \hat{x}_j} - \frac{\partial \hat{u}_j}{\partial \hat{x}_i} = \sum_{k,l} b_{ki}\frac{\partial u_k}{\partial x_l}b_{lj} - \sum_{k,l} b_{kj}\frac{\partial u_k}{\partial x_l}b_{li} = \sum_{k,l} b_{ki}\left(\frac{\partial u_k}{\partial x_l} - \frac{\partial u_l}{\partial x_k}\right)b_{lj}.$$

This yields

$$\hat{\mathcal{C}}_{ij} = \sum_{k,l} b_{ki}\mathcal{C}_{kl}b_{lj} \qquad \text{and hence} \qquad \hat{\mathcal{C}} = B_K^T(\mathcal{C} \circ F_K)B_K,$$

that is, (8.11) holds.

Lemma 8.4. *We have*

(i) $\mathbf{u} \in \mathcal{P}(K) \Leftrightarrow \hat{\mathbf{u}} \in \hat{\mathcal{P}}(\hat{K})$;

(ii) $\nabla \times \mathbf{u} = 0 \Leftrightarrow \hat{\nabla} \times \hat{\mathbf{u}} = 0, \quad \forall\, \mathbf{u} \in \mathcal{P}(K)$;

(iii) $M_e(\mathbf{u}) = 0 \Leftrightarrow M_{\hat{e}}(\hat{\mathbf{u}}) = 0, \quad \forall\, \mathbf{u} \in \mathcal{P}(K)$;

(iv) *Let* $\mathbf{u} \in \mathcal{P}(K)$ *and* F *be a face of* K*. If* $M_e(\mathbf{u}) = 0$ *for any edge* $e \subset \partial F$*, then* $\mathbf{u} \times \mathbf{n} = 0$ *on* F*;*

(v) *If* $\mathbf{u} \in \mathcal{P}(K)$ *and* $M_e(\mathbf{u}) = 0$ *for any edge* e*, then* $\mathbf{u} = 0$ *in* K*.*

PROOF. (i) For $\mathbf{u} = \mathbf{a}_K + \mathbf{b}_K \times x \in \mathcal{P}(K)$, we have

$$\begin{aligned}\hat{\mathbf{u}}(\hat{x}) &= B_K^T(\mathbf{a}_K + \mathbf{b}_K \times (B_K\hat{x} + \mathbf{b}_K)) \\ &= B_K^T(\mathbf{a}_K + \mathbf{b}_K \times \mathbf{b}_K) + B_K^T(\mathbf{b}_K \times B_K\hat{x}) \\ &= B_K^T\mathbf{a}_K + \hat{\mathbf{b}}_{\hat{K}} \times \hat{x} \qquad (\text{by } B_K^T(\mathbf{b}_K \times B_K\hat{x}) \cdot \hat{x} = 0) \\ &\in \hat{\mathcal{P}}(\hat{K}).\end{aligned}$$

(ii) From (8.11), it is obvious.

(iii) By definition,

$$\begin{aligned}M_e(\mathbf{u}) = \int_e \mathbf{u} \cdot \mathbf{t}\, \mathrm{d}l &= \frac{|e|}{|\hat{e}|}\int_{\hat{e}} \mathbf{u}(F_K(\hat{x})) \cdot \frac{B_K\hat{\mathbf{t}}}{|B_K\hat{\mathbf{t}}|}\mathrm{d}\hat{l} \\ &= \frac{|e|}{|\hat{e}|}\frac{1}{|B_K\hat{\mathbf{t}}|}\int_{\hat{e}} \hat{\mathbf{u}} \cdot \hat{\mathbf{t}}\mathrm{d}\hat{l} \\ &= \frac{|e|}{|\hat{e}|}\frac{1}{|B_K\hat{\mathbf{t}}|}M_{\hat{e}}(\hat{\mathbf{u}}). \qquad (8.12)\end{aligned}$$

(iv) Without loss of generality, we may assume

$$F \subset \{x = (x_1, x_2, x_3) \in \mathbb{R}^3 : \quad x_3 = 0\}.$$

First by Stokes theorem

$$\int_F \nabla \times \mathbf{u} \cdot \mathbf{n}\, \mathrm{d}s = \int_{\partial F} \mathbf{u} \cdot \mathbf{t}\, \mathrm{d}l = 0$$

which implies $\nabla \times \mathbf{u} \cdot \mathbf{n} = 0$ on F, i.e.,

$$\frac{\partial u_1}{\partial x_2} - \frac{\partial u_2}{\partial x_1} = 0 \qquad \text{on } \{x_3 = 0\}. \qquad (8.13)$$

Let $\mathbf{u}' = (u_1(x_1, x_2, 0), u_2(x_1, x_2, 0))$ and

$$\mathbf{u} = \mathbf{a}_K + \mathbf{b}_K \times x = (b_2x_3 - b_3x_2, b_3x_1 - b_1x_3, b_1x_2 - b_2x_1) + \mathbf{a}_K.$$

(8.13) implies $b_3 = 0$. Thus $\mathbf{u}'$ is constant. Note that $\mathbf{u} \cdot \mathbf{t} = \mathbf{u}' \cdot \mathbf{t}$ on F. By assumption, we have, $M_e(\mathbf{u}) = 0$, for any edge $e \subset F$, which implies $\mathbf{u}' \cdot \mathbf{t} = 0$ on any edge $e \subset F$. Thus $\mathbf{u}' = 0$ in F. This shows $\mathbf{u} \times \mathbf{n} = (u_2(x_1, x_2, 0), -u_1(x_1, x_2, 0), 0) = 0$ on F.

(v) At first, by (iv) we know that $\mathbf{u} \times \mathbf{n} = 0$ on F for any face F. Thus, from Theorem 8.2,

$$\int_K \nabla \times \mathbf{u} \cdot \mathbf{b}_K \,\mathrm{d}x = -\int_{\partial K} (\mathbf{u} \times \mathbf{n}) \cdot \mathbf{b}_K \,\mathrm{d}s = 0$$

which implies $\mathbf{b}_K = 0$, that is, $\mathbf{u} = \mathbf{a}_K$. Now $\mathbf{a}_K \times \mathbf{n} = 0$ for any face F yields $\mathbf{a}_K = 0$. □

This lemma induces a natural interpolation operator on K.

Definition 8.8. Let K be an arbitrary tetrahedron in $\mathbb{R}^3$ and $\mathbf{u} \in W^{1,p}(K)^3$ for some $p > 2$. Its interpolant $\gamma_K \mathbf{u}$ is a unique polynomial in $\mathcal{P}(K)$ that has the same moments as $\mathbf{u}$ on K. In other words, $M_e(\gamma_K \mathbf{u} - \mathbf{u}) = 0$.

Recall that for any bounded Lipschitz domain, the trace theorem says that the trace of any function in $W^{s,p}(\Omega)$ is in $W^{s-1/p,p}(\partial\Omega)$, where $s > 1/p$. Thus if $\mathbf{u} \in W^{1,p}(K)^3$ for some $p > 2$, $\mathbf{u} \in W^{1-1/p,p}(\partial F)$ for any face F of K. Again by the trace theorem $\mathbf{u} \in W^{1-2/p,p}(\partial F)$. Therefore, $M_e(\mathbf{u})$ is well-defined for functions $\mathbf{u}$ in $W^{1,p}(K)$, $p > 2$.

The following lemma indicates that the interpolation operator γ_K can also be defined for functions with weaker regularity.

Lemma 8.5. *For any $p > 2$, the operator γ_K is continuous on the space*

$$\{\mathbf{v} \in L^p(K)^3 : \nabla \times \mathbf{v} \in L^p(K)^3 \quad \textit{and}\; \mathbf{v} \times \mathbf{n} \in L^p(\partial K)^3\}.$$

Proof. Let p' be such that $1/p + 1/p' = 1$. For an edge e of a face F of K, we let φ be the function which equals to 1 on e and 0 on the other edges of F. Then $\varphi \in W^{1-1/p',p'}(\partial F)$ since $1 - 1/p' < 1/2$. Denote $\bar{\varphi}$ its lifting from $W^{1-1/p',p'}(\partial F)$ to $W^{1,p'}(F)$. Next, we extend $\bar{\varphi}$ to be zero on the other faces of ∂K and denote $\bar{\bar{\varphi}}$ its lifting from $W^{1-1/p',p'}(\partial K)$ to $W^{1,p'}(K)$. By Stokes theorem and Green formula

$$\begin{aligned} M_e(\mathbf{v}) &= \int_{\partial F} (\mathbf{v} \cdot \mathbf{t})\varphi \,\mathrm{d}l \\ &= \int_F \nabla \times (\bar{\varphi}\mathbf{v}) \cdot \mathbf{n} \,\mathrm{d}s \\ &= \int_F \bar{\varphi} \nabla \times \mathbf{v} \cdot \mathbf{n} \,\mathrm{d}s + \int_F \nabla\bar{\varphi} \times \mathbf{v} \cdot \mathbf{n} \,\mathrm{d}s \\ &= \int_K \nabla \times \mathbf{v} \cdot \nabla\bar{\bar{\varphi}} \,\mathrm{d}x + \int_F \nabla\bar{\varphi} \times \mathbf{v} \cdot \mathbf{n} \,\mathrm{d}s. \end{aligned}$$

This implies

$$|M_e(\mathbf{v})| \leqslant C(\| \nabla \times \mathbf{v} \|_{L^p(K)} + \| \mathbf{v} \times \mathbf{n} \|_{L^p(F)})\| \varphi \|_{W^{1-1/p',p'}(e)}$$

with a constant C depending only on K and p. This completes the proof. □

Let Ω be a bounded polyhedron and $\mathcal{M}_h$ be a regular mesh of Ω. We set

$$\mathbf{X}_h = \{\mathbf{u}_h \in H(\mathrm{curl};\Omega) : \ \mathbf{u}_h|_K \in \mathcal{P}(K) \ \ \forall K \in \mathcal{M}_h\}.$$

For any function $\mathbf{u}$ whose moments are defined on all edges of the mesh $\mathcal{M}_h$, we define the interpolation operator γ_h by

$$\gamma_h \mathbf{u}|_K = \gamma_K \mathbf{u} \quad \text{on } K, \ \ \forall K \in \mathcal{M}_h.$$

THEOREM 8.9. *Let $\mathbf{u} \in H^1(\mathrm{curl};\Omega)$, that is, $\mathbf{u} \in H^1(\Omega)^3$ and $\nabla \times \mathbf{u} \in H^1(\Omega)^3$. We have*

$$\|\mathbf{u} - \gamma_h \mathbf{u}\|_{H(\mathrm{curl};\Omega)} \leqslant Ch \left(|\mathbf{u}|_{H^1(\Omega)} + |\nabla \times \mathbf{u}|_{H^1(\Omega)} \right).$$

PROOF. First it follows from (8.12) that

$$\widehat{\gamma_K \mathbf{u}} = \gamma_{\hat{K}} \hat{\mathbf{u}}, \quad \text{i.e.,} \quad B_K^T[(\gamma_K \mathbf{u}) \circ F_K] = \gamma_{\hat{K}}[B_K^T(\mathbf{u} \circ F_K)].$$

From (8.10), we have

$$\|\mathbf{u} - \gamma_h \mathbf{u}\|_{L^2(K)} \leqslant |\det(B_K)|^{1/2} \, \|B_K^{-1}\| \, \|\hat{\mathbf{u}} - \gamma_{\hat{K}} \hat{\mathbf{u}}\|_{L^2(\hat{K})}.$$

Since $P_0(\hat{K})^3 \subset \mathcal{P}(\hat{K})$, for any $\hat{\mathbf{p}} \in P_0(\hat{K})^3$, we have

$$\|\hat{\mathbf{u}} - \gamma_{\hat{K}} \hat{\mathbf{u}}\|_{L^2(\hat{K})} = \|(I - \gamma_{\hat{K}})(\hat{\mathbf{u}} + \hat{\mathbf{p}})\|_{L^2(\hat{K})}.$$

But the degrees of freedom of $\mathbf{u}$ may be estimated by using Lemma 8.5 and by using the Sobolev imbedding theorem to get

$$\begin{aligned} \|(I - \gamma_{\hat{K}})(\hat{\mathbf{u}} + \hat{\mathbf{p}})\|_{L^2(\hat{K})} &\leqslant C \left(\|\hat{\mathbf{u}} + \hat{\mathbf{p}}\|_{H^1(\hat{K})} + \|\hat{\nabla} \times (\hat{\mathbf{u}} + \hat{\mathbf{p}})\|_{H^1(\hat{K})} \right) \\ &= C \left(\|\hat{\mathbf{u}} + \hat{\mathbf{p}}\|_{H^1(\hat{K})} + \|\hat{\nabla} \times \hat{\mathbf{u}}\|_{H^1(\hat{K})} \right) \end{aligned}$$

Now, by using Theorem 3.1,

$$\inf_{\hat{\mathbf{p}} \in P_0(\hat{K})^3} \|(I - \gamma_{\hat{K}})(\hat{\mathbf{u}} + \hat{\mathbf{p}})\|_{L^2(\hat{K})} \leqslant C \left(|\hat{\mathbf{u}}|_{H^1(\hat{K})} + |\hat{\nabla} \times \hat{\mathbf{u}}|_{H^1(\hat{K})} \right).$$

Mapping back to the original element K and using (8.11) we obtain

$$\begin{aligned} \|\mathbf{u} - \gamma_h \mathbf{u}\|_{L^2(K)} &\leqslant C|\det(B_K)|^{1/2} \|B_K^{-1}\| (|\hat{\mathbf{u}}|_{H^1(\hat{K})} + |\hat{\nabla} \times \hat{\mathbf{u}}|_{H^1(\hat{K})}) \\ &\leqslant C\|B_K^{-1}\| \, \|B_K\|^2 (|\mathbf{u}|_{H^1(K)} + \|B_K\| \, |\nabla \times \mathbf{u}|_{H^1(K)}) \\ &\leqslant Ch_K (|\mathbf{u}|_{H^1(K)} + |\nabla \times \mathbf{u}|_{H^1(K)}). \end{aligned}$$

To show the curl estimate, we use the $H(\mathrm{div},\Omega)$ conforming finite element space $\mathbf{W}_h$ and the interpolation operator $\tau_h : H(\mathrm{div},\Omega) \to \mathbf{W}_h$ in Exercise 8.4 to obtain

$$\|\nabla \times (\mathbf{u} - \gamma_h \mathbf{u})\|_{L^2(\Omega)} \leqslant \|(I - \tau_h)\nabla \times \mathbf{u}\|_{L^2(\Omega)} \leqslant Ch\|\nabla \times \mathbf{u}\|_{H^1(\Omega)}.$$

This proves the theorem. □

8.3. Finite element methods for time harmonic Maxwell equations

Let Ω be bounded polyhedral domain in $\mathbb{R}^3$. In this section we consider the finite element approximation to the time harmonic Maxwell equation

$$\nabla \times (\alpha(x)\nabla \times \mathbf{E}) - k^2\beta(x)\mathbf{E} = \mathbf{f} \quad \text{in } \Omega,$$

with the boundary condition

$$\mathbf{E} \times \mathbf{n} = 0 \qquad \text{on} \quad \partial\Omega.$$

Here $k > 0$ is the wave number. We assume $\mathbf{f} \in L^2(\Omega)^3$, $\alpha, \beta \in L^\infty(\Omega)$ such that $\alpha \geqslant \alpha_0 > 0$ and $\beta \geqslant \beta_0 > 0$.

The variational formulation is to find $\mathbf{E} \in H_0(\text{curl}; \Omega)$ such that

$$(\alpha\nabla \times \mathbf{E}, \nabla \times \mathbf{v}) - k^2(\beta\mathbf{E}, \mathbf{v}) = (\mathbf{f}, \mathbf{v}) \quad \forall \mathbf{v} \in H_0(\text{curl}; \Omega). \tag{8.14}$$

The problem (8.14) is not necessarily coercive and thus its uniqueness and existence is not guaranteed. Here we will not elaborate on this issue and simply assume (8.14) has a unique solution $\mathbf{E} \in H_0(\text{curl}; \Omega)$ for any given $\mathbf{f} \in L^2(\Omega)^3$.

Let $\mathbf{X}_h^0 = \mathbf{X}_h \cap H_0(\text{curl}; \Omega)$. Then the finite element approximation to (8.14) is to find $\mathbf{E}_h \in \mathbf{X}_h^0$ such that

$$(\alpha\nabla \times \mathbf{E}_h, \nabla \times \mathbf{v}_h) - k^2(\beta\mathbf{E}_h, \mathbf{v}_h) = (\mathbf{f}, \mathbf{v}_h) \qquad \forall\, \mathbf{v}_h \in \mathbf{X}_h^0. \tag{8.15}$$

The discrete problem (8.15) may not have a unique solution. It can be proved that for sufficiently small mesh size h, the problem (8.15) indeed has a unique solution under fairly general conditions on the domain and the coefficients α, β. Here we only consider a special case when the domain is a convex polyhedron and α, β are constants.

Theorem 8.10. *Let Ω be a convex polyhedral domain in $\mathbb{R}^3$ and $\alpha = 1, \beta = 1$. Then there exists a constant $h_0 > 0$ such that for $h < h_0$, the discrete problem* (8.15) *has a unique solution $\mathbf{E}_h$. Moreover, assume that the solution $\mathbf{E}$ of* (8.14) *satisfies $\mathbf{E} \in H^1(\Omega)^3, \nabla \times \mathbf{E} \in H^1(\Omega)^3$, then the following error estimate holds*

$$\|\mathbf{E} - \mathbf{E}_h\|_{H(\text{curl};\Omega)} \leqslant Ch\left(|\mathbf{E}|_{H^1(\Omega)} + |\nabla \times \mathbf{E}|_{H^1(\Omega)}\right).$$

Proof. The proof is divided into several steps.

1°) Let $a(\cdot,\cdot) : H_0(\text{curl};\Omega)\times H_0(\text{curl};\Omega)\to\mathbb{R}$ be the bilinear form defined by

$$a(\mathbf{u},\mathbf{v}) = (\nabla\times\mathbf{u},\nabla\times\mathbf{v}) - k^2(\mathbf{u},\mathbf{v}).$$

From (8.14) and (8.15) we know that

$$a(\mathbf{E}-\mathbf{E}_h,\mathbf{v}_h) = 0 \quad \forall \mathbf{v}_h \in \mathbf{X}_h^0. \tag{8.16}$$

Let $P_h\mathbf{E}\in\mathbf{X}_h^0$ be the projection of $\mathbf{E}$ to $\mathbf{X}_h^0$ defined by

$$(\nabla\times P_h\mathbf{E},\nabla\times\mathbf{v}) + (P_h\mathbf{E},\mathbf{v}) = (\nabla\times\mathbf{E},\nabla\times\mathbf{v}) + (\mathbf{E},\mathbf{v}) \quad \forall\mathbf{v}\in\mathbf{X}_h^0.$$

Thus

$$\begin{aligned}\|\,\mathbf{E}-\mathbf{E}_h\,\|_{H(\text{curl};\Omega)}^2 &= a(\mathbf{E}-\mathbf{E}_h,\mathbf{E}-\mathbf{E}_h) + (1+k^2)\|\,\mathbf{E}-\mathbf{E}_h\,\|_{L^2(\Omega)}^2\\ &= a(\mathbf{E}-\mathbf{E}_h,\mathbf{E}-P_h\mathbf{E}) + (1+k^2)\|\,\mathbf{E}-\mathbf{E}_h\,\|_{L^2(\Omega)}^2\\ &= (\nabla\times(\mathbf{E}-\mathbf{E}_h),\nabla\times(\mathbf{E}-P_h\mathbf{E}) + (\mathbf{E}-\mathbf{E}_h,\mathbf{E}-P_h\mathbf{E})\\ &\quad + (1+k^2)(\mathbf{E}-\mathbf{E}_h,P_h\mathbf{E}-\mathbf{E}_h).\end{aligned}$$

This yields

$$\begin{aligned}&\|\,\mathbf{E}-\mathbf{E}_h\,\|_{H(\text{curl};\Omega)}\\ &\leqslant \|\,\mathbf{E}-P_h\mathbf{E}\,\|_{H(\text{curl};\Omega)} + (1+k^2)\sup_{0\neq\mathbf{v}_h\in\mathbf{X}_h^0}\frac{|(\mathbf{E}-\mathbf{E}_h,\mathbf{v}_h)|}{\|\,\mathbf{v}_h\,\|_{H(\text{curl};\Omega)}}.\end{aligned} \tag{8.17}$$

2°) Now we estimate the second term in above estimate. First since $\mathbf{E}-\mathbf{E}_h\in H_0(\text{curl};\Omega)$, there exists a $\mathbf{w}\in H_0(\text{curl};\Omega)$ and $\varphi\in H_0^1(\Omega)$ such that

$$\mathbf{E}-\mathbf{E}_h = \mathbf{w}+\nabla\varphi, \quad \text{div}\mathbf{w} = 0. \tag{8.18}$$

In fact we can define $\varphi\in H_0^1(\Omega)$ as the solution of the following problem

$$(\nabla\varphi,\nabla v) = (\mathbf{E}-\mathbf{E}_h,\nabla v) \quad \forall v\in H_0^1(\Omega),$$

and let $\mathbf{w} = \mathbf{E}-\mathbf{E}_h-\nabla\varphi$. Clearly $\|\,\nabla\varphi\,\|_{L^2(\Omega)} \leqslant \|\,\mathbf{E}-\mathbf{E}_h\,\|_{L^2(\Omega)}$.

For any $\mathbf{v}_h\in\mathbf{X}_h^0$, we use the following decomposition

$$\mathbf{v}_h = \mathbf{w}_h+\nabla\varphi_h, \quad \mathbf{w}_h\in\mathbf{X}_{h,0}^0, \varphi_h\in V_h^0, \tag{8.19}$$

where $V_h^0\subset H_0^1(\Omega)$ is the conforming linear finite element space having zero trace on the boundary and $\mathbf{X}_{h,0}^0$ is the subspace of $\mathbf{X}_h^0$ whose functions are discrete divergence free

$$\mathbf{X}_{h,0}^0 = \{\mathbf{u}_h\in\mathbf{X}_h^0 : (\mathbf{u}_h,\nabla v_h) = 0 \;\; \forall v_h\in V_h^0\}.$$

In fact we can construct $\varphi_h \in V_h^0$ as the solution of the following discrete problem

$$(\nabla\varphi_h, \nabla v_h) = (\mathbf{v}_h, \nabla v_h) \qquad \forall v_h \in V_h^0,$$

and let $\mathbf{w}_h = \mathbf{v}_h - \nabla\varphi_h$. Clearly $\|\nabla\varphi_h\|_{L^2(\Omega)} \leqslant \|\mathbf{v}_h\|_{L^2(\Omega)}$ and thus $\|\mathbf{w}_h\|_{L^2(\Omega)} \leqslant C\|\mathbf{v}_h\|_{L^2(\Omega)}$.

Since $\mathbf{E}-\mathbf{E}_h$ is discrete divergence free by (8.16), we have by (8.18)-(8.19) that

$$(\mathbf{E}-\mathbf{E}_h, \mathbf{v}_h) = (\mathbf{E}-\mathbf{E}_h, \mathbf{w}_h) = (\mathbf{w}, \mathbf{w}_h) + (\nabla\varphi, \mathbf{w}_h). \tag{8.20}$$

3°) We will use the duality argument to estimate $\|\mathbf{w}\|_{L^2(\Omega)}$. Let $\mathbf{z} \in H_0(\mathrm{curl};\Omega)$ be the solution of the following problem

$$\begin{aligned} \nabla\times\nabla\times\mathbf{z} - k^2\mathbf{z} &= \mathbf{w} \quad \text{in } \Omega, \\ \mathbf{z}\times\mathbf{n} &= 0 \quad \text{on } \Gamma. \end{aligned}$$

By the assumption that k^2 is not the eigenvalue of the Dirichlet problem for the Maxwell system, we know that $\mathbf{z} \in H_0(\mathrm{curl};\Omega)$ is well-defined and $\|\mathbf{z}\|_{H(\mathrm{curl};\Omega)} \leqslant C\|\mathbf{w}\|_{L^2(\Omega)}$. Moreover, $\operatorname{div}\mathbf{z} = 0$ in Ω as the consequence of $\operatorname{div}\mathbf{w} = 0$ in Ω. Thus $\mathbf{z} \in \mathbf{X}_N(\Omega)$. Since Ω is convex, by Theorem 8.5, we have $\mathbf{z} \in H^1(\Omega)^3$ and $\|\mathbf{z}\|_{H^1(\Omega)} \leqslant C\|\mathbf{w}\|_{L^2(\Omega)}$. Similarly, noting that

$$\int_\Gamma \nabla\times\mathbf{z}\cdot\mathbf{n}\,\varphi = \int_\Omega \nabla\times\mathbf{z}\cdot\nabla\varphi = \int_\Gamma \mathbf{n}\times\mathbf{z}\cdot\varphi = 0, \quad \forall\varphi \in H^1(\Omega),$$

we have $\nabla\times\mathbf{z} \in \mathbf{X}_T(\Omega)$, and hence by Theorem 8.5, $\nabla\times\mathbf{z} \in H^1(\Omega)^3$ and $\|\nabla\times\mathbf{z}\|_{H^1(\Omega)} \leqslant C\|\mathbf{w}\|_{L^2(\Omega)}$.

Now, since $\mathbf{z}$ is divergence free, by (8.16),

$$\|\mathbf{w}\|_{L^2(\Omega)}^2 = a(\mathbf{w}, \mathbf{z}) = a(\mathbf{E}-\mathbf{E}_h, \mathbf{z}) = a(\mathbf{E}-\mathbf{E}_h, \mathbf{z}-\gamma_h\mathbf{z}),$$

which implies, by Theorem 8.9,

$$\|\mathbf{w}\|_{L^2(\Omega)}^2 \leqslant Ch\|\mathbf{w}\|_{L^2(\Omega)}\|\mathbf{E}-\mathbf{E}_h\|_{H(\mathrm{curl};\Omega)}.$$

Therefore

$$\begin{aligned} |(\mathbf{w}, \mathbf{w}_h)| &\leqslant Ch\|\mathbf{E}-\mathbf{E}_h\|_{H(\mathrm{curl};\Omega)}\|\mathbf{w}_h\|_{L^2(\Omega)} \\ &\leqslant Ch\|\mathbf{E}-\mathbf{E}_h\|_{H(\mathrm{curl};\Omega)}\|\mathbf{v}_h\|_{L^2(\Omega)}. \end{aligned} \tag{8.21}$$

4°) To estimate $(\nabla\varphi, \mathbf{w}_h)$, we define $\mathbf{v} \in H_0(\mathrm{curl};\Omega)$ such that

$$\nabla\times\mathbf{v} = \nabla\times\mathbf{w}_h, \qquad \operatorname{div}\mathbf{v} = 0.$$

Note that $\mathbf{v}$ is the divergence free part in the Helmholtz decomposition of $\mathbf{v}_h$. Since Ω is convex, we have $\mathbf{v} \in H^1(\Omega)^3$ and $\|\mathbf{v}\|_{H^1(\Omega)} \leqslant C\|\nabla\times\mathbf{w}_h\|_{L^2(\Omega)}$. On the other hand, since $\nabla\times\mathbf{v} = \nabla\times\mathbf{w}_h \in L^p(\Omega)$ for any $p > 2$, we know from Lemma 8.5 that $\gamma_h\mathbf{v}$ is well-defined. By using the operator τ_h in Exercise 8.3, we have

$$\nabla\times\gamma_h\mathbf{v}=\tau_h\nabla\times\mathbf{v}=\tau_h\nabla\times\mathbf{w}_h=\nabla\times\mathbf{w}_h.$$

Thus $\gamma_n\mathbf{v}-\mathbf{w}_h=\nabla\psi_h$ for some $\psi_h\in H_0^1(\Omega)$. Since $\gamma_n\mathbf{v}-\mathbf{w}_h\in\mathcal{P}(K)$, $\psi_h\in V_h^0$. Now since div$\mathbf{v}=0$ and $\mathbf{w}_h$ is discrete divergence free,

$$\begin{aligned}\|\,\mathbf{w}_h-\mathbf{v}\,\|_{L^2(\Omega)}^2&=(\mathbf{w}_h-\mathbf{v},\mathbf{w}_h-\gamma_h\mathbf{v})+(\mathbf{w}_h-\mathbf{v},\gamma_h\mathbf{v}-\mathbf{v})\\&=(\mathbf{w}_h-\mathbf{v},\gamma_h\mathbf{v}-\mathbf{v})\leqslant\|\,\mathbf{w}_h-\mathbf{v}\,\|_{L^2(\Omega)}\|\,\gamma_h\mathbf{v}-\mathbf{v}\,\|_{L^2(\Omega)},\end{aligned}$$

which implies

$$\|\,\mathbf{w}_h-\mathbf{v}\,\|_{L^2(\Omega)}\leqslant\|\,\gamma_h\mathbf{v}-\mathbf{v}\,\|_{L^2(\Omega)}.$$

By using Lemma 8.5 we can prove as in Theorem 8.9 that

$$\|\,\gamma_h\mathbf{v}-\mathbf{v}\,\|_{L^2(\Omega)}\leqslant Ch\left(|\mathbf{v}|_{H^1(\Omega)}+\|\nabla\times\mathbf{v}\|_{L^2(\Omega)}\right)\leqslant Ch\|\,\nabla\times\mathbf{w}_h\,\|_{L^2(\Omega)}.$$

Therefore, since $\operatorname{div}\mathbf{v}=0$,

$$\begin{aligned}|(\nabla\varphi,\mathbf{w}_h)|=|(\nabla\varphi,\mathbf{w}_h-\mathbf{v})|&\leqslant Ch\|\,\nabla\times\mathbf{w}_h\,\|_{L^2(\Omega)}\|\,\nabla\varphi\,\|_{L^2(\Omega)}\\&=Ch\|\,\nabla\times\mathbf{v}_h\,\|_{L^2(\Omega)}\|\,\nabla\varphi\,\|_{L^2(\Omega)}.\end{aligned}\tag{8.22}$$

5°) Combining (8.21)-(8.22) with (8.20) we obtain

$$|(\mathbf{E}-\mathbf{E}_h,\mathbf{v}_h)|\leqslant Ch\|\,\mathbf{E}-\mathbf{E}_h\,\|_{H(\mathrm{curl};\Omega)}\|\,\mathbf{v}_h\,\|_{H(\mathrm{curl};\Omega)}.$$

Substitute it into (8.17) we know that for sufficiently small h, $\mathbf{E}_h$ is uniquely existent, and by Theorem 8.9

$$\|\,\mathbf{E}-\mathbf{E}_h\,\|_{H(\mathrm{curl};\Omega)}\leqslant Ch\left(|\mathbf{E}|_{H^1(\Omega)}+|\nabla\times\mathbf{E}|_{H^1(\Omega)}\right).$$

This completes the proof. □

8.4. A posteriori error analysis

In this section we consider the a posteriori error estimate for the time-harmonic Maxwell equation with homogeneous Dirichlet boundary condition (8.14). We start with the following theorem on the interpolation of non-smooth functions [**8**].

THEOREM 8.11. *There exists a linear projection* $\Pi_h : H^1(\Omega)^3 \cap H_0(\mathrm{curl};\Omega) \mapsto \mathbf{X}_h^0$ *such that for all* $\mathbf{v} \in H^1(\Omega)^3$

$$\begin{aligned}
&\|\Pi_h \mathbf{v}\|_{L^2(K)} \leqslant C\big(\|\mathbf{v}\|_{L^2(\widetilde{K})} + h_K |\mathbf{v}|_{H^1(\widetilde{K})}\big) \quad \forall K \in \mathcal{M}_h,\\
&\|\nabla \times \Pi_h \mathbf{v}\|_{L^2(K)} \leqslant C|\mathbf{v}|_{H^1(\widetilde{K})} \quad \forall K \in \mathcal{M}_h,\\
&\|\mathbf{v} - \Pi_h \mathbf{v}\|_{L^2(K)} \leqslant Ch_K|\mathbf{v}|_{H^1(\widetilde{K})} \quad \forall K \in \mathcal{M}_h\\
&\|\mathbf{v} - \Pi_h \mathbf{v}\|_{L^2(F)} \leqslant Ch_F^{1/2}|\mathbf{v}|_{H^1(\widetilde{F})} \quad \forall \textit{ face } F \in \mathcal{F}_h,
\end{aligned}$$

where $\mathcal{F}_h$ *is the set of all interior faces of the mesh* $\mathcal{M}_h$*,* $\widetilde{K}$ *and* $\widetilde{F}$ *are the union of the elements in* $\mathcal{M}_h$ *having having nonempty intersection with* K *and* F*, respectively.*

PROOF. For any edge $e \in \mathcal{E}_h$, let $\mathbf{w}_e \in \mathbf{X}_h$ be the associated canonical basis function of $\mathbf{X}_h$, that is, $\{\mathbf{w}_e\}_{e\in\mathcal{E}_h}$ be the basis of $\mathbf{X}_h$ satisfying

$$\int_e \mathbf{w}_e \cdot \mathbf{t}_e \,\mathrm{d}l = 1, \qquad \int_{e'} \mathbf{w}_e \cdot \mathbf{t}_{e'} \,\mathrm{d}l = 0 \qquad \forall e, e' \in \mathcal{E}_h, \quad e' \neq e.$$

On each face $F \in \mathcal{F}_h$ with edges $\{e_1, e_2, e_3\}$, we construct a dual basis $\{\mathbf{q}_j\}$ of $\{\mathbf{w}_i \times \mathbf{n}\}$ as follows

$$\int_F (\mathbf{w}_i \times \mathbf{n}) \cdot \mathbf{q}_j \,\mathrm{d}s = \delta_{ij}, \qquad i, j = 1, 2, 3. \tag{8.23}$$

We claim that

$$\|\mathbf{q}_i\|_{L^\infty(F)} \leqslant Ch_F^{-1} \tag{8.24}$$

which implies that $\|\mathbf{q}_i\|_{L^2(F)} \leqslant C$. Without loss of generality, we will prove that (8.24) holds for $i = 1$. We first find $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^T$ such that

$$\mathbf{q}_1 = \alpha_1 \mathbf{w}_1 \times \mathbf{n} + \alpha_2 \mathbf{w}_2 \times \mathbf{n} + \alpha_3 \mathbf{w}_3 \times \mathbf{n}, \qquad \int_F (\mathbf{w}_i \times \mathbf{n}) \cdot \mathbf{q}_1 \,\mathrm{d}s = \delta_{i1}, \quad i = 1, 2, 3.$$

It is clear that $\boldsymbol{\alpha}$ is the solution of the linear system

$$A_F \boldsymbol{\alpha} = (1, 0, 0)^T, \qquad \text{where} \quad A_F = \Big(\int_F (\mathbf{w}_i \times \mathbf{n}) \cdot (\mathbf{w}_j \times \mathbf{n}) \,\mathrm{d}s \Big)_{3\times 3}. \tag{8.25}$$

We will show that A_F is invertible. Let F be the face F_{123} of a tetrahedron K with vertices $A_i, i = 1, 2, 3, 4$ and let e_1, e_2, e_3 be the edges A_2A_3, A_3A_1, A_1A_2, respectively. From Lemma 8.3,

$$\mathbf{w}_1 = \lambda_2 \nabla \lambda_3 - \lambda_3 \nabla \lambda_2, \quad \mathbf{w}_2 = \lambda_3 \nabla \lambda_1 - \lambda_1 \nabla \lambda_3, \quad \mathbf{w}_3 = \lambda_1 \nabla \lambda_2 - \lambda_2 \nabla \lambda_1.$$

Let $b_{ij} = (\nabla\lambda_i \times \mathbf{n}) \cdot (\nabla\lambda_j \times \mathbf{n})$. Since $\sum_{i=1}^{4} \nabla\lambda_i = 0$ and $\nabla\lambda_4$ is perpendicular to the face F_{123}, we have

$$\sum_{j=1}^{3} b_{ij} = 0 \quad \text{and} \quad b_{ij} = b_{ji}.$$

Therefore, A_F can be rewritten as

$$A_F = \frac{|F|}{12} \begin{pmatrix} 3b_{22} + 3b_{33} - b_{11} & -3b_{33} + b_{11} + b_{22} & -3b_{22} + b_{33} + b_{11} \\ -3b_{33} + b_{11} + b_{22} & 3b_{11} + 3b_{33} - b_{22} & -3b_{11} + b_{22} + b_{33} \\ -3b_{22} + b_{33} + b_{11} & -3b_{11} + b_{22} + b_{33} & 3b_{11} + 3b_{22} - b_{33} \end{pmatrix}.$$

It follows from $\nabla\lambda_1 \perp F_{234}$ that

$$|\nabla\lambda_1| = 1/\text{the height of } K \text{ to the face } F_{234},$$

which implies that

$$b_{11} = |\nabla\lambda_1 \times \mathbf{n}|^2 = \frac{|e_1|^2}{4\,|F|^2}.$$

Similarly,

$$b_{22} = \frac{|e_2|^2}{4\,|F|^2}, \quad b_{33} = \frac{|e_3|^2}{4\,|F|^2}.$$

Straightforward computation shows that

$$\det A_F = \frac{|e_1|^2 + |e_2|^2 + |e_3|^2}{576\,|F|} \geqslant c_0,$$

where c_0 is a positive constant that depends only on the minimum angle of the elements in the mesh. Thus A_F is invertible. Since $A_F = O(1)$, we have $A_F^{-1} = O(1)$ which implies $\boldsymbol{\alpha} = A_F^{-1}(1,0,0)^T = O(1)$, that is, (8.24) holds.

Now for each $e \in \mathcal{E}_h$, we assign one of those faces with edge e and call it $F_e \in \mathcal{F}_h$. We have to comply with the restriction that for e on the boundary, F_e also on the boundary. Then we can define

$$\Pi_h \mathbf{v} = \sum_{e \in \mathcal{E}_h} \Big(\int_{F_e} (\mathbf{v} \times \mathbf{n}) \cdot \mathbf{q}_e^{F_e} \,\mathrm{d}s \Big) \mathbf{w}_e.$$

By virtue of (8.23) this defines a projection. Obviously the boundary condition is respected. By (8.24)

$$\left| \int_{F_e} (\mathbf{v} \times \mathbf{n}) \cdot \mathbf{q}_e^{F_e} \,\mathrm{d}\sigma \right| \leqslant \|\mathbf{v}\|_{L^2(F_e)} \left\|\mathbf{q}_e^{F_e}\right\|_{L^2(F_e)} \leqslant C\, \|\mathbf{v}\|_{L^2(F_e)} \,.$$

Let $K_e \in \mathcal{M}_h$ be an element with F_e as one of its faces. By the scaled trace inequality, we have

$$\|\mathbf{v}\|_{L^2(F_e)}^2 \leqslant C\big(h_e^{-1}\,\|\mathbf{v}\|_{L^2(K_e)}^2 + h_e\,|\mathbf{v}|_{H^1(K_e)}^2\big).$$

Therefore,

$$\begin{aligned}\|\Pi_h \mathbf{v}\|_{L^2(K)}^2 &\leqslant C h_K \sum_{\substack{e\in\mathcal{E}_h\\ e\subset\partial K}} \left|\int_{F_e} (\mathbf{v}\times\mathbf{n})\mathbf{q}_e^{F_e}\,\mathrm{d}\sigma\right|^2 \\ &\leqslant C h_K \sum_{\substack{e\in\mathcal{E}_h\\ e\subset\partial K}} \big(h_e^{-1}\,\|\mathbf{v}\|_{L^2(K_e)}^2 + h_e\,|\mathbf{v}|_{H^1(K_e)}^2\big) \\ &\leqslant C\big(\,\|\mathbf{v}\|_{L^2(\tilde{K})}^2 + h_K^2\,|\mathbf{v}|_{H^1(\tilde{K})}^2\big).\end{aligned}$$

This proves the first estimate in the theorem.

Since Π_h is a projection, we know that $\Pi_h \mathbf{c}_K = \mathbf{c}_K$ for any constant $\mathbf{c}_K$. Thus

$$\begin{aligned}\|\mathbf{v} - \Pi_h\mathbf{v}\|_{L^2(K)} &= \inf_{\mathbf{c}_K} \|(\mathbf{v}+\mathbf{c}_K) - \Pi_h(\mathbf{v}+\mathbf{c}_K)\|_{L^2(K)} \\ &\leqslant C\inf_{\mathbf{c}_K}\big(\,\|\mathbf{v}+\mathbf{c}_K\|_{L^2(\tilde{K})} + h_K\,|\mathbf{v}+c_K|_{H^1(\tilde{K})}\big) \\ &\leqslant C h_K\,|\mathbf{v}|_{H^1(\tilde{K})}\,,\end{aligned}$$

where we have used the scaling argument and Theorem 3.1 in the last inequality. This proves the third inequality. The last inequality can be proved similarly. The proof of the second inequality is left as an exercise. □

The following regular decomposition theorem is due to Birman-Solomyak.

Lemma 8.6. *Let Ω be a bounded Lipschitz domain. Then for any $\mathbf{v} \in H_0(\mathrm{curl};\Omega)$, there exists a $\psi \in H_0^1(\Omega)$ and a $\mathbf{v}_s \in H^1(\Omega)^3 \cap H_0(\mathrm{curl};\Omega)$ such that $\mathbf{v} = \nabla\psi + \mathbf{v}_s$ in Ω, and*

$$\|\psi\|_{H^1(\Omega)} + \|\mathbf{v}_s\|_{H^1(\Omega)} \leqslant C\,\|\mathbf{v}\|_{H(\mathrm{curl};\Omega)}\,,$$

where the constant C depends only on Ω.

Proof. Let O be a ball containing Ω. We extend $\mathbf{v}$ by zero to the exterior of Ω and denote the extension by $\widetilde{\mathbf{v}}$. Clearly $\widetilde{\mathbf{v}} \in H_0(\mathrm{curl};O)$ with compact support in O. By Theorem 8.4 there exists a $\mathbf{w} \in H^1(O)^3$ such that

$$\nabla\times\mathbf{w} = \nabla\times\widetilde{\mathbf{v}},\quad \operatorname{div}\mathbf{w} = 0 \ \text{ in } O, \tag{8.26}$$

and

$$\|\mathbf{w}\|_{H^1(O)} \leqslant C\,\|\nabla\times\widetilde{\mathbf{v}}\|_{L^2(O)} = C\,\|\nabla\times\mathbf{v}\|_{L^2(\Omega)}\,. \tag{8.27}$$

Now since O is simply-connected, $\nabla\times(\mathbf{w}-\widetilde{\mathbf{v}})=0$, by Theorem 8.3, there exists a $\varphi\in H^1(O)/\mathbb{R}$ such that $\tilde{\mathbf{v}}=\mathbf{w}+\nabla\varphi$ in O, and, from (8.27),

$$\begin{aligned}&\|\varphi\|_{H^1(O)}\leqslant C\,|\varphi|_{H^1(O)}\leqslant C\big(\,\|\widetilde{\mathbf{v}}\|_{L^2(O)}+\|\mathbf{w}\|_{L^2(O)}\,\big)\leqslant C\,\|\mathbf{v}\|_{H(\mathrm{curl};\Omega)}\,,\\&|\varphi|_{H^2(O\setminus\bar{\Omega})}\leqslant|\mathbf{w}|_{H^1(O)}\leqslant C\,\|\nabla\times\mathbf{v}\|_{L^2(\Omega)}\,.\end{aligned}$$

Since $O\setminus\bar{\Omega}$ is a Lipschitz domain, by the extension theorem of Nečas, there exists an extension of $\varphi|_{O\setminus\bar{\Omega}}$, denoted by $\widetilde{\varphi}\in H^2(\mathbb{R}^3)$, such that

$$\widetilde{\varphi}=\varphi\ \text{ in }\ O\setminus\bar{\Omega},\qquad\|\widetilde{\varphi}\|_{H^2(\mathbb{R}^3)}\leqslant C\,\|\varphi\|_{H^2(O\setminus\bar{\Omega})}\leqslant C\,\|\mathbf{v}\|_{H(\mathrm{curl};\Omega)}\,.$$

This completes the proof by letting $\psi=\varphi-\widetilde{\varphi}\in H_0^1(\Omega)$ and $\mathbf{v}_s=\mathbf{w}+\nabla\widetilde{\varphi}$. Remember that $\widetilde{\mathbf{v}}=\mathbf{v}_s+\nabla\psi$ in O and $\mathbf{v}_s=\widetilde{\mathbf{v}}=0$ in $O\setminus\bar{\Omega}$. Thus $\mathbf{v}_s\in H_0^1(\Omega)^3$. □

THEOREM 8.12. *Let* $\mathbf{E}\in H_0(\mathrm{curl};\Omega)$ *and* $\mathbf{E}_h\in\mathbf{X}_h^0$ *be respectively the solutions of* (8.14) *and* (8.15). *We have the following a posteriori error estimate*

$$\|\mathbf{E}-\mathbf{E}_h\|_{H(\mathrm{curl};\Omega)}\leqslant C\left(\sum_{K\in\mathcal{M}_h}\eta_K^2\right)^{1/2},$$

where

$$\begin{aligned}\eta_K^2=&h_K^2\left\|\mathbf{f}-\nabla\times(\alpha\nabla\times\mathbf{E}_h)+k^2\beta\mathbf{E}_h\right\|_{L^2(K)}^2+h_K^2\left\|\mathrm{div}\,(\mathbf{f}+k^2\beta\mathbf{E}_h)\right\|_{L^2(K)}^2\\&+\sum_{F\subset\partial K}\Big(h_F\,\|[\![\mathbf{n}\times(\alpha\nabla\times\mathbf{E}_h)]\!]\|_{L^2(F)}^2+h_F\left\|[\![(\mathbf{f}+k^2\beta\mathbf{E}_h)\cdot\mathbf{n}]\!]\right\|_{L^2(F)}^2\Big).\end{aligned}$$

Here $[\![\cdot]\!]$ *denotes the jump across the interior face* F.

PROOF. Let $a(\cdot,\cdot):H_0(\mathrm{curl};\Omega)\times H_0(\mathrm{curl};\Omega)\to\mathbb{R}$ be the bilinear form defined by

$$a(\mathbf{u},\mathbf{v})=(\alpha\nabla\times\mathbf{u},\nabla\times\mathbf{v})-k^2(\beta\mathbf{u},\mathbf{v}).$$

From (8.14) and (8.15) we know that

$$a(\mathbf{E}-\mathbf{E}_h,\mathbf{v}_h)=0\quad\forall\mathbf{v}_h\in\mathbf{X}_h^0.\tag{8.28}$$

The unique existence of the weak solution of the problem (8.14) implies that there exists a constant $C>0$ such that

$$\inf_{0\neq\mathbf{u}\in H_0(\mathrm{curl};\Omega)}\ \sup_{0\neq\mathbf{v}\in H_0(\mathrm{curl};\Omega)}\frac{|a(\mathbf{u},\mathbf{v})|}{\|\mathbf{u}\|_{H(\mathrm{curl};\Omega)}\|\mathbf{v}\|_{H(\mathrm{curl},\Omega)}}\geqslant C.\tag{8.29}$$

For any $\mathbf{v}\in H_0(\mathrm{curl};\Omega)$, by Lemma 8.6, there exists a $\psi\in H_0^1(\Omega)$ and a $\mathbf{v}_s\in H^1(\Omega)^3\cap H_0(\mathrm{curl};\Omega)$ such that $\mathbf{v}=\nabla\psi+\mathbf{v}_s$ in Ω, and

$$\|\psi\|_{H^1(\Omega)}+\|\mathbf{v}_s\|_{H^1(\Omega)}\leqslant C\,\|\mathbf{v}\|_{H(\mathrm{curl};\Omega)}\,.\tag{8.30}$$

Let $r_h : H^1(\Omega) \mapsto V_h^0$ be the Clément interpolant defined in Chapter 4, Section 4.2.1, and define

$$\mathbf{v}_h = \nabla r_h \psi + \Pi_h \mathbf{v}_s \in \mathbf{X}_h^0.$$

Then by the inf-sup condition (8.29) and (8.28)

$$\begin{aligned}\|\mathbf{E}-\mathbf{E}_h\|_{H(\mathrm{curl};\Omega)} &\leqslant C \sup_{0\neq \mathbf{v}\in H_0(\mathrm{curl},\Omega)} \frac{|a(\mathbf{E}-\mathbf{E}_h,\mathbf{v})|}{\|\mathbf{v}\|_{H(\mathrm{curl},\Omega)}} \\ &= C \sup_{0\neq \mathbf{v}\in H_0(\mathrm{curl},\Omega)} \frac{|a(\mathbf{E}-\mathbf{E}_h,\mathbf{v}-\mathbf{v}_h)|}{\|\mathbf{v}\|_{H(\mathrm{curl},\Omega)}}.\end{aligned}$$

On the other hand, by integrating by parts, we have

$$\begin{aligned}&a(\mathbf{E}-\mathbf{E}_h,\mathbf{v}-\mathbf{v}_h)\\ &= (\mathbf{f},\mathbf{v}-\mathbf{v}_h) - (\alpha\,\nabla\times\mathbf{E}_h, \nabla\times(\mathbf{v}-\mathbf{v}_h)) + k^2\big(\beta\mathbf{E}_h,\mathbf{v}-\mathbf{v}_h\big)\\ &= \big(\mathbf{f},(\nabla\psi+\mathbf{v}_s)-(\nabla r_h\psi+\Pi_h\mathbf{v}_s)\big) - \big(\alpha\,\nabla\times\mathbf{E}_h,\nabla\times(\mathbf{v}_s-\Pi_h\mathbf{v}_s)\big)\\ &\quad + k^2\big(\beta\mathbf{E}_h,\mathbf{v}_s-\Pi_h\mathbf{v}_s\big) + k^2\big(\beta\mathbf{E}_h,\nabla(\psi-r_h\psi)\big)\\ &= \sum_{K\in\mathcal{M}_h}\big(\mathbf{f}-\nabla\times(\alpha\,\nabla\times\mathbf{E}_h)+k^2\beta\mathbf{E}_h,\mathbf{v}_s-\Pi_h\mathbf{v}_s\big)\\ &\quad + \sum_{F\in\mathcal{F}_h}\int_F [\![\mathbf{n}\times(\alpha\nabla\times\mathbf{E}_h)]\!]\cdot(\mathbf{v}_s-\Pi_h\mathbf{v}_s)\\ &\quad - \sum_{K\in\mathcal{M}_h}\big(\mathrm{div}\,(\mathbf{f}+k^2\beta\mathbf{E}_h),\psi-r_h\psi\big)\\ &\quad + \sum_{F\in\mathcal{F}_h}\int_F [\![(\mathbf{f}+k^2\beta\mathbf{E}_h)\cdot\mathbf{n}]\!](\psi-r_h\psi).\end{aligned}$$

Now by Theorem 8.11 and (8.30)

$$\begin{aligned}|a(\mathbf{E}-\mathbf{E}_h,\mathbf{v}-\mathbf{v}_h)| &\leqslant C\Big(\sum_{K\subset\mathcal{M}_h}\eta_K^2\Big)^{1/2}\big(\,\|\mathbf{v}_s\|_{H^1(\Omega)}+|\psi|_{H^1(\Omega)}\,\big)\\ &\leqslant C\Big(\sum_{k\subset\mathcal{M}_h}\eta_K^2\Big)^{1/2}\|\mathbf{v}\|_{H(\mathrm{curl};\Omega)}\,.\end{aligned}$$

This completes the proof. □

Bibliographic notes. The results in Section 8.1 are taken from Girault and Raviart [**34**]. Further results on vector potentials on nonsmooth domains can be found in Amrouche et al [**2**]. The full characterization of the trace for functions in $H(\mathrm{curl};\Omega)$ can be found in Buffa et al [**17**]. The Nédélec edge elements are introduced in Nédédec [**43, 44**]. Lemma 8.5 is taken from [**2**].

Further properties of edge elements can be found in Hiptmair [**36**] and Monk [**42**]. The error analysis in Section 8.3 follows the development in [**42**] which we refer to for further results. The interpolation operator in Theorem 8.11 is from Beck et al [**8**] although the proof here is slightly different. In [**8**] the a posteriori error estimate is derived for smooth or convex domains. Lemma 8.6, which is also known as regular decomposition theorem, is from Birman and Solomyak [**10**]. Theorem 8.12 is from Chen et al [**21**] in which the adaptive multilevel edge element method for time-harmonic Maxwell equations based on a posteriori error estimate is also considered.

8.5. Exercises

Exercise 8.1. Prove $\mathcal{D}(\bar{\Omega})^3$ is dense in $H(\mathrm{div};\Omega)$.

Exercise 8.2. The lowest order divergence conforming finite element is a triple $(K,\mathcal{D},\mathcal{N})$ with the following properties:

(i) $K\subset\mathbb{R}^3$ is a tetrahedron;
(ii) $\mathcal{D}=\{\mathbf{u}=\mathbf{a}_K+b_K\,x\quad\forall\,\mathbf{a}_K\in\mathbb{R}^3,b_K\in\mathbb{R}\}$;
(iii) $\mathcal{N}=\{M_F:M_F(\mathbf{u})=\int_F(\mathbf{u}\cdot\mathbf{n})\,\mathrm{d}s\quad\forall$ face F of K, $\forall\mathbf{u}\in\mathcal{D}\}$.

For any vector field $\mathbf{u}$ defined on K, let

$$\hat{\mathbf{u}}\circ F_K(\hat{x})=B_K\mathbf{u}(B_K\hat{x}+\mathbf{b}_K).$$

Prove that

(i) $\mathbf{u}\in\mathcal{D}(K)\Leftrightarrow\hat{\mathbf{u}}\in\hat{\mathcal{D}}(\hat{K})$;
(ii) $\mathrm{div}\,\mathbf{u}=0\Leftrightarrow\hat{\mathrm{div}}\,\hat{\mathbf{u}}=0\quad\forall\;\mathbf{u}\in\mathcal{D}(K)$;
(iii) $M_F(\mathbf{u})=0\Leftrightarrow M_{\hat{F}}(\hat{\mathbf{u}})=0\quad\forall\;\mathbf{u}\in\mathcal{D}(K)$;
(iv) If $\mathbf{u}\in\mathcal{D}(K)$ and $M_F(\mathbf{u})=0$, then $\mathbf{u}\times\mathbf{n}=0$ on F;
(v) If $\mathbf{u}\in\mathcal{D}(K)$ and $M_F(\mathbf{u})=0$ for any face F, then $\mathbf{u}=0$ in K.

Exercise 8.3. For any function $\mathbf{u}$ defined on K such that $M_F(\mathbf{u})$ is defined on each face F of K. Let $\tau_K\mathbf{u}$ be the unique polynomial in $\mathcal{D}(K)$ that has the same moments as $\mathbf{u}$ on K: $M_F(\tau_K\mathbf{u}-\mathbf{u})=0$. Prove that for any $p>2$, τ_K is continuous on the space

$$\{\mathbf{u}\in L^p(K)^3:\mathrm{div}\,\mathbf{u}\in L^2(K)\}.$$

Exercise 8.4. Let $\mathcal{D}(K)$ be the finite element space in Exercise 8.2 and

$$\mathbf{W}_h=\{\mathbf{u}_h\in H(\mathrm{div};\Omega):\mathbf{u}_h|_K\in\mathcal{D}(K)\ \ \forall K\in\mathcal{M}_h\}.$$

Let τ_h be the global interpolation operator

$$\tau_h\mathbf{u}|_K=\tau_K\mathbf{u}\ \text{ on } K,\ \ \forall K\in\mathcal{M}_h.$$

Prove that $\nabla \times \mathbf{X}_h \subset \mathbf{W}_h$ and $\nabla \times \gamma_h \mathbf{u} = \tau_h \nabla \times \mathbf{u}$.

Exercise 8.5. Prove that

$$\begin{aligned}
&\|\mathbf{u} - \tau_h \mathbf{u}\|_{L^2(\Omega)} \leqslant Ch\|\mathbf{u}\|_{H^1(\Omega)} \qquad \forall \mathbf{u} \in H^1(\Omega)^3,\\
&\|\mathrm{div}\,(\mathbf{u} - \tau_h \mathbf{u})\|_{L^2(\Omega)} \leqslant Ch\|\mathrm{div}\,\mathbf{u}\|_{H^1(\Omega)} \qquad \forall \mathbf{u} \in H^1(\Omega)^3, \mathrm{div}\,\mathbf{u} \in H^1(\Omega).
\end{aligned}$$

Exercise 8.6. Prove the second estimate in Theorem 8.11.

CHAPTER 9

Multiscale Finite Element Methods for Elliptic Equations

In this chapter we consider finite element methods for solving the following elliptic equation with oscillating coefficients

$$\begin{aligned} -\nabla \cdot (a(x/\varepsilon)\nabla u) &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{9.1}$$

where $\Omega \subset \mathbb{R}^2$ is a bounded Lipschitz domain and $f \in L^2(\Omega)$. We assume $a(x/\varepsilon) = (a_{ij}(x/\varepsilon))$ is a symmetric matrix and $a_{ij}(y)$ are $W^{1,p}(p > 2)$ periodic functions in y with respect to a unit cube Y. We assume $a_\varepsilon = a(x/\varepsilon)$ is elliptic, that is, there exists a constant $\gamma > 0$ such that

$$a_{ij}(y)\xi_i\xi_j \geqslant \gamma|\xi|^2 \quad \forall \xi \in \mathbb{R}^2, \ \text{ for a.e. } y \in Y.$$

Here and throughout this chapter, the Einstein convention for repeated indices are assumed.

The problem (9.1) a model multiscale problem which arises in the modeling of composite materials and the flow transport in heterogeneous porous media. The main difficulty in solving it by standard finite element method is that when ε is small, the underlying finite element mesh h must be much less than ε which makes the computational costs prohibitive. The multiscale finite element method allows to solve the problem with mesh size h greater than ε.

9.1. The homogenization result

In this section we introduce the homogenization result that will be used in the subsequent analysis. We start with using the method of asymptotic expansion to derive the homogenized equation for (9.1). Assume that the solution of (9.1) has the following expansion

$$u(x) = u_0(x, x/\varepsilon) + \varepsilon u_1(x, x/\varepsilon) + \varepsilon^2 u_2(x, x/\varepsilon) + o(\varepsilon^2),$$

where $u_i(x,y)$ is periodic in Y with respect to the second variable y. By $\nabla = \varepsilon^{-1}\nabla_y + \nabla_x$ we know that

$$\nabla u = \varepsilon^{-1}\nabla_y u_0 + (\nabla_x u_0 + \nabla_y u_1) + \varepsilon(\nabla_x u_1 + \nabla_y u_2) + o(\varepsilon),$$

and

$$\begin{aligned}
\nabla \cdot (a_\varepsilon \nabla u) &= \frac{\partial}{\partial x_i}\left(a_{ij}(x/\varepsilon)\frac{\partial u}{\partial x_j}\right) \\
&= \varepsilon^{-1}\frac{\partial a_{ij}}{\partial y_i}\cdot\frac{\partial u}{\partial x_j} + a_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j} \\
&= \varepsilon^{-2}\frac{\partial}{\partial y_i}\left(a_{ij}\frac{\partial u_0}{\partial y_j}\right) \\
&\quad + \varepsilon^{-1}\left[\frac{\partial}{\partial y_i}\left(a_{ij}\frac{\partial u_0}{\partial x_j}\right) + \frac{\partial}{\partial y_i}\left(a_{ij}\frac{\partial u_1}{\partial y_j}\right) + a_{ij}\frac{\partial^2 u_0}{\partial y_i \partial x_j}\right] \\
&\quad + \varepsilon^0\left[\frac{\partial}{\partial y_i}\left(a_{ij}\frac{\partial u_2}{\partial y_j}\right) + \frac{\partial}{\partial y_i}\left(a_{ij}\frac{\partial u_1}{\partial x_j}\right) + a_{ij}\frac{\partial^2 u_0}{\partial x_i \partial x_j} + a_{ij}\frac{\partial u_1}{\partial x_i \partial y_j}\right] \\
&\quad + o(1).
\end{aligned}$$

Substitute the above equation into (9.1) and compare the coefficient of ε^{-2} we know that

$$-\frac{\partial}{\partial y_i}\left(a_{ij}\frac{\partial u_0}{\partial y_j}\right) = 0 \quad \text{in } \Omega.$$

By the boundary condition we have $u_0 = 0$ on $\partial\Omega$. Thus we deduce u_0 is independent of y, that is, $u_0(x,y) = u_0(x)$ in Ω.

Now we compare the coefficient of ε^{-1} and obtain

$$\frac{\partial}{\partial y_i}\left(a_{ij}\frac{\partial u_1}{\partial y_j}\right) + \frac{\partial a_{ij}}{\partial y_i}\frac{\partial u_0}{\partial x_j} = 0.$$

If we assume χ^j is the periodic solution of

$$\nabla_y \cdot (a(y)\nabla_y \chi^j) = \frac{\partial}{\partial y_i}a_{ij}(y) \quad \text{in } Y \tag{9.2}$$

with zero mean, i.e., $\int_Y \chi^j dy = 0$, then

$$u_1(x,y) = -\chi^j(y)\frac{\partial u_0}{\partial x_j}. \tag{9.3}$$

Finally we compare the coefficient of ε^0 to get

$$\frac{\partial}{\partial y_i}\left(a_{ij}\frac{\partial u_2}{\partial y_j}\right) + \frac{\partial}{\partial y_i}\left(a_{ij}\frac{\partial u_1}{\partial x_j}\right) + a_{ij}\frac{\partial^2 u_0}{\partial x_i \partial x_j} + a_{ij}\frac{\partial^2 u_1}{\partial x_i \partial y_j} = -f.$$

Integrating the above equation in y over the cell Y and using the periodicity, we obtain the following homogenized equation

$$\begin{aligned} -\nabla\cdot(a^*\nabla u_0) &= f \quad \text{in } \Omega,\\ u_0 &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{9.4}$$

where $a^* = (a^*_{ij})$ is the homogenized coefficient

$$a^*_{ij} = \frac{1}{|Y|}\int_Y a_{ik}(y)(\delta_{kj} - \partial\chi^j/\partial y_k)\,\mathrm{d}y. \tag{9.5}$$

In summary we have the following asymptotic expansion

$$u(x) = u_0(x) + \varepsilon u_1(x, x/\varepsilon) + o(\varepsilon),$$

where u_0 satisfies (9.4) and $u_1(x,y)$ is given by (9.3).

The above argument is heuristic. Our purpose now is to show the convergence of the asymptotic expansion. Let θ_ε denote the boundary corrector which is the solution of

$$\begin{aligned} -\nabla\cdot(a_\varepsilon\nabla\theta_\varepsilon) &= 0 && \text{in } \Omega,\\ \theta_\varepsilon &= u_1(x, x/\varepsilon) && \text{on } \partial\Omega. \end{aligned} \tag{9.6}$$

The variational form of the problem (9.1) is to find $u(x) \in H^1_0(\Omega)$ such that

$$a(u,v) := (a(x/\varepsilon)\nabla u, \nabla v) = (f, v) \quad \forall v \in H^1_0(\Omega). \tag{9.7}$$

Similarly, the variational form of the problem (9.4) is to find $u_0(x) \in H^1_0(\Omega)$ such that

$$(a^*\nabla u_0, \nabla v) = (f, v) \quad \forall v \in H^1_0(\Omega). \tag{9.8}$$

It can be shown that a^* satisfies

$$a^*_{ij}\xi_i\xi_j \geqslant \gamma|\xi|^2 \qquad \forall \xi \in \mathbb{R}^2.$$

Thus by Lax-Milgram lemma, (9.8) has a unique solution.

Lemma 9.1. *Let* $\mathbf{p} \in L^2_{\mathrm{loc}}(\mathbb{R}^d)^d$ $(d \geqslant 1)$ *be* Y*-periodic and* $\operatorname{div}\mathbf{p} = 0$ *in* $\mathbb{R}^d$. *Then there exists a skew-symmetric matrix* $\alpha = (\alpha_{ij}) \in \mathbb{R}^{d\times d}$ *such that* $\alpha_{ij} \in H^1_{\mathrm{loc}}(\mathbb{R}^d)$, α_{ij} *is* Y*-periodic with zero mean, and*

$$p_j = \frac{1}{|Y|}\int_Y p_j(y)\,\mathrm{d}y + \frac{\partial}{\partial y_i}\alpha_{ij}.$$

This lemma extends the classical result in Theorem 8.4 that a divergence free vector must be a curl field. The proof is left as an exercise.

The following theorem plays an important role in our analysis.

THEOREM 9.1. *Assume that* $u_0 \in H^2(\Omega)$. *There exists a constant* C *independent of* u_0, ε, Ω *such that*

$$\|u - u_0 - \varepsilon u_1 + \varepsilon\theta_\varepsilon\|_{H^1(\Omega)} \leqslant C\varepsilon|u_0|_{H^2(\Omega)}$$

PROOF. By simple calculation, we have

$$\begin{aligned}
&a_{ij}(x/\varepsilon)\frac{\partial}{\partial x_j}\left(u_0 - \varepsilon\chi^k\frac{\partial u_0}{\partial x_k}\right)\\
&= a_{ij}(x/\varepsilon)\frac{\partial u_0}{\partial x_j} - \varepsilon a_{ij}(x/\varepsilon)\frac{\partial}{\partial x_j}\Big(\chi^k(x/\varepsilon)\frac{\partial u_0}{\partial x_k}\Big)\\
&= a_{ij}(x/\varepsilon)\frac{\partial u_0}{\partial x_j} - a_{ij}(x/\varepsilon)\frac{\partial\chi^k(y)}{\partial y_j}\frac{\partial u_0}{\partial x_k} - \varepsilon a_{ij}(x/\varepsilon)\chi^k(x/\varepsilon)\frac{\partial^2 u_0}{\partial x_j\partial x_k}\\
&= a^*_{ij}\frac{\partial u_0}{\partial x_j} - G^k_i(x/\varepsilon)\frac{\partial u_0}{\partial x_k} - \varepsilon a_{ij}(x/\varepsilon)\chi^k(x/\varepsilon)\frac{\partial^2 u_0}{\partial x_j\partial x_k},
\end{aligned}$$

where

$$G^k_i = a^*_{ik} - a_{ij}\left(\delta_{kj} - \frac{\partial\chi^k}{\partial y_j}\right).$$

From the definitions of a^*_{ik} and $\chi^k(y)$, it follows that

$$\int_Y G^k_i(y)\,\mathrm{d}y = 0 \quad \text{and} \quad \frac{\partial G^k_i}{\partial y_i} = 0.$$

By Lemma 9.1 there exist skew-symmetric matrices $\alpha^k(x/\varepsilon) = (\alpha^k_{ij}(x/\varepsilon))$ such that

$$G^k_i(y) = \frac{\partial}{\partial y_j}(\alpha^k_{ij}(y)), \qquad \int_Y \alpha^k_{ij}(y)\,\mathrm{d}y = 0.$$

With this notation, we can rewrite

$$G^k_i(x/\varepsilon)\frac{\partial u_0}{\partial x_k} = \varepsilon\frac{\partial}{\partial x_j}\left(\alpha^k_{ij}(x/\varepsilon)\frac{\partial u_0}{\partial x_k}\right) - \varepsilon\alpha^k_{ij}(x/\varepsilon)\frac{\partial^2 u_0}{\partial x_j\partial x_k}.$$

For any $\varphi \in H^1_0(\Omega)$, from (9.6)–(9.8) and (9.3), we have

$$\begin{aligned}
&(a(x/\varepsilon)\nabla(u - u_0 - \varepsilon u_1 + \varepsilon\theta_\varepsilon), \nabla\varphi)\\
&= (a^*\nabla u_0, \nabla\varphi) - \left(a(x/\varepsilon)\nabla\left(u_0 - \varepsilon\chi^k\frac{\partial u_0}{\partial x_k}\right), \nabla\varphi\right)\\
&= \varepsilon\int_\Omega a_{ij}(x/\varepsilon)\chi^k\frac{\partial^2 u_0}{\partial x_j\partial x_k}\frac{\partial\varphi}{\partial x_i}\,\mathrm{d}x - \varepsilon\int_\Omega \alpha^k_{ij}(x/\varepsilon)\frac{\partial^2 u_0}{\partial x_j\partial x_k}\frac{\partial\varphi}{\partial x_i}\,\mathrm{d}x.
\end{aligned}$$

Notice here we have used $\frac{\partial}{\partial x_j}\big(\alpha^k_{ij}(x/\varepsilon)\frac{\partial u_0}{\partial x_k}\big)$ is divergence free. Thus by taking $\varphi = u - u_0 - \varepsilon u_1 + \varepsilon\theta_\varepsilon$ yields the result. □

9.2. The multiscale finite element method

Let Ω be a bounded convex polygonal domain and $\mathcal{M}_h$ be a regular mesh over Ω. We denote by $\{x_j\}_{j=1}^J$ the interior nodes of the mesh $\mathcal{M}_h$ and $\{\psi_j\}_{j=1}^J$ the canonical basis of the H^1-conforming linear finite element space $W_h \subset H^1(\Omega)$. Let $S_i = \operatorname{supp}(\psi_i)$ and define ϕ^i with support in S_i as follows

$$\begin{aligned} -\nabla \cdot (a_\varepsilon(x)\nabla\phi^i) &= 0 \quad \text{in } K \subset S_i, \\ \phi^i &= \psi_i \quad \text{on } \partial K. \end{aligned} \tag{9.9}$$

It is clear that $\phi^i \in H_0^1(S_i) \subset H_0^1(\Omega)$. Introduce the multiscale finite element space

$$V_h = \operatorname{span}\{\phi^i : i = 1, \cdots, J\} \subset H_0^1(\Omega).$$

In the following, we study the approximate solution of (9.4) in V_h, i.e., $u_h \in V_h$ such that

$$a(u_h, v_h) = (f, v_h) \quad \forall v_h \in V_h. \tag{9.10}$$

9.2.1. Error estimate when $h < \varepsilon$. We first introduce the interpolation operator $I_h : C(\bar{\Omega}) \to V_h$

$$I_h u = \sum_{j=1}^J u(x_j)\phi^j(x)$$

and the usual Lagrange interpolation operator $\Pi_h : C(\bar{\Omega}) \to W_h$

$$\Pi_h u = \sum_{j=1}^J u(x_j)\psi_j(x).$$

It is obvious that

$$\begin{aligned} -\nabla \cdot (a(x/\varepsilon)\nabla I_h u) &= 0 \qquad \text{in } K, \\ I_h u &= \Pi_h u \quad \text{on } \partial K. \end{aligned} \tag{9.11}$$

Lemma 9.2. *Let $u \in H^2(\Omega) \cap H_0^1(\Omega)$ be the solution of (9.1). There exists a constant C independent of h, ε such that*

$$\|u - I_h u\|_{L^2(\Omega)} + h\|u - I_h u\|_{H^1(\Omega)} \leqslant Ch^2(|u|_{H^2(\Omega)} + \|f\|_{L^2(\Omega)}).$$

Proof. By Theorem 3.6, we have

$$\|u - \Pi_h u\|_{L^2(\Omega)} + h\|u - \Pi_h u\|_{H^1(\Omega)} \leqslant Ch^2|u|_{H^2(\Omega)}. \tag{9.12}$$

Since $\Pi_h u - I_h u = 0$ on ∂K, by the scaling argument and the Poincaré-Friedrichs inequality we get

$$\|\Pi_h u - I_h u\|_{0,K} \leqslant Ch\|\Pi_h u - I_h u\|_{1,K}. \tag{9.13}$$

From (9.11), it follows that

$$(a(x/\varepsilon)\nabla I_h u, \nabla(I_h u - \Pi_h u))_K = 0.$$

Then

$$\begin{aligned}
&(a(x/\varepsilon)\nabla(I_h u - \Pi_h u), \nabla(I_h u - \Pi_h u))_K \\
&= (a(x/\varepsilon)\nabla(u - \Pi_h u), \nabla(I_h u - \Pi_h u))_K - (a(x/\varepsilon)\nabla u, \nabla(I_h u - \Pi_h u))_K \\
&= (a(x/\varepsilon)\nabla(u - \Pi_h u), \nabla(I_h u - \Pi_h u))_K - (f, \nabla(I_h u - \Pi_h u))_K \\
&\leqslant C|u - \Pi_h u|_{1,K}|I_h u - \Pi_h u|_{1,K} + \|f\|_{0,K}\|I_h u - \Pi_h u\|_{0,K},
\end{aligned}$$

which implies by using (9.13) that

$$|I_h u - \Pi_h u|_{1,K} \leqslant Ch(|u|_{2,K} + \|f\|_{0,K}).$$

This completes the proof. □

THEOREM 9.2. *Let u and u_h be the solutions of (9.1) and (9.10), respectively. Then there exists a constant C, independent of h and ε, such that*

$$\|u - u_h\|_{H^1(\Omega)} \leqslant Ch(|u|_{H^2(\Omega)} + \|f\|_{L^2(\Omega)}). \tag{9.14}$$

Moreover,

$$\|u - u_h\|_{L^2(\Omega)} \leqslant C(h/\varepsilon)^2\|f\|_{L^2(\Omega)}.$$

PROOF. (9.14) follows easily from the Céa Lemma and Lemma 9.2. To show the error estimate in L^2, we use the Aubin-Nitsche trick. By the regularity estimate for the elliptic equation, we know that

$$|u|_{H^2(\Omega)} \leqslant C\varepsilon^{-1}\|f\|_{L^2(\Omega)}.$$

Thus, by (9.14)

$$\|u - u_h\|_{H^1(\Omega)} \leqslant C\,(h/\varepsilon)\|f\|_{L^2(\Omega)}. \tag{9.15}$$

For any $\varphi \in L^2(\Omega)$, let $w \in H_0^1(\Omega) \cap H^2(\Omega)$ be the solution of the dual problem

$$\begin{aligned}
-\nabla\cdot(a(x/\varepsilon)\nabla w) &= \varphi \quad \text{in } \Omega, \\
w &= 0 \quad \text{on } \partial\Omega.
\end{aligned}$$

Then, for any $v_h \in V_h$, we have

$$\begin{aligned}
(u_h - u, \varphi) &= (a(x/\varepsilon)\nabla(u_h - u), \nabla w) \\
&= (a(x/\varepsilon)\nabla(u_h - u), \nabla(w - v_h)) \\
&\leqslant C\|u_h - u\|_{H^1(\Omega)}\|w - v_h\|_{H^1(\Omega)}.
\end{aligned}$$

Hence by (9.15)

$$(u_h - u, \varphi) \leqslant C(h/\varepsilon)\|f\|_{L^2(\Omega)} \inf_{v_h\in V_h} \|w - v_h\|_{H^1(\Omega)}.$$

Choosing v_h as $I_h w$ yields

$$\begin{aligned}(u_h - u, \varphi) &\leqslant C(h/\varepsilon)\|f\|_{L^2(\Omega)}\|w - I_h w\|_{H^1(\Omega)}\\ &\leqslant C(h/\varepsilon)\|f\|_{L^2(\Omega)}(h/\varepsilon)\|\varphi\|_{L^2(\Omega)}.\end{aligned}$$

Hence

$$\|u^h - u\|_{L^2(\Omega)} = \sup_{0\neq\varphi\in L^2(\Omega)} \frac{(u_h - u, \varphi)}{\|\varphi\|_{L^2(\Omega)}} \leqslant C(h/\varepsilon)^2\|f\|_{L^2(\Omega)}.$$

This completes the proof. □

9.2.2. Error estimate when $h > \varepsilon$. Now we consider the error estimate when $h > \varepsilon$ which is the main attraction of the multiscale finite element method.

THEOREM 9.3. *Let u and u_h be the solutions of (9.1) and (9.10), respectively. Then there exists a constant C, independent of h and ε, such that*

$$\|u - u_h\|_{H^1(\Omega)} \leqslant C(h + \varepsilon)\|f\|_{L^2(\Omega)} + C((\varepsilon/h)^{1/2} + \varepsilon^{1/2})\|u_0\|_{W^{1,\infty}(\Omega)}.$$

PROOF. Denote $u_\varepsilon = u_0 + \varepsilon u_1 - \varepsilon\theta_\varepsilon$. Let

$$u_I = I_h u_0 = \sum_{j=1}^{J} u_0(x_j)\phi^j(x).$$

It follows that

$$\begin{aligned}-\nabla\cdot(a_\varepsilon\nabla u_I) &= 0 \qquad \text{in } K,\\ u_I &= \Pi_h u_0 \quad \text{on } \partial K.\end{aligned}$$

Let u_{I0} be the solution of the homogenized problem

$$\begin{aligned}-\nabla\cdot(a^*\nabla u_{I0}) &= 0 \qquad \text{in } K,\\ u_{I0} &= \Pi_h u_0 \quad \text{on } \partial K,\end{aligned} \tag{9.16}$$

and

$$u_{I1} = -\chi^j\frac{\partial u_{I0}}{\partial x_j} \quad \text{in } K. \tag{9.17}$$

Let $\theta_{I\varepsilon}$ be the boundary corrector

$$\begin{aligned}-\nabla\cdot(a(x/\varepsilon)\nabla\theta_{I\varepsilon}) &= 0 \qquad\qquad \text{in } K,\\ \theta_{I\varepsilon} &= u_{I1}(x, x/\varepsilon) \quad \text{on } \partial K.\end{aligned} \tag{9.18}$$

Clearly

$$u_{I0} = \Pi_h u_0 \quad \text{in} \quad K, \tag{9.19}$$

that is, u_{I0} is linear in K which implies $|u_{I0}|_{2,K} = 0$. Thus by following the proof of Theorem 9.1,

$$\|u_I - u_{I0} - \varepsilon u_{I1} + \varepsilon\theta_{I\varepsilon}\|_{H^1(K)} = 0.$$

Again by Theorem 9.1

$$\begin{aligned}\|u - u_I\|_{H^1(\Omega)} \leqslant \|u_0 - u_{I0}\|_{H^1(\Omega)} &+ \varepsilon\|u_1 - u_{I1}\|_{H^1(\Omega)} \\ &+ \|\varepsilon(\theta_\varepsilon - \theta_{I\varepsilon})\|_{H^1(\Omega)} + C\varepsilon|u_0|_{H^2(\Omega)}.\end{aligned} \tag{9.20}$$

It is clear that

$$\|u_0 - u_{I0}\|_{H^1(\Omega)} = \|u_0 - \Pi_h u_0\|_{H^1(\Omega)} \leqslant Ch|u_0|_{H^2(\Omega)}. \tag{9.21}$$

Simple calculation shows that

$$\begin{aligned}\|\varepsilon\nabla(u_1 - u_{I1})\|_{L^2(K)} &= \|\varepsilon\nabla(\chi^j\partial(u_0 - \Pi_h u_0)/\partial x_j\|_{L^2(K)} \\ &\leqslant C\|\nabla(u_0 - \Pi_h u_0)\|_{L^2(K)} + C\varepsilon|u_0|_{H^2(K)} \\ &\leqslant C(h+\varepsilon)|u_0|_{H^2(K)}.\end{aligned}$$

Here we have used the fact that for any $K \in M_h$

$$\|\chi^j\|_{L^\infty(K)} + \varepsilon\|\nabla\chi^j\|_{L^\infty(K)} \leqslant C,$$

where C is independent of K, h, ε. Hence

$$\|\varepsilon\nabla(u_1 - u_{I1})\|_{L^2(\Omega)} \leqslant C(h+\varepsilon)|u_0|_{H^2(\Omega)}.$$

On the other hand

$$\|\varepsilon(u_1 - u_{I1})\|_{L^2(\Omega)} = \varepsilon\|\chi^j\partial(u_0 - \Pi_h u_0)/\partial x_j\|_{L^2(\Omega)} \leqslant Ch\varepsilon|u_0|_{H^2(\Omega)}.$$

Thus, we have

$$\|\varepsilon(u_1 - u_{I1})\|_{1,\Omega} \leqslant C(h+\varepsilon)|u_0|_{H^2(\Omega)} \leqslant C(h+\varepsilon)\|f\|_{L^2(\Omega)}. \tag{9.22}$$

Next we estimate $\|\varepsilon\theta_\varepsilon\|_{H^1(\Omega)}$ and $\|\varepsilon\theta_{I\varepsilon}\|_{H^1(\Omega)}$ respectively. Let $\xi \in C_0^\infty(R^2)$ be the cut-off function such that $0 \leqslant \xi \leqslant 1$, $\xi = 1$ in $\Omega \setminus \Omega_{\varepsilon/2}$, $\xi = 0$ in Ω_ε, and $|\nabla\xi| \leqslant C/\varepsilon$ in Ω, where $\Omega_\varepsilon := \{x : \mathrm{dist}\{x, \partial\Omega\} \geqslant \varepsilon\}$. Then

$$\theta_\varepsilon + \xi(\chi^j\partial u_0/\partial x_j) \in H_0^1(\Omega).$$

Thus, from (9.6) we get

$$(a(x/\varepsilon)\nabla\theta_\varepsilon, \nabla(\theta_\varepsilon + \xi\chi^j\partial u_0/\partial x_j)) = 0.$$

Hence

$$\begin{aligned}\|\nabla\theta_\varepsilon\|_{L^2(\Omega)} &\leqslant C\|\nabla(\xi\chi^j\partial u_0/\partial x_j)\|_{L^2(\Omega)}\\ &\leqslant C\|\nabla\xi\chi^j\partial u_0/\partial x_j\|_{L^2(\Omega)} + C\|\xi\nabla\chi^j\partial u_0/\partial x_j\|_{L^2(\Omega)}\\ &\quad + C\|\xi\chi^j\nabla(\partial u_0/\partial x_j)\|_{L^2(\Omega)}\\ &\leqslant C\|\nabla u_0\|_{L^\infty(\Omega)}\sqrt{|\partial\Omega|\varepsilon}/\varepsilon + C|u_0|_{H^2(\Omega)} \qquad (9.23)\end{aligned}$$

On the other hand, from the maximum principle, we have

$$\|\theta_\varepsilon\|_{L^\infty(\Omega)} \leqslant \|\chi^j\partial u_0/\partial x_j\|_{L^\infty(\partial\Omega)} \leqslant C\|u_0\|_{W^{1,\infty}(\Omega)}.$$

Thus, we obtain

$$\|\varepsilon\theta_\varepsilon\|_{H^1(\Omega)} \leqslant C\sqrt{\varepsilon}\|u_0\|_{W^{1,\infty}(\Omega)} + C\varepsilon|u_0|_{H^2(\Omega)}. \qquad (9.24)$$

Finally, we estimate $\|\varepsilon\theta_{I\varepsilon}\|_{H^1(\Omega)}$. From the maximum principle, we have

$$\begin{aligned}\|\theta_{I\varepsilon}\|_{L^\infty(K)} \leqslant \|\chi^j\partial(\Pi_h u_0)/\partial x_j\|_{L^\infty(\partial K)} &\leqslant C\|\Pi_h u_0\|_{W^{1,\infty}(K)}\\ &\leqslant C\|u_0\|_{W^{1,\infty}(K)}.\end{aligned}$$

Hence

$$\|\varepsilon\theta_{I\varepsilon}\|_{L^2(\Omega)} \leqslant C\varepsilon\|u_0\|_{W^{1,\infty}(\Omega)}.$$

Similar to (9.23), we have

$$\begin{aligned}\|\varepsilon\nabla\theta_{I\varepsilon}\|_{L^2(K)} &\leqslant C\|\nabla u_0\|_{L^\infty(K)}\sqrt{|\partial K|\varepsilon} + C\varepsilon|\Pi_h u_0|_{H^2(K)}\\ &\leqslant C\sqrt{h\varepsilon}\|u_0\|_{W^{1,\infty}(K)},\end{aligned}$$

which implies

$$\|\varepsilon\nabla\theta_{I\varepsilon}\|_{L^2(\Omega)} \leqslant C(\varepsilon/h)^{1/2}\|u_0\|_{W^{1,\infty}(\Omega)}.$$

Hence

$$\|\varepsilon\theta_{I\varepsilon}\|_{1,\Omega} \leqslant C((\varepsilon/h)^{1/2} + \varepsilon)\|u_0\|_{W^{1,\infty}(\Omega)}. \qquad (9.25)$$

Combing (9.21)–(9.22), (9.24)–(9.25) and using Céa Lemma, we obtain

$$\|u - u_h\|_{1,\Omega} \leqslant C(h+\varepsilon)\|f\|_{L^2(\Omega)} + C((\varepsilon/h)^{1/2} + \varepsilon^{1/2})\|u_0\|_{W^{1,\infty}(\Omega)}.$$

This completes the proof. □

We remark that the error estimate in Theorem 9.3 is uniform when $\varepsilon \to 0$ which suggests that one can take the mesh size h larger than ε in using the multiscale finite element methods. The term $(\varepsilon/h)^{1/2}$ in the error estimate is due to the mismatch of the multiscale finite element basis functions with the solution u of the original problem inside the domain. One way to improve this error is the over-sampling finite element method that we introduce in the next section.

9.3. The over-sampling multiscale finite element method

Let $\mathcal{M}_H$ be a regular and quasi-uniform triangulation of Ω and W_H the H^1-conforming linear finite element space over $\mathcal{M}_H$. For any $K \in \mathcal{M}_H$ with nodes $\{x_i^K\}_{i=1}^3$, let $\{\varphi_i^K\}_{i=1}^3$ be the basis of $P_1(K)$ satisfying $\varphi_i^K(x_j^K) = \delta_{ij}$. For any $K \in \mathcal{M}_H$, we denote by $S = S(K)$ a macro-element which contains K and satisfies that $H_S \leqslant C_1 H_K$ and $\operatorname{dist}(\partial K, \partial S) \geqslant \delta_0 H_K$ for some positive constants C_1, δ_0 indipendent of H. The minimum angle of $S(K)$ is bounded below by some positive constant θ_0 independent of H.

Let $M_S(K)$ be the multiscale finite element space spanned by $\psi_i^K, i = 1, 2, 3$, with $\psi_i^S \in H^1(S)$ being the solution of the problem

$$-\operatorname{div}(a_\varepsilon \nabla \psi_i^S) = 0 \qquad \text{in } S, \qquad \psi_i^S|_{\partial S} = \varphi_i^S.$$

Here $\{\varphi_i^S\}_{i=1}^3$ is the nodal basis of $P_1(S)$ such that $\varphi_i^S(x_j^S) = \delta_{ij}, i, j = 1, 2, 3$. The over-sampling multiscale finite element base functions over K is defined by

$$\bar{\psi_i}^K = c_{ij}^K \psi_j^S|_K \qquad \text{in } K,$$

with the constants so chosen that

$$\varphi_i^K = c_{ij}^K \varphi_j^S|_K \qquad \text{in } K.$$

The existence of the constants c_{ij}^K is guaranteed because $\left\{\varphi_j^S\right\}_{j=1}^3$ also forms the basis of $P_1(K)$.

Let $\operatorname{OMS}(K) = \operatorname{span}\{\bar{\psi_i}^K\}_{i=1}^3$ and $\Pi_K : \operatorname{OMS}(K) \to P_1(K)$ the projection

$$\Pi_K \psi = c_i \varphi_i^K \qquad \text{if} \qquad \psi = c_i \bar{\psi_i}^K \in \operatorname{OMS}(K).$$

Let $\bar{X}_H$ be the finite element space

$$\bar{X}_H = \{\psi_H : \psi_H|_K \in \operatorname{OMS}(K) \ \ \forall K \in \mathcal{M}_H\}$$

and define $\Pi_H : \bar{X}_H \to \Pi_{K \in \mathcal{M}_H} P_1(K)$ through the relation

$$\Pi_H \psi_H|_K = \Pi_K \psi_H \ \ \text{for any } K \in \mathcal{M}_H, \psi_H \in \bar{X}_H.$$

The over-sampling multiscale finite element space is then defined as

$$X_H = \left\{\psi_H \in \bar{X}_H : \Pi_H \psi_H \in W_H \subset H^1(\Omega)\right\}.$$

In general, $X_H \not\subset H^1(\Omega)$ and the requirement $\Pi_H \psi_H \in W_H$ is to impose certain continuity of the functions $\psi_H \in X_H$ across the inter-element boundaries. Here we have an example of nonconforming finite element method.

The multiscale finite element method is then to find $u_H \in X_H^0$, where

$$X_H^0 = \{\psi_H \in X_H : \Pi_H \psi_H = 0 \quad \text{on } \partial\Omega\},$$

such that

$$\sum_{K \in \mathcal{M}_H} \int_K a_\varepsilon \nabla u_H \nabla \psi_H \,\mathrm{d}x = \int_\Omega f \psi_H \,\mathrm{d}x \qquad \forall \psi_H \in X_H^0.$$

We introduce the bilinear form $a_H(\cdot,\cdot) : \prod_{K\in\mathcal{M}_H} H^1(K) \times \prod_{K\in\mathcal{M}_H} H^1(K) \to \mathbb{R}$

$$a_H(\varphi,\psi) = \sum_{K\in\mathcal{M}_H} \int_K a_\varepsilon \nabla\varphi\nabla\psi \,\mathrm{d}x \quad \forall \varphi,\psi \in \prod_{K\in\mathcal{M}_H} H^1(K),$$

and the discrete norm

$$\|\varphi\|_{h,\Omega} = \left(\sum_{K\in\mathcal{M}_H} \|\nabla\varphi\|^2_{L^2(K)}\right)^{1/2} \qquad \forall\varphi \in \prod_{K\in\mathcal{M}_H} H^1(K).$$

Lemma 9.3. *We have*

$$\begin{aligned}&\|u_\varepsilon - u_H\|_{h,\Omega}\\ &\leqslant C \inf_{\psi_H \in X_H^0} \|u_\varepsilon - \psi_H\|_{h,\Omega} + C \sup_{0\neq\psi_H\in X_H^0} \frac{\left|\int_\Omega f\psi_H \,\mathrm{d}x - a_H(u_\varepsilon,\psi_H)\right|}{\|\psi_H\|_{h,\Omega}}.\end{aligned}$$

Proof. Define $\langle R, \psi_H\rangle = \int_\Omega f\psi_H - a_H(u_\varepsilon,\psi_H)$, then we have

$$a_H(u_\varepsilon - u_H, \psi_H) = -\langle R,\psi_H\rangle \quad \forall\psi_H \in X_H^0.$$

The lemma follows easily by taking $\psi_H = v_H - u_H$ for any $v_H \in X_H^0$. □

Lemma 9.4. *Let $N \in L^\infty(R^2)$ be a periodic function with respect to the cell Y and assume $\int_Y N(y)\,\mathrm{d}y = 0$. Then for any $\zeta \in H^1(K)\cap L^\infty(K), K \in \mathcal{M}_H$, we have*

$$\left|\int_K \zeta(x) N\big(\frac{x}{\varepsilon}\big)\,\mathrm{d}x\right| \leqslant C h_K \varepsilon \|\nabla\zeta\|_{L^2(K)} + C\varepsilon h_K \|\zeta\|_{L^\infty(K)}.$$

Proof. Define $\zeta_i = \int_{Y_i} \zeta \,\mathrm{d}x$, where Y_i is a periodic cell of $N(x/\varepsilon), Y_i \subset K$. Then

$$\|\zeta - \zeta_i\|_{L^2(Y_i)} \leqslant C\varepsilon \|\nabla\zeta\|_{L^2(Y_i)}.$$

Denote $K' = \cup_{Y_i \subset K} Y_i$, we have

$$\begin{aligned}\left|\int_K \zeta(x) N\big(\frac{x}{\varepsilon}\big)\,\mathrm{d}x\right| \leqslant& \left|\sum_{Y_i \subset K}\int_{Y_i}(\zeta-\zeta_i)N\big(\frac{x}{\varepsilon}\big)\,\mathrm{d}x\right| + \left|\sum_{Y_i\subset K}\int_{Y_i}\zeta_i N\big(\frac{x}{\varepsilon}\big)\,\mathrm{d}x\right| \\ &+\left|\int_{K\setminus K'}\zeta(x)N\big(\frac{x}{\varepsilon}\big)\,\mathrm{d}x\right| \\ \leqslant& C\varepsilon\,\|\nabla\zeta\|_{L^2(K')}\,\|N\|_{L^2(K')} + C\,\|\zeta\|_{L^\infty(K)}\,\big|K\setminus K'\big| \\ \leqslant& Ch_K\varepsilon\,\|\nabla\zeta\|_{L^2(K)} + C\varepsilon h_K\,\|\zeta\|_{L^\infty(K)}\,.\end{aligned}$$

This completes the proof. □

Lemma 9.5. *There exist constants γ_2 and C independent of H and ε such that if $H_K \leqslant \gamma_2$ and $\varepsilon/H_K \leqslant \gamma_2$ for all $K \in \mathcal{M}_H$, the following estimates are valid*

$$C^{-1}\,\|\nabla\Pi_H\chi_H\|_{L^2(K)} \leqslant \|\nabla\chi_H\|_{L^2(K)} \leqslant C\,\|\nabla\Pi_H\chi_H\|_{L^2(K)} \qquad \forall\chi_H \in X_H^0.$$

Proof. We know that $\chi_H \in H^1(S)$ satisfies

$$-\nabla\cdot(a_\varepsilon\nabla\chi_H) = 0 \quad \text{in } S, \qquad \chi_H = \Pi_H\chi_H \quad \text{on } \partial S. \tag{9.26}$$

For any $\varphi \in H_0^1(S)$ we have

$$(a_\varepsilon\nabla\chi_H, \nabla\varphi)_S = 0.$$

By taking $\varphi = \chi_H - \Pi_H\chi_H \in H_0^1(S)$ we obtain easily that

$$\|\nabla\chi_H\|_{L^2(K)} \leqslant \|\nabla\chi_H\|_{L^2(S)} \leqslant C\,\|\nabla\Pi_H\chi_H\|_{L^2(S)} \leqslant C\,\|\nabla\Pi_H\chi_H\|_{L^2(K)}\,.$$

Next by Theorem 9.1 we have the asymptotic expansion

$$\chi_H = \chi_H^0 - \varepsilon\chi^j\frac{\partial\chi_H^0}{\partial x_j} - \varepsilon\theta_\varepsilon^S, \tag{9.27}$$

where $\chi_H^0 = \Pi_H\chi_H$ and $\theta_\varepsilon^S \in H^1(S)$ is the boundary corrector defined by

$$-\nabla\cdot(a_\varepsilon\nabla\theta_\varepsilon^S) = 0 \quad \text{in } S, \qquad \theta_\varepsilon^S\big|_{\partial S} = -\chi^j\frac{\partial\chi_H^0}{\partial x_j}.$$

By simple calculations

$$\begin{aligned} a_{ij}\frac{\partial\chi_H}{\partial x_j} &= a_{ij}^*\frac{\partial\chi_H^0}{\partial x_j} - G_i^k\frac{\partial\chi_H^0}{\partial x_k} - \varepsilon a_{ij}\chi^k\frac{\partial^2\chi_H^0}{\partial x_j\partial x_k} - \varepsilon(a\nabla\theta_\varepsilon^S)_i \\ &= a_{ij}^*\frac{\partial\chi_H^0}{\partial x_j} - G_i^k\big(\frac{x}{\varepsilon}\big)\frac{\partial\chi_H^0}{\partial x_k} - \varepsilon(a\nabla\theta_\varepsilon^S)_i, \end{aligned} \tag{9.28}$$

where G_i^K satisfies

$$\int_Y G_i^k(y)\,\mathrm{d}y = 0 \qquad \text{and} \qquad \frac{\partial G_i^k}{\partial y_i} = 0.$$

Multiplying (9.28) by $\nabla\chi_H^0$ and integrating over K we see

$$\int_K a_{ij}^* \frac{\partial \chi_H^0}{\partial x_j}\frac{\partial \chi_H^0}{\partial x_i}\,\mathrm{d}x = \int_K a_{ij}\frac{\partial \chi_H}{\partial x_j}\frac{\partial \chi_H^0}{\partial x_i}\,\mathrm{d}x + \int_K G_i^k(\frac{x}{\varepsilon})\frac{\partial \chi_H^0}{\partial x_k}\frac{\partial \chi_H^0}{\partial x_i}\,\mathrm{d}x$$
$$- \varepsilon \int_K a_{ij}\frac{\partial \theta_\varepsilon^S}{\partial x_j}\frac{\partial \chi_H^0}{\partial x_i}\,\mathrm{d}x.$$

From the interior estimate due to Avellaneda and Lin [**4**, Lemma 16]

$$\left\|\nabla\theta_\varepsilon^S\right\|_{L^\infty(K)} \leqslant C h_K^{-1}\left\|\theta_\varepsilon^S\right\|_{L^\infty(S)}.$$

Therefore by the maximum principle and the finite element inverse estimate

$$\left|\varepsilon\int_K a_{ij}\frac{\partial \theta_\varepsilon^S}{\partial x_j}\frac{\partial \chi_H^0}{\partial x_i}\,\mathrm{d}x\right| \leqslant C\varepsilon h_K\left\|\nabla\theta_\varepsilon^S\right\|_{L^\infty(K)}\left\|\nabla\chi_H^0\right\|_{L^2(K)}$$
$$\leqslant C\frac{\varepsilon}{h_K}\left\|\nabla\chi_H^0\right\|_{L^2(K)}^2.$$

By Lemma 9.4 we have

$$\left|\int_K G_i^k(\frac{x}{\varepsilon})\frac{\partial \chi_H^0}{\partial x_k}\frac{\partial \chi_H^0}{\partial x_i}\,\mathrm{d}x\right| \leqslant C\varepsilon h_K\left\|\nabla\chi_H^0\right\|_{L^\infty(K)}^2 \leqslant C\frac{\varepsilon}{h_K}\left\|\nabla\chi_H^0\right\|_{L^2(K)}^2.$$

Thus

$$\alpha^*\left\|\nabla\chi_H^0\right\|_{L^2(K)}^2 \leqslant C\left\|\nabla\chi_H\right\|_{L^2(K)}\left\|\nabla\chi_H^0\right\|_{L^2(K)} + C\frac{\varepsilon}{h_K}\left\|\nabla\chi_H^0\right\|_{L^2(K)}^2.$$

This completes the proof. □

THEOREM 9.4. *we have*

$$\|u_\varepsilon - u_H\|_{h,\Omega} \leqslant C(h+\varepsilon)\|f\|_{L^2(\Omega)} + C\Big(\frac{\varepsilon}{h}+\sqrt{\varepsilon}\Big)\Big(\|u_0\|_{W^{1,\infty}(\Omega)} + \|f\|_{L^2(\Omega)}\Big).$$

PROOF. First we notice that by Theorem 9.1

$$\left\|\nabla\Big[u_\varepsilon - \big(u_0 - \varepsilon x^j\frac{\partial u_0}{\partial x_j} - \varepsilon\theta_\varepsilon\big)\Big]\right\|_{L^2(\Omega)} \leqslant C\varepsilon\,|u|_{H^2(\Omega)}.$$

By the estimate (9.24)

$$\varepsilon\left\|\nabla\theta_\varepsilon\right\|_{L^2(\Omega)} \leqslant C\sqrt{\varepsilon}\|u_0\|_{W^{1,\infty}(\Omega)} + C\varepsilon\,|u_0|_{H^2(\Omega)},$$

we get

$$\left\|\nabla\big(u_\varepsilon - u_0 + \varepsilon x^j\frac{\partial u_0}{\partial x_j}\big)\right\|_{L^2(\Omega)} \leqslant C\varepsilon\,|u_0|_{H^2(\Omega)} + C\sqrt{\varepsilon}\|u_0\|_{W^{1,\infty}(\Omega)}. \tag{9.29}$$

We take

$$\psi_H = \sum_{x_j \text{ interior node}} u_0(x_j)\bar{\psi}_j(x) \in X_H^0,$$

then

$$\Pi_H \psi_H|_K = I_H u_0 \qquad \forall K \in \mathcal{M}_H.$$

where $I_H : C(\bar{\Omega}) \to W_H$ is the standard Lagrange interpolation operator over linear finite element space. By (9.27) we know that

$$\psi_H = (I_H u_0) - \varepsilon \chi^j \frac{\partial(I_H u_0)}{\partial x_j} - \varepsilon \theta_\varepsilon^S,$$

where $\theta_\varepsilon^S \in H^1(S)$ is the boundary corrector given by

$$-\nabla \cdot (a_\varepsilon \nabla \theta_\varepsilon^S) = 0 \quad \text{in } S, \qquad \theta_\varepsilon^S|_{\partial S} = -\chi^j \frac{\partial(I_H u_0)}{\partial x_j}.$$

By the interior estimate in Avellaneda and Lin [**4**, Lemma 16]

$$\left\|\nabla \theta_\varepsilon^S\right\|_{L^\infty(K)} \leqslant C h_K^{-1} \left\|\theta_\varepsilon^S\right\|_{L^\infty(S)} \leqslant C h_K^{-1} \left|I_H u_0\right|_{W^{1,\infty}(S)} \leqslant C h_K^{-1} \left|u_0\right|_{W^{1,\infty}(K)}.$$

Therefore

$$\begin{aligned}\left\|\nabla\Big(\psi_H - \big(I_H u_0 + \varepsilon \chi^j \frac{\partial(I_H u_0)}{\partial x_j}\big)\Big)\right\|_{L^2(K)} &\leqslant C\varepsilon h_K^{-1} \left|u_0\right|_{W^{1,\infty}(K)} |K|^{1/2} \\ &\leqslant C\varepsilon \left|u_0\right|_{W^{1,\infty}(K)}. \qquad (9.30)\end{aligned}$$

Since

$$\begin{aligned}&\left\|\nabla(u_0 - I_H u_0)\right\|_{L^2(K)} \leqslant C h_K \left|u_0\right|_{H^2(K)}, \\ &\left\|\varepsilon \nabla\Big(\chi^j \frac{\partial(I_H u_0 - u_0)}{\partial x_j}\Big)\right\|_{L^2(K)} \leqslant C(h + \varepsilon)\left|u_0\right|_{H^2(K)},\end{aligned}$$

we finally obtain

$$\left\|\nabla(u_\varepsilon - \psi_H)\right\|_{h,\Omega} \leqslant C(h+\varepsilon)\left|u_0\right|_{H^2(K)} + C\left(\frac{\varepsilon}{h} + \sqrt{\varepsilon}\right)\left\|u_0\right\|_{W^{1,\infty}(\Omega)}.$$

It remains to estimate the non-conforming error. Since $\Pi_H \psi_H \in W_H \subset H^1(\Omega)$ we know that

$$\begin{aligned}&\left|\int_\Omega f \psi_H \,\mathrm{d}x - a_H(u_\varepsilon, \psi_H)\right| \\ &= \left|\int_\Omega f(\psi_H - \Pi_H \psi_H)\,\mathrm{d}x - \sum_{K \in \mathcal{M}_h} \int_K a_\varepsilon \nabla u_\varepsilon \nabla(\psi_H - \Pi_H \psi_H)\,\mathrm{d}x\right|.\end{aligned}$$

By (9.27) and Lemma 9.5 we have

$$
\begin{aligned}
&\left|\int_K f(\psi_H - \Pi_H\psi_H)\,\mathrm{d}x\right| = \left|\int_K f\left(-\varepsilon\chi^j\frac{\partial\Pi_H\psi_H}{\partial x_j} - \varepsilon\theta_\varepsilon^S\right)\mathrm{d}x\right| \\
&\leqslant C\varepsilon\,\|f\|_{L^2(K)}\big(\,\|\nabla\Pi_H\psi_H\|_{L^2(K)} + \|\nabla\Pi_H\psi_H\|_{L^\infty(K)}\big) \\
&\leqslant C\frac{\varepsilon}{h}\,\|f\|_{L^2(K)}\,\|\nabla\psi_H\|_{L^2(K)}\,.
\end{aligned}
$$

Thus

$$
\left|\int_\Omega f(\psi_H - \Pi_H\psi_H)\,\mathrm{d}x\right| \leqslant C\frac{\varepsilon}{h}\,\|f\|_{L^2(\Omega)}\,\|\psi_H\|_{h,\Omega}\,.
$$

Furthermore, by (9.29)

$$
\begin{aligned}
&\sum_{K\in\mathcal{M}_H}\left|\int_K a_\varepsilon\nabla u_\varepsilon\nabla(\psi_H - \Pi_h\psi_H)\right| \\
&\leqslant C\big(\varepsilon\,|u_0|_{H^2(\Omega)} + \sqrt{\varepsilon}\|u_0\|_{W^{1,\infty}(\Omega)}\big)\,\|\psi_H\|_{h,\Omega} \\
&\quad + \sum_{K\in\mathcal{M}_H}\left|\int_K a_\varepsilon\nabla\left(u_0 + \varepsilon\chi^j\frac{\partial u_0}{\partial x_j}\right)\cdot\nabla(\psi_H - \Pi_H\psi_H)\,\mathrm{d}x\right| \\
&= \mathrm{I} + \mathrm{II}.
\end{aligned}
$$

But

$$
\begin{aligned}
|\mathrm{II}| \leqslant \sum_{K\in\mathcal{M}_H}\Bigg(&\left|\int_K a^*\nabla u_0\nabla(\psi_H - \Pi_H\psi_H)\,\mathrm{d}x\right| \\
&+\left|\int_K G_i^k\frac{\partial u_0}{\partial x_k}\frac{\partial(\psi_H - \Pi_H\psi_H)}{\partial x_i}\,\mathrm{d}x\right| \\
&+\varepsilon\,|u_0|_{H^2(K)}\,\|\nabla(\psi_H - \Pi_H\psi_H)\|_{L^2(K)}\Bigg) \\
=&\mathrm{II}_1 + \mathrm{II}_2 + \mathrm{II}_3.
\end{aligned}
$$

By (9.27) and Lemma 9.4

$$
\begin{aligned}
&\left|\int_K a_{ij}^*\frac{\partial u_0}{\partial x_i}\frac{\partial}{\partial x_j}(\psi_H - \Pi_H\psi_H)\,\mathrm{d}x\right| \\
&= \left|\int_K a_{ij}^*\frac{\partial u_0}{\partial x_i}\frac{\partial}{\partial x_j}\left(\varepsilon\chi^k\frac{\partial\Pi_H\psi_H}{\partial x_k} - \varepsilon\theta_\varepsilon^S\right)\mathrm{d}x\right| \\
&\leqslant \left|\int_K a_{ij}^*\frac{\partial u_0}{\partial x_i}\frac{\partial\chi^k}{\partial y_j}\frac{\partial\Pi_H\psi_H}{\partial x_k}\,\mathrm{d}x\right| + \varepsilon\,\|\nabla u_0\|_{L^2(K)}\,\left\|\nabla\theta_\varepsilon^S\right\|_{L^2(K)} \\
&\leqslant C\varepsilon\big(\,|u_0|_{H^2(K)} + |u_0|_{W^{1,\infty}(K)}\,\big)\,\|\nabla\psi_H\|_{L^2(K)}\,.
\end{aligned}
$$

That is,

$$|\mathrm{II}_1| \leqslant C\Big(\varepsilon\,|u_0|_{H^2(\Omega)} + C\frac{\varepsilon}{h}\,|u_0|_{W^{1,\infty}(\Omega)}\Big)\,\|\psi_H\|_{h,\Omega}\,.$$

Similarly, we know that

$$|\mathrm{II}_2| \leqslant C\Big(\varepsilon\,|u_0|_{H^2(\Omega)} + C\frac{\varepsilon}{h}\,|u_0|_{W^{1,\infty}(\Omega)}\Big)\,\|\psi_H\|_{h,\Omega}\,.$$

It is obvious that

$$|\mathrm{II}_3| \leqslant C\varepsilon\,|u_0|_{H^2(\Omega)}\,\|\psi_H\|_{h,\Omega}\,.$$

This shows that the non-conforming error in Lemma 9.3

$$\sup_{0\neq\psi_H\in X_H^0}\frac{\big|\int_\Omega f\psi_H\,\mathrm{d}x - a_H(u_\varepsilon,\psi_H)\big|}{\|\psi_H\|_{h,\Omega}}$$
$$\leqslant C\varepsilon\,|u_0|_{H^2(\Omega)} + C\big(\frac{\varepsilon}{h}+\sqrt{\varepsilon}\big)\Big(\,\|f\|_{L^2(\Omega)} + \|u_0\|_{W^{1,\infty}(\Omega)}\Big).$$

This completes the proof. □

Bibliographic notes. Homogenization theory for elliptic equations with highly oscillatory coefficients is a topic of intensive studies. We refer to the monographs Bensoussan et al [**9**] and Jikov et al [**39**] for further results. Theorem 9.1 is taken from [**39**]. The multiscale finite element method is introduced in Hou and Wu [**37**] and Hou et al [**38**]. The over-sampling multiscale finite element is introduced in Efendiev et al [**29**]. Further development of multiscale finite elements can be found in Chen and Hou [**19**] for the mixed multiscale finite element method and in Chen and Yue [**22**] for the multiscale finite element method dealing with well singularities.

9.4. Exercises

Exercise 9.1. Show that the homogenized coefficient a^* satisfies

$$a^*_{ij}\xi_i\xi_j \geqslant \gamma|\xi|^2 \quad \forall\xi\in\mathbb{R}^2.$$

Exercise 9.2. Prove Lemma 9.1.

CHAPTER 10

Implementations

In this chapter we talk about some implementation issues. First we give a brief introduction to the MATLAB PDE Toolbox. Then we show how to solve the L-shaped domain problem on uniform meshes and adaptive meshes by MATLAB. Finally we introduce the implementation of the multigrid V-cycle algorithm.

10.1. A brief introduction to the MATLAB PDE Toolbox

The MATLAB Partial Differential Equation (PDE) Toolbox is a tool for solving partial differential equations in two space dimensions and time by linear finite element methods on triangular meshes. The PDE Toolbox can solve linear or nonlinear elliptic PDE

$$-\nabla \cdot (c\nabla u) + au = f, \tag{10.1}$$

the linear parabolic PDE

$$d\frac{\partial u}{\partial t} - \nabla \cdot (c\nabla u) + au = f, \tag{10.2}$$

the linear hyperbolic PDE

$$d\frac{\partial^2 u}{\partial t^2} - \nabla \cdot (c\nabla u) + au = f, \tag{10.3}$$

or the linear eigenvalue problem

$$-\nabla \cdot (c\nabla u) + au = \lambda du, \tag{10.4}$$

in a plane region Ω, with boundary condition

$$h\,u = r \quad \text{on } \Gamma_1, \tag{10.5}$$

$$(c\nabla u) \cdot \mathbf{n} + q\,u = g \quad \text{on } \Gamma_2, \tag{10.6}$$

where $\overline{\Gamma_1 \cup \Gamma_2} = \partial\Omega$, $\Gamma_1 \cap \Gamma_2 = \emptyset$. The PDE Toolbox can also solve the PDE systems. The PDE Toolbox includes tools that:

- Define a PDE problem, i.e., define 2-D regions, boundary conditions, and PDE coefficients;

- Numerically solve the PDE problem, i.e., generate unstructured meshes, discretize the equations, and produce an approximation to the solution;
- Visualize the results.

There are two approaches to define and solve a PDE problem: by using a graphical user interface (GUI) or by MATLAB programming. The GUI can be started by typing

```
pdetool
```

at the MATLAB command line. From the command line (or M-files) you can call functions from the toolbox to do the hard work, e.g., generate meshes, discretize your problem, perform interpolation, plot data on unstructured grids, etc., while you retain full control over the global numerical algorithm.

One advantage of the PDE Toolbox is that it is written using the MATLAB open system philosophy. There are no black-box functions, although some functions may not be easy to understand at first glance. The data structures and formats are documented. You can examine existing functions and create your own as needed.

10.1.1. A first example—Poisson equation on the unit disk. We consider the Poisson equation

$$\begin{aligned} -\nabla \cdot (\nabla u) &= 1 \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{10.7}$$

on the unit disk Ω. For this problem, you can compare the exact solution $u = (1 - x^2 - y^2)/4$ with the numerical solution at the nodal points on the mesh. To set up the PDE on the command line follow these steps (cf. `pdedemo1.m`):

1. Create a unit circle centered at the origin using the geometry M-file "circleg.m":

```
g='circleg';
```

2. The `initmesh` function creates a triangular mesh on the geometry defined in `g`:

```
[p,e,t]=initmesh(g);
pdemesh(p,e,t); axis equal; %Plot the mesh.
```

3. Specify the PDE coefficients:

```
c=1;
a=0;
```

```
f=1;
```

4. Specify the boundary condition:

```
b='circleb1';
```

5. Solve the PDE and plot the solution:

```
u=assempde(b,p,e,t,c,a,f);
pdesurf(p,t,u);
```

6. Compute the maximum error:

```
exact=(1-p(1,:).^2-p(2,:).^2)'/4;
error=max(abs(u-exact));
fprintf('Error: %e. Number of nodes: %d\n',...
        error,size(p,2));
pdesurf(p,t,u-exact); %Plot the error.
```

7. If the error is not sufficiently small, refine the mesh:

```
[p,e,t]=refinemesh(g,p,e,t);
```

You can then solve the problem on the new mesh, plot the solution, and recompute the error by repeating Steps 5 and 6.

10.1.2. The mesh data structure. A triangular mesh is described by the mesh data which consists of a Point matrix, an Edge matrix, and a Triangle matrix.

In the mesh vertex matrix (for example, denoted by p), the first and second row contain x- and y-coordinates of the mesh vertices in the mesh.

```
p = [x     % x coordinates for mesh vertices
     y];   % y coordinates for mesh vertices
```

In the boundary element matrix (for example, denoted by e), the first and second row contain indices of the starting and ending point, the third and fourth row contain the starting and ending parameter values, the fifth row contains the boundary segment number, and the sixth and seventh row contain the left- and right-hand side subdomain numbers.

```
e = [p1;p2    % index to column in p
     s1;s2    % arc-length parameters
     en       % geometry boundary number
     l        % left-subdomain number
     r];      % right-subdomain number
```

In the element matrix (for example, denoted by t), the first three rows contain indices to the corner points, given in counter-clockwise order, and the fourth row contains the subdomain number.

```
t = [p1; p2; p3    % index to column in p
     sd];          % subdomain number
```

We remark that the (global) indices to the nodal points are indicated by the column numbers of the point matrix p, i.e., the coordinates of the i-th point is p(:,i). The edge matrix e contains only the element sides on the boundary of the (sub)domain(s). In the j-th element (the triangle defined by the j-th column of t), the 1st–3rd rows gives the global indices of the 1st–3rd vertices of the element. It is clear that, the first three rows of element matrix t defines a map from the local indices of the nodal points to their global indices. This relationship is important in assembling the global stiffness matrix from the element stiffness matrices in the finite element discretization (see Section 2.3).

For example, we consider the unit square described by the decomposed geometry matrix

```
g = [2     2     2     2
     0     1     1     0
     1     1     0     0
     1     1     0     0
     1     0     0     1
     0     0     0     0
     1     1     1     1];
```

Here the decomposed geometry matrix g is obtained as follows. We first draw the geometry (the unit square) in the GUI, then export it by selecting "Export the Decomposed Geometry, Boundary Cond's" from the "Boundary" menu. For details on the decomposed geometry matrix we refer to the help on the function "decsg.m". Figure 1 shows a standard triangulation of the unit square obtained by running

```
[p,e,t] = poimesh(g,2);
```

Here the output mesh data

```
p = [0   0.5 1   0   0.5 1   0   0.5 1
     0   0   0   0.5 0.5 0.5 1   1   1];

e = [1   2   3   6   9   8   7   4
```

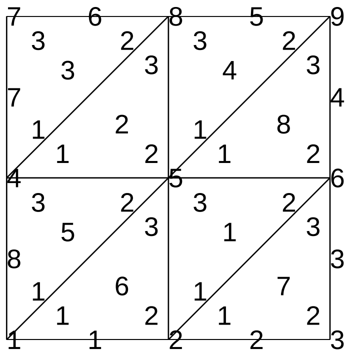

FIGURE 1. A standard triangulation of the unit square. The numbers give the global and local indices to the points, the indices to the elements, and the indices to the edges, respectively.

```
          2   3   6   9   8   7   4   1
          1   0.5 1   0.5 1   0.5 1   0.5
          0.5 0   0.5 0   0.5 0   0.5 0
          3   3   2   2   1   1   4   4
          1   1   1   1   1   1   1   1
          0   0   0   0   0   0   0   0];

     t = [2   4   4   5   1   1   2   5
          6   5   8   9   5   2   3   6
          5   8   7   8   4   5   6   9
          1   1   1   1   1   1   1   1];
```

Figure 1 also shows the global and local indices to the points, the indices to the elements, and the indices to the edges, respectively.

EXAMPLE 10.1. *Assemble the stiffness matrix for the Poisson equation* (10.7) *on a given mesh* `p, e, t`. *The following function assembles the stiffness matrix from the element stiffness matrices which is analogous to the function "pdeasmc.m".*

CODE 10.1. (Assemble the Poission equation)

```
function [A,F,B,ud]=pdeasmpoi(p,e,t)
% Assemble the Poission's equation -div(grad u)=1
% with homogeneous Dirichlet boundary condition.
```

```
%
% A is the stiffness matrix, F is the right-hand side vector.
% UN=A\F returns the solution on the non-Dirichlet points.
% The solution to the full PDE problem can be obtained by the
% MATLAB command U=B*UN+ud.

% Corner point indices
it1=t(1,:);
it2=t(2,:);
it3=t(3,:);

np=size(p,2); % Number of points

% Areas and partial derivatives of nodal basis functions
[ar,g1x,g1y,g2x,g2y,g3x,g3y]=pdetrg(p,t);

% The element stiffness matrices AK.
c3=((g1x.*g2x+g1y.*g2y)).*ar; % AK(1,2)=AK(2,1)=c3
c1=((g2x.*g3x+g2y.*g3y)).*ar; % AK(2,3)=AK(3,2)=c1
c2=((g3x.*g1x+g3y.*g1y)).*ar; % AK(1,3)=AK(3,1)=c2
% AK(1,1)=-AK(1,2)-AK(1,3)=-c2-c3
% AK(2,2)=-AK(2,1)-AK(2,3)=-c3-c1
% AK(3,3)=-AK(3,1)-AK(3,2)=-c1-c2

% Assemble the stiffness matrix
A=sparse(it1,it2,c3,np,np);
A=A+sparse(it2,it3,c1,np,np);
A=A+sparse(it3,it1,c2,np,np);
A=A+A.';
A=A+sparse(it1,it1,-c2-c3,np,np);
A=A+sparse(it2,it2,-c3-c1,np,np);
A=A+sparse(it3,it3,-c1-c2,np,np);

% Assmeble the right-hand side
f=ar/3;
F=sparse(it1,1,f,np,1);
F=F+sparse(it2,1,f,np,1);
F=F+sparse(it3,1,f,np,1);
```

```
% We have A*U=F.

% Assemble the boundary condition
[Q,G,H,R]=assemb('circleb1',p,e); % H*U=R

% Eliminate the Dirichlet boundary condition:
% Orthonormal basis for nullspace of H and its complement
[null,orth]=pdenullorth(H);
% Decompose U as U=null*UN+orth*UM. Then, from H*U=R, we have
% H*orth*UM=R which implies UM=(H*orth)\R.
% The linear system A*U=F becomes
%   null'*A*null*UN+null'*A*orth*((H*orth)\R)=null'*F
ud=full(orth*((H*orth)\R));
F=null'*(F-A*ud);
A=null'*A*null;
B=null;
```

10.1.3. A quick reference. Here is a brief table that tell you where to find help information on constructing geometries, writing boundary conditions, generating and refining meshes, and so on.

Decomposed geometry g that is specified by either a Decomposed Geometry matrix, or by a Geometry M-file:	See *decsg, pdegeom, initmesh.*
Boundary condition b that is specified by either a Boundary Condition matrix, or a Boundary M-file:	See *assemb, pdebound.*
Coefficients c, a, f:	See *assempde.*
Mesh structure p, e, t:	See *initmesh.*
Mesh generation:	See *initmesh.*
Mesh refinement:	See *refinemesh.*
Solvers:	See *assempde, adaptmesh, parabolic, hyperbolic, pdeeig, pdenonlin, ...*

Table 1. A brief reference for the PDE Toolbox.

10.2. Codes for Example 4.1—L-shaped domain problem on uniform meshes

10.2.1. The main script.

CODE 10.2. (L-shaped domain problem — uniform meshes)

```
% lshaped_uniform.m
% Solve Poisson equation -div(grad(u))=0 on the L-shaped
% membrane with Dirichlet boundary condition.
% The exact solution is ue(r,theta)=r^(2/3)*sin(2/3*theta).

% The exact solution and its partial derivatives
ue='(x.^2+y.^2).^(1/3).*sin(2/3*(atan2(y,x)+2*pi*(y<0)))';
uex='-2/3*(x.^2+y.^2).^(-1/6).*sin(1/3*(atan2(y,x)+2*pi*(y<0)))';
uey='2/3*(x.^2+y.^2).^(-1/6).*cos(1/3*(atan2(y,x)+2*pi*(y<0)))';

% Geometry
g = [2  2  2  2  2  2
     0  1  1 -1 -1  0
     1  1 -1 -1  0  0
     0  0  1  1 -1 -1
     0  1  1 -1 -1  0
     1  1  1  1  1  1
     0  0  0  0  0  0];

% Boundary conditions
r='(x.^2+y.^2).^(1/3).*sin(2/3*(atan2(y,x)+2*pi*(y<0)))';
b=[1 1 1 1 1 length(r) '0' '0' '1' r]';
b=repmat(b,1,6);

% PDE coefficients
c=1;
a=0;
f=0;

% Initial mesh
[p,e,t]=initmesh(g);

% Do iterative refinement, solve PDE, estimate the error.
```

```
error=[];
J=6;
for j=1:J
    u=assempde(b,p,e,t,c,a,f);
    err=pdeerrH1(p,t,u,ue,uex,uey);
    error=[error err];
    if j<J,
        [p,e,t]=refinemesh(g,p,e,t);
    end
end

% Plot the error versus 2^j in log-log coordinates and
% the reference line with slope -2/3.
n=2.^(0:J-1);
figure;
loglog(n,error,'k');
hold on;
loglog(n,error(end)*(n./n(end)).^(-2/3),'k:');
xlabel('2^j');
ylabel('H^1 error');
hold off;
```

10.2.2. H^1 **error.** The following function estimates the H^1 error of the linear finite element approximation.

CODE 10.3. (H^1 error)

```
function error=pdeerrH1(p,t,u,ue,uex,uey)
% Evaluate the H^1 error of "u"

it1=t(1,:); % Vertices of triangles.
it2=t(2,:);
it3=t(3,:);
% Areas, gradients of linear basis functions.
[ar,g1x,g1y,g2x,g2y,g3x,g3y]=pdetrg(p,t);
% The finite element approximation and its gradient
u=u.';
ux=u(it1).*g1x+u(it2).*g2x+u(it3).*g3x; % ux
uy=u(it1).*g1y+u(it2).*g2y+u(it3).*g3y; % uy
f=['(' ue '-(xi*u(it1)+eta*u(it2)+(1-xi-eta)*u(it3))).^2'];
```

```
err=quadgauss(p,t,f,u);
errx=quadgauss(p,t,['(' uex '-repmat(u,7,1)).^2'],ux);
erry=quadgauss(p,t,['(' uey '-repmat(u,7,1)).^2'],uy);
error=sqrt(err+errx+erry);
```

10.2.3. Seven-point Gauss quadrature rule. The following function integrates a function over a triangular mesh.

Code 10.4. (*Seven-point Gauss quadrature rule*)

```
function q=quadgauss(p,t,f,par)
% Integrate 'f' over the domain with triangulation 'p, t' using
% seven-point Gauss quadrature rule.
%
% q=quadgauss(p,t,f) evaluate the integral of 'f' where f can
% be a expression of x and y.
% q=quadgauss(p,t,f,par) evaluate the integral of 'f' where f
% can be a expression of x, y ,u, xi, and eta, where xi and eta
% are 7 by 1 vector such that (xi, eta, 1-xi-eta) gives the
% barycentric coordinates of the Gauss nodes. The parameter
% 'par' will be passed to 'u'. The evaluation of 'f' should
% give a matrix of 7 rows and size(t,2) columns.

% Nodes and weigths on the reference element
xi=[1/3;(6+sqrt(15))/21;(9-2*sqrt(15))/21;(6+sqrt(15))/21
     (6-sqrt(15))/21;(9+2*sqrt(15))/21;(6-sqrt(15))/21];
eta=[1/3;(6+sqrt(15))/21;(6+sqrt(15))/21;(9-2*sqrt(15))/21
        (6-sqrt(15))/21;(6-sqrt(15))/21;(9+2*sqrt(15))/21];
w=[9/80;(155+15^(1/2))/2400;(155-15^(1/2))/2400];

it1=t(1,:); % Vertices of triangles.
it2=t(2,:);
it3=t(3,:);
ar=pdetrg(p,t); % Areas of triangles
% Quadrature nodes on triangles
x=xi*p(1,it1)+eta*p(1,it2)+(1-xi-eta)*p(1,it3);
y=xi*p(2,it1)+eta*p(2,it2)+(1-xi-eta)*p(2,it3);
if nargin==4,
    u=par;
end
```

```
f=eval(f);
qt=2*ar.*(w(1)*f(1,:)+w(2)*sum(f(2:4,:))+w(3)*sum(f(5:7,:)));
q=sum(qt);
```

10.3. Codes for Example 4.6—L-shaped domain problem on adaptive meshes

Code 10.5. (L-shaped domain problem — adaptive meshes)

```
% Solve the L-shaped domain problem by the adaptive finite
% element algorithm based on the greedy strategy.
% See "lshaped_uniform.m" for a description of the L-shaped
% domain problem.

% Parameters for the a posteriori error estimates
alfa=0.15;beta=0.15;mexp=1;
J=17; % Maximum number of iterations.

% The exact solution and its partial derivatives
ue='(x.^2+y.^2).^(1/3).*sin(2/3*(atan2(y,x)+2*pi*(y<0)))';
uex='-2/3*(x.^2+y.^2).^(-1/6).*sin(1/3*(atan2(y,x)+2*pi*(y<0)))';
uey='2/3*(x.^2+y.^2).^(-1/6).*cos(1/3*(atan2(y,x)+2*pi*(y<0)))';

% Geometry
g = [2  2  2  2  2  2
     0  1  1 -1 -1  0
     1  1 -1 -1  0  0
     0  0  1  1 -1 -1
     0  1  1 -1 -1  0
     1  1  1  1  1  1
     0  0  0  0  0  0];

% Boundary conditions
b=[1 1 1 1 1 length(ue) '0' '0' '1' ue]';
b=repmat(b,1,6);

% PDE coefficients
c=1;
a=0;
f=0;
```

```
% Initial mesh
[p,e,t]=initmesh(g);

% Do iterative adaptive refinement, solve PDE,
% estimate the error.
error=[];
N_k=[];
for k=1:J+1
    fprintf('Number of triangles: %g\n',size(t,2))
    u=assempde(b,p,e,t,c,a,f);
    err=pdeerrH1(p,t,u,ue,uex,uey); % H^1 error
    error=[error err];
    N_k=[N_k,size(p,2)]; % DoFs
    if k<J+1,
        % A posteriori error estimate
        [cc,aa,ff]=pdetxpd(p,t,u,c,a,f);
        eta_k=pdejmps(p,t,cc,aa,ff,u,alfa,beta,mexp);
        % Mark triangles
        it=pdeadworst(p,t,cc,aa,ff,u,eta_k,0.5);
        tl=it';
        % Kludge: tl must be a column vector
        if size(tl,1)==1,
            tl=[tl;tl];
        end
        % Refine mesh
        [p,e,t]=refinemesh(g,p,e,t,tl);
    end
end

% Plot the error versus DOFs in log-log coordinates and
% the reference line with slope -1/2.
figure;
loglog(N_k,error,'k');
hold on;
loglog(N_k,error(end)*(N_k./N_k(end)).^(-1/2),'k:');
xlabel('DOFs');
ylabel('H^1 error');
```

```
hold off;
```

10.4. Implementation of the multigrid V-cycle algorithm

In this section, we first introduce the matrix versions of multigrid V-cycle algorithm 5.1 and the FMG algorithm 5.2, then provide the MATLAB codes for the FMG algorithm (fmg.m), the multigrid V-cycle iterator algorithm (mgp_vcycle.m), the V-cycle algorithm(mg_vcycle.m), and “newest vertex bisection” algorithm (refinemesh_mg.m). We remark that “mgp_vcycle.m” is an implementation of the adaptive multigrid V-cycle iterator algorithm that can be applied to adaptive finite element methods.

10.4.1. Matrix versions for the multigrid V-cycle algorithm and FMG. Recall that $\left\{\phi_k^1, \cdots, \phi_k^{n_k}\right\}$ is the nodal basis for V_k, we define the so called *prolongation matrix* $I_{k-1}^k \in \mathbb{R}^{n_k \times n_{k-1}}$ as follows

$$\phi_{k-1}^j = \sum_{i=1}^{n_k} (I_{k-1}^k)_{ij} \phi_k^i. \tag{10.8}$$

It follows from the definition (5.6) of $\widetilde{v}_k$ and $\widetilde{\widetilde{v}}_k$ that

$$\begin{aligned} \widetilde{v}_k &= I_{k-1}^k \widetilde{v}_{k-1} \quad \forall v_k = v_{k-1}, v_k \in V_k, v_{k-1} \in V_{k-1}, \\ \widetilde{\widetilde{Q_{k-1}r_k}} &= (I_{k-1}^k)^t \widetilde{\widetilde{r}}_k \quad \forall r_k \in V_k. \end{aligned} \tag{10.9}$$

Notice that $A_{k-1}v_{k-1} = Q_{k-1}A_k v_{k-1}, \forall v_{k-1} \in V_{k-1}$, we have

$$\widetilde{A}_{k-1}\widetilde{v}_{k-1} = \widetilde{\widetilde{A_{k-1}v_{k-1}}} = (I_{k-1}^k)^t \widetilde{\widetilde{A_k v_{k-1}}} = (I_{k-1}^k)^t \widetilde{A}_k I_{k-1}^k \widetilde{v}_{k-1},$$

that is,

$$\widetilde{A}_{k-1} = (I_{k-1}^k)^t \widetilde{A}_k I_{k-1}^k. \tag{10.10}$$

Algorithm 10.1. (Matrix version for V-cycle iterator). Let $\widetilde{\mathbb{B}}_1 = \widetilde{A}_1^{-1}$. Assume that $\widetilde{\mathbb{B}}_{k-1} \in \mathbb{R}^{n_{k-1} \times n_{k-1}}$ is defined, then $\widetilde{\mathbb{B}}_k \in \mathbb{R}^{n_k \times n_k}$ is defined as follows: Let $\widetilde{\widetilde{g}} \in \mathbb{R}^{n_k}$.

(1) Pre-smoothing: For $\widetilde{y}_0 = 0$ and $j = 1, \cdots, m$,

$$\widetilde{y}_j = \widetilde{y}_{j-1} + \widetilde{R}_k(\widetilde{\widetilde{g}} - \widetilde{A}_k \widetilde{y}_{j-1}).$$

(2) Coarse grid correction: $\widetilde{e} = \widetilde{\mathbb{B}}_{k-1}(I_{k-1}^k)^t(\widetilde{\widetilde{g}} - \widetilde{A}_k \widetilde{y}_m)$, $\widetilde{y}_{m+1} = \widetilde{y}_m + I_{k-1}^k \widetilde{e}$.

(3) Post-smoothing: For $j = m+2, \cdots, 2m+1$,

$$\widetilde{y}_j = \widetilde{y}_{j-1} + \widetilde{R}_k^t(\widetilde{\widetilde{g}} - \widetilde{A}_k \widetilde{y}_{j-1}).$$

Define $\widetilde{\mathbb{B}}_k \widetilde{\widetilde{g}} = \widetilde{y}_{2m+1}$.

Then the multigrid V-cycle iteration for (5.7) read as:

$$\widetilde{u}_k^{(n+1)} = \widetilde{u}_k^{(n)} + \widetilde{\mathbb{B}}_k(\widetilde{\widetilde{f}}_k - \widetilde{A}_k \widetilde{u}_k^{(n)}), \quad n = 0, 1, 2, \cdots, \tag{10.11}$$

Algorithm 10.2. (Matrix version for FMG).

For $k = 1, \widetilde{u}_1 = \widetilde{A}_1^{-1}\widetilde{\widetilde{f}}_1$.

For $k \geqslant 2$, let $\widetilde{u}_k = I_{k-1}^k \widetilde{u}_{k-1}$, and iterate $\widetilde{u}_k \leftarrow \widetilde{u}_k + \widetilde{\mathbb{B}}_k(\widetilde{\widetilde{f}}_k - \widetilde{A}_k \widetilde{u}_k)$ for l times.

10.4.2. Code for FMG. The following code is an implementation of the above FMG algorithm.

Code 10.6. (FMG)

```
function [u,p,e,t]=fmg(g,b,c,a,f,p0,e0,t0,nmg,nsm,nr)
% Full multigrid solver for "-div(c*grad(u))+a*u=f".
%
% "nmg": Number of multigrid iterations.
% "nsm": Number of smoothing iterations.
% "nr":  Number of refinements.

p=p0;
e=e0;
t=t0;
fprintf('k = %g. Number of triangles = %g\n',1,size(t,2));
[A,F,Bc,ud]=assempde(b,p,e,t,c,a,f);
u=Bc*(A\F)+ud;
I={};
for k=2:nr+1
    % Mesh and prolongation matrix
    [p,e,t,C]=refinemesh_mg(g,p,e,t);
    [A,F,Bf,ud]=assempde(b,p,e,t,c,a,f);
    fprintf('k = %g. Number of triangles = %g\n',k,size(t,2));
    I=[I {Bf'*C*Bc}];  % Eliminate the Dirichlet boundary nodes
    Bc=Bf;
    u=Bf'*C*u; % Initial value
    for l=1:nmg % Multigrid iteration
        r=F-A*u;
        Br=mgp_vcycle(A,r,I,nsm,k); % Multigrid precondtioner
        u=u+Br;
    end
    u=Bf*u+ud;
end
```

10.4.3. Code for the multigrid V-cycle algorithm. The following code is an implementation of Algorithm 10.1 for multigrid iterator.

CODE 10.7. (V-cycle iterator)

```
function Br=mgp_vcycle(A,r,I,m,k)
% Multigrid V-cycle precondtioner.
%
% "A" is stiffness matrix at level k,
% "I" is a cell of matrics such that: [I;I{k-1}]=I_{k-1}^k.
% "m" is the number of smoothing iterations. Br=B_k*r.

if(k==1),
    Br=A\r;
else
    Ik=I{k-1}; % Prolongation matrix
    [np,np1]=size(Ik);
    ns=find(sum(Ik)>1);
    ns=[ns, np1+1:np]; % Nodes to be smoothed.
    y=zeros(np,1);
    y=mgs_gs(A,r,y,ns,m); % Pre-smoothing
    r1=r-A(:,ns)*y(ns);
    r1=Ik'*r1;
    B=Ik'*A*Ik;
    Br1=mgp_vcycle(B,r1,I,m,k-1);
    y=y+Ik*Br1;
    y=mgs_gs(A,r,y,ns(end:-1:1),m); % Post-smoothing
    Br=y;
end
```

The following code is an implementation of the V-cycle algorithm (10.11).

CODE 10.8. (V-cycle)

```
function [u,steps]=mg_vcycle(A,F,I,u0,m,k,tol)
% Multigrid V-cycle iteration

if k==1,
    u=A\F;
    steps=1;
else
    u=u0;
    r=F-A*u;
    error0=max(abs(r));
    error=error0;
    steps=0;
    fprintf('Number of Multigrid iterations: ');
    while error>tol*error0,
        Br=mgp_vcycle(A,r,I,m,k); % Multigrid precondtioner
        u=u+Br;
        r=F-A*u;
        error=max(abs(r));
        steps=steps+1;
        for j=1:floor(log10(steps-0.5))+1,
            fprintf('\b');
        end
        fprintf('%g',steps);
    end
    fprintf('\n');
end
```

The following code is the Gauss-Seidel smoother.

Code 10.9. (Gauss-Seidel smoother)

```
function x=mgs_gs(A,r,x0,ns,m)
% (Local) Gauss-Seidel smoother for multigrid method.
%
% "A*x=r": The equation.
% "x0": The initial guess.
% "ns": The set of nodes to be smoothed.
% "m": Number of iterations.

A1=A(ns,ns);
```

```
ip=ones(size(r));
ip(ns)=0;
ip=ip==1;
A2=A(ns,ip);
y1=x0(ns);
y2=x0;
y2(ns)=[];
r1=r(ns)-A2*y2;
L=tril(A1);
U=triu(A1,1);
for k=1:m
    y1=L\(r1-U*y1);
end
x=x0;
x(ns)=y1;
```

10.4.4. The "newest vertex bisection" algorithm for mesh refinements. We first recall the "newest vertex bisection" algorithm for the mesh refinements which consists of two steps:

1. The marked triangles for refinements are bisected by the edge opposite to the newest vertex a fixed number of times (the newest vertex of an element in the initial mesh is the vertex opposite to the longest edge). The resultant triangulation may have nodes that are not the common vertices of two triangles. Such nodes are called *hanging nodes*.

2. All triangles with hanging nodes are bisected by the edge opposite to the newest vertex, this process is repeated until there are no hanging nodes.

It is known that the iteration in the second step to remove the hanging nodes can be completed in finite number of steps. An important property of the newest vertex bisection algorithm is that the algorithm generates a sequence of meshes that all the descendants of an original triangle fall into four similarity classes indicated in Figure 2. Therefore, let $\mathcal{M}_j, j = 1, 2, \cdots$, be a sequence of nested meshes generated by the newest vertex bisection algorithm, then there exists a constant $\theta > 0$ such that

$$\theta_K \geqslant \theta \quad \forall K \in \mathcal{M}_j, \quad j = 1, 2, \cdots, \tag{10.12}$$

where θ_K is the minimum angle of the element K.

Next, we provide a code for the "newest vertex bisection" algorithm for mesh refinements.

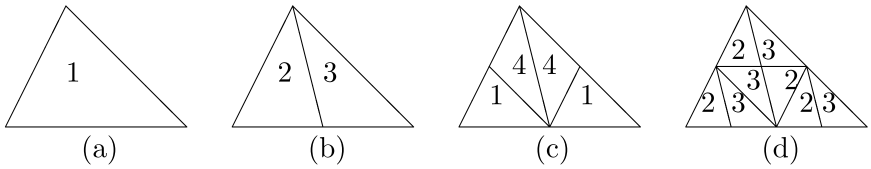

FIGURE 2. *Four similarity classes of triangles generated by "newest vertex bisection".*

CODE 10.10. (Newest vertex bisection)

```
function [p1,e1,t1,Icr]=refinemesh_mg(g,p,e,t,it)
% The "newest-vertex-bisection algorithm" for mesh refinements.
% Output also the prolongation matrix for multigrid iteration.
%
% G describes the geometry of the PDE problem. See either
% DECSG or PDEGEOM for details.
% The triangular mesh is given by the mesh data P, E, and T.
% Details can be found under INITMESH.
% The matrix Icr is the prolongation matrix from the coarse
% mesh to the fine mesh.
% 'it' is a list of triangles to be refined.
%
% This function is a modification of the 'refinemesh.m' from
% the MATLAB PDE Toolbox.

np=size(p,2);
nt=size(t,2);

if nargin==4,
  it=(1:nt)'; % All triangles
end

itt1=ones(1,nt);
itt1(it)=zeros(size(it));
it1=find(itt1);   % Triangles not yet to be refined
it=find(itt1==0); % Triangles whose side opposite to
                  % the newest vertex is to be bisected

% Make a connectivity matrix, with edges to be refined.
% -1 means no point is yet allocated
ip1=t(1,it);
ip2=t(2,it);
ip3=t(3,it);
A=sparse(ip1,ip2,-1,np,np)+sparse(ip2,ip3,-1,np,np)...
                          +sparse(ip3,ip1,-1,np,np);
A=-((A+A.')<0);
newpoints=1;

% Loop until no additional hanging nodes are introduced
while newpoints,
  newpoints=0;
  ip1=t(1,it1);
  ip2=t(2,it1);
  ip3=t(3,it1);
  m1 = aij(A,ip2,ip3);%A(ip2(i),ip3(i)), i=1:length(it1).
  m2 = aij(A,ip3,ip1);
  m3 = aij(A,ip1,ip2);
  ii=find(m3);
  if ~isempty(ii),
    itt1(it1(ii))=zeros(size(ii));
  end
  ii=find((m1 | m2) & (~m3));
  if ~isempty(ii),
    A=A+sparse(ip1(ii),ip2(ii),-1,np,np);
    A=-((A+A.')<0);
    newpoints=1;
    itt1(it1(ii))=zeros(size(ii));
  end
  it1=find(itt1);           % Triangles not yet fully refined
  it=find(itt1==0);         % Triangles fully refined
end

% Find edges to be refined
ie=(aij(A,e(1,:),e(2,:))==-1);

ie1=find(ie==0);            % Edges not to be refined
ie=find(ie);                % Edges to be refined
```

```
% Get the edge "midpoint" coordinates
[x,y]=pdeigeom(g,e(5,ie),(e(3,ie)+e(4,ie))/2);
% Create new points
p1=[p [x;y]];

% Prolongation matrix.
if nargout == 4,
  nie = length(ie);
  Icr = [sparse(1:nie,e(1,ie),1/2,nie,np)+...
             sparse(1:nie,e(2,ie),1/2,nie,np)];
end

ip=(np+1):(np+length(ie));
np1=np+length(ie);
% Create new edges
e1=[e(:,ie1) ...
        [e(1,ie);ip;e(3,ie);(e(3,ie)+e(4,ie))/2;e(5:7,ie)] ...
        [ip;e(2,ie);(e(3,ie)+e(4,ie))/2;e(4,ie);e(5:7,ie)]];
% Fill in the new points
A=sparse(e(1,ie),e(2,ie),ip+1,np,np)...
    +sparse(e(2,ie),e(1,ie),ip+1,np,np)+A;

% Generate points on interior edges
[i1,i2]=find(A==-1 & A.'==-1);
i=find(i2>i1);
i1=i1(i);
i2=i2(i);
p1=[p1 [(p(1:2,i1)+p(1:2,i2))/2]];

% Prolongation matrix.
if nargout == 4,
  ni=length(i);
  Icr = [Icr;sparse(1:ni,i1,1/2,ni,np)+...
                 sparse(1:ni,i2,1/2,ni,np)];
  Icr = [speye(size(Icr,2));Icr];
end
```

```
ip=(np1+1):(np1+length(i));
% Fill in the new points
A=sparse(i1,i2,ip+1,np,np)+sparse(i2,i1,ip+1,np,np)+A;

% Lastly form the triangles
ip1=t(1,it);
ip2=t(2,it);
ip3=t(3,it);
mp1 = aij(A,ip2,ip3); % A(ip2(i),ip3(i)), i=1:length(it).
mp2 = aij(A,ip3,ip1);
mp3 = aij(A,ip1,ip2);

% Find out which sides are refined
bm=1*(mp1>0)+2*(mp2>0);
% The number of new triangles
nnt1=length(it1)+length(it)+sum(mp1>0)+sum(mp2>0)+sum(mp3>0);
t1=zeros(4,nnt1);
t1(:,1:length(it1))=t(:,it1);      % The unrefined triangles
nt1=length(it1);
i=find(bm==3);       % All sides are refined
li=length(i); iti=it(i);
t1(:,(nt1+1):(nt1+li))=[t(1,iti);mp3(i);mp2(i);t(4,iti)];
nt1=nt1+length(i);
t1(:,(nt1+1):(nt1+li))=[mp3(i);t(2,iti);mp1(i);t(4,iti)];
nt1=nt1+length(i);
t1(:,(nt1+1):(nt1+li))=[t(3,iti);mp3(i);mp1(i);t(4,iti)];
nt1=nt1+length(i);
t1(:,(nt1+1):(nt1+li))=[mp3(i);t(3,iti);mp2(i);t(4,iti)];
nt1=nt1+li;
i=find(bm==2);       % Sides 2, 3 are refined
li=length(i); iti=it(i);
t1(:,(nt1+1):(nt1+li))=[t(1,iti);mp3(i);mp2(i);t(4,iti)];
nt1=nt1+length(i);
t1(:,(nt1+1):(nt1+li))=[t(2,iti);t(3,iti);mp3(i);t(4,iti)];
nt1=nt1+length(i);
t1(:,(nt1+1):(nt1+li))=[mp3(i);t(3,iti);mp2(i);t(4,iti)];
nt1=nt1+li;
i=find(bm==1);       % Sides 3 and 1 are refined
```

```
li=length(i); iti=it(i);
t1(:,(nt1+1):(nt1+li))=[mp3(i);t(2,iti);mp1(i);t(4,iti)];
nt1=nt1+li;
t1(:,(nt1+1):(nt1+li))=[t(3,iti);t(1,iti);mp3(i);t(4,iti)];
nt1=nt1+length(i);
t1(:,(nt1+1):(nt1+li))=[t(3,iti);mp3(i);mp1(i);t(4,iti)];
nt1=nt1+li;
i=find(bm==0);       % Side 3 is refined
li=length(i); iti=it(i);
t1(:,(nt1+1):(nt1+li))=[t(3,iti);t(1,iti);mp3(i);t(4,iti)];
nt1=nt1+li;
t1(:,(nt1+1):(nt1+li))=[t(2,iti);t(3,iti);mp3(i);t(4,iti)];
```

The following code “aij.c” in C language should be built into a MATLAB “mex” file that is used by the above function.

Code 10.11. (Find A(i,j))

```
/*=============================================================
 *
 * aij.c, aij.mex:
 *
 * The calling syntax is:
 *
 *      b = aij(A,vi,vj)
 *
 * where A should be a sparse matrix, vi and vj be integer
 * vertors. b is a row vector satisfying b_m=A(vi_m,vj_m).
 * This is a MEX-file for MATLAB.
 *
 *===========================================================*/
/* $Revision: 1.0 $ */
#include "mex.h"

/* Input Arguments */

#define A_IN    prhs[0]
#define vi_IN   prhs[1]
#define vj_IN   prhs[2]
```

```
/* Output Arguments */

#define b_OUT   plhs[0]

void mexFunction( int nlhs, mxArray *plhs[],
          int nrhs, const mxArray*prhs[] )
{
    double  *pr, *pi, *br, *bi, *vi, *vj;
    mwIndex  *ir, *jc;
    mwSize  ni, nj, l, row, col, k;

    /* Check for proper number of arguments */

    if (nrhs != 3) {
    mexErrMsgTxt("Three input arguments required.");
    } else if (nlhs > 1) {
    mexErrMsgTxt("Too many output arguments.");
    }

    ni = mxGetN(vi_IN)*mxGetM(vi_IN);
    nj = mxGetN(vj_IN)*mxGetM(vj_IN);
    if (ni != nj)
        mexErrMsgTxt("The lengths of vi and vi must be equal");

    pr = mxGetPr(A_IN);
    pi = mxGetPi(A_IN);
    ir = mxGetIr(A_IN);
    jc = mxGetJc(A_IN);
    vi = mxGetPr(vi_IN);
    vj = mxGetPr(vj_IN);

    if (!mxIsComplex(A_IN)){
    /* Create a matrix for the return argument */
        b_OUT = mxCreateDoubleMatrix(1, ni, mxREAL);

    /* Assign pointers to the various parameters */
        br = mxGetPr(b_OUT);
        for (l=0; l<ni; l++){
            row = *(vi+l);row--;
            col = *(vj+l);
            for (k=*(jc+col-1); k<*(jc+col); k++){
                if (*(ir+k)==row)
                    *(br+l) = *(pr+k);
            }
        }
    }else{
    /* Create a matrix for the return argument */
        b_OUT = mxCreateDoubleMatrix(1, ni, mxCOMPLEX);

    /* Assign pointers to the various parameters */
        br = mxGetPr(b_OUT);
        bi = mxGetPi(b_OUT);

        for (l=0; l<ni; l++){
            row = *(vi+l);row--;
            col = *(vj+l);
            for (k=*(jc+col-1); k<*(jc+col); k++){
                if (*(ir+k)==row){
                    *(br+l) = *(pr+k);
                    *(bi+l) = *(pi+k);
                }
            }
        }
    }
    return;
}
```

Bibliographic notes. The “newest vertex bisection” algorithm is introduced in Bänsch [**7**], Mitchell [**41**]. Further details of the algorithm can be found in Schmidt and Siebert [**48**].

10.5. Exercises

EXERCISE 10.1. Solve the following problem by the linear finite element method.

$$\begin{cases} -\Delta u = x, \quad -\infty < x < \infty, 0 < y < 1, \\ u(x,0) = u(x,1) = 0, \quad -\infty < x < \infty, \\ u(x,y) \text{ is periodic in the } x \text{ direction with period 1.} \end{cases}$$

EXERCISE 10.2. Solve the L-shaped domain problem in Example 4.1 by using the adaptive finite element algorithm base on the Dörfler marking strategy and verify the quasi-optimality of the algorithm.

EXERCISE 10.3. Solve the Poisson equation on the unit disk with homogeneous Dirichlet boundary condition by using the full muligrid algorithm 10.2 with Gauss-Seidel smoother and verify Theorem 5.5 numerically.

Bibliography

[1] R. Adams, *Sobolev Spaces*, Academic Press, 1975.

[2] C. Amrouche, C. Bernardi, M. Dauge, and V. Girault, *Vector potentials in three-diemnsional nonsmooth domains*, Math. Meth. Appl. Sci., 21 (1998), pp. 823–864.

[3] D. Arnold, R. Falk, and R. Winther, *Preconditioning in $h(\mathrm{div})$ and apploications*, Math. Comp., 66 (1997).

[4] M. Avellaneda and F.-H. Lin, *Compactness methods in the theory of homogenization*, Comm. Pure Appl. Math., 40 (1987), pp. 803–847.

[5] I. Babuška and A. Miller, *A feedback finite element method with a posteriori error estimation: Part i. the finite element method and some basic properties of the a posteriori error estimator*, Comput. Meth. Appl. Mech. Engrg., 61 (1987), pp. 1–40.

[6] I. Babuška and C. Rheinboldt, *Error estimates for adaptive finite element computations*, SIAM J. Numer. Anal., 15 (1978), pp. 736–754.

[7] E. Bänsch, *Local mesh refinement in 2 and 3 dimensions*, Impact of Computing in Science and Engineering, 3 (1991), pp. 181–191.

[8] R. Beck, R. Hiptmair, R. Hoppe, and B. Wohlmuth, *Residual based a posteriori error estimators for eddy current computation*, M2AN, 34 (2000), pp. 159–182.

[9] A. Bensoussan, J. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, 1978.

[10] M. Birman and M. Solomyak, *L^2-theory of the maxwell operator in arbitary domains*, Russian Math. Surveys, 43 (1987), pp. 75–96.

[11] D. Braess, *Finite Elements*, Cambridge University Press, Cambridge, 1997.

[12] J. Bramble, *Multigrid Methods*, no. 294 in Pitman Research Notes in Mathematical Sciences, Longman, Essex, 1993.

[13] J. Bramble and S. Hilbert, *Estimation of linear functionals on sobolev spaces with application to fourier transforms and spline inerpolation*, SIAM J. Numer. Anal., 7 (1970), pp. 112–124.

[14] A. Brandt, *Multi-level adaptive solutions to boundary-value problems*, Math. Comp., 31 (1977), pp. 333–390.

[15] ———, *Multigrid Techniques: 1984 Guide, with applications to uid dynamics*, Weizmann Institute of Science, Rehovot, Israel, 1984.

[16] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[17] A. Buffa, M. Costabel, and S. D., *On traces for $h(curl;\Omega)$ in Lipschitz domains*, J. Math. Anal. Appl., 276 (2002), pp. 845–867.

[18] J. M. Cascon, C. Kreuzer, R. Nochetto, and K. Siebert, *Quasi-optimal convergence rate for an adaptive finite element method*, SIAM J. Numer. Anal., 46 (2008), pp. 2524–2550.

[19] Z. Chen and T. Hou, *A mixed multiscale finite element method for elliptic problems with oscillating coefficients*, Math. Comp., 72 (2003), pp. 541–576.

[20] Z. Chen and F. Jia, *An adaptive finite element method with reliable and efficient error control for linear parabolic problems*, Math. Comp., 73 (2004), pp. 1163–1197.

[21] Z. Chen, L. Wang, and W. Zheng, *An adaptive multilevel method for time-harmonic maxwell equations with singularities*, SIAM J. Sci. Comput., 29 (2007), pp. 118–138.

[22] Z. Chen and X. Yue, *Numerical homogenization of well singularities in the flow transport through heterogeneous porous media*, Multiscale Modeling and Simulation, 1 (2003), pp. 260–303.

[23] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, vol. 4 of Studies in Mathematics and its Applications, North-Holland, New York, 1978.

[24] P. Clément, *Approximation by finite element functions using local regularization*, RARIO Numer. Anal., 2 (1975), pp. 77–84.

[25] M. Dauge, *Elliptic Bounary Value Problems on Corner Domains*, vol. 1341 of Lecture Notes in Mathematics, Springer, Berlin, 1988.

[26] J. Deny and J.-L. Lions, *Les espaces du type de deppo levi*, Ann. Inst. Fourier, Grenoble, 5 (1955), pp. 305–370.

[27] W. Dörfler, *A convergent adaptive algorithm for possion's equations*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.

[28] G. Duvaut and J.-L. Lions, *Les Inéquations en Mécanique et en Physique*, Dunod, 1972.

[29] Y. Efendiev, T. Hou, and X. Wu, *The convergence of non-conforming multiscale finite element methods*, SIAM J. Numer. Anal., 37 (2000), pp. 888–910.

[30] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems i: A linear model problem*, SIAM J. Numer. Anal., 28 (1991), pp. 43–77.

[31] ———, *Adaptive finite element methods for parabolic problems iv: Nonlinear problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1729–1749.

[32] L. Evans, *Partial Differential Equations*, vol. 19 of Graduate Studies in Mathematics, American Mathematical Society, Providence, 1998.

[33] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer, Berlin, 2001.

[34] V. Girault and A. Raviart, P, *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*, vol. 5 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 1986.

[35] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, London, 1985.

[36] R. Hiptmair, *Finite elements in computational electromagnetism*, Acta Numerica, 11 (2002), pp. 237–339.

[37] T. Hou and X. Wu, *A multiscale finite element method for elliptic problems in composite materials and porous media*, J. Comput. Phys., 134 (1997), pp. 169–189.

[38] T. Hou, X. Wu, and Z. Cai, *Convergence of a multiscale finite element method for elliptic problems with rapidly oscillating coefficients*, Math. Comp., 68 (1999), pp. 913–943.

[39] V. Jikov, S. Kozlov, and O. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer, Berlin, 1994.

[40] O. Ladyzhenskaya, V. Solonnikov, and N. Uraltseva, *Linear and Quasilinear Equations of Parabolic Type*, Americal Mathematical Society, Providence, 1968.

[41] W. Mitchell, *Optimal multilevel iterative methods for adaptive grids*, SIAM J. Sci. Stat. Comput., 13 (1992), pp. 146–167.

[42] P. Monk, *Finite Element Methods for Maxwell's Equations*, Clarendon Press, Oxford, 2003.

[43] J. Nédélec, *Mixed finite lements in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.

[44] ———, *A new family of mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 50 (1986), pp. 57–81.

[45] J. Nečas, *Sur une méthode pour résoudre les équations aux dérivées partielles du type elliptique, voisine de la variationnelle*, Ann. Sc. Norm. Sup. Pisa Sér. 3, 16 (1962), pp. 305–326.

[46] ———, *Equations aux Dérivées Partielles*, Presses de l'Université de Montréal, 1965.

[47] M. Picasso, *Adaptive finite elements for a linear parabolic problem*, Comput. Method Appl. Mech. Engrg., 167 (1998), pp. 223–237.

[48] A. Schmidt and K. Siebert, *Albert: An adaptive hierarchical finite element toolbox.* IAM, University of Freiburg, http://www.mathematik.uni-freiburg.de /IAM/Research/projectsdz/albert, 2000.

[49] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, no. 25, Springer-Verlag, New York, 1997.

[50] R. Verfürth, *A posteriori error estimates for the stokes equations*, Numer. Math., 55 (1989), pp. 309–325.

[51] ———, *A Review of A Posteriori Error Estimation and adaptive Mesh Refinement Techniques*, Teubner, 1996.

[52] H. Wu and Z. Chen, *Uniform convergence of multigrid v-cycle on adaptively refined meshes for second order elliptic problems*, Science on China, (Series A), 49 (2006), pp. 1405–1429.

[53] J. Xu, *Iterative methods by space decomposition and subspace correction*, SIAM Rev, 34 (1992), pp. 581–613.

Index