Finite Elements II: Galerkin Approximation, Elliptic and Mixed PDEs

Alexandre Ern, Jean-Luc Guermond

Finite Elements II:
Galerkin approximation, elliptic and mixed PDEs

Alexandre Ern   Jean-Luc Guermond

May 13, 2021

Contents

Part VII. Elliptic PDEs: conforming approximation

Part VIII. Elliptic PDEs: nonconforming approximation

Part IX. Vector-valued elliptic PDEs

Appendix

Chapter 24

Weak formulation of model problems

In Part V, composed of Chapters 24 and 25, we introduce the notion of weak formulations and state two well-posedness results: the Lax–Milgram lemma and the more fundamental Banach–Nečas–Babuška theorem. Weak formulations are useful for building finite element approximations to partial differential equations (PDEs). This chapter presents a step-by-step derivation of weak formulations. We start by considering a few simple PDEs posed over a bounded subset D of $\mathbb{R}^d$. Our goal is to reformulate these problems in weak form using the important notion of *test functions*. We show by examples that there are many ways to write weak formulations. Choosing one can be guided, e.g., by the smoothness of the data and the quantities of interest (e.g., the solution or its gradient). The reader who is not familiar with functional analysis arguments is invited to review the four chapters composing Part I before reading Part V.

24.1 A second-order PDE

Let D be a Lipschitz domain in $\mathbb{R}^d$ (see §3.1) and consider a function $f : D \to \mathbb{R}$. The problem we want to solve consists of seeking a function $u : D \to \mathbb{R}$ with some appropriate smoothness yet to be clearly defined such that

$$-\Delta u = f \ \text{ in } D \qquad u = 0 \ \text{ on } \partial D, \tag{24.1}$$

where the Laplace operator is defined by $\Delta u := \nabla \cdot (\nabla u)$. In Cartesian coordinates, we have $\Delta u := \sum_{i \in \{1:d\}} \frac{\partial^2 u}{\partial x_i^2}$.

The PDE $-\Delta u = f$ in D is called *Poisson equation* (and *Laplace equation* when $f = 0$). The Laplace operator is ubiquitous in physics since it is the prototypical operator modelling diffusion processes. Applications include heat transfer (where u is the temperature and f the heat source), mass transfer (where u is the concentration of a species and f the mass source), porous media flow (where u is the hydraulic head and f the mass source), electrostatics (where u is the electrostatic potential and f the charge density), and static equilibria of membranes (where u is the transverse membrane displacement and f the transverse load).

The condition enforced on ∂D in (24.1) is called *boundary condition*. A condition prescribing the value of the solution at the boundary is called *Dirichlet condition*, and when the prescribed

value is zero, the condition is called *homogeneous Dirichlet condition*. In the context of the above models, the Dirichlet condition means that the temperature (the concentration, the hydraulic head, the electrostatic potential, or the transverse membrane displacement) is prescribed at the boundary. Other boundary conditions can be prescribed for the Poisson equation, as reviewed in Chapter 31 in the more general context of second-order elliptic PDEs.

To sum up, (24.1) is the Poisson equation (or problem) with a homogeneous Dirichlet condition. We now present three weak formulations of (24.1).

24.1.1 First weak formulation

We derive a weak formulation of (24.1) by proceeding informally. Consider an arbitrary test function $\varphi \in C_0^\infty(D)$, where $C_0^\infty(D)$ is the space of infinitely differentiable functions compactly supported in D. As a first step, we multiply the PDE in (24.1) by φ and integrate over D to obtain

$$-\int_D (\Delta u)\varphi \, \mathrm{d}x = \int_D f\varphi \, \mathrm{d}x. \tag{24.2}$$

Equation (24.2) is equivalent to the PDE in (24.1) if Δu is smooth enough (e.g., integrable over D). Indeed, if an integrable function g satisfies $\int_D g\varphi \, \mathrm{d}x = 0$ for all $\varphi \in C_0^\infty(D)$, Theorem 1.32 implies that $g = 0$ a.e. in D.

As a second step, we use the *divergence formula* stating that for any smooth vector-valued function $\boldsymbol{\Phi}$,

$$\int_D \nabla\cdot\boldsymbol{\Phi} \, \mathrm{d}x = \int_{\partial D} \boldsymbol{\Phi}\cdot\boldsymbol{n} \, \mathrm{d}s, \tag{24.3}$$

where $\boldsymbol{n}$ is the outward unit normal to D. We apply this formula to the function $\boldsymbol{\Phi} := w\nabla v$, where v and w are two scalar-valued smooth functions. Since $\nabla\cdot\boldsymbol{\Phi} = \nabla w\cdot\nabla v + w\Delta v$, we infer that

$$-\int_D (\Delta v)w \, \mathrm{d}x = \int_D \nabla v\cdot\nabla w \, \mathrm{d}x - \int_{\partial D} (\boldsymbol{n}\cdot\nabla v)w \, \mathrm{d}s. \tag{24.4}$$

This is *Green's formula*, which is a very useful tool to derive weak formulations of PDEs involving the Laplace operator. This formula is valid for instance if $v \in C^2(D) \cap C^1(\overline{D})$ and $w \in C^1(D) \cap C^0(\overline{D})$, and it can be extended to functions in the usual Sobolev spaces. In particular, it remains valid for all $v \in H^2(D)$ and all $w \in H^1(D)$. We apply Green's formula to the functions $v := u$ and $w := \varphi$, assuming enough smoothness for u. Since φ vanishes at the boundary, we transform (24.2) into

$$\int_D \nabla u\cdot\nabla\varphi \, \mathrm{d}x = \int_D f\varphi \, \mathrm{d}x, \qquad \forall \varphi \in C_0^\infty(D). \tag{24.5}$$

We now recast (24.5) into a functional framework. Let us take $f \in L^2(D)$. We observe that a natural solution space is

$$H^1(D) := \{v \in L^2(D) \mid \nabla v \in \boldsymbol{L}^2(D)\}. \tag{24.6}$$

Recall from Proposition 2.9 that $H^1(D)$ is a Hilbert space when equipped with the inner product $(u,v)_{H^1(D)} := \int_D uv \, \mathrm{d}x + \ell_D^2 \int_D \nabla u\cdot\nabla v \, \mathrm{d}x$ with associated norm $\|v\|_{H^1(D)} := (\int_D v^2 \, \mathrm{d}x + \ell_D^2 \int_D \|\nabla v\|_{\ell^2}^2 \, \mathrm{d}x)^{\frac{1}{2}}$, where $\|\cdot\|_{\ell^2}$ denotes the Euclidean norm in $\mathbb{R}^d$ and ℓ_D is a length scale associated with the domain D, e.g., $\ell_D := \mathrm{diam}(D)$ (one can take $\ell_D := 1$ when working in nondimensional form). In order to account for the boundary condition in (24.1), we consider the subspace spanned by those functions in $H^1(D)$ that vanish at the boundary. It turns out that this space is $H_0^1(D)$; see Theorem 3.10. Finally, we can extend the space of the test functions in (24.5) to the closure of $C_0^\infty(D)$ in $H^1(D)$, which is by definition $H_0^1(D)$ (see Definition 3.9). To see this,

we consider any test function $w \in H_0^1(D)$, observe that there is a sequence $(\varphi_n)_{n \in \mathbb{N}}$ in $C_0^\infty(D)$ converging to w in $H_0^1(D)$, and pass to the limit in (24.5) with φ_n used as the test function. To sum up, a weak formulation of the Poisson equation with homogeneous Dirichlet condition is as follows:

$$\begin{cases} \text{Find } u \in V := H_0^1(D) \text{ such that} \\ \int_D \nabla u \cdot \nabla w \, \mathrm{d}x = \int_D f w \, \mathrm{d}x, \quad \forall w \in V. \end{cases} \tag{24.7}$$

A function u solving (24.7) is called *weak solution* to (24.1).

We now investigate whether a solution to (24.7) (i.e., a weak solution to (24.1)) satisfies the PDE and the boundary condition in (24.1). Similarly to Definition 2.3, we say that a vector-valued field $\boldsymbol{\sigma} \in \boldsymbol{L}^1_{\mathrm{loc}}(D) := L^1_{\mathrm{loc}}(D; \mathbb{R}^d)$ has a weak divergence $\psi \in L^1_{\mathrm{loc}}(D)$ if

$$\int_D \boldsymbol{\sigma} \cdot \nabla \varphi \, \mathrm{d}x = - \int_D \psi \varphi \, \mathrm{d}x, \qquad \forall \varphi \in C_0^\infty(D), \tag{24.8}$$

and we write $\nabla \cdot \boldsymbol{\sigma} := \psi$. The argument of Lemma 2.4 shows that the weak divergence of a vector-valued field, if it exists, is uniquely defined.

Proposition 24.1 (Weak solution). *Assume that u solves (24.7) with $f \in L^2(D)$. Then $-\nabla u$ has a weak divergence equal to f, the PDE in (24.1) is satisfied a.e. in D, and the boundary condition a.e. in ∂D.*

Proof. Let u be a weak solution. Then $\nabla u \in \boldsymbol{L}^2(D) \subset \boldsymbol{L}^1_{\mathrm{loc}}(D)$. Taking as a test function in (24.7) an arbitrary function $\varphi \in C_0^\infty(D) \subset H_0^1(D)$ and observing that $f \in L^2(D) \subset L^1_{\mathrm{loc}}(D)$, we infer from the definition (24.8) of the weak divergence that the vector-valued field $\boldsymbol{\sigma} := -\nabla u$ has a weak divergence equal to f. Hence, the PDE is satisfied in the sense that $-\nabla \cdot (\nabla u) = f$ in $L^2(D)$, i.e., both functions are equal a.e. in D. Since $u \in H_0^1(D)$, u vanishes a.e. in ∂D owing to the trace theorem (Theorem 3.10). $\square$

The crucial advantage of the weak formulation (24.7) with respect to the original formulation (24.1) is that, as we will see in the next chapter, there exist powerful tools that allow us to assert the existence and uniqueness of weak solutions. It is noteworthy that uniqueness is not a trivial property in spaces larger than $H^1(D)$, and existence is nontrivial in spaces smaller than $H^1(D)$. For instance, one can construct domains in which uniqueness does not hold in $L^2(D)$, and existence does not hold in $H^2(D)$; see Exercise 24.2.

24.1.2 Second weak formulation

To derive our second formulation, we introduce the vector-valued function $\boldsymbol{\sigma} := -\nabla u$. To avoid notational collisions, we use the letter p instead of u to denote the scalar-valued unknown function, and we use the symbol u to denote the pair $(\boldsymbol{\sigma}, p)$. In many applications, p plays the role of a potential and $\boldsymbol{\sigma}$ plays the role of a (diffusive) flux. More generally, p is called *primal variable* and $\boldsymbol{\sigma}$ *dual variable*.

Since $\boldsymbol{\sigma} = -\nabla p$ and $-\Delta p = f$, we obtain $\nabla \cdot \boldsymbol{\sigma} = f$. Therefore, the model problem is now written as follows:

$$\boldsymbol{\sigma} + \nabla p = 0 \text{ in } D, \qquad \nabla \cdot \boldsymbol{\sigma} = f \text{ in } D, \qquad p = 0 \text{ on } \partial D. \tag{24.9}$$

This is the *mixed formulation* of the original problem (24.1). The PDEs in (24.9) are often called *Darcy's equations* (in the context of porous media flows, p is the hydraulic head and $\boldsymbol{\sigma}$ the filtration velocity).

We multiply the first PDE in (24.9) by a vector-valued test function $\boldsymbol{\tau}$ and integrate over D to obtain

$$\int_D \boldsymbol{\sigma}{\cdot}\boldsymbol{\tau}\,\mathrm{d}x + \int_D \nabla p{\cdot}\boldsymbol{\tau}\,\mathrm{d}x = 0. \tag{24.10}$$

We multiply the second PDE in (24.9) by a scalar-valued test function q and integrate over D to obtain

$$\int_D (\nabla{\cdot}\boldsymbol{\sigma})q\,\mathrm{d}x = \int_D fq\,\mathrm{d}x. \tag{24.11}$$

No integration by parts is performed in this approach.

We now specify a functional framework. We consider $H^1(D)$ as the solution space for p (so that $\nabla p \in \boldsymbol{L}^2(D)$ and $p \in L^2(D)$), and $\boldsymbol{H}(\mathrm{div};D)$ as the solution space for $\boldsymbol{\sigma}$ with $\|\boldsymbol{\sigma}\|_{\boldsymbol{H}(\mathrm{div};D)} :=$ $(\|\boldsymbol{\sigma}\|_{\boldsymbol{L}^2(D)}^2 + \ell_D^2\|\nabla{\cdot}\boldsymbol{\sigma}\|_{L^2(D)}^2)^{\frac{1}{2}}$ (recall that ℓ_D is a characteristic length associated with D, e.g., $\ell_D := \mathrm{diam}(D)$). Moreover, we enforce the boundary condition explicitly by restricting p to be in the space $H_0^1(D)$. With this setting, the test function $\boldsymbol{\tau}$ can be taken in $\boldsymbol{L}^2(D)$ and the test function q in $L^2(D)$. To sum up, a second weak formulation is as follows:

$$\begin{cases} \text{Find } u := (\boldsymbol{\sigma}, p) \in V \text{ such that} \\ \int_D (\boldsymbol{\sigma}{\cdot}\boldsymbol{\tau} + \nabla p{\cdot}\boldsymbol{\tau} + (\nabla{\cdot}\boldsymbol{\sigma})q)\,\mathrm{d}x = \int_D fq\,\mathrm{d}x, \quad \forall w := (\boldsymbol{\tau}, q) \in W, \end{cases} \tag{24.12}$$

with the functional spaces $V := \boldsymbol{H}(\mathrm{div};D){\times}H_0^1(D)$ and $W := \boldsymbol{L}^2(D){\times}L^2(D)$. Note that the space where the solution is expected to be (trial space) differs from the space where the test functions are taken (test space).

Proposition 24.2 (Weak solution). *Assume that u solves* (24.12) *with $f \in L^2(D)$. Then the PDEs in* (24.9) *are satisfied a.e. in D, and the boundary condition a.e. in ∂D.*

Proof. Left as an exercise. □

24.1.3　Third weak formulation

We start with the mixed formulation (24.9), and we now perform an integration by parts on the term involving $\nabla{\cdot}\boldsymbol{\sigma}$. Proceeding informally, we obtain

$$-\int_D \boldsymbol{\sigma}{\cdot}\nabla q\,\mathrm{d}x + \int_{\partial D} (\boldsymbol{n}{\cdot}\boldsymbol{\sigma})q\,\mathrm{d}s = \int_D fq\,\mathrm{d}x. \tag{24.13}$$

We take the test function q in $H^1(D)$ for the first integral to make sense. Moreover, to eliminate the boundary integral, we restrict q to be in the space $H_0^1(D)$. Now the dual variable $\boldsymbol{\sigma}$ can be taken in $\boldsymbol{L}^2(D)$. To sum up, a third weak formulation is as follows:

$$\begin{cases} \text{Find } u := (\boldsymbol{\sigma}, p) \in V \text{ such that} \\ \int_D (\boldsymbol{\sigma}{\cdot}\boldsymbol{\tau} + \nabla p{\cdot}\boldsymbol{\tau} + \boldsymbol{\sigma}{\cdot}\nabla q)\,\mathrm{d}x = -\int_D fq\,\mathrm{d}x, \quad \forall w := (\boldsymbol{\tau}, q) \in V, \end{cases} \tag{24.14}$$

with the same functional space $V := \boldsymbol{L}^2(D){\times}H_0^1(D)$ for the trial and test spaces. The change of sign on the right-hand side has been introduced to make the left-hand side symmetric with respect to $(\boldsymbol{\sigma}, p)$ and $(\boldsymbol{\tau}, q)$.

Proposition 24.3. *Let u solve* (24.14) *with $f \in L^2(D)$. Then the PDEs in* (24.9) *are satisfied a.e. in D, and the boundary condition a.e. in ∂D.*

Proof. Left as an exercise. □

24.2   A first-order PDE

For simplicity, we consider a one-dimensional model problem (a more general setting is covered in Chapter 56). Let $D := (0, 1)$ and let $f : D \to \mathbb{R}$ be a smooth function. The problem we want to solve consists of seeking a function $u : D \to \mathbb{R}$ such that

$$u' = f \quad \text{in } D, \qquad u(0) = 0. \tag{24.15}$$

Proceeding informally, the solution to this problem is the function defined as follows:

$$u(x) := \int_0^x f(t)\, \mathrm{d}t, \qquad \forall x \in D. \tag{24.16}$$

To give a precise mathematical meaning to this statement, we assume that $f \in L^1(D)$, and we introduce the Sobolev space (see Definition 2.8)

$$W^{1,1}(D) := \{v \in L^1(D) \mid v' \in L^1(D)\}, \tag{24.17}$$

where as usual we interpret the derivatives in the weak sense.

Lemma 24.4 (Solution in $W^{1,1}(D)$). *If $f \in L^1(D)$, the problem* (24.15) *has a unique solution in $W^{1,1}(D)$ which is given by* (24.16).

Proof. Let u be defined in (24.16).
(1) Let us first show that $u \in C^0(\overline{D})$ (recall that $\overline{D} = [0, 1]$). Let $x \in \overline{D}$ and let $(x_n)_{n \in \mathbb{N}}$ be a sequence converging to x in $\overline{D}$. This gives

$$u(x) - u(x_n) = \int_0^x f(t)\, \mathrm{d}t - \int_0^{x_n} f(t)\, \mathrm{d}t = \int_{x_n}^x f(t)\, \mathrm{d}t = \int_D \mathbb{1}_{[x_n, x]}(t) f(t)\, \mathrm{d}t,$$

where $\mathbb{1}_{[x_n, x]}$ is the indicator function of the interval $[x_n, x]$. Since $\mathbb{1}_{[x_n, x]} f \to 0$ and $|\mathbb{1}_{[x_n, x]} f| \leq |f|$ a.e. in D, Lebesgue's dominated convergence theorem (Theorem 1.23) implies that $u(x_n) \to u(x)$. This shows that $u \in C^0(\overline{D})$. Hence, the boundary condition $u(0) = 0$ is meaningful.
(2) Let us now prove that $u' = f$ a.e. in D. One can verify (see Exercise 24.7) that

$$\int_0^1 \left(\int_0^x f(t)\, \mathrm{d}t \right) \varphi'(x)\, \mathrm{d}x = - \int_0^1 f(x)\varphi(x)\, \mathrm{d}x, \quad \forall \varphi \in C_0^\infty(D). \tag{24.18}$$

Since the left-hand side is equal to $\int_0^1 u(x)\varphi'(x)\, \mathrm{d}x$ and $f \in L^1(D) \subset L_{\mathrm{loc}}^1(D)$, we infer that u has a weak derivative in $L_{\mathrm{loc}}^1(D)$ equal to f. This implies that the PDE in (24.15) is satisfied a.e. in D.
(3) Uniqueness of the solution is a consequence of Lemma 2.11 since the difference of two weak solutions is constant on D (since it has zero weak derivative) and vanishes at $x = 0$. $\qquad \square$

We now present two possible mathematical settings for the weak formulation of the problem (24.15).

24.2.1   Formulation in $L^1(D)$

Since $f \in L^1(D)$ and $u \in W^{1,1}(D)$ with $u(0) = 0$, a first weak formulation is obtained by just multiplying the PDE in (24.15) by a test function w and integrating over D:

$$\int_D u'w\, \mathrm{d}t = \int_D fw\, \mathrm{d}t. \tag{24.19}$$

This equality is meaningful for all $w \in W^{(\infty)} := L^{\infty}(D)$. Moreover, the boundary condition $u(0) = 0$ can be explicitly enforced by considering the solution space $V^{(1)} := \{v \in W^{1,1}(D) \mid v(0) = 0\}$. Thus, a first weak formulation of (24.15) is as follows:

$$\begin{cases} \text{Find } u \in V^{(1)} \text{ such that} \\ \int_D u'w \, \mathrm{d}t = \int_D fw \, \mathrm{d}t, \quad \forall w \in W^{(\infty)}. \end{cases} \tag{24.20}$$

Remark 24.5 (Literature). Solving first-order PDEs using L^1-based formulations has been introduced by Lavery [276, 277]; see also Guermond [227], Guermond and Popov [228], and the references therein. □

24.2.2 Formulation in $L^2(D)$

Although the weak formulation (24.20) gives a well-posed problem (as we shall see in §25.4.2), the dominant viewpoint in the literature consists of using L^2-based formulations. This leads us to consider a second weak formulation where the source term f has slightly more smoothness, i.e., $f \in L^2(D)$ instead of just $f \in L^1(D)$, thereby allowing us to work in a Hilbertian setting. Since $L^2(D) \subset L^1(D)$, we have $f \in L^1(D)$, and we can still consider the function u defined in (24.16). This function turns out to be in $H^1(D)$ if $f \in L^2(D)$. Indeed, the Cauchy–Schwarz inequality and Fubini's theorem imply that

$$\int_0^1 |u(x)|^2 \, \mathrm{d}x = \int_0^1 \left| \int_0^x f(t) \, \mathrm{d}t \right|^2 \mathrm{d}x \leq \int_0^1 \left(\int_0^x |f(t)|^2 \, \mathrm{d}t \right) x \, \mathrm{d}x$$
$$= \int_0^1 \left(\int_t^1 \mathrm{d}x \right) |f(t)|^2 dt = \int_0^1 (1-t)|f(t)|^2 dt \leq \int_0^1 |f(t)|^2 dt,$$

which shows that $\|u\|_{L^2(D)} \leq \|f\|_{L^2(D)}$. Moreover, $\|u'\|_{L^2(D)} = \|f\|_{L^2(D)}$. Hence, $u \in H^1(D)$. We can then restrict the test functions to the Hilbert space $W^{(2)} := L^2(D)$ and use the Hilbert space $V^{(2)} := \{v \in H^1(D) \mid v(0) = 0\}$ as the solution space. Thus, a second weak formulation of (24.20), provided $f \in L^2(D)$, is as follows:

$$\begin{cases} \text{Find } u \in V^{(2)} \text{ such that} \\ \int_D u'w \, \mathrm{d}t = \int_D fw \, \mathrm{d}t, \quad \forall w \in W^{(2)}. \end{cases} \tag{24.21}$$

The main change with respect to (24.20) is in the trial and test spaces.

24.3 A complex-valued model problem

Some model problems are formulated using complex-valued functions. A salient example is Maxwell's equations in the time-harmonic regime; see §43.1. For simplicity, let us consider here the PDE

$$\mathrm{i}u - \nu \Delta u = f \quad \text{in } D, \tag{24.22}$$

with $u : D \to \mathbb{C}$, $f : D \to \mathbb{C}$, $\mathrm{i}^2 = -1$, and a real number $\nu > 0$. To fix the ideas, we enforce a homogeneous Dirichlet condition on u at the boundary.

When working with complex-valued functions, one uses the complex conjugate of the test function in the weak problem, i.e., the starting point of the weak formulation is the identity

$$\int_D \mathrm{i}u\overline{w} \, \mathrm{d}x + \nu \int_D \nabla u \cdot \nabla \overline{w} \, \mathrm{d}x = \int_D f\overline{w} \, \mathrm{d}x. \tag{24.23}$$

One can then proceed as in §24.1.1 (for instance). The functional setting uses the functional space $V := H_0^1(D;\mathbb{C})$, and the weak formulation is as follows:

$$\begin{cases} \text{Find } u \in V \text{ such that} \\ \int_D \mathrm{i}u\overline{w}\,\mathrm{d}x + \nu \int_D \nabla u \cdot \nabla\overline{w}\,\mathrm{d}x = \int_D f\overline{w}\,\mathrm{d}x, \quad \forall w \in V. \end{cases} \tag{24.24}$$

Proposition 24.1 is readily adapted to this setting.

The reason for using the complex conjugate of test functions is that it allows us to infer positivity properties on the real and imaginary parts of the quantity $a(u,w) := \int_D \mathrm{i}u\overline{w}\,\mathrm{d}x + \nu \int_D \nabla u \cdot \nabla\overline{w}\,\mathrm{d}x$ by taking $w := u$ as the test function. Indeed, we obtain

$$a(u,u) = \mathrm{i}\int_D |u|^2\,\mathrm{d}x + \nu\int_D \|\nabla u\|_{\ell^2(\mathbb{C}^d)}^2\,\mathrm{d}x = \mathrm{i}\|u\|_{L^2(D;\mathbb{C})}^2 + \nu\|\nabla u\|_{L^2(D;\mathbb{C}^d)}^2.$$

This means that $\Re(a(u,u)) = \nu\|\nabla u\|_{L^2(D;\mathbb{C}^d)}^2$ and $\Im(a(u,u)) = \|u\|_{L^2(D;\mathbb{C})}^2$. These results imply that

$$\Re(e^{-\mathrm{i}\frac{\pi}{4}}a(u,u)) \geq \frac{1}{\sqrt{2}}\min(1,\nu\ell_D^{-2})\|u\|_{H^1(D;\mathbb{C})}^2, \tag{24.25}$$

where we recall that the Hilbert space $L^2(D;\mathbb{C})$ is equipped with the inner product $(v,w)_{L^2(D)} := \int_D v\overline{w}\,\mathrm{d}x$ and the Hilbert space $H^1(D;\mathbb{C})$ is equipped with the inner product $(v,w)_{H^1(D)} := \int_D v\overline{w}\,\mathrm{d}x + \ell_D^2\int_D \nabla v \cdot \nabla\overline{w}\,\mathrm{d}x$, where ℓ_D is a characteristic length associated with D, e.g., $\ell_D := \mathrm{diam}(D)$.

24.4 Toward an abstract model problem

We conclude this chapter by casting all of the above weak formulations into a unified setting. We consider complex-valued functions since it is in general simpler to go from complex to real numbers than the other way around. Whenever relevant, we indicate the (minor) changes to apply in this situation (apart from replacing $\mathbb{C}$ by $\mathbb{R}$).

The above weak formulations fit into the following abstract model problem:

$$\begin{cases} \text{Find } u \in V \text{ such that} \\ a(u,w) = \ell(w), \quad \forall w \in W, \end{cases} \tag{24.26}$$

with maps $a : V \times W \to \mathbb{C}$ and $\ell : W \to \mathbb{C}$, where V, W are complex vector spaces whose elements are functions defined on D. V is called *trial space* or *solution space*, and W is called *test space*. Members of V are called *trial functions* and members of W are called *test functions*. The maps a and ℓ are called *forms* since their codomain is $\mathbb{C}$ (or $\mathbb{R}$ in the real case).

Recall that a map $A : V \to \mathbb{C}$ is said to be *linear* if $A(v_1 + v_2) = A(v_1) + A(v_2)$ for all $v_1, v_2 \in V$ and $A(\lambda v) = \lambda A(v)$ for all $\lambda \in \mathbb{C}$ and all $v \in V$, whereas a map $B : W \to \mathbb{C}$ is said to be *antilinear* if $B(w_1 + w_2) = B(w_1) + B(w_2)$ for all $w_1, w_2 \in W$ and $B(\lambda w) = \overline{\lambda}B(w)$ for all $\lambda \in \mathbb{C}$ and all $w \in W$. Then ℓ in (24.26) is an *antilinear form*, whereas a is a *sesquilinear form* (that is, the map $a(\cdot,w)$ is linear for all $w \in W$, and the map $a(v,\cdot)$ is antilinear for all $v \in V$). In the real case, ℓ is a *linear form* and a is a *bilinear form* (that is, it is linear in each of its arguments).

Remark 24.6 (Linearity). The linearity of a w.r.t. to its first argument is a consequence of the linearity of the problem, whereas the (anti)linearity of a w.r.t. its second argument results from the weak formulation. $\square$

Remark 24.7 (Bilinearity). Bilinear forms and linear forms on $V \times W$ are different objects. For instance, the action of a linear form on $(v, 0) \in V \times W$ is not necessarily zero, whereas $a(v, 0) = 0$ if a is a bilinear form. □

Remark 24.8 (Test functions). The role of the test functions in the weak formulations (24.20) and (24.26) are somewhat different. Since $L^\infty(D)$ is the dual space of $L^1(D)$ (the reverse is not true), the test functions $w \in L^\infty(D)$ in (24.20) act on the function $f \in L^1(D)$. Hence, in principle it should be more appropriate to write $w(\ell)$ instead of $\ell(w)$ in (24.26). Although this alternative viewpoint is not often considered in the literature, it actually allows for a more general setting regarding well-posedness. We return to this point in §25.3.2. This distinction is not relevant for model problems set in a Hilbertian framework. □

Exercises

Exercise 24.1 (Forms). Let $D := (0, 1)$. Which of these maps are linear or bilinear forms on $L^2(D) \times L^2(D)$: $a_1(f, g) := \int_D (f + g + 1) \, dx$, $a_2(f, g) := \int_D x(f - g) \, dx$, $a_3(f, g) := \int_D (1 + x^2) f g \, dx$, $a_4(f, g) := \int_D (f + g)^2 \, dx$?

Exercise 24.2 ((Non)-uniqueness). Consider the domain D in $\mathbb{R}^2$ whose definition in polar coordinates is $D := \{(r, \theta) \mid r \in (0, 1), \theta \in (\frac{\pi}{\alpha}, 0)\}$ with $\alpha \in (-1, -\frac{1}{2})$. Let $\partial D_1 := \{(r, \theta) \mid r = 1, \theta \in (\frac{\pi}{\alpha}, 0)\}$ and $\partial D_2 := \partial D \backslash \partial D_1$. Consider the PDE $-\Delta u = 0$ in D with the Dirichlet conditions $u = \sin(\alpha \theta)$ on ∂D_1 and $u = 0$ on ∂D_2. (i) Let $\varphi_1 := r^\alpha \sin(\alpha \theta)$ and $\varphi_2 := r^{-\alpha} \sin(\alpha \theta)$. Prove that φ_1 and φ_2 solve the above problem. (*Hint*: in polar coordinates $\Delta \varphi = \frac{1}{r} \partial_r(r \partial_r \varphi) + \frac{1}{r^2} \partial_{\theta\theta} \varphi$.) (ii) Prove that φ_1 and φ_2 are in $L^2(D)$ if $\alpha \in (-1, -\frac{1}{2})$. (iii) Consider the problem of seeking $u \in H^1(D)$ s.t. $u = \sin(\alpha \theta)$ on ∂D_1, $u = 0$ on ∂D_2, and $\int_D \nabla u \cdot \nabla v = 0$ for all $v \in H_0^1(D)$. Prove that φ_2 solves this problem, but φ_1 does not. Comment.

Exercise 24.3 (Poisson in 1D). Let $D := (0, 1)$ and $f(x) := \frac{1}{x(1-x)}$. Consider the PDE $-\partial_x((1 + \sin(x)^2)\partial_x u) = f$ in D with the Dirichlet conditions $u(0) = u(1) = 0$. Write a weak formulation of this problem with both trial and test spaces equal to $H_0^1(D)$ and show that the linear form on the right-hand side is bounded on $H_0^1(D)$. (*Hint*: notice that $f(x) = \frac{1}{x} + \frac{1}{1-x}$.)

Exercise 24.4 (Weak formulations). Prove Propositions 24.2 and 24.3.

Exercise 24.5 (Darcy). (i) Derive another variation on (24.12) and (24.14) with the functional spaces $V = W := \boldsymbol{H}(\mathrm{div}; D) \times L^2(D)$. (*Hint*: use Theorem 4.15.) (ii) Derive yet another variation with the functional spaces $V := \boldsymbol{L}^2(D) \times L^2(D)$ and $W := \boldsymbol{H}(\mathrm{div}; D) \times H_0^1(D)$.

Exercise 24.6 (Variational formulation). Prove that u solves (24.7) if and only if u minimizes over $H_0^1(D)$ the energy functional

$$\mathfrak{E}(v) := \frac{1}{2} \int_D |\nabla v|^2 \, dx - \int_D f v \, dx.$$

(*Hint*: show first that $\mathfrak{E}(v + tw) = \mathfrak{E}(v) + t \left\{ \int_D \nabla v \cdot \nabla w \, dx - \int_D f w \, dx \right\} + \frac{1}{2} t^2 \int_D |\nabla w|^2 \, dx$ for all $v, w \in H_0^1(D)$ and all $t \in \mathbb{R}$.)

Exercise 24.7 (Derivative of primitive). Prove (24.18). (*Hint*: use Theorem 1.38 and Lebesgue's dominated convergence theorem.)

Exercise 24.8 (Biharmonic problem). Let D be an open, bounded, set in $\mathbb{R}^d$ with smooth boundary. Derive a weak formulation for the biharmonic problem

$$\Delta(\Delta u) = f \ \text{ in } D, \qquad u = \partial_n u = 0 \ \text{ on } \partial D,$$

with $f \in L^2(D)$. (*Hint*: use Theorem 3.16.)

Chapter 25

Main results on well-posedness

The starting point of this chapter is the model problem derived in §24.4. Our goal is to specify conditions under which this problem is well-posed. Two important results are presented: the Lax–Milgram lemma and the more fundamental Banach–Nečas–Babuška theorem. The former provides a *sufficient* condition for well-posedness, whereas the latter, relying on slightly more sophisticated assumptions, provides *necessary and sufficient* conditions. The reader is invited to review the material of Appendix C on bijective operators in Banach spaces before reading this chapter.

25.1 Mathematical setting

To stay general, we consider complex vector spaces. The case of real vector spaces is recovered by replacing the field $\mathbb{C}$ by $\mathbb{R}$, by removing the real part symbol $\Re(\cdot)$ and the complex conjugate symbol $\bar{\cdot}$, and by interpreting the symbol $|\cdot|$ as the absolute value instead of the modulus.

We consider the following model problem:

$$\begin{cases} \text{Find } u \in V \text{ such that} \\ a(u, w) = \ell(w), \quad \forall w \in W. \end{cases} \tag{25.1}$$

The spaces V and W are complex Banach spaces equipped with norms denoted by $\|\cdot\|_V$ and $\|\cdot\|_W$, respectively. In many applications, V and W are Hilbert spaces. The map $a : V \times W \to \mathbb{C}$ is a sesquilinear form (bilinear in the real case). We assume that a is *bounded*, which means that

$$\|a\|_{V \times W} := \sup_{v \in V} \sup_{w \in W} \frac{|a(v, w)|}{\|v\|_V \|w\|_W} < \infty. \tag{25.2}$$

It is henceforth implicitly understood that this type of supremum is taken over nonzero arguments (notice that the order in which the suprema are taken in (25.2) does not matter). Furthermore, the map $\ell : W \to \mathbb{C}$ is an antilinear form (linear in the real case). We assume that ℓ is *bounded*, and we write $\ell \in W'$. The boundedness of ℓ means that

$$\|\ell\|_{W'} := \sup_{w \in W} \frac{|\ell(w)|}{\|w\|_W} < \infty. \tag{25.3}$$

Notice that it is possible to replace the modulus by the real part in (25.2) and (25.3) (replace w by ξw with a unitary complex number ξ), and in the real case, the absolute value is not needed (replace w by $\pm w$).

Definition 25.1 (Well-posedness, Hadamard [236]). *We say that the problem* (25.1) *is* well-posed *if it admits one and only one solution for all* $\ell \in W'$, *and there is* c, *uniform with respect to* ℓ, *s.t. the a priori estimate* $\|u\|_V \leq c \|\ell\|_{W'}$ *holds true.*

The goal of this chapter is to study the well-posedness of (25.1). The key idea is to introduce the bounded linear operator $A \in \mathcal{L}(V; W')$ that is naturally associated with the bilinear form a on $V \times W$ by setting

$$\langle A(v), w \rangle_{W', W} := a(v, w), \qquad \forall (v, w) \in V \times W. \tag{25.4}$$

This definition implies that A is linear and bounded with norm $\|A\|_{\mathcal{L}(V;W')} = \|a\|_{V \times W}$. The problem (25.1) can be reformulated as follows: Find $u \in V$ such that $A(u) = \ell$ in W'. Hence, proving the existence and uniqueness of the solution to (25.1) amounts to proving that the operator A is bijective. Letting $A^* : W'' \to V'$ be the adjoint of A, the way to do this is to prove the following three conditions:

$$\overbrace{\underbrace{\text{(i) } A \text{ is injective,}}_{\Longleftrightarrow\, \exists \alpha > 0,\, \|A(v)\|_{W'} \geq \alpha \|v\|_V,\, \forall v \in V} \quad \underbrace{\text{(ii) } \operatorname{im}(A) \text{ is closed,}}_{} \quad \text{(iii) } A^* \text{ is injective}}^{\Longleftrightarrow\, A \text{ is surjective}}. \tag{25.5}$$

The conditions (ii)-(iii) in (25.5) are equivalent to A being surjective since the closure of $\operatorname{im}(A)$ is $(\ker(A^*))^\perp \subset W'$ owing to Lemma C.34 (see also (C.14b)). That the conditions (i)-(ii) are equivalent to the existence of some $\alpha > 0$ s.t. $\|A(v)\|_{W'} \geq \alpha \|v\|_V$, for all $v \in V$, is established in Lemma C.39 (these two conditions are also equivalent to the surjectivity of A^*).

25.2   Lax–Milgram lemma

The Lax–Milgram lemma is applicable only if the solution and the test spaces are *identical*. Assuming $W = V$, the model problem (25.1) becomes

$$\begin{cases} \text{Find } u \in V \text{ such that} \\ a(u, w) = \ell(w), \quad \forall w \in V. \end{cases} \tag{25.6}$$

Lemma 25.2 (Lax–Milgram). *Let* V *be a Hilbert space, let* a *be a bounded sesquilinear form on* $V \times V$, *and let* $\ell \in V'$. *Assume the following* coercivity *property: There is a real number* $\alpha > 0$ *and a complex number* ξ *with* $|\xi| = 1$ *such that*

$$\Re\left(\xi a(v, v)\right) \geq \alpha \|v\|_V^2, \qquad \forall v \in V. \tag{25.7}$$

Then (25.6) *is well-posed with the a priori estimate* $\|u\|_V \leq \frac{1}{\alpha} \|\ell\|_{V'}$.

Proof. Although this lemma is a consequence of the more abstract BNB theorem (Theorem 25.9), we present a direct proof for completeness. Let $A : V \to V'$ be the bounded linear operator defined in (25.4) and let us prove the three conditions (i)-(ii)-(iii) in (25.5). Since $\xi a(v, v) = a(v, \overline{\xi}v)$, the coercivity property (25.7) implies that

$$\alpha \|v\|_V \leq \frac{\Re(a(v, \overline{\xi}v))}{\|v\|_V} \leq \sup_{w \in V} \frac{\Re(a(v, \overline{\xi}w))}{\|w\|_V} = \sup_{w \in V} \frac{|a(v, w)|}{\|w\|_V} = \|A(v)\|_{V'},$$

so that the conditions (i)-(ii) hold true. Since V is reflexive, we identify V and V'', so that the adjoint operator $A^* : V \to V'$ is such that $\langle A^*(v), w \rangle_{V', V} = \overline{\langle A(w), v \rangle_{V', V}}$ for all $v, w \in V$. Let

$v \in V$ and assume that $A^*(v) = 0$. Then $0 = \overline{0} = \overline{\langle A^*(v), \xi v \rangle_{V',V}} = \xi a(v,v)$. We then infer from (25.7) that $\alpha \|v\|_V^2 \leq \Re(\xi a(v,v)) = 0$, i.e., $v = 0$. This proves that A^* is injective. Hence, the condition (iii) also holds true. Finally, the a priori estimate follows from $\alpha \|u\|_V \leq \frac{\Re(a(u,\overline{\xi}u))}{\|u\|_V} = \frac{\Re(\ell(\overline{\xi}u))}{\|u\|_V} \leq \|\ell\|_{V'}$. $\qquad\square$

Remark 25.3 (Hilbertian setting). An important observation is that the Lax–Milgram lemma relies on the notion of coercivity which is applicable only in Hilbertian settings; see Proposition C.59. $\qquad\square$

Example 25.4 (Laplacian). Consider the weak formulation (24.7) of the Poisson equation with homogeneous Dirichlet condition. The functional setting is $V = W := H_0^1(D)$ equipped with the norm $\|\cdot\|_{H^1(D)}$, the bilinear form is $a(v,w) := \int_D \nabla v \cdot \nabla w \, dx$, and the linear form is $\ell(w) := \int_D fw \, dx$. Owing to the Cauchy–Schwarz inequality, the forms a and ℓ are bounded on $V \times V$ and V, respectively. Moreover, the Poincaré–Steklov inequality (3.11) (with $p := 2$) implies that (see Remark 3.29)

$$a(v,v) = \|\nabla v\|_{\boldsymbol{L}^2(D)}^2 = |v|_{H^1(D)}^2 \geq \ell_D^{-2} \frac{C_{\mathrm{PS}}^2}{1 + C_{\mathrm{PS}}^2} \|v\|_{H^1(D)}^2,$$

for all $v \in V$. Hence, (25.7) holds true with $\alpha := \ell_D^{-2} \frac{C_{\mathrm{PS}}^2}{1+C_{\mathrm{PS}}^2}$ and $\xi := 1$, and by the Lax–Milgram lemma, the problem (24.7) is well-posed. Alternatively one can equip V with the norm $\|v\|_V := \ell_D^{-1} \|\nabla v\|_{\boldsymbol{L}^2(D)}$ which is equivalent to the norm $\|\cdot\|_{H^1(D)}$ owing to the Poincaré–Steklov inequality. The coercivity constant of a is then $\alpha := \ell_D^{-2}$. $\qquad\square$

Example 25.5 (Complex case). Consider the PDE $\mathrm{i}u - \nu\Delta u = f$ in D with $\mathrm{i}^2 = -1$, a real number $\nu > 0$, a source term $f \in L^2(D;\mathbb{C})$, and a homogeneous Dirichlet condition. The functional setting is $V = W := H_0^1(D;\mathbb{C})$ equipped with the norm $\|\cdot\|_{H^1(D;\mathbb{C})}$, the sesquilinear form is $a(v,w) := \int_D \mathrm{i}v\overline{w} \, dx + \nu \int_D \nabla v \cdot \nabla \overline{w} \, dx$, and the antilinear form is $\ell(w) := \int_D f\overline{w} \, dx$. Then (24.25) shows that the coercivity property (25.7) holds true with $\xi := e^{-\mathrm{i}\frac{\pi}{4}}$ and $\alpha := \frac{1}{\sqrt{2}} \min(1, \nu\ell_D^{-2})$. $\qquad\square$

Remark 25.6 (Definition of coercivity). The coercivity property can also be defined in the following way: There is a real number $\alpha > 0$ such that $|a(v,v)| \geq \alpha \|v\|_V^2$ for all $v \in V$. It is shown in Lemma C.58 that this definition and (25.7) are equivalent. $\qquad\square$

Definition 25.7 (Hermitian/symmetric form). *Let V be a Hilbert space. In the complex case, we say that a sesquilinear form $a : V \times V \to \mathbb{C}$ is* Hermitian *whenever $a(v,w) = \overline{a(w,v)}$ for all $v, w \in V$. In the real case, we say that a bilinear form a is* symmetric *whenever $a(v,w) = a(w,v)$ for all $v, w \in V$.*

Whenever the sesquilinear form a is Hermitian and coercive (with $\xi := 1$ for simplicity), setting $((\cdot,\cdot))_V := a(\cdot,\cdot)$ one defines an inner product in V, and the induced norm is equivalent to $\|\cdot\|_V$ owing to the coercivity and the boundedness of a. Then solving the problem (25.6) amounts to finding the representative $u \in V$ of the linear form $\ell \in V'$, i.e., $((u,w))_V = \ell(w)$ for all $w \in V$. This problem is well-posed by the Riesz–Fréchet theorem (Theorem C.24). Thus, the Lax–Milgram lemma can be viewed as an extension of the Riesz–Fréchet theorem to non-Hermitian forms.

Whenever V is a real Hilbert space and the bilinear form a is symmetric and coercive with $\xi := 1$, the problem (25.6) can be interpreted as a minimization problem (or a maximization problem if $\xi := -1$). In this context, (25.6) is called *variational formulation*.

Proposition 25.8 (Variational formulation). *Let V be a real Hilbert space, let a be a bounded bilinear form on $V \times V$, and let $\ell \in V'$. Assume that a is coercive with $\xi := 1$. Assume that a is symmetric, i.e.,*

$$a(v,w) = a(w,v), \qquad \forall v, w \in V. \tag{25.8}$$

Then introducing the energy functional $\mathfrak{E} : V \to \mathbb{R}$ *such that*

$$\mathfrak{E}(v) := \frac{1}{2}a(v,v) - \ell(v), \tag{25.9}$$

u solves (25.6) *iff u minimizes* $\mathfrak{E}$ *over V.*

Proof. The proof relies on the fact that for all $u, w \in V$ and all $t \in \mathbb{R}$,

$$\mathfrak{E}(u + tw) = \mathfrak{E}(u) + t(a(u,w) - \ell(w)) + \frac{1}{2}t^2 a(w,w), \tag{25.10}$$

which results from the symmetry of a. (i) Assume that u solves (25.6). Since $a(w,w) \geq 0$ owing to the coercivity of a with $\xi := 1$, (25.10) implies that u minimizes $\mathfrak{E}$ over V. (ii) Conversely, assume that u minimizes $\mathfrak{E}$ over V. The right-hand side of (25.10) is a quadratic polynomial in t reaching its minimum value at $t = 0$. Hence, the derivative of this polynomial vanishes at $t = 0$, which amounts to $a(u,w) - \ell(w) = 0$. Since w is arbitrary in V, we conclude that u solves (25.6). $\qquad\square$

25.3   Banach–Nečas–Babuška (BNB) theorem

The BNB theorem plays a fundamental role in this book. We use this terminology since, to our knowledge, the BNB theorem was stated by Nečas in 1962 [310] and Babuška in 1970 in the context of finite element methods [33]. From a functional analysis point of view, the BNB theorem is a rephrasing of two fundamental results by Banach: the closed range theorem and the open mapping theorem. We present two settings for the BNB theorem depending on whether the test functions in the model problem belong to a reflexive Banach space or to the dual of a Banach space. Recall from Definition C.18 that a Banach space W is said to be reflexive if the canonical isometry $J_W : W \to W''$ is an isomorphism. This is always the case if W is a Hilbert space.

25.3.1   Test functions in reflexive Banach space

Theorem 25.9 (Banach–Nečas–Babuška (BNB)). *Let V be a Banach space and let W be a reflexive Banach space. Let a be a bounded sesquilinear form on $V \times W$ and let $\ell \in W'$. Then the problem* (25.1) *is well-posed iff:*

$$(\textsc{bnb}1) \qquad \inf_{v \in V} \sup_{w \in W} \frac{|a(v,w)|}{\|v\|_V \|w\|_W} =: \alpha > 0, \tag{25.11a}$$

$$(\textsc{bnb}2) \qquad \forall w \in W, \quad [\forall v \in V,\ a(v,w) = 0] \implies [w = 0]. \tag{25.11b}$$

(It is implicitly understood that the argument is nonzero in the above infimum and supremum.) Moreover, we have the a priori estimate $\|u\|_V \leq \frac{1}{\alpha}\|\ell\|_{W'}$.

Proof. Let $A \in \mathcal{L}(V; W')$ be defined by (25.4) and let us prove that the three conditions (i)-(ii)-(iii) in (25.5) are equivalent to (BNB1)-(BNB2). The conditions (i)-(ii) are equivalent to (BNB1) since for all $v \in V$,

$$\|Av\|_{W'} = \sup_{w \in W} \frac{|\langle A(v), w\rangle_{W',W}|}{\|w\|_W} = \sup_{w \in W} \frac{|a(v,w)|}{\|w\|_W}.$$

Since $\langle A^*(J_W(w)), v\rangle_{V',V} = \langle J_W(w), A(v)\rangle_{W'',W'} = \overline{\langle A(v), w\rangle_{W',W}} = \overline{a(v,w)}$ for all $(v,w) \in V \times W$, stating that $a(v,w) = 0$ for all $v \in V$ is equivalent to stating that $(A^* \circ J_W)(w) = 0$. Hence,

(BNB2) is equivalent to stating that $A^* {\circ} J_W$ is injective. Furthermore, since W is reflexive, the canonical isometry $J_W : W \to W''$ from Proposition C.17 is an isomorphism. Hence, (BNB2) is equivalent to stating that $A^* : W'' \to V'$ is injective, which is the condition (iii) in (25.5). Finally, the a priori estimate follows from the inequalities $\alpha\|u\|_V \leq \sup_{w \in W} \frac{|a(u,w)|}{\|w\|_W} = \sup_{w \in W} \frac{|\ell(w)|}{\|w\|_W} = \|\ell\|_{W'}$. $\qquad\square$

Remark 25.10 ((BNB1)). Condition (BNB1) is called *inf-sup condition* and it is equivalent to the following statement:

$$\exists \alpha > 0, \qquad \alpha\|v\|_V \leq \sup_{w \in W} \frac{|a(v,w)|}{\|w\|_W}, \qquad \forall v \in V. \tag{25.12}$$

Establishing (25.12) is usually done by finding two positive real numbers c_1, c_2 s.t. for all $v \in V$, one can find a "partner" $w_v \in W$ s.t. $\|w_v\|_W \leq c_1\|v\|_V$ and $|a(v, w_v)| \geq c_2\|v\|_V^2$. If this is indeed the case, then (25.12) holds true with $\alpha := \frac{c_2}{c_1}$. Establishing coercivity amounts to asserting that $w_v = \zeta v$ is a suitable partner for some $\zeta \in \mathbb{C}$ with $|\zeta| = 1$. $\qquad\square$

Remark 25.11 ((BNB2)). The statement in (BNB2) is equivalent to asserting that for all w in W, either there exists v in V such that $a(v, w) \neq 0$ or $w = 0$. In view of the proof Theorem 25.9, (BNB2) says that for all w in W, either $A^* {\circ} J_W(w) \neq 0$ or $w = 0$. $\qquad\square$

Remark 25.12 (Two-sided bound). Since $\|\ell\|_{W'} = \|A(u)\|_{W'} \leq \omega\|u\|_V$ where $\omega := \|a\|_{V \times W} = \|A\|_{\mathcal{L}(V;W')}$ is the boundedness constant of the sesquilinear form a on $V {\times} W$, we infer the two-sided bound

$$\frac{1}{\|a\|_{V \times W}}\|\ell\|_{W'} \leq \|u\|_V \leq \frac{1}{\alpha}\|\ell\|_{W'}.$$

Since $\alpha^{-1} = \|A^{-1}\|_{\mathcal{L}(W';V)}$ owing to Lemma C.51, the quantity

$$\kappa(a) = \frac{\|a\|_{V \times W}}{\alpha} = \|A\|_{\mathcal{L}(V;W')}\|A^{-1}\|_{\mathcal{L}(W';V)} \geq 1$$

can be viewed as the *condition number* of the sesquilinear form a (or of the associated operator A). A similar notion of conditioning is developed for matrices in §28.2.1. $\qquad\square$

Remark 25.13 (Link with Lax–Milgram). Let V be a Hilbert space and let a be a bounded and coercive bilinear form on $V {\times} V$. The proof of the Lax–Milgram lemma shows that a satisfies the conditions (BNB1) and (BNB2) (with $W = V$). The converse is false: the conditions (BNB1) and (BNB2) do not imply coercivity. Hence, (25.7) is *not necessary* for well-posedness, whereas (BNB1)-(BNB2) are *necessary and sufficient*. However, coercivity is both necessary and sufficient for well-posedness when the bilinear form a is Hermitian and positive semidefinite; see Exercise 25.7. $\qquad\square$

Remark 25.14 (T-coercivity). Let V, W be Hilbert spaces. Then (BNB1)-(BNB2) are equivalent to the existence of a bijective operator $T \in \mathcal{L}(V;W)$ and a positive real number η such that

$$\Re(a(v, T(v))) \geq \eta\|v\|_V^2, \qquad \forall v \in V.$$

This property is called T-*coercivity* in Bonnet-Ben Dhia et al. [72, 73]; see Exercise 25.10. The advantage of this notion over coercivity is the possibility of treating different trial and test spaces and using a test function different from $v \in V$ to estimate $\|v\|_V^2$. Note that the bilinear form $(u, v) \mapsto a(u, T(v))$ is bounded and coercive on $V {\times} V$. Proposition C.59 then implies that V is necessarily a Hilbert space. This argument proves that T-coercivity is a notion relevant in Hilbert spaces only. The BNB theorem is more general than T-coercivity since it also applies to Banach spaces. $\qquad\square$

25.3.2 Test functions in dual Banach space

The requirement on the reflexivity of the space W in the BNB theorem can be removed if the model problem is reformulated in such a way that the test functions act on the problem data instead of the data acting on the test functions. Assume that we are given a bounded operator $A \in \mathcal{L}(V; W)$ and some data $f \in W$, and we want to assert that there is a unique $u \in V$ s.t. $A(u) = f$. To recast this problem in the general setting of (25.1) using test functions, we define the bounded sesquilinear form on $V \times W'$ such that

$$a(v, w') := \overline{\langle w', A(v) \rangle_{W',W}}, \qquad \forall (v, w') \in V \times W', \tag{25.13}$$

and we consider the following model problem:

$$\begin{cases} \text{Find } u \in V \text{ such that} \\ a(u, w') = \overline{\langle w', f \rangle_{W',W}}, \quad \forall w' \in W'. \end{cases} \tag{25.14}$$

Then $u \in V$ solves (25.14) iff $\langle w', A(u) - f \rangle_{W',W} = 0$ for all $w' \in W'$, that is, iff $A(u) = f$. In (25.14), the data is f is in W and the test functions belong to W', whereas in the original model problem (25.1) the data is $\ell \in W'$ and the test functions belong to W. The functional setting of (25.14) is useful, e.g., when considering first-order PDEs; see §24.2.1.

Theorem 25.15 (Banach–Nečas–Babuška (BNB)). *Let V, W be Banach spaces. Let $A \in \mathcal{L}(V; W)$ and let $f \in W$. Let a be the bounded sesquilinear form on $V \times W'$ defined in (25.13). The problem (25.14) is well-posed iff:*

$$(\text{BNB1'}) \qquad \inf_{v \in V} \sup_{w' \in W'} \frac{|a(v, w')|}{\|v\|_V \|w'\|_{W'}} := \alpha > 0, \tag{25.15}$$

$$(\text{BNB2'}) \qquad \forall w' \in W', \quad [\forall v \in V, \ a(v, w') = 0] \implies [w' = 0]. \tag{25.16}$$

Moreover, we have the a priori estimate $\|u\|_V \leq \frac{1}{\alpha} \|f\|_W$.

Proof. The well-posedness of (25.14) is equivalent to the bijectivity of $A : V \to W$, and this property is equivalent to the three conditions (i)-(ii)-(iii) in (25.5) with W in lieu of W' and $A^* : W' \to V'$. Since $\|A(v)\|_W = \sup_{w' \in W'} \frac{|\langle w', A(v) \rangle_{W',W}|}{\|w'\|_{W'}}$ owing to Corollary C.14, the condition (BNB1') means that $\|A(v)\|_W \geq \alpha \|v\|_V$ for all $v \in V$. This condition is therefore equivalent to the conditions (i)-(ii). Moreover, since $a(v, w') = \overline{\langle w', A(v) \rangle_{W',W}} = \overline{\langle A^*(w'), v \rangle_{V',V}}$, (BNB2') amounts to the condition (iii) (i.e., the injectivity of A^*). $\qquad \square$

Remark 25.16 (A vs. a). In the first version of the BNB theorem (Theorem 25.9), it is equivalent to assume that we are given an operator $A \in \mathcal{L}(V; W')$ or a bounded sesquilinear form a on $V \times W$. But, in the second version of the BNB theorem (Theorem 25.15), we are given an operator $A \in \mathcal{L}(V; W)$, and the bounded sesquilinear form a on $V \times W'$ is defined *from A*. If we were given instead a bounded sesquilinear form a on $V \times W'$, proceeding as in (25.4) would be awkward since it would lead to an operator $\tilde{A} \in \mathcal{L}(V; W'')$ s.t. $\langle \tilde{A}(v), w' \rangle_{W'',W'} := a(v, w')$ for all $(v, w') \in V \times W'$. $\qquad \square$

Remark 25.17 (Literature). Inf-sup conditions in nonreflexive Banach spaces are discussed in Amrouche and Ratsimahalo [9]. $\qquad \square$

25.4 Two examples

In this section, we present two examples illustrating the above abstract results.

25.4.1 Darcy's equations

The weak formulation (24.12) fits the setting of the model problem (25.1) with

$$V := \boldsymbol{H}(\mathrm{div}; D) \times H_0^1(D), \qquad W := \boldsymbol{L}^2(D) \times L^2(D),$$

where $\|\boldsymbol{\sigma}\|_{\boldsymbol{H}(\mathrm{div};D)} := (\|\boldsymbol{\sigma}\|_{\boldsymbol{L}^2(D)}^2 + \ell_D^2 \|\nabla{\cdot}\boldsymbol{\sigma}\|_{L^2(D)}^2)^{\frac{1}{2}}$ (recall that ℓ_D is a characteristic length scale associated with D, e.g., $\ell_D := \mathrm{diam}(D)$), and with the bilinear and linear forms

$$a(v, w) := \int_D \left(\boldsymbol{\sigma}{\cdot}\boldsymbol{\tau} + \nabla p{\cdot}\boldsymbol{\tau} + (\nabla{\cdot}\boldsymbol{\sigma})q \right) \mathrm{d}x, \qquad \ell(w) := \int_D fq \, \mathrm{d}x, \tag{25.17}$$

with $v := (\boldsymbol{\sigma}, p) \in V$ and $w := (\boldsymbol{\tau}, q) \in W$.

Proposition 25.18. *Problem* (24.12) *is well-posed.*

Proof. We equip the Hilbert spaces V and W with the norms $\|v\|_V := \left(\|\boldsymbol{\sigma}\|_{\boldsymbol{H}(\mathrm{div};D)}^2 + |p|_{H^1(D)}^2 \right)^{\frac{1}{2}}$ and $\|w\|_W := \left(\|\boldsymbol{\tau}\|_{\boldsymbol{L}^2(D)}^2 + \ell_D^{-2} \|q\|_{L^2(D)}^2 \right)^{\frac{1}{2}}$ with $v := (\boldsymbol{\sigma}, p)$ and $w := (\boldsymbol{\tau}, q)$, respectively. That $\|{\cdot}\|_V$ is indeed a norm follows from the Poincaré–Steklov inequality (3.11) (see Remark 3.29). Since the bilinear form a and the linear form ℓ are obviously bounded, it remains to check the conditions (BNB1) and (BNB2).
(1) Proof of (BNB1). Let $(\boldsymbol{\sigma}, p) \in V$ and define $\mathbb{S} := \sup_{(\boldsymbol{\tau}, q) \in W} \frac{|a((\boldsymbol{\sigma}, p), (\boldsymbol{\tau}, q))|}{\|(\boldsymbol{\tau}, q)\|_W}$. Since $V \subset W$, we can take $(\boldsymbol{\sigma}, p)$ as the test function. Since p vanishes at the boundary, $a((\boldsymbol{\sigma}, p), (\boldsymbol{\sigma}, p)) = \|\boldsymbol{\sigma}\|_{\boldsymbol{L}^2(D)}^2$, whence we infer that

$$\|\boldsymbol{\sigma}\|_{\boldsymbol{L}^2(D)}^2 = \frac{a((\boldsymbol{\sigma}, p), (\boldsymbol{\sigma}, p))}{\|(\boldsymbol{\sigma}, p)\|_W} \|(\boldsymbol{\sigma}, p)\|_W \le \mathbb{S}\, \|(\boldsymbol{\sigma}, p)\|_W.$$

Since $\|{\cdot}\|_W \le \gamma \|{\cdot}\|_V$ on V with $\gamma := \max(1, C_{\mathrm{PS}}^{-1})$, we infer that $\|\boldsymbol{\sigma}\|_{\boldsymbol{L}^2(D)}^2 \le \gamma \mathbb{S}\, \|(\boldsymbol{\sigma}, p)\|_V$. Moreover, we have

$$\left(\|\nabla p\|_{\boldsymbol{L}^2(D)}^2 + \ell_D^2 \|\nabla{\cdot}\boldsymbol{\sigma}\|_{L^2(D)}^2 \right)^{\frac{1}{2}} = \sup_{(\boldsymbol{\tau}, q) \in W} \frac{|\int_D \{\nabla p{\cdot}\boldsymbol{\tau} + (\nabla{\cdot}\boldsymbol{\sigma})q\}\, \mathrm{d}x|}{\|(\boldsymbol{\tau}, q)\|_W}$$

$$\le \sup_{(\boldsymbol{\tau}, q) \in W} \frac{|a((\boldsymbol{\sigma}, p), (\boldsymbol{\tau}, q))|}{\|(\boldsymbol{\tau}, q)\|_W} + \sup_{(\boldsymbol{\tau}, q) \in W} \frac{|\int_D \boldsymbol{\sigma}{\cdot}\boldsymbol{\tau}\, \mathrm{d}x|}{\|(\boldsymbol{\tau}, q)\|_W}.$$

Hence, $\left(\|\nabla p\|_{\boldsymbol{L}^2(D)}^2 + \ell_D^2 \|\nabla{\cdot}\boldsymbol{\sigma}\|_{L^2(D)}^2 \right)^{\frac{1}{2}} \le \mathbb{S} + \|\boldsymbol{\sigma}\|_{\boldsymbol{L}^2(D)}$. Squaring this inequality and combining it with the above bound on $\|\boldsymbol{\sigma}\|_{\boldsymbol{L}^2(D)}$, we infer that

$$\|(\boldsymbol{\sigma}, p)\|_V^2 = \|\nabla p\|_{\boldsymbol{L}^2(D)}^2 + \|\boldsymbol{\sigma}\|_{\boldsymbol{H}(\mathrm{div};D)}^2 \le 2\mathbb{S}^2 + 3\|\boldsymbol{\sigma}\|_{\boldsymbol{L}^2(D)}^2 \le 2\mathbb{S}^2 + 3\gamma \mathbb{S}\, \|(\boldsymbol{\sigma}, p)\|_V.$$

Hence, the inf-sup condition (BNB1) holds true with $\alpha \ge (4 + 9\gamma^2)^{-\frac{1}{2}}$.
(2) Proof of (BNB2). Let $(\boldsymbol{\tau}, q) \in W$ be such that $a((\boldsymbol{\sigma}, p), (\boldsymbol{\tau}, q)) = 0$ for all $(\boldsymbol{\sigma}, p) \in V$. This means on the one hand that $\int_D \nabla p{\cdot}\boldsymbol{\tau}\, \mathrm{d}x = 0$ for all $p \in H_0^1(D)$, so that $\nabla{\cdot}\boldsymbol{\tau} = 0$. On the other hand we obtain that $\int_D \{\boldsymbol{\sigma}{\cdot}\boldsymbol{\tau} + (\nabla{\cdot}\boldsymbol{\sigma})q\}\, \mathrm{d}x = 0$ for all $\boldsymbol{\sigma} \in \boldsymbol{H}(\mathrm{div}; D)$. Taking $\boldsymbol{\sigma} \in \boldsymbol{C}_0^\infty(D)$ we infer that $q \in H^1(D)$ and $\nabla q = \boldsymbol{\tau}$. Observing that $\boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div}; D)$ and taking $\boldsymbol{\sigma} := \boldsymbol{\tau}$, we infer that $0 = \int_D \{\boldsymbol{\tau}{\cdot}\boldsymbol{\tau} + (\nabla{\cdot}\boldsymbol{\tau})q\}\, \mathrm{d}x = \|\boldsymbol{\tau}\|_{\boldsymbol{L}^2(D)}^2$ since $\nabla{\cdot}\boldsymbol{\tau} = 0$. Hence, $\boldsymbol{\tau} = \boldsymbol{0}$. Finally, $\nabla q = \boldsymbol{\tau} = \boldsymbol{0}$, which implies that q is constant on D. Since $\int_D (\nabla{\cdot}\boldsymbol{\sigma})q\, \mathrm{d}x = 0$ for all $\boldsymbol{\sigma} \in \boldsymbol{H}(\mathrm{div}; D)$, q is identically zero in D (take for instance $\boldsymbol{\sigma}(\boldsymbol{x}) := \boldsymbol{x}$). □

25.4.2   First-order PDE

Consider the weak formulation (24.20) on $D := (0,1)$. This formulation fits the setting of the model problem (25.14) with the spaces

$$V := \{v \in W^{1,1}(D) \mid v(0) = 0\}, \qquad W := L^1(D). \tag{25.18}$$

The data is $f \in W$ and we consider the bounded operator $A : V \to W$ s.t. $A(v) := \frac{dv}{dt}$ for all $v \in V$. (Here, we denote derivatives by $\frac{d}{dt}$ and reserve the primes to duality.) Recalling that $W' = L^\infty(D)$, the bilinear form a associated with the operator A is s.t.

$$a(v, w') := \int_0^1 \frac{dv}{dt} w' \, dt, \qquad \forall (v, w') \in V \times W', \tag{25.19}$$

and the right-hand side is $\langle w', f \rangle_{W',W} := \int_0^1 w' f \, dt$ with $f \in W$.

Proposition 25.19. *Problem* (24.20) *is well-posed.*

Proof. We equip the Banach spaces V and W' with the norms $\|v\|_V := \|v\|_{L^1(D)} + \|\frac{dv}{dt}\|_{L^1(D)}$ and $\|w'\|_{W'} := \|w'\|_{L^\infty(D)}$, and we verify the conditions (BNB1') and (BNB2') from Theorem 25.15.
(1) Proof of (BNB1'). Let $v \in V$ and set $D^\pm := \{t \in D \mid \pm \frac{dv}{dt}(t) > 0\}$. Taking $w'_v := \mathbb{1}_{D^+} - \mathbb{1}_{D^-}$, where $\mathbb{1}_S$ denotes the indicator function of a measurable set S, we infer that

$$\sup_{w' \in W'} \frac{|a(v, w')|}{\|w'\|_{W'}} \geq \frac{|a(v, w'_v)|}{\|w'_v\|_{W'}} = \frac{|\int_0^1 \frac{dv}{dt} w'_v \, dt|}{\|w'_v\|_{L^\infty(D)}} = \int_0^1 \left|\frac{dv}{dt}\right| \, dt = \left\|\frac{dv}{dt}\right\|_{L^1(D)}.$$

Invoking the extended Poincaré–Steklov inequality on V (with $p := 1$ and the bounded linear form $v \mapsto v(0)$ in (3.13)) yields (BNB1').
(2) Proof of (BNB2'). Let $w' \in W'$ be such that $\int_0^1 \frac{dv}{dt} w' \, dt = 0$ for all $v \in V$. Taking v in $C_0^\infty(D)$, we infer that the weak derivative of w' vanishes. Lemma 2.11 implies that w' is a constant. Choosing $v(t) := t$ as a test function leads to $\int_0^1 w' \, dt = 0$. Hence, we have $w' = 0$. $\qquad\square$

Exercises

Exercise 25.1 (Riesz–Fréchet). The objective is to prove the Riesz–Fréchet theorem (Theorem C.24) by using the BNB theorem. Let V be a Hilbert space with inner product $(\cdot, \cdot)_V$. (i) Show that for every $v \in V$, there is a unique $J_V^{\mathrm{RF}}(v) \in V'$ s.t. $\langle J_V^{\mathrm{RF}}(v), w \rangle_{V',V} := (v, w)_V$ for all $w \in V$. (ii) Show that $J_V^{\mathrm{RF}} : V' \to V$ is a linear isometry.

Exercise 25.2 (Reflexivity). Let V, W be two Banach spaces such that there is an isomorphism $A \in \mathcal{L}(V; W)$. Assume that V is reflexive. Prove that W is reflexive. (*Hint*: consider the map $A^{**} \circ J_V \circ A^{-1}$.)

Exercise 25.3 (Space $V_\mathbb{R}$). Let V be a set and assume that V has a vector space structure over the field $\mathbb{C}$. By restricting the scaling λv to $\lambda \in \mathbb{R}$ and $v \in V$, V has also a vector space structure over the field $\mathbb{R}$, which we denote by $V_\mathbb{R}$ (V and $V_\mathbb{R}$ are the same sets, but they are equipped with different vector space structures); see Remark C.11. Let V' be the set of the bounded anti-linear forms on V and $V'_\mathbb{R}$ be the set of the bounded linear forms on $V_\mathbb{R}$. Prove that the map $I : V' \to V'_\mathbb{R}$ such that for all $\ell \in V'$, $I(\ell)(v) := \Re(\ell(v))$ for all $v \in V$, is a bijective isometry. (*Hint*: for $\psi \in V'_\mathbb{R}$, set $\ell(v) := \psi(v) + i\psi(iv)$ with $i^2 = -1$.)

Exercise 25.4 (Orthogonal projection). Let V be a Hilbert space with inner product $(\cdot,\cdot)_V$ and induced norm $\|\cdot\|_V$. Let U be a nonempty, closed, and convex subset of V. Let $f \in V$. (i) Show that there is a unique u in U such that $\|f - u\|_V = \min_{v \in U} \|f - v\|_V$. (*Hint*: recall that $\frac{1}{4}(a-b)^2 = \frac{1}{2}(c-a)^2 + \frac{1}{2}(c-b)^2 - (c - \frac{1}{2}(a+b))^2$ and show that a minimizing sequence is a Cauchy sequence.) (ii) Show that $u \in U$ is the minimizer if and only if $\Re((f-u, v-u)_V) \le 0$ for all $v \in U$. (*Hint*: proceed as in the proof of Proposition 25.8.) (iii) Assuming that U is a (nontrivial) subspace of V, prove that the unique minimizer is characterized by $(f-u, v)_V = 0$ for all $v \in U$, and prove that the map $\Pi_U : V \ni f \mapsto u \in U$ is linear and $\|\Pi_U\|_{\mathcal{L}(V;U)} = 1$. (iv) Let a be a bounded, Hermitian, and coercive sesquilinear form (with $\xi := 1$ for simplicity). Let $\ell \in V'$. Set $\mathfrak{E}(v) := \frac{1}{2}a(v,v) - \ell(v)$. Show that there is a unique $u \in V$ such that $\mathfrak{E}(u) = \min_{v \in U} \mathfrak{E}(v)$ and that u is the minimizer if and only if $\Re(a(u, v-u) - \ell(v-u)) \ge 0$ for all $v \in U$.

Exercise 25.5 (Inf-sup constant). Let V be a Hilbert space, U a subset of V, and W a closed subspace of V. Let $\beta := \inf_{u \in U} \sup_{w \in W} \frac{|(u,w)_V|}{\|u\|_V \|w\|_W}$. (i) Prove that $\beta \in [0,1]$. (ii) Prove that $\beta = \inf_{u \in U} \frac{\|\Pi_W(u)\|_V}{\|u\|_V}$, where Π_W is the orthogonal projection onto W. (*Hint*: use Exercise 25.4.) (iii) Prove that $\|u - \Pi_W(u)\|_V \le (1 - \beta^2)^{\frac{1}{2}} \|u\|_V$. (*Hint*: use the Pythagorean identity.)

Exercise 25.6 (Fixed-point argument). The goal of this exercise is to derive another proof of the Lax–Milgram lemma. Let $A \in \mathcal{L}(V;V)$ be defined by $(A(v), w)_V := a(v,w)$ for all $v, w \in V$ (note that we use an inner product to define A). Let L be the representative in V of the linear form $\ell \in V'$. Let λ be a positive real number. Consider the map $T_\lambda : V \to V$ s.t. $T_\lambda(v) := v - \lambda\xi(A(v) - L)$ for all $v \in V$. Prove that if λ is small enough, $\|T_\lambda(v) - T_\lambda(w)\|_V \le \rho_\lambda \|v - w\|_V$ for all $v, w \in V$ with $\rho_\lambda \in (0,1)$, and show that (25.6) is well-posed. (*Hint*: use Banach's fixed-point theorem.)

Exercise 25.7 (Coercivity as necessary condition). Let V be a reflexive Banach space and let $A \in \mathcal{L}(V;V')$ be a monotone self-adjoint operator; see Definition C.31. Prove that A is bijective if and only if A is coercive (with $\xi := 1$). (*Hint*: prove that $\Re(\langle A(v), w\rangle_{V',V}) \le \langle A(v), v\rangle_{V',V}^{\frac{1}{2}} \langle A(w), w\rangle_{V',V}^{\frac{1}{2}}$ for all $v, w \in V$.)

Exercise 25.8 (Darcy). Prove that the problem (24.14) is well-posed. (*Hint*: adapt the proof of Proposition 25.18.)

Exercise 25.9 (First-order PDE). Prove that the problem (24.21) is well-posed. (*Hint*: adapt the proof of Proposition 25.19.)

Exercise 25.10 (T-coercivity). Let V, W be Hilbert spaces. Prove that (BNB1)-(BNB2) are equivalent to the existence of a bijective operator $T \in \mathcal{L}(V;W)$ and a real number $\eta > 0$ such that $\Re(a(v, T(v))) \ge \eta\|v\|_V^2$ for all $v \in V$. (*Hint*: use J_W^{-1}, $(A^{-1})^*$, and the map J_V^{RF} from the Riesz–Fréchet theorem to construct T.)

Exercise 25.11 (Sign-changing diffusion). Let D be a Lipschitz domain D in $\mathbb{R}^d$ partitioned into two disjoint Lipschitz subdomains D_1 and D_2. Set $\Sigma := \partial D_1 \cap \partial D_2$, each having an intersection with ∂D of positive measure. Let κ_1, κ_2 be two real numbers s.t. $\kappa_1 > 0$ and $\kappa_2 < 0$. Set $\kappa(x) := \kappa_1 \mathbb{1}_{D_1}(x) + \kappa_2 \mathbb{1}_{D_2}(x)$ for all $x \in D$. Let $V := H_0^1(D)$ be equipped with the norm $\|\nabla v\|_{\boldsymbol{L}^2(D)}$. The goal is to show that the bilinear form $a(v,w) := \int_D \kappa \nabla v \cdot \nabla w$ satisfies conditions (BNB1)-(BNB2) on $V \times V$; see Chesnel and Ciarlet [118]. Set $V_m := \{v|_{D_m} \mid v \in V\}$ for all $m \in \{1,2\}$, equipped with the norm $\|\nabla v_m\|_{\boldsymbol{L}^2(D_m)}$ for all $v_m \in V_m$, and let $\gamma_{0,m}$ be the traces of functions in V_m on Σ. (i) Assume that there is $S_1 \in \mathcal{L}(V_1;V_2)$ s.t. $\gamma_{0,2}(S_1(v_1)) = \gamma_{0,1}(v_1)$. Define $T : V \to V$ s.t. for all $v \in V$, $T(v)(x) := v(x)$ if $x \in D_1$ and $T(v)(x) := -v(x) + 2S_1(v|_{D_1})(x)$ if $x \in D_2$. Prove that $T \in \mathcal{L}(V)$ and that T is an isomorphism. (*Hint*: verify that $T \circ T = I_V$, the identity in V.) (ii) Assume that $\frac{\kappa_1}{|\kappa_2|} > \|S_1\|_{\mathcal{L}(V_1;V_2)}^2$. Prove that the conditions (BNB1)-(BNB2) are satisfied.

(*Hint*: use T-coercivity from Remark 25.14.) (iii) Let $D_1 := (-a, 0) \times (0, 1)$ and $D_2 := (0, b) \times (0, 1)$ with $a > b > 0$. Show that if $\frac{\kappa_1}{|\kappa_2|} \notin [1, \frac{a}{b}]$, (BNB1)-(BNB2) are satisfied. (*Hint*: consider the map $S_1 \in \mathcal{L}(V_1; V_2)$ s.t. $S_1(v_1)(x, y) := v_1(-\frac{a}{b}x, y)$ for all $v_1 \in V_1$, and the map $S_2 \in \mathcal{L}(V_2; V_1)$ s.t. $S_2(v_2)(x, y) := v_2(-x, y)$ if $x \in (-b, 0)$ and $S_2(v_2)(x, y) := 0$ otherwise, for all $v_2 \in V_2$.)

Chapter 26

Basic error analysis

In Part VI, composed of Chapters 26 to 30, we introduce the Galerkin approximation technique and derive fundamental stability results and error estimates. We also investigate implementation aspects of the method (quadratures, linear algebra, assembling, storage). In this chapter, we consider the following problem, introduced in Chapter 25, and study its approximation by the Galerkin method:

$$\begin{cases} \text{Find } u \in V \text{ such that} \\ a(u, w) = \ell(w), \quad \forall w \in W. \end{cases} \tag{26.1}$$

Here, V and W are Banach spaces, a is a bounded sesquilinear form on $V \times W$, and ℓ is a bounded antilinear form on W. We focus on the well-posedness of the approximate problem, and we derive a bound on the approximation error in a simple setting. This bound is known in the literature as Céa's lemma. We also characterize the well-posedness of the discrete problem by using the notion of Fortin operator.

To stay general, we consider complex vector spaces. The case of real vector spaces is recovered by replacing the field $\mathbb{C}$ by $\mathbb{R}$, by removing the real part symbol $\Re(\cdot)$ and the complex conjugate symbol $\bar{\cdot}$, and by interpreting the symbol $|\cdot|$ as the absolute value instead of the modulus. Moreover, sesquilinear forms become bilinear forms, and antilinear forms are just linear forms. We denote by α and $\|a\|_{V \times W}$ the inf-sup and the boundedness constants of the sesquilinear form a on $V \times W$, i.e.,

$$\alpha := \inf_{v \in V} \sup_{w \in W} \frac{|a(v, w)|}{\|v\|_V \|w\|_W} \leq \sup_{v \in V} \sup_{w \in W} \frac{|a(v, w)|}{\|v\|_V \|w\|_W} =: \|a\|_{V \times W}. \tag{26.2}$$

We assume that (26.1) is well-posed, i.e., $0 < \alpha$ and $\|a\|_{V \times W} < \infty$. Whenever the context is unambiguous, we write $\|a\|$ instead of $\|a\|_{V \times W}$.

26.1 The Galerkin method

The central idea in the Galerkin method is to replace in (26.1) the infinite-dimensional spaces V and W by *finite-dimensional* spaces V_h and W_h (we always assume that $V_h \neq \{0\}$ and $W_h \neq \{0\}$). The subscript $h \in \mathcal{H}$ refers to the fact that these spaces are constructed as explained in Volume I using finite elements and a mesh $\mathcal{T}_h$ belonging to some sequence of meshes $(\mathcal{T}_h)_{h \in \mathcal{H}}$. The discrete

problem takes the following form:

$$\begin{cases} \text{Find } u_h \in V_h \text{ such that} \\ a_h(u_h, w_h) = \ell_h(w_h), \quad \forall w_h \in W_h, \end{cases} \tag{26.3}$$

where a_h is a bounded sesquilinear form on $V_h \times W_h$ and ℓ_h is a bounded antilinear form on W_h. Notice that a_h and ℓ_h possibly differ from a and ℓ, respectively. Since the spaces V_h and W_h are finite-dimensional, (26.3) is called *discrete problem*. The space V_h is called *discrete trial space* (or *discrete solution space*), and W_h *discrete test space*.

Definition 26.1 (Standard Galerkin, Petrov–Galerkin). *The discrete problem* (26.3) *is called* standard Galerkin approximation *when $W_h = V_h$ and* Petrov–Galerkin approximation *otherwise.*

Definition 26.2 (Conforming setting). *The approximation is said to be* conforming *if $V_h \subset V$ and $W_h \subset W$.*

There are circumstances when considering nonconforming approximations is useful. Two important examples are discontinuous Galerkin methods where discrete functions are discontinuous across the mesh interfaces (see Chapters 38 and 60) and boundary penalty methods where boundary conditions are enforced weakly (see Chapters 37 for elliptic PDEs and Chapters 57–59 for Friedrichs' systems). Very often, nonconforming approximations make it necessary to work with discrete forms that differ from their continuous counterparts. For instance, the bilinear form $\int_D \nabla v \cdot \nabla w \, dx$ does not make sense if the functions v and w are discontinuous. Another important example leading to a modification of the forms at the discrete level is the use of quadratures (see Chapter 30).

26.2 Discrete well-posedness

Our goal in this section is to study the well-posedness of the discrete problem (26.3). We equip V_h and W_h with norms denoted by $\|\cdot\|_{V_h}$ and $\|\cdot\|_{W_h}$, respectively. These norms can differ from those of V and W. One reason can be that the approximation is nonconforming and the norm $\|\cdot\|_V$ is meaningless on V_h. This is the case for instance if the norm $\|\cdot\|_V$ includes the H^1-norm and the discrete functions are allowed to jump across the mesh interfaces.

26.2.1 Discrete Lax–Milgram

Lemma 26.3 (Discrete Lax–Milgram). *Let V_h be a finite-dimensional space. Assume that $W_h = V_h$ in (26.3). Let a_h be a bounded sesquilinear form on $V_h \times V_h$ and let $\ell_h \in V_h'$. Assume that a_h is* coercive *on V_h, i.e., there is a real number $\alpha_h > 0$ and a complex number ξ with $|\xi| = 1$ such that*

$$\Re(\xi a_h(v_h, v_h)) \geq \alpha_h \|v_h\|_{V_h}^2, \qquad \forall v_h \in V_h. \tag{26.4}$$

Then (26.3) is well-posed with the a priori estimate $\|u_h\|_{V_h} \leq \frac{1}{\alpha_h} \|\ell_h\|_{V_h'}$.

Proof. A simple proof just consists of invoking the Lax–Milgram lemma (see Lemma 25.2). We now propose an elementary proof that relies on V_h being finite-dimensional. Let $A_h : V_h \to V_h'$ be the linear operator such that $\langle A_h(v_h), w_h \rangle_{V_h', V_h} := a_h(v_h, w_h)$ for all $v_h, w_h \in V_h$. Problem (26.3) amounts to seeking $u_h \in V_h$ such that $A_h(u_h) = \ell_h$ in V_h'. Hence, (26.3) is well-posed iff A_h is an isomorphism. Since $\dim(V_h) = \dim(V_h') < \infty$ this is equivalent to require that A_h be injective, i.e.,

$\ker(A_h) = \{0\}$. Let $v_h \in \ker(A_h)$ so that $0 = \xi\langle A_h(v_h), v_h\rangle_{V_h', V_h} = \xi a_h(v_h, v_h)$. From coercivity, we deduce that $0 \geq \alpha_h \|v_h\|_{V_h}^2$, which proves that $v_h = 0$. Hence, $\ker(A_h) = \{0\}$, thereby proving that A_h is bijective. $\qquad\square$

Example 26.4 (Sufficient condition). (26.4) holds true if $V_h \subset V$ (conformity), $a_h := a_{|V_h \times V_h}$, and a is coercive on $V \times V$. $\qquad\square$

Remark 26.5 (Variational formulation). As in the continuous setting (see Proposition 25.8), if V_h is a real Hilbert space and if a_h is symmetric and coercive (with $\xi := 1$ and $W_h = V_h$), then u_h solves (26.3) iff u_h minimizes the functional $\mathfrak{E}_h(v_h) := \frac{1}{2}a_h(v_h, v_h) - \ell_h(v_h)$ over V_h. If $V_h \subset V$, $a_h := a_{|V_h \times V_h}$, and $\ell_h := \ell_{|V_h}$, then $\mathfrak{E}_h = \mathfrak{E}_{|V_h}$ ($\mathfrak{E}$ is the exact energy functional), and $\mathfrak{E}(u_h) \geq \mathfrak{E}(u)$ since u minimizes $\mathfrak{E}$ over the larger space V. $\qquad\square$

26.2.2 Discrete BNB

Theorem 26.6 (Discrete BNB). *Let V_h, W_h be finite-dimensional spaces. Let a_h be a bounded sesquilinear form on $V_h \times W_h$ and let $\ell_h \in W_h'$. Then the problem (26.3) is well-posed iff*

$$\inf_{v_h \in V_h} \sup_{w_h \in W_h} \frac{|a_h(v_h, w_h)|}{\|v_h\|_{V_h} \|w_h\|_{W_h}} =: \alpha_h > 0, \tag{26.5a}$$

$$\dim(V_h) = \dim(W_h). \tag{26.5b}$$

(Recall that arguments in the above infimum and supremum are understood to be nonzero.) Moreover, we have the a priori estimate $\|u_h\|_{V_h} \leq \frac{1}{\alpha_h}\|\ell_h\|_{W_h'}$.

Proof. Let $A_h : V_h \to W_h'$ be the linear operator such that

$$\langle A_h(v_h), w_h\rangle_{W_h', W_h} := a_h(v_h, w_h), \qquad \forall(v_h, w_h) \in V_h \times W_h. \tag{26.6}$$

The well-posedness of (26.3) is equivalent to A_h being an isomorphism, which owing to the finite-dimensional setting and the rank nullity theorem, is equivalent to (i) $\ker(A_h) = \{0\}$ (i.e., A_h is injective) and (ii) $\dim(V_h) = \dim(W_h')$. Since $\dim(W_h) = \dim(W_h')$, (26.5b) is equivalent to (ii). Let us prove that (i) is equivalent to the inf-sup condition (26.5a). By definition, we have

$$\sup_{w_h \in W_h} \frac{|a_h(v_h, w_h)|}{\|w_h\|_{W_h}} = \sup_{w_h \in W_h} \frac{|\langle A_h(v_h), w_h\rangle_{W_h', W_h}|}{\|w_h\|_{W_h}} =: \|A_h(v_h)\|_{W_h'}.$$

Assume first that (26.5a) holds true and let $v_h \in V_h$ be s.t. $A_h(v_h) = 0$. Then we have $\alpha_h \|v_h\|_{V_h} \leq \|A_h(v_h)\|_{W_h'} = 0$, which shows that $v_h = 0$. Hence, (26.5a) implies the injectivity of A_h. Conversely, assume $\ker(A_h) = \{0\}$ and let us prove (26.5a). An equivalent statement of (26.5a) is that there is $n_0 \in \mathbb{N}^*$ such that for all $v_h \in V_h$ with $\|v_h\|_{V_h} = 1$, one has $\|A_h(v_h)\|_{W_h'} > \frac{1}{n_0}$. Reasoning by contradiction, consider a sequence $(v_{hn})_{n \in \mathbb{N}^*}$ in V_h with $\|v_{hn}\|_{V_h} = 1$ and $\|A_h(v_{hn})\|_{W_h'} \leq \frac{1}{n}$. Since V_h is finite-dimensional, its unit sphere is compact. Hence, there is $v_h \in V_h$ such that, up to a subsequence, $v_{hn} \to v_h$. The limit v_h satisfies $\|v_h\|_{V_h} = 1$ and $A_h(v_h) = 0$, i.e., $v_h \in \ker(A_h) = \{0\}$, which contradicts $\|v_h\|_{V_h} = 1$. Hence, the injectivity of A_h implies (26.5a). In conclusion, $\ker(A_h) = \{0\}$ iff (26.5a) holds true. Finally, the a priori estimate follows from $\alpha_h\|u_h\|_{V_h} \leq \|A_h(u_h)\|_{W_h'} = \|\ell_h\|_{W_h'}$. $\qquad\square$

Remark 26.7 (Link with BNB theorem). Condition (26.5a) is identical to (BNB1) from Theorem 25.9 applied to (26.3), and it is equivalent to the following *inf-sup condition*:

$$\exists\alpha_h > 0, \qquad \alpha_h\|v_h\|_{V_h} \leq \sup_{w_h \in W_h} \frac{|a_h(v_h, w_h)|}{\|w_h\|_{W_h}}, \qquad \forall v_h \in V_h. \tag{26.7}$$

Condition (26.5b) seemingly differs from (BNB2) applied to (26.3), which reads

$$\forall w_h \in W_h, \quad [\, a_h(v_h, w_h) = 0, \ \forall v_h \in V_h \,] \implies [\, w_h = 0 \,]. \tag{26.8}$$

To see that (26.5b) is equivalent to (26.8) provided (26.5a) holds true, let us introduce the adjoint operator $A_h^* : W_h \to V_h'$ (note that the space W_h is reflexive since it is finite-dimensional) such that

$$\overline{\langle A_h^*(w_h), v_h \rangle_{V_h', V_h}} = a_h(v_h, w_h), \qquad \forall (v_h, w_h) \in V_h \times W_h. \tag{26.9}$$

Then (26.8) says that A_h^* is injective, and this statement is equivalent to (26.5b) if $\ker(A_h) = \{0\}$; see Exercise 26.1. In summary, when the setting is finite-dimensional, the key property guaranteeing well-posedness is (26.5a), whereas the other condition (26.5b) is very simple to verify. $\square$

Remark 26.8 (A_h^*). A_h is an isomorphism iff A_h^* is an isomorphism; see Exercise 26.2. Moreover, owing to Lemma C.53 (note that the space V_h is reflexive since it is finite-dimensional), A_h and A_h^* satisfy the inf-sup condition (26.5a) with *the same constant* α_h, i.e.,

$$\inf_{v_h \in V_h} \sup_{w_h \in W_h} \frac{|\langle A_h(v_h), w_h \rangle_{W_h', W_h}|}{\|v_h\|_{V_h} \|w_h\|_{W_h}} = \inf_{w_h \in W_h} \sup_{v_h \in V_h} \frac{|\langle A_h(v_h), w_h \rangle_{W_h', W_h}|}{\|v_h\|_{V_h} \|w_h\|_{W_h}}. \tag{26.10}$$

Note that $\langle A_h(v_h), w_h \rangle_{W_h', W_h} = \overline{\langle A_h^*(w_h), v_h \rangle_{V_h', V_h}}$. As shown in Remark C.54, the identity (26.10) may fail if A_h is not an isomorphism. $\square$

26.2.3   Fortin's lemma

We focus on a conforming approximation, i.e., $V_h \subset V$ and $W_h \subset W$, we equip the spaces V_h and W_h with the norms of V and W, respectively, and we assume that $a_h := a_{|V_h \times W_h}$. Our goal is to devise a criterion to ascertain that a_h satisfies the inf-sup condition (26.5a). To this purpose, we would like to use the inf-sup condition (26.2) satisfied by a on $V \times W$. Unfortunately, this condition does not imply its discrete counterpart on $V_h \times W_h$. Since $V_h \subset V$, (26.2) implies that $\alpha \|v_h\|_V \le \sup_{w \in W} \frac{|a(v_h, w)|}{\|w\|_W}$ for all $v_h \in V_h$, but it is not clear that the bound still holds when restricting the supremum to the subspace W_h. The Fortin operator provides the missing ingredient.

Lemma 26.9 (Fortin). *Let V, W be Hilbert spaces and let a be a bounded sesquilinear form on $V \times W$. Let α and $\|a\|$ be the inf-sup and boundedness constants of a defined in (26.2). Let $V_h \subset V$ and let $W_h \subset W$ be equipped with the norms of V and W, respectively. Consider the following two statements:*

  (i) *There exists a map $\Pi_h : W \to W_h$, called* Fortin operator *such that: (i.a) $a(v_h, \Pi_h(w) - w) = 0$ for all $(v_h, w) \in V_h \times W$; (i.b) There is $\gamma_{\Pi_h} > 0$ such that $\gamma_{\Pi_h} \|\Pi_h(w)\|_W \le \|w\|_W$ for all $w \in W$.*

  (ii) *The discrete inf-sup condition (26.5a) holds true.*

Then (i) $\implies$ (ii) *with $\alpha_h \ge \gamma_{\Pi_h} \alpha$. Conversely,* (ii) $\implies$ (i) *with $\gamma_{\Pi_h} \ge \frac{\alpha_h}{\|a\|}$ and Π_h can be constructed to be linear and idempotent ($\Pi_h \circ \Pi_h = \Pi_h$).*

Proof. (1) Let us assume (i). Let $\epsilon > 0$. We have for all $v_h \in V_h$,

$$\sup_{w_h \in W_h} \frac{|a(v_h, w_h)|}{\|w_h\|_W} \ge \sup_{w \in W} \frac{|a(v_h, \Pi_h(w))|}{\|\Pi_h(w)\|_W + \epsilon \|w\|_W} = \sup_{w \in W} \frac{|a(v_h, w)|}{\|\Pi_h(w)\|_W + \epsilon \|w\|_W}$$

$$\ge \gamma_{\Pi_h} \sup_{w \in W} \frac{|a(v_h, w)|}{\|w\|_W (1 + \epsilon \gamma_{\Pi_h})} \ge \frac{\gamma_{\Pi_h}}{1 + \epsilon \gamma_{\Pi_h}} \alpha \, \|v_h\|_V,$$

since a satisfies (BNB1) and $V_h \subset V$. This proves (26.5a) with $\alpha_h \geq \gamma_{\Pi_h}\alpha$ since ϵ can be taken arbitrarily small. (Since Π_h cannot be injective, we introduced $\epsilon > 0$ to avoid dividing by zero whenever $w \in \ker(\Pi_h)$.)

(2) Conversely, let us assume that a satisfies (26.5a). Let $A_h : V_h \to W_h'$ be defined in (26.6). Condition (26.5a) means that $\|A_h(v_h)\|_{W_h'} \geq \alpha_h\|v_h\|_V$ for all $v_h \in V_h$ ($\|\cdot\|_{W_h'}$ should not be confused with $\|\cdot\|_{W'}$). Hence, the operator $B := A_h$ satisfies the assumptions of Lemma C.44 with $Y := V_h$, $Z := W_h'$, and $\beta := \alpha_h$. We infer that $A_h^* : W_h \to V_h'$ has a (linear) right inverse $A_h^{*\dagger} : V_h' \to W_h$ such that $\|A_h^{*\dagger}\|_{\mathcal{L}(V_h', W_h)} \leq \alpha_h^{-1}$. Let us now consider the operator $B : W \to V_h'$ s.t. $\langle B(w), v_h \rangle_{V_h', V_h} := \overline{a(v_h, w)}$ for all $(v_h, w) \in V_h \times W$, and let us set $\Pi_h := A_h^{*\dagger} \circ B : W \to W_h$. We have

$$a(v_h, \Pi_h(w)) = \langle A_h(v_h), A_h^{*\dagger}(B(w)) \rangle_{W_h', W_h} = \overline{\langle B(w), v_h \rangle_{V_h', V_h}} = a(v_h, w),$$

so that $a(v_h, \Pi_h(w) - w) = 0$. Moreover, we have $\|\Pi_h(w)\|_W \leq \frac{\|a\|}{\alpha_h}\|w\|_W$ since $\|A_h^{*\dagger}\|_{\mathcal{L}(V_h'; W_h)} = \alpha_h^{-1}$ and $\|B\|_{\mathcal{L}(W; V_h')} \leq \|a\|$. Finally, since $B_{|W_h} = A_h^*$, we have $\Pi_h \circ \Pi_h = (A_h^{*\dagger} \circ B) \circ (A_h^{*\dagger} \circ B) = A_h^{*\dagger} \circ (A_h^* \circ A_h^{*\dagger}) \circ B = \Pi_h$, which proves that Π_h is idempotent. □

Remark 26.10 (Dimension, equivalence). We did not assume that V_h and W_h have the same dimension. This level of generality is useful to apply Lemma 26.9 to mixed finite element approximations; see Chapter 50. The implication (i) $\Longrightarrow$ (ii) in Lemma 26.9 is known in the literature as Fortin's lemma [201], and is useful to analyze mixed finite element approximations (see, e.g., Chapter 54 on the Stokes equations). The converse implication can be found in Girault and Raviart [217, p. 117]. This statement is useful in the analysis of Petrov–Galerkin methods; see Carstensen et al. [111], Muga and van der Zee [308], and also Exercise 50.7. Note that the gap in the stability constant γ_{Π_h} between the direct and the converse statements is equal to the condition number $\kappa(a) := \frac{\|a\|}{\alpha}$ of the sesquilinear form a (see Remark 25.12). Finally, we observe that the Fortin operator is not uniquely defined. □

Remark 26.11 (Banach spaces). Lemma 26.9 can be extended to Banach spaces. Such a construction is done in [187], where Lemma C.42 is invoked to build a (bounded) right inverse of A_h^*, and where the proposed map Π_h is nonlinear. Whether one can always construct a Fortin operator Π_h that is linear in Banach spaces seems to be an open question. □

26.3 Basic error estimates

In this section, we assume that the exact problem (26.1) and the discrete problem (26.3) are well-posed. Our goal is to bound the approximation error $(u - u_h)$ in the simple setting where the approximation is conforming ($V_h \subset V$, $W_h \subset W$, $a_h := a_{|V_h \times W_h}$, and $\ell_h := \ell_{|W_h}$).

26.3.1 Strong consistency: Galerkin orthogonality

The starting point of the error analysis is to make sure that the discrete problem (26.3) is *consistent* with the original problem (26.1). Loosely speaking one way of checking consistency is to insert the exact solution into the discrete problem and to verify that the discrepancy is small. We say that there is *strong consistency* whenever this operation is possible and the discrepancy is actually zero. A more general definition of consistency is given in the next chapter. The following result, known as the Galerkin orthogonality property, expresses the fact that strong consistency holds true in the present setting.

Lemma 26.12 (Galerkin orthogonality). *Assume that $V_h \subset V$, $W_h \subset W$, $a_h := a_{|V_h \times W_h}$, and $\ell_h := \ell_{|W_h}$. The following holds true:*

$$a(u, w_h) = \ell(w_h) = a(u_h, w_h), \qquad \forall w_h \in W_h. \tag{26.11}$$

In particular, we have $a(u - u_h, w_h) = 0$ for all $w_h \in W_h$.

Proof. The first equality follows from $W_h \subset W$ and the second one from $a_h := a_{|V_h \times W_h}$ and $\ell_h := \ell_{|W_h}$. $\qquad\square$

26.3.2   Céa's and Babuška's lemmas

Lemma 26.13 (Céa). *Assume that $W_h = V_h \subset V = W$, $a_h := a_{|V_h \times V_h}$, and $\ell_h := \ell_{|V_h}$. Assume that the sesquilinear form a is V-coercive with constant $\alpha > 0$ and let $\|a\|$ be its boundedness constant defined in (26.2) (with $W = V$). Then the following error estimate holds true:*

$$\|u - u_h\|_V \leq \frac{\|a\|}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_V. \tag{26.12}$$

Moreover, if the sesquilinear form a is Hermitian, the error estimate (26.12) can be sharpened as follows:

$$\|u - u_h\|_V \leq \left(\frac{\|a\|}{\alpha}\right)^{\frac{1}{2}} \inf_{v_h \in V_h} \|u - v_h\|_V. \tag{26.13}$$

Proof. Invoking the coercivity of a (stability), followed by the Galerkin orthogonality property (strong consistency) and the boundedness of a, gives

$$\begin{aligned}
\alpha \|u - u_h\|_V^2 &\leq \Re(\xi a(u - u_h, u - u_h)) \\
&= \Re(\xi a(u - u_h, u - v_h)) \\
&\leq \|a\| \, \|u - u_h\|_V \|u - v_h\|_V,
\end{aligned}$$

for all v_h in V_h. This proves the error estimate (26.12). Assume now that the sesquilinear form a is Hermitian. Let v_h be arbitrary in V_h. Let us set $e := u - u_h$ and $\eta_h := u_h - v_h$. The Galerkin orthogonality property and the Hermitian symmetry of a imply that $a(e, \eta_h) = a(\eta_h, e) = 0$. Hence, we have

$$a(u - v_h, u - v_h) = a(e + \eta_h, e + \eta_h) = a(e, e) + a(\eta_h, \eta_h),$$

and the coercivity of a implies that $\Re(\xi a(e, e)) \leq \Re(\xi a(u - v_h, u - v_h))$. Combining this bound with the stability and boundedness properties of a yields

$$\begin{aligned}
\alpha \|u - u_h\|_V^2 &\leq \Re(\xi a(u - u_h, u - u_h)) = \Re(\xi a(e, e)) \\
&\leq \Re(\xi a(u - v_h, u - v_h)) \leq \|a\| \, \|u - v_h\|_V^2.
\end{aligned}$$

Taking the infimum over $v_h \in V_h$ proves the error estimate (26.13). $\qquad\square$

We now extend Céa's lemma to the more general case where stability relies on a discrete inf-sup condition rather than a coercivity argument. Thus, the discrete spaces V_h and W_h can differ.

Lemma 26.14 (Babuška). *Assume that $V_h \subset V$, $W_h \subset W$, $a_h := a_{|V_h \times W_h}$, $\ell_h := \ell_{|W_h}$, and $\dim(V_h) = \dim(W_h)$. Equip V_h and W_h with the norms of V and W, respectively. Assume the following discrete inf-sup condition:*

$$\inf_{v_h \in V_h} \sup_{w_h \in W_h} \frac{|a(v_h, w_h)|}{\|v_h\|_V \|w_h\|_W} =: \alpha_h > 0. \tag{26.14}$$

Let $\|a\|$ be the boundedness constant of a defined in (26.2)*. The following error estimate holds true:*

$$\|u - u_h\|_V \leq \left(1 + \frac{\|a\|}{\alpha_h}\right) \inf_{v_h \in V_h} \|u - v_h\|_V. \qquad (26.15)$$

Proof. Let $v_h \in V_h$. Using stability (i.e., (26.14)), strong consistency (i.e., the Galerkin orthogonality property), and the boundedness of a, we infer that

$$\alpha_h \|u_h - v_h\|_V \leq \sup_{w_h \in W_h} \frac{|a(u_h - v_h, w_h)|}{\|w_h\|_W}$$

$$= \sup_{w_h \in W_h} \frac{|a(u - v_h, w_h)|}{\|w_h\|_W} \leq \|a\| \, \|u - v_h\|_V,$$

and (26.15) follows from the triangle inequality. $\qquad \square$

The error estimates from Lemma 26.13 and from Lemma 26.14 are said to be *quasi-optimal* since $\inf_{v_h \in V_h} \|u - v_h\|_V$ is the best-approximation error of u by an element in V_h, and by definition $\|u - u_h\|_V$ cannot be smaller than the best-approximation error, i.e., the following two-sided error bound holds:

$$\inf_{v_h \in V_h} \|u - v_h\|_V \leq \|u - u_h\|_V \leq c \inf_{v_h \in V_h} \|u - v_h\|_V, \qquad (26.16)$$

with $c := \frac{\|a\|}{\alpha}$ for Céa's lemma and $c := 1 + \frac{\|a\|}{\alpha_h}$ for Babuška's lemma. One noteworthy consequence of (26.16) is that $u_h = u$ whenever the exact solution turns out to be in V_h.

Corollary 26.15 (Convergence). *We have $\lim_{h \to 0} \|u - u_h\|_V = 0$ if the assumptions of Lemma 26.14 hold true together with the following properties:*

(i) *Uniform stability: $\alpha_h \geq \alpha_0 > 0$ for all $h \in \mathcal{H}$.*

(ii) *Approximability: $\lim_{h \to 0}(\inf_{v_h \in V_h} \|v - v_h\|_V) = 0$ for all $v \in V$.*

Proof. Direct consequence of the assumptions. $\qquad \square$

Remark 26.16 (Céa). In the context of Céa's lemma, uniform stability follows from coercivity. Thus, approximability implies convergence. $\qquad \square$

Remark 26.17 (Literature). Lemma 26.13 is derived in [114, Prop. 3.1] and is usually called Céa's lemma in the literature; see, e.g., Ciarlet [124, Thm. 2.4.1], Brenner and Scott [87, Thm. 2.8.1]. Lemma 26.14 is derived in Babuška [33, Thm. 2.2]. $\qquad \square$

26.3.3 Approximability by finite elements

Let us present an important example where the approximability property identified in Corollary 26.15 holds true. Let $V := H^1(D)$ where D is a Lipschitz polyhedron in $\mathbb{R}^d$. Let $V_h := P_k^{\text{g}}(\mathcal{T}_h) \subset H^1(D)$ be the H^1-conforming finite element space of degree $k \geq 1$ (see (20.1)), where $(\mathcal{T}_h)_{h \in \mathcal{H}}$ is a shape-regular sequence of affine meshes so that each mesh covers D exactly. One way to prove approximability is to consider the Lagrange interpolation operator or the canonical interpolation operator (see §19.3), i.e., let us set either $\mathcal{I}_h := \mathcal{I}_h^{\text{L}}$ or $\mathcal{I}_h := \mathcal{I}_h^{\text{g}}$ (we omit the subscript k for simplicity), so that $\mathcal{I}_h : V^{\text{g}}(D) \to P_k^{\text{g}}(\mathcal{T}_h)$ with domain $V^{\text{g}}(D) := H^s(D)$, $s > \frac{d}{2}$ (see (19.19) with $p := 2$). Let l be the smallest integer s.t. $l > \frac{d}{2}$. Setting $r := \min(l - 1, k)$, Corollary 19.8 with $m := 1$ (note that $r \geq 1$) implies that

$$\inf_{v_h \in V_h} \|v - v_h\|_{H^1(D)} \leq \|v - \mathcal{I}_h(v)\|_{H^1(D)} \leq c \, h^r \ell_D |v|_{H^{1+r}(D)},$$

for all $v \in H^{1+r}(D)$, where ℓ_D is a characteristic length of D, e.g., $\ell_D := \mathrm{diam}(D)$. Another possibility consists of using the quasi-interpolation operator $\mathcal{I}_h^{\mathrm{g,av}} : L^1(D) \to V_h$ from Chapter 22 since Theorem 22.6 implies that

$$\inf_{v_h \in V_h} \|v - v_h\|_{H^1(D)} \le \|v - \mathcal{I}_h^{\mathrm{g,av}}(v)\|_{H^1(D)} \le c\, h^r \ell_D |v|_{H^{1+r}(D)},$$

for all $v \in H^{1+r}(D)$ and all $r \in (0, k]$. We now establish approximability by invoking a density argument. Let $v \in V$ and let $\epsilon > 0$. Since $H^{1+r}(D)$ is dense in V for all $r > 0$, there is $v_\epsilon \in H^{1+r}(D)$ s.t. $\|v - v_\epsilon\|_{H^1(D)} \le \epsilon$. Using the triangle inequality and the above interpolation estimates, we infer that

$$\begin{aligned}
\inf_{v_h \in V_h} \|v - v_h\|_{H^1(D)} &\le \|v - \mathcal{I}_h^{\mathrm{g,av}}(v_\epsilon)\|_{H^1(D)} \\
&\le \|v - v_\epsilon\|_{H^1(D)} + \|v_\epsilon - \mathcal{I}_h^{\mathrm{g,av}}(v_\epsilon)\|_{H^1(D)} \\
&\le \epsilon + c\, h^r \ell_D |v_\epsilon|_{H^{1+r}(D)}.
\end{aligned}$$

Letting $h \to 0$ shows that $\limsup_{h \to 0} (\inf_{v_h \in V_h} \|v - v_h\|_{H^1(D)}) \le \epsilon$, and since $\epsilon > 0$ is arbitrary, we conclude that approximability holds true, i.e., the best-approximation error in V_h of any function $v \in V$ tends to zero as $h \to 0$. The above arguments can be readily adapted when homogeneous Dirichlet conditions are strongly enforced.

26.3.4 Sharper error estimates

We now sharpen the constant appearing in the error estimate (26.15) from Lemma 26.14. Let $V_h \subset V$ and $W_h \subset W$ with $\dim(V_h) = \dim(W_h)$, and let a be a bounded sesquilinear form on $V \times W$ satisfying the discrete inf-sup condition (26.14) on $V_h \times W_h$. We define the *discrete solution map* $G_h : V \to V_h$ s.t. for all $v \in V$, $G_h(v)$ is the unique element in V_h satisfying

$$a(G_h(v) - v, w_h) = 0, \qquad \forall w_h \in W_h. \tag{26.17}$$

Note that $G_h(v)$ is well defined owing to the discrete inf-sup condition (26.14) and since $a(v, \cdot) : W_h \to \mathbb{C}$ is a bounded antilinear form on W_h. Moreover, G_h is linear and V_h is pointwise invariant under G_h.

Lemma 26.18 (Xu–Zikatanov). *Let $\{0\} \subsetneq V_h \subsetneq V$ and $W_h \subset W$ with $\dim(V_h) = \dim(W_h)$ where V, W are Hilbert spaces, and let a be a bounded sesquilinear form on $V \times W$ with constant $\|a\|$ defined in (26.2) satisfying the discrete inf-sup condition (26.14) on $V_h \times W_h$ with constant α_h. Then,*

$$\|u - u_h\|_V \le \frac{\|a\|}{\alpha_h} \inf_{v_h \in V_h} \|u - v_h\|_V. \tag{26.18}$$

Proof. Since G_h is linear and V_h is pointwise invariant under G_h, we have

$$u - u_h = u - G_h(u) = (u - v_h) - G_h(u - v_h),$$

for all $v_h \in V_h$. We infer that

$$\|u - u_h\|_V \le \|I - G_h\|_{\mathcal{L}(V)} \|u - v_h\|_V = \|G_h\|_{\mathcal{L}(V)} \|u - v_h\|_V,$$

where the last equality follows from the fact that in any Hilbert space H, any operator $T \in \mathcal{L}(H)$ such that $0 \ne T \circ T = T \ne I$ verifies $\|T\|_{\mathcal{L}(H)} = \|I - T\|_{\mathcal{L}(H)}$ (see the proof of Theorem 5.14). We can apply this result to the discrete solution map since $G_h \ne 0$ (since $V_h \ne \{0\}$), $G_h \circ G_h = G_h$

(since V_h is pointwise invariant under G_h), and $G_h \neq I$ (since $V_h \neq V$). To conclude the proof, we bound $\|G_h\|_{\mathcal{L}(V)}$ as follows: For all $v \in V$,

$$\alpha_h \, \|G_h(v)\|_V \leq \sup_{w_h \in W_h} \frac{|a(G_h(v), w_h)|}{\|w_h\|_W} = \sup_{w_h \in W_h} \frac{|a(v, w_h)|}{\|w_h\|_W} \leq \|a\| \, \|v\|_V,$$

which shows that $\|G_h\|_{\mathcal{L}(V)} \leq \frac{\|a\|}{\alpha_h}$. □

Let Λ be the smallest c so that the inequality $\frac{\|u - u_h\|_V}{\inf_{v_h \in V_h} \|u - v_h\|_V} \leq c$ holds for every $u \in V$. Then $\Lambda = \sup_{u \in V} \sup_{v_h \in V_h} \frac{\|u - G_h(u)\|_V}{\|u - v_h\|_V}$ since $u_h = G_h(u)$. But the proof of Lemma 26.18 shows that $\Lambda = \|I - G_h\|_{\mathcal{L}(V)} = \|G_h\|_{\mathcal{L}(V)}$. Hence, $\|G_h\|_{\mathcal{L}(V)}$ is the smallest constant such that the following quasi-optimal error estimate holds:

$$\|u - u_h\|_V \leq \|G_h\|_{\mathcal{L}(V)} \inf_{v_h \in V_h} \|u - v_h\|_V.$$

Thus, sharp estimates on $\|G_h\|_{\mathcal{L}(V)}$ are important to determine whether the approximation error is close or not to the best-approximation error. The following result shows in particular that $\|G_h\|_{\mathcal{L}(V)}$ is, up to a factor in the interval $[\alpha, \|a\|]$, proportional to the inverse of the discrete inf-sup constant α_h.

Lemma 26.19 (Tantardini–Veeser). *Under the assumptions of Lemma 26.18, the following holds true:*

$$\|G_h\|_{\mathcal{L}(V)} = \sup_{w_h \in W_h} \frac{\left(\sup_{v \in V} \dfrac{|a(v, w_h)|}{\|v\|_V} \right)}{\left(\sup_{v_h \in V_h} \dfrac{|a(v_h, w_h)|}{\|v_h\|_V} \right)} \geq 1, \tag{26.19a}$$

$$\frac{\alpha}{\alpha_h} \leq \|G_h\|_{\mathcal{L}(V)} \leq \frac{\|a\|}{\alpha_h}. \tag{26.19b}$$

Proof. (1) Let $A \in \mathcal{L}(V; W')$ be the operator associated with the sesquilinear form a, i.e.,

$$\langle A(v), w \rangle_{W', W} := a(v, w), \qquad \forall (v, w) \in V \times W,$$

and let $A^* \in \mathcal{L}(W; V')$ be its adjoint (where we used the reflexivity of W). We have

$$\alpha \, \|w\|_W \leq \|A^*(w)\|_{V'} = \sup_{v \in V} \frac{|a(v, w)|}{\|v\|_V} \leq \|a\| \, \|w\|_W, \tag{26.20}$$

for all $w \in W$. Indeed, the first bound follows from Lemma C.53 and the inf-sup stability of a, and the second one follows from the boundedness of a. This shows that the norms $\|\cdot\|_W$ and $\|A^*(\cdot)\|_{V'}$ are equivalent on W.
(2) Since $W_h \subset W$, we have $A^*(w_h) \in V'$ for all $w_h \in W_h$. Upon setting

$$\gamma_h := \inf_{w_h \in W_h} \sup_{v_h \in V_h} \frac{|a(v_h, w_h)|}{\|v_h\|_V \|A^*(w_h)\|_{V'}},$$

we have $\gamma_h \geq \frac{\alpha_h}{\|a\|} > 0$ owing to the inf-sup condition satisfied by a on $V_h \times W_h$, the norm equivalence (26.20), and Lemma C.53. Recalling that $\|A^*(w_h)\|_{V'} = \sup_{v \in V} \frac{|a(v, w_h)|}{\|v\|_V}$, the assertion (26.19a) amounts to $\|G_h\|_{\mathcal{L}(V)} = \gamma_h^{-1} \geq 1$.

(3) Let $w_h \in W_h$. Using the definition (26.17) of the discrete solution map and the definition of the dual norm $\|A^*(w_h)\|_{V'}$, we have

$$
\begin{aligned}
\|A^*(w_h)\|_{V'} &= \sup_{v \in V} \frac{|a(G_h(v), w_h)|}{\|v\|_V} \\
&\le \sup_{v \in V} \frac{|a(G_h(v), w_h)|}{\|G_h(v)\|_V} \sup_{v \in V} \frac{\|G_h(v)\|_V}{\|v\|_V} \\
&\le \sup_{v_h \in V_h} \frac{|a(v_h, w_h)|}{\|v_h\|_V} \|G_h\|_{\mathcal{L}(V)}.
\end{aligned}
$$

Rearranging the terms and taking the infimum over $w_h \in W_h$ shows that $\gamma_h \ge \|G_h\|_{\mathcal{L}(V)}^{-1}$, i.e., $\|G_h\|_{\mathcal{L}(V)} \ge \gamma_h^{-1}$.
(4) Since $\gamma_h > 0$, Remark 26.8 implies that

$$
\gamma_h = \inf_{v_h \in V_h} \sup_{w_h \in W_h} \frac{|a(v_h, w_h)|}{\|v_h\|_V \|A^*(w_h)\|_{V'}}. \tag{26.21}
$$

Let $v \in V$. Applying the above identity to $G_h(v) \in V_h$, we infer that

$$
\gamma_h \|G_h(v)\|_V \le \sup_{w_h \in W_h} \frac{|a(G_h(v), w_h)|}{\|A^*(w_h)\|_{V'}} = \sup_{w_h \in W_h} \frac{|a(v, w_h)|}{\|A^*(w_h)\|_{V'}} \le \|v\|_V,
$$

since $|a(v, w_h)| = |\langle A^*(w_h), v\rangle_{V',V}| \le \|A^*(w_h)\|_{V'} \|v\|_V$. Taking the supremum over $v \in V$ shows that $\|G_h\|_{\mathcal{L}(V)} \le \gamma_h^{-1}$. Thus, we have proved that $\|G_h\|_{\mathcal{L}(V)} = \gamma_h^{-1}$, and the lower bound in (26.19a) is a direct consequence of $V_h \subset V$.
(5) It remains to prove (26.19b). Using the norm equivalence (26.20) in (26.21) to bound from below and from above $\|A^*(w_h)\|_{V'}$, we infer that

$$
\frac{1}{\|a\|} \inf_{v_h \in V_h} \sup_{w_h \in W_h} \frac{|a(v_h, w_h)|}{\|v_h\|_V \|w_h\|_W} \le \gamma_h \le \frac{1}{\alpha} \inf_{v_h \in V_h} \sup_{w_h \in W_h} \frac{|a(v_h, w_h)|}{\|v_h\|_V \|w_h\|_W},
$$

so that $\frac{\alpha_h}{\|a\|} \le \gamma_h \le \frac{\alpha_h}{\alpha}$, and (26.19b) follows from $\|G_h\|_{\mathcal{L}(V)} = \gamma_h^{-1}$. $\qquad\square$

Remark 26.20 (Literature). Lemma 26.18 is proved in Xu and Zikatanov [397, Thm. 2], and Lemma 26.19 in Tantardini and Veeser [361, Thm. 2.1]. See also Arnold et al. [18] for the lower bound $\frac{\alpha}{\alpha_h} \le \|G_h\|_{\mathcal{L}(V)}$. $\qquad\square$

Remark 26.21 (Discrete dual norm). For all $w_h \in W_h$, $A^*(w_h) \in V'$ can be viewed as a member of V_h' by restricting its action to the subspace $V_h \subset V$. We use the same notation and simply write $A^*(w_h) \in V_h'$. The statement (26.19a) in Lemma 26.19 can be rewritten as follows:

$$
\|G_h\|_{\mathcal{L}(V)} = \sup_{w_h \in W_h} \frac{\|A^*(w_h)\|_{V'}}{\|A^*(w_h)\|_{V_h'}}, \tag{26.22}
$$

where $\|A^*(w_h)\|_{V_h'} := \sup_{v_h \in V_h} \frac{|\langle A^*(w_h), v_h \rangle_{V',V}|}{\|v_h\|_V} = \sup_{v_h \in V_h} \frac{|a(v_h, w_h)|}{\|v_h\|_V}$. $\qquad\square$

Example 26.22 (Orthogonal projection). Let $V \hookrightarrow L$ be two Hilbert spaces with continuous and dense embedding. Using the Riesz–Fréchet theorem (Theorem C.24), we identify L with its dual L' by means of the inner product $(\cdot, \cdot)_L$ in L. This allows us to define the continuous embedding $E_{V'} : V \to V'$ s.t. $\langle E_{V'}(v), w \rangle_{V',V} := (v, w)_L$ for all $v, w \in V$. Note that $E_{V'}$ is self-adjoint. Consider a subspace $\{0\} \subsetneq V_h \subsetneq V$. Let $\mathcal{P}_h$ be the discrete solution map associated with

the sesquilinear form $a(v,w) := \langle E_{V'}(v), w \rangle_{V',V}$ for all $v, w \in V$. Note that $\mathcal{P}_h$ is the L-orthogonal projection onto V_h since

$$(\mathcal{P}_h(v), w_h)_L = \langle E_{V'}(\mathcal{P}_h(v)), w_h \rangle_{V',V} := \langle E_{V'}(v), w_h \rangle_{V',V} = (v, w_h)_L,$$

for all $v \in V$ and all $w_h \in V_h$. Then Lemma 26.19 provides a precise estimate on the V-stability of $\mathcal{P}_h$ in the form

$$\|\mathcal{P}_h\|_{\mathcal{L}(V)}^{-1} = \inf_{w_h \in V_h} \frac{\|E_{V'}(w_h)\|_{V_h'}}{\|E_{V'}(w_h)\|_{V'}} = \inf_{w_h \in V_h} \sup_{v_h \in V_h} \frac{|(w_h, v_h)_L|}{\|E_{V'}(w_h)\|_{V'}\|v_h\|_V}. \tag{26.23}$$

See also Tantardini and Veeser [361, Prop. 2.5], Andreev [11, Lem. 6.2]. An important example is $V := H_0^1(D)$ and $L := L^2(D)$. The reader is referred to §22.5 for further discussion on the L^2-orthogonal projection onto conforming finite element spaces (see in particular Remark 22.23 for sufficient conditions on the underlying mesh to ensure H^1-stability). $\qquad\square$

Exercises

Exercise 26.1 ((BNB2)). Prove that (26.8) is equivalent to (26.5b) provided (26.5a) holds true. (*Hint*: use that $\dim(W_h) = \operatorname{rank}(A_h) + \dim(\ker(A_h^*))$ (A_h^* is defined in (26.9)) together with the rank nullity theorem.)

Exercise 26.2 (Bijectivity of A_h^*). Prove that A_h is an isomorphism if and only if A_h^* is an isomorphism. (*Hint*: use $\dim(V_h) = \operatorname{rank}(A_h^*) + \dim(\ker(A_h))$ and $\dim(W_h) = \operatorname{rank}(A_h) + \dim(\ker(A_h^*))$.)

Exercise 26.3 (Petrov–Galerkin). Let V, W be real Hilbert spaces, let $A \in \mathcal{L}(V; W')$ be an isomorphism, and let $\ell \in W'$. Consider a conforming Petrov–Galerkin approximation with a finite-dimensional subspace $V_h \subset V$ and $W_h := (J_W^{\mathrm{RF}})^{-1} A V_h \subset W$, where $J_W^{\mathrm{RF}} : W \to W'$ is the Riesz–Fréchet isomorphism. The discrete bilinear form is $a_h(v_h, w_h) := \langle A(v_h), w_h \rangle_{W',W}$, and the discrete linear form is $\ell_h(w_h) := \ell(w_h)$ for all $v_h \in V_h$ and all $w_h \in W_h$. (i) Prove that the discrete problem (26.3) is well-posed. (ii) Show that its unique solution minimizes the residual functional $\mathfrak{R}(v) := \|A(v) - \ell\|_{W'}$ over V_h.

Exercise 26.4 (Fortin's lemma). (i) Prove that Π_h in the converse statement of Lemma 26.9 is idempotent. (*Hint*: prove that $B \circ A_h^{*\dagger} = I_{V_h'}$.) (ii) Assume that there are two maps $\Pi_{1,h}, \Pi_{2,h} : W \to W_h$ and two uniform constants $c_1, c_2 > 0$ such that $\|\Pi_{1,h}(w)\|_W \le c_1 \|w\|_W$, $\|\Pi_{2,h}((I - \Pi_{1,h})(w))\|_W \le c_2 \|w\|_W$ and $a(v_h, \Pi_{2,h}(w) - w) = 0$ for all $v_h \in V_h$, $w \in W$. Prove that $\Pi_h := \Pi_{1,h} + \Pi_{2,h}(I - \Pi_{1,h})$ is a Fortin operator. (iii) Write a variant of the direct statement in Lemma 26.9 assuming V, W reflexive, $A \in \mathcal{L}(V; W')$ bijective, and using this time an operator $\Pi_h : V \to V_h$ such that $a(\Pi_h(v) - v, w_h) = 0$ for all $(v, w_h) \in V \times W_h$ and $\gamma_{\Pi_h} \|\Pi_h(v)\|_V \le \|v\|_V$ for all $v \in V$ for some $\gamma_{\Pi_h} > 0$. (*Hint*: use (26.10) and Lemma C.53.)

Exercise 26.5 (Compact perturbation). Let V, W be Banach spaces with W reflexive. Let $A_0 \in \mathcal{L}(V; W')$ be bijective, let $T \in \mathcal{L}(V; W')$ be compact, and assume that $A := A_0 + T$ is injective. Let $a_0(v, w) := \langle A_0(v), w \rangle_{W',W}$ and $a(v, w) := \langle A(v), w \rangle_{W',W}$ for all $(v, w) \in V \times W$. Let $V_h \subset V$ and $W_h \subset W$ be s.t. $\dim(V_h) = \dim(W_h)$ for all $h \in \mathcal{H}$. Assume that approximability holds, and that the sesquilinear form a_0 satisfies the inf-sup condition

$$\inf_{v_h \in V_h} \sup_{w_h \in W_h} \frac{|a_0(v_h, w_h)|}{\|v_h\|_V \|w_h\|_W} =: \alpha_0 > 0, \qquad \forall h \in \mathcal{H}.$$

Following Wendland [392], the goal is to show that there is $h_0 > 0$ s.t.

$$\inf_{v_h \in V_h} \sup_{w_h \in W_h} \frac{|a(v_h, w_h)|}{\|v_h\|_V \|w_h\|_W} =: \alpha > 0, \qquad \forall h \in \mathcal{H} \cap (0, h_0].$$

(i) Prove that $A \in \mathcal{L}(V; W')$ is bijective. (*Hint*: recall that a compact operator is bijective iff it is injective; this follows from the Fredholm alternative, Theorem 46.13.) (ii) Consider $R_h \in \mathcal{L}(V; V_h)$ s.t. for all $v \in V$, $R_h(v) \in V_h$ satisfies $a_0(R_h(v) - v, w_h) = 0$ for all $w_h \in W_h$. Prove that $R_h \in \mathcal{L}(V; V_h)$ and that $R_h(v)$ converges to v as $h \downarrow 0$ for all $v \in V$. (*Hint*: proceed as in the proof of Céa's lemma.) (iii) Set $L := I_V + A_0^{-1} T$ and $L_h := I_V + R_h A_0^{-1} T$ where I_V is the identity operator in V (observe that both L and L_h are in $\mathcal{L}(V)$). Prove that L_h converges to L in $\mathcal{L}(V)$. (*Hint*: use Remark C.5.) (iv) Show that if $h \in \mathcal{H}$ is small enough, L_h is bijective and there is C, independent of $h \in \mathcal{H}$, such that $\|L_h^{-1}\|_{\mathcal{L}(V)} \le C$. (*Hint*: observe that $L^{-1} L_h = I_V - L^{-1}(L - L_h)$ and consider the Neumann series.) (v) Conclude.

Chapter 27

Error analysis with variational crimes

We have shown in the previous chapter how the Galerkin method can be used to approximate the solution to the model problem (26.1), and we have derived an error estimate in the simple setting where $V_h \subset V$, $W_h \subset W$, $a_h := a_{|V_h \times W_h}$, and $\ell_h := \ell_{|W_h}$. Departures from this setting are often called *variational crimes* in the literature. In this chapter, we perform the error analysis when variational crimes are committed. The main results, Lemma 27.5 and Lemma 27.8, will be invoked frequently in this book. They give an upper bound on the approximation error in terms of the best-approximation error of the exact solution by members of the discrete trial space. These error estimates are based on the notions of stability and consistency/boundedness. Combined with an approximability property, they allow us to conclude that the approximation method is convergent. Two simple examples illustrate the theory: a first-order PDE approximated by the Galerkin/least-squares technique and a second-order PDE approximated by a boundary penalty method.

27.1 Setting

In the entire chapter, we suppose that the assumptions of the BNB theorem (Theorem 25.9 or its variant Theorem 25.15) are satisfied, so that the exact problem (26.1) is well-posed. The inf-sup and boundedness constants on $V \times W$ of the exact sesquilinear form a are denoted by α and $\|a\|$; see (26.2). The exact solution is denoted by $u \in V$.

Recall that the Galerkin approximation (26.3) relies on the discrete trial space V_h and the discrete test space W_h. These spaces are equipped with the norms $\|\cdot\|_{V_h}$ and $\|\cdot\|_{W_h}$, respectively. The discrete problem uses a discrete sesquilinear form a_h defined on $V_h \times W_h$ and a discrete antilinear form ℓ_h defined on W_h. The sesquilinear form a_h and the antilinear form ℓ_h must be viewed, respectively, as some approximations to a and ℓ. The solution to the discrete problem (26.3) is denoted by $u_h \in V_h$. We always assume that $\dim(V_h) = \dim(W_h)$, so that the well-posedness of the discrete problem is equivalent to the following inf-sup condition:

$$\inf_{v_h \in V_h} \sup_{w_h \in W_h} \frac{|a_h(v_h, w_h)|}{\|v_h\|_{V_h} \|w_h\|_{W_h}} =: \alpha_h > 0. \tag{27.1}$$

We say that the approximation (26.3) is *stable* whenever (27.1) holds true, i.e., $\alpha_h > 0$.

The goal of this chapter is to bound the error, i.e., we want to estimate how far the discrete solution $u_h \in V_h$ lies from the exact solution $u \in V$. We say that the method *converges* if the error tends to zero as the approximation capacity of the discrete trial space V_h increases. The approximation capacity of V_h increases by refining an underlying mesh. We will see that there are three key properties to establish convergence: (i) stability, (ii) consistency/boundedness, and (iii) approximability. Stability and approximability have already emerged as important notions in the error analysis presented in §26.3. The notion of consistency was present in the simple form of the Galerkin orthogonality property, and the boundedness of the sesquilinear form a on $V \times W$ was also invoked.

Remark 27.1 (Lax principle). A loose principle in numerical analysis, known as *Lax Principle*, is that stability and consistency imply convergence. The fact that boundedness and approximability are not mentioned does not mean that these properties should be taken for granted. We refer the reader to the upcoming chapters for numerous examples. $\qquad\square$

Remark 27.2 (Norms). Since all the norms are equivalent in finite-dimensional vector spaces, if (27.1) holds true for one choice of norms in V_h and W_h, it holds true also for every other choice. The goal is to select norms s.t. (i) a_h is *uniformly stable*, i.e., $\alpha_h \geq \alpha_0 > 0$ for all $h \in \mathcal{H}$, and (ii) a_h is *uniformly bounded* on $V_h \times W_h$ with respect to $h \in \mathcal{H}$. $\qquad\square$

27.2   Main results

This section contains our two main abstract error estimates.

27.2.1   The spaces V_{s} and $V_\sharp$

In a nonconforming approximation setting where $V_h \not\subset V$, the exact solution u and the discrete solution u_h may be objects of different nature. This poses the question of how to measure the approximation error. For instance, does the expression $(u - u_h)$ make sense? We are going to assume that it is possible to define a common ground between u and u_h to evaluate the error. A simple way to do this is to assume that it is meaningful to define the linear space $(V + V_h)$. If it is indeed the case, then the error belongs to this space.

However, we will see in numerous examples that the error analysis often requires to assume that the exact solution has slightly more smoothness than just being a member of V. We formalize this assumption by introducing a functional space V_{s} such that $u \in V_{\mathsf{s}} \subseteq V$. Our setting for the error analysis is therefore as follows:

$$u \in V_{\mathsf{s}} \subseteq V, \qquad u - u_h \in V_\sharp := V_{\mathsf{s}} + V_h. \tag{27.2}$$

Note that this setting allows for $V_{\mathsf{s}} := V$, and in the conforming setting, where $V_h \subset V$, this then implies that $V_\sharp := V$.

27.2.2   Consistency/boundedness

A crucial notion in the error analysis is that of consistency/boundedness. Loosely speaking the idea behind consistency is to insert the exact solution into the discrete equations and to verify that the discrepancy is small. This may not be possible in a nonconforming approximation setting because it may turn out that the discrete sesquilinear form a_h is not meaningful when its first

argument is the exact solution. To stay general, we are going to define a consistency error for every discrete trial function $v_h \in V_h$ with the expectation that this error is small if the difference $(u - v_h) \in V_\sharp$ is small. Let us now formalize this idea. Recall that the norm of any antilinear form $\phi_h \in W_h' := \mathcal{L}(W_h; \mathbb{C})$ is defined by $\|\phi_h\|_{W_h'} := \sup_{w_h \in W_h} \frac{|\phi_h(w_h)|}{\|w_h\|_{W_h}}$.

Definition 27.3 (Consistency/boundedness). *Let $\delta_h : V_h \to W_h'$ be defined by setting*

$$\langle \delta_h(v_h), w_h \rangle_{W_h', W_h} := \ell_h(w_h) - a_h(v_h, w_h) = a_h(u_h - v_h, w_h). \qquad (27.3)$$

The quantity $\|\delta_h(v_h)\|_{W_h'}$ is called consistency error *for the discrete trial function $v_h \in V_h$. We say that* consistency/boundedness *holds true if the space $V_\sharp$ can be equipped with a norm $\|\cdot\|_{V_\sharp}$ such that there is a real number $\omega_{\sharp h}$, uniform w.r.t. $u \in V_S$, such that for all $v_h \in V_h$ and all $h \in \mathcal{H}$,*

$$\|\delta_h(v_h)\|_{W_h'} \leq \omega_{\sharp h} \|u - v_h\|_{V_\sharp}. \qquad (27.4)$$

Example 27.4 (Simple setting). Assume conformity (i.e., $V_h \subset V$ and $W_h \subset W$), $a_h := a_{|V_h \times W_h}$, and $\ell_h := \ell_{|W_h}$. Take $V_S := V$, so that $V_\sharp := V$, and take $\|\cdot\|_{V_\sharp} := \|\cdot\|_V$. The consistency error (27.3) is such that

$$\langle \delta_h(v_h), w_h \rangle_{W_h', W_h} = \ell(w_h) - a(v_h, w_h) = a(u - v_h, w_h),$$

where we used that $\ell(w_h) = a(u, w_h)$ (i.e., the Galerkin orthogonality property). Since a is bounded on $V \times W$, (27.4) holds true with $\omega_{\sharp h} := \|a\|$. $\qquad \square$

27.2.3 Error estimate using one norm

We can now establish our first abstract error estimate. This estimate will be applied to various nonconforming approximation settings of elliptic PDEs. It hinges on the assumption that there is a real number $c_\sharp$, uniform w.r.t. $h \in \mathcal{H}$, s.t.

$$\|v_h\|_{V_\sharp} \leq c_\sharp \|v_h\|_{V_h}, \qquad \forall v_h \in V_h. \qquad (27.5)$$

Recall that $\|\cdot\|_{V_h}$ is the stability norm on V_h used in (27.1) and $\|\cdot\|_{V_\sharp}$ is the consistency/boundedness norm on $V_\sharp$ used in (27.4).

Lemma 27.5 (Quasi-optimal error estimate). *Assume the following:* (i) *Stability, i.e., (27.1) holds true;* (ii) *Consistency/boundedness, i.e., $u \in V_S$ and (27.4) holds true. Assume that (27.5) holds true. Then we have*

$$\|u - u_h\|_{V_\sharp} \leq \left(1 + c_\sharp \frac{\omega_{\sharp h}}{\alpha_h} \right) \inf_{v_h \in V_h} \|u - v_h\|_{V_\sharp}. \qquad (27.6)$$

Proof. Owing to the assumptions, we infer that for all $v_h \in V_h$,

$$\begin{aligned}
\|u - u_h\|_{V_\sharp} &\leq \|u - v_h\|_{V_\sharp} + \|v_h - u_h\|_{V_\sharp} \\
&\leq \|u - v_h\|_{V_\sharp} + c_\sharp \|v_h - u_h\|_{V_h} \\
&\leq \|u - v_h\|_{V_\sharp} + \frac{c_\sharp}{\alpha_h} \sup_{w_h \in W_h} \frac{|a_h(u_h - v_h, w_h)|}{\|w_h\|_{W_h}} \\
&= \|u - v_h\|_{V_\sharp} + \frac{c_\sharp}{\alpha_h} \|\delta_h(v_h)\|_{W_h'} \\
&\leq \|u - v_h\|_{V_\sharp} + \frac{c_\sharp \omega_{\sharp h}}{\alpha_h} \|u - v_h\|_{V_\sharp}.
\end{aligned}$$

Taking the infimum over $v_h \in V_h$ yields (27.6). $\qquad \square$

Example 27.6 (Simple setting). In the setting of Example 27.4, we can equip V_h and $V_\sharp$ with the norm $\|\cdot\|_V$, so that $c_\sharp = 1$. Since $\omega_{\sharp h} = \|a\|$, the error estimate (27.6) coincides with the error estimate in Lemma 26.14. □

Remark 27.7 (Literature). A general framework for the error analysis of nonconforming methods for elliptic PDEs can be found in Veeser and Zanotti [373]. This framework introduces a different notion of consistency and leads to quasi-optimal error estimates in the $\|\cdot\|_V$-norm without any smoothness assumption on the exact solution $u \in V$ (or equivalently for all data $\ell \in V'$), i.e., the space V_S and the norm $\|\cdot\|_{V_\sharp}$ are not invoked. This remarkable result is achieved at the expense of a specific design of the discrete form ℓ_h. We also refer the reader to the gradient discretization method discussed in Droniou et al. [172] which can be used to analyze nonconforming methods. □

27.2.4 Error estimate using two norms

It turns out that the assumption (27.5) on the $\|\cdot\|_{V_\sharp}$-norm cannot be satisfied when one considers the approximation of first-order PDEs using stabilization techniques. A more general setting consists of introducing a second norm on $V_\sharp$, say $\|\cdot\|_{V_\flat}$, and assuming that there exists a real number $c_\flat$ s.t.

$$\|v_h\|_{V_\flat} \le c_\flat \|v_h\|_{V_h}, \ \forall v_h \in V_h, \qquad \|v\|_{V_\flat} \le c_\flat \|v\|_{V_\sharp}, \ \forall v \in V_\sharp, \tag{27.7}$$

where $\|\cdot\|_{V_h}$ is the stability norm on V_h used in (27.1) and $\|\cdot\|_{V_\sharp}$ is the consistency/boundedness norm on $V_\sharp$ used in (27.4).

Lemma 27.8 (Error estimate). *Assume the following:* (i) *Stability, i.e., (27.1) holds true;* (ii) *Consistency/boundedness, i.e., $u \in V_S$ and (27.4) holds true. Assume that (27.7) holds true. Then we have*

$$\|u - u_h\|_{V_\flat} \le c_\flat \left(1 + \frac{\omega_{\sharp h}}{\alpha_h}\right) \inf_{v_h \in V_h} \|u - v_h\|_{V_\sharp}. \tag{27.8}$$

Proof. The proof is similar to that of Lemma 27.5. Owing to the assumptions, we infer that for all $v_h \in V_h$,

$$\begin{aligned}
\|u - u_h\|_{V_\flat} &\le \|u - v_h\|_{V_\flat} + \|v_h - u_h\|_{V_\flat} \\
&\le c_\flat \|u - v_h\|_{V_\sharp} + c_\flat \|v_h - u_h\|_{V_h} \\
&\le c_\flat \|u - v_h\|_{V_\sharp} + \frac{c_\flat}{\alpha_h} \sup_{w_h \in W_h} \frac{|a_h(u_h - v_h, w_h)|}{\|w_h\|_{W_h}} \\
&= c_\flat \|u - v_h\|_{V_\sharp} + \frac{c_\flat}{\alpha_h} \|\delta_h(v_h)\|_{W_h'} \\
&\le c_\flat \|u - v_h\|_{V_\sharp} + \frac{c_\flat \omega_{\sharp h}}{\alpha_h} \|u - v_h\|_{V_\sharp}.
\end{aligned}$$

Taking the infimum over $v_h \in V_h$ yields (27.8). □

Remark 27.9 (Lemma 27.5 vs. Lemma 27.8). Lemma 27.5 estimates the approximation error by the best-approximation error using the same norm $\|\cdot\|_{V_\sharp}$. We say that this estimate is quasi-optimal *over the whole computational range.* In contrast, Lemma 27.8 estimates the approximation error in the $\|\cdot\|_{V_\flat}$-norm by the best-approximation error in the stronger $\|\cdot\|_{V_\sharp}$-norm. We will see numerous examples where the best-approximation errors in both norms actually exhibit the same decay rate in terms of the meshsize $h \in \mathcal{H}$ for smooth solutions. In this situation, we say that the error estimate from Lemma 27.8 is quasi-optimal *in the asymptotic range.* □

27.2.5   Convergence

We are now ready to state a convergence result. The last missing ingredient that we introduce now is approximability.

Corollary 27.10 (Convergence). *We have* $\lim_{h\to 0} \|u - u_h\|_{V_\sharp} = 0$ *in the setting of Lemma* 27.5 *and* $\lim_{h\to 0} \|u - u_h\|_{V_\flat} = 0$ *in the setting of Lemma* 27.8, *provided the following properties hold true:*

(i) *Uniform stability:* $\alpha_h \geq \alpha_0 > 0$ *for all* $h \in \mathcal{H}$;

(ii) *Uniform consistency/boundedness:* $\omega_{\sharp h} \leq \omega_{\sharp 0} < \infty$ *for all* $h \in \mathcal{H}$;

(iii) *Approximability:* $\lim_{h\to 0}(\inf_{v_h \in V_h} \|v - v_h\|_{V_\sharp}) = 0$ *for all* $v \in V_{\mathrm{s}}$.

Proof. Direct consequence of the assumptions. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\square$

27.3   Two simple examples

This section presents two one-dimensional examples illustrating how to use the above error estimates: (i) a boundary penalty method applied to an elliptic PDE where Lemma 27.5 is applied; (ii) a stabilized approximation applied to a first-order PDE where Lemma 27.8 is applied.

27.3.1   Boundary penalty method for an elliptic PDE

Consider the PDE $-u'' = f$ in $D := (0,1)$ with $u(0) = u(1) = 0$, $f \in L^2(D)$. The trial and test spaces are $V = W := H_0^1(D)$. The corresponding bilinear and linear forms are $a(v,w) := \int_0^1 v'w' \, dt$ and $\ell(w) := \int_0^1 fw \, dt$. Consider the standard Galerkin approximation using as discrete trial and test spaces the spaces $V_h = W_h$ built using continuous $\mathbb{P}_1$ Lagrange finite elements on a uniform mesh $\mathcal{T}_h$ of step $h \in \mathcal{H}$. We do not enforce any boundary condition on V_h. As a result, the approximation setting is *nonconforming*. Let us define the discrete forms

$$a_h(v_h, w_h) := \int_0^1 v_h' w_h' \, dt - (v_h'(1)w_h(1) - v_h'(0)w_h(0)) + h^{-1}(v_h(1)w_h(1) + v_h(0)w_h(0)),$$

$$\ell_h(w_h) := \int_0^1 fw_h \, dt.$$

One can show that coercivity holds true with the stability norm

$$\|v_h\|_{V_h}^2 := \|v_h'\|_{L^2(D)}^2 + h^{-1}|v_h(0)|^2 + h^{-1}|v_h(1)|^2,$$

i.e., $a_h(v_h, v_h) \geq \alpha_0 \|v_h\|_{V_h}^2$ with $\alpha_0 := \frac{3}{8}$ for all $v_h \in V_h$; see Exercise 27.2 and Chapter 37.

Let us perform the error analysis using Lemma 27.5. The assumption $u \in V_{\mathrm{s}} := H^2(D) \cap H_0^1(D)$ is natural here since $f \in L^2(D)$ and $-u'' = f$. We equip the space $V_\sharp := V_{\mathrm{s}} + V_h$ with the norm

$$\|v\|_{V_\sharp}^2 := \|v'\|_{L^2(D)}^2 + h^{-1}|v(0)|^2 + h^{-1}|v(1)|^2 + h|v'(0)|^2 + h|v'(1)|^2.$$

(Recall that $H^2(D) \hookrightarrow C^1(\overline{D})$ in one dimension.) Using a discrete trace inequality shows that the norms $\|\cdot\|_{V_h}$ and $\|\cdot\|_{V_\sharp}$ are equivalent on V_h uniformly w.r.t. $h \in \mathcal{H}$. Hence, (27.5) holds true. It remains to establish consistency/boundedness. Since $u \in H^2(D)$, integrating by parts leads to

$$\ell_h(w_h) = -\int_0^1 u'' w_h \, dt = \int_0^1 u' w_h' \, dt - (u'(1)w_h(1) - u'(0)w_h(0)),$$

so that letting $\eta := u - v_h$ and since $u(0) = u(1) = 0$, we obtain

$$\langle \delta_h(v_h), w_h \rangle_{V_h', V_h} = \ell_h(w_h) - a_h(v_h, w_h)$$

$$= \int_0^1 \eta' w_h' \, \mathrm{d}t - (\eta'(1)w_h(1) - \eta'(0)w_h(0)) + h^{-1}(\eta(1)w_h(1) + \eta(0)w_h(0)).$$

Using the Cauchy–Schwarz inequality, we conclude that (27.4) holds true with $\omega_{\sharp h} = 1$. In conclusion, Lemma 27.5 implies that

$$\|u - u_h\|_{V_\sharp} \le c \inf_{v_h \in V_h} \|u - v_h\|_{V_\sharp}. \tag{27.9}$$

Since $u \in H^2(D)$, we use the approximation properties of finite elements to obtain $\inf_{v_h \in V_h} \|u - v_h\|_{V_\sharp} \le ch|u|_{H^2(D)}$, so that

$$\|u - u_h\|_{V_\sharp} \le c\, h|u|_{H^2(D)}. \tag{27.10}$$

This shows that the error in the $\|\cdot\|_{V_\sharp}$-norm tends to zero at rate h.

27.3.2   Stabilized approximation of a first-order PDE

Consider the PDE $u' = f$ in $D := (0,1)$ with $u(0) = 0$ and $f \in L^2(D)$. Following §24.2.2, we consider the L^2-based weak formulation with the trial space $V := \{v \in H^1(D) \mid v(0) = 0\}$ and the test space $W := L^2(D)$. The exact forms are $a(v, w) := \int_0^1 v'w \, \mathrm{d}t$ and $\ell(w) := \int_0^1 fw \, \mathrm{d}t$. The model problem consists of seeking $u \in V$ such that $a(u, w) = \ell(w)$ for all $w \in W$. This problem is well-posed; see Exercise 25.9.

Consider the standard Galerkin approximation using as discrete trial and test spaces the space V_h built by using continuous $\mathbb{P}_1$ Lagrange finite elements on a uniform mesh $\mathcal{T}_h$ of step $h \in \mathcal{H}$ and by enforcing the boundary condition $v_h(0) = 0$. The discrete problem consists of seeking $u_h \in V_h$ such that $a(u_h, w_h) = \ell(w_h)$ for all $w_h \in V_h$. (The reader is invited to verify that the resulting linear system is identical to that obtained with centered finite differences.) The approximation setting is *conforming* since $V_h \subset V$ and $W_h = V_h \subset W$. Unfortunately, it turns out that the bilinear form a is not uniformly stable on $V_h \times V_h$. Indeed, one can show (see Exercise 27.3) that there are $0 < c_1 \le c_2$ s.t. for all $h \in \mathcal{H}$,

$$c_1 h \le \inf_{v_h \in V_h} \sup_{w_h \in V_h} \frac{|a(v_h, w_h)|}{\|v_h\|_{H^1(D)} \|w_h\|_{L^2(D)}} =: \alpha_h \le c_2 h. \tag{27.11}$$

This result shows that the above naive Galerkin approximation of first-order PDEs cannot produce optimal error estimates, even though it yields an invertible linear system ($c_1 \ne 0$). In practice, this problem manifests itself through the presence of spurious wiggles in the approximate solution. To circumvent this difficulty, let us define the discrete bilinear and linear forms

$$a_h(v_h, w_h) := \int_0^1 (v_h' w_h + h v_h' w_h') \, \mathrm{d}t, \qquad \ell_h(w_h) := \int_0^1 f(w_h + h w_h') \, \mathrm{d}t,$$

for all $v_h, w_h \in V_h$. Referring to Exercise 27.4 (see also §57.3 and §61.4), one can establish the uniform inf-sup condition

$$\inf_{v_h \in V_h} \sup_{w_h \in V_h} \frac{|a_h(v_h, w_h)|}{\|v_h\|_{V_h} \|w_h\|_{V_h}} \ge \alpha_0 > 0, \tag{27.12}$$

with the stability norm

$$\|v_h\|_{V_h}^2 := \ell_D^{-1} \|v_h\|_{L^2(D)}^2 + |v_h(1)|^2 + h\|v_h'\|_{L^2(D)}^2,$$

where we introduced the length scale $\ell_D := 1$ to be dimensionally consistent.

Let us perform the error analysis using Lemma 27.8. We set $V_s := V$ so that $V_\sharp = V + V_h = V$, and we equip $V_\sharp$ with the following norms (recall that $H^1(D) \hookrightarrow C^0(\overline{D})$ in one dimension):

$$\|v\|_{V_\flat} := \ell_D^{-1}\|v\|_{L^2(D)}^2 + |v(1)|^2 + h\|v'\|_{L^2(D)}^2, \tag{27.13}$$

$$\|v\|_{V_\sharp}^2 := h^{-1}\|v\|_{L^2(D)}^2 + |v(1)|^2 + h\|v'\|_{L^2(D)}^2, \tag{27.14}$$

so that (27.7) holds true with $c_\flat := 1$ since $h \le \ell_D$. Notice that there is no uniform constant $c_\sharp$ s.t. (27.5) holds true, i.e., we cannot apply Lemma 27.5. To apply Lemma 27.8, it remains to establish consistency/boundedness. Since $u' = f$ in D, letting $\eta := u - v_h$, we infer that

$$\langle \delta_h(v_h), w_h \rangle_{V_h', V_h} = \ell_h(w_h) - a_h(v_h, w_h)$$
$$= \int_0^1 f(w_h + hw_h')\,dt - \int_0^1 (v_h' w_h + hv_h' w_h')\,dt$$
$$= \int_0^1 (\eta' w_h + h\eta' w_h')\,dt =: \mathfrak{T}_1 + \mathfrak{T}_2.$$

Integrating by parts, we obtain

$$\mathfrak{T}_1 = \int_0^1 \eta' w_h\,dt = -\int_0^1 \eta w_h'\,dt + \eta(1)w_h(1),$$

since $\eta(0) = 0$. Using the Cauchy–Schwarz inequality, we infer that

$$|\mathfrak{T}_1| \le h^{-\frac{1}{2}}\|\eta\|_{L^2(D)} h^{\frac{1}{2}}\|w_h'\|_{L^2(D)} + |\eta(1)|\,|w_h(1)| \le \|\eta\|_{V_\sharp}\|w_h\|_{V_h},$$
$$|\mathfrak{T}_2| \le h^{\frac{1}{2}}\|\eta'\|_{L^2(D)} h^{\frac{1}{2}}\|w_h'\|_{L^2(D)} \le \|\eta\|_{V_\sharp}\|w_h\|_{V_h},$$

which shows that (27.4) holds true with $\omega_{\sharp h} := 2$. In conclusion, Lemma 27.8 implies that

$$\|u - u_h\|_{V_\flat} \le c \inf_{v_h \in V_h} \|u - v_h\|_{V_\sharp}. \tag{27.15}$$

Assuming that $u \in H^{1+r}(D)$, $r \in [0, 1]$, we use the approximation properties of finite elements to obtain $\inf_{v_h \in V_h} \|u - v_h\|_{V_\sharp} \le ch^{\frac{1}{2}+r}|u|_{H^{1+r}(D)}$, so that

$$\|u - u_h\|_{V_\flat} \le c\,h^{\frac{1}{2}+r}|u|_{H^{1+r}(D)}. \tag{27.16}$$

The error estimate (27.15) is quasi-optimal in the asymptotic range since the best-approximation errors in the $\|\cdot\|_{V_\flat}$- and $\|\cdot\|_{V_\sharp}$-norms converge to zero at the same rate (see Remark 27.9 for the terminology).

27.4   Strang's lemmas

We review in this section results due to Strang [358] and often called Strang's lemmas in the literature. These lemmas are historically important for the development of the analysis of finite element methods. In this book, we are going to use systematically Lemma 27.5 and Lemma 27.8 and only use Strang's lemmas at a few instances.

There are two Strang's lemmas: the first one is tailored to conforming approximations but allows for $a_h \neq a$ and $\ell_h \neq \ell$, and the second one can be applied to nonconforming approximations. Both lemmas can be seen as variants of Lemma 27.5 and Lemma 27.8, where the consistency error $\|\delta_h(v_h)\|_{W_h'}$ is further decomposed by adding/subtracting some terms so as to separate the approximation of a by a_h and the approximation of ℓ by ℓ_h (these contributions are sometimes called consistency error in the literature) from the best-approximation error of u by a function in V_h.

Remark 27.11 (Consistency). One should bear in mind that the notion of consistency in Strang's lemmas is somewhat arbitrary. This is illustrated in §27.4.3, where each lemma leads to a different notion of consistency for the same approximation method. We think that it is preferable to use the quantity $\|\delta_h(v_h)\|_{W_h'}$ defined in (27.3) as the only notion of consistency. This is the convention we are going to follow in the rest of the book. $\square$

27.4.1   Strang's first lemma

Strang's first lemma is tailored to conforming approximations. It has been devised to estimate the error due to quadratures when approximating elliptic PDEs by H^1-conforming finite elements (see §33.3).

Lemma 27.12 (Strang 1). *Assume:* (i) Conformity: $V_h \subset V$ and $W_h \subset W$, and set $V_s := V$ so that $V_\sharp := V + V_h = V$; (ii) Stability: (27.1) *holds true;* (iii) Boundedness: *the sesquilinear form a is bounded on $V \times W_h$, and set*

$$\|a\|_{\sharp h} := \sup_{v \in V} \sup_{w_h \in W_h} \frac{|a(v, w_h)|}{\|v\|_{V_\sharp} \|w_h\|_{W_h}}, \tag{27.17}$$

where the norm $\|\cdot\|_{V_\sharp}$ satisfies (27.5). Let $\delta_h^{\mathrm{St1}} : V_h \to W_h'$ be defined by

$$\langle \delta_h^{\mathrm{St1}}(v_h), w_h \rangle_{W_h', W_h} := \ell_h(w_h) - \ell(w_h) + a(v_h, w_h) - a_h(v_h, w_h). \tag{27.18}$$

Then the following holds true:

$$\|u - u_h\|_{V_\sharp} \leq \inf_{v_h \in V_h} \left[\left(1 + c_\sharp \frac{\|a\|_{\sharp h}}{\alpha_h} \right) \|u - v_h\|_{V_\sharp} + \frac{c_\sharp}{\alpha_h} \|\delta_h^{\mathrm{St1}}(v_h)\|_{W_h'} \right]. \tag{27.19}$$

Proof. Proceeding as in the proof of Lemma 27.5 leads to

$$\|u - u_h\|_{V_\sharp} \leq \|u - v_h\|_{V_\sharp} + \frac{c_\sharp}{\alpha_h} \|\delta_h(v_h)\|_{W_h'}.$$

We write the consistency error as follows:

$$\begin{aligned}
\langle \delta_h(v_h), w_h \rangle_{W_h', W_h} &:= \ell_h(w_h) - a_h(v_h, w_h) \\
&= \ell_h(w_h) - \ell(w_h) + a(u, w_h) - a_h(v_h, w_h) \\
&= \ell_h(w_h) - \ell(w_h) + a(u, w_h) - a_h(v_h, w_h) + [a(v_h, w_h) - a(v_h, w_h)] \\
&= \langle \delta_h^{\mathrm{St1}}(v_h), w_h \rangle_{W_h', W_h} + a(u - v_h, w_h),
\end{aligned}$$

where we used that $a(u, w_h) = \ell(w_h)$ since $W_h \subset W$. Using the triangle inequality and the boundedness property (27.17), we infer that

$$\|\delta_h(v_h)\|_{W_h'} \leq \|\delta_h^{\mathrm{St1}}(v_h)\|_{W_h'} + \|a\|_{\sharp h} \|u - v_h\|_{V_\sharp}.$$

Rearranging the terms leads to the expected estimate. $\square$

Remark 27.13 (Comparison). In the original statement of Strang's first lemma, one takes $\|\cdot\|_{V_\sharp} := \|\cdot\|_V$, and one equips V_h with the $\|\cdot\|_V$-norm, so that the error estimate (27.19) holds true with $\|a\|_{\sharp h} := \|a\|$. Moreover, the terms $\ell_h(w_h) - \ell(w_h)$ and $a(v_h, w_h) - a_h(v_h, w_h)$ composing $\langle \delta_h^{\mathrm{St1}}(v_h), w_h \rangle_{W_h', W_h}$ are separated, and the term $\|\ell_h - \ell\|_{W_h'}$ is taken out of the infimum over $v_h \in V_h$ in (27.19). The original statement is sufficient to analyze quadrature errors in the H^1-conforming approximation of elliptic PDEs, but as illustrated in §27.4.3, Strang's first lemma is not well adapted to analyze stabilized finite element approximations of first-order PDEs, since in this case one needs to invoke the two norms $\|\cdot\|_{V_\flat}$ and $\|\cdot\|_{V_\sharp}$ defined in (27.13)-(27.14). □

Remark 27.14 (Nonconforming setting). It is possible to derive an error estimate in the spirit of Strang's first lemma in some nonconforming settings. Following Gudi [226], the idea is to introduce an operator $T : W_h \to W$ acting on the discrete test functions. This operator can be built using the averaging operators analyzed in §22.2. We refer the reader to [226] and Exercise 27.5 for error estimates obtained with this technique. □

27.4.2 Strang's second lemma

Contrary to Strang's first lemma, the second lemma is applicable to nonconforming approximation settings.

Lemma 27.15 (Strang 2). *Let $V_\mathrm{s} := V$ so that $V_\sharp := V + V_h$. Assume:* (i) Stability: (27.1) *holds true;* (ii) Bounded extendibility*: There exists a bounded sesquilinear form $a_\sharp$ on $V_\sharp \times W_h$ that extends a_h originally defined on $V_h \times W_h$, i.e., $a_\sharp(v_h, w_h) = a_h(v_h, w_h)$ for all $(v_h, w_h) \in V_h \times W_h$ and*

$$\|a_\sharp\|_{\sharp h} := \sup_{v \in V_\sharp} \sup_{w_h \in W_h} \frac{|a_\sharp(v, w_h)|}{\|v\|_{V_\sharp} \|w_h\|_{W_h}} < \infty, \tag{27.20}$$

with a norm $\|\cdot\|_{V_\sharp}$ satisfying (27.5). The following holds true:

$$\|u - u_h\|_{V_\sharp} \leq \left(1 + c_\sharp \frac{\|a_\sharp\|_{\sharp h}}{\alpha_h}\right) \inf_{v_h \in V_h} \|u - v_h\|_{V_\sharp} + \frac{c_\sharp}{\alpha_h} \|\delta_h^{\mathrm{St2}}(u)\|_{W_h'}, \tag{27.21}$$

with $\langle \delta_h^{\mathrm{St2}}(u), w_h \rangle_{W_h', W_h} := \ell_h(w_h) - a_\sharp(u, w_h)$.

Proof. The starting point is again the bound

$$\|u - u_h\|_{V_\sharp} \leq \|u - v_h\|_{V_\sharp} + \frac{c_\sharp}{\alpha_h} \|\delta_h(v_h)\|_{W_h'}.$$

Now we write the consistency error as follows:

$$\begin{aligned}
\langle \delta_h(v_h), w_h \rangle_{W_h', W_h} &:= \ell_h(w_h) - a_h(v_h, w_h) = \ell_h(w_h) - a_\sharp(v_h, w_h) \\
&= \ell_h(w_h) - a_\sharp(v_h, w_h) + [a_\sharp(u, w_h) - a_\sharp(u, w_h)] \\
&= a_\sharp(u - v_h, w_h) + \langle \delta_h^{\mathrm{St2}}(u), w_h \rangle_{W_h', W_h}.
\end{aligned}$$

Using the triangle inequality and the boundedness property (27.20), we infer that

$$\|\delta_h(v_h)\|_{W_h'} \leq \|\delta_h^{\mathrm{St2}}(u)\|_{W_h'} + \|a_\sharp\|_{\sharp h} \|u - v_h\|_{V_\sharp}.$$

Rearranging the terms leads to the expected estimate.   □

Remark 27.16 (Strong consistency, quasi-optimality). Recalling the Galerkin orthogonality terminology introduced in the context of conforming approximations (see §26.3.1), we say that *strong consistency* holds true if $\delta_h^{\mathrm{St2}}(u)$ vanishes identically on W_h, i.e., if the exact solution satisfies the discrete equations rewritten using the extended sesquilinear form $a_\sharp$. In this case, (27.21) leads to a quasi-optimal error estimate.   □

Remark 27.17 (Bounded extendibility). Lemma 27.15 has been originally devised to analyze the Crouzeix–Raviart approximation of elliptic PDEs (see Chapter 36). In this context, the bounded extendibility assumption is indeed reasonable. However, it is no longer satisfied if a boundary penalty method or a discontinuous Galerkin method is used (see Chapters 37 and 38). For such methods, it is possible to recover the bounded extendibility assumption (and to prove strong consistency) provided the exact solution satisfies an additional smoothness assumption which is typically of the form $u \in H^{1+r}(D)$ with regularity pickup $r > \frac{1}{2}$. We will see that the error analysis based on Lemma 27.5 is more general since it only requires a regularity pickup $r > 0$ in the Sobolev scale. There are also other situations where the bounded extendibility assumption is simply not reasonable, e.g., when considering quadratures using point values or for stabilization techniques based on a two-scale hierarchical decomposition of the discrete spaces that is not meaningful for nondiscrete functions (see Chapter 59).   □

27.4.3 Example: first-order PDE

Let us consider the first-order PDE and the discrete setting introduced in §27.3.2, and let us briefly illustrate how to estimate the error using Strang's lemmas in this context. Using Strang's first lemma, one finds that

$$\langle \delta_h^{\mathrm{St1}}(v_h), w_h \rangle_{V_h', V_h} := \ell_h(w_h) - \ell(w_h) + a(v_h, w_h) - a_h(v_h, w_h)$$
$$= \int_0^1 h(f - v_h') w_h' \, \mathrm{d}t = \int_0^1 h\eta' w_h' \, \mathrm{d}t,$$

since $f = u'$ and $\eta := u - v_h$, so that $\|\delta_h^{\mathrm{St1}}(v_h)\|_{V_h'} \leq h^{\frac{1}{2}} \|\eta'\|_{L^2(D)} \leq \|\eta\|_{V_\sharp}$, where $\|\cdot\|_{V_\sharp}$ is defined in (27.14). One also has $\|a\|_{\sharp h} \leq \ell_D^{\frac{1}{2}} h^{-\frac{1}{2}}$. In conclusion, $\|u - u_h\|_{V_\sharp} \leq (1 + \alpha_0^{-1}(\ell_D^{\frac{1}{2}} h^{-\frac{1}{2}} + 1)) \inf_{v_h \in V_h} \|u - v_h\|_{V_\sharp}$, which yields the suboptimal error estimate $\|u - u_h\|_{V_\sharp} \leq ch^r \ell_D^{\frac{1}{2}} |u|_{H^{1+r}(D)}$ for all $r \in [0,1]$ (compare with (27.16)). Using instead Strang's second lemma, one finds that $\langle \delta_h^{\mathrm{St2}}(u), w_h \rangle_{V_h', V_h} := \ell_h(w_h) - a_\sharp(u, w_h) = 0$ for all $w_h \in V_h$, i.e., strong consistency holds true, and one obtains again the suboptimal error estimate $\|u - u_h\|_{V_\sharp} \leq ch^r \ell_D^{\frac{1}{2}} |u|_{H^{1+r}(D)}$. This example shows that the two Strang lemmas may lead to different notions of consistency, and, if applied blindly, they may yield suboptimal error estimates.

Exercises

Exercise 27.1 (Error identity). Assume stability, i.e., (27.1) holds true. Let $V_\sharp$ be defined in (27.2) and equip this space with a norm $\|\cdot\|_{V_\flat}$ s.t. there is $c_\flat$ s.t. $\|v_h\|_{V_\flat} \leq c_\flat \|v_h\|_{V_h}$ for all $v_h \in V_h$. Prove that

$$\|u - u_h\|_{V_\flat} = \inf_{v_h \in V_h} \left[\|u - v_h\|_{V_\flat} + \frac{c_\flat}{\alpha_h} \|\delta_h(v_h)\|_{W_h'} \right].$$

Exercise 27.2 (Boundary penalty). (i) Prove that $x^2 - 2\beta x y + \eta_0 y^2 \geq \frac{\eta_0 - \beta^2}{1+\eta_0}(x^2 + y^2)$ for all real numbers x, y, $\eta_0 \geq 0$ and $\beta \geq 0$. (ii) Using the notation of §27.3.1, prove that $a_h(v_h, v_h) \geq \frac{3}{8}\|v_h\|_{V_h}^2$ for all $v_h \in V_h$. (*Hint*: prove that $|v_h'(0)v_h(0)| \leq \|v_h'\|_{L^2(0,h)} h^{-\frac{1}{2}}|v_h(0)|$.)

Exercise 27.3 (First-order PDE). The goal is to prove (27.11). (i) Prove that

$$h^{-\frac{1}{2}}\|G(v_h)\|_{\ell^2(\mathbb{R}^I)} \leq \sup_{w_h \in V_h} \frac{|a(v_h, w_h)|}{\|w_h\|_{L^2(D)}} \leq \sqrt{6}h^{-\frac{1}{2}}\|G(v_h)\|_{\ell^2(\mathbb{R}^I)},$$

where $G_i(v_h) := a(v_h, \varphi_i)$ for all $i \in \{1:I\}$ with $I := \dim(V_h)$. (*Hint*: use Simpson's rule to compare Euclidean norms of component vectors and L^2-norms of functions.) (ii) Assume that I is even (the odd case is treated similarly). Prove that $\alpha_h \leq c_2 h$. (*Hint*: consider the oscillating function v_h s.t. $v_h(x_{2i}) := 2ih$ for all $i \in \{1:\frac{I}{2}\}$ and $v_h(x_{2i+1}) := 1$ for all $i \in \{0:\frac{I}{2}-1\}$.) (iii) Prove that $\alpha_h \geq c_1 h$. (*Hint*: prove that $\max_{i \in \{1:I\}} |v_h(x_i)| \leq 2\sum_{k \in \{1:I\}} |G_k(v_h)|$.) (iv) Prove that

$$\inf_{v_h \in V_h} \sup_{w_h \in W_h} \frac{|a(v_h, w_h)|}{\|v_h\|_{W^{1,1}(D)}\|w_h\|_{L^\infty(D)}} \geq \alpha_0 > 0$$

with $W_h := \{w_h \in L^\infty(D) \mid \forall i \in \{0:I-1\}, w_{h|[x_i, x_{i+1}]} \in \mathbb{P}_0\}$. (*Hint*: see Proposition 25.19.)

Exercise 27.4 (GaLS 1D). The goal is to prove (27.12). Let $v_h \in V_h$. (i) Compute $a_h(v_h, v_h)$. (ii) Let $\zeta(x) := -2x/\ell_D$, set $\zeta_h := \mathcal{I}_h^{\mathrm{b}}(\zeta)$, and show that $a_h(v_h, \mathcal{J}_h^{\mathrm{av}}(\zeta_h v_h)) \geq \frac{1}{2}\ell_D^{-1}\|v_h\|_{L^2(D)}^2 - c_1 a(v_h, v_h)$ uniformly w.r.t. $h \in \mathcal{H}$, $\mathcal{J}_h^{\mathrm{av}}$ is the averaging operator defined in (22.9), and $\mathcal{I}_h^{\mathrm{b}}$ is the L^2-projection on the functions that are piecewise constant over the mesh. (iii) Prove (27.12). (*Hint*: use the test function $z_h := 2\mathcal{J}_h^{\mathrm{av}}(\zeta_h v_h) + 2(c_1 + 1)v_h$.)

Exercise 27.5 (Nonconforming Strang 1). Let $T : W_h \to W \cap W_h$. Let $V_{\mathrm{s}} := V$ so that $V_\sharp := V + V_h$, and assume that $V_\sharp$ is equipped with a norm $\|\cdot\|_{V_\sharp}$ satisfying (27.5). (i) Assume that a_h can be extended to $V_h \times (W + W_h)$. Assume that there is $\|a\|_{\sharp h}$ s.t. consistency/boundedness holds true in the form $|a(u, T(w_h)) - a_h(v_h, T(w_h))| \leq \|a\|_{\sharp h}\|u - v_h\|_{V_\sharp}\|w_h\|_{W_h}$. Prove that

$$\|u - u_h\|_{V_\sharp} \leq \inf_{v_h \in V_h}\left[\left(1 + c_\sharp \frac{\|a\|_{\sharp h}}{\alpha_h}\right)\|u - v_h\|_{V_\sharp} + \frac{c_\sharp}{\alpha_h}\|\hat{\delta}_h^{\mathrm{St1}}(v_h)\|_{W_h'}\right],$$

with $\|\hat{\delta}_h^{\mathrm{St1}}(v_h)\|_{W_h'} := \|\ell_h - \ell \circ T + a_h(v_h, T(\cdot)) - a_h(v_h, \cdot)\|_{W_h'}$. (*Hint*: add/subtract $a_h(v_h, T(w_h))$.) (ii) We now derive another error estimate that avoids extending a_h but restricts the discrete trial functions to $V_h \cap V$ (this is reasonable provided the subspace $V_h \cap V$ has approximation properties that are similar to those of V_h). Assuming that there is $\|a\|_{V \times W_h}$ s.t. boundedness holds true in the form $|a(u - v_h, T(w_h))| \leq \|a\|_{V \times W_h}\|u - v_h\|_{V_\sharp}\|w_h\|_{W_h}$, prove that

$$\|u - u_h\|_{V_\sharp} \leq \inf_{v_h \in V_h \cap V}\left[\left(1 + c_\sharp \frac{\|a\|_{V \times W_h}}{\alpha_h}\right)\|u - v_h\|_{V_\sharp} + \frac{c_\sharp}{\alpha_h}\|\check{\delta}_h^{\mathrm{St1}}(v_h)\|_{W_h'}\right],$$

with $\|\check{\delta}_h^{\mathrm{St1}}(v_h)\|_{W_h'} := \|\ell_h - \ell \circ T + a(v_h, T(\cdot)) - a_h(v_h, \cdot)\|_{W_h'}$. (*Hint*: add/subtract $a(v_h, T(w_h))$.)

Exercise 27.6 (Orthogonal projection). Consider the setting of Exercise 25.4 with real vector spaces and coercivity with $\xi := 1$ for simplicity. Let u be the unique element in V such that $a(u, v - u) \geq \ell(v - u)$ for all $v \in U$. Let V_h be a finite-dimensional subspace of V, and let U_h be a nonempty, closed, and convex subset of V_h. We know from Exercise 25.4 that there is a unique u_h in V_h such that $a(u_h, v_h - u_h) \geq \ell(v_h - u_h)$ for all $v_h \in U_h$. (i) Show that there is $c_1(u)$ such that for all $(v, v_h) \in U \times V_h$,

$$\|u - u_h\|_V^2 \leq c_1(u)\big(\|u - v_h\|_V + \|u_h - v\|_V + \|u - u_h\|_V\|u - v_h\|_V\big).$$

(*Hint*: prove $\alpha \|u - u_h\|_V^2 \leq a(u, v - u_h) + \ell(u_h - v) + a(u_h, v_h - u) + \ell(u - v_h)$.) (ii) Show that there is $c_2(u)$ such that

$$\|u - u_h\|_V \leq c_2(u) \left(\inf_{v_h \in U_h} \left(\|u - v_h\|_V + \|u - v_h\|_V^2 \right) + \inf_{v \in U} \|u_h - v\|_V \right)^{\frac{1}{2}}.$$

Chapter 28

Linear algebra

In this chapter, we first show that the discrete problem generated by the Galerkin approximation can be reformulated as a linear system once bases for the discrete trial space and the discrete test space are chosen. Then, we investigate important properties of the system matrix, which is called *stiffness matrix*, and we also introduce the *mass matrix*, which is relevant when computing L^2-orthogonal projections. We derive various estimates on the norm, the spectrum, and the condition number of both matrices. Finally, we give a brief overview of direct and iterative solution methods for linear systems.

28.1 Stiffness and mass matrices

Recall that the discrete problem (26.3) consists of seeking $u_h \in V_h$ s.t. $a_h(u_h, w_h) = \ell_h(w_h)$ for all $w_h \in W_h$, where a_h is sesquilinear (bilinear in the real case) and ℓ_h is antilinear (linear in the real case). We assume that the discrete problem is well-posed, i.e., the inf-sup condition (26.5a) holds true and $\dim(V_h) = \dim(W_h) =: I$. We show in this section that the discrete problem can be reformulated as a linear system once bases for V_h and W_h are chosen.

28.1.1 Main definitions

Let $\{\varphi_i\}_{i \in \{1:I\}}$ be a basis of V_h and $\{\psi_i\}_{i \in \{1:I\}}$ be a basis of W_h. Let $\mathsf{R}_\varphi : \mathbb{C}^I \to V_h$ be the isomorphism that reconstructs functions in V_h from coordinate vectors, i.e., $\mathsf{R}_\varphi(\mathsf{V}) := \sum_{i \in \{1:I\}} \mathsf{V}_i \varphi_i$ for all $\mathsf{V} := (\mathsf{V}_i)_{i \in \{1:I\}} \in \mathbb{C}^I$. A similar isomorphism R_ψ is considered for the discrete space W_h equipped with the basis $\{\psi_i\}_{i \in \{1:I\}}$. These isomorphisms are instrumental to go back and forth from the functional viewpoint to the algebraic viewpoint.

Let $\mathcal{A} \in \mathbb{C}^{I \times I}$ be the *stiffness matrix* with entries

$$\mathcal{A}_{ij} := a_h(\varphi_j, \psi_i), \qquad \forall i, j \in \{1:I\}, \tag{28.1}$$

(note the position of the indices i and j in (28.1)) and let $\mathsf{B} \in \mathbb{C}^I$ be the column vector with components

$$\mathsf{B}_i := \ell_h(\psi_i), \qquad \forall i \in \{1:I\}. \tag{28.2}$$

The link between the discrete sesquilinear form a_h and the stiffness matrix $\mathcal{A}$ can be formalized as follows:

$$\mathsf{W}^\mathsf{H} \mathcal{A} \mathsf{V} = a_h(\mathsf{R}_\varphi(\mathsf{V}), \mathsf{R}_\psi(\mathsf{W})), \qquad \forall \mathsf{V}, \mathsf{W} \in \mathbb{C}^I. \tag{28.3}$$

Similarly, we have $\mathsf{W}^{\mathsf{H}}B = \ell_h(\mathsf{R}_\psi(\mathsf{W}))$. The above definitions imply that

$$[\, u_h \text{ solves } (26.3)) \,] \iff [\, \mathcal{A}\mathsf{U} = \mathsf{B} \,] \quad \text{with } u_h := \mathsf{R}_\varphi(\mathsf{U}). \tag{28.4}$$

We observe that the number of equations and the number of unknowns in the linear system $\mathcal{A}\mathsf{U} = \mathsf{B}$ is equal to $\dim(W_h)$ and $\dim(V_h)$, respectively. Thus, the linear system is *square* if and only if $\dim(V_h) = \dim(W_h)$.

Remark 28.1 (Well-posedness). One easily verifies the following: (i) The inf-sup condition (26.5a) holds true if and only if $\ker(\mathcal{A}) = \{0\}$. (ii) The condition (26.8) is equivalent to $\mathrm{rank}(\mathcal{A}) = I$. (iii) The sesquilinear form a_h is coercive on V_h if and only if the matrix $\mathcal{A}$ is Hermitian positive definite, i.e., $\Re(\xi\mathsf{V}^{\mathsf{H}}\mathcal{A}\mathsf{V}) \geq 0$ for all $\mathsf{V} \in \mathbb{C}^I$ and $\mathsf{V}^{\mathsf{H}}\mathcal{A}\mathsf{V} = 0$ implies that $\mathsf{V} = 0$. □

In many applications, V_h and W_h are discrete subspaces of $L^2(D)$. It is then meaningful to define the following *mass matrices*:

$$\mathcal{M}_{\varphi,ij} := (\varphi_j, \varphi_i)_{L^2(D)}, \quad \mathcal{M}_{\psi,ij} := (\psi_j, \psi_i)_{L^2(D)}, \quad \forall i,j \in \{1{:}I\}. \tag{28.5}$$

Notice that both matrices are Hermitian positive definite. When the basis functions are real-valued, these matrices are symmetric positive definite. The mass matrices are useful to evaluate L^2-orthogonal projections. For instance, consider the L^2-orthogonal projection $\Pi_{V_h} : L^2(D) \to V_h$ such that for all $v \in L^2(D)$, $\Pi_{V_h}(v)$ is the unique function in V_h satisfying $(\Pi_{V_h}(v)-v, y_h)_{L^2(D)} = 0$ for all $y_h \in V_h$. One easily verifies that $\Pi_{V_h}(v) = \mathsf{R}_\varphi(\mathsf{X})$ where $\mathsf{X} \in \mathbb{C}^I$ solves the linear system $\mathcal{M}_\varphi\mathsf{X} = \mathsf{Y}$ with right-hand side vector $\mathsf{Y} := ((v, \varphi_i)_{L^2(D)})_{i \in \{1{:}\,I\}} \in \mathbb{C}^I$.

Example 28.2 ($\mathbb{P}_1$ Lagrange, 1D). Consider the bilinear form $a(v,w) := \int_D v'w'\,\mathrm{d}x$ for all $v, w \in H_0^1(D)$ with $D := (0,1)$. Consider a uniform mesh $\mathcal{T}_h$ of D specified by its vertices $x_i := ih$ for all $i \in \{0{:}(I+1)\}$ with $h := \frac{1}{I+1}$. Let V_h be spanned by piecewise affine functions on $\mathcal{T}_h$ vanishing at the two endpoints of D. The global shape functions in V_h are the hat basis functions s.t. $\varphi_i(x) := 1 - h^{-1}|x - x_i|$ for $x \in [x_{i-1}, x_{i+1}]$ and $\varphi_i(x) := 0$ otherwise, for all $i \in \{1{:}I\}$. With $W_h := V_h$ and $a_h := a_{|V_h \times V_h}$, the stiffness matrix $\mathcal{A} \in \mathbb{R}^{I \times I}$ is tridiagonal. The diagonal entries are equal to $2h^{-1}$ and the upper- and lower-diagonal entries are equal to $-h^{-1}$. We write $\mathcal{A} = h^{-1}\,\mathrm{tridiag}(-1, 2, -1)$. The mass matrix $\mathcal{M} \in \mathbb{R}^{I \times I}$ is also tridiagonal with $\mathcal{M} = \frac{h}{6}\,\mathrm{tridiag}(1, 4, 1)$. □

28.1.2   Static condensation

The idea behind static condensation is that one can eliminate from the linear system in (28.4) all the unknowns corresponding to the global basis functions that are supported in one mesh cell only. This elimination is a cost-effective approach to reduce the size of the linear system since it can be realized by performing only local computations in each mesh cell.

For simplicity, we present the technique in the case where $V_h = W_h$. Recall that for every cell $K \in \mathcal{T}_h$, the degrees of freedom (dofs) of the finite element (K, P_K, Σ_K) are enumerated by using the set $\mathcal{N}$. We consider the partition $\mathcal{N} = \mathcal{N}^\circ \cup \mathcal{N}^\partial$ where $\mathcal{N}^\partial := \bigcup_{F \in \mathcal{F}_K} \mathcal{N}_{K,F}$ (recall that $n \in \mathcal{N}_{K,F}$ iff the local shape function $\theta_{K,n}$ has a nonzero γ-trace on the face $F \in \mathcal{F}_K$; see §20.1). Thus, $n \in \mathcal{N}^\circ$ iff the γ-trace of the local shape function $\theta_{K,n}$ vanishes on the boundary of K. Note that both sets $\mathcal{N}^\circ$ and $\mathcal{N}^\partial$ only depend on the reference finite element. Let us now partition the global set of dofs $\mathcal{I}_h := \{1{:}I\}$ in the form $\mathcal{I}_h = \mathcal{I}_{\mathcal{T}_h} \cup \mathcal{I}_{\mathcal{F}_h}$ where

$$\mathcal{I}_{\mathcal{T}_h} := \{i \in \mathcal{I}_h \mid i = \mathtt{j\_dof}(K,n),\ K \in \mathcal{T}_h,\ n \in \mathcal{N}^\circ\}, \tag{28.6a}$$

$$\mathcal{I}_{\mathcal{F}_h} := \{i \in \mathcal{I}_h \mid i = \mathtt{j\_dof}(K,n),\ K \in \mathcal{T}_h,\ n \in \mathcal{N}^\partial\}, \tag{28.6b}$$

where j_dof is the connectivity array introduced in §19.1 (the sets $\mathcal{I}_{\mathcal{T}_h}$ and $\mathcal{I}_{\mathcal{F}_h}$ are disjoint owing to the injectivity of the map j_dof$(K, \cdot)$; see (19.3)). We first enumerate the global dofs in $\mathcal{I}_{\mathcal{T}_h}$ (the associated global shape functions are often called bubble functions), and then we enumerate those in $\mathcal{I}_{\mathcal{F}_h}$. This leads to the following block-decomposition of the linear system (with obvious notation):

$$\begin{bmatrix} \mathcal{A}_{\mathcal{T}_h \mathcal{T}_h} & \mathcal{A}_{\mathcal{T}_h \mathcal{F}_h} \\ \mathcal{A}_{\mathcal{F}_h \mathcal{T}_h} & \mathcal{A}_{\mathcal{F}_h \mathcal{F}_h} \end{bmatrix} \begin{bmatrix} \mathsf{U}_{\mathcal{T}_h} \\ \mathsf{U}_{\mathcal{F}_h} \end{bmatrix} = \begin{bmatrix} \mathsf{B}_{\mathcal{T}_h} \\ \mathsf{B}_{\mathcal{F}_h} \end{bmatrix}. \tag{28.7}$$

The key observation is that $\mathcal{A}_{\mathcal{T}_h \mathcal{T}_h}$ is block-diagonal and each block has size card$(\mathcal{N}^\circ)$, If the method is conforming, the entries of each block are $a(\theta_{K,n'}, \theta_{K,n})$ for all $n, n' \in \mathcal{N}^\circ$, and it can be shown that each of these small matrices is invertible. Hence, $\mathcal{A}_{\mathcal{T}_h \mathcal{T}_h}$ is easy to invert, and this leads to

$$\mathsf{U}_{\mathcal{T}_h} = -(\mathcal{A}_{\mathcal{T}_h \mathcal{T}_h})^{-1}\mathcal{A}_{\mathcal{T}_h \mathcal{F}_h}\mathsf{U}_{\mathcal{F}_h} + (\mathcal{A}_{\mathcal{T}_h \mathcal{T}_h})^{-1}\mathsf{B}_{\mathcal{T}_h}. \tag{28.8}$$

Substituting this expression into the second equation of (28.7), we infer that

$$\left(\mathcal{A}_{\mathcal{F}_h \mathcal{F}_h} - \mathcal{A}_{\mathcal{F}_h \mathcal{T}_h}(\mathcal{A}_{\mathcal{T}_h \mathcal{T}_h})^{-1}\mathcal{A}_{\mathcal{T}_h \mathcal{F}_h}\right)\mathsf{U}_{\mathcal{F}_h} = \mathsf{B}'_{\mathcal{F}_h}, \tag{28.9}$$

where $\mathsf{B}'_{\mathcal{F}_h} := \mathsf{B}_{\mathcal{F}_h} - \mathcal{A}_{\mathcal{F}_h \mathcal{T}_h}(\mathcal{A}_{\mathcal{T}_h \mathcal{T}_h})^{-1}\mathsf{B}_{\mathcal{T}_h}$. The matrix on the left-hand side of (28.9) is called *Schur complement* of $\mathcal{A}_{\mathcal{T}_h \mathcal{T}_h}$. One proceeds as follows to solve the linear system (28.7): one first computes $\mathsf{U}_{\mathcal{F}_h}$ by solving (28.9), then one computes $\mathsf{U}_{\mathcal{T}_h}$ by solving (28.8). This technique is called *static condensation*; see Guyan [232], Irons [252]. Static condensation makes sense only if $\mathcal{N}^\circ$ is nonempty, and it is computationally effective if card$(\mathcal{N}^\circ) \sim$ card$(\mathcal{N})$. Referring to §29.1 for further insight, we note that static condensation reduces the size of the stiffness matrix without altering its sparsity pattern.

Example 28.3 (Lagrange elements). For Lagrange finite elements of degree $k \geq 1$, the internal dofs are evaluations at the nodes located inside K. Then if $k \geq d+1$, the set $\mathcal{N}^\circ$ is nonempty and card$(\mathcal{N}^\circ) = \binom{k-1}{d}$. Static condensation is effective when k is large. □

28.2   Bounds on the stiffness and mass matrices

In this section, we introduce the notion of condition number for a nonsingular matrix, and we derive various bounds on the spectrum and the norm of the stiffness and mass matrices.

28.2.1   Condition number

We denote by $\|\cdot\|_{\ell^2(\mathbb{C}^I)}$ the Euclidean norm in $\mathbb{C}^I$. The induced matrix norm is denoted similarly. Recall that for every square matrix $\mathcal{Z} \in \mathbb{C}^{I \times I}$, we have $\|\mathcal{Z}\|_{\ell^2(\mathbb{C}^I)} := \rho(\mathcal{Z}^\mathsf{H}\mathcal{Z})^{\frac{1}{2}} = \rho(\mathcal{Z}\mathcal{Z}^\mathsf{H})^{\frac{1}{2}}$, where $\rho(\cdot)$ denotes the spectral radius and $\mathcal{Z}^\mathsf{H}$ the Hermitian transpose of $\mathcal{Z}$, i.e., $(\mathcal{Z}^\mathsf{H})_{ij} := \overline{\mathcal{Z}_{ji}}$. The Euclidean *condition number* of any nonsingular matrix $\mathcal{Z} \in \mathbb{C}^{I \times I}$ is defined by

$$\kappa_{\ell^2}(\mathcal{Z}) := \|\mathcal{Z}\|_{\ell^2(\mathbb{C}^I)}\|\mathcal{Z}^{-1}\|_{\ell^2(\mathbb{C}^I)}. \tag{28.10}$$

(Condition numbers can be defined for every matrix norm induced by a vector norm.) Observe that $\kappa_{\ell^2}(\mathcal{Z}) \geq \|\mathcal{Z}\mathcal{Z}^{-1}\|_{\ell^2(\mathbb{C}^I)} = 1$. The Euclidean condition number of $\mathcal{Z}$ is the ratio of the maximal and the minimal singular values of $\mathcal{Z}$. In particular, if $\mathcal{Z}$ is Hermitian (or symmetric in the real case), i.e., $\mathcal{Z}^\mathsf{H} = \mathcal{Z}$ (or $\mathcal{Z}^\mathsf{T} = \mathcal{Z}$), its eigenvalues are real and $\kappa_{\ell^2}(\mathcal{Z})$ is the ratio of the maximal and the minimal eigenvalues of $\mathcal{Z}$ (in absolute value). In other words, letting $\sigma(\mathcal{Z}) \subset \mathbb{R}$

denote the spectrum of $\mathcal{Z}$ and setting $\lambda_{\min} := \min_{\lambda \in \sigma(\mathcal{Z})} |\lambda|$ and $\lambda_{\max} := \max_{\lambda \in \sigma(\mathcal{Z})} |\lambda|$, we have $\kappa_{\ell^2}(\mathcal{Z}) := \frac{\lambda_{\max}}{\lambda_{\min}}$.

Definition 28.4 (Ill-conditioning). *A matrix $\mathcal{Z} \in \mathbb{C}^{I \times I}$ is said to be* ill-conditioned *whenever $\kappa_{\ell^2}(\mathcal{Z}) \gg 1$.*

A large condition number often indicates that numerical difficulties are to be expected when solving a linear system. We refer the reader to Exercise 28.6 for further insight into the influence of the condition number on the sensitivity to perturbations and to Proposition 28.21 for further insight into the convergence rate of iterative methods.

28.2.2   Spectrum of the mass matrix

In this section, we investigate the spectrum of the mass matrix $\mathcal{M}_\varphi$. The results are the same for $\mathcal{M}_\psi$. Since the mass matrix is Hermitian positive definite, its spectrum lies on the positive real half-line, i.e., $\sigma(\mathcal{M}_\varphi) \subset [0, \infty)$. Let $\mu_{\min}^\varphi := \min_{\mu \in \sigma(\mathcal{M}_\varphi)} \mu$ and $\mu_{\max}^\varphi := \max_{\mu \in \sigma(\mathcal{M}_\varphi)} \mu$ be the smallest and the largest eigenvalue of $\mathcal{M}_\varphi$, respectively. Since $\mathsf{V}^\mathsf{H} \mathcal{M}_\varphi \mathsf{V} = \|\mathsf{R}_\varphi(\mathsf{V})\|_{L^2(D)}^2$, we infer that

$$\mu_{\min}^\varphi = \min_{\mathsf{V} \in \mathbb{C}^I} \frac{\|\mathsf{R}_\varphi(\mathsf{V})\|_{L^2(D)}^2}{\|\mathsf{V}\|_{\ell^2(\mathbb{C}^I)}^2} \leq \max_{\mathsf{V} \in \mathbb{C}^I} \frac{\|\mathsf{R}_\varphi(\mathsf{V})\|_{L^2(D)}^2}{\|\mathsf{V}\|_{\ell^2(\mathbb{C}^I)}^2} = \mu_{\max}^\varphi. \tag{28.11}$$

Moreover, we have

$$\mu_{\max}^\varphi = \|\mathcal{M}_\varphi\|_{\ell^2(\mathbb{C}^I)}, \quad \mu_{\min}^\varphi = \|\mathcal{M}_\varphi^{-1}\|_{\ell^2(\mathbb{C}^I)}^{-1}, \quad \kappa_{\ell^2}(\mathcal{M}_\varphi) = \frac{\mu_{\max}^\varphi}{\mu_{\min}^\varphi}. \tag{28.12}$$

We assume that the basis functions $\{\varphi_i\}_{i \in \{1:I\}}$ spanning V_h are finite element global shape functions built using a sequence of affine meshes $(\mathcal{T}_h)_{h \in \mathcal{H}}$ and a reference finite element with shape functions $\{\widehat{\theta}_n\}_{n \in \mathcal{I}}$. In every mesh cell $K \in \mathcal{T}_h$, the local shape functions are defined by $\theta_{K,n} := \psi_K^{-1}(\widehat{\theta}_n)$ for all $n \in \mathcal{N}$, and we assume that the functional transformation is such that $\psi_K(v) := \mathbb{A}_K(v \circ \boldsymbol{T}_K)$, where $\boldsymbol{T}_K$ is the geometric mapping and $\mathbb{A}_K \in \mathbb{R}^{q \times q}$ for some integer $q \geq 1$. The global basis functions φ_i are such that

$$\varphi_{\mathtt{j\_dof}(K,n)|K} = \theta_{K,n}, \qquad \forall (K, n) \in \mathcal{T}_h \times \mathcal{N}, \tag{28.13}$$

where $\mathtt{j\_dof}$ is the connectivity array introduced in Chapter 19.

Proposition 28.5 (Local spectrum). *Assume that the mesh $\mathcal{T}_h$ is affine (the regularity of the mesh sequence is not needed). Let $K \in \mathcal{T}_h$ and let $\mathcal{M}^K \in \mathbb{R}^{n_{\mathrm{sh}} \times n_{\mathrm{sh}}}$ be the local mass matrix with entries $\mathcal{M}_{nn'}^K = (\theta_{K,n'}, \theta_{K,n})_{L^2(K)}$ for all $n, n' \in \mathcal{N}$. Let $\mu_{\min}^K$ and $\mu_{\max}^K$ be the smallest and the largest eigenvalue of $\mathcal{M}^K$. Then there are $0 < c_\flat \leq c_\sharp$ s.t. for all $K \in \mathcal{T}_h$ and all $h \in \mathcal{H}$,*

$$c_\flat \|\mathbb{A}_K\|_{\ell^2}^{-2} |K| \leq \mu_{\min}^K \leq \mu_{\max}^K \leq c_\sharp \|\mathbb{A}_K^{-1}\|_{\ell^2}^2 |K|. \tag{28.14}$$

Proof. Norm equivalence in $\mathbb{C}^{n_{\mathrm{sh}}}$ implies that there are $0 < \widehat{c}_\flat \leq \widehat{c}_\sharp$ s.t.

$$\widehat{c}_\flat \|\mathsf{V}\|_{\ell^2(\mathbb{C}^{n_{\mathrm{sh}}})} \leq \|\mathsf{R}_{\widehat{\theta}}(\mathsf{V})\|_{L^2(\widehat{K})} \leq \widehat{c}_\sharp \|\mathsf{V}\|_{\ell^2(\mathbb{C}^{n_{\mathrm{sh}}})}, \qquad \forall \mathsf{V} \in \mathbb{C}^{n_{\mathrm{sh}}},$$

with the reconstructed function $\mathsf{R}_{\widehat{\theta}}(\mathsf{V}) := \sum_{n \in \mathcal{N}} \mathsf{V}_n \widehat{\theta}_n$ in $\widehat{K}$. Let $K \in \mathcal{T}_h$ be a mesh cell and consider the reconstructed function $\mathsf{R}_{\theta_K}(\mathsf{V}) := \sum_{n \in \mathcal{N}} \mathsf{V}_n \theta_{K,n}$ in K. Owing to the linearity of the map ψ_K, we have $\psi_K(\mathsf{R}_{\theta_K}(\mathsf{V})) = \mathsf{R}_{\widehat{\theta}}(\mathsf{V})$. Lemma 11.7 and the above norm equivalence imply that

$$c_\flat \|\mathbb{A}_K\|_{\ell^2}^{-1} |K|^{\frac{1}{2}} \leq \frac{\|\mathsf{R}_{\theta_K}(\mathsf{V})\|_{L^2(K)}}{\|\mathsf{V}\|_{\ell^2(\mathbb{C}^{n_{\mathrm{sh}}})}} \leq c_\sharp \|\mathbb{A}_K^{-1}\|_{\ell^2} |K|^{\frac{1}{2}},$$

with $0 < c_\flat \leq c_\sharp$ uniform w.r.t. $K \in \mathcal{T}_h$ and $h \in \mathcal{H}$. We conclude the proof by invoking (28.11). $\quad\square$

Proposition 28.6 (Global spectrum). *Assume that the mesh sequence is shape-regular and that $\|\mathbb{A}_K\|_{\ell^2}$ is uniformly equivalent to h_K^r for some $r \geq 0$. Then there are $0 < c_1 \leq c_2$ s.t. for all $K \in \mathcal{T}_h$ and all $h \in \mathcal{H}$,*

$$c_1 \min_{K \in \mathcal{T}_h} h_K^{d-2r} \leq \mu_{\min}^{\varphi} \leq \mu_{\max}^{\varphi} \leq c_2 \max_{K \in \mathcal{T}_h} h_K^{d-2r}. \tag{28.15}$$

Moreover, if the mesh sequence is quasi-uniform (see Definition 22.20), there are $0 < c_1' \leq c_2'$, uniform w.r.t. $h \in \mathcal{H}$, such that

$$c_1' h^{d-2r} \leq \mu_{\min}^{\varphi} \leq \mu_{\max}^{\varphi} \leq c_2' h^{d-2r}, \tag{28.16}$$

implying the bound $\kappa_{\ell^2}(\mathcal{M}_\varphi) \leq \frac{c_2'}{c_1'}$.

Proof. For all $\mathsf{V} \in \mathbb{C}^I$ and all $K \in \mathcal{T}_h$, let $\mathsf{V}^K \in \mathbb{C}^{n_{\mathrm{sh}}}$ be the components of V associated with the local dofs in K, i.e., $\mathsf{V}_n^K := \mathsf{V}_{\mathtt{j\_dof}(K,n)}$ for all $n \in \mathcal{N}$. The regularity of the mesh sequence implies that there is c s.t. $\mathrm{card}(\{(K,n) \in \mathcal{T}_h \times \mathcal{N} \mid i = \mathtt{j\_dof}(K,n)\}) \leq c$ for all $i \in \{1{:}I\}$ and all $h \in \mathcal{H}$. We infer that $\|\mathsf{V}\|_{\ell^2(\mathbb{C}^I)}^2 \leq \sum_{K \in \mathcal{T}_h} \|\mathsf{V}^K\|_{\ell^2(\mathbb{C}^{n_{\mathrm{sh}}})}^2 \leq c\|\mathsf{V}\|_{\ell^2(\mathbb{C}^I)}^2$. Since $\mathsf{R}_\varphi(\mathsf{V})_{|K} = \mathsf{R}_{\theta_K}(\mathsf{V}^K)$ owing to (28.13), we infer that

$$\|\mathsf{R}_\varphi(\mathsf{V})\|_{L^2(D)}^2 = \sum_{K \in \mathcal{T}_h} \|\mathsf{R}_{\theta_K}(\mathsf{V}^K)\|_{L^2(K)}^2 \geq c_\flat \sum_{K \in \mathcal{T}_h} \|\mathbb{A}_K\|_{\ell^2}^{-2} |K| \|\mathsf{V}^K\|_{\ell^2(\mathbb{C}^{n_{\mathrm{sh}}})}^2$$

$$\geq c_\flat \Big(\min_{K \in \mathcal{T}_h} \|\mathbb{A}_K\|_{\ell^2}^{-2} |K| \Big) \sum_{K \in \mathcal{T}_h} \|\mathsf{V}^K\|_{\ell^2(\mathbb{C}^{n_{\mathrm{sh}}})}^2 \geq c_\flat \Big(\min_{K \in \mathcal{T}_h} h_K^{d-2r} \Big) \|\mathsf{V}\|_{\ell^2(\mathbb{C}^I)}^2,$$

by our assumption on $\|\mathbb{A}_K\|_{\ell^2}$ and since $|K|$ is uniformly equivalent to h_K^d for shape-regular mesh sequences. This yields the lower bound in (28.15). The upper bound is proved similarly using that $\|\mathbb{A}_K\|_{\ell^2}\|\mathbb{A}_K^{-1}\|_{\ell^2}$ is uniformly bounded. The upper and lower bounds in (28.16) follow from the quasi-uniformity assumption on the mesh sequence. $\qquad\square$

Example 28.7 ($\mathbb{P}_1$ Lagrange, 1D). Recall from Example 28.2 the mass matrix

$$\mathcal{M} = \tfrac{h}{6} \operatorname{tridiag}(1, 4, 1).$$

Letting $\eta := \frac{1}{I+1}$, one can show that the eigenvalues of a $I \times I$ tridiagonal matrix $\operatorname{tridiag}(b, a, b)$, $a, b \in \mathbb{R}$, are $\lambda_l := a + 2b\cos(\pi l \eta)$ with associated eigenvectors $\mathsf{V}_l := (\sin(\pi l m \eta))_{m \in \{1{:}I\}}$, for all $l \in \{1{:}I\}$. Hence, the eigenvalues of the mass matrix are $\mu_l = \frac{1}{3}h\,(2 + \cos(\pi l \eta))$, for all $l \in \{1{:}I\}$. This implies that $\mu_{\min} = \mu_I = \frac{1}{3}h\,(2 - \cos(\pi\eta)) \approx \frac{1}{3}h$ if I is large, and $\mu_{\max} = \mu_1 = \frac{1}{3}h\,(2 + \cos(\pi\eta)) \approx h$ if I is large. $\qquad\square$

Remark 28.8 (Exponent r). For Lagrange and canonical hybrid elements, one has $r = 0$ in (28.15) since ψ_K is just the pullback by $\boldsymbol{T}_K$. For Nédélec and Raviart–Thomas elements, one has $r = 1$ and $r = 2$, respectively. $\qquad\square$

Remark 28.9 (Broken spaces). If V_h is a broken finite element space, the support of the basis functions φ_i is localized to a single mesh cell. This implies that the mass matrix $\mathcal{M}_\varphi$ is *block-diagonal*, each block being of size n_{sh}. Thus, $\mathcal{M}_\varphi$ is easy to invert. Although this special structure is lost if V_h is a conforming finite element space, the mass matrix $\mathcal{M}_\varphi$ remains in general easy to invert. In particular, Proposition 28.6 shows that $\mathcal{M}_\varphi$ is well-conditioned (at least on quasi-uniform mesh sequences). $\qquad\square$

Remark 28.10 (Literature). Other bounds on the eigenvalues of the mass matrix are derived in Wathen [389]. $\qquad\square$

28.2.3   Bounds on the stiffness matrix

Let us introduce the following real numbers:

$$\alpha_{\ell^2} := \inf_{\mathsf{V} \in \mathbb{C}^I} \frac{\|\mathcal{A}\mathsf{V}\|_{\ell^2(\mathbb{C}^I)}}{\|\mathsf{V}\|_{\ell^2(\mathbb{C}^I)}} = \inf_{\mathsf{V} \in \mathbb{C}^I} \sup_{\mathsf{W} \in \mathbb{C}^I} \frac{|\mathsf{W}^{\mathsf{H}} \mathcal{A}\mathsf{V}|}{\|\mathsf{V}\|_{\ell^2(\mathbb{C}^I)} \|\mathsf{W}\|_{\ell^2(\mathbb{C}^I)}}, \tag{28.17a}$$

$$\omega_{\ell^2} := \sup_{\mathsf{V} \in \mathbb{C}^I} \frac{\|\mathcal{A}\mathsf{V}\|_{\ell^2(\mathbb{C}^I)}}{\|\mathsf{V}\|_{\ell^2(\mathbb{C}^I)}} = \sup_{\mathsf{V} \in \mathbb{C}^I} \sup_{\mathsf{W} \in \mathbb{C}^I} \frac{|\mathsf{W}^{\mathsf{H}} \mathcal{A}\mathsf{V}|}{\|\mathsf{V}\|_{\ell^2(\mathbb{C}^I)} \|\mathsf{W}\|_{\ell^2(\mathbb{C}^I)}}. \tag{28.17b}$$

We have $\omega_{\ell^2} = \|\mathcal{A}\|_{\ell^2(\mathbb{C}^I)}$, and one can verify that $\alpha_{\ell^2} = \|\mathcal{A}^{-1}\|_{\ell^2(\mathbb{C}^I)}^{-1}$ (see Exercise 28.2). The real numbers α_{ℓ^2} and ω_{ℓ^2} are called *smallest and largest singular values* of $\mathcal{A}$, respectively. Our goal is derive upper and lower bounds on ω_{ℓ^2} and α_{ℓ^2}. To this purpose, we introduce the following real numbers:

$$\alpha_{L^2} := \inf_{v_h \in V_h} \sup_{w_h \in W_h} \frac{|a_h(v_h, w_h)|}{\|v_h\|_{L^2(D)} \|w_h\|_{L^2(D)}}, \tag{28.18a}$$

$$\omega_{L^2} := \sup_{v_h \in V_h} \sup_{w_h \in W_h} \frac{|a_h(v_h, w_h)|}{\|v_h\|_{L^2(D)} \|w_h\|_{L^2(D)}}. \tag{28.18b}$$

Note that we are not using the natural norms in V_h and W_h but the L^2-norm.

Proposition 28.11 (Bounds on $\mathcal{A}$ and $\mathcal{A}^{-1}$). *The following holds true:*

$$(\mu_{\min}^{\varphi} \mu_{\min}^{\psi})^{\frac{1}{2}} \omega_{L^2} \le \omega_{\ell^2} = \|\mathcal{A}\|_{\ell^2(\mathbb{C}^I)} \le (\mu_{\max}^{\varphi} \mu_{\max}^{\psi})^{\frac{1}{2}} \omega_{L^2},$$

$$(\mu_{\max}^{\varphi} \mu_{\max}^{\psi})^{-\frac{1}{2}} \alpha_{L^2}^{-1} \le \alpha_{\ell^2}^{-1} = \|\mathcal{A}^{-1}\|_{\ell^2(\mathbb{C}^I)} \le (\mu_{\min}^{\varphi} \mu_{\min}^{\psi})^{-\frac{1}{2}} \alpha_{L^2}^{-1},$$

where the μ's denote (with obvious notation) the minimal/maximal eigenvalues of the mass matrices $\mathcal{M}_{\varphi}$ and $\mathcal{M}_{\psi}$.

Proof. Since R_{φ} and R_{ψ} are isomorphisms, we infer that

$$\alpha_{L^2} = \inf_{\mathsf{V} \in \mathbb{C}^I} \sup_{\mathsf{W} \in \mathbb{C}^I} \frac{|\mathsf{W}^{\mathsf{H}} \mathcal{A}\mathsf{V}|}{\|\mathsf{R}_{\varphi}(\mathsf{V})\|_{L^2(D)} \|\mathsf{R}_{\psi}(\mathsf{W})\|_{L^2(D)}},$$

$$\omega_{L^2} = \sup_{\mathsf{V} \in \mathbb{C}^I} \sup_{\mathsf{W} \in \mathbb{C}^I} \frac{|\mathsf{W}^{\mathsf{H}} \mathcal{A}\mathsf{V}|}{\|\mathsf{R}_{\varphi}(\mathsf{V})\|_{L^2(D)} \|\mathsf{R}_{\psi}(\mathsf{W})\|_{L^2(D)}}.$$

Let $\xi_{\ell^2}(\mathsf{V}, \mathsf{W}) := \frac{|\mathsf{W}^{\mathsf{H}} \mathcal{A}\mathsf{V}|}{\|\mathsf{V}\|_{\ell^2(\mathbb{C}^I)} \|\mathsf{W}\|_{\ell^2(\mathbb{C}^I)}}$, $\xi_{L^2}(\mathsf{V}, \mathsf{W}) := \frac{|\mathsf{W}^{\mathsf{H}} \mathcal{A}\mathsf{V}|}{\|\mathsf{R}_{\varphi}(\mathsf{V})\|_{L^2(D)} \|\mathsf{R}_{\psi}(\mathsf{W})\|_{L^2(D)}}$. We have $\xi_{\ell^2}(\mathsf{V}, \mathsf{W}) = \xi_{L^2}(\mathsf{V}, \mathsf{W}) \frac{\|\mathsf{R}_{\varphi}(\mathsf{V})\|_{L^2(D)}}{\|\mathsf{V}\|_{\ell^2(\mathbb{C}^I)}} \frac{\|\mathsf{R}_{\psi}(\mathsf{W})\|_{L^2(D)}}{\|\mathsf{W}\|_{\ell^2(\mathbb{C}^I)}}$. Owing to (28.11), we infer that

$$\xi_{L^2}(\mathsf{V}, \mathsf{W})(\mu_{\min}^{\varphi} \mu_{\min}^{\psi})^{\frac{1}{2}} \le \xi_{\ell^2}(\mathsf{V}, \mathsf{W}) \le \xi_{L^2}(\mathsf{V}, \mathsf{W})(\mu_{\max}^{\varphi} \mu_{\max}^{\psi})^{\frac{1}{2}}.$$

The expected bounds follow by taking the supremum over W and then the infimum or the supremum over V. $\square$

Recalling the definition (28.10) of the Euclidean condition number, Proposition 28.10 implies that

$$c_{\mathcal{M}}^{-1} \frac{\omega_{L^2}}{\alpha_{L^2}} \le \kappa_{\ell^2}(\mathcal{A}) \le c_{\mathcal{M}} \frac{\omega_{L^2}}{\alpha_{L^2}}, \tag{28.19}$$

with $c_{\mathcal{M}} := (\kappa_{\ell^2}(\mathcal{M}_{\varphi}) \kappa_{\ell^2}(\mathcal{M}_{\psi}))^{\frac{1}{2}}$. Since the mass matrices $\mathcal{M}_{\varphi}$ and $\mathcal{M}_{\psi}$ are expected to be relatively well-conditioned (see in particular Proposition 28.6), it is reasonable to expect that sharp bounds for $\kappa_{\ell^2}(\mathcal{A})$ can be obtained once sharp estimates of the real numbers α_{L^2} and ω_{L^2} are available.

Example 28.12 (Elliptic PDEs). Consider the bilinear form $a_h(v_h, w_h) := \int_D \nabla v_h \cdot \nabla w_h \, dx$ on $V_h \times V_h$ with quasi-uniform meshes. The global Poincaré–Steklov inequality together with the existence of large-scale discrete functions in V_h (i.e., some interpolant of the distance to the boundary of D) imply that α_{L^2} is uniformly equivalent to ℓ_D^{-2}, where ℓ_D is a characteristic length of D, e.g., $\ell_D := \mathrm{diam}(D)$. Moreover, a discrete inverse inequality together with the existence of small-scale functions in V_h (e.g., the global shape functions) implies that ω_{L^2} is uniformly equivalent to h^{-2}. Hence, the Euclidean condition number of the stiffness matrix is uniformly equivalent to $\ell_D^2 h^{-2}$. In dimension one, this can be verified explicitly for $\mathbb{P}_1$ Lagrange elements on a uniform mesh since the eigenvalues of $\mathcal{A} = \frac{1}{h}\mathrm{tridiag}(-1, 2, -1)$ are $\{\frac{2}{h}(1 - \cos(\pi l h))\}_{l \in \{1:I\}}$ (compare with Example 28.7 for the eigenvalues of the 1D mass matrix). See Strang and Fix [359] for early results on elliptic PDEs and [186] for further insight and more examples. □

Remark 28.13 (Ill-conditioning). In general, the Euclidean condition number of the stiffness matrix grows as the mesh is refined. This growth may have an important impact on the efficiency of iterative solvers when it comes to solving the linear system $\mathcal{A}\mathsf{U} = \mathsf{B}$; see §28.3.2. Note though that the sensitivity to perturbations induced by the growth of the condition number is usually not a major concern since a discrete stability property can be formulated by using suitable norms; see Exercise 28.7. □

Remark 28.14 (Choice of basis). The condition number of the stiffness matrix and that of the mass matrix depend on the choice made for the global shape functions. Using well-chosen hierarchical bases leads to a stiffness matrix having a condition number $\kappa_{\ell^2}(\mathcal{A})$ uniformly bounded in $h \in \mathcal{H}$; see, e.g., Hackbusch [235], Bramble et al. [81]. However, if $\kappa_{\ell^2}(\mathcal{A})$ is bounded, then $\kappa_{\ell^2}(\mathcal{M})$ must explode as $h \to 0$, i.e., it is not possible to find bases for which both $\mathcal{A}$ and $\mathcal{M}$ are well-conditioned. □

Remark 28.15 (Dependence on polynomial degree). Estimates on the condition number of the stiffness matrix on a single mesh cell with high-order polynomials can be found in Olsen and Douglas [320], Hu et al. [249]. □

28.2.4 Max-norm estimates

The notion of M-matrix is important in the real case when discretizing a PDE that enjoys a maximum principle; see §33.2 for an example.

Definition 28.16 (M-matrix and Z-matrix). *A matrix $\mathcal{A} \in \mathbb{R}^{I \times I}$ is said to be a Z-matrix if $\mathcal{A}_{ij} \leq 0$ for all $i, j \in \{1:I\}$ with $i \neq j$. A matrix $\mathcal{A}$ is said to be a nonsingular M-matrix if it is a Z-matrix, invertible, and $(\mathcal{A}^{-1})_{ij} \geq 0$ for all $i, j \in \{1:I\}$.*

A nonsingular M-matrix $\mathcal{A}$ enjoys several interesting properties: $\mathcal{A}\mathsf{V} \geq 0$ implies $\mathsf{V} \geq 0$ for all $V \in \mathbb{R}^I$ (where $\mathsf{V} \geq 0$ means $\mathsf{V}_i \geq 0$ for all $i \in \{1:I\}$); all the eigenvalues of $\mathcal{A}$ have positive real part; and all the diagonal entries of $\mathcal{A}$ are positive; see e.g., in Plemmons [326].

Lemma 28.17 (Majorizing vector). *Let $\mathcal{A} \in \mathbb{R}^{I \times I}$ be a Z-matrix. Then $\mathcal{A}$ is a nonsingular M-matrix iff there is a vector $\mathsf{Y} \in \mathbb{R}^I$ called majorizing vector s.t. $\mathsf{Y} > 0$ and $\mathcal{A}\mathsf{Y} > 0$, i.e., $\mathsf{Y}_i > 0$ and $(\mathcal{A}\mathsf{Y})_i := \sum_{j \in \{1:I\}} \mathcal{A}_{ij}\mathsf{Y}_j > 0$ for all $i \in \{1:I\}$.*

Proof. See Grossmann and Roos [225, p. 70]. □

Let us recall that $\|\mathsf{V}\|_{\ell^\infty(\mathbb{R}^I)} := \max_{j \in \{1:I\}} |\mathsf{V}_j|$ for all $V \in \mathbb{R}^I$, and that the induced matrix norm, which we also denote by $\|\cdot\|_{\ell^\infty(\mathbb{R}^I)}$, is such that $\|\mathcal{Z}\|_{\ell^\infty(\mathbb{R}^I)} = \max_{i \in \{1:I\}} \sum_{j \in \{1:I\}} |\mathcal{Z}_{ij}|$ for all $\mathcal{Z} \in \mathbb{R}^{I \times I}$. It is possible to estimate $\|\mathcal{A}^{-1}\|_{\ell^\infty(\mathbb{R}^I)}$ as follows if $\mathcal{A}$ is a nonsingular M-matrix.

Proposition 28.18 (Bound on $\|\cdot\|_{\ell^\infty(\mathbb{R}^I)}$-norm). *Let $\mathcal{A}$ be a nonsingular M-matrix. Let Y be a majorizing vector for $\mathcal{A}$. The following holds true:*

$$\|\mathcal{A}^{-1}\|_{\ell^\infty(\mathbb{R}^I)} \leq \frac{\|\mathsf{Y}\|_{\ell^\infty(\mathbb{R}^I)}}{\min_{i\in\{1:I\}}(\mathcal{A}\mathsf{Y})_i}. \tag{28.20}$$

Proof. See Exercise 28.8. $\square$

28.3   Solution methods

This section briefly reviews some methods to solve the linear system $\mathcal{A}\mathsf{U} = \mathsf{B}$ resulting from the Galerkin approximation. We will see in Chapter 29 that finite element-based matrices are generally sparse. This means that the number of nonzero entries in $\mathcal{A}$ is significantly smaller than the total number of entries. It is important to keep this property in mind when considering solution methods.

28.3.1   Direct methods

The best-known example of direct method for solving the linear system $\mathcal{A}\mathsf{U} = \mathsf{B}$ consists of constructing the *LU factorization* of $\mathcal{A}$. Recall that a matrix $\mathcal{Z} \in \mathbb{C}^{I\times I}$ is said to be lower (resp., upper) triangular if $\mathcal{Z}_{ij} = 0$ for all $1 \leq i < j \leq I$ (resp., $\mathcal{Z}_{ij} = 0$ for all $1 \leq j < i \leq I$). Let σ be any permutation of the set $\{1{:}I\}$, the matrix $\mathcal{P} \in \mathbb{C}^{I\times I}$ with entries $\mathcal{P}_{ij} = \delta_{\sigma(i)j}$ (δ is the Kronecker symbol) is called *permutation matrix*. The LU factorization of $\mathcal{A}$ with complete pivoting takes the form

$$\mathcal{P}\mathcal{A}\mathcal{Q} = \mathcal{T}_{\mathrm{L}}\,\mathcal{T}_{\mathrm{U}}, \tag{28.21}$$

where $\mathcal{P}$ and $\mathcal{Q}$ are permutation matrices, $\mathcal{T}_{\mathrm{L}}$ is lower triangular, and $\mathcal{T}_{\mathrm{U}}$ is upper triangular. These matrices can be constructed by using *Gaussian elimination*; see, e.g., Golub and van Loan [218, pp. 96-119]. The permutation matrices $\mathcal{P}$ and $\mathcal{Q}$ are needed to avoid divisions by zero or divisions by quantities that are very small compared to the other entries of the matrix. When the matrix $\mathcal{A}$ is Hermitian (symmetric in the real case), the right-hand side of (28.21) can be obtained in the form $\mathcal{T}_{\mathrm{L}}\,\mathcal{D}\,\mathcal{T}_{\mathrm{L}}^{\mathsf{H}}$, where $\mathcal{T}_{\mathrm{L}}$ is lower triangular with unit diagonal and $\mathcal{D}$ is diagonal; see [218, p. 137]. When $\mathcal{A}$ is Hermitian positive definite (symmetric positive definite in the real case), the matrix $\mathcal{D}$ can be incorporated in the matrix $\mathcal{T}_{\mathrm{L}}$ by means of *Choleski's factorization* so that $\mathcal{T}_{\mathrm{L}}$ has real and positive diagonal entries; see [218, p. 141].

Once the LU factorization (28.21) of $\mathcal{A}$ has been constructed, the linear system $\mathcal{A}\mathsf{U} = \mathsf{B}$ is solved by performing the following three steps: (i) Solve the lower triangular system $\mathcal{T}_{\mathrm{L}}\mathsf{U}' = \mathcal{P}\mathsf{B}$. (ii) Solve the upper triangular system $\mathcal{T}_{\mathrm{U}}\mathsf{U}'' = \mathsf{U}'$. (iii) Set $\mathsf{U} := \mathcal{Q}\mathsf{U}''$. In practice, the cost of computing the LU factorization dominates that of solving the triangular systems (for a dense $I\times I$ matrix, the former scales as $\frac{1}{3}I^3$ and the latter as $\frac{1}{2}I^2$).

An important issue in the context of sparse matrices is the fill-in induced by the LU decomposition. The (l,u)-bandwidth of a sparse matrix $\mathcal{A}$ is obtained from the two integers l,u such that $l := \max\{p \mid \mathcal{A}_{ij} = 0,\ i > j + p\}$ and $u := \max\{q \mid \mathcal{A}_{ij} = 0,\ j > i + q\}$. For instance, $l = u = 0$ for a diagonal matrix, $l = u = 1$ for a tridiagonal matrix, and so on. It can be shown that if $\mathcal{A}$ has (l,u)-bandwidth and if $\mathcal{A}$ has a LU factorization without pivoting, then $\mathcal{T}_{\mathrm{L}}$ has $(l,0)$-bandwidth and $\mathcal{T}_{\mathrm{U}}$ has $(0,u)$-bandwidth; see [218, Thm. 4.3.1]. Hence, the LU factorization does not increase the bandwidth, but the matrices $\mathcal{T}_{\mathrm{L}}$ and $\mathcal{T}_{\mathrm{U}}$ have more nonzero entries than $\mathcal{A}$. This fill-in can be partly tamed by using reordering techniques; see §29.3. Broadly speaking *sparse direct solvers* can be competitive alternatives to iterative methods (see below) for linear systems obtained by

approximating PDEs posed in dimension two, but this is no longer the case in dimension three on sufficiently fine meshes.

Remark 28.19 (Literature). We refer the reader to George and Liu [214], Duff et al. [177], Demmel et al. [160], Davis [156], Björck [58] for an overview of sparse direct solvers. □

28.3.2 Iterative methods

Using an iterative method for solving a large sparse linear system presents the twofold advantage of avoiding additional matrix storage and taking full advantage of sparsity by only performing matrix-vector products. An iterative method is initialized by some vector $U_0 \in \mathbb{C}^I$ and then produces a sequence $(U_m)_{m\geq1}$ of vectors in $\mathbb{C}^I$ that is expected to converge to the solution of the linear system.

When the stiffness matrix is Hermitian positive definite, the *conjugate gradient (CG)* method designed by Hestenes and Stiefel in 1952 [242] is particularly effective. The CG method is presented in Algorithm 28.1. One matrix-vector product needs to be performed at each iteration (as well as

---

Algorithm 28.1 Conjugate gradient.

    choose $U_0 \in \mathbb{C}^I$, set $R_0 := B - \mathcal{A}U_0$ and $P_0 := R_0$
    choose a tolerance `tol` and set $m := 0$
    while $\|R_m\|_{\ell^2(\mathbb{C}^I)} > $ `tol` **do**
        $\alpha_m := R_m^H R_m / P_m^H \mathcal{A} P_m$
        $U_{m+1} := U_m + \alpha_m P_m$
        $R_{m+1} := R_m - \alpha_m \mathcal{A} P_m$
        $\beta_m := R_{m+1}^H R_{m+1} / R_m^H R_m$
        $P_{m+1} := R_{m+1} + \beta_m P_m$
        $m \leftarrow m + 1$
    end while

---

two inner products and three vector updates, but these operations induce a marginal computational cost compared to the matrix-vector product). One can show by induction that $R_m = B - \mathcal{A}U_m$, that $\{R_0, \dots, R_{m-1}\}$ is an ℓ^2-orthogonal set, and that $\{P_0, \dots, P_{m-1}\}$ is an $\mathcal{A}$-orthogonal set; see Saad [339, Prop. 6.13]. The crucial property of CG is the following [339, Prop. 5.2].

Proposition 28.20 (Optimality of CG). *Let $\mathbb{C}^I$ be equipped with the energy norm $\|\cdot\|_{\mathcal{A}} := (\mathcal{A}\cdot, \cdot)_{\ell^2(\mathbb{C}^I)}^{\frac{1}{2}}$. Then, at step $m \geq 1$ of CG (provided no termination has occurred), U_m satisfies the following optimality property:*

$$\|U - U_m\|_{\mathcal{A}} = \min_{Y \in U_0 + K_m} \|U - Y\|_{\mathcal{A}}, \tag{28.22}$$

with the Krylov subspace $K_m := \mathrm{span}\{R_0, \mathcal{A}R_0, \dots, \mathcal{A}^{m-1}R_0\}$.

The optimality property guarantees that CG terminates in at most I steps (in the absence of roundoff errors). In practice, termination often occurs much earlier. It is a remarkable fact that CG provides an optimality property over the whole affine subspace $U_0 + K_m$ without needing to store an entire basis of K_m. This nice property is unfortunately lost if $\mathcal{A}$ is not Hermitian; see Faber and Manteuffel [198], Voevodin [379]. When $\mathcal{A}$ is symmetric but indefinite, it is still possible to achieve some optimality property over $U_0 + K_m$; see §50.3.2 for mixed finite element approximations. Solving the normal equations $\mathcal{A}^H \mathcal{A} U = \mathcal{A}^H B$ by CG is generally not a good idea since it leads to very poor convergence rates.

There are two broad classes of Krylov subspace methods for non-Hermitian matrices. On the one hand there are those that guarantee an optimality property over a subspace but require to store a complete basis, thereby making storage and computational costs grow linearly with the number of iterations. An important example is the generalized minimal residual (GMRES) method (Marchuk and Kuznetsov [292, 293], Saad and Schultz [340]) where U_m minimizes the Euclidean norm of the residual over $U_0 + K_m$. The computational costs of GMRES can often be tamed by using a restarted version. On the other hand there are those methods that give up optimality, but employ short-term recurrences to compute the iterates. Examples are the conjugate gradient squared (CGS) (Sonneveld [350]) and the bi-conjugate gradient stabilized (BI-CGSTAB) (van der Vorst [370]) methods. These methods often work well in practice, although convergence is not guaranteed.

The convergence rate of Krylov subspace methods depends on the spectrum of $\mathcal{A}$. Sharp bounds can be derived in the normal case (i.e., $\mathcal{A}$ commutes with $\mathcal{A}^{\mathsf{H}}$), whereas bounds in the nonnormal case are more delicate and involve also the eigenvectors. We state the following result for CG when $\mathcal{A}$ is Hermitian positive definite (see Saad [339, p. 193] or Elman et al. [185, p. 75]).

Proposition 28.21 (Convergence rate of CG). *Let $\mathcal{A} \in \mathbb{C}^{I \times I}$ be a Hermitian positive definite matrix and let $\kappa(\mathcal{A})$ be its (Euclidean) condition number. The following holds true for the CG iterates:*

$$\|U - U_m\|_{\mathcal{A}} \le 2 \left(\frac{\kappa(\mathcal{A})^{\frac{1}{2}} - 1}{\kappa(\mathcal{A})^{\frac{1}{2}} + 1} \right)^m \|U - U_0\|_{\mathcal{A}}. \tag{28.23}$$

Remark 28.22 (Clustering of eigenvalues). A sharper bound is $\|U - U_m\|_{\mathcal{A}} \le c_m \|U - U_0\|_{\mathcal{A}}$ with $c_m := \min_{p \in \mathbb{P}_m, p(0)=1} \max_{\lambda \in \sigma(\mathcal{A})} |p(\lambda)|$, showing that clustering of the eigenvalues around a few points (even spread out) is favorable to fast convergence. Note that (28.23) is derived from this bound by writing $c_m \le \min_{p \in \mathbb{P}_m, p(0)=1} \|p\|_{C^0([s_\flat, s_\sharp])}$ with $\sigma(\mathcal{A}) \subset [s_\flat, s_\sharp]$ and constructing a suitable minimizing polynomial (recall that $\sigma(\mathcal{A}) \subset [0, \infty)$ denotes the spectrum of $\mathcal{A}$). □

Preconditioning can be very effective to speed-up the convergence of Krylov subspace methods, the ideal goal being to achieve computational costs that grow linearly with the size of the linear system. Let $\mathcal{P}_{\mathrm{L}}, \mathcal{P}_{\mathrm{R}} \in \mathbb{C}^{I \times I}$ be two nonsingular matrices and assume that linear systems of the form $\mathcal{P}_{\mathrm{L}} X = Y$ and $\mathcal{P}_{\mathrm{R}} X' = Y'$ are relatively inexpensive to solve. Then U solves $\mathcal{A}U = B$ if and only if $\widetilde{U} := \mathcal{P}_{\mathrm{R}} U$ solves $\widetilde{\mathcal{A}}\widetilde{U} = \widetilde{B}$, where $\widetilde{\mathcal{A}} := \mathcal{P}_{\mathrm{L}}^{-1} \mathcal{A} \mathcal{P}_{\mathrm{R}}^{-1}$ and $\widetilde{B} := \mathcal{P}_{\mathrm{L}}^{-1} B$. When $\mathcal{A}$ is Hermitian, one can take $\mathcal{P}_{\mathrm{R}} = \mathcal{P}_{\mathrm{L}}^{\mathsf{H}}$, and CG can be implemented by just considering the matrix $\mathcal{P} := \mathcal{P}_{\mathrm{L}} \mathcal{P}_{\mathrm{L}}^{\mathsf{H}}$; see Exercise 28.10.

Choosing a preconditioner is a compromise between computational cost per iteration and improving the spectral properties of the preconditioned matrix by clustering its eigenvalues. Relatively simple preconditioners can be derived by using the splitting $\mathcal{A} = \mathcal{A}_+ - \mathcal{A}_-$ (this type of splitting arises naturally in the context of stationary fixed-point iterations) and by using $\mathcal{A}_+$ as a preconditioner. *Incomplete LU (ILU)* preconditioning is generally a robust choice, the idea being to discard entries in the LU factorization that do not match the sparsity pattern of $\mathcal{A}$; see [339, §10.3]. Many other preconditioning techniques are available in the literature. A particularly important class is that of the multilevel (or multigrid) preconditioners where the solution is expanded over a more or less hierarchical basis; see Bramble et al. [80, 81], Briggs [94], Elman et al. [185], Hackbusch [235], Trottenberg et al. [365], Wesseling [393] for further insight into this topic.

Remark 28.23 (From complex to real linear systems). Solving the complex linear system $\mathcal{A}U = B$ can be avoided by rewriting it as a linear system of twice the size of the real (or imaginary) part of U. Using the obvious notation $(\mathcal{R} + \mathrm{i}\mathcal{S})(U_1 + \mathrm{i}U_2) = B_1 + \mathrm{i}B_2$ with $\mathrm{i}^2 = -1$ and $\mathcal{R}, \mathcal{S} \in \mathbb{R}^{I \times I}$,

two possible rewritings of $\mathcal{A}\mathsf{U} = \mathsf{B}$ are

$$\begin{pmatrix} \mathcal{R} & -\mathcal{S} \\ \mathcal{S} & \mathcal{R} \end{pmatrix} \begin{pmatrix} \mathsf{U}_1 \\ \mathsf{U}_2 \end{pmatrix} = \begin{pmatrix} \mathsf{B}_1 \\ \mathsf{B}_2 \end{pmatrix}, \qquad \begin{pmatrix} \mathcal{R} & \mathcal{S} \\ \mathcal{S} & -\mathcal{R} \end{pmatrix} \begin{pmatrix} \mathsf{U}_1 \\ -\mathsf{U}_2 \end{pmatrix} = \begin{pmatrix} \mathsf{B}_1 \\ \mathsf{B}_2 \end{pmatrix}.$$

Let us denote by $\mathcal{A}_*$ and $\mathcal{A}_{**}$ the above two real matrices of size $2I \times 2I$. Note that in general both matrices are nonsymmetric and/or indefinite. Unfortunately, the distribution of the eigenvalues in the spectra $\sigma(\mathcal{A}_*)$ and $\sigma(\mathcal{A}_{**})$ are typically unfavorable for the convergence of Krylov subspace methods as the eigenvalues embrace the origin of the complex plane with a large number of eigenvalues straddling the origin. Specifically, one can show that $\sigma(\mathcal{A}_*) = \sigma(\mathcal{A}) \cup \overline{\sigma(\mathcal{A})}$ is symmetric with respect to the real line, whereas $\sigma(\mathcal{A}_{**})$ is symmetric with respect to both the real and imaginary lines and we have $\sigma(\mathcal{A}_{**}) = \{\lambda \in \mathbb{C} \mid \lambda^2 \in \sigma(\overline{\mathcal{A}}\mathcal{A})\}$; see Freund [208, Prop. 5.1]. Hence, it is in general preferable to deploy Krylov subspace methods on the complex linear system than on the equivalent real ones. Effective algorithms can be devised by exploiting some particular structure of $\mathcal{A}$, e.g., if $\mathcal{A}$ is complex symmetric (i.e., $\mathcal{A} = \mathcal{A}^\mathsf{T}$ instead of $\mathcal{A} = \mathcal{A}^\mathsf{H}$); see Freund [208], Axelsson and Kucherov [31]. $\square$

Exercises

Exercise 28.1 (Matrix representation of operators). Let H be a (complex) Hilbert space with inner product $(\cdot, \cdot)_H$. Let V_h be a finite-dimensional subspace of H with basis $\{\varphi_i\}_{i \in \{1:I\}}$. Let $Z : V_h \to V_h$ be a linear operator. Let $\mathcal{M} \in \mathbb{C}^{I \times I}$ be the mass matrix s.t. $\mathcal{M}_{ij} := (\varphi_j, \varphi_i)_H$, and let $\mathcal{B}, \mathcal{D} \in \mathbb{C}^{I \times I}$ be s.t. $\mathcal{B}_{ij} := (Z(\varphi_j), \varphi_i)_H$, $\mathcal{D}_{ij} := (Z(\varphi_j), Z(\varphi_i))_H$ for all $i, j \in \{1:I\}$. Prove that $\mathcal{D} = \mathcal{B}^\mathsf{H} \mathcal{M}^{-1} \mathcal{B}$. (*Hint*: use $\mathcal{Z} \in \mathbb{C}^{I \times I}$ s.t. $Z(\varphi_j) := \sum_{k \in \{1:I\}} \mathcal{Z}_{kj} \varphi_k$.)

Exercise 28.2 (Smallest singular value). Prove that the real number α_{ℓ^2} defined (28.17a) is equal to $\|\mathcal{A}^{-1}\|_{\ell^2(\mathbb{C}^I)}^{-1}$. (*Hint*: to bound α_{ℓ^2}, consider a vector $\mathsf{V}_* \in \mathbb{C}^I$ s.t. $\|\mathcal{A}^{-1}\mathsf{V}_*\|_{\ell^2(\mathbb{C}^I)} = \|\mathcal{A}^{-1}\|_{\ell^2(\mathbb{C}^I)}\|\mathsf{V}_*\|_{\ell^2(\mathbb{C}^I)}$.)

Exercise 28.3 (ℓ^2-condition number). Let $\mathcal{Z} \in \mathbb{R}^{I \times I}$ be the upper triangular matrix such that $\mathcal{Z}_{ii} := 1$ for all $i \in \{1:I\}$, and $\mathcal{Z}_{ij} := -1$ for all $i, j \in \{1:I\}$, $i \neq j$. Let $\mathsf{X} \in \mathbb{R}^I$ have coordinates $\mathsf{X}_i := 2^{1-i}$ for all $i \in \{1:I\}$. Compute $\mathcal{Z}\mathsf{X}$, $\|\mathcal{Z}\mathsf{X}\|_{\ell^2(\mathbb{R}^I)}$, and $\|\mathsf{X}\|_{\ell^2(\mathbb{R}^I)}$. Show that $\|\mathcal{Z}\|_{\ell^2(\mathbb{R}^I)} \geq 1$ and derive a lower bound for $\kappa_{\ell^2}(\mathcal{Z})$. What happens if I is large?

Exercise 28.4 (Local mass matrix, 1D). Evaluate the local mass matrix for one-dimensional $\mathbb{P}_1$ and $\mathbb{P}_2$ Lagrange finite elements on a cell of length h.

Exercise 28.5 (Stiffness matrix). (i) Let $\{\widehat{\lambda}_1, \widehat{\lambda}_2, \widehat{\lambda}_3\}$ be the shape functions of the $\mathbb{P}_1$ Lagrange element with the cell $\widehat{K}$ shown on the leftmost part of Figure 28.1. Here, $\widehat{\lambda}_1$ is associated with the vertex $(1,0)$, $\widehat{\lambda}_2$ with the vertex $(0,1)$, and $\widehat{\lambda}_3$ with the vertex $(0,0)$. Evaluate the stiffness matrix for $\int_{\widehat{K}} \nabla v \cdot \nabla w \, \mathrm{d}x$. Same question for the $\mathbb{Q}_1$ Lagrange element with the shape functions $\{\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3, \widehat{\theta}_4\}$ associated with the vertices $(1,0), (1,1), (0,1), (0,0)$, respectively (see the central part of Figure 28.1). (ii) Consider the meshes of $D := (0,3) \times (0,2)$ shown in the right part of Figure 28.1. Evaluate the stiffness matrix for $\int_D \nabla v \cdot \nabla w \, \mathrm{d}x$.

Exercise 28.6 (Sensitivity to perturbations). Let $\mathcal{Z} \in \mathbb{C}^{I \times I}$ be invertible and let $\mathsf{X} \in \mathbb{C}^I$ solve $\mathcal{Z}\mathsf{X} = \mathsf{B}$ with $\mathsf{B} \neq 0$. Set $\kappa := \kappa_{\ell^2}(\mathcal{Z})$. (i) Let $\tilde{\mathsf{X}} \in \mathbb{C}^I$ solve $\mathcal{Z}\tilde{\mathsf{X}} = \tilde{\mathsf{B}}$. Prove that $\frac{\|\tilde{\mathsf{X}}-\mathsf{X}\|_{\ell^2(\mathbb{C}^I)}}{\|\mathsf{X}\|_{\ell^2(\mathbb{C}^I)}} \leq \kappa \frac{\|\tilde{\mathsf{B}}-\mathsf{B}\|_{\ell^2(\mathbb{C}^I)}}{\|\mathsf{B}\|_{\ell^2(\mathbb{C}^I)}}$. (ii) Let $\check{\mathsf{X}} \in \mathbb{C}^I$ solve $\check{\mathcal{Z}}\check{\mathsf{X}} = \mathsf{B}$. Prove that $\frac{\|\check{\mathsf{X}}-\mathsf{X}\|_{\ell^2(\mathbb{C}^I)}}{\|\check{\mathsf{X}}\|_{\ell^2(\mathbb{C}^I)}} \leq \kappa \frac{\|\check{\mathcal{Z}}-\mathcal{Z}\|_{\ell^2(\mathbb{C}^I)}}{\|\mathcal{Z}\|_{\ell^2(\mathbb{C}^I)}}$. (iii) Explain why the above bounds are sharp.
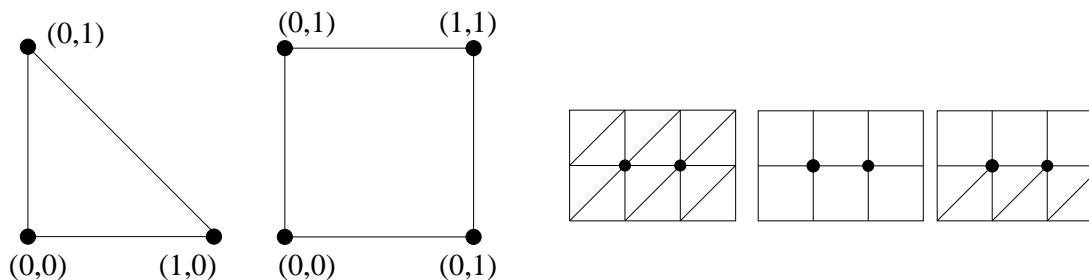
Figure 28.1: Illustration for Exercise 28.5. Left and central panels: reference triangle and square considered in Step (i). Right panel: three meshes for Step (ii).

Exercise 28.7 (Stability). Let $\mathcal{A}\mathsf{U} = \mathsf{B}$ be the linear system resulting from the Galerkin approximation. Equip the vector space $\mathbb{C}^I$ with the norm $\|\mathsf{V}\|_* := \sup_{\mathsf{Y}\in\mathbb{C}^I} \frac{|\mathsf{V}^\mathsf{H}\mathsf{Y}|}{\|\mathsf{R}_\psi(\mathsf{Y})\|_{W_h}}$. Show that $\frac{\|u_h - v_h\|_{V_h}}{\|u_h\|_{V_h}} \leq \frac{\|a_h\|}{\alpha_h} \frac{\|\mathsf{B}-\mathcal{A}\mathsf{V}\|_*}{\|\mathsf{B}\|_*}$ for all $\mathsf{V}\in\mathbb{C}^I$, where $u_h := \mathsf{R}_\varphi(\mathsf{U})$ and $v_h := \mathsf{R}_\varphi(\mathsf{V})$. (*Hint*: show that $\alpha_h\|u_h - v_h\|_{V_h} \leq \|\mathcal{A}(\mathsf{U} - \mathsf{V})\|_*$ and that $\|\mathsf{B}\|_* \leq \|a_h\|\|u_h\|_{V_h}$, where α_h and $\|a_h\|$ are the stability and boundedness constants of a_h on $V_h{\times}W_h$.)

Exercise 28.8 (ℓ^∞-norm). (i) Prove Proposition 28.18. (*Hint*: use that $\mathcal{A}\mathsf{Y} \geq \min_{j\in\{1:I\}}(\mathcal{A}\mathsf{Y})_j\mathsf{U}$, where $\mathsf{U}\in\mathbb{R}^I$ has all entries equal to 1.) (ii) Derive a bound on $\|\mathcal{A}^{-1}\|_{\ell^\infty(\mathbb{R}^I)}$ with $\mathcal{A} := h^{-1}\operatorname{tridiag}(-1, 2, -1)$. (*Hint*: consider the function $x \mapsto x(1-x)$ on $(0,1)$ to build a majorizing vector.) (iii) Let $(\mathsf{E}_1, \ldots, \mathsf{E}_I)$ be the canonical basis of $\mathbb{R}^I$. Let $\alpha\in\mathbb{R}$ and consider the matrix $\mathcal{Z} := \mathcal{I} + \alpha\mathsf{E}_1\otimes\mathsf{E}_I$ with entries $\mathcal{Z}_{ij} := \delta_{ij} + \alpha\delta_{i1}\delta_{jI}$. Verify that $\mathcal{Z}^{-1} = \mathcal{I} - \alpha\mathsf{E}_1\otimes\mathsf{E}_I$ and evaluate the condition number $\kappa_{\ell^\infty}(\mathcal{Z})$. What happens if α is large?

Exercise 28.9 (Lumped mass matrix). Let D be a two-dimensional polygonal set and consider an affine mesh $\mathcal{T}_h$ of D composed of triangles and $\mathbb{P}_1$ Lagrange elements. (i) Let K be a cell in $\mathcal{T}_h$. Compute the local mass matrix $\mathcal{M}^K$ with entries $\mathcal{M}_{ij}^K := \int_K \theta_{K,i}(\boldsymbol{x})\theta_{K,j}(\boldsymbol{x})\,\mathrm{d}x$, $i, j \in \{1:3\}$. (ii) Compute the lumped local mass matrix $\overline{\mathcal{M}}^K$ with $\overline{\mathcal{M}}_{ij}^K := \delta_{ij}\sum_{l\in\{1:3\}}\mathcal{M}_{il}^K$. (iii) Compute the eigenvalues of $(\overline{\mathcal{M}}^K)^{-1}(\overline{\mathcal{M}}^K - \mathcal{M}^K)$. (iv) Let $\mathcal{M}$ be the global mass matrix and $\overline{\mathcal{M}}$ be the lumped mass matrix. Show that the largest eigenvalue of $(\overline{\mathcal{M}})^{-1}(\overline{\mathcal{M}} - \mathcal{M})$ is $\frac{3}{4}$.

Exercise 28.10 (CG). Let $\mathcal{A} \in \mathbb{R}^{I\times I}$ be a real symmetric positive definite matrix and let $\mathfrak{J} : \mathbb{R}^I \to \mathbb{R}$ be such that $\mathfrak{J}(\mathsf{V}) := \frac{1}{2}\mathsf{V}^\mathsf{T}\mathcal{A}\mathsf{V} - \mathsf{B}^\mathsf{T}\mathsf{V}$. Let U_m be the iterate at step $m \geq 1$ of the CG method. (i) Prove that U_m minimizes $\mathfrak{J}$ over $\mathsf{U}_0 + K_m$. (*Hint*: use Proposition 28.20.) (ii) Let $\eta_m := \arg\min_{\eta\in\mathbb{C}}\mathfrak{J}(\mathsf{U}_m + \eta\mathsf{P}_m)$. Show that $\eta_m = \alpha_m$ in the CG method. (iii) Write the preconditioned CG method by just invoking the matrix $\mathcal{P} := \mathcal{P}_\mathrm{L}\mathcal{P}_\mathrm{L}^\mathsf{T}$.

Exercise 28.11 (Complex symmetric system). Let $\mathcal{A} := \mathcal{T} + \mathrm{i}\sigma\mathcal{I}$ where $\mathcal{T}$ is symmetric real, $\sigma > 0$, and $\mathcal{I}$ is the identity matrix of size $I{\times}I$. Let $\mathcal{A}_*$ and $\mathcal{A}_{**}$ be the two rewritings of $\mathcal{A}$ as a real matrix of size $2I{\times}2I$ (see Remark 28.23). Determine the spectra $\sigma(\mathcal{A})$, $\sigma(\mathcal{A}_*)$, and $\sigma(\mathcal{A}_{**})$, and comment on their position with respect to the origin. What happens if one considers the rotated linear system $-\mathrm{i}\mathcal{A}U = -\mathrm{i}B$ instead?

Chapter 29

Sparse matrices

A matrix is said to be sparse if the number of its nonzero entries is significantly smaller than the total number of its entries. The stiffness matrix is generally sparse as a consequence of the global shape functions having local support. This chapter deals with important computational aspects related to sparsity: storage, assembling, and reordering.

29.1 Origin of sparsity

Let us assume for simplicity that the discrete trial and the test spaces coincide. Recalling (28.1), the entries of the stiffness matrix $\mathcal{A} \in \mathbb{C}^{I \times Nglob}$ are given by $\mathcal{A}_{ij} := a_h(\varphi_j, \varphi_i)$ for all $i, j \in \{1{:}I\}$, for some sesquilinear form a_h evaluated by computing an integral over D. Let $\mathcal{T}_h$ be a mesh of D. Decomposing the integral as a sum over the mesh cells we write

$$\mathcal{A}_{ij} = \sum_{K \in \mathcal{T}_h} \int_K A_K(\boldsymbol{x}, \varphi_{j|K}, \varphi_{i|K}) \, dx, \qquad \forall i, j \in \{1{:}I\}, \tag{29.1}$$

for some local functional A_K acting on the restriction of the global shape functions to K. A crucial consequence of (29.1) is that

$$[\, \mathcal{A}_{ij} \neq 0 \,] \implies [\, |\operatorname{supp}(\varphi_i) \cap \operatorname{supp}(\varphi_j)| > 0 \,], \tag{29.2}$$

where $\operatorname{supp}(f)$ denotes the support in D of the function $f : D \to \mathbb{R}$ (i.e., the closure in D of the subset $\{\boldsymbol{x} \in D \mid f(\boldsymbol{x}) \neq 0\}$).

The support of a global shape function φ_i depends on the associated global degree of freedom, which is typically an evaluation at a node or an integral over an edge, a face, or a cell of the mesh. As a result, $\operatorname{supp}(\varphi_i)$ coincides with the set of the mesh cells containing the corresponding vertex, edge, face, or cell, respectively. Let $n_{\text{mesh}} := \max_{i \in \{1{:}I\}} \operatorname{card}\{K \subset \operatorname{supp}(\varphi_i)\}$. This number is bounded uniformly w.r.t. $h \in \mathcal{H}$ owing to the regularity of the mesh sequence. Let n_{sh} be the number of local shape functions. Let N_{row} be the maximum number of nonzero entries per row of $\mathcal{A}$. A consequence of (29.2) is that N_{row} is bounded from above as follows:

$$N_{\text{row}} \leq n_{\text{mesh}} \times (n_{\text{sh}} - 1) + 1. \tag{29.3}$$

The right-hand side of (29.3) being independent of $h \in \mathcal{H}$, we infer that the stiffness matrix becomes sparser as the mesh is refined. The bound (29.3) can be made sharper by considering the type of

support for the various global shape functions. For instance, the support of a global shape function associated with a vertex is larger than the support of a global shape function associated with an edge or a face.

Example 29.1 (Lagrange $\mathbb{P}_1$ and $\mathbb{P}_2$, 2D). Consider a two-dimensional matching simplicial mesh. Let n_{vtx} be the maximum number of edges arriving at a vertex. For Lagrange $\mathbb{P}_1$ finite elements, the global shape functions are attached to the mesh vertices, and each shape function interacts with at most n_{vtx} other shape functions, so that $N_{\mathrm{row}} = n_{\mathrm{vtx}}+1$ (left panel of Figure 29.1). For Lagrange $\mathbb{P}_2$ finite elements, the global shape functions are attached either to the mesh vertices or to the edge midpoints. The vertex shape functions interact with at most $3n_{\mathrm{vtx}}$ other shape functions, whereas the edge shape functions interact with at most 8 other shape functions. Hence, $N_{\mathrm{row}} = 3n_{\mathrm{vtx}} + 1$ since $n_{\mathrm{vtx}} \geq 3$, (central and right panels of Figure 29.1).   □
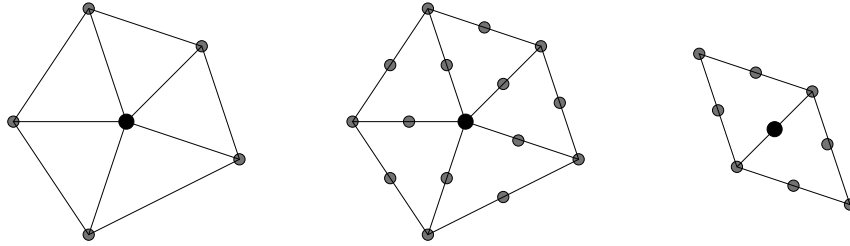
Figure 29.1: Left (Lagrange $\mathbb{P}_1$ element): the global shape function attached to the vertex in black interacts with the shape functions attached to the vertices in gray. Center and right (Lagrange $\mathbb{P}_2$ element): the global shape function attached to the vertex (center) or edge midpoint (right) in black interacts with the shape functions attached to the nodes in gray.

Example 29.2 (Sparsity pattern, structured mesh). The sparsity pattern of a $16{\times}16$ stiffness matrix using $\mathbb{P}_1$ Lagrange finite elements on a two-dimensional structured mesh is shown in the left panel of Figure 29.2. The mesh is shown in the right panel. The black squares in the sparsity pattern are the nonzero entries. There are at most seven nonzero entries per row, i.e., $N_{\mathrm{row}} = 7$. More generally, on a structured mesh that is built by using M nodes in each direction ($M = 4$ above), the stiffness matrix is of size $M^2{\times}M^2$, its entries are organized into a tridiagonal block-structure with M blocks of size $M{\times}M$, and each block is tridiagonal or bidiagonal. As a result, we also have $N_{\mathrm{row}} = 7$ in this case, independently of M.   □
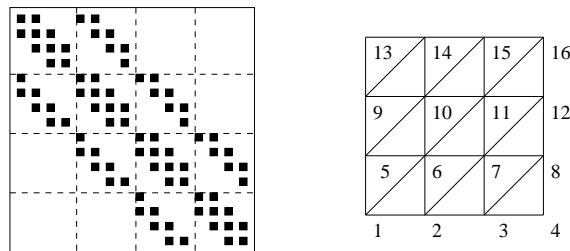
Figure 29.2: Sparsity pattern (left) and underlying mesh (right).

29.2   Storage and assembling

Compressed storage formats are crucial in large-scale applications to avoid wasting memory space by just storing zero entries. Assembling refers to the procedure in a finite element code where the entries of the stiffness matrix (and those of the right-hand side) are computed.

29.2.1   CSR and CSC formats

One of the most frequently used storage techniques is probably the *Compressed sparse rows* or *Compressed row storage* format (resp., *columns*), usually referred to as CSR or CRS format (resp., CSC or CCS). The CSR and CSC formats are very similar, the role played by rows and columns being simply interchanged. We only present the CSR format for brevity.

Let $\mathcal{A}(1{:}I, 1{:}I')$ be a sparse matrix not necessarily square and containing nnz nonzero entries. We define three arrays to store this matrix in the CSR format: $\mathtt{ia}(1{:}I{+}1)$, $\mathtt{ja}(1{:}nnz)$, and $\mathtt{aa}(1{:}nnz)$.

Array ia. The integer array $\mathtt{ia}$ stores the number of nonzero entries in each row. More precisely, conventionally setting $\mathtt{ia}(1) := 1$, the value of $\mathtt{ia}(i+1)$ is defined such that $\mathtt{ia}(i+1) - \mathtt{ia}(i)$ is equal to the number of nonzero entries in the i-th row of the matrix $\mathcal{A}$, $i \in \{1{:}I\}$. Note that $nnz = \mathtt{ia}(I+1) - \mathtt{ia}(1)$.

Array ja. The integer array $\mathtt{ja}$ gives the column indices of the nonzero entries. More precisely, for all $i \in \{1{:}I\}$, the list $(\mathtt{ja}(p))_{p \in \{\mathtt{ia}(i){:}\,\mathtt{ia}(i+1)-1\}}$ contains the column indices of the nonzero entries in row i. A usual convention is to store the column indices in $\mathtt{ja}$ in increasing order for every row.

Array aa. The array $\mathtt{aa}$ contains the nonzero entries of the matrix. For every row i, the list $(\mathtt{aa}(p))_{p \in \{\mathtt{ia}(i){:}\,\mathtt{ia}(i+1)-1\}}$ contains all the nonzero entries of the row i. The same ordering is used for $\mathtt{ja}$ and $\mathtt{aa}$, so that $\mathtt{aa}(p) := \mathcal{A}_{i,\mathtt{ja}(p)}$.

Example 29.3. The CSR arrays for the following 5×5 matrix:

$$
\mathcal{A} := \begin{bmatrix}
1. & 0. & 0. & 2. & 0. \\
3. & 4. & 0. & 5. & 0. \\
6. & 0. & 7. & 8. & 9. \\
0. & 0. & 10. & 11. & 0. \\
0. & 0. & 0. & 0. & 12.
\end{bmatrix}
\tag{29.4}
$$

are $\mathtt{aa} = [1.\ 2.\ 3.\ 4.\ 5.\ 6.\ 7.\ 8.\ 9.\ 10.\ 11.\ 12.]$, $\mathtt{ja} = [1\ 4\ 1\ 2\ 4\ 1\ 3\ 4\ 5\ 3\ 4\ 5]$, and $\mathtt{ia} = [1\ 3\ 6\ 10\ 12\ 13]$. Note that $nnz = \mathtt{ia}(6) - \mathtt{ia}(1) = 12$.   □

Matrix-vector multiplication. Matrix-vector multiplication is an operation that is invoked repeatedly in iterative solution methods (see §28.3.2). Algorithm 29.1 shows how to perform the matrix-vector multiplication with the CSR format. The technique is optimal in the sense that it involves only the number of operations that are necessary.

29.2.2   Ellpack format

The CSR format has some drawbacks since the rows of the compressed matrix are not of constant length. Moreover, the fact that the index of the first nonzero entry has to be computed for each row may hamper vectorization. The purpose of the *Ellpack format* is to solve these difficulties. This format, which is based on the hypothesis that each row of the matrix contains almost the same number of nonzero entries, is well-adapted for meshes that are almost structured. Let $\mathcal{A}$ be

Algorithm 29.1 Matrix-vector multiplication in CSR format.

for $i \in \{1{:}I\}$ **do**; $\; yi := 0$
  for $p \in \{ia(i){:}\, \mathtt{ia}(i+1)-1\}$ **do**
    $yi := yi + \mathtt{aa}(p) * x(\mathtt{ja}(p))$
  end for
  $y(i) := yi$
end for

a matrix of size $I{\times}I'$ and let $n_{\mathrm{row}}(i)$ be the number of nonzero entries in the i-th row of $\mathcal{A}$ for all $i \in \{1{:}I\}$. Let $N_{\mathrm{row}} := \max_{i\in\{1{:}I\}} n_{\mathrm{row}}(i)$ be the maximum number of nonzero entries per row in $\mathcal{A}$. For instance, for continuous $\mathbb{Q}_1$ finite elements, $N_{\mathrm{row}} = 9$ in dimension two and $N_{\mathrm{row}} = 27$ in dimension three. The storage is done with two arrays $\mathtt{aa}(1{:}I, 1{:}N_{\mathrm{row}})$ and $\mathtt{ja}(1{:}I, 1{:}N_{\mathrm{row}})$ as follows:

Array aa. The array $\mathtt{aa}$ contains the nonzero entries of $\mathcal{A}$. For every row i, $\mathtt{aa}(i, 1{:}n_{\mathrm{row}}(i))$ contains all the nonzero entries in row i, and if $n_{\mathrm{row}}(i) < N_{\mathrm{row}}$, the entries $\mathtt{aa}(i, (n_{\mathrm{row}}(i)+1){:}N_{\mathrm{row}})$ are set to zero by convention.

Array ja. The integer array $\mathtt{ja}$ contains the column indices of the nonzero entries in the matrix $\mathcal{A}$. For every row i, $\mathtt{ja}(i, 1{:}n_{\mathrm{row}}(i))$ contains all the column indices of the nonzero entries in row i. The ordering of the indices in $\mathtt{ja}$ is the same as that in $\mathtt{aa}$. The simplest convention consists of ordering the column indices in increasing order. If $n_{\mathrm{row}}(i) < N_{\mathrm{row}}$, the entries $\mathtt{ja}(i, (n_{\mathrm{row}}(i)+1){:}N_{\mathrm{row}})$ are given an arbitrary value, say $\mathtt{ja}(i, n_{\mathrm{row}}(i))$.

29.2.3 Assembling

Let us see how the formula (29.1) can be implemented to evaluate the entries of the stiffness matrix when it is stored in some compressed format, e.g., the CSR format. Let $\mathtt{j\_dof}(1{:}N_{\mathrm{c}}, 1{:}n_{\mathrm{sh}})$ be the double-entry connectivity array introduced in Chapter 19. Recall that this array is defined such that

$$\varphi_{\mathtt{j\_dof}(m,n)|K_m} = \theta_{K_m,n}, \tag{29.5}$$

for every integers $n \in \{1{:}n_{\mathrm{sh}}\}$ and $m \in \{1{:}N_{\mathrm{c}}\}$. The assembling of the matrix $\mathcal{A}$ stored in the CSR format is described in Algorithm 29.2. The temporary array $\mathtt{tmp}$ in each mesh cell stores the local stiffness matrix. We will see in §30.3 how to compute the entries of this array by means of quadratures.

29.3 Reordering

Reordering a square matrix $\mathcal{A}$ means replacing $\mathcal{A}$ by the matrix $\mathcal{B} := \mathcal{P}\mathcal{A}\mathcal{P}^{\mathsf{T}}$, where the permutation $\mathcal{P}$ has entries $\mathcal{P}_{ij} := \delta_{\sigma(i)j}$ and σ is a permutation of the set $\{1{:}I\}$, so that $\mathcal{B}_{ij} = \mathcal{A}_{\sigma(i)\sigma(j)}$. Since $\mathcal{P}^{\mathsf{T}} = \mathcal{P}^{-1}$, the Euclidean condition number of a matrix is invariant by reordering. The goal of reordering techniques modify the sparsity pattern of the matrix by clustering nonzero entries as close as possible to the diagonal to reduce the bandwidth. We present in this section various reordering techniques based on the concept of adjacency graph.

Example 29.4 ($8{\times}8$ matrix). To illustrate how reordering can affect the fill-in resulting from the LU factorization (see §28.3.1), let us consider the $8{\times}8$ matrix $\mathcal{A}$ whose sparsity pattern is shown

Algorithm 29.2 Matrix assembling in CSR format.

```
aa := 0
for m ∈ {1:N_c} do
  for ni ∈ {1:n_sh} do
    for nj ∈ {1:n_sh} do
      tmp(ni, nj) := ∫_{K_m} A_{K_m}(x, θ_{K_m,nj}, θ_{K_m,ni}) dx
    end for
  end for
  for ni ∈ {1:n_sh} do;  i := j_dof(m, ni)
    for nj ∈ {1:n_sh} do;  j := j_dof(m, nj)
      for p ∈ {ia(i): ia(i+1)−1} do
        if ja(p) := j then
          aa(p) := aa(p) + tmp(ni, nj); Exit loop on p
        end if
      end for
    end for
  end for
end for
```

in the left panel of Figure 29.3. One can verify that the LU factorization of $\mathcal{A}$ (without pivoting) results in complete fill-in, i.e., the lower and upper triangular matrices $\mathcal{T}_{\mathrm{L}}$ and $\mathcal{T}_{\mathrm{U}}$ in (28.21) (such that $\mathcal{A} = \mathcal{T}_{\mathrm{L}}\mathcal{T}_{\mathrm{U}}$) are filled. Let us now consider the permutation $\sigma : (1, 2, 3, 4, 5, 6, 7, 8) \mapsto (8, 7, 6, 5, 4, 3, 2, 1)$. Let $\mathcal{B}$ be the 8×8 matrix such that $\mathcal{B}_{ij} := \mathcal{A}_{\sigma(i)\sigma(j)}$ for all $i, j \in \{1:8\}$. The sparsity pattern of $\mathcal{B}$ is shown in the right panel of Figure 29.3. It is straightforward to check that no fill-in occurs when computing the LU factorization of $\mathcal{B}$. This simple example illustrates that significant savings in memory and computational time can be achieved by enumerating properly the degrees of freedom (dofs) in a finite element code. □
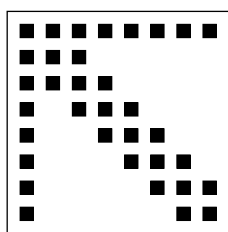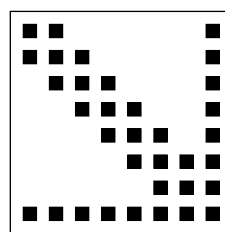
Initial matrix $\mathcal{A}$ Reordered matrix $\mathcal{B}$

Figure 29.3: Two different orderings for a sparse matrix.

Remark 29.5 (Literature). We refer the reader to Saad [339] for reordering techniques applied to iterative solvers and to George and Liu [214], George et al. [215], Davis [156, Chap. 7] for direct solution methods. □

29.3.1 Adjacency graph

As illustrated in Example 29.4, it is often important to reorder the unknowns and the equations before solving a linear system. Of course, the reordering technique to be used depends on the

strategy chosen to solve the linear system (direct, iterative, parallel, etc.). Choosing optimal reordering strategies is a difficult branch of graph theory.

To better understand the enumeration issue, it is convenient to introduce the notion of adjacency graph. Let V be a set, let $\Re$ be a binary relation on V, and denote $E := \{(x, y) \in V \times V \mid x \Re y\}$. The pair $G := (V, E)$ is called *graph*. The elements of V are called *graph vertices* or nodes and the members of E are called *graph edges*. We say that G is an *undirected graph* if $\Re$ is symmetric. A vertex y is said to be adjacent to x if $(x, y) \in E$. For a subset $X \subset V$, the *adjacent set* of X is defined as $\mathrm{Adj}(X) := \{y \in V \backslash X \mid \exists x \in X, (x, y) \in E\}$. The set $\mathrm{Adj}(x)$, defined as $\mathrm{Adj}(\{x\})$, is called *neighborhood* of x. The cardinality of $\mathrm{Adj}(x)$ is called *degree* of x. A common way of representing graphs is to associate with each vertex in V a point in the plane and to draw a directed line between two points (possibly identical) whenever their associated vertices are in E.

Let $\mathcal{A}$ be a $I \times I$ matrix. The *adjacency graph* of $\mathcal{A}$ is the pair (V, E) where $V := \{1{:}I\}$ and $E := \{(i, j) \in V \times V \mid \mathcal{A}_{ij} \neq 0\}$. Thus, we have

$$\mathrm{Adj}(i) = \{j \in \{1{:}I\} \backslash \{i\} \mid \mathcal{A}_{ij} \neq 0\}, \qquad \forall i \in \{1{:}I\}. \tag{29.6}$$

We say that (E, V) is the *undirected adjacency graph* of $\mathcal{A}$ when $(i, j) \in E$ iff $\mathcal{A}_{ij} \neq 0$ or $\mathcal{A}_{ji} \neq 0$. Figure 29.4 shows the adjacency graph of a 8×8 sparse matrix. A circle around a number means that the corresponding diagonal entry in the matrix is not zero.
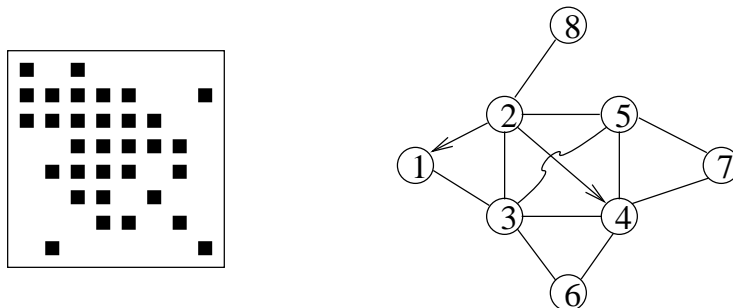
Figure 29.4: Sparsity pattern (left) and adjacency graph (right) of a 8×8 sparse matrix.

29.3.2   Level-set ordering

Assume that V is finite, let $G := (V, E)$ be a graph, and let $x \in V$ be a vertex. The elements of an indexed collection of disjoint subsets of V, say $L_1, L_2, L_3, \ldots$, are said to be *level sets* associated with x if $L_1 := \{x\}$ and $L_{k+1} := \mathrm{Adj}(L_k) \backslash (\bigcup_{l \in \{1:k\}} L_l)$ for all $k \geq 1$. L_k is said to be the k-th level set. The list $L_1, L_2, L_3, \ldots$ is finite since V is finite. Moreover, $L_1, L_2, L_3, \ldots$ forms a partition of V if G is a *strongly connected graph*, i.e., if there exists a path from each vertex to every other vertex.

For every vertex y in V, we define the distance from x to y, $d_x(y)$, as follows: If there is k such that $y \in L_k$, then $d_x(y) := k - 1$. Otherwise, $d_x(y) := \infty$ (note that $d_x(y) < \infty$ if the graph is strongly connected). In general, $d_x(y) \neq d_y(x)$ unless the graph is undirected (think of $V := \{x, y\}$ and $E := \{(x, y)\}$ so that $d_x(y) = 1$ and $d_y(x) = \infty$).

Algorithm 29.3 shows a possible way to evaluate the level sets associated with a vertex i_1 in the adjacency graph of a $I \times I$ sparse matrix $\mathcal{A}$. The integer `max_levelset` is the number of level sets associated with the vertex i_1. The level sets are deduced from the arrays `perm` and `stride` as

Algorithm 29.3 Evaluation of the level sets of i_1.

Input: i_1
Output: perm, max_levelset, stride
$k := 2$; count := 1; virgin(1:I) :=.true.
perm(1) := i_1; stride(1) := 1; stride(2) := 2
loop
  nb_vert_in_Lk := 0
  for $l \in \{$stride$(k-1)$:stride$(k)-1\}$ **do**
    for all $j \in \mathrm{Adj}($perm$(l))$ **do**
      if (virgin(j)) **then**
        virgin(j) := .false.; nb_vert_in_Lk = nb_vert_in_Lk + 1
        count := count + 1; perm(count) := j
      end if
    end for
  end for
  if (nb_vert_in_Lk = 0) **then**
    max_levelset := $k - 1$; exit **loop**
  end if
  stride$(k + 1)$:= stride(k) + nb_vert_in_Lk; $k := k + 1$
end loop
if count $\neq I$ **then** G is not strongly connected

follows:

$$\underbrace{\{\texttt{perm}(1)\}}_{=:L_1}, \dots, \underbrace{\{\texttt{perm}(\texttt{stride}(k)), \dots, \texttt{perm}(\texttt{stride}(k+1)-1)\}}_{=:L_k}, \dots.$$

If there is a vertex i_2 which is not in any of the level sets associated with i_1, i.e., i_1 is not connected to i_2, then the level sets associated with i_2 are constructed by using Algorithm 29.3 again. The process is repeated until the union of all the level sets forms a partition of V. At the end all the permutation arrays are collected in a single array still denoted by perm.

Example 29.6. Let us consider the undirected graph shown in Figure 29.5 to illustrate the level set concept. The level sets associated with vertex 2 are

$$L_1 = \{2\}, \quad L_2 = \{5, 7\}, \quad L_3 = \{9, 11, 14\}, \quad L_4 = \{1, 3, 12, 15\},$$
$$L_5 = \{8, 10, 13\}, \quad L_6 = \{4, 6\}.$$

Hence, max_levelset = 6, stride = $(1, 2, 4, 7, 11, 14, 16)$, and a possible choice for perm is perm = $(2, 5, 7, 9, 11, 14, 1, 3, 12, 15, 8, 10, 13, 4, 6)$. □

The simplest reordering for $\mathcal{A}$ consists of setting $\mathcal{B}_{ij} := \mathcal{A}_{\texttt{perm}(i)\texttt{perm}(j)}$. This technique is known as the *breadth-first-search* (BFS) reordering. One interest of this reordering is the following result.

Proposition 29.7. *Assume G is undirected and* max_levelset ≥ 3. *The array* stride *defines a tridiagonal block structure of $\mathcal{B}$, i.e.,*

$$\mathcal{B}_{ij} = 0 \quad \text{if} \quad \begin{cases} i \in \{\texttt{stride}(k):\texttt{stride}(k+1)-1\} & \text{i.e., } \texttt{perm}(i) \in L_k, \\ j \in \{\texttt{stride}(k'):\texttt{stride}(k'+1)-1\} & \text{i.e., } \texttt{perm}(j) \in L_{k'}, \\ |k - k'| \geq 2. \end{cases}$$
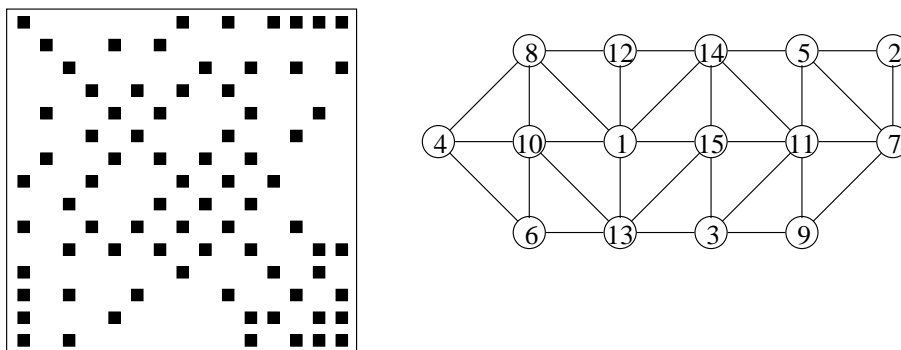
Figure 29.5: Sparsity pattern and adjacency graph.

Proof. The proof proceeds by contradiction. Assume that $\mathcal{B}_{ij} \neq 0$ and $|k - k'| \geq 2$ with $\texttt{perm}(i) \in L_k$ and $\texttt{perm}(j) \in L_{k'}$. Then $\mathcal{A}_{\texttt{perm}(i)\texttt{perm}(j)} \neq 0$. This means $\texttt{perm}(i) \in \text{Adj}(\texttt{perm}(j))$ and $\texttt{perm}(j) \in \text{Adj}(\texttt{perm}(i))$, since G is undirected and $i \neq j$. Assume further that $k' \geq k$. Then $\texttt{perm}(j) \notin \bigcup_{l \in \{1:k\}} L_l$ since the level sets are disjoint. Moreover, $\texttt{perm}(j) \in \text{Adj}(\texttt{perm}(i))$ and $\texttt{perm}(i) \in L_k$ means $\texttt{perm}(j) \in \text{Adj}(L_k)$. Combining the above two statements yields $\texttt{perm}(j) \in L_{k+1} = \text{Adj}(L_k)\backslash(\bigcup_{l \in \{1:k\}} L_l)$. This means $k' = k + 1$, which contradicts $|k - k'| \geq 2$. The argument applies also if $k' \leq k$ since the graph is undirected.   □

Proposition 29.7 shows that choosing level sets with $\texttt{max\_levelset}$ as large as possible minimizes the bandwidth of $\mathcal{B}$. This can be achieved by picking the initial vertex i_1 such that $\max_{y \in V} d_{i_1}(y)$ is maximal.

The ordering depends on the way the vertices are traversed in each level set. In the BFS reordering, the vertices are traversed in the natural order. Another strategy consists of ordering the vertices in each level set by increasing degree. This ordering technique is known as the *Cuthill–McKee* (CMK) ordering. Another popular strategy consists of reversing the CMK ordering. It has been observed that the reversing strategy yields a better scheme for sparse Gaussian elimination. We refer the reader to George and Liu [214], George et al. [215] for further insight into these techniques and their many generalizations.

Example 29.8 (CMK reordering). Figure 29.6 shows the adjacency graph and the sparsity pattern of the CMK-reordered matrix corresponding to the matrix shown in Figure 29.5. The reordering has been done by using the level sets associated with vertex 2. In each level set, the nodes are ordered by increasing degree. The permutation array is

$$\texttt{perm} = (2, 5, 7, 9, 14, 11, 12, 3, 15, 1, 8, 10, 13, 4, 6).$$

The reordered matrix has a tridiagonal block structure, and the size of the k-th block is $\texttt{stride}(k+1) - \texttt{stride}(k)$ with the array $\texttt{stride}$ evaluated in Example 29.6.   □

29.3.3   Independent set ordering (ISO)

The aim of ISO techniques is to find a permutation of the vertices such that the reordered matrix has the following 2×2 block structure:

$$\mathcal{B} = \left[\begin{array}{c|c} \mathcal{D} & \mathcal{E} \\ \hline \mathcal{F} & \mathcal{H} \end{array} \right],$$
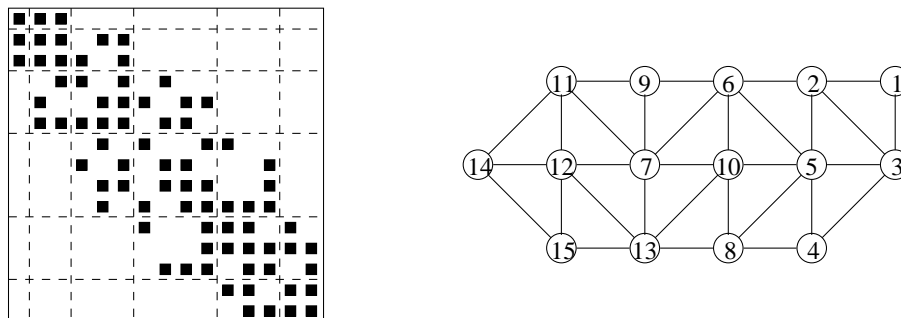
Figure 29.6: Sparsity pattern and adjacency graph after using the CMK reordering for the matrix shown in Figure 29.5.

Algorithm 29.4 Independent set ordering.

$S := \emptyset$; virgin :=.true.
for $i \in \{1{:}I\}$ **do**
  $j := \texttt{traverse}(i)$
  if (virgin(j)) **then**
    $S := S \cup \{j\}$; virgin(j) :=.false.
    for all $k \in \text{Adj}(j)$ **do**
      virgin(k) :=.false.
    end for
  end if
end for

---

where $\mathcal{D}$ is diagonal and as large as possible. To this purpose, we introduce the notion of *independent set*. Let $G := (V, E)$ be a graph. $S \subset V$ is said to be an independent set if for all $x \neq y \in S$, the edge (x, y) is not in E. An independent set is said to be *maximal* if it is maximal with respect to the inclusion order.

Assume that G is the adjacency graph of a square matrix $\mathcal{A}$. Let S be an independent set. Let perm be any permutation array of $\{1{:}I\}$ such that $S = \{\texttt{perm}(1), \ldots, \texttt{perm}(\text{card}(S))\}$. Define the reordered matrix $\mathcal{B}$ such that $\mathcal{B}_{ij} = \mathcal{A}_{\texttt{perm}(i)\texttt{perm}(j)}$ for all $i, j \in \{1{:}I\}$. We readily infer the following result.

Proposition 29.9. *The triple $(1, \text{card}(S), I)$ defines a $2{\times}2$ block structure of $\mathcal{B}$ where the top left block is diagonal.*

Let traverse be a permutation array of $\{1{:}I\}$. Algorithm 29.4 presents a simple strategy to construct an independent set. A possible choice for traverse consists of setting $\texttt{traverse}(i) := i$, but in general traverse is set to maximize the cardinality of S. Since $\text{card}(S)$ is equal to the number of times the statement (virgin(j)) is true in Algorithm 29.4, a possible technique to maximize this number is to choose j s.t. $\text{card}(\text{Adj}(j))$ is small, i.e., among all the nodes left, j must be one of those having the lowest degree. A simple strategy consists of sorting the nodes in increasing degree in traverse.

29.3.4 Multicolor ordering

A third standard reordering method uses *graph coloring*. Assume that G is an undirected and strongly connected graph. Then the map $C : V \to \mathbb{N}$ is said to be a *graph coloring* if $C(x) \neq C(y)$ for all $(x, y) \in E$ s.t. $x \neq y$. For $x \in V$, $C(x)$ is referred to as the color of x. The goal of graph coloring is to find a map C s.t. the cardinality of the range of C is as small as possible, i.e., the number of colors to color V is as small as possible.

In the context of linear algebra, optimality is not a major issue and one is usually satisfied by using simple heuristics. For instance, given a permutation array `traverse` of $\{1:I\}$, Algorithm 29.5 describes a basic coloring strategy. The simplest choice consists of setting $\texttt{traverse}(i) := i$, but more sophisticated choices are possible. For instance, it can be shown that if the graph can be colored with two colors only and if BFS is used to initialize `traverse`, then Algorithm 29.5 finds a two-color partitioning; see Exercise 29.5. Independently of `traverse`, the number of colors found by Algorithm 29.5 is at most equal to 1 plus the largest degree in the graph; see Exercise 29.5.

Let G be the undirected adjacency graph of a matrix $\mathcal{A}$. Assume that we have colored G. Denote by `k_max` the number of colors that are used, and let $C : V \to \{1:\texttt{k\_max}\}$ be the corresponding color mapping. Let $\texttt{col\_part}(1 : \texttt{k\_max})$ be the array such that $\texttt{col\_part}(1) := 1$ and $\texttt{col\_part}(k+1) := \texttt{col\_part}(k) + \text{card}(C^{-1}(k))$ for all $k \in \{1:\texttt{k\_max}\}$. Let `perm` be any permutation array s.t. the color of the vertices in the set $\{\texttt{perm}(\texttt{col\_part}(k)), \ldots, \texttt{perm}(\texttt{col\_part}(k + 1) - 1)\}$ is k. Define the reordered matrix $\mathcal{B}$ such that $\mathcal{B}_{ij} = \mathcal{A}_{\texttt{perm}(i)\texttt{perm}(j)}$. Multicolor ordering partially finds its justification in the following result.

Proposition 29.10. *The array* `col_part` *defines a* `k_max`×`k_max` *block structure of* $\mathcal{B}$ *where the diagonal blocks are diagonal.*

Proof. Left as an exercise; see also Adams and Jordan [5]. $\qquad\qquad\square$

Exercises

Exercise 29.1 (Retrieving a nonzero entry in CSR format). Write an algorithm to retrieve the value $\mathcal{A}_{ij}$ from the array `aa` stored in CSR format.

Exercise 29.2 (Ellpack (ELL)). Write the arrays needed to store the matrix from Example 29.3 in the Ellpack format. Write an algorithm that performs a matrix-vector multiplication in this format.

Exercise 29.3 (Coordinate format (COO)). Let $\mathcal{A}$ be a $I{\times}I$ sparse matrix. Consider the storage format where one stores the nonzero entries $\mathcal{A}_{ij}$ in the array $\texttt{aa}(1:nnz)$ and stores in the same order the row and columns indices in the integer arrays $\texttt{ia}(1:nnz)$ and $\texttt{ja}(1:nnz)$, respectively. (i) Use this format to store the matrix defined in (29.4). (ii) Write an algorithm to perform a matrix-vector product in this format. Compare with the CSR format.

Algorithm 29.5 Greedy coloring.

```
color := 0
for i ∈ {1:I} do
    j := traverse(i) {Since G is strongly connected Adj(j) cannot be empty.}
    color(j) := min{k > 0 | k ∉ color(Adj(j))}
end for
```

Exercise 29.4 (Storage). Consider the storage format for sparse $I \times I$ matrices where one stores the nonzero entries $\mathcal{A}_{ij}$ in the array aa(1:nnz) and stores in the same order the integer $(i-1)I + j$ in the integer array ja(1:nnz). (i) Use this format for the matrix defined in (29.4). (ii) Write an algorithm to do matrix-vector products in this format. Compare with the CSR format.

Exercise 29.5 (Greedy coloring). (i) Prove that the total number of colors found by Algorithm 29.5 is at most equal to 1 plus the largest degree in the graph. (ii) Assume that a graph G can be colored with two colors only. Prove that if the BFS reordering is used to initialize traverse, then Algorithm 29.5 finds a two-color partitioning. (*Hint*: by induction on the number of level sets.)

Exercise 29.6 (Multicolor ordering). Prove Proposition 29.10.

Exercise 29.7 (CMK reordering). Give the sparsity pattern and the CMK reordering for the matrix shown in Figure 29.4.

Chapter 30

Quadratures

Implementing the finite element method requires evaluating the entries of the stiffness matrix and the right-hand side vector, which in turn requires computing integrals over the cells and (possibly) the faces of the mesh. In practice, these integrals must often be evaluated approximately by means of quadratures. In this chapter, we review multidimensional quadratures that are frequently used in finite element codes, and we derive bounds on the quadrature error. We also describe the implementation of quadratures in conjunction with the assembling of the stiffness matrix. Recall that one-dimensional quadratures are presented in Chapter 6.

30.1   Definition and examples

Let D be a Lipschitz polyhedron in $\mathbb{R}^d$, $d \in \{2, 3\}$, let $\mathcal{T}_h$ be a mesh of D that covers D exactly, and let $\phi : D \to \mathbb{R}$ be a smooth function. Suppose that we want to evaluate the integral $\int_D \phi(\boldsymbol{x}) \, \mathrm{d}x$. Since

$$\int_D \phi(\boldsymbol{x}) \, \mathrm{d}x = \sum_{K \in \mathcal{T}_h} \int_K \phi(\boldsymbol{x}) \, \mathrm{d}x,$$

this problem reduces to evaluating the integral $\int_K \phi(\boldsymbol{x}) \, \mathrm{d}x$ over each mesh cell $K \in \mathcal{T}_h$. Since computing integrals exactly is often impossible, one needs to use quadratures to approximate $\int_K \phi(\boldsymbol{x}) \, \mathrm{d}x$.

Definition 30.1 (Quadrature nodes and weights). *Let K be a compact, connected, Lipschitz subset of $\mathbb{R}^d$ with nonempty interior. Let $l_\mathcal{Q} \geq 1$ be an integer. A quadrature in K with $l_\mathcal{Q}$ nodes is specified through a set of $l_\mathcal{Q}$ points $\{\boldsymbol{\xi}_l\}_{l \in \{1:l_\mathcal{Q}\}}$ in K, called* quadrature nodes *or* Gauss nodes *and a set of $l_\mathcal{Q}$ real numbers $\{\omega_l\}_{l \in \{1:l_\mathcal{Q}\}}$, called* quadrature weights. *The quadrature consists of the approximation*

$$\int_K \phi(\boldsymbol{x}) \, \mathrm{d}x \approx \sum_{l \in \{1:l_\mathcal{Q}\}} \omega_l \, \phi(\boldsymbol{\xi}_l). \tag{30.1}$$

The largest integer k such that (30.1) is an equality for every polynomial in $\mathbb{P}_{k,d}$ is called quadrature order *and is denoted by $k_\mathcal{Q}$.*

Given a quadrature on the reference element $\widehat{K}$ and a mesh $\mathcal{T}_h$, a quadrature on every cell $K \in \mathcal{T}_h$ can be generated by using the geometric mapping $\boldsymbol{T}_K : \widehat{K} \to K$. Let $\mathbb{J}_K$ denote the Jacobian matrix of $\boldsymbol{T}_K$.

Proposition 30.2 (Quadrature generation). *Consider a quadrature in $\widehat{K}$ with nodes $\{\widehat{\boldsymbol{\xi}}_l\}_{l\in\{1:l_Q\}}$ and weights $\{\widehat{\omega}_l\}_{l\in\{1:l_Q\}}$. Setting*

$$\boldsymbol{\xi}_{lK} := \boldsymbol{T}_K(\widehat{\boldsymbol{\xi}}_l) \qquad and \qquad \omega_{lK} := \widehat{\omega}_l |\det(\mathbb{J}_K(\widehat{\boldsymbol{\xi}}_l))|, \tag{30.2}$$

for all $l \in \{1{:}l_Q\}$, generates a quadrature on K. If the quadrature on $\widehat{K}$ is of order k_Q and the geometric mapping $\boldsymbol{T}_K$ is affine, then the quadrature on K is also of order k_Q.

Proof. Since $\boldsymbol{T}_K$ is a C^1-diffeomorphism, the change of variables $\boldsymbol{x} = \boldsymbol{T}_K(\widehat{\boldsymbol{x}})$ yields $\int_K \phi(\boldsymbol{x})\,\mathrm{d}x = \int_{\widehat{K}} \phi\big(\boldsymbol{T}_K(\widehat{\boldsymbol{x}})\big)|\det(\mathbb{J}_K(\widehat{\boldsymbol{x}}))|\,\mathrm{d}\widehat{x}$, and we can apply the quadrature over $\widehat{K}$ to the right-hand side. The statement on the quadrature order is immediate to verify. Indeed, if $\boldsymbol{T}_K$ is affine, $\mathbb{J}_K$ is constant and $\phi \circ \boldsymbol{T}_K$ is in $\mathbb{P}_{k,d}$ iff $\phi \in \mathbb{P}_{k,d}$. $\qquad\square$

Remark 30.3 (Surface quadrature). When generating a surface quadrature from a quadrature on a reference surface, Lemma 9.12 must be used to account for the transformation of the surface measure; see Exercise 30.5. $\qquad\square$

Example 30.4 (Literature). The literature on quadratures is abundant; see Abramowitz and Stegun [3, Chap. 25], Hammer and Stroud [237], Stroud [360], Davis and Rabinowitz [155], Brass and Petras [85]. We refer the reader to §6.2 for a review of one-dimensional quadratures using the Gauss–Legendre, Gauss–Lobatto, and Gauss–Radau nodes. $\qquad\square$

Example 30.5 (Cuboids). Quadratures on cuboids can be deduced from one-dimensional quadratures by taking the Gauss nodes in tensor-product form. Note though that tensor-product formulas are not optimal in the sense of using the fewest function evaluations for a given order. Although no general formula for non-tensor-product quadratures for the cube is known, many quasi-optimal quadratures are available in the literature; see, e.g., Cools [139, §2.3.1] and Cools and Rabinowitz [140, §4.1]. $\qquad\square$

Example 30.6 (Quadratures on the triangle). Table 30.1 lists some quadratures on the triangle (see, e.g., Dunavant [178]). In this table, we call multiplicity the number of permutations to be performed on the barycentric coordinates to obtain the list of all the Gauss nodes of the quadrature. For instance, the first-order formula in the second line has three Gauss nodes with barycentric coordinates and weights $\{1, 0, 0; \frac{1}{3}S\}$, $\{0, 1, 0; \frac{1}{3}S\}$, $\{0, 0, 1; \frac{1}{3}S\}$, where S denotes the surface of the triangle. $\qquad\square$

Example 30.7 (Quadratures on the tetrahedron). Table 30.2 lists some quadratures on the tetrahedron (see, e.g., Keast [265]). As above, the multiplicity is the number of permutations to perform on the barycentric coordinates to obtain all the Gauss nodes of the quadrature. For instance, the third-order formula has five Gauss nodes which are the node $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ with the weight $-\frac{4}{5}V$ and the four nodes $(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{2})$, $(\frac{1}{6}, \frac{1}{6}, \frac{1}{2}, \frac{1}{6})$, $(\frac{1}{6}, \frac{1}{2}, \frac{1}{6}, \frac{1}{6})$, $(\frac{1}{2}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$ with the weight $\frac{9}{20}V$, where V denotes the volume of the tetrahedron. $\qquad\square$

Example 30.8 (Integral of barycentric coordinates). Let $\{\lambda_i\}_{i\in\{0:d\}}$ be the barycentric coordinates in a simplex K in $\mathbb{R}^d$. We have

$$\int_K \lambda_0^{\alpha_0} \dots \lambda_d^{\alpha_d}\,\mathrm{d}x = |K|\frac{\alpha_0!\dots\alpha_d!d!}{(\alpha_0 + \dots + \alpha_d + d)!}, \tag{30.3}$$

for every natural numbers $\alpha_0, \dots, \alpha_d$. This formula is useful to verify numerically the order of a quadrature. $\qquad\square$

$k_\mathcal{Q}$	$l_\mathcal{Q}$	Barycentric coord.	Multiplicity	Weights ω_l
1	1	$\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)$	1	S
1	3	$(1,0,0)$	3	$\frac{1}{3}S$
2	3	$\left(\frac{1}{6},\frac{1}{6},\frac{2}{3}\right)$	3	$\frac{1}{3}S$
2	3	$\left(\frac{1}{2},\frac{1}{2},0\right)$	3	$\frac{1}{3}S$
3	4	$\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)$ $\left(\frac{1}{5},\frac{1}{5},\frac{3}{5}\right)$	1 3	$-\frac{9}{16}S$ $\frac{25}{48}S$
3	7	$\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)$ $\left(\frac{1}{2},\frac{1}{2},0\right)$ $(1,0,0)$	1 3 3	$\frac{9}{20}S$ $\frac{2}{15}S$ $\frac{1}{20}S$
4	6	$(a_i,a_i,1-2a_i)$ for $i=1,2$ $a_1 = 0.445948490915965$ $a_2 = 0.091576213509771$	3	ω_i for $i=1,2$ $\omega_1 = S{\times}0.223381589678010$ $\omega_2 = S{\times}0.109951743655322$
5	7	$\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)$ $(a_i,a_i,1-2a_i)$ for $i=1,2$ $a_1 = \frac{6-\sqrt{15}}{21}$ $a_2 = \frac{6+\sqrt{15}}{21}$	1 3	$\frac{9}{40}S$ $\frac{155-\sqrt{15}}{1200}S$ $\frac{155+\sqrt{15}}{1200}S$
6	12	$(a_i,a_i,1-2a_i)$ for $i=1,2$ $a_1 = 0.063089014491502$ $a_2 = 0.249286745170910$ $(a,b,1-a-b)$ $a = 0.310352451033785$ $b = 0.053145049844816$	3 6	 $S{\times}0.050844906370206$ $S{\times}0.116786275726378$ $S{\times}0.082851075618374$

Table 30.1: Nodes and weights for quadratures on a triangle of area S.

30.2 Quadrature error

Let $(\mathcal{T}_h)_{h\in\mathcal{H}}$ be a shape-regular sequence of affine meshes. Proposition 30.2 allows us to generate a quadrature in each mesh cell $K \in \mathcal{T}_h$ from a reference quadrature by using the geometric mapping $\boldsymbol{T}_K : \widehat{K} \to K$. Let $\{\boldsymbol{\xi}_{lK}\}_{l\in\{1:l_\mathcal{Q}\}}$ and $\{\omega_{lK}\}_{l\in\{1:l_\mathcal{Q}\}}$ be the nodes and weights of the quadrature on K thus obtained. Let $k_\mathcal{Q} \geq 0$ be the order of the quadrature. For every function ϕ that is smooth enough to have point values, say $\phi \in C^0(K)$, we define the quadrature error in the mesh cell K as follows:

$$E_K(\phi) := \int_K \phi(\boldsymbol{x})\,\mathrm{d}x - \sum_{l\in\{1:l_\mathcal{Q}\}} \omega_{lK}\,\phi(\boldsymbol{\xi}_{lK}). \tag{30.4}$$

Lemma 30.9 (Quadrature error). *Let $p \in [1,\infty]$ and let $m \in \mathbb{N}$ be such that $m > \frac{d}{p}$. Assume*

$k_{\mathcal{Q}}$	$l_{\mathcal{Q}}$	Barycentric coord.	Multiplicity	Weights ω_l
1	1	$\left(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\right)$	1	V
1	4	$(1,0,0,0)$	4	$\frac{1}{4}V$
2	4	$(a,a,a,1-3a)$ $a=\frac{5-\sqrt{5}}{20}$	4	$\frac{1}{4}V$
2	10	$\left(\frac{1}{2},\frac{1}{2},0,0\right)$ $(1,0,0,0)$	6 4	$\frac{1}{5}V$ $-\frac{1}{20}V$
3	5	$\left(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\right)$ $\left(\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{2}\right)$	1 4	$-\frac{4}{5}V$ $\frac{9}{20}V$
5	15	$\left(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\right)$ $(a_i,a_i,a_i,1-3a_i)$ for $i=1,2$ $a_1=\frac{7-\sqrt{15}}{34}$ $a_2=\frac{7+\sqrt{15}}{34}$ $\left(a,a,\frac{1}{2}-a,\frac{1}{2}-a\right)$ $a=\frac{10-2\sqrt{15}}{40}$	1 4 6	$\frac{16}{135}V$ $\frac{2665+14\sqrt{15}}{37800}V$ $\frac{2665-14\sqrt{15}}{37800}V$ $\frac{10}{189}V$

Table 30.2: Nodes and weights for quadratures on a tetrahedron of volume V.

that $k_{\mathcal{Q}}+1\geq m$. There is c s.t. for all $\phi\in W^{m,p}(K)$, all $K\in\mathcal{T}_h$, and all $h\in\mathcal{H}$,

$$|E_K(\phi)|\leq c\,h_K^{m+d(1-\frac{1}{p})}|\phi|_{W^{m,p}(K)}. \tag{30.5}$$

Proof. Let $\phi\in W^{m,p}(K)$. Since $m>\frac{d}{p}$, the embedding $W^{m,p}(K)\hookrightarrow C^0(K)$ from Theorem 2.31 implies that the function ϕ is continuous. Moreover, since the meshes are affine, we have $E_K(\phi)=|\det(\mathbb{J}_K)|\,\widehat{E}(\widehat{\phi})$, where $\widehat{\phi}:=\phi\circ\boldsymbol{T}_K$ and $\widehat{E}(\widehat{\phi})$ is the quadrature error on $\widehat{K}$. By definition, $\widehat{E}:C^0(\widehat{K})\to\mathbb{R}$ is a bounded linear form, i.e., $|\widehat{E}(\widehat{\phi})|\leq c\|\widehat{\phi}\|_{C^0(\widehat{K})}$. Using the embedding $W^{m,p}(\widehat{K})\hookrightarrow C^0(\widehat{K})$, we infer that $|\widehat{E}(\widehat{\phi})|\leq c\|\widehat{\phi}\|_{W^{m,p}(\widehat{K})}$. Since $\widehat{E}(\widehat{p})=0$ for all $\widehat{p}$ in $\mathbb{P}_{m-1,d}\subset\mathbb{P}_{k_{\mathcal{Q}},d}$ (since $k_{\mathcal{Q}}+1\geq m$), we deduce from the Bramble–Hilbert/Deny–Lions lemma (more precisely Corollary 11.11 with $k:=m-1$) that $|\widehat{E}(\widehat{\phi})|\leq c|\widehat{\phi}|_{W^{m,p}(\widehat{K})}$. Since the geometric mapping is affine and the mesh sequence is shape-regular, we infer from (11.7a) in Lemma 11.7 that

$$|\widehat{\phi}|_{W^{m,p}(\widehat{K})}\leq c\,\|\mathbb{J}_K\|^m_{\ell^2(\mathbb{R}^d)}\,|\det(\mathbb{J}_K)|^{-\frac{1}{p}}|\phi|_{W^{m,p}(K)}.$$

We conclude by using (11.3), i.e., $\|\mathbb{J}_K\|_{\ell^2(\mathbb{R}^d)}\leq\frac{h_K}{\rho_{\widehat{K}}}$ and $|\det(\mathbb{J}_K)|\leq\frac{|K|}{|\widehat{K}|}$. $\square$

In the analysis of finite element methods with quadrature, it is useful to estimate the quadrature error $E_K(\phi p)$, where $p\in\mathbb{P}_{n,d}\circ\boldsymbol{T}_K^{-1}$ for some integer $n\geq 0$; see §33.3 for an application.

Lemma 30.10 (Quadrature error with polynomial factor). *Let $m\in\mathbb{N}$, let $n\in\mathbb{N}$.* (i) *Assume that $k_{\mathcal{Q}}\geq m+n-1$. There is c s.t.*

$$|E_K(\phi p)|\leq c\,h_K^m|\phi|_{W^{m,\infty}(K)}\|p\|_{L^1(K)}, \tag{30.6}$$

for all $\phi \in W^{m,\infty}(K)$, all $p \in \mathbb{P}_{n,d} \circ \boldsymbol{T}_K^{-1}$, all $K \in \mathcal{T}_h$, and all $h \in \mathcal{H}$. (ii) *Assume that* $n \geq 1$, $m \geq 1$, *and* $k_{\mathcal{Q}} \geq n + m - 2$. *There is* c *s.t.*

$$|E_K(\phi p)| \leq c\, h_K^m \Big(|\phi|_{W^{m,\infty}(K)} \|p\|_{L^1(K)} + |\phi|_{W^{m-1,\infty}(K)} \|\nabla p\|_{\boldsymbol{L}^1(K)} \Big), \tag{30.7}$$

for all $\phi \in W^{m,\infty}(K)$, all $p \in \mathbb{P}_{n,d} \circ \boldsymbol{T}_K^{-1}$, all $K \in \mathcal{T}_h$, and all $h \in \mathcal{H}$.

Proof. We only prove (30.6), and the reader is referred to Exercise 30.4 for the proof of (30.7). Let $\widehat{p} := p \circ \boldsymbol{T}_K \in \mathbb{P}_{n,d}$ and $\widehat{\phi} := \phi \circ \boldsymbol{T}_K$. After making a change of variable, we obtain $E_K(\phi p) = |\det(\mathbb{J}_K)| \widehat{E}(\widehat{\phi}\widehat{p})$ since $\mathcal{T}_h$ is affine (with obvious notation). Assuming first $m \geq 1$, we infer that $\widehat{E}(\widehat{g}\widehat{p}) = 0$ for all $\widehat{g} \in \mathbb{P}_{m-1,d}$ since $\widehat{p} \in \mathbb{P}_{n,d}$ and $k_{\mathcal{Q}} \geq n + m - 1$. Hence, $\widehat{E}(\widehat{\phi}\widehat{p}) = \widehat{E}((\widehat{\phi} - \widehat{g})\widehat{p})$ for all $\widehat{g} \in \mathbb{P}_{m-1,d}$. Therefore, we obtain

$$|E_K(\phi p)| = |\det(\mathbb{J}_K)| |\widehat{E}(\widehat{\phi}\widehat{p})| = |\det(\mathbb{J}_K)| \inf_{\widehat{g} \in \mathbb{P}_{m-1,d}} |\widehat{E}((\widehat{\phi} - \widehat{g})\widehat{p})|$$

$$\leq c\, |\det(\mathbb{J}_K)| \Big(\inf_{\widehat{g} \in \mathbb{P}_{m-1,d}} \|\widehat{\phi} - \widehat{g}\|_{C^0(\widehat{K})} \Big) \|\widehat{p}\|_{C^0(\widehat{K})}$$

$$\leq c\, |\det(\mathbb{J}_K)| \Big(\inf_{\widehat{g} \in \mathbb{P}_{m-1,d}} \|\widehat{\phi} - \widehat{g}\|_{W^{m,\infty}(\widehat{K})} \Big) \|\widehat{p}\|_{L^1(\widehat{K})},$$

where we used norm equivalence in $\mathbb{P}_{n,d}$ for $\widehat{p}$. Since

$$\inf_{\widehat{g} \in \mathbb{P}_{m-1,d}} \|\widehat{\phi} - \widehat{g}\|_{W^{m,\infty}(\widehat{K})} \leq c\, |\widehat{\phi}|_{W^{m,\infty}(\widehat{K})},$$

owing to the Bramble–Hilbert/Deny–Lions lemma (see Lemma 11.9), we infer that

$$|E_K(\phi p)| \leq c\, |\widehat{\phi}|_{W^{m,\infty}(\widehat{K})} |\det(\mathbb{J}_K)|\, \|\widehat{p}\|_{L^1(\widehat{K})}.$$

We conclude by using $|\det(\mathbb{J}_K)|\, \|\widehat{p}\|_{L^1(\widehat{K})} = \|p\|_{L^1(K)}$ and $|\widehat{\phi}|_{W^{m,\infty}(\widehat{K})} \leq c\|\mathbb{J}_K\|_{\ell^2}^m |\phi|_{W^{m,\infty}(K)}$ with $\|\mathbb{J}_K\|_{\ell^2} \leq c h_K$ (see Lemma 11.7). Finally, if $m = 0$, we have

$$|E_K(\phi p)| \leq c\|\widehat{\phi}\|_{L^\infty(\widehat{K})} |\det(\mathbb{J}_K)|\, \|\widehat{p}\|_{L^1(\widehat{K})},$$

and we conclude by using that $\|\widehat{\phi}\|_{L^\infty(\widehat{K})} = \|\phi\|_{L^\infty(K)}$. $\qquad\square$

30.3 Implementation

This section addresses practical implementation aspects of quadratures in the assembling modules of a finite element code.

30.3.1 Nodes and weights

Let $\{\widehat{\boldsymbol{\xi}}_l\}_{l \in \{1:l_{\mathcal{Q}}\}}$ and $\{\widehat{\omega}_l\}_{l \in \{1:l_{\mathcal{Q}}\}}$ be the quadrature nodes and weights on the reference element. Let $m \in \{1:N_c\}$ and let K_m be the corresponding mesh cell. The nodes and weights of the quadrature on K_m are defined in Proposition 30.2. Recall from Definition 8.1 that the geometric mapping $\boldsymbol{T}_{K_m}$ is built from the reference shape functions of the geometric element, $\{\widehat{\psi}_n\}_{n \in \{1:n_{\text{geo}}\}}$.

This leads us to define the two-dimensional array $\texttt{psi}(1{:}n_{\text{geo}}, 1{:}l_{\mathcal{Q}})$ that contains the values of the geometric shape functions at the quadrature nodes:

$$\texttt{psi}(n,l) := \widehat{\psi}_n(\widehat{\boldsymbol{\xi}}_l).$$

The k-th Cartesian component of the Gauss node $\boldsymbol{\xi}_{lK_m} := \boldsymbol{T}_{K_m}(\widehat{\boldsymbol{\xi}}_l)$ is then given by

$$(\boldsymbol{\xi}_{lK_m})_k = \sum_{n \in \{1:n_{\text{geo}}\}} \texttt{coord}(k, \texttt{j\_geo}(n,m))\, \texttt{psi}(n,l),$$

where the arrays $\texttt{coord}$ and $\texttt{j\_geo}$ are defined in §8.3. We also need the three-dimensional array $\texttt{dpsi\_dhatK}(1{:}d, 1{:}n_{\text{geo}}, 1{:}l_{\mathcal{Q}})$ providing the derivatives of the geometric shape functions at the Gauss nodes:

$$\texttt{dpsi\_dhatK}(k,n,l) := \frac{\partial \widehat{\psi}_n}{\partial \widehat{x}_k}(\widehat{\boldsymbol{\xi}}_l).$$

Then the entries of the Jacobian matrix $\mathbb{J}_{K_m}$ at $\widehat{\boldsymbol{\xi}}_l$ can be computed as follows:

$$\left(\mathbb{J}_{K_m}(\widehat{\boldsymbol{\xi}}_l)\right)_{k_1,k_2} = \sum_{n \in \{1:n_{\text{geo}}\}} \texttt{coord}(k_1, \texttt{j\_geo}(n,m))\, \texttt{dpsi\_dhatK}(k_2, n, l),$$

for all $k_1, k_2 \in \{1{:}d\}$. Since $\det(\mathbb{J}_{K_m}(\widehat{\boldsymbol{\xi}}_l))$ is always multiplied by the weight $\widehat{\omega}_l$ in the quadratures, it can be useful to store this product once and for all in the two-dimensional array of weights $\texttt{weight\_K}(1{:}l_{\mathcal{Q}}, 1{:}N_{\text{c}})$:

$$\texttt{weight\_K}(l,m) := \widehat{\omega}_l\, |\det(\mathbb{J}_{K_m}(\widehat{\boldsymbol{\xi}}_l))|.$$

Notice that when the mesh is affine, the partial derivatives of the shape functions $\widehat{\psi}_n$ are constant on $\widehat{K}$, so that the size of the array $\texttt{dpsi\_dhatK}$ can be reduced to $d \times n_{\text{geo}}$. Further memory space can be saved by storing separately the reference quadrature weights $\widehat{\omega}_l$ and the determinants $|\det(\mathbb{J}_{K_m})|$. The choice between storing and recomputing on the fly depends on the hardware at hand. For instance, it is preferable to recompute a quantity if accessing the memory is slower than the actual computing.

30.3.2   Shape functions

Let $\{\widehat{\theta}_n\}_{n \in \mathcal{N}}$ be the reference shape functions. Since these functions and their derivatives need to be evaluated many times at the Gauss nodes in $\widehat{K}$, it can be useful to compute these values once and for all and store them in the two-dimensional array $\texttt{theta}(1{:}n_{\text{sh}}, 1{:}l_{\mathcal{Q}})$ and in the three-dimensional array $\texttt{dtheta\_dhatK}(1{:}d, 1{:}n_{\text{sh}}, 1{:}l_{\mathcal{Q}})$ such that

$$\texttt{theta}(n,l) := \widehat{\theta}_n(\widehat{\boldsymbol{\xi}}_l), \qquad \texttt{dtheta\_dhatK}(k,n,l) := \frac{\partial \widehat{\theta}_n}{\partial \widehat{x}_k}(\widehat{\boldsymbol{\xi}}_l).$$

Let us assume for simplicity that the linear bijective map used to generate the local shape functions is the pullback by the geometric mapping. Then the values of the local shape functions at the Gauss nodes in the mesh cell K_m are given by

$$\theta_n(\boldsymbol{\xi}_{lK_m}) := \widehat{\theta}_n(\widehat{\boldsymbol{\xi}}_l) = \texttt{theta}(n,l),$$

for all $n \in \{1{:}n_{\text{sh}}\}$, all $l \in \{1{:}l_{\mathcal{Q}}\}$, and all $m \in \{1{:}N_{\text{c}}\}$ (notice that the value of $\theta_n(\boldsymbol{\xi}_{lK_m})$ is independent of K_m). Let us now consider the first-order derivatives of the local shape functions

at the Gauss nodes. Using the chain rule, we infer that for all $k_1 \in \{1{:}d\}$,

$$\frac{\partial \theta_n}{\partial x_{k_1}}(\boldsymbol{\xi}_{lK_m}) = \sum_{k_2 \in \{1{:}d\}} \frac{\partial \widehat{\theta}_n}{\partial \widehat{x}_{k_2}}(\widehat{\boldsymbol{\xi}}_l) \left(\mathbb{J}_{K_m}^{-1}(\widehat{\boldsymbol{\xi}}_l) \right)_{k_2, k_1},$$

where we used that $\frac{\partial (\boldsymbol{T}_{K_m}^{-1})_{k_2}}{\partial x_{k_1}}(\boldsymbol{\xi}_{lK_m}) = (\mathbb{J}_{K_m}^{-1}(\widehat{\boldsymbol{\xi}}_l))_{k_2, k_1}$. One can evaluate the partial deriva-
tives of the local shape functions once and for all and store them in the four-dimensional array
$\mathtt{dtheta\_dK}(1{:}d, 1{:}n_{\mathrm{sh}}, 1{:}l_{\mathcal{Q}}, 1{:}N_{\mathrm{c}})$ such that

$$\mathtt{dtheta\_dK}(k, n, l, m) := \frac{\partial \theta_n}{\partial x_k}(\boldsymbol{\xi}_{lK_m}).$$

Notice that the size of this array, $d \times n_{\mathrm{sh}} \times l_{\mathcal{Q}} \times N_{\mathrm{c}}$, can be very large. If the mesh is affine, one can
adopt another strategy since the Jacobian matrix $\mathbb{J}_{K_m}$ and its inverse do not depend on the Gauss
nodes. In this case, one can store the inverse of the Jacobian matrix in the three-dimensional array
$\mathtt{inv\_jac\_K}(1{:}d, 1{:}d, 1{:}N_{\mathrm{c}})$ such that

$$\mathtt{inv\_jac\_K}(k_1, k_2, m) := \left(\left[\mathbb{J}_{K_m} \right]^{-1} \right)_{k_1, k_2}.$$

Then the following operations must be performed each time the quantity $\frac{\partial \theta_n}{\partial x_{k_1}}(\boldsymbol{\xi}_{lK_m})$ is needed:

$$\frac{\partial \theta_n}{\partial x_{k_1}}(\boldsymbol{\xi}_{lK_m}) = \sum_{k_2 \in \{1{:}d\}} \mathtt{dtheta\_dhatK}(k_2, n, l)\, \mathtt{inv\_jac\_K}(k_2, k_1, m).$$

The array $\mathtt{inv\_jac\_K}$ has $d \times d \times N_{\mathrm{c}}$ entries, which is smaller than the number of entries of $\mathtt{dtheta\_dK}$
if $d \ll n_{\mathrm{sh}} \times l_{\mathcal{Q}}$. In this situation, storing $\mathtt{inv\_jac\_K}$ will save memory space at the prize of some
additional computations. But again, depending of the hardware at hand, one must be aware that
a balance must be struck between storage and recomputing on the fly.

30.3.3   Assembling

For simplicity, we assume that a standard Galerkin formulation is considered with the bilinear form
a and the linear form ℓ. The discrete trial and test spaces are identical. Let $\{\varphi_i\}_{i \in \{1{:}I\}}$ be the
global shape functions. In the absence of quadratures, the entries of the stiffness matrix $\mathcal{A} \in \mathbb{R}^{I \times I}$
are $\mathcal{A}_{ij} := a(\varphi_j, \varphi_i)$ for all $i, j \in \{1{:}I\}$, and those of the right-hand side vector are $\mathsf{B}_i := \ell(\varphi_i)$ for
all $i \in \{1{:}I\}$; see §28.1.1. The goal of this section is to revisit the assembling of $\mathcal{A}$ and B when
quadratures are employed.

   Let us first consider the assembling of the stiffness matrix. To fix the ideas, we consider the
bilinear form associated with a diffusion-advection-reaction model problem for which $a(v_h, w_h) :=
\int_D A(\boldsymbol{x}, v_h, w_h)\, \mathrm{d}\boldsymbol{x}$, with

$$A(\boldsymbol{x}, \varphi, \psi) := \mathrm{d}(\boldsymbol{x}) \nabla \varphi(\boldsymbol{x}) \cdot \nabla \psi(\boldsymbol{x}) + \psi(\boldsymbol{x}) \boldsymbol{\beta}(\boldsymbol{x}) \cdot \nabla \varphi(\boldsymbol{x}) + \varphi(\boldsymbol{x}) \mu(\boldsymbol{x}) \psi(\boldsymbol{x}),$$

with smooth fields d, $\boldsymbol{\beta}$, and μ taking values in $\mathbb{R}^{d \times d}$, $\mathbb{R}^d$, and $\mathbb{R}$, respectively. This model problem
is considered, e.g., in Chapter 31. In Chapter 24, we considered the simpler setting where $\mathrm{d} := \mathbb{I}_d$,
$\boldsymbol{\beta} := \boldsymbol{0}$, and $\mu := 0$; see (24.7). In Cartesian notation, the quantity $A(\boldsymbol{x}, \varphi, \psi)$ is expressed as
follows:

$$A(\boldsymbol{x}, \varphi, \psi) := \sum_{k_1, k_2 \in \{1{:}d\}} \frac{\partial \varphi}{\partial x_{k_1}}(\boldsymbol{x}) \mathrm{d}_{k_1 k_2}(\boldsymbol{x}) \frac{\partial \psi}{\partial x_{k_2}}(\boldsymbol{x})$$

$$+ \psi(\boldsymbol{x}) \sum_{k_1 \in \{1{:}d\}} \boldsymbol{\beta}_{k_1}(\boldsymbol{x}) \frac{\partial \varphi}{\partial x_{k_1}}(\boldsymbol{x}) + \varphi(\boldsymbol{x}) \mu(\boldsymbol{x}) \psi(\boldsymbol{x}).$$

Using quadratures to integrate $A(\boldsymbol{x}, \varphi, \psi)$ over all the mesh cells, we obtain the approximate bilinear form $a_{\mathcal{Q}}$ s.t.

$$a_{\mathcal{Q}}(v_h, w_h) := \sum_{m \in \{1:N_c\}} \sum_{l \in \{1:l_{\mathcal{Q}}\}} \omega_{lK_m} A(\boldsymbol{\xi}_{lK_m}, v_{h|K_m}, w_{h|K_m}).$$

This leads to the approximate stiffness matrix $\mathcal{A}_{\mathcal{Q}}$ with entries $\mathcal{A}_{\mathcal{Q},ij} := a_{\mathcal{Q}}(\varphi_j, \varphi_i)$ for all $i, j \in \{1:I\}$.

Algorithm 30.1 Assembling of $\mathcal{A}_{\mathcal{Q}}$ for analytic data.

$\mathcal{A}_{\mathcal{Q}} = 0$
for $m \in \{1:N_c\}$ **do**
   for $l \in \{1:l_{\mathcal{Q}}\}$ **do**; $\mathtt{tmp} := 0$
      for $k \in \{1:d\}$ **do**
         $\mathtt{xi\_l}(k) := \displaystyle\sum_{n \in \{1:n_{\mathrm{geo}}\}} \mathtt{coord}(k, \mathtt{j\_geo}(n, m)) * \mathtt{psi}(n, l)$
      end for
      for $ni \in \{1:n_{\mathrm{sh}}\}$ **do**
         for $nj \in \{1:n_{\mathrm{sh}}\}$ **do**
            $x_1 := \displaystyle\sum_{k_1,k_2 \in \{1:d\}} \mathtt{dtheta\_dK}(k_1, nj, l, m) * \mathbb{d}_{k_1 k_2}(\mathtt{xi\_l}) * \mathtt{dtheta\_dK}(k_2, ni, l, m)$
            $x_2 := \mathtt{theta}(ni, l) \displaystyle\sum_{k_1 \in \{1:d\}} \boldsymbol{\beta}_{k_1}(\mathtt{xi\_l}) * \mathtt{dtheta\_dK}(k_1, nj, l, m)$
            $x_3 := \mathtt{theta}(ni, l) * \mu(\mathtt{xi\_l}) * \mathtt{theta}(nj, l)$
            $\mathtt{tmp}(ni, nj) := \mathtt{tmp}(ni, nj) + [x_1 + x_2 + x_3] * \mathtt{weight\_K}(l, m)$
         end for
      end for
   end for
   Accumulate $\mathtt{tmp}$ in $\mathcal{A}_{\mathcal{Q}}$ as in Algorithm 29.2
end for

A general assembling procedure for the stiffness matrix $\mathcal{A}$ (stored in the CSR format) has been outlined in Algorithm 29.2. Our goal is now to detail the evaluation of the array $\mathtt{tmp}$ used in this algorithm by means of quadratures. We assume for simplicity that the coefficients $(\mathbb{d}_{k_1 k_2})_{k_1, k_2 \in \{1:d\}}$, $(\boldsymbol{\beta}_{k_1})_{k_1 \in \{1:d\}}$, and μ are known analytically; see Exercise 30.7 for discrete data. The assembling procedure of the approximate stiffness matrix $\mathcal{A}_{\mathcal{Q}}$ is shown in Algorithm 30.1. Notice that we first evaluate and store the coordinates of the Gauss nodes $\boldsymbol{\xi}_{lK_m}$ since we need to evaluate the values of the coefficients at these nodes.

The assembling of the right-hand side vector can be performed similarly. To fix the ideas, we consider a linear form such that $\ell(w_h) := \int_D F(\boldsymbol{x}, w_h) \, \mathrm{d}x$, with $F(\boldsymbol{x}, \psi) := f(\boldsymbol{x})\psi(\boldsymbol{x})$ and $f : D \to \mathbb{R}$ is a smooth function. Using quadratures to integrate $F(\boldsymbol{x}, \psi)$ over all the mesh cells, we obtain the approximate linear form $\ell_{\mathcal{Q}}$ s.t.

$$\ell_{\mathcal{Q}}(w_h) := \sum_{m \in \{1:N_c\}} \sum_{l \in \{1:l_{\mathcal{Q}}\}} \omega_{lK_m} F(\boldsymbol{\xi}_{lK_m}, w_{h|K_m}).$$

This leads to the approximate right-hand side vector $\mathsf{B}_{\mathcal{Q}}$ with entries $\mathsf{B}_{\mathcal{Q},i} := \ell_{\mathcal{Q}}(\varphi_i)$ for all $i \in \{1:I\}$. The assembling procedure of the vector $\mathsf{B}_{\mathcal{Q}}$ is presented in Algorithm 30.2.

Algorithm 30.2 Assembling of $\mathsf{B}_{\mathcal{Q}}$ for analytic data.

```
B_Q = 0
for m ∈ {1:N_c} do
    for l ∈ {1:l_Q} do;  tmp := 0
        for k_1 ∈ {1:d} do
            xi_l(k_1) :=      Σ       coord(k_1, j_geo(n, m)) psi(n, l)
                        n∈{1:n_geo}
        end for
        for ni ∈ {1:n_sh} do
            tmp(ni) := tmp(ni) + f(xi_l) * theta(ni, l) * weight_K(l, m)
        end for
    end for
    for ni ∈ {1:n_sh} do;  i := j_dof(m, ni)
        B_{Q,i} := B_{Q,i} + tmp(ni)
    end for
end for
```

Exercises

Exercise 30.1 (Quadratures on simplices). Let K be a simplex in $\mathbb{R}^d$. Let $\boldsymbol{z}_K$ be the barycenter of K, let $\{\boldsymbol{z}_i\}_{i\in\{0:d\}}$ be the vertices of K, and let $\{\boldsymbol{m}_i\}_{i\in\{0:d\}}$ be the midpoints of the edges of K. Consider the following quadratures: $\{\boldsymbol{z}_K\}$, $\{|K|\}$; $\{\boldsymbol{z}_i\}_{i\in\{0:d\}}$, $\{\frac{1}{d+1}|K|\}$; $\{\boldsymbol{m}_i\}_{i\in\{0:d\}}$, $\{\frac{1}{d+1}|K|\}$. (i) Prove that the first and the second quadratures are of order one. (ii) Prove that the third one is of order two for $d = 2$.

Exercise 30.2 (Quadrature for $\mathbb{Q}_{2,d}$). Let $\widehat{K} := [0,1]^d$ be the unit hypercube. Let $\widehat{\boldsymbol{a}}_{i_1\ldots i_d} := (\frac{i_1}{2}, \ldots, \frac{i_d}{2})$, $i_1, \ldots, i_d \in \{0:2\}$. Show that the quadrature $\int_{\widehat{K}} f(\widehat{\boldsymbol{x}}) \, d\widehat{x} \approx \sum_{i_1,\ldots i_d} w_{i_1\ldots i_d} f(\widehat{\boldsymbol{a}}_{i_1\ldots i_d})$ where $w_{i_1\ldots i_d} := \frac{1}{6^d} \prod_{k=1}^d (3 i_k(2 - i_k) + 1)$ is exact for all $f \in \mathbb{Q}_{2,d}$. (*Hint*: write the $\mathbb{Q}_{2,d}$ Lagrange shape functions in tensor-product form and use Simpson's rule in each direction.)

Exercise 30.3 (Global quadrature error). Prove that

$$\left| \int_D \phi(\boldsymbol{x}) \, dx - \sum_{K\in\mathcal{T}_h} \sum_{l\in\{1:l_Q\}} \omega_{lK} \, \phi(\boldsymbol{\xi}_{lK}) \right| \leq c h^m |D|^{1-\frac{1}{p}} |\phi|_{W^{m,p}(D)},$$

for all $\phi \in W^{m,p}(D)$ and all $h \in \mathcal{H}$. (*Hint*: use Lemma 30.9.)

Exercise 30.4 (Quadrature error with polynomial). The goal is to prove (30.7). We are going to make use of (30.6) formulated as follows: $|E_K(\psi q)| \leq c h_K^\mu |\psi|_{W^{\mu,\infty}(K)} \|q\|_{L^1(K)}$ for all $q \in \mathbb{P}_{\nu,d} \circ \boldsymbol{T}_K$ where $\mu + \nu - 1 \leq k_Q$, $\mu, \nu \in \mathbb{N}$. (i) Prove that $|E_K(\phi \underline{p}_K)| \leq c h_K^m |\phi|_{W^{m,\infty}(K)} \|p\|_{L^1(K)}$, where $\underline{p}_K$ is the mean value of p over K. (ii) Prove (30.7). (*Hint*: use Step (i) with $\mu := m - 1$.)

Exercise 30.5 (Surface quadrature). Assume $d = 3$. Let F be a face of a mesh cell. Let $\widehat{F} \subset \mathbb{R}^2$ be a reference face and let $\boldsymbol{T}_F : \widehat{F} \to F$ be the geometric mapping for F. Let $\boldsymbol{t}_1(\widehat{\boldsymbol{s}}), \boldsymbol{t}_2(\widehat{\boldsymbol{s}})$ be the two column vectors of the Jacobian matrix of $\boldsymbol{T}_F(\widehat{\boldsymbol{s}})$, say $\mathbb{J}_F(\widehat{\boldsymbol{s}}) := [\boldsymbol{t}_1(\widehat{\boldsymbol{s}}), \boldsymbol{t}_2(\widehat{\boldsymbol{s}})] \in \mathbb{R}^{3\times 2}$. (i) Compute the metric tensor $\mathbb{g}_F := \mathbb{J}_F^\mathsf{T} \mathbb{J}_F \in \mathbb{R}^{2\times 2}$ in terms of the dot products $\boldsymbol{t}_i \cdot \boldsymbol{t}_j$, $i, j \in \{1, 2\}$. (ii) Show that $ds = \|\boldsymbol{t}_1(\widehat{\boldsymbol{s}}) \times \boldsymbol{t}_2(\widehat{\boldsymbol{s}})\|_{\ell^2(\mathbb{R}^3)} \, d\widehat{s}$. (*Hint*: use Lagrange's identity, that is, $\|\boldsymbol{a}\|_{\ell^2(\mathbb{R}^3)}^2 \|\boldsymbol{b}\|_{\ell^2(\mathbb{R}^3)}^2 - (\boldsymbol{a} \cdot \boldsymbol{b})^2 = \|\boldsymbol{a} \times \boldsymbol{b}\|_{\ell^2(\mathbb{R}^3)}^2$ for any pair of vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^3$, and recall that $ds = \sqrt{\det(\mathbb{g}_F)} \, d\widehat{s}$). (iii) Given a quadrature $\{\widehat{\boldsymbol{s}}_l, \widehat{w}_l\}_{l\in\{1:l_Q^\partial\}}$ on $\widehat{F}$, generate the quadrature on F.

Exercise 30.6 (Asssembling). Let $D := (0,1)^2$. Consider the problem $-\Delta u + u = 1$ in D and $u_{|\partial D} = 0$. (i) Approximate its solution with $\mathbb{P}_1$ H^1-conforming finite elements on the two meshes shown in Figure 30.1. (ii) Evaluate the discrete solution in both cases. (*Hint*: there is only one degree of freedom in both cases, see Exercise 28.5 for computing the gradient part of the stiffness coefficient and use a quadrature from Table 30.1 for the zero-order term.) (iii) For a fine mesh composed of 800 elements, we have $u_h(\frac{1}{2}, \frac{1}{2}) \approx 0.0702$. Comment.
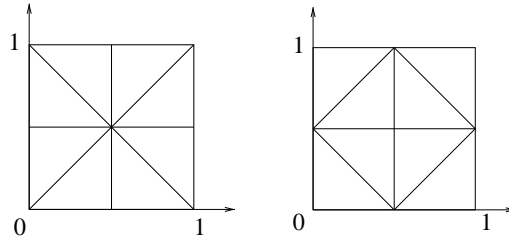
Figure 30.1: Illustration for Exercise 30.6.

Exercise 30.7 (Discrete data). Adapt Algorithm 30.1 to the case where $(\mathbb{d}_{k_1 k_2})_{k_1, k_2 \in \{1:d\}}$, $(\boldsymbol{\beta}_{k_1})_{k_1 \in \{1:d\}}$, and μ are known in the discrete space V_h. (*Hint*: let `dif`, `beta`, and `mu` be the corresponding coordinate vectors, and observe that $\mu(\boldsymbol{\xi}_{lK_m}) = \sum_{n \in \{1:n_{sh}\}}$ `mu(j_dof(m,i))` $\times$ `theta(n,l)`, etc.)

Exercise 30.8 (Assembling of RHS). Write the assembling algorithm for the right-hand side vector in the case where $F(\boldsymbol{\xi}, w_h) := f(\boldsymbol{\xi}) w_h(\boldsymbol{\xi}) + \sum_{k_1 \in \{1:d\}} \boldsymbol{\beta}_{k_1}(\boldsymbol{\xi}) \frac{\partial w_h}{\partial x_{k_1}}(\boldsymbol{\xi})$ with analytically known data.

Chapter 31

Scalar second-order elliptic PDEs

In Part VII, composed of Chapters 31 to 35, we study the approximation of scalar second-order elliptic PDEs by H^1-conforming finite elements. Among the topics we address in this part are weak formulations and well-posedness, a priori error analysis, the discrete maximum principle, the impact of quadratures, and a posteriori error analysis. In Chapters 31 to 34, we focus on weak formulations endowed with a *coercivity property*, so that well-posedness hinges on the Lax–Milgram lemma and the error analysis on Céa's lemma (and its variants). In Chapter 35, we study the Helmholtz problem as an example of elliptic PDE without coercivity.

The present chapter addresses fundamental properties of scalar-valued second-order elliptic PDEs endowed with a coercivity property. The prototypical example is the Laplacian with homogeneous Dirichlet conditions. More generally, we consider PDEs including lower-order terms, such as the diffusion-advection-reaction equation, where the lower-order terms are small enough so as not to pollute the coercivity provided by the diffusion operator. We also study in some detail how various boundary conditions (Dirichlet, Neumann, Robin) can be enforced in the weak formulation. Moreover, important smoothness properties of the solutions to scalar second-order elliptic PDEs are listed at the end of the chapter. These results will be useful later to establish error estimates for the finite element approximation.

31.1 Model problem

Let D be a domain in $\mathbb{R}^d$, i.e., D is a nonempty, open, bounded, connected subset of $\mathbb{R}^d$ (see Definition 3.1). Let $\mathbb{d}$, $\boldsymbol{\beta}$, and μ be functions defined on D that take values in $\mathbb{R}^{d \times d}$, $\mathbb{R}^d$, and $\mathbb{R}$, respectively. Given a function $f : D \to \mathbb{R}$, we look for a function $u : D \to \mathbb{R}$ that solves the following linear PDE:

$$-\nabla\cdot(\mathbb{d}\nabla u) + \boldsymbol{\beta}\cdot\nabla u + \mu u = f \quad \text{in } D. \tag{31.1}$$

Boundary conditions are discussed later. Using Cartesian coordinates, the PDE (31.1) amounts to $\sum_{i,j\in\{1:d\}} \frac{\partial}{\partial x_i}\left(\mathbb{d}_{ij}\frac{\partial u}{\partial x_j}\right) + \sum_{i\in\{1:d\}} \beta_i\frac{\partial u}{\partial x_i} + \mu u = f$. The PDE reduces to the *Poisson equation* $-\Delta u = f$ studied in §24.1 if $\mathbb{d}$ is the identity tensor in $\mathbb{R}^d$ and $\boldsymbol{\beta}$ and μ vanish identically. More generally, (31.1) is a diffusion-advection-reaction equation modeling for instance heat or mass transfer or flows in porous media. The first term on the left-hand side of (31.1) accounts for diffusion processes, the second one for advection processes, and the third one for reaction processes (depletion occurs when μ is positive).

31.1.1   Ellipticity and assumptions on the data

We assume that $\mathbb{d} \in \mathbb{L}^\infty(D) := L^\infty(D; \mathbb{R}^{d \times d})$ and that $\mathbb{d}$ takes symmetric values. We also assume that $\boldsymbol{\beta} \in \boldsymbol{W}^{1,\infty}(D) := W^{1,\infty}(D; \mathbb{R}^d)$, $\mu \in L^\infty(D)$, and $f \in L^2(D)$. For dimensional consistency, we equip the space $H^1(D)$ with the norm $\|v\|_{H^1(D)} := (\|v\|_{L^2(D)}^2 + \ell_D^2 \|\nabla v\|_{\boldsymbol{L}^2(D)}^2)^{\frac{1}{2}}$, where ℓ_D is a length scale associated with D, e.g., $\ell_D := \text{diam}(D)$. A key notion for the second-order PDE (31.1) is that of ellipticity.

Definition 31.1 (Ellipticity). *For a.e. $\boldsymbol{x} \in D$, let $[\lambda_{\min}(\boldsymbol{x}), \lambda_{\max}(\boldsymbol{x})]$ be the smallest interval containing the eigenvalues of $\mathbb{d}(\boldsymbol{x})$. We say that the PDE* (31.1) *is elliptic if*

$$0 < \lambda_\flat := \operatorname*{ess\,inf}_{\boldsymbol{x} \in D} \lambda_{\min}(\boldsymbol{x}) \le \operatorname*{ess\,sup}_{\boldsymbol{x} \in D} \lambda_{\max}(\boldsymbol{x}) =: \lambda_\sharp < \infty. \tag{31.2}$$

Example 31.2 (Anisotropic diffusion). We say that the diffusion process is *anisotropic* if the diffusion matrix is not proportional to the identity, as in the PDE $-\frac{\partial^2 u}{\partial x_1^2} + 2\kappa \frac{\partial^2 u}{\partial x_1 \partial x_2} - \frac{\partial^2 u}{\partial x_2^2} = f$ which is is elliptic if $\kappa \in (-1, 1)$. □

Remark 31.3 (Divergence form, Cordes condition). The PDE (31.1) is said to be in divergence form because of the way the second-order term is written. One can also consider the PDE in nondivergence form $-\mathbb{d}:D^2 u + \boldsymbol{\beta} \cdot \nabla u + \mu u = f$, where $\mathbb{d}:D^2 u = \sum_{i,j \in \{1:d\}} \mathbb{d}_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j}$. In this case, one usually adds the Cordes condition [141] to the ellipticity assumption: There is $\epsilon \in (0, 1]$ s.t. $\frac{\|\mathbb{d}\|_F^2}{(\text{tr}(\mathbb{d}))^2} \le \frac{1}{d-1+\epsilon}$ uniformly in D, where $\|\mathbb{d}\|_F = (\mathbb{d}:\mathbb{d})^{\frac{1}{2}}$ is the Frobenius norm of $\mathbb{d}$ and $\text{tr}(\mathbb{d})$ its trace (note that $\text{tr}(\mathbb{d}) > 0$ owing to the ellipticity condition). We refer the reader to Smears and Süli [348] for further insight in the context of Hamilton–Jacobi–Bellman equations. □

The following important result, which is similar to the unique continuation principle for real analytic functions, hinges on the ellipticity property.

Theorem 31.4 (Unique continuation principle). *Let D be a connected subset of $\mathbb{R}^d$ with $\boldsymbol{0} \in D$. Assume that $\mathbb{d}$ satisfies the ellipticity condition* (31.2), *$\mathbb{d}_{ij} \in C^0(D; \mathbb{R})$, $\mathbb{d}_{ij}$ is Lipschitz continuous in $D \backslash \{\boldsymbol{0}\}$, and there are $c > 0$ and $\delta > 0$ such that $\|\nabla \mathbb{d}_{ij}(\boldsymbol{x})\|_{\ell^2} \le c \|\boldsymbol{x}\|_{\ell^2}^{\delta-1}$ for all $\boldsymbol{x} \in D$. Let $u \in H^1_{\text{loc}}(D)$ and assume that*

$$|\mathbb{d}:D^2 u| \le c \sum_{|\alpha| \le 1} \|\boldsymbol{x}\|_{\ell^2}^{\delta+|\alpha|-2} |\partial^\alpha u|, \tag{31.3a}$$

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon^d} \int_{\|\boldsymbol{x}\|_{\ell^2} < \epsilon} u^2(\boldsymbol{x}) \, \mathrm{d}x = 0. \tag{31.3b}$$

Then $u = 0$ in D.

Proof. See Hörmander [247, Thm. 17.2.6]. We also refer the reader to Reed and Simon [332, Thm. XIII.57&63] for variations on the unique continuation principle that are somewhat easier to grasp. □

Example 31.5 (Application to (31.1)**).** The above result, known in the literature as the Aronszajn–Cordes uniqueness theorem, can be used to establish the uniqueness of the solution to the PDE (31.1). Assume that u_1, u_2 are two solutions of (31.1), and assume that one can show that $u_1 \in H^1_{\text{loc}}(D)$, $u_2 \in H^1_{\text{loc}}(D)$, and there is an open set $S \subset D$ s.t. $(u_1 - u_2)_{|S} = 0$. One can always assume that $\boldsymbol{0} \in S$. Setting $u := u_1 - u_2$, one has $-\mathbb{d}:D^2 u = (\nabla \cdot \mathbb{d} - \boldsymbol{\beta}) \cdot \nabla u - \mu u$. Let us assume that $\mathbb{d}$ satisfies the assumptions of Theorem 31.4. Then one immediately deduces that (31.3a) holds true with some appropriate constant c. Using that $u_{|S} = 0$, the second condition (31.3b) is trivially satisfied, and uniqueness follows. □

31.1.2 Toward a weak formulation

Proceeding informally as in §24.1, e.g., assuming $u \in H^2(D)$, we multiply (31.1) by a test function $w \in H^1(D)$ and integrate over D to obtain

$$\int_D \left(-\nabla\cdot(d\nabla u)w + (\boldsymbol{\beta}\cdot\nabla u)w + \mu uw \right) dx = \int_D fw \, dx. \tag{31.4}$$

Integrating by parts the first term on the left-hand side leads to

$$\int_D -\nabla\cdot(d\nabla u)w \, dx = \int_D (d\nabla u)\cdot\nabla w \, dx - \int_{\partial D} (\boldsymbol{n}\cdot(d\nabla u))w \, ds, \tag{31.5}$$

where $\boldsymbol{n}$ denotes the outward unit normal to D. We then arrive at

$$a(u, w) - \int_{\partial D} (\boldsymbol{n}\cdot(d\nabla u))w \, ds = \int_D fw \, dx, \qquad \forall w \in H^1(D), \tag{31.6}$$

where a is defined for all $(v, w) \in H^1(D) \times H^1(D)$ as follows:

$$a(v, w) := \int_D \left((d\nabla v)\cdot\nabla w + (\boldsymbol{\beta}\cdot\nabla v)w + \mu vw \right) dx. \tag{31.7}$$

Notice in passing that using Cartesian coordinates, the symmetry of d implies $(d\nabla v)\cdot\nabla w = \sum_{i,j\in\{1:d\}} d_{ij}\frac{\partial v}{\partial x_j}\frac{\partial w}{\partial x_i} = \sum_{i,j\in\{1:d\}} d_{ij}\frac{\partial v}{\partial x_i}\frac{\partial w}{\partial x_j} = \nabla v\cdot(d\nabla w)$, that is, $(d\nabla v)\cdot\nabla w = \nabla v\cdot(d\nabla w)$. Moreover, using the Cauchy–Schwarz inequality for the three integrals leads to

$$|a(v, w)| \leq (\lambda_\sharp \ell_D^{-2} + \beta_\sharp \ell_D^{-1} + \mu_\sharp)\|v\|_{H^1(D)}\|w\|_{H^1(D)}, \tag{31.8}$$

for all $v, w \in H^1(D)$, with $\beta_\sharp := \|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(D)}$ and $\mu_\sharp := \|\mu\|_{L^\infty(D)}$, which proves that the bilinear form a is bounded on $H^1(D)\times H^1(D)$. Equation (31.6) is the starting point to derive weak formulations for the PDE (31.1) with various types of boundary conditions.

31.2 Dirichlet boundary condition

Our goal is now to prove the well-posedness of the weak formulation when a Dirichlet boundary condition is enforced. In what follows, we identify $L^2(D)$ with its dual space $L^2(D)'$ so that we are in the situation where

$$H_0^1(D) \hookrightarrow L^2(D) \equiv L^2(D)' \hookrightarrow H^{-1}(D) = H_0^1(D)', \tag{31.9}$$

with bounded and densely defined embeddings (recall that the notation $V \hookrightarrow W$ means that the embedding of V into W is bounded).

31.2.1 Homogeneous Dirichlet condition

We consider the homogeneous Dirichlet condition

$$u = 0 \quad \text{on } \partial D, \tag{31.10}$$

which we are going to enforce strongly by using the space $H_0^1(D)$ for both the trial and the test spaces. Recall from the trace theorem (Theorem 3.10) that $u \in H_0^1(D)$ implies that $\gamma^g(u) = 0$,

where $\gamma^{\mathrm{g}} : H^1(D) \to H^{\frac{1}{2}}(\partial D)$ is the trace map such that $\gamma^{\mathrm{g}}(v) = v_{|\partial D}$ if the function v is smooth. Since the test functions vanish at the boundary, we can drop the boundary term on the left-hand side of (31.6), leading to the following weak formulation:

$$\begin{cases} \text{Find } u \in V := H_0^1(D) \text{ such that} \\ a(u,w) = \int_D fw \, dx, \quad \forall w \in V. \end{cases} \tag{31.11}$$

Proposition 31.6 (Weak solution). *Let $f \in L^2(D)$. If the function $u \in H_0^1(D)$ solves (31.11), then it satisfies the PDE (31.1) a.e. in D and the boundary condition (31.10) a.e. on ∂D.*

Proof. Let u be a weak solution. Testing the weak formulation (31.11) against an arbitrary function $\varphi \in C_0^\infty(D) \subset H_0^1(D)$ and using the notion of weak derivatives leads to $\langle -\nabla\cdot(\mathrm{d}\nabla u), \varphi \rangle = \int_D (f - \boldsymbol{\beta}\cdot\nabla u - \mu u)\varphi \, dx$ since $f \in L^2(D)$ and $\boldsymbol{\beta}\cdot\nabla u + \mu u \in L^2(D)$ owing to the assumptions on the data. Hence, $-\nabla\cdot(\mathrm{d}\nabla u)$ defines a bounded linear form on $L^2(D)$ with Riesz–Fréchet representative equal to $f - \boldsymbol{\beta}\cdot\nabla u - \mu u$. This means that u solves the PDE (31.1) a.e. in D. Moreover, $u \in H_0^1(D)$ implies that $\gamma^{\mathrm{g}}(u) = 0$ in $H^{\frac{1}{2}}(\partial D) \hookrightarrow L^2(\partial D)$, i.e., the boundary condition (31.10) holds a.e. on ∂D. $\qquad\square$

Remark 31.7 ($f \in H^{-1}(D)$). When $f \in H^{-1}(D)$, the term $\int_D fw \, dx$ in (31.11) must be understood as $\langle f, w \rangle_{H^{-1}(D), H_0^1(D)}$. More specifically, recalling from Theorem 4.12 that the assumption $f \in H^{-1}(D)$ is equivalent to assuming that there are $g_0 \in L^2(D)$ and $\boldsymbol{g}_1 \in \boldsymbol{L}^2(D)$ such that $\langle f, w \rangle_{H^{-1}(D), H_0^1(D)} = \int_D (g_0 w + \boldsymbol{g}_1\cdot\nabla w) \, dx$, each time we write $a(u,w) = \langle f, w \rangle_{H^{-1}(D), H_0^1(D)}$, we actually mean $a(u,w) = \int_D (g_0 w + \boldsymbol{g}_1\cdot\nabla w) \, dx$, and the PDE we actually solve is $-\nabla\cdot(\mathrm{d}\nabla u) + \boldsymbol{\beta}\cdot\nabla u + \mu u = g_0 - \nabla\cdot\boldsymbol{g}_1$ in $H^{-1}(D)$. $\qquad\square$

We now make assumptions on the PDE coefficients that are *sufficient* to prove the well-posedness of (31.11) by invoking a coercivity property. Recall the Poincaré–Steklov inequality (3.11) (with $p := 2$), i.e., there is $C_{\mathrm{PS}} > 0$ such that

$$C_{\mathrm{PS}} \|v\|_{L^2(D)} \leq \ell_D \|\nabla v\|_{\boldsymbol{L}^2(D)}, \qquad \forall v \in H_0^1(D). \tag{31.12}$$

Owing to (31.12), we can equip the space $V := H_0^1(D)$ with the norm

$$\|v\|_V := \|\nabla v\|_{\boldsymbol{L}^2(D)} = |v|_{H^1(D)}. \tag{31.13}$$

The space V equipped with this norm is a Hilbert space since $\|v\|_V \leq \ell_D^{-1}\|v\|_{H^1(D)} \leq (1 + C_{\mathrm{PS}}^{-2})^{\frac{1}{2}}\|v\|_V$ for all $v \in V$.

Proposition 31.8 (Well-posedness). *Assume the ellipticity condition (31.2). Assume that there exists $\theta > 0$ such that*

$$\mu_\flat := \operatorname*{ess\,inf}_{\boldsymbol{x}\in D} \left(\mu - \frac{1}{2}\nabla\cdot\boldsymbol{\beta}\right)(\boldsymbol{x}) \geq -(1-\theta)C_{\mathrm{PS}}^2 \ell_D^{-2}\lambda_\flat. \tag{31.14}$$

(i) *The bilinear form a is V-coercive:*

$$a(v,v) \geq \lambda_\flat \min(1,\theta)\|v\|_V^2, \qquad \forall v \in V. \tag{31.15}$$

(ii) *The problem (31.11) is well-posed.*

Proof. The boundedness property (31.8) of a can be rewritten as

$$|a(v,w)| \leq \left(\lambda_\sharp + C_{\mathrm{PS}}^{-1}\ell_D\beta_\sharp + C_{\mathrm{PS}}^{-2}\ell_D^2\mu_\sharp\right)\|v\|_V\|w\|_V,$$

for all $v, w \in V$. Moreover, the linear form $\ell(w) := \int_D fw \, dx$ is bounded on V since $|\ell(w)| \leq \|f\|_{L^2(D)} \|w\|_{L^2(D)} \leq \|f\|_{L^2(D)} C_{\mathrm{PS}}^{-1} \ell_D \|w\|_V$ for all $w \in V$. Let us now prove the coercivity property (31.15). Using the divergence formula for the field $(\frac{1}{2}v^2)\boldsymbol{\beta}$, we infer that

$$\int_D v(\boldsymbol{\beta}\cdot\nabla v) \, dx = -\frac{1}{2} \int_D (\nabla\cdot\boldsymbol{\beta}) v^2 \, dx + \frac{1}{2} \int_{\partial D} (\boldsymbol{\beta}\cdot\boldsymbol{n}) v^2 \, ds, \tag{31.16}$$

for all smooth functions $v \in C^\infty(\overline{D})$. A density argument then shows that the formula (31.16) remains valid for all $v \in H^1(D)$. Using the definition of $\mu_\flat$, the identity (31.16), and that v vanishes at the boundary, we obtain $a(v,v) \geq \int_D \left(\lambda_\flat \|\nabla v\|_{\ell^2(\mathbb{R}^d)}^2 + \mu_\flat |v|^2 \right) dx$ for all $v \in V$. The assumptions on $\lambda_\flat$ and $\mu_\flat$ imply that

$$a(v,v) \geq \lambda_\flat \left(\|\nabla v\|_{\boldsymbol{L}^2(D)}^2 - (1-\theta) C_{\mathrm{PS}}^2 \ell_D^{-2} \|v\|_{L^2(D)}^2 \right).$$

If $\theta > 1$, the last term is positive, whereas if $\theta \in (0,1]$, we have

$$\|\nabla v\|_{\boldsymbol{L}^2(D)}^2 - (1-\theta) C_{\mathrm{PS}}^2 \ell_D^{-2} \|v\|_{L^2(D)}^2$$
$$= \theta \|\nabla v\|_{\boldsymbol{L}^2(D)}^2 + (1-\theta)(\|\nabla v\|_{\boldsymbol{L}^2(D)}^2 - C_{\mathrm{PS}}^2 \ell_D^{-2} \|v\|_{L^2(D)}^2) \geq \theta \|\nabla v\|_{\boldsymbol{L}^2(D)}^2,$$

where the last bound follows from the Poincaré–Steklov inequality. The coercivity property (31.15) then results from $\|v\|_V := \|\nabla v\|_{\boldsymbol{L}^2(D)}$. Finally, the well-posedness of (31.11) follows from the Lax–Milgram lemma. $\square$

Example 31.9 (Pure diffusion). Coercivity for a purely diffusive problem (so that $\mu_\flat = 0$) holds true with $\theta := 1$. $\square$

Remark 31.10 (Variational formulation). Assume that $\boldsymbol{\beta} = \boldsymbol{0}$ in D. Then u solves (31.11) iff u minimizes in $H_0^1(D)$ the energy functional $\mathfrak{E}_{\mathrm{D}}(v) := \frac{1}{2} \int_D \left(\nabla v \cdot (\mathrm{d}\nabla v) + \mu v^2 - 2fv \right) dx$; see Proposition 25.8. $\square$

Remark 31.11 (Helmholtz). The condition (31.14) is only sufficient to ensure the well-posedness of (31.11) by means of a coercivity argument. We will see in Chapter 35, which deals with the Helmholtz problem, that well-posedness can also hold without invoking (31.14). In this case, we will establish well-posedness by means of an inf-sup argument. $\square$

31.2.2 Non-homogeneous Dirichlet condition

Let $g \in H^{\frac{1}{2}}(\partial D)$. We consider the non-homogeneous Dirichlet condition

$$u = g \quad \text{on } \partial D. \tag{31.17}$$

Since the map $\gamma^{\mathrm{g}} : H^1(D) \to H^{\frac{1}{2}}(\partial D)$ is surjective, there is a uniform constant $C_{\gamma^{\mathrm{g}}}$ and $u_g \in H^1(D)$ such $\gamma^{\mathrm{g}}(u_g) = g$ and $\|u_g\|_{H^1(D)} \leq C_{\gamma^{\mathrm{g}}} \|g\|_{H^{\frac{1}{2}}(\partial D)}$; see Theorem 3.10(iii). Setting $u_0 := u - u_g$, we obtain $\gamma^{\mathrm{g}}(u - u_g) = g - g = 0$, i.e., $u_0 \in H_0^1(D)$. This leads to the following weak formulation:

$$\begin{cases} \text{Find } u \in H^1(D) \text{ such that } u_0 := u - u_g \in V := H_0^1(D) \text{ satisfies} \\ a(u_0, w) = \int_D fw \, dx - a(u_g, w), \quad \forall w \in V. \end{cases} \tag{31.18}$$

The right-hand side in (31.18) defines a bounded linear form on V owing to the boundedness of a on $H^1(D) \times H^1(D)$ and the above bound on u_g. Proceeding as in the homogeneous case, one can prove the following result.

Proposition 31.12 (Well-posedness). *Let $f \in L^2(D)$ and $g \in H^{\frac{1}{2}}(\partial D)$. (i) If the function $u \in H^1(D)$ solves (31.18), then it satisfies the PDE (31.1) a.e. in D and the boundary condition (31.17) a.e. on ∂D. (ii) Under the assumptions of Proposition 31.8, (31.18) is well-posed.*

31.3 Robin/Neumann conditions

The Dirichlet conditions are called *essential boundary conditions* since they are imposed explicitly in the solution space. The Robin and the Neumann conditions belong to the class of *natural boundary conditions*. These conditions are not explicitly enforced in the solution space, but they are enforced in the weak formulations by using test functions that are not zero at the boundary.

31.3.1 Robin condition

Let $\rho \in L^\infty(\partial D)$ and $g \in L^2(\partial D)$. We consider the Robin boundary condition

$$\rho u + \boldsymbol{n} \cdot (\mathbb{d}\nabla u) = g \quad \text{on } \partial D. \tag{31.19}$$

Starting from (31.6) and still proceeding informally, we consider test functions in $H^1(D)$ (i.e., they are no longer in $H_0^1(D)$ as for the Dirichlet conditions), and we use the Robin condition in the boundary integral on the left-hand side of (31.6), thereby replacing $\boldsymbol{n} \cdot (\mathbb{d}\nabla u)$ by $g - \rho u$. This leads to $a(u, w) + \int_{\partial D} (g - \rho u) w \, \mathrm{d}s = \int_D f w \, \mathrm{d}x$. Introducing the trace map γ^{g} in the boundary term and rearranging the expression, we obtain the following weak formulation:

$$\begin{cases} \text{Find } u \in V := H^1(D) \text{ such that} \\ a_\rho(u, w) = \int_D f w \, \mathrm{d}x + \int_{\partial D} g \gamma^{\mathrm{g}}(w) \, \mathrm{d}s, \quad \forall w \in V, \end{cases} \tag{31.20}$$

with the bilinear form a_ρ on $H^1(D) \times H^1(D)$ s.t.

$$a_\rho(v, w) := a(v, w) + \int_{\partial D} \rho \gamma^{\mathrm{g}}(v) \gamma^{\mathrm{g}}(w) \, \mathrm{d}s. \tag{31.21}$$

The boundedness of the trace map (see Theorem 3.10) implies that there is $M_{\gamma^{\mathrm{g}}}$ s.t. $\|\gamma^{\mathrm{g}}(v)\|_{L^2(\partial D)} \leq M_{\gamma^{\mathrm{g}}} \ell_D^{-\frac{1}{2}} \|v\|_{H^1(D)}$. Using the Cauchy–Schwarz inequality yields

$$\int_{\partial D} \rho \gamma^{\mathrm{g}}(v) \gamma^{\mathrm{g}}(w) \, \mathrm{d}s \leq \rho_\sharp M_{\gamma^{\mathrm{g}}}^2 \ell_D^{-1} \|v\|_{H^1(D)} \|w\|_{H^1(D)}$$

M with $\rho_\sharp := \|\rho\|_{L^\infty(\partial D)}$. Since a is bounded on $H^1(D) \times H^1(D)$, so is a_ρ. Similarly, the right-hand side in (31.20) defines a bounded linear form in $H^1(D)$.

We identify $L^2(\partial D)$ with its dual space $L^2(\partial D)'$ in order to interpret the boundary condition satisfied by weak solutions to (31.20). Hence, we have $H^{\frac{1}{2}}(\partial D) \hookrightarrow L^2(\partial D) \equiv L^2(\partial D)' \hookrightarrow H^{-\frac{1}{2}}(\partial D)$ with dense embeddings, where $H^{-\frac{1}{2}}(\partial D)$ is the dual space of $H^{\frac{1}{2}}(\partial D)$. Recall from Theorem 4.15 the normal trace map $\gamma^{\mathrm{d}} : \boldsymbol{H}(\mathrm{div}; D) \to H^{-\frac{1}{2}}(\partial D)$ defined such that the following identity holds true for all $\boldsymbol{\phi} \in \boldsymbol{H}(\mathrm{div}; D)$ and all $w \in H^1(D)$:

$$\langle \gamma^{\mathrm{d}}(\boldsymbol{\phi}), \gamma^{\mathrm{g}}(w) \rangle_{H^{-\frac{1}{2}}, H^{\frac{1}{2}}} = \int_D \left(\boldsymbol{\phi} \cdot \nabla w + (\nabla \cdot \boldsymbol{\phi}) w \right) \mathrm{d}x. \tag{31.22}$$

We have $\gamma^{\mathrm{d}}(\boldsymbol{\phi}) = \boldsymbol{n} \cdot \boldsymbol{\phi}$ whenever $\boldsymbol{\phi}$ is smooth, e.g., if $\boldsymbol{\phi} \in \boldsymbol{H}^s(D)$, $s > \frac{1}{2}$.

Proposition 31.13 (Weak solution). *Let $f \in L^2(D)$, $\rho \in L^\infty(\partial D)$, and $g \in L^2(\partial D)$. If the function $u \in H^1(D)$ solves (31.20), then it satisfies the PDE (31.1) a.e. in D, and the boundary condition (31.19) a.e. in ∂D in the sense that $\rho\gamma^g(u) + \gamma^d(\mathrm{d}\nabla u) = g$ in $L^2(\partial D)$.*

Proof. As in the proof of Proposition 31.6, one can show that the PDE (31.1) is satisfied a.e. in D. In particular, introducing the diffusive flux $\boldsymbol{\sigma} := -\mathrm{d}\nabla u$, we obtain $\boldsymbol{\sigma} \in \boldsymbol{L}^2(D)$ and $\nabla\cdot\boldsymbol{\sigma} = f - \boldsymbol{\beta}\cdot\nabla u - \mu u \in L^2(D)$, i.e., $\boldsymbol{\sigma} \in \boldsymbol{H}(\mathrm{div}; D)$. Using the weak formulation, we infer that

$$-\langle\gamma^d(\boldsymbol{\sigma}),\gamma^g(w)\rangle_{H^{-\frac{1}{2}},H^{\frac{1}{2}}} + \int_{\partial D}(\rho\gamma^g(u) - g)\gamma^g(w)\,\mathrm{d}s = 0,$$

for all $w \in H^1(D)$. Since the trace operator $\gamma^g : H^1(D) \to H^{\frac{1}{2}}(\partial D)$ is surjective and the above equality is valid for all $w \in H^1(D)$, we infer that $\gamma^d(\boldsymbol{\sigma})$ defines a bounded linear form on $L^2(\partial D)$ with Riesz–Fréchet representative equal to $\rho\gamma^g(u) - g$. Hence, the boundary condition is satisfied a.e. on ∂D. $\qquad\square$

Remark 31.14 (Data smoothness). Notice that $f \in L^2(D)$ is needed to establish that $\nabla\cdot\boldsymbol{\sigma} \in L^2(D)$. It is possible to assume that g is only in $H^{-\frac{1}{2}}(\partial D)$. Then the boundary term in (31.20) becomes $\langle g,\gamma^g(w)\rangle_{H^{-\frac{1}{2}},H^{\frac{1}{2}}}$, and the Robin boundary condition is satisfied only in $H^{-\frac{1}{2}}(\partial D)$. $\quad\square$

We now address the well-posedness of (31.20). One can show (see Exercise 31.2 and (3.15)) that there is $\check{C}_{\mathrm{PS}} > 0$ such that for all $v \in H^1(D)$,

$$\check{C}_{\mathrm{PS}}\|v\|_{L^2(D)} \leq \ell_D\|v\|_V,$$
$$\text{with} \quad \|v\|_V := \left\{\|\nabla v\|^2_{\boldsymbol{L}^2(D)} + \ell_D^{-1}\|\gamma^g(v)\|^2_{L^2(\partial D)}\right\}^{\frac{1}{2}}. \tag{31.23}$$

Thus, $(1 + \check{C}_{\mathrm{PS}}^{-2})^{-\frac{1}{2}}\|v\|_{H^1(D)} \leq \ell_D\|v\|_V \leq (1 + M_{\gamma^g}^2)^{\frac{1}{2}}\|v\|_{H^1(D)}$, so that the space $V := H^1(D)$ equipped with the norm $\|v\|_V$ is a Hilbert space. Let $\mu_\flat := \operatorname{ess\,inf}_{\boldsymbol{x}\in D}(\mu - \frac{1}{2}\nabla\cdot\boldsymbol{\beta})(\boldsymbol{x})$ and $\nu_\flat := \operatorname{ess\,inf}_{\boldsymbol{x}\in\partial D}(\rho + \frac{1}{2}\boldsymbol{\beta}\cdot\boldsymbol{n})(\boldsymbol{x})$.

Proposition 31.15 (Coercivity, well-posedness). *Assume that the ellipticity assumption (31.2) holds. Assume that either $\mu_\flat > 0$, $\nu_\flat \geq 0$ or $\mu_\flat \geq 0$, $\nu_\flat > 0$. (i) The bilinear form a_ρ is V-coercive. (ii) The problem (31.20) is well-posed.*

Proof. The boundedness of a_ρ follows from

$$|a_\rho(v,w)| \leq (\lambda_\sharp + \beta_\sharp\check{C}_{\mathrm{PS}}^{-1}\ell_D + \mu_\sharp\check{C}_{\mathrm{PS}}^{-2}\ell_D^2 + \rho_\sharp\ell_D)\|v\|_V\|w\|_V,$$

for all $v, w \in V$. Moreover, the linear form $\ell(w) := \int_D fw\,\mathrm{d}x + \int_{\partial D} g\gamma^g(w)\,\mathrm{d}s$ is bounded on V since $|\ell(w)| \leq (\check{C}_{\mathrm{PS}}^{-1}\ell_D\|f\|_{L^2(D)} + \ell_D^{\frac{1}{2}}\|g\|_{L^2(\partial D)})\|w\|_V$ for all $w \in V$. Let us now prove the V-coercivity of a_ρ. Let $v \in V$. Using (31.16), we infer that

$$a_\rho(v,v) \geq \lambda_\flat\|\nabla v\|^2_{\boldsymbol{L}^2(D)} + \mu_\flat\|v\|^2_{L^2(D)} + \nu_\flat\|\gamma^g(v)\|^2_{L^2(\partial D)}. \tag{31.24}$$

If $\mu_\flat > 0$ and $\nu_\flat \geq 0$, we can drop the term multiplied by $\nu_\flat$ in (31.24), and coercivity follows from

$$a_\rho(v,v) \geq \min(\lambda_\flat, \mu_\flat\ell_D^2)\ell_D^{-2}\|v\|^2_{H^1(D)} \geq \min(\lambda_\flat, \mu_\flat\ell_D^2)(1 + M_{\gamma^g}^2)^{-1}\|v\|^2_V.$$

If $\nu_\flat > 0$ and $\mu_\flat \geq 0$, we can drop the term multiplied by $\mu_\flat$ in (31.24), and coercivity follows from

$$a_\rho(v,v) \geq \min(\lambda_\flat, \nu_\flat\ell_D)\|v\|^2_V.$$

That (31.20) is well-posed follows from the Lax–Milgram lemma. $\qquad\square$

Example 31.16 (Pure diffusion). For a purely diffusive problem, coercivity holds if ρ is uniformly bounded from below away from zero. $\qquad\square$

Remark 31.17 (Variational formulation). Assume that $\boldsymbol{\beta}$ is identically zero in D. Owing to Proposition 25.8, u solves (31.20) iff u minimizes in $H^1(D)$ the energy functional $\mathfrak{E}_{\mathrm{R}}(v) := \frac{1}{2}\int_D \left(\nabla v{\cdot}(\mathbb{d}\nabla v) + \mu v^2 - 2fv\right)\mathrm{d}x + \frac{1}{2}\int_{\partial D}\left(\rho\gamma^{\mathrm{g}}(v)^2 - 2g\gamma^{\mathrm{g}}(v)\right)\mathrm{d}s.$ $\qquad\square$

31.3.2 Neumann condition

The Neumann condition is a particular case of the Robin condition in which ρ vanishes identically on ∂D, i.e., we want to enforce

$$\boldsymbol{n}{\cdot}(\mathbb{d}\nabla u) = g \quad \text{on } \partial D. \tag{31.25}$$

The following weak formulation is obtained by setting ρ to zero in (31.20):

$$\begin{cases} \text{Find } u \in V := H^1(D) \text{ such that} \\ a(u,w) = \int_D fw\,\mathrm{d}x + \int_{\partial D} g\gamma^{\mathrm{g}}(w)\,\mathrm{d}s, \quad \forall w \in V. \end{cases} \tag{31.26}$$

Proposition 31.18 (Weak solution, well-posedness). *Let $f \in L^2(D)$ and $g \in L^2(\partial D)$. (i) If the function $u \in H^1(D)$ solves (31.26), then it satisfies the PDE (31.1) a.e. in D and the boundary condition (31.25) a.e. in ∂D in the sense that $\gamma^{\mathrm{d}}(\mathbb{d}\nabla u) = g$ in $L^2(\partial D)$. (ii) If the ellipticity assumption (31.2) holds true and if $\mu_\flat > 0$ and $\operatorname{ess\,inf}_{\boldsymbol{x}\in\partial D}(\boldsymbol{\beta}{\cdot}\boldsymbol{n})(\boldsymbol{x}) \geq 0$, the bilinear form a is V-coercive. (iii) The problem (31.26) is well-posed.*

Proof. Set $\rho := 0$ in Propositions 31.13 and 31.15. $\qquad\square$

The coercivity assumption invoked in Proposition 31.18 fails when μ and $\boldsymbol{\beta}$ vanish identically in D, i.e., for the purely diffusive problem

$$-\nabla{\cdot}(\mathbb{d}\nabla u) = f \quad \text{in } D, \qquad \boldsymbol{n}{\cdot}(\mathbb{d}\nabla u) = g \quad \text{on } \partial D. \tag{31.27}$$

Indeed, we observe that if u is a solution, then $u + c$ is also a solution for all $c \in \mathbb{R}$. A simple way to deal with this arbitrariness is to restrict the solution space to functions whose mean value over D is zero, i.e., we consider the space $H^1_*(D) := \left\{v \in H^1(D) \mid \underline{v}_D = 0\right\}$ where $\underline{v}_D := |D|^{-1}\int_D v\,\mathrm{d}x$. Note that a necessary condition for a solution to exist is the following compatibility condition on f and g:

$$\int_D f\,\mathrm{d}x + \int_{\partial D} g\,\mathrm{d}s = 0. \tag{31.28}$$

Indeed, (31.22) implies that if (31.27) has a solution u, then $\int_D f\,\mathrm{d}x + \int_{\partial D} g\,\mathrm{d}s = -\int_D \nabla{\cdot}(\mathbb{d}\nabla u)\,\mathrm{d}x + \langle\gamma^{\mathrm{d}}(\mathbb{d}\nabla u), 1\rangle_{H^{-\frac{1}{2}}, H^{\frac{1}{2}}} = 0$.

We now consider the following weak formulation:

$$\begin{cases} \text{Find } u \in V := H^1_*(D) \text{ such that} \\ a_{\mathbb{d}}(u,w) = \int_D fw\,\mathrm{d}x + \int_{\partial D} g\gamma^{\mathrm{g}}(w)\,\mathrm{d}s, \quad \forall w \in V, \end{cases} \tag{31.29}$$

with the bilinear form $a_{\mathbb{d}}(v,w) := \int_D (\mathbb{d}\nabla v){\cdot}\nabla w\,\mathrm{d}x$. Note that the test functions in (31.29) have also zero mean value over D.

Proposition 31.19 (Well-posedness). *Let $f \in L^2(D)$ and $g \in L^2(\partial D)$ satisfy (31.28). (i) If the function $u \in H^1_*(D)$ solves (31.29), then it satisfies the PDE (31.27) a.e. in D and the boundary condition a.e. in ∂D in the sense that $\gamma^{\mathrm{d}}(\mathbb{d}\nabla u) = g$ in $L^2(\partial D)$. (ii) Under the ellipticity condition (31.2), $a_{\mathbb{d}}$ is V-coercive. (iii) The problem (31.29) is well-posed.*

Proof. See Exercise 31.4. ☐

Recall from (4.12) that the normal trace operator $\gamma^{\mathrm{d}} : \boldsymbol{H}(\mathrm{div}; D) \to H^{-\frac{1}{2}}(\partial D)$ is defined by setting $\langle \gamma^{\mathrm{d}}(\boldsymbol{v}), \gamma^{\mathrm{g}}(w) \rangle_{\partial D} := \int_D (\boldsymbol{v} \cdot \nabla w + w \nabla \cdot \boldsymbol{v}) \, \mathrm{d}x$ for all $w \in H^1(D)$, where $\langle \cdot, \cdot \rangle_{\partial D}$ denotes the duality pairing between $H^{-\frac{1}{2}}(\partial D)$ and $H^{-\frac{1}{2}}(\partial D)$. This definition makes sense since the full trace operator $\gamma^{\mathrm{g}} : H^1(D) \to H^{\frac{1}{2}}(\partial D)$ is surjective (see Theorem 3.10(iii)) and $\int_D (\boldsymbol{v} \cdot \nabla w + w \nabla \cdot \boldsymbol{v}) \, \mathrm{d}x = 0$ for all $w \in H^1_0(D)$ and all $\boldsymbol{v} \in \boldsymbol{H}(\mathrm{div}; D)$, i.e., we have $\langle \gamma^{\mathrm{d}}(\boldsymbol{v}), \gamma^{\mathrm{g}}(w_1) \rangle_{\partial D} = \langle \gamma^{\mathrm{d}}(\boldsymbol{v}), \gamma^{\mathrm{g}}(w_2) \rangle_{\partial D}$ if $\gamma^{\mathrm{g}}(w_1) = \gamma^{\mathrm{g}}(w_2)$.

Corollary 31.20 (Surjectivity of normal trace operator). *Let D be a Lipschitz domain. The normal trace operator $\gamma^{\mathrm{d}} : \boldsymbol{H}(\mathrm{div}; D) \to H^{-\frac{1}{2}}(\partial D)$ is surjective.*

Proof. Let $a_{\mathrm{n}} \in H^{-\frac{1}{2}}(\partial D)$ and $\phi \in H^1(D)$ solve $\int_D (\nabla \phi \cdot \nabla w + \ell_D^{-2} \phi w) \, \mathrm{d}x = \langle a_{\mathrm{n}}, \gamma^{\mathrm{g}}(w) \rangle_{\partial D}$ for all $w \in H^1(D)$. We have seen above that this problem has a unique solution. Since $\int_D (\nabla \phi \cdot \nabla w + \ell_D^{-2} \phi w) \, \mathrm{d}x = 0$ for all $w \in C_0^\infty(D)$, we infer that $\Delta \phi = \ell_D^{-2} \phi$ a.e. in D. Hence, $\nabla \phi \in \boldsymbol{H}(\mathrm{div}; D)$. Moreover, $\langle \gamma^{\mathrm{d}}(\nabla \phi), \gamma^{\mathrm{g}}(w) \rangle_{\partial D} := \int_D (\nabla \phi \cdot \nabla w + w \Delta \phi) \, \mathrm{d}x = \int_D (\nabla \phi \cdot \nabla w + \ell_D^{-2} w \phi) \, \mathrm{d}x = \langle a_{\mathrm{n}}, \gamma^{\mathrm{g}}(w) \rangle_{\partial D}$ for all $w \in H^1(D)$. This proves that $\langle \gamma^{\mathrm{d}}(\nabla \phi) - a_{\mathrm{n}}, l \rangle_{\partial D} = 0$ for all $l \in H^{\frac{1}{2}}(\partial D)$ since γ^{g} is surjective. In conclusion, we have established $\gamma^{\mathrm{d}}(\nabla \phi) = a_{\mathrm{n}}$, i.e., γ^{d} is surjective. ☐

31.3.3 Mixed Dirichlet–Neumann conditions

It is possible to combine the Dirichlet and the Neumann conditions. Let ∂D_{d} be a closed subset of ∂D and set $\partial D_{\mathrm{n}} := \partial D \backslash \partial D_{\mathrm{d}}$. We assume that both subsets ∂D_{d} and ∂D_{n} have positive (surface) measures, and we enforce a Dirichlet and a Neumann condition on ∂D_{d} and ∂D_{n}, respectively:

$$u = g_{\mathrm{d}} \text{ on } \partial D_{\mathrm{d}}, \qquad \boldsymbol{n} \cdot (\mathrm{d} \nabla u) = g_{\mathrm{n}} \text{ on } \partial D_{\mathrm{n}}, \tag{31.30}$$

with g_{d} and g_{n} defined on ∂D_{d} and ∂D_{n}, respectively. We assume that there exists a bounded extension operator $H^{\frac{1}{2}}(\partial D_{\mathrm{d}}) \to H^{\frac{1}{2}}(\partial D)$, i.e., there exists $C_{\partial D_{\mathrm{d}}} > 0$ s.t. for all $\alpha \in H^{\frac{1}{2}}(\partial D_{\mathrm{d}})$, there is $\check{\alpha} \in H^{\frac{1}{2}}(\partial D)$ s.t. $\check{\alpha}_{|\partial D_{\mathrm{d}}} := \alpha$ and $C_{\partial D_{\mathrm{d}}} \|\check{\alpha}\|_{H^{\frac{1}{2}}(\partial D)} \leq \|\alpha\|_{H^{\frac{1}{2}}(\partial D_{\mathrm{d}})}$. Owing to Theorem 2.30, this assumption holds true if the interface between ∂D_{d} and ∂D_{n} is Lipschitz. Then let $\check{u}_{\mathrm{d}} \in H^1(D)$ be s.t. $\gamma^{\mathrm{g}}(\check{u}_{\mathrm{d}}) = \check{g}_{\mathrm{d}}$ and let $V := \{v \in H^1(D) \mid \gamma^{\mathrm{g}}(v) = 0 \text{ a.e. on } \partial D_{\mathrm{d}}\}$. Consider the weak formulation:

$$\begin{cases} \text{Find } u_0 \in V \text{ such that} \\ a(u_0, w) = \int_D f w \, \mathrm{d}x + \int_{\partial D_{\mathrm{n}}} g_{\mathrm{n}} \gamma^{\mathrm{g}}(w) \, \mathrm{d}s - a(\check{u}_{\mathrm{d}}, w), \quad \forall w \in V. \end{cases} \tag{31.31}$$

Let $\widetilde{H}^{\frac{1}{2}}(\partial D_{\mathrm{n}}) := \{v \in H^{\frac{1}{2}}(\partial D_{\mathrm{n}}) \mid \widetilde{v} \in H^{\frac{1}{2}}(\partial D)\}$, where $\widetilde{v}$ is the zero-extension of v to ∂D.

Proposition 31.21 (Well-posedness). *Let $f \in L^2(D)$, $g_{\mathrm{d}} \in H^{\frac{1}{2}}(\partial D_{\mathrm{d}})$, and $g_n \in L^2(\partial D_{\mathrm{n}})$.* (i) *Under the above assumptions, if the function $u_0 \in H^1_0(D)$ solves (31.31), then the function $u := u_0 + \check{u}_{\mathrm{d}} \in H^1(D)$ satisfies the PDE (31.1) a.e. in D. It also satisfies the Dirichlet condition a.e. on ∂D_{d} and the Neumann condition a.e. on ∂D_{n} in the sense that $\langle \gamma^{\mathrm{d}}(\mathrm{d} \nabla u), \widetilde{v} \rangle_{H^{-\frac{1}{2}}, H^{\frac{1}{2}}} = \int_{\partial D_{\mathrm{n}}} gv \, \mathrm{d}s$ for all $v \in \widetilde{H}^{\frac{1}{2}}(\partial D_{\mathrm{n}})$.* (ii) *The problem (31.31) is well-posed under the assumptions of Proposition 31.8.*

Proof. We only sketch the proof.
(i) That $u := u_0 + \check{u}_{\mathrm{d}}$ satisfies the PDE in D is shown as above. The Dirichlet condition results from $\gamma^{\mathrm{g}}(u)_{|\partial D_{\mathrm{d}}} = \gamma^{\mathrm{g}}(u_0)_{|\partial D_{\mathrm{d}}} + \gamma^{\mathrm{g}}(\check{u}_{\mathrm{d}})_{|\partial D_{\mathrm{d}}} = \gamma^{\mathrm{g}}(\check{u}_{\mathrm{d}})_{|\partial D_{\mathrm{d}}} = \check{g}_{\mathrm{d}|\partial D_{\mathrm{d}}} = g_{\mathrm{d}}$. To obtain the Neumann condition, we observe that for all $v \in \widetilde{H}^{\frac{1}{2}}(\partial D_{\mathrm{n}})$, there is $w \in V$ s.t. $\gamma^{\mathrm{g}}(w) = \widetilde{v}$. Using w as a test

function in (31.31), we infer that $\langle \gamma^{\mathrm{d}}(\mathrm{d}\nabla u), \tilde{v}\rangle_{H^{-\frac{1}{2}}, H^{\frac{1}{2}}} = \int_{\partial D_{\mathrm{n}}} gv\,\mathrm{d}s$.

(ii) To prove the well-posedness of (31.31), we first notice that V is a closed subspace of $H^1(D)$. Indeed, if $(v_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in V, then $v_n \to v$ in $H^1(D)$ as $n \to \infty$. This implies that $\gamma^{\mathrm{g}}(v) = 0$ a.e. on ∂D_{d} since $\gamma^{\mathrm{g}}(v_n) \to \gamma^{\mathrm{g}}(v)$ in $H^{\frac{1}{2}}(\partial D)$. To conclude the proof, we use the following Poincaré–Steklov inequality in V: There is $\tilde{C}_{\mathrm{PS}} > 0$ such that $\tilde{C}_{\mathrm{PS}}\|v\|_{L^2(D)} \leq \ell_D \|\nabla v\|_{\boldsymbol{L}^2(D)}$ for all $v \in V$. This inequality is a consequence of Lemma 3.30 applied with the linear form $f(v) := \int_{\partial D_{\mathrm{d}}} \gamma^{\mathrm{g}}(v)\,\mathrm{d}s$ and $p := 2$ (notice that $V \ni v \longmapsto \int_{\partial D_{\mathrm{d}}} \gamma^{\mathrm{g}}(v)\,\mathrm{d}s$ restricted to constant functions is nonzero since ∂D_{d} has positive measure). $\qquad\square$

Remark 31.22 (Data in $\widetilde{H}^{\frac{1}{2}}(\partial D_{\mathrm{n}})'$). The weak formulation (31.31) still makes sense if the boundary integral $\int_{\partial D_{\mathrm{n}}} g_{\mathrm{n}}\gamma^{\mathrm{g}}(w)\,\mathrm{d}s$ is replaced by $g_{\mathrm{n}}(\gamma^{\mathrm{g}}(w)_{|\partial D_{\mathrm{n}}})$ where $g_{\mathrm{n}} \in \widetilde{H}^{\frac{1}{2}}(\partial D_{\mathrm{n}})'$, since the map $V \ni w \mapsto \gamma^{\mathrm{g}}(w)_{|\partial D_{\mathrm{n}}} \in \widetilde{H}^{\frac{1}{2}}(\partial D_{\mathrm{n}})$ is bounded. $\qquad\square$

Remark 31.23 ($\widetilde{H}^{\frac{1}{2}}(\partial D_{\mathrm{n}})$ vs. $H_{00}^{\frac{1}{2}}(\partial D_{\mathrm{n}})$). In the literature, the interpolation space $H_{00}^{\frac{1}{2}}(\partial D_{\mathrm{n}})$ introduced in Lions and Magenes [286, Thm. 11.7] is sometimes invoked instead of $\widetilde{H}^{\frac{1}{2}}(\partial D_{\mathrm{n}})$. More precisely, if U is a Lipschitz domain in $\mathbb{R}^{d'}$ (think of $d' := d - 1$), we define $H_{00}^{\frac{1}{2}}(U) := [L^2(U), H_0^1(U)]_{\frac{1}{2}, 2}$ (see Definition A.22). Then $H_{00}^{\frac{1}{2}}(U) \hookrightarrow \widetilde{H}^{\frac{1}{2}}(U)$ follows from Theorem A.27, since the zero-extension operator maps boundedly $L^2(U)$ to $L^2(\mathbb{R}^d)$ and $H_0^1(D)$ to $H^1(\mathbb{R}^d)$ (since $1 - \frac{1}{2} \notin \mathbb{N}$). Moreover, as observed in [286, Thm. 11.7] and Tartar [362, p. 160], "$H_{00}^{\frac{1}{2}}(U)$ is characterized as the space of functions u in $H^{\frac{1}{2}}(U)$ such that $u/\sqrt{d(\boldsymbol{x})} \in L^2(U)$, where $d(\boldsymbol{x})$ is the distance to the boundary ∂U", which according to Theorem 3.18 is also the characterization of $\widetilde{H}^{\frac{1}{2}}(U)$. Hence, $\widetilde{H}^{\frac{1}{2}}(U) = H_{00}^{\frac{1}{2}}(U)$. We also refer the reader to Chandler-Wilde et al. [115, Cor. 4.10], where it is shown that $\{\widetilde{H}^s(U) \mid s \in \mathbb{R}\}$ is an interpolation scale (i.e., for all $s_1 < s_2$ and all $s \in (s_1, s_2)$, we have $\widetilde{H}^s(U) = [\widetilde{H}^{s_1}(U), \widetilde{H}^{s_2}(U)]_{\theta, 2}$ with $\theta := (s - s_1)/(s_2 - s_1)$). The above argument leads us to conjecture that the spaces $\widetilde{H}^{\frac{1}{2}}(\partial D_{\mathrm{n}})$ and $H_{00}^{\frac{1}{2}}(\partial D_{\mathrm{n}})$ are identical provided the interface between ∂D_{n} and ∂D_{d} is smooth enough. Since we do not know any precise result from the literature establishing this equality, we prefer to work with the space $\widetilde{H}^{\frac{1}{2}}(\partial D_{\mathrm{n}})$. $\qquad\square$

31.4 Elliptic regularity

The solution space V for scalar second-order elliptic PDEs is such that $H_0^1(D) \subseteq V \subseteq H^1(D)$ depending on the type of boundary condition that is enforced. Since functions in V may not have weak second-order derivatives, a natural question is whether it is possible to prove that the weak solution enjoys higher regularity. The *elliptic regularity* theory provides theoretical results allowing one to assert that under suitable assumptions on the smoothness of the domain and the data, the weak solution sits indeed in a Sobolev space with higher regularity, e.g., in $H^{1+r}(D)$ with $r > 0$. We say that r is the index of *elliptic regularity pickup*. These results are important for the finite element error analysis since convergence rates depend on the smoothness of the weak solution. In this section, we consider elliptic regularity results in the interior of the domain and then up to the boundary, with a particular attention paid to the case of Lipschitz domains. Most of the results are just stated and we provide pointers to the literature for the proofs.

Besides the hypotheses on the PDE coefficients from §31.1.1, we implicitly assume that the lower-order terms $\boldsymbol{\beta}$ and μ are s.t. the advection-reaction term $\boldsymbol{\beta}\cdot\nabla v + \mu v$ has the same smoothness as that requested for the source f for all $v \in H^1(D)$. For instance, when we assume $f \in L^2(D)$,

we also implicitly assume that $\boldsymbol{\beta} \in \boldsymbol{L}^\infty(D)$ and $\mu \in L^r(D)$, with $r > 2$ and $r \geq d$, so that $\boldsymbol{\beta} \cdot \nabla v + \mu v \in L^2(D)$ for all $v \in H^1(D)$.

31.4.1 Interior regularity

We first present a general result concerning interior regularity, i.e., regularity in any subset $S \subset\subset D$ (meaning that $\overline{S} \subsetneq D$). Notice that we do not make any assumption on the boundary condition satisfied by the weak solution or on the smoothness of D.

Theorem 31.24 (Interior regularity). *Let D be a bounded open set. Assume that $\mathrm{d} \in \mathbb{C}^1(\overline{D})$ and $f \in L^2(D)$. Let $u \in H^1(D)$ be any of the above weak solutions. Then for every open subset $S \subset\subset D$, there are C_1, C_2 (depending on S, D, and the PDE coefficients) such that*

$$\|u\|_{H^2(S)} \leq C_1 \|f\|_{L^2(D)} + C_2 \|u\|_{L^2(D)}. \tag{31.32}$$

Proof. See Evans [196, §6.3.1]. The main tool for the proof is the technique of difference quotients by Nirenberg [312], Agmon et al. [6]. $\qquad\square$

Remark 31.25 (Sharper bound). If the weak formulation is well-posed, the bound (31.32) takes the form $\|u\|_{H^2(S)} \leq C \|f\|_{L^2(D)}$ owing to the a priori estimate $\|u\|_{H^1(D)} \leq C' \|f\|_{L^2(D)}$. $\quad\square$

Remark 31.26 (Higher-order interior regularity). Let m be a nonnegative integer. Assume that $\mathrm{d} \in \mathbb{C}^{m+1}(\overline{D})$, that the coefficients $\{\beta_i\}_{i \in \{1:d\}}$ and μ are in $C^m(\overline{D})$, and that $f \in H^m(D)$. Let $u \in H^1(D)$ be any of the above weak solutions. Then for every open subset $S \subset\subset D$, there are C_1, C_2 (depending on S, D, m, and the PDE coefficients) s.t. $\|u\|_{H^{m+2}(S)} \leq C_1 \|f\|_{H^m(D)} + C_2 \|u\|_{L^2(D)}$; see [196, §6.3.1]. $\qquad\square$

31.4.2 Regularity up to the boundary

We are now concerned with the smoothness of the weak solution up to the boundary. In this context, the smoothness of ∂D and the nature of the boundary condition enforced on ∂D play a role. The following theorems gather results established over the years by many authors. We refer the reader to the textbooks by Grisvard [223, 224], Dauge [152] for more detailed presentations. We consider three situations: domains having a smooth boundary, convex domains, and Lipschitz domains. In what follows, we assume that the weak formulations are well-posed.

Theorem 31.27 (Smooth domain). *Let D be a domain in $\mathbb{R}^d$ with a $C^{1,1}$-boundary. Assume that d is Lipschitz in D, i.e., there is L s.t.*

$$\|\mathrm{d}(\boldsymbol{x}) - \mathrm{d}(\boldsymbol{y})\|_{\ell^2(\mathbb{R}^{d \times d})} \leq L \|\boldsymbol{x} - \boldsymbol{y}\|_{\ell^2(\mathbb{R}^d)}, \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \overline{D}. \tag{31.33}$$

Let $p \in (1, \infty)$ and assume that $f \in L^p(D)$. (i) The weak solution to the Dirichlet problem with boundary data $g \in W^{2-\frac{1}{p}, p}(\partial D)$ is in $W^{2,p}(D)$. (ii) The weak solution to the Neumann problem with boundary data $g \in W^{1-\frac{1}{p}, p}(\partial D)$ is in $W^{2,p}(D)$. The same conclusion holds true for the Robin problem if ρ is Lipschitz on ∂D.

Proof. See [223, Thm. 2.4.2.5-2.4.2.7] (see also [196, §6.3.2] for the Dirichlet problem and $p := 2$). $\qquad\square$

Remark 31.28 (Neumann problem). Elliptic regularity for the Neumann problem is often established under the assumptions of Proposition 31.18, i.e., the coefficient μ is uniformly bounded from below away from zero. The Neumann problem (31.27) with the compatibility condition (31.28)

can be treated by observing that if u is the weak solution to this problem, then u is also the weak solution to the Neumann problem set in $H^1(D)$ with the coefficient $\mu := \mu_0 > 0$ and the source term f replaced by $f + \mu_0 u$, where μ_0 is any nonzero constant with the appropriate units.   □

Theorem 31.29 (Higher-order regularity). *Let m be a positive integer. Assume that ∂D is of class $C^{m+1,1}$, $\mathrm{d} \in \mathbb{C}^{m,1}(D)$, and $f \in W^{m,p}(D)$. Assume that the coefficients $\{\beta_i\}_{i \in \{1:d\}}$ and μ are in $C^m(\overline{D})$. Then the weak solution to the Dirichlet problem with $g \in W^{m+2-\frac{1}{p},p}(\partial D)$ is in $W^{m+2,p}(D)$. The same conclusion holds true for the Robin and Neumann problems if $g \in W^{m+1-\frac{1}{p},p}(\partial D)$ and $\rho \in C^{m,1}(\partial D)$.*

Proof. See [223, Thm. 2.5.1.1].   □

The smoothness assumption on ∂D can be relaxed if the domain D is convex. Notice that a convex domain is Lipschitz; see [223, Cor. 1.2.2.3].

Theorem 31.30 (Convex domain). *Let D be a convex domain. Assume that d is Lipschitz in D. Let $f \in L^2(D)$. (i) The weak solution to the Dirichlet problem with $g := 0$ is in $H^2(D)$. (ii) The weak solution to the Robin or Neumann problem with $g := 0$ is in $H^2(D)$.*

Proof. See [223, Thm. 3.2.1.2, 3.2.1.3, 3.2.3.1].   □

Elliptic regularity in Lipschitz domains is widely studied in the literature; see, e.g., Kondrat′ev [270], Maz′ja and Plamenevskiĭ [296], Jerison and Kenig [255, 256]. We first consider polygons in $\mathbb{R}^2$ and quote results from [223, Chap. 4].

Theorem 31.31 (Polygon, $\mathrm{d} = \mathbb{I}$). *Let $D \subset \mathbb{R}^2$ be a polygon with boundary vertices $\{S_j\}_{1 \leq j \leq J}$ where the segment joining S_j to S_{j+1} corresponds to the boundary face denoted by F_j (setting conventionally $J + 1 := 1$). Let $\theta_j \in (0, 2\pi)$ be the interior angle formed by the faces F_j and F_{j+1}. Assume that d is the identity matrix and that $\theta_j \neq \pi$ for all $j \in \{1:J\}$. Let $f \in L^2(D)$. (i) There is $s_0 \in (\frac{1}{2}, 1]$ such that the weak solution to the Dirichlet problem enforcing $u_{|F_j} = g_j$, with $g_j \in H^{\frac{3}{2}}(F_j)$ and $g_j(S_j) = g_{j+1}(S_j)$ for all $j \in \{1:J\}$, is in $H^{1+s}(D)$ for all $s \in [0, s_0]$ and $s_0 = 1$ if D is convex. (ii) The same conclusion holds true for Neumann problem enforcing $\frac{\partial u}{\partial n}_{|F_j} = g_j$ with $g_j \in H^{\frac{1}{2}}(F_j)$ for all $j \in \{1:J\}$.*

Proof. See [223, Cor. 4.4.4.14] (which treats mixed Dirichlet–Neumann conditions and L^p-Sobolev spaces). The weak solution is in $H^2(D)$ up to singular perturbations that behave in radial coordinates as $r^{\frac{\pi}{2\theta_j}} \sin(\theta \frac{\pi}{2\theta_j} + \varphi_j)$ in the vicinity of S_j with $\varphi_j \in \mathbb{R}$; see Exercise 31.5.   □

Remark 31.32 (Variable coefficients). This case can be treated by freezing the diffusion tensor at each polygon vertex and applying locally a coordinate transformation to recover the Laplace operator; see [223, §5.2].   □

The analysis of elliptic regularity in a polyhedron is more intricate since vertex, edge, and edge-vertex singularities can occur; see Grisvard [223, §8.2], Dauge [152, §5], Lubuma and Nicaise [288, 289], Nicaise [311], Guo and Babuška [229, 230], Costabel et al. [147, 148]. For $s \in (0, 1)$, let us define the space $H^{-1+s}(D)$ either by interpolation between $L^2(D)$ and $H^{-1}(D)$ or as the dual of $H_0^{1-s}(D)$ (the subspace of $H^{1-s}(D)$ spanned by functions with zero trace on ∂D for $s \in (0, \frac{1}{2})$). These two definitions give the same space with equivalent norms.

Theorem 31.33 (Polyhedron, $\mathrm{d} = \mathbb{I}$, Dirichlet). *Let $D \subset \mathbb{R}^3$ be a Lipschitz polyhedron. There exists $s_0 > \frac{1}{2}$, depending on D, such that the Laplace operator is an isomorphism from $H^{1+s}(D) \cap H_0^1(D)$ to $H^{-1+s}(D)$ for all $s \in [0, s_0]$.*

Proof. This is a consequence of Theorem 18.13 in Dauge [152, p. 158]. □

Theorem 31.34 (Lipschitz domain, Lipschitz diffusion). *Let D be a Lipschitz domain in $\mathbb{R}^d$. Assume that $\mathbb{d}$ is Lipschitz in $\overline{D}$. There is $s_0 \in (0, \frac{1}{2})$ such that the following holds true for all $s \in [0, s_0]$: (i) The weak solution to the Dirichlet problem with $f \in L^2(D)$ and $g \in H^{\frac{1}{2}+s}(\partial D)$ is in $H^{1+s}(D)$. (ii) The weak solution to the Neumann problem with $f \in L^2(D)$ and $g \in H^{-\frac{1}{2}+s}(\partial D)$ is in $H^{1+s}(D)$.*

Proof. See Theorems 3 and 4 in Savaré [341]. Notice also that the lowest-order terms in the PDE are in $L^2(D)$ and that $L^2(D) \subset H^{-1+s}(D)$ for all $s \leq 1$, so that f can be replaced by $f - \boldsymbol{\beta}{\cdot}\nabla u - \mu u$. □

Remark 31.35 (Very weak solution). It is possible to extend the notion of elliptic regularity to the very weak solutions. Such solutions do not necessarily belong to the space $H^1(D)$. For instance, using the transposition technique from Lions and Magenes [286, Chap. 2], it is shown in Savaré [341] that the statement of Theorem 31.34 also holds true for all $s \in (-\frac{1}{2}, 0)$. □

The Lipschitz property of $\mathbb{d}$ is rather restrictive since it excludes domains composed of different materials. Following Jochmann [258], it is possible to replace this hypothesis by a (usually called) *multiplier assumption*, which consists of assuming that there is $s_0 \in (0, \frac{1}{2})$ such that

$$\text{the map } \boldsymbol{H}^{s_0}(D) \ni \boldsymbol{\xi} \longmapsto \mathbb{d}\boldsymbol{\xi} \in \boldsymbol{H}^{s_0}(D) \text{ is bounded.} \tag{31.34}$$

It is shown in Jochmann [258, Lem. 2] (see also Bonito et al. [70, Prop. 2.1]) that this property holds true if D is partitioned into M disjoint Lipschitz subdomains $\{D_m\}_{m\in\{1:M\}}$ and if there is a real number $\alpha \in (s_0, 1]$ and there are diffusion tensors $\mathbb{d}_m \in \mathbb{C}^{0,\alpha}(D_m)$ for all $m \in \{1{:}M\}$, s.t. $\mathbb{d} := \sum_{m\in\{1:M\}} \mathbb{1}_{D_m}\mathbb{d}_m$, where $\mathbb{1}_{D_m}$ is the indicator function of D_m.

Theorem 31.36 (Piecewise smooth diffusion). *Assume that there is $s_0 \in (0, \frac{1}{2})$ such that the multiplier assumption (31.34) holds true. Then there is $s \in (0, s_0)$, depending on D and $\mathbb{d}$, s.t. the weak solution to the homogeneous Dirichlet problem or to the Neumann problem with $f \in L^2(D)$ (and $g := 0$) is in $H^{1+s}(D)$.*

Proof. See Theorem 3 in [258] or Lemma 3.2 in [70]. The statement also holds true for f in the dual space of $H_0^{1-s}(D)$ for the Dirichlet problem and for f in the dual space of $H^{1-s}(D)$ for the Neumann problem. See also Bernardi and Verfürth [55] for Dirichlet conditions and piecewise constant (or pcw. twice continuously differentiable) isotropic diffusion. □

Theorem 31.36 also holds true for the mixed Dirichlet–Neumann problem. We refer the reader to Jochmann [258] for more details on this question.

Exercises

Exercise 31.1 (Cordes). Prove that ellipticity implies the Cordes condition if $d = 2$. (*Hint*: use that $\|\mathbb{d}\|_F^2 = (\text{tr}(\mathbb{d}))^2 - 2\det(\mathbb{d})$.)

Exercise 31.2 (Poincaré–Steklov). Prove (31.23). (*Hint*: use (3.12).)

Exercise 31.3 (Potential flow). Consider the PDE $\nabla{\cdot}(-\kappa\nabla u + \boldsymbol{\beta}u) = f$ in D with homogeneous Dirichlet conditions and assume that κ is a positive real number. Assume that $\boldsymbol{\beta} := \nabla\psi$ for some smooth function ψ (we say that $\boldsymbol{\beta}$ is a potential flow). Find a functional $\mathfrak{E} : H_0^1(D) \to \mathbb{R}$ of which the weak solution u is a minimizer on $H_0^1(D)$. (*Hint*: consider the function $e^{-\psi/\kappa}u$.)

Exercise 31.4 (Purely diffusive Neumann). Prove Proposition 31.19. (*Hint*: for all $w \in H^1(D)$, the function $\tilde{w} := w - \underline{w}_D$ is in $H^1_*(D)$, use also the Poincaré–Steklov inequality from Lemma 3.24.)

Exercise 31.5 (Mixed Dirichlet–Neumann). The goal is to show by a counterexample that one cannot assert that the weak solution is in $H^2(D)$ for the mixed Dirichlet–Neumann problem even if the domain and the boundary data are smooth. Using polar coordinates, set $D := \{(r, \theta) \in (0, 1) \times (0, \pi)\}$, $\partial D_{\mathrm{n}} := \{r \in (0, 1), \ \theta = \pi\}$, and $\partial D_{\mathrm{d}} := \partial D \backslash \partial D_{\mathrm{n}}$. Verify that the function $u(r, \theta) := r^{\frac{1}{2}} \sin(\frac{1}{2}\theta)$ satisfies $-\Delta u = 0$ in D, $\frac{\partial u}{\partial n}|_{D_{\mathrm{n}}} = 0$, and $u_{|D_{\mathrm{d}}} = r^{\frac{1}{2}} \sin(\frac{1}{2}\theta)$. (*Hint*: in polar coordinates, $\Delta u = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2}$.) Verify that $u \notin H^2(D)$.

Exercise 31.6 ($H^2(\mathbb{R}^d)$-seminorm). Prove that $|\phi|_{H^2(\mathbb{R}^d)} = \|\Delta\phi\|_{L^2(\mathbb{R}^d)}$ for all $\phi \in C_0^\infty(\mathbb{R}^d)$. (*Hint*: use Theorem B.3.)

Exercise 31.7 (Counterexample to elliptic regularity in $W^{2,\infty}(D)$). Let D be the unit disk in $\mathbb{R}^2$. Consider the function $u(x_1, x_2) := x_1 x_2 \ln(r)$ with $r^2 := x_1^2 + x_2^2$ (note that $u_{|\partial D} = 0$). Verify that $\Delta u \in L^\infty(D)$, but that $u \notin W^{2,\infty}(D)$. (*Hint*: consider the cross-derivative.)

Exercise 31.8 (Domain with slit). Let $D := \{r \in (0, 1), \ \theta \in (0, 2\pi)\}$, where (r, θ) are the polar coordinates, i.e., $\overline{D}$ is the closed ball of radius 1 centered at 0. Let $u(r, \theta) := r \cos(\frac{1}{2}\theta)$ for all $r > 0$ and $\theta \in [0, 2\pi)$. (i) Let $p \in [1, \infty)$. Is $u_{|D}$ in $W^{1,p}(D)$? Is $u_{|\mathrm{int}(\overline{D})}$ in $W^{1,p}(\mathrm{int}(\overline{D}))$? (*Hint*: recall Example 4.3.) (ii) Is the restriction to D of the functions in $C^1(\overline{D})$ dense in $W^{1,p}(D)$? (*Hint*: argue by contradiction and use that $\|v_{|D}\|_{W^{1,p}(D)} = \|v_{|\mathrm{int}(\overline{D})}\|_{W^{1,p}(\mathrm{int}(\overline{D}))}$ for all $v \in C^1(\overline{D})$.)

Exercise 31.9 (A priori estimate). Consider the PDE $-\kappa_0\Delta u + \boldsymbol{\beta}\cdot\nabla u + \mu_0 u = f$ with homogeneous Dirichlet conditions. Assume that $\kappa_0, \mu_0 \in \mathbb{R}$, $\kappa_0 > 0$, $\nabla\cdot\boldsymbol{\beta} = 0$, $\boldsymbol{\beta}_{|\partial D} = \mathbf{0}$, and $f \in H_0^1(D)$. Let $\nabla_{\mathrm{s}}\boldsymbol{\beta} := \frac{1}{2}(\nabla\boldsymbol{\beta} + (\nabla\boldsymbol{\beta})^{\mathsf{T}})$ denote the symmetric part of the gradient of $\boldsymbol{\beta}$, and assume that there is $\mu_0' > 0$ s.t. $\nabla_{\mathrm{s}}\boldsymbol{\beta} + \mu_0\mathbb{I}_d \geq \mu_0'\mathbb{I}_d$ in the sense of quadratic forms. Prove that $|u|_{H^1(D)} \leq (\mu_0')^{-1}|f|_{H^1(D)}$ and $\|\Delta u\|_{L^2(D)} \leq (4\mu_0'\kappa_0)^{-\frac{1}{2}}|f|_{H^1(D)}$. (*Hint*: use $-\Delta u$ as a test function.) *Note*: these results are established in Beirão da Veiga [49], Burman [97].

Exercise 31.10 (Complex-valued diffusion). Assume that the domain D is partitioned into two disjoint subdomains D_1 and D_2. Let κ_1, κ_2 be two complex numbers, both with positive modulus and such that $\frac{\kappa_1}{\kappa_2} \notin \mathbb{R}_-$. Set $\kappa(x) := \kappa_1\mathbb{1}_{D_1}(\boldsymbol{x}) + \kappa_2\mathbb{1}_{D_2}(\boldsymbol{x})$ for all $\boldsymbol{x} \in D$. Let $f \in L^2(D)$. Show that the problem of seeking $u \in V := H_0^1(D; \mathbb{C})$ such that $a(u, w) := \int_D \kappa\nabla u\cdot\nabla\overline{w}\,\mathrm{d}x = \int_D f\overline{w}\,\mathrm{d}x$ for all $w \in V$ is well-posed. (*Hint*: use (25.7).)

Exercise 31.11 (Dependence on diffusion coefficient). Consider two numbers $0 < \lambda_\flat \leq \lambda_\sharp < \infty$ and define the set $K := \{\kappa \in L^\infty(D; \mathbb{R}) \mid \kappa(\boldsymbol{x}) \in [\lambda_\flat, \lambda_\sharp], \text{a.e. } \boldsymbol{x} \in D\}$. Let $V := H_0^1(D)$ equipped with the norm $\|v\|_V := \|\nabla v\|_{\boldsymbol{L}^2(D)}$ and $V' = H^{-1}(D)$. Consider the operator $T_\kappa : V \to V'$ s.t. $T_\kappa(v) := -\nabla\cdot(\kappa\nabla v)$ for all $v \in V$ and all $\kappa \in K$. (i) Prove that $\lambda_\flat \leq \|T_\kappa\|_{\mathcal{L}(V;V')} \leq \lambda_\sharp$ and that T_κ is an isomorphism. (*Hint*: use Proposition 31.8 with $\theta := 1$ and the bilinear form $a(v, w) := \int_D \kappa\nabla v\cdot\nabla w\,\mathrm{d}x$ on $V \times V$.) (ii) Prove that $\|T_\kappa - T_{\kappa'}\|_{\mathcal{L}(V;V')} = \|\kappa - \kappa'\|_{L^\infty(D)}$ for all $\kappa, \kappa' \in K\cap C^0(D; \mathbb{R})$. (*Hint*: if $\|\kappa - \kappa'\|_{L^\infty(D)} > 0$, for all $\epsilon > 0$ there is an open subset $D_\epsilon \subset D$ such that the sign of $(\kappa - \kappa')_{|D_\epsilon}$ is constant and $|\kappa - \kappa'| \geq \|\kappa - \kappa'\|_{L^\infty(D)} - \epsilon$ in D_ϵ; then consider functions in $H_0^1(D_\epsilon)$.) (iii) Let $S_\kappa := T_\kappa^{-1} \in \mathcal{L}(V'; V)$. Prove that $\lambda_\flat^2\|S_\kappa - S_{\kappa'}\|_{\mathcal{L}(V';V)} \leq \|\kappa - \kappa'\|_{L^\infty(D)} \leq \lambda_\sharp^2\|S_\kappa - S_{\kappa'}\|_{\mathcal{L}(V';V)}$ for all $\kappa, \kappa' \in K \cap C^0(D; \mathbb{R})$. (*Hint*: $S_\kappa - S_{\kappa'} = S_\kappa(T_{\kappa'} - T_\kappa)S_{\kappa'}$.)

Chapter 32

H^1-conforming approximation (I)

The goal of this chapter is to analyze the approximation of second-order elliptic PDEs using H^1-conforming finite elements. We focus the presentation on homogeneous Dirichlet boundary conditions for simplicity. The well-posedness of the discrete problem follows from the Lax–Milgram lemma and the error estimate in the H^1-norm from Céa's lemma. We also introduce a duality argument due to Aubin and Nitsche to derive an improved error estimate in the (weaker) L^2-norm. Some further topics on the H^1-conforming approximation of second-order elliptic PDEs are covered in the next chapter.

32.1 Continuous and discrete problems

Let D be a Lipschitz domain in $\mathbb{R}^d$ and let $f \in L^2(D)$. We assume for simplicity that D is a polyhedron. The model problem we want to approximate is the homogeneous Dirichlet problem:

$$-\nabla\cdot(\mathbb{d}\nabla u) + \boldsymbol{\beta}\cdot\nabla u + \mu u = f \quad \text{in } D, \qquad u = 0 \quad \text{on } \partial D, \tag{32.1}$$

with $\mathbb{d} \in \mathbb{L}^\infty(D)$, $\boldsymbol{\beta} \in \boldsymbol{W}^{1,\infty}(D)$, $\mu \in L^\infty(D)$. We assume that $\mathbb{d}$ is a symmetric second-order tensor field and that its smallest eigenvalue is uniformly bounded from below by $\lambda_\flat > 0$. We also assume that $(\mu - \frac{1}{2}\nabla\cdot\boldsymbol{\beta})$ takes nonnegative values a.e. in D. The model problem is formulated as follows:

$$\begin{cases} \text{Find } u \in V := H_0^1(D) \text{ such that} \\ a(u, w) = \ell(w), \quad \forall w \in V, \end{cases} \tag{32.2}$$

with the following bilinear and linear forms on $V \times V$ and V, respectively:

$$a(v, w) := \int_D \left((\mathbb{d}\nabla v)\cdot\nabla w + (\boldsymbol{\beta}\cdot\nabla v)w + \mu vw\right) \mathrm{d}x, \qquad \ell(w) := \int_D fw \,\mathrm{d}x.$$

This problem is well-posed owing to the Lax–Milgram lemma. We equip the space V with the norm $\|v\|_V := \|\nabla v\|_{\boldsymbol{L}^2(D)} = |v|_{H^1(D)}$. This is legitimate owing to the Poincaré–Steklov inequality $C_{\mathrm{PS}}\|v\|_{L^2(D)} \le \ell_D\|\nabla v\|_{\boldsymbol{L}^2(D)}$ for all $v \in H_0^1(D)$ (see (3.11) with $p := 2$), where ℓ_D is a characteristic length of D, e.g., $\ell_D := \mathrm{diam}(D)$. With this choice of norm, the coercivity and the boundedness constants of the bilinear form a on $V \times V$ are

$$\alpha := \lambda_\flat, \qquad \|a\| := \lambda_\sharp + \beta_\sharp C_{\mathrm{PS}}^{-1}\ell_D + \mu_\sharp C_{\mathrm{PS}}^{-2}\ell_D^2, \tag{32.3}$$

with $\lambda_\sharp := \|\mathrm{d}\|_{\mathbb{L}^\infty(D)}$, $\beta_\sharp := \|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(D)}$, and $\mu_\sharp := \|\mu\|_{L^\infty(D)}$.

Let $(\mathcal{T}_h)_{h\in\mathcal{H}}$ be a shape-regular sequence of affine meshes so that each mesh covers D exactly. We approximate (32.2) with H^1-conforming finite elements of some degree $k \geq 1$. Let $(\widehat{K}, \widehat{P}, \widehat{\Sigma})$ be the reference finite element, e.g., a $\mathbb{Q}_{k,d}$ Lagrange element if $\widehat{K}$ is a cuboid (see Chapter 6) or a $\mathbb{P}_{k,d}$ Lagrange element or the canonical hybrid element if $\widehat{K}$ is a simplex (see Chapter 7). For all $K \in \mathcal{T}_h$, let $\boldsymbol{T}_K : \widehat{K} \to K$ be the geometric mapping and let $\psi_K^{\mathrm{g}}(v) := v \circ \boldsymbol{T}_K$ be the pullback by the geometric mapping. Let us define the local (polynomial) space $P_K := (\psi_K^{\mathrm{g}})^{-1}(\widehat{P})$. Let $P_k^{\mathrm{b}}(\mathcal{T}_h)$ be the broken finite element space, $P_k^{\mathrm{g}}(\mathcal{T}_h)$ the H^1-conforming subspace, and $P_{k,0}^{\mathrm{g}}(\mathcal{T}_h)$ its zero-trace subspace. Recalling the construction from Chapter 19, we have

$$P_k^{\mathrm{b}}(\mathcal{T}_h) := \{v_h \in L^\infty(D) \mid v_{h|K} \in P_K, \forall K \in \mathcal{T}_h\}, \tag{32.4a}$$

$$P_k^{\mathrm{g}}(\mathcal{T}_h) := \{v_h \in P_k^{\mathrm{b}}(\mathcal{T}_h) \mid [\![v_h]\!]_F = 0, \forall F \in \mathcal{F}_h^\circ\}, \tag{32.4b}$$

$$P_{k,0}^{\mathrm{g}}(\mathcal{T}_h) := \{v_h \in P_k^{\mathrm{g}}(\mathcal{T}_h) \mid v_{h|\partial D} = 0\}, \tag{32.4c}$$

where $\mathcal{F}_h^\circ$ (resp., $\mathcal{F}_h^\partial$) is the collection of the mesh interfaces (resp., boundary faces) and $[\![v_h]\!]_F$ denotes the jump of v_h across F. In other words, $P_{k,0}^{\mathrm{g}}(\mathcal{T}_h)$ is composed of functions that are piecewise in P_K, that are continuous across the mesh interfaces, and that vanish at the boundary. Recalling Theorem 18.8, we have $P_k^{\mathrm{g}}(\mathcal{T}_h) \subset H^1(D)$ and $P_{k,0}^{\mathrm{g}}(\mathcal{T}_h) \subset H_0^1(D)$.

The discrete problem is as follows:

$$\begin{cases} \text{Find } u_h \in V_h := P_{k,0}^{\mathrm{g}}(\mathcal{T}_h) \text{ such that} \\ a(u_h, w_h) = \ell(w_h), \quad \forall w_h \in V_h. \end{cases} \tag{32.5}$$

Since $P_{k,0}^{\mathrm{g}}(\mathcal{T}_h) \subset H_0^1(D)$, this problem is well-posed owing to the Lax–Milgram lemma. Note that we enforce the homogeneous Dirichlet condition in an essential manner in (32.5). An alternative technique weakly enforcing the Dirichlet condition by means of a boundary penalty method is studied in Chapter 37. Using the notation from §26.3.4, we introduce the *discrete solution map* $G_h : V \to V_h$ so that for all $v \in V$,

$$a(G_h(v) - v, w_h) := 0, \qquad \forall w_h \in V_h. \tag{32.6}$$

It follows from the Lax–Milgram lemma that $G_h(v)$ is uniquely defined, and since u solves (32.2), one readily sees that $u_h = G_h(u)$ iff u_h solves (32.5). The main properties of the discrete solution map are investigated in §26.3.4 in the abstract context of Galerkin methods. It is observed therein that G_h is a projection and that $\|G_h\|_{\mathcal{L}(V)} \leq \frac{\|a\|}{\alpha}$.

Remark 32.1 (Variants). One must use the entire space $P_k^{\mathrm{g}}(\mathcal{T}_h)$ to enforce Robin/Neumann conditions; see §31.3. When working with Dirichlet–Neumann conditions (see §31.3.3), one must construct meshes that are compatible with the boundary partition $\partial D = \partial D_{\mathrm{d}} \cup \partial D_{\mathrm{n}}$, i.e., boundary faces cannot be split, they must belong either to ∂D_{d} or to ∂D_{n}. $\qquad\square$

32.2 Error analysis and best approximation in H^1

Let u solve (32.2) and let u_h solve (32.5). Our goal is to bound the approximation error $(u - u_h)$. Recall that $\|v\|_V := \|\nabla v\|_{\boldsymbol{L}^2(D)} = |v|_{H^1(D)}$.

Theorem 32.2 (H^1-error estimate). *The following holds true:*

$$\|\nabla(u - u_h)\|_{\boldsymbol{L}^2(D)} \le \frac{\|a\|}{\alpha} \inf_{v_h \in V_h} \|\nabla(u - v_h)\|_{\boldsymbol{L}^2(D)}, \tag{32.7}$$

with the coercivity and boundedness constants α and $\|a\|$ defined in (32.3). Moreover, $\lim_{h \to 0} \|\nabla(u - u_h)\|_{\boldsymbol{L}^2(D)} = 0$, and assuming $u \in H^{1+r}(D)$ with $r \in (0, k]$, there is c s.t. for all $h \in \mathcal{H}$,

$$\|\nabla(u - u_h)\|_{\boldsymbol{L}^2(D)} \le c \left(\sum_{K \in \mathcal{T}_h} h_K^{2r} |u|_{H^{1+r}(K)}^2 \right)^{\frac{1}{2}} \le c\, h^r |u|_{H^{1+r}(D)}. \tag{32.8}$$

Proof. The bound (32.7) follows from Céa's lemma. We use a density argument and proceed as in §26.3.3 to prove that $\lim_{h \to 0} \|\nabla(u - u_h)\|_{\boldsymbol{L}^2(D)} = 0$. To prove (32.8), we start from (32.7) and estimate the infimum from above by taking $v_h := \mathcal{I}_{h0}^{\mathrm{g,av}}(u)$, where $\mathcal{I}_{h0}^{\mathrm{g,av}} : L^1(D) \to H_0^1(D)$ is the quasi-interpolation operator with zero boundary trace introduced in §22.4.2. Using the estimate (22.29) from Theorem 22.14 (with $m := 1$ and $p := 2$), we infer that $\|\nabla(u - \mathcal{I}_{h0}^{\mathrm{g,av}}(u))\|_{\boldsymbol{L}^2(K)} \le ch_K^r |u|_{H^{1+r}(\check{\mathcal{T}}_K)}$, where $|\cdot|_{H^{1+r}(\check{\mathcal{T}}_K)}^2 := \sum_{K' \in \check{\mathcal{T}}_K} |\cdot|_{H^{1+r}(K')}^2$ and $\check{\mathcal{T}}_K$ is the collection of all the mesh cells sharing at least one vertex with K. We obtain (32.8) by invoking the regularity of the mesh sequence which implies that all the cells in $\check{\mathcal{T}}_K$ have a diameter uniformly equivalent to h_K and that $\mathrm{card}(\check{\mathcal{T}}_K)$ is uniformly bounded. $\qquad\square$

Remark 32.3 (Canonical or Lagrange interpolant). If $1 + r > \frac{d}{2}$, one can also prove (32.8) by replacing v_h in (32.7) by either the canonical interpolant of u or the Lagrange interpolant of u, both with zero boundary trace (see §19.4). This leads to $\|\nabla(u - v_h)\|_{\boldsymbol{L}^2(K)} \le ch_K^r |u|_{H^{1+r}(K)}$ for all $K \in \mathcal{T}_h$, i.e., this argument circumvents the use of the subset $\check{\mathcal{T}}_K$. $\qquad\square$

Remark 32.4 (Condition number). The ratio $\frac{\|a\|}{\alpha}$, which represents the condition number of the bilinear form a (see Remark 25.12), can become very large when the lower-order terms in the PDE (32.1) dominate the diffusive term. One then says that the PDE is *singularly perturbed*. In this situation, one needs to use stabilized finite elements to obtain an accurate approximate solution on a reasonably fine mesh. Examples can be found in Chapter 61. $\qquad\square$

Remark 32.5 ($W^{1,p}$-estimate). The reader is referred to the seminal work by Rannacher and Scott [329] for $W^{1,p}$-error estimates on convex polygonal domains ($d = 2$) and quasi-uniform mesh families with $p \in [2, \infty]$. Extensions to dimension three can be found in Guzmán et al. [234], and extensions to graded meshes can be found in Demlow et al. [159]. $\qquad\square$

Theorem 32.2 shows that the approximation error in the H^1-seminorm is controlled by the best-approximation error of u in V_h in the same norm, that is, by the quantity $\inf_{v_h \in V_h} \|\nabla(u - v_h)\|_{\boldsymbol{L}^2(D)}$. It is therefore interesting to investigate the behavior of this quantity. A question one may ask is whether the broken finite element space $P_k^{\mathrm{b}}(\mathcal{T}_h)$ and its $H_0^1(D)$-conforming counterpart $V_h := P_{k,0}^{\mathrm{g}}(\mathcal{T}_h)$ have the same capacity to approximate a given function $v \in H_0^1(D)$. In other words, did we sacrifice anything in terms of best-approximation error by working with V_h rather than with $P_k^{\mathrm{b}}(\mathcal{T}_h)$? We are going to show that, remarkably, this is not the case.

To better understand the above question, let us look at how the best-approximation errors in V_h and in $P_k^{\mathrm{b}}(\mathcal{T}_h)$ are evaluated for a given function $v \in H_0^1(D)$. When working in V_h, we need to find a function $v_h^{\mathrm{g}} \in V_h$ s.t.

$$\|\nabla(v - v_h^{\mathrm{g}})\|_{\boldsymbol{L}^2(D)}^2 = \min_{v_h \in V_h} \|\nabla(v - v_h)\|_{\boldsymbol{L}^2(D)}^2. \tag{32.9}$$

Since $\|\nabla(v - v_h)\|_{\boldsymbol{L}^2(D)}^2 = \|\nabla v_h\|_{\boldsymbol{L}^2(D)}^2 - 2(\nabla v, \nabla v_h)_{\boldsymbol{L}^2(D)} + \|\nabla v\|_{\boldsymbol{L}^2(D)}^2$ and the function v is kept fixed in our reasoning, we want to minimize over V_h the functional $\mathfrak{E} : V_h \to \mathbb{R}$ defined by $\mathfrak{E}(v_h) := \|\nabla v_h\|_{\boldsymbol{L}^2(D)}^2 - 2(\nabla v, \nabla v_h)_{\boldsymbol{L}^2(D)}$. Owing to Proposition 25.8, this problem has a unique minimizer in V_h characterized by the equations $(\nabla(v_h^{\mathrm{g}} - v), \nabla w_h)_{\boldsymbol{L}^2(D)} = 0$ for all $w_h \in V_h$. (Note that uniqueness follows from the Poincaré–Steklov inequality since $V_h \subset H_0^1(D)$.) In practice, one can find v_h^{g} by inverting the *global* stiffness matrix associated with the *global* shape functions in V_h (see §28.1.1). On the other hand, when working in $P_k^{\mathrm{b}}(\mathcal{T}_h)$, we need to find a function $v_h^{\mathrm{b}} \in P_k^{\mathrm{b}}(\mathcal{T}_h)$ such that

$$\sum_{K \in \mathcal{T}_h} \|\nabla(v - v_h^{\mathrm{b}})\|_{\boldsymbol{L}^2(K)}^2 = \min_{v_h \in P_k^{\mathrm{b}}(\mathcal{T}_h)} \sum_{K \in \mathcal{T}_h} \|\nabla(v - v_h)\|_{\boldsymbol{L}^2(K)}^2. \tag{32.10}$$

(We sum over the mesh cells since functions in $P_k^{\mathrm{b}}(\mathcal{T}_h)$ do not necessarily have a weak gradient in $\boldsymbol{L}^2(D)$.) Since for all $v_h \in P_k^{\mathrm{b}}(\mathcal{T}_h)$ and all $K \neq K' \in \mathcal{T}_h$, the restrictions $v_{h|K}$ and $v_{h|K'}$ can be chosen independently in the local polynomial spaces P_K and $P_{K'}$, we have

$$\sum_{K \in \mathcal{T}_h} \|\nabla(v - v_h^{\mathrm{b}})\|_{\boldsymbol{L}^2(K)}^2 = \sum_{K \in \mathcal{T}_h} \min_{q \in P_K} \|\nabla(v - q)\|_{\boldsymbol{L}^2(K)}^2, \tag{32.11}$$

and thus we need to find a function $v_K^{\mathrm{b}} := v_{h|K}^{\mathrm{b}} \in P_K$ for all $K \in \mathcal{T}_h$ s.t.

$$\|\nabla(v - v_K^{\mathrm{b}})\|_{\boldsymbol{L}^2(K)}^2 = \min_{q \in P_K} \|\nabla(v - q)\|_{\boldsymbol{L}^2(K)}^2. \tag{32.12}$$

Invoking Proposition 25.8, the above argument shows that the function v_K^{b} is such that $(\nabla(v_K^{\mathrm{b}} - v), \nabla q)_{\boldsymbol{L}^2(K)} = 0$ for all $q \in P_K$, and it is therefore uniquely defined up to an additive constant. It is convenient to require that $(v_K^{\mathrm{b}} - v, 1)_{L^2(K)} := 0$. In practice, one finds each function v_K^{b} by inverting the *local* stiffness matrix associated with the *local* shape functions in P_K.

Theorem 32.6 (Best-approximation error). *There is a constant c such that the following two-sided bounds hold true for all $v \in H_0^1(D)$ and all $h \in \mathcal{H}$:*

$$\min_{v_h \in P_k^{\mathrm{b}}(\mathcal{T}_h)} \sum_{K \in \mathcal{T}_h} \|\nabla(v - v_h)\|_{\boldsymbol{L}^2(K)}^2 \leq \min_{v_h \in V_h} \|\nabla(v - v_h)\|_{\boldsymbol{L}^2(D)}^2$$

$$\leq c \min_{v_h \in P_k^{\mathrm{b}}(\mathcal{T}_h)} \sum_{K \in \mathcal{T}_h} \|\nabla(v - v_h)\|_{\boldsymbol{L}^2(K)}^2.$$

Proof. Let $v \in H_0^1(D)$. The first inequality follows from V_h being a subspace of $P_k^{\mathrm{b}}(\mathcal{T}_h)$ and the identity $\sum_{K \in \mathcal{T}_h} \|\nabla(v - v_h)\|_{\boldsymbol{L}^2(K)}^2 = \|\nabla(v - v_h)\|_{\boldsymbol{L}^2(D)}^2$ if $v_h \in V_h$. Let us prove the second inequality. Recalling that the minimizers are denoted by v_h^{g} and v_h^{b} respectively, we need to prove that

$$\|\nabla(v - v_h^{\mathrm{g}})\|_{\boldsymbol{L}^2(D)}^2 \leq c \sum_{K \in \mathcal{T}_h} \|\nabla(v - v_h^{\mathrm{b}})\|_{\boldsymbol{L}^2(K)}^2.$$

Let $\mathcal{J}_{h,0}^{\mathrm{g,av}} : P_k^{\mathrm{b}}(\mathcal{T}_h) \to V_h$ be the averaging operator defined in §22.4.1. Owing to Lemma 22.12 (with $p := 2$, $r := 2$, and $m := 1$) and since $[\![v]\!]_F = 0$ for all $F \in \mathcal{F}_h$ because $v \in H_0^1(D)$, we have

$$\|\nabla(v_h^{\mathrm{b}} - \mathcal{J}_{h,0}^{\mathrm{g,av}}(v_h^{\mathrm{b}}))\|_{\boldsymbol{L}^2(K)} \leq c \, h_K^{-\frac{1}{2}} \sum_{F \in \check{\mathcal{F}}_K} \|[\![v - v_h^{\mathrm{b}}]\!]_F\|_{L^2(F)},$$

where $\check{\mathcal{F}}_K$ is the collection of the mesh faces (interfaces and boundary faces) sharing at least one vertex with K. Since the jump is the difference of the values from both sides of the interface (the

jump is the actual value for the boundary faces), we bound the jump by the triangle inequality. Then we apply the multiplicative trace inequality (12.16) (with $p := 2$) and invoke the local Poincaré–Steklov inequality (12.13) in all the cells having a face in $\check{\mathcal{F}}_K$ (recall that v_h^{b} and v share the same mean value in every mesh cell). This leads to

$$\|\nabla(v_h^{\mathrm{b}} - \mathcal{J}_h^{\mathrm{g,av}}(v_h^{\mathrm{b}}))\|_{\boldsymbol{L}^2(K)} \leq c\, \|\nabla(v - v_h^{\mathrm{b}})\|_{\boldsymbol{L}^2(D_K)},$$

where D_K is the set of the points composing the cells sharing at least one vertex with K. Since $\mathcal{J}_h^{\mathrm{g,av}}(v_h^{\mathrm{b}}) \in V_h$, the minimization property of v_h^{g} over V_h implies that $\|\nabla(v - v_h^{\mathrm{g}})\|_{\boldsymbol{L}^2(D)}^2 \leq \|\nabla(v - \mathcal{J}_h^{\mathrm{g,av}}(v_h^{\mathrm{b}}))\|_{\boldsymbol{L}^2(D)}^2$. Hence, we have

$$\|\nabla(v - v_h^{\mathrm{g}})\|_{\boldsymbol{L}^2(D)}^2 \leq 2 \sum_{K \in \mathcal{T}_h} \left(\|\nabla(v - v_h^{\mathrm{b}})\|_{\boldsymbol{L}^2(K)}^2 + \|\nabla(v_h^{\mathrm{b}} - \mathcal{J}_h^{\mathrm{g,av}}(v_h^{\mathrm{b}}))\|_{\boldsymbol{L}^2(K)}^2 \right)$$

$$\leq 2 \sum_{K \in \mathcal{T}_h} \left(\|\nabla(v - v_h^{\mathrm{b}})\|_{\boldsymbol{L}^2(K)}^2 + c\, \|\nabla(v - v_h^{\mathrm{b}})\|_{\boldsymbol{L}^2(D_K)}^2 \right).$$

We conclude by invoking the regularity of the mesh sequence. $\qquad\square$

Remark 32.7 (Literature). Theorem 32.6 is due to Veeser [371]. The present proof makes a direct use of the averaging operator from §22.4.1. $\qquad\square$

32.3 L^2-error analysis: the duality argument

The goal of this section is to derive an improved error estimate in a norm that is weaker than that of $V := H_0^1(D)$. This type of estimate is important, in particular, in the approximation of eigenvalue problems (see Chapter 48). More precisely, the question we want to investigate is whether it is possible to find some exponent $\gamma > 0$, uniform w.r.t. $h \in \mathcal{H}$ and u, such that $\|u - u_h\|_{L^2(D)} \leq c h^\gamma \ell_D^{1-\gamma} \|\nabla(u - u_h)\|_{\boldsymbol{L}^2(D)}$? (The length scale $\ell_D := \mathrm{diam}(D)$ is introduced to make the constant c dimensionless.)

32.3.1   Abstract duality argument

The above question can be formulated in a context more general than that of the boundary value problem (32.1). Let us for a moment adopt an abstract point of view. Let V and L be two Banach spaces such that V embeds continuously in L, i.e., $V \hookrightarrow L$. Let $a : V \times V \to \mathbb{C}$ be a bounded sesquilinear form satisfying the assumptions of the BNB theorem (Theorem 25.9). Let $V_h \subset V$ be a finite-dimensional subspace equipped with the norm of V, and assume that the restriction of a to $V_h \times V_h$ satisfies a uniform inf-sup condition with constant α_h, i.e., $\alpha_h \geq \alpha_0 > 0$ for all $h \in \mathcal{H}$. A first important step toward answering the above question for the error measured in the L- and V-norms is given by the following result due to Sayas [342].

Theorem 32.8 (Improved estimate ⇔ compactness). *Let $G_h : V \to V_h \subset L$ be the discrete solution map defined in (32.6), i.e., $a(G_h(v) - v, w_h) := 0$ for all $w_h \in V_h$. Then the following holds true if and only if the embedding $V \hookrightarrow L$ is compact:*

$$\lim_{h \to 0} \left(\sup_{v \in V \setminus V_h} \frac{\|G_h(v) - v\|_L}{\|G_h(v) - v\|_V} \right) = 0. \tag{32.13}$$

Proof. See [342, Thm. 1.1] and Exercise 32.3.   □

An immediate consequence of Theorem 32.8 is that it is necessary that the embedding $V \hookrightarrow L$ be compact to get a better convergence rate on $\|G_h(u) - u\|_L$ than on $\|G_h(u) - u\|_V$. We now present a result due to Aubin [28, 29] and Nitsche [313] that gives an estimate of the gain in convergence rate that one should expect. Let $\iota_{L,V}$ denote the operator norm of the above embedding, i.e., $\|v\|_L \leq \iota_{L,V} \|v\|_V$ for all $v \in V$. Recall that α and $\|a\|$ denote the coercivity and boundedness constants of a on $V \times V$.

Definition 32.9 (Adjoint problem). *Assume that L is a Hilbert space with inner product $(\cdot, \cdot)_L$. For any $g \in L$, we denote by $\zeta_g \in V$ the unique solution to the following adjoint problem:*

$$a(v, \zeta_g) = (v, g)_L, \qquad \forall v \in V. \tag{32.14}$$

The adjoint problem is well-posed since $a : V \times V \to \mathbb{C}$ is a bounded sesquilinear form satisfying the assumptions of the BNB theorem.

Lemma 32.10 (Aubin–Nitsche, abstract setting). *Let $\zeta_{u-G_h(u)}$ solve the adjoint problem* (32.14) *with data $g := u - G_h(u)$, i.e., $a(v, \zeta_{u-G_h(u)}) = (v, u - G_h(u))_L$ for all $v \in V$. The following holds true:*

$$\|u - G_h(u)\|_L \leq \|a\| \left(\inf_{w_h \in V_h} \frac{\|\zeta_{u-G_h(u)} - w_h\|_V}{\|u - G_h(u)\|_L} \right) \|u - G_h(u)\|_V. \tag{32.15}$$

Proof. Using $g := u - G_h(u)$ and the test function $v := u - G_h(u)$ in (32.14), and using the definition of $G_h(u)$ (that is, the Galerkin orthogonality property), we obtain

$$\|u - G_h(u)\|_L^2 = a(u - G_h(u), \zeta_{u-G_h(u)}) = a(u - G_h(u), \zeta_{u-G_h(u)} - w_h),$$

for all $w_h \in V_h$. The assertion follows readily.   □

The factor that leads to an improved rate of convergence on the L-error is the infimum on the right-hand side of (32.15). Assume that there is a subspace $Y \hookrightarrow V$ composed of functions that can be approximated at a rate h^γ in the V-norm by a function in V_h, that is, $\inf_{w_h \in V_h} \|\zeta - w_h\|_V \leq c_{\mathrm{app}} h^\gamma \ell_D^{-\gamma} \iota_{V,Y} \|\zeta\|_Y$ for all $\zeta \in Y$, where $\iota_{V,Y}$ is the operator norm of the above embedding, i.e., $\|y\|_V \leq \iota_{V,Y} \|y\|_Y$ for all $y \in Y$. Assume that the adjoint solution to (32.14) enjoys a smoothness property of the form $\alpha \|\zeta_g\|_Y \leq c_{\mathrm{smo}} \frac{\iota_{L,V}}{\iota_{V,Y}} \|g\|_L$. Setting $c := c_{\mathrm{app}} c_{\mathrm{smo}} \frac{\|a\|}{\alpha}$, we conclude from (32.15) that

$$\|u - G_h(u)\|_L \leq c \, h^\gamma \ell_D^{-\gamma} \iota_{L,V} \|u - G_h(u)\|_V. \tag{32.16}$$

32.3.2 L^2-error estimate

Let us now return to the H^1-conforming approximation of the elliptic PDE (32.1). Let u solve (32.2) and let u_h solve (32.5). Since the embedding $H^1(D) \hookrightarrow L^2(D)$ is compact (this is the Rellich–Kondrachov theorem), Theorem 32.8 says that it is possible to obtain a convergence rate on $\|u - u_h\|_{L^2(D)}$ that is better than that on $\|\nabla(u - u_h)\|_{\boldsymbol{L}^2(D)}$. It is important to realize that the compactness property is essential here (see Theorem 2.35).

Let us apply Lemma 32.10 with $L := L^2(D)$ and $V := H_0^1(D)$ equipped with the norms $\|v\| := \|v\|_{L^2(D)}$ and $\|v\|_V := \|\nabla v\|_{\boldsymbol{L}^2(D)}$, respectively, so that $\iota_{L,V} := C_{\mathrm{PS}}^{-1} \ell_D$. For all $g \in L^2(D)$, the adjoint problem consists of seeking the function $\zeta_g \in H_0^1(D)$ s.t.

$$a(v, \zeta_g) = (v, g)_{L^2(D)}, \qquad \forall v \in H_0^1(D). \tag{32.17}$$

We assume that there is $s \in (0,1]$ and a constant c_{smo} s.t. the following *smoothing property* holds true:

$$\|\zeta_g\|_{H^{1+s}(D)} \leq c_{\text{smo}}\, \alpha^{-1}\ell_D^2\|g\|_{L^2(D)}, \qquad \forall g \in L^2(D). \tag{32.18}$$

In the setting of §32.3.1, we have $Y := H^{1+s}(D)$ with $\|\zeta\|_Y := \|\zeta\|_{H^{1+s}(D)}$ so that $\iota_{V,Y} := \ell_D^{-1}$. Since d is symmetric, a distribution argument shows that $-\nabla\cdot(\mathrm{d}\nabla\zeta_g) - \boldsymbol{\beta}\cdot\nabla\zeta_g + \mu\zeta_g = g$ in D and $\zeta_g = 0$ on ∂D. Sufficient conditions for the smoothness property (32.18) to hold true then follow from the elliptic regularity theory of §31.4. For instance, this property holds true with $s \in (\frac{1}{2},1]$ if D is a Lipschitz polyhedron and the fields d, $\boldsymbol{\beta}$, and μ are smooth. The maximal value $s = 1$ is obtained for convex domains.

Lemma 32.11 (Aubin–Nitsche). *Let u solve (32.2) and let u_h solve (32.5). Assume that the smoothing property (32.18) holds true for some $s \in (0,1]$. There is c, depending linearly on $C_{\text{PS}}^{-1}\frac{\|a\|}{\alpha}$, s.t. for all $h \in \mathcal{H}$,*

$$\|u - u_h\|_{L^2(D)} \leq c\, h^s \ell_D^{1-s}\|\nabla(u - u_h)\|_{\boldsymbol{L}^2(D)}. \tag{32.19}$$

Proof. Direct consequence of (32.16) since $\iota_{L,V} := C_{\text{PS}}^{-1}\ell_D$. $\qquad\square$

Remark 32.12 (Adjoint operator). Let $A \in \mathcal{L}(V;V')$ be the operator associated with the bilinear form a, i.e., $\langle A(v), w\rangle_{V',V} := a(v,w)$ for all $(v,w) \in V{\times}V$. The adjoint operator $A^* \in \mathcal{L}(V;V')$ is s.t. $\langle A^*(v), w\rangle_{V',V} = \langle A(w), v\rangle_{V',V} = a(w,v)$. Hence, the adjoint solution solves $A^*(\zeta_g) = g$. $\qquad\square$

Remark 32.13 (Best approximation in L^2 and L^∞). It is in general not true that there is c s.t. $\|u - u_h\|_{L^2(D)} \leq c \inf_{w_h \in V_h} \|u - w_h\|_{L^2(D)}$ for all $h \in \mathcal{H}$; see Babuška and Osborn [37, p. 58] for a one-dimensional counterexample. However, if the mesh sequence is quasi-uniform, it is shown in Schatz and Wahlbin [344, Thm. 5.1] that $\|u - u_h\|_{L^\infty(D)} \leq c \inf_{w_h \in V_h} \|u - w_h\|_{L^\infty(D)}$ if $k \geq 2$, and $\|u - u_h\|_{L^\infty(D)} \leq c \ln(\ell_D/h) \inf_{w_h \in V_h} \|u - w_h\|_{L^\infty(D)}$ if $k = 1$. $\qquad\square$

32.4 Elliptic projection

The operator defined below is a useful tool we are going to invoke often; see, e.g., §66.3.1 for parabolic problems.

Definition 32.14 (Elliptic projection). *Let $V \subset H^1(D)$ be a Hilbert space and assume that $v \mapsto \|\nabla v\|_{\boldsymbol{L}^2(D)}$ is a norm on V. The discrete solution map $G_h : V \to V_h$ defined in (32.6) with $a(v,w) := (\nabla v, \nabla w)_{\boldsymbol{L}^2(D)}$ is called* elliptic projection *and is denoted by $\Pi_h^{\text{E}} : V \to V_h$. Thus, for all $v \in V$, we have*

$$(\nabla(v - \Pi_h^{\text{E}}(v)), \nabla w_h)_{\boldsymbol{L}^2(D)} = 0, \qquad \forall w_h \in V_h. \tag{32.20}$$

The two main properties of Π_h^{E} are the following: (i) Π_h^{E} is a projection, i.e., $\Pi_h^{\text{E}}(\Pi_h^{\text{E}}(v)) = \Pi_h^{\text{E}}(v)$ for all $v \in V$; (ii) Since by definition $a(\Pi_h^{\text{E}}(v) - v, w_h) = 0$ for all $w_h \in V_h$, one always has $|\Pi_h^{\text{E}}(v)|_{H^1(D)} \leq |v|_{H^1(D)}$.

Theorem 32.15 (Approximation). *Let $s \in (0,1]$ be the elliptic regularity index (i.e., there is c_{smo} s.t. for all $\xi \in L^2(D)$, the solution to the adjoint problem $a(v, z(\xi)) = (v,\xi)_{L^2(D)}$ for all $v \in V$, satisfies $\|z\|_{H^{1+s}(D)} \leq c_{\text{smo}}\ell_D^2\|\xi\|_{L^2(D)}$). There is c such that for all $h \in \mathcal{H}$,*

$$\|\nabla(\Pi_h^{\text{E}}(v) - v)\|_{\boldsymbol{L}^2(D)} \leq \inf_{v_h \in V_h} \|\nabla(v - v_h)\|_{\boldsymbol{L}^2(D)}, \tag{32.21a}$$

$$\|\Pi_h^{\text{E}}(v) - v\|_{L^2(D)} \leq c\, h^s \ell_D^{1-s}\|\nabla(\Pi_h^{\text{E}}(v) - v)\|_{\boldsymbol{L}^2(D)}. \tag{32.21b}$$

Proof. The estimate (32.21a) is a consequence of

$$|\Pi_h^{\mathrm{E}}(v) - v|_{H^1(D)}^2 + |v_h - \Pi_h^{\mathrm{E}}(v)|_{H^1(D)}^2 = |v - v_h|_{H^1(D)}^2, \qquad \forall v_h \in V_h.$$

The estimate (32.21b) follows from (32.19) since Π_h^{E} is the solution operator associated with the bilinear form $a(v, w) := (\nabla v, \nabla w)_{\boldsymbol{L}^2(D)}$ on V_h. $\qquad\square$

Remark 32.16 (Other BCs). The elliptic projection is unambiguously defined when Dirichlet conditions are applied on some part of ∂D with positive measure. In the case of Neumann conditions, Π_h^{E} acts only on $H_*^1(D) := \{v \in H^1(D) \mid \underline{v} = 0\}$, where $\underline{v}$ denotes the average of v over D. We can extend Π_h^{E} to $H^1(D)$ by setting $\Pi_{*h}^{\mathrm{E}}(v) := \Pi_h^{\mathrm{E}}(v - \underline{v}) + \underline{v}$ for all $v \in H^1(D)$. The approximation properties of Π_{*h}^{E} in $H^1(D)$ are exactly the same as those of Π_h^{E} in $H_*^1(D)$. $\qquad\square$

Exercises

Exercise 32.1 (Discrete solution map). Let G_h be defined in (32.6). (i) Prove that $\|\nabla(v - G_h(v))\|_{\boldsymbol{L}^2(D)} \leq ch^r|v|_{H^{1+r}(D)}$ for all $r \in (0, k]$, all $v \in H^{1+r}(D)$, and all $h \in \mathcal{H}$. (*Hint*: observe that $G_h(\mathcal{I}_{h0}^{\mathrm{g,av}}(v)) = \mathcal{I}_{h0}^{\mathrm{g,av}}(v)$.) (ii) Assume that the adjoint operator A^* has a smoothing property in $H^{1+s}(D)$ for some real number $s \in (0, 1]$. Prove that $\|v - G_h(v)\|_{L^2(D)} \leq ch^{r+s}\ell_D^{1-s}|v|_{H^{1+r}(D)}$. (*Hint*: consider the adjoint problem $A^*(\zeta) = v - G_h(v)$.)

Exercise 32.2 (H^{-1}-estimate). Assume that for all $g \in H^1(D)$, the adjoint solution $\zeta \in H_0^1(D)$ s.t. $A^*(\zeta) = g$ satisfies $\|\zeta\|_{H^{2+s}(D)} \leq c_{\mathrm{smo}}\alpha^{-1}\ell_D^2\|g\|_{H^1(D)}$ with $s \in (\frac{1}{2}, 1]$. Assume that $k \geq 1 + s$. Let $\|v\|_{H^{-1}(D)} := \sup_{z \in H_0^1(D)} \frac{(v, z)_{L^2(D)}}{|z|_{H^1(D)}}$ for all $v \in L^2(D)$. Prove that $\|u - u_h\|_{H^{-1}(D)} \leq ch^{1+s}\ell_D^{1-s}\|\nabla(u - u_h)\|_{\boldsymbol{L}^2(D)}$. (*Hint*: consider the adjoint problem $A^*(\zeta) = z$.)

Exercise 32.3 (Compactness). The goal is to prove Theorem 32.8. Let $I : V \to L$ be the natural embedding and define $\epsilon(h) := \sup_{v \in V \setminus V_h} \frac{\|G_h(v) - v\|_L}{\|G_h v - v\|_V}$. (i) Prove that $\|G_h - I\|_{\mathcal{L}(V;L)} \leq \frac{\|a\|}{\alpha}\epsilon(h)$, where α and $\|a\|$ are the coercivity and the boundedness constants of a on $V \times V$. (ii) Assume that $\lim_{h \to 0} \epsilon(h) = 0$. Prove that I is compact. (*Hint*: use (i).) (iii) Let $R : L \to V$ be s.t. $a(y, R(f)) := (y, f)_L$ for all $y \in V$ and all $f \in L$. Assuming that I is compact, prove that R is compact. (*Hint*: prove that $R = (A^*)^{-1}I^*$ and use Schauder's theorem; see Theorem C.48.) (iv) Let $P_h^V : V \to V_h$ be the V-orthogonal projection onto V_h. Let $R_h : L \to V_h$ be the operator defined by $a(v_h, R_h(f)) := (v_h, f)_L$, for all $v_h \in V_h$ and all $f \in L$. Prove that $\|R - R_h\|_{\mathcal{L}(L;V)} \leq \frac{\|a\|}{\alpha}\|R - P_h^V \circ R\|_{\mathcal{L}(L;V)}$. (v) Assuming that I is compact, prove that $\lim_{h \to 0} \|R - R_h\|_{\mathcal{L}(L;V)} = 0$. (*Hint*: use (iii)-(iv) and proceed as in Remark C.5.) (vi) Assuming that I is compact, prove that $\lim_{h \to 0} \epsilon(h) = 0$.

Exercise 32.4 (Source approximation). Let $f \in L^2(D)$, let $\mathcal{I}_h^{\mathrm{b}}(f)$ be the L^2-projection of f onto $P_{k'}^{\mathrm{b}}(\mathcal{T}_h)$. Consider the discrete problem (32.5) with the right-hand side $\int_D \mathcal{I}_h^{\mathrm{b}}(f)w_h\,\mathrm{d}x$, that is: Find $u_h \in V_h := P_{k,0}^{\mathrm{g}}(\mathcal{T}_h)$ s.t. $a(u_h, w_h) = \ell_h(w_h) := \int_D \mathcal{I}_h^{\mathrm{b}}(f)w_h\,\mathrm{d}x$ for all $w_h \in V_h$. (i) How should (32.7) be rewritten? Show that $k' := k - 1$ leads to an optimal H^1-norm error estimate. (ii) How should (32.19) be rewritten? Assuming full elliptic regularity, show that $k' := k$ leads to an optimal L^2-norm error estimate.

Exercise 32.5 (Advection-diffusion, 1D). Let $D := (0, 1)$. Let ν, b be positive real numbers. Let $f : D \to \mathbb{R}$ be a smooth function. Consider the model problem $-\nu u'' + bu' = f$ in D, $u(0) = 0$, $u(1) = 0$. Consider H^1-conforming $\mathbb{P}_1$ Lagrange finite elements on the uniform grid $\mathcal{T}_h$ with nodes

$x_i := ih$, $\forall i \in \{0{:}I\}$, and meshsize $h := \frac{1}{I+1}$. (i) Evaluate the stiffness matrix. (*Hint*: factor out the ratio $\frac{\nu}{h}$ and introduce the local Péclet number $\gamma := \frac{bh}{\nu}$.) (ii) Solve the linear system when $f := 1$ and plot the solutions for $h := 10^{-2}$ and $\gamma \in \{0.1, 1, 10\}$. (*Hint*: wrtite $\mathsf{U} = \mathsf{U}^0 + \tilde{\mathsf{U}} \in \mathbb{R}^I$ with $\mathsf{U}_i^0 := b^{-1}ih$ and $\tilde{\mathsf{U}}_i := \varrho + \theta\delta^i$ for some constants ϱ, θ, δ.) (iii) Consider now the boundary conditions $u(0) = 0$ and $u'(1) = 0$. Write the weak formulation and show its well-posedness. Evaluate the stiffness matrix. (*Hint*: the matrix is of order $(I + 1)$.) Derive the equation satisfied by $h^{-1}(\mathsf{U}_{I+1} - \mathsf{U}_I)$, and find the limit values as $h \to 0$ with fixed $\nu > 0$ and as $\nu \to 0$ with fixed $h \in \mathcal{H}$.

Chapter 33

H^1-conforming approximation (II)

In this chapter, we study the following questions regarding the approximation of second-order elliptic PDEs by H^1-conforming finite elements: (i) How can non-homogeneous Dirichlet conditions be taken into account in the error analysis, and how can they be implemented in practice; (ii) Can the discrete problem reproduce the maximum principle, which is an important property enjoyed by the exact problem; (iii) How quadratures impact the well-posedness and error analysis of the discrete problem. Two other important topics treated in the forthcoming chapters are: (iv) The derivation of a posteriori error estimates and their use for mesh adaptation (Chapter 34); (v) A local post-processing technique to recover an $\boldsymbol{H}(\mathrm{div}; D)$-conforming flux approximating the exact flux $\boldsymbol{\sigma} := -\nabla u$ (Chapter 52).

33.1 Non-homogeneous Dirichlet conditions

In this section, we consider the PDE (32.1), i.e.,

$$-\nabla\cdot(\mathbb{d}\nabla u) + \boldsymbol{\beta}\cdot\nabla u + \mu u = f \quad \text{in } D, \tag{33.1}$$

with the same assumptions on $\mathbb{d}$, $\boldsymbol{\beta}$, μ, and f as in §32.1, but with the non-homogeneous Dirichlet condition $\gamma^{\mathrm{g}}(u) = g$ on ∂D with $g \in H^{\frac{1}{2}}(\partial D)$, where $\gamma^{\mathrm{g}} : H^1(D) \to H^{\frac{1}{2}}(\partial D)$ is the trace map such that $\gamma^{\mathrm{g}}(v) = v_{|\partial D}$ for every smooth function v. Following §31.2.2, we invoke the surjectivity of the trace map γ^{g} to infer that there is $C_{\gamma^{\mathrm{g}}}$ such that for all $g \in H^{\frac{1}{2}}(\partial D)$, there is $u_g \in H^1(D)$ satisfying $\gamma^{\mathrm{g}}(u_g) = g$ and $\|u_g\|_{H^1(D)} \leq C_{\gamma^{\mathrm{g}}}\|g\|_{H^{\frac{1}{2}}(\partial D)}$; see Theorem 3.10(iii). The function u_g is called *lifting of the Dirichlet condition*. By making the change of variable $u_0 := u - u_g$, we now look for a function u_0 satisfying the homogeneous Dirichlet condition $\gamma^{\mathrm{g}}(u_0) = 0$. Let $\widetilde{V} := H^1(D)$ and $V := H_0^1(D)$. The above considerations lead to the following weak formulation:

$$\begin{cases} \text{Find } u \in \widetilde{V} \text{ such that } u_0 := u - u_g \in V \text{ satisfies} \\ a(u_0, w) = \ell(w) - \widetilde{a}(u_g, w), \quad \forall w \in V, \end{cases} \tag{33.2}$$

where

$$\widetilde{a}(v, w) := \int_D \left((\mathbb{d}\nabla v)\cdot\nabla w + (\boldsymbol{\beta}\cdot\nabla v)w + \mu vw\right) \mathrm{d}x, \quad \forall(v, w) \in \widetilde{V}\times V, \tag{33.3}$$

$a := \widetilde{a}_{|V\times V}$, and $\ell(w) := \int_D fw \, \mathrm{d}x$ for all $w \in V$. We equip the space V with the norm $\|v\|_V := \|\nabla v\|_{\boldsymbol{L}^2(D)}$. Let ℓ_D be a characteristic length of D, e.g., $\ell_D := \mathrm{diam}(D)$, and let C_{PS} be the

Poincaré–Steklov constant s.t. $C_{\text{PS}}\|w\|_{L^2(D)} \leq \ell_D\|\nabla w\|_{\boldsymbol{L}^2(D)} = \ell_D|w|_{H^1(D)}$ for all $w \in H_0^1(D)$ (see (3.11) with $p := 2$). Then the assumptions on d, $\boldsymbol{\beta}$, and μ from §32.1 imply that a is V-coercive, i.e., there is $\alpha > 0$ s.t. $a(v,v) \geq \alpha\|\nabla v\|_{\boldsymbol{L}^2(D)}^2$ for all $v \in V$. Moreover, we have $|\widetilde{a}(v,w)| \leq \|\widetilde{a}\|\ell_D^{-1}\|v\|_{H^1(D)}\|\nabla w\|_{\boldsymbol{L}^2(D)}$ for all $(v,w) \in \widetilde{V} \times V$, with $\|\widetilde{a}\| := \lambda_\sharp + \beta_\sharp C_{\text{PS}}^{-1}\ell_D + \mu_\sharp C_{\text{PS}}^{-1}\ell_D^2$, $\lambda_\sharp := \|\text{d}\|_{\mathbb{L}^\infty(D)}$, $\beta_\sharp := \|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(D)}$, $\mu_\sharp := \|\mu\|_{L^\infty(D)}$.

33.1.1 Discrete problem and well-posedness

We want to approximate the model problem (33.2) using the H^1-conforming finite element space $P_k^{\text{g}}(\mathcal{T}_h)$ defined in (32.4b) and its zero-trace subspace $P_{k,0}^{\text{g}}(\mathcal{T}_h)$ defined in (32.4c). Since the function g may not be in $\gamma^{\text{g}}(P_k^{\text{g}}(\mathcal{T}_h))$, we need to approximate the non-homogeneous Dirichlet condition in the discrete problem. To this purpose, we assume for simplicity that $g \in C^0(\partial D)$, and we define an approximation g_h of g by using the boundary degrees of freedom (dofs) of the finite element. Recall that the dofs are point-values for Lagrange elements, whereas they are point-values or integrals over edges, faces, or cells for the canonical hybrid element. Let $\{\varphi_a\}_{a\in\mathcal{A}_h}$ and $\{\sigma_a\}_{a\in\mathcal{A}_h}$ be, respectively, the global shape functions and dofs in $P_k^{\text{g}}(\mathcal{T}_h)$ (see §19.2.1). Let $s > \frac{d}{2}$ and $\mathcal{I}_h : V^{\text{g}}(D) := H^s(D) \to P_k^{\text{g}}(\mathcal{T}_h)$ denote either the canonical interpolation operator $\mathcal{I}_h^{\text{g}}$ from §19.3, or the Lagrange interpolation operator $\mathcal{I}_h^{\text{L}}$. We have $\mathcal{I}_h(v) := \sum_{a\in\mathcal{A}_h} \sigma_a(v)\varphi_a$ for all $v \in V^{\text{g}}(D)$. Recall from Definition 19.11 that the set $\mathcal{A}_h^\partial$ is the collection of the boundary dofs, i.e., $a \in \mathcal{A}_h^\partial$ iff $\gamma^{\text{g}}(\varphi_a) = \varphi_{a|\partial D} \neq 0$. Then $\sigma_a(v)$ only depends on $\gamma^{\text{g}}(v)$ for all $a \in \mathcal{A}_h^\partial$, i.e., we can write $\sigma_a(v) = (\sigma_a^\partial \circ \gamma^{\text{g}})(v)$ for all $v \in V^{\text{g}}(D)$, where σ_a^∂ can be a value at a boundary point or an integral over a boundary edge or a boundary face. Let us set

$$g_h := \sum_{a\in\mathcal{A}_h^\partial} \sigma_a^\partial(g)\varphi_{a|\partial D}. \tag{33.4}$$

We consider the following discrete problem:

$$\begin{cases} \text{Find } u_h \in \widetilde{V}_h := P_k^{\text{g}}(\mathcal{T}_h) \text{ such that } u_{h|\partial D} = g_h \text{ and} \\ \widetilde{a}(u_h, w_h) = \ell(w_h), \quad \forall w_h \in V_h := P_{k,0}^{\text{g}}(\mathcal{T}_h). \end{cases} \tag{33.5}$$

At this stage, the discrete trial space includes boundary dofs (these dofs are fixed for u_h by setting $u_{h|\partial D} = g_h$), whereas the boundary dofs vanish for the discrete test functions. Upon introducing the discrete lifting $u_{hg} := \sum_{a\in\mathcal{A}_h^\partial} \sigma_a^\partial(g)\varphi_a \in \widetilde{V}_h$ and making the change of variable $u_{h0} := u_h - u_{hg}$, we notice that $u_{h0} \in V_h$ since $u_{h|\partial D} = g_h = u_{hg|\partial D}$, i.e., $u_{h0|\partial D} = 0$. The discrete problem (33.5) can then be recast in a form that is similar to (33.2), that is,

$$\begin{cases} \text{Find } u_h \in \widetilde{V}_h \text{ such that } u_{h0} := u_h - u_{hg} \in V_h \text{ satisfies} \\ a(u_{h0}, w_h) = \ell(w_h) - \widetilde{a}(u_{hg}, w_h), \quad \forall w_h \in V_h. \end{cases} \tag{33.6}$$

Lemma 33.1 (Well-posedness). *The discrete problems* (33.5) *and* (33.6) *are well-posed.*

Proof. Since $V_h \subset V$, the bilinear form a is bounded and coercive on V_h, and the linear form $\ell_{hg}(\cdot) := \ell(\cdot) - \widetilde{a}(u_{hg}, \cdot)$ is bounded on V_h. The Lax–Milgram lemma implies that the discrete solution $u_{h0} \in V_h$ is uniquely defined. Since the problems (33.5) and (33.6) are equivalent, the discrete solution $u_h \in \widetilde{V}_h$ is also uniquely defined. $\qquad\square$

33.1.2   Error analysis

The approximation setting leading to (33.5) is conforming since $\widetilde{V}_h \subset \widetilde{V}$ and $V_h \subset V$. However, there is a consistency error resulting from the fact that the non-homogeneous Dirichlet condition is interpolated in (33.5).

Theorem 33.2 (H^1-estimate). *Let u solve (33.2) and let u_h solve (33.5). Assume that $k+1 > \frac{d}{2}$ and $u \in H^{1+r}(D)$ with $r \in (\frac{d}{2} - 1, k]$. There is c s.t. for all $h \in \mathcal{H}$,*

$$\|\nabla(u - u_h)\|_{\boldsymbol{L}^2(D)} \leq c \left(\sum_{K \in \mathcal{T}_h} h_K^{2r} |u|_{H^{1+r}(K)}^2 \right)^{\frac{1}{2}} \leq c\, h^r |u|_{H^{1+r}(D)}. \tag{33.7}$$

Proof. The proof is similar to that of Céa's Lemma, except that we need to account for the interpolation of the Dirichlet condition. Since $1 + r > \frac{d}{2}$ by assumption, we have $u \in V^{\mathrm{g}}(D)$ and $\mathcal{I}_h(u)$ is well defined. Owing to (33.4) and since $\sigma_a(u) = \sigma_a^\partial(\gamma^{\mathrm{g}}(u)) = \sigma_a^\partial(g)$ for all $a \in \mathcal{A}_h^\partial$, we infer that

$$\mathcal{I}_h(u)_{|\partial D} = \sum_{a \in \mathcal{A}_h^\partial} \sigma_a(u) \varphi_{a|\partial D} = \sum_{a \in \mathcal{A}_h^\partial} \sigma_a^\partial(g) \varphi_{a|\partial D} = g_h = u_{h|\partial D}.$$

Hence, $\mathcal{I}_h(u) - u_h \in V_h$. Moreover, (33.2) and (33.5) imply that $\widetilde{a}(u - u_h, w_h) = 0$ for all $w_h \in V_h \subset V := H_0^1(D)$. The coercivity of a on V and the boundedness of $\widetilde{a}$ on $\widetilde{V} \times V$ imply that

$$\alpha \|\nabla(\mathcal{I}_h(u) - u_h)\|_{\boldsymbol{L}^2(D)} \leq \sup_{w_h \in V_h} \frac{|a(\mathcal{I}_h(u) - u_h, w_h)|}{\|\nabla w_h\|_{\boldsymbol{L}^2(D)}} = \sup_{w_h \in V_h} \frac{|\widetilde{a}(\mathcal{I}_h(u) - u_h, w_h)|}{\|\nabla w_h\|_{\boldsymbol{L}^2(D)}}$$

$$= \sup_{w_h \in V_h} \frac{|\widetilde{a}(\mathcal{I}_h(u) - u, w_h)|}{\|\nabla w_h\|_{\boldsymbol{L}^2(D)}} \leq \|\widetilde{a}\| \ell_D^{-1} \|u - \mathcal{I}_h(u)\|_{H^1(D)}.$$

The triangle inequality leads to $\|\nabla(u - u_h)\|_{\boldsymbol{L}^2(D)} \leq c\ell_D^{-1} \|u - \mathcal{I}_h(u)\|_{H^1(D)}$ with $c := 1 + \frac{\|\widetilde{a}\|}{\alpha}$. Finally, Corollary 19.8 yields $\ell_D^{-1} \|u - \mathcal{I}_h(u)\|_{H^1(D)} \leq (\sum_{K \in \mathcal{T}_h} (\ell_D^{-2} h_K^{2(r+1)} + h_K^{2r}) |u|_{H^{r+1}(K)}^2)^{\frac{1}{2}}$, and (33.7) follows since $h_K \leq \ell_D$ for all $K \in \mathcal{T}_h$. $\qquad\square$

We now use duality techniques to derive an improved L^2-norm error estimate. Recall from §32.3 that for all $g \in L^2(D)$, the adjoint solution $\zeta_g \in H_0^1(D)$ is s.t. $a(v, \zeta_g) = (v, g)_{L^2(D)}$ for all $v \in H_0^1(D)$. Notice that ζ satisfies a homogeneous Dirichlet condition.

Theorem 33.3 (L^2-estimate). *Assume that there is $s \in (\frac{1}{2}, 1]$ and $c_{\mathrm{smo}} > 0$ s.t. the adjoint solution satisfies $\|\zeta_g\|_{H^{1+s}(D)} \leq c_{\mathrm{smo}} \alpha^{-1} \ell_D^2 \|g\|_{L^2(D)}$. Assume that d is Lipschitz. Assume that $k + 1 > \frac{d}{2}$ and $u \in H^{1+r}(D)$ with $r \in (\frac{d}{2} - 1, k]$. There is c s.t. for all $h \in \mathcal{H}$,*

$$\|u - u_h\|_{L^2(D)} \leq c \left(h^{r+s} \ell_D^{1-s} |u|_{H^{1+r}(D)} + \ell_D^{\frac{1}{2}} \|g - g_h\|_{L^2(\partial D)} \right), \tag{33.8}$$

where c depends linearly on $\frac{\|\widetilde{a}\|}{\alpha}$ and the Lipschitz constant of $\ell_D \lambda_\sharp^{-1} \mathrm{d}$.

Proof. The proof is similar to that of the Aubin–Nitsche lemma, except that we need to account for the interpolation of the Dirichlet condition. Let $\zeta \in H_0^1(D)$ be the adjoint solution s.t. $\widetilde{a}(v, \zeta) = (v, \mathcal{I}_h(u) - u_h)_{L^2(D)}$ for all $v \in H_0^1(D)$. The smoothness property implies that $|\zeta|_{H^{1+s}(D)} \leq \ell_D^{-1-s} \|\zeta\|_{H^{1+s}(D)} \leq c_{\mathrm{smo}} \alpha^{-1} \ell_D^{1-s} \|\mathcal{I}_h(u) - u_h\|_{L^2(D)}$. Since $\mathcal{I}_h(u) - u_h \in V_h \subset H_0^1(D)$, we obtain

$$\|\mathcal{I}_h(u) - u_h\|_{L^2(D)}^2 = a(\mathcal{I}_h(u) - u_h, \zeta) = \widetilde{a}(\mathcal{I}_h(u) - u, \zeta) + \widetilde{a}(u - u_h, \zeta) =: \mathfrak{T}_1 + \mathfrak{T}_2.$$

The term $\mathfrak{T}_1$ is bounded using Lemma 33.4, leading to

$$|\mathfrak{T}_1| \leq c\,\|\widetilde{a}\|\ell_D^{-2}\big(\|u - \mathcal{I}_h(u)\|_{H^{1-s}(D)} + \ell_D^{\frac{1}{2}}\|g - g_h\|_{L^2(\partial D)}\big)\|\zeta\|_{H^{1+s}(D)},$$

where we used that $\gamma^{\mathrm{g}}(\mathcal{I}_h(u) - u) = g_h - g$. For the term $\mathfrak{T}_2$, letting $e := u - u_h$, we have $\widetilde{a}(e, v_h) = 0$ for all $v_h \in V_h$. The boundedness of $\widetilde{a}$ on $\widetilde{V} \times V$ and the approximation properties of $\mathcal{I}_{h0}^{\mathrm{g,av}}$ then give

$$
\begin{aligned}
|\mathfrak{T}_2| &\leq \|\widetilde{a}\|\ell_D^{-1}\|e\|_{H^1(D)}\|\nabla(\zeta - \mathcal{I}_{h0}^{\mathrm{g,av}}(\zeta))\|_{\boldsymbol{L}^2(D)} \\
&\leq c\,h^s\|\widetilde{a}\|\ell_D^{-1}\|e\|_{H^1(D)}|\zeta|_{H^{1+s}(D)} \leq c\,h^s\|a\|\ell_D^{-2-s}\|e\|_{H^1(D)}\|\zeta\|_{H^{1+s}(D)} \\
&\leq c'\,h^s\|\widetilde{a}\|\ell_D^{-2-s}\big(\ell_D\|\nabla e\|_{\boldsymbol{L}^2(D)} + \ell_D^{\frac{1}{2}}\|g - g_h\|_{L^2(\partial D)}\big)\|\zeta\|_{H^{1+s}(D)} \\
&\leq c'\,\|\widetilde{a}\|\ell_D^{-2}\big(h^s\ell_D^{1-s}\|\nabla e\|_{\boldsymbol{L}^2(D)} + \ell_D^{\frac{1}{2}}\|g - g_h\|_{L^2(\partial D)}\big)\|\zeta\|_{H^{1+s}(D)},
\end{aligned}
$$

where we used that $\|e\|_{H^1(D)} \leq c(\ell_D\|\nabla e\|_{\boldsymbol{L}^2(D)} + \ell_D^{\frac{1}{2}}\|g - g_h\|_{L^2(\partial D)})$ owing to the Poincaré–Steklov inequality (31.23), and $h \leq \ell_D$ for the boundary term in the last line. Using the above bounds on $\mathfrak{T}_1$ and $\mathfrak{T}_2$, and the smoothness property of ζ, we infer that

$$\|\mathcal{I}_h(u) - u_h\|_{L^2(D)} \leq c\,\big(h^s\ell_D^{1-s}\|\nabla e\|_{\boldsymbol{L}^2(D)} + \|u - \mathcal{I}_h(u)\|_{H^{1-s}(D)} + \ell_D^{\frac{1}{2}}\|g - g_h\|_{L^2(\partial D)}\big),$$

where c depends linearly on $\frac{\|\widetilde{a}\|}{\alpha}$. We now use Theorem 33.2 to bound $\|\nabla e\|_{\boldsymbol{L}^2(D)}$, and the estimate $\|u - \mathcal{I}_h(u)\|_{H^{1-s}(D)} \leq c\ell_D^{1-s}h^{r+s}|u|_{H^{1+r}(D)}$ which results from the Riesz–Thorin theorem (Theorem A.27). We conclude by using the triangle inequality. $\qquad\square$

Lemma 33.4 (H^{1+s}-boundedness). *Assume that d is Lipschitz. There is c, depending linearly on the Lipschitz constant of $\ell_D\lambda_\sharp^{-1}\mathrm{d}$, s.t.*

$$\widetilde{a}(v, \zeta) \leq c\,\|\widetilde{a}\|\ell_D^{-2}\big(\|v\|_{H^{1-s}(D)} + \ell_D^{\frac{1}{2}}\|\gamma^{\mathrm{g}}(v)\|_{L^2(\partial D)}\big)\|\zeta\|_{H^{1+s}(D)}, \qquad (33.9)$$

for all $v \in \widetilde{V} := H^1(D)$ and all $\zeta \in H_0^1(D) \cap H^{1+s}(D)$ with $s \in (\frac{1}{2}, 1]$.

Proof. Since d is Lipschitz and $\nabla\zeta \in \boldsymbol{H}^s(D)$, we infer that $\mathrm{d}\nabla\zeta \in \boldsymbol{H}^s(D)$, i.e., there is c (depending linearly on the Lipschitz constant of $\ell_D\lambda_\sharp^{-1}\mathrm{d}$) s.t. $\|\mathrm{d}\nabla\zeta\|_{\boldsymbol{H}^s(D)} \leq c\lambda_\sharp\|\nabla\zeta\|_{\boldsymbol{H}^s(D)}$. This implies that $\mathrm{d}\nabla\zeta$ has a trace in $\boldsymbol{H}^{s-\frac{1}{2}}(\partial D)$, and hence that $\boldsymbol{n}\cdot(\mathrm{d}\nabla\zeta) \in L^2(\partial D)$ (since $s > \frac{1}{2}$). Moreover, $\nabla\cdot(\mathrm{d}\nabla\zeta) \in H^{-1+s}(D)$ and

$$
\begin{aligned}
\|\nabla\cdot(\mathrm{d}\nabla\zeta)\|_{H^{-1+s}(D)} &\leq c\,\ell_D^{-1}\|\mathrm{d}\nabla\zeta\|_{\boldsymbol{H}^s(D)} \leq c'\ell_D^{-1}\lambda_\sharp\|\nabla\zeta\|_{\boldsymbol{H}^s(D)} \\
&\leq c''\,\lambda_\sharp\ell_D^{-2}\|\zeta\|_{H^{1+s}(D)} \leq c'''\,\|\widetilde{a}\|\ell_D^{-2}\|\zeta\|_{H^{1+s}(D)}.
\end{aligned}
$$

Here, we used that $\nabla: H^{1+s}(D) \to \boldsymbol{H}^s(D)$ and $\nabla\cdot: \boldsymbol{H}^s(D) \to H^{-1+s}(D)$ are bounded owing to the Riesz–Thorin theorem. Observing that $H^{1-s}(D) = H_0^{1-s}(D)$ (since $1 - s < \frac{1}{2}$), the linear form $\nabla\cdot(\mathrm{d}\nabla\zeta)$ can act on any $v \in H^{1-s}(D)$ even if v does not have a zero trace at the boundary. Denoting by $\langle\cdot, \cdot\rangle$ the corresponding duality product between $H^{-1+s}(D)$ and $H^{1-s}(D)$, we infer that

$$\langle\nabla\cdot(\mathrm{d}\nabla\zeta), v\rangle + \int_D \nabla v\cdot(\mathrm{d}\nabla\zeta)\,\mathrm{d}x = \int_{\partial D}(\boldsymbol{n}\cdot(\mathrm{d}\nabla\zeta))\gamma^{\mathrm{g}}(v)\,\mathrm{d}s.$$

As a result, the bilinear form $\widetilde{a}$ can be rewritten as

$$
\begin{aligned}
\widetilde{a}(v, \zeta) = {}&- \langle\nabla\cdot(\mathrm{d}\nabla\zeta), v\rangle + \int_{\partial D}(\boldsymbol{n}\cdot(\mathrm{d}\nabla\zeta))\gamma^{\mathrm{g}}(v)\,\mathrm{d}s \\
&+ \int_D(-\boldsymbol{\beta}\cdot\nabla\zeta + (\mu - \nabla\cdot\boldsymbol{\beta})\zeta)v\,\mathrm{d}x =: \mathfrak{T}_1 + \mathfrak{T}_2 + \mathfrak{T}_3.
\end{aligned}
$$

The three terms on the right-hand side can be bounded as follows:

$$|\mathfrak{T}_1| \leq \|\nabla{\cdot}(\mathbb{d}\nabla\zeta)\|_{H^{-1+s}(D)}\|v\|_{H^{1-s}(D)} \leq c\,\|\widetilde{a}\|\ell_D^{-2}\|\zeta\|_{H^{1+s}(D)}\|v\|_{H^{1-s}(D)},$$

$$|\mathfrak{T}_2| \leq \|\boldsymbol{n}{\cdot}(\mathbb{d}\nabla\zeta)\|_{L^2(\partial D)}\|\gamma^{\mathrm{g}}(v)\|_{L^2(\partial D)} \leq c\,\|\widetilde{a}(\|\ell_D^{-\frac{3}{2}}\|\zeta\|_{H^{1+s}(D)}\|\gamma^{\mathrm{g}}(v)\|_{L^2(\partial D)},$$

$$|\mathfrak{T}_3| \leq c\,\|\widetilde{a}\|\ell_D^{-2}\|\zeta\|_{H^1(D)}\|v\|_{L^2(D)} \leq c\,\|\widetilde{a}\|\ell_D^{-2}\|\zeta\|_{H^{1+s}(D)}\|v\|_{H^{1-s}(D)}. \qquad \square$$

33.1.3 Algebraic viewpoint

Let us enumerate the dofs using the set $\mathcal{I}_h := \{1{:}I\}$. We identify a block structure by enumerating first the internal dofs by using the set $\mathcal{I}_h^{\circ} := \{1{:}I^{\circ}\}$, then we enumerate the boundary dofs by using the set $\mathcal{I}_h^{\partial} := \{1{:}I^{\partial}\}$. Notice that $I = I^{\circ} + I^{\partial}$. Introducing the decomposition of the discrete solution $u_h := \sum_{i\in\mathcal{I}_h} \mathsf{U}_i\varphi_i$, the algebraic realization of (33.5) is the linear system $\mathcal{A}\mathsf{U} = \mathsf{B}$, where the stiffness matrix $\mathcal{A}$ and the load vector B have entries given by $\mathcal{A}_{ij}^{\circ\circ} := a(\varphi_j,\varphi_i)$ and $\mathsf{B}_i^{\circ} := \ell(\varphi_i)$ for all $(i,j) \in \mathcal{I}_h^{\circ}{\times}\mathcal{I}_h^{\circ}$, and $\mathcal{A}_{ij}^{\circ\partial} := a(\varphi_j,\varphi_i)$ for all $(i,j) \in \mathcal{I}_h^{\circ}{\times}\mathcal{I}_h^{\partial}$. Note that the row index associated with the test function takes values in $\mathcal{I}_h^{\circ}$ only. Moreover, the boundary prescription $u_{h|\partial D} = g_h$ in (33.5) leads to $\mathsf{U}_i = \mathsf{B}_i^{\partial} := \sigma_i^{\partial}(g)$ for all $i \in \mathcal{I}_h^{\partial}$. Thus, we obtain the following block-decomposition (with obvious notation)

$$\begin{bmatrix} \mathcal{A}^{\circ\circ} & \mathcal{A}^{\circ\partial} \\ \mathbb{O} & \mathbb{I}_{I^{\partial}} \end{bmatrix} \begin{bmatrix} \mathsf{U}^{\circ} \\ \mathsf{U}^{\partial} \end{bmatrix} = \begin{bmatrix} \mathsf{B}^{\circ} \\ \mathsf{B}^{\partial} \end{bmatrix}, \tag{33.10}$$

where $\mathbb{O}$ is a zero rectangular matrix of order $I^{\partial} \times I^{\circ}$ and $\mathbb{I}_{I^{\partial}}$ is the identity matrix of order I^{∂}. The matrix $\mathcal{A}^{\circ\circ}$ is of size I° and is invertible owing to the H_0^1-coercivity of a.

A first option to solve (33.10) is to eliminate U^{∂}, i.e., to solve the linear system $\mathcal{A}^{\circ\circ}\mathsf{U}^{\circ} = \mathsf{B}^{\circ} - \mathcal{A}^{\circ\partial}\mathsf{B}^{\partial}$. The advantage is that the final size of the linear system is optimal since only the internal dofs are unknown. However, this technique requires assembling two matrices, $\mathcal{A}^{\circ\circ}$ and $\mathcal{A}^{\circ\partial}$, instead of one, and the two matrices have a different sparsity profile. An alternative technique consists of assembling first the stiffness matrix for all the dofs in $\mathcal{A}_h$ (this is the stiffness matrix for the Neumann problem) and then correcting the rows for $a \in \mathcal{A}_h^{\partial}$ by setting the entries to zero except the diagonal ones which are set to 1. The right-hand side of (33.10) is assembled similarly. Despite the slight increase in the number of unknowns, this technique is computationally effective. It has the apparent drawback of breaking the symmetry of the model problem (recall that $\mathcal{A}^{\circ\circ}$ is symmetric if the advective velocity is zero) since the matrix in (33.10) is not symmetric. Actually, when using an iterative solution method based on a Krylov space, if the initial residual is zero for the boundary dofs, it is always zero during the iterations, and the iterative algorithm behaves exactly as if the boundary dofs are eliminated; see Exercise 33.2. Of course, in practice the way the boundary and interior dofs are enumerated does not matter.

Remark 33.5 (Penalty method). Another way of enforcing Dirichlet conditions without elimination is to use a penalty method. First, one assembles the matrix and the right-hand side of the homogeneous Neumann problem. Then, for each row associated with $a \in \mathcal{A}_h^{\partial}$, one adds ϵ^{-1} to the diagonal entry of the stiffness matrix and $\epsilon^{-1}\sigma_a^{\partial}(g)$ to the right-hand side; see Lions [285], Babuška [35]. If ϵ^{-1} is not large enough, the method suffers from a lack of consistency. As shown in Chapter 37, this problem can be avoided by adding extra boundary terms ensuring consistency; see Nitsche [314]. $\qquad \square$

33.2   Discrete maximum principle

The maximum principle is an important property of scalar second-order elliptic PDEs that sets them apart from higher-order PDEs and systems of PDEs. We focus here on the PDE (33.1) equipped with Dirichlet boundary conditions. Thus, the weak formulation is again (33.2).

Theorem 33.6 (Maximum principle). *Let D be a Lipschitz domain in $\mathbb{R}^d$. Let $\mathbb{d}$, $\boldsymbol{\beta}$, and μ satisfy the assumptions in §31.1.1. Let $f \in L^2(D)$ and let $u \in H^1(D)$ satisfy (33.1). (i) If $\mu = 0$ in D, then $f \leq 0$ a.e. in D implies that $u \leq \operatorname{ess\,sup}_{\partial D} u$ a.e. in D, and $f = 0$ in D implies that $\operatorname{ess\,inf}_{\partial D} u \leq u \leq \operatorname{ess\,sup}_{\partial D} u$ a.e. in D. (ii) Assume $\min(\operatorname{ess\,inf}_D(\mu - \nabla{\cdot}\boldsymbol{\beta}), \operatorname{ess\,inf}_D \mu) > 0$. Then the following holds true a.e. in D:*

$$\min(\operatorname*{ess\,inf}_{\partial D} u, \operatorname*{ess\,inf}_{D}(\mu^{-1}f)) \leq u \leq \max(\operatorname*{ess\,sup}_{\partial D} u, \operatorname*{ess\,sup}_{D}(\mu^{-1}f)). \tag{33.11}$$

Proof. For the proof of (i), see Gilbarg and Trudinger [216, Thm. 8.1], Brezis [89, Prop. 9.29], or Evans [196, §6.4]. Let us prove (ii) by following Brezis [89, Prop. 9.29], i.e., we use *Stampacchia's truncation technique*. Let $G \in C^1(\mathbb{R})$ be such that $G(t) = 0$ for all $t \leq 0$, and $0 < G(t)$, $0 < G'(t) < M$ for all $t > 0$. Let $\overline{K} := \max(\operatorname{ess\,sup}_{\partial D} u, \operatorname{ess\,sup}_D \mu^{-1}f)$ and assume that $\overline{K} < \infty$, otherwise there is nothing to prove. Note that $\zeta_{\overline{K}}(u) := G(u - \overline{K}) \in H^1(D)$ and $\zeta_{\overline{K}}(u)_{|\partial D} = 0$ a.e. (since $(u - \overline{K})_{|\partial D} \leq 0$ a.e. in ∂D), so that $\zeta_{\overline{K}}(u) \in H^1_0(D)$. Testing the weak formulation (33.2) with $\zeta_{\overline{K}}(u)$, we infer that

$$\int_D \left(\|\mathbb{d}^{\frac{1}{2}}\nabla u\|^2_{\ell^2(\mathbb{R}^d)} G'(u - \overline{K}) + (\boldsymbol{\beta}{\cdot}\nabla u + \mu u)\zeta_{\overline{K}}(u) \right) \mathrm{d}x = \int_D f\zeta_{\overline{K}}(u)\,\mathrm{d}x.$$

This proves that

$$\int_D \left((\boldsymbol{\beta}{\cdot}\nabla u + \mu(u - \overline{K}))\zeta_{\overline{K}}(u) \right)\,\mathrm{d}x \leq \int_D (f - \mu\overline{K})\zeta_{\overline{K}}(u)\,\mathrm{d}x \leq 0,$$

where the last bound follows from $f - \mu\overline{K} \leq 0$ and $0 \leq \zeta_{\overline{K}}(u)$ a.e. in D by definition of G. Let $F(t) := \int_0^t G(z)\,\mathrm{d}z$. We have $\int_D (\boldsymbol{\beta}{\cdot}\nabla u)\zeta_{\overline{K}}(u)\,\mathrm{d}x = \int_D (\boldsymbol{\beta}{\cdot}\nabla(\eta_{\overline{K}}(u))\,\mathrm{d}x$ with $\eta_{\overline{K}}(u) := F(u - \overline{K})$. Integrating by parts the advective derivative and since $\eta_{\overline{K}}(u)|_{\partial D} = 0$ (because $F(t) = 0$ for all $t \leq 0$), we infer that

$$\int_D \left(-\eta_{\overline{K}}(u)\nabla{\cdot}\boldsymbol{\beta} + \mu(u - \overline{K})\zeta_{\overline{K}}(u) \right)\,\mathrm{d}x \leq 0. \tag{33.12}$$

The definition of F implies that $F(t) + \int_0^t zG'(z)\,\mathrm{d}z = tG(t)$, which applied to $t := u - \overline{K}$ yields $\eta_{\overline{K}}(u) + \int_0^{u-\overline{K}} zG'(z)\,\mathrm{d}z = (u - \overline{K})\zeta_{\overline{K}}(u)$. Using this identity in (33.12) implies that

$$\int_D \left(\eta_{\overline{K}}(u)(\mu - \nabla{\cdot}\boldsymbol{\beta}) + \mu \int_0^{u-\overline{K}} zG'(z)\,\mathrm{d}z \right)\,\mathrm{d}x \leq 0.$$

Using the assumption $\min(\operatorname{ess\,inf}_D(\mu - \nabla{\cdot}\boldsymbol{\beta}), \operatorname{ess\,inf}_D \mu) > 0$ together with $\eta_{\overline{K}}(u) \geq 0$ and $\int_0^{u-\overline{K}} zG'(z)\,\mathrm{d}z \geq 0$ a.e. in D, we conclude that $\int_D \eta_{\overline{K}}(u)\,\mathrm{d}x = 0$. This means that $\eta_{\overline{K}}(u) = 0$ a.e. in D, i.e., $u - \overline{K} \leq 0$ a.e. in D. □

Remark 33.7 (Sign change). Owing to the linearity of the PDE, Theorem 33.6(i) can be adapted to a sign change, e.g., if $\mu = 0$ in D, $f \geq 0$ a.e. in D implies that $u \geq \operatorname{ess\,inf}_{\partial D} u$ a.e. in D. □

The discrete analogue of Item (i) in Theorem 33.6 is called discrete maximum principle (DMP) (see Ciarlet and Raviart [127] for one of the pioneering works on this topic). As in [127], we only consider linear finite elements on simplicial meshes with homogeneous Dirichlet conditions (see Vejchodský and Šolín [374] for a 1D example where the DMP is shown to hold with higher-order elements). Let $P_{1,0}^{\mathrm{g}}(\mathcal{T}_h)$ be the $H_0^1(D)$-conforming finite element space using linear finite elements. Let $\{\varphi_i\}_{i\in\{1:I\}}$ denote the global shape functions in $P_{1,0}^{\mathrm{g}}(\mathcal{T}_h)$, where I now denotes the number of interior mesh vertices, and let $\mathcal{A} \in \mathbb{R}^{I \times I}$ be the stiffness matrix. For a vector $\mathsf{V} \in \mathbb{R}^I$, the notation $\mathsf{V} \leq 0$ means that $V_i \leq 0$ for all $i \in \{1:I\}$. Then $u_h := \sum_{i\in\{1:I\}} U_i \varphi_i \leq 0$ on D iff $\mathsf{U} \leq 0$ in $\mathbb{R}^I$ since the linear shape functions are nonnegative (this equivalence is no longer valid for higher-order finite elements).

Definition 33.8 (DMP). *We say that the DMP holds true for the discrete problem* (32.5) *with* $V_h := P_{1,0}^{\mathrm{g}}(\mathcal{T}_h)$ *and* $\mathsf{B}_i := \ell(\varphi_i) := \int_D f\varphi_i \, \mathrm{d}x$ *for all* $i \in \{1:I\}$, *if*

$$[\, \mathsf{B} \leq 0 \text{ in } \mathbb{R}^I \,] \implies [\, \mathsf{U} \leq 0 \text{ in } \mathbb{R}^I \,]. \tag{33.13}$$

We now formulate conditions on $\mathcal{A}$ that are equivalent to, or imply, the DMP. Let us recall from Definition 28.16 the notions of Z-matrix and M-matrix. A matrix $\mathcal{A} \in \mathbb{R}^{I \times I}$ is said to be a Z-matrix if $\mathcal{A}_{ij} \leq 0$ for all $i, j \in \{1:I\}$ with $i \neq j$. A matrix $\mathcal{A}$ is said to be a *nonsingular M-matrix* if it is a Z-matrix, invertible, and $(\mathcal{A}^{-1})_{ij} \geq 0$ for all $i, j \in \{1:I\}$.

Lemma 33.9 (Stiffness matrix). (i) *The DMP holds iff* $(\mathcal{A}^{-1})_{ij} \geq 0$ *for all* $i, j \in \{1:I\}$. (ii) *The DMP holds if* $\mathcal{A}$ *is a Z-matrix.*

Proof. The statement (i) follows from the fact that $\mathcal{A}^{-1}\mathsf{B} \leq 0$ for all $\mathsf{B} \leq 0$ iff $(\mathcal{A}^{-1})_{ij} \geq 0$ for all $i, j \in \{1:I\}$. Let us now prove the statement (ii). We follow Jiang and Nochetto [257]. Assume that $\mathcal{A}$ is a Z-matrix. Letting $z^+ := \max(0, z)$, $z^- := z - z^+ = \min(0, z)$ for all $z \in \mathbb{R}$, and $\Pi^+ : \mathbb{R}^I \to \mathbb{R}^I$ be such that $(\Pi^+(\mathsf{V}))_i = V_i^+$ for all $i \in \{1:I\}$, we have for all $\mathsf{V} \in \mathbb{R}^I$,

$$\Pi^+(\mathsf{V})^{\mathsf{T}}\mathcal{A}(\mathsf{V} - \Pi^+(\mathsf{V})) = \sum_{i,j\in\{1:I\}} V_i^+ \mathcal{A}_{ij} V_j^- = \sum_{i,j\in\{1:I\}, i\neq j} V_i^+ \mathcal{A}_{ij} V_j^- \geq 0,$$

since $\mathcal{A}$ is a Z-matrix, i.e., $\mathcal{A}_{ij} \leq 0$ for all $i \neq j$. Let now $\mathsf{B} \leq 0$ and assume that $\mathsf{U} \in \mathbb{R}^I$ solves $\mathcal{A}\mathsf{U} = \mathsf{B}$. We want to prove that $\mathsf{U} \leq 0$. We have

$$0 \geq \Pi^+(\mathsf{U})^{\mathsf{T}}\mathsf{B} = \Pi^+(\mathsf{U})^{\mathsf{T}}\mathcal{A}\mathsf{U} \geq \Pi^+(\mathsf{U})^{\mathsf{T}}\mathcal{A}\Pi^+(\mathsf{U}),$$

which implies that $\Pi^+(\mathsf{U}) = 0$ since $\mathcal{A}$ is positive definite (owing to the coercivity of the bilinear form a). In other words, $\mathsf{U} \leq 0$. $\qquad\square$

Remark 33.10 (Nonsingular M-matrix). A consequence of Lemma 33.9 is that if $\mathcal{A}$ is a Z-matrix, then it is a nonsingular M-matrix. Indeed, if $\mathcal{A}$ is a Z-matrix, Item (ii) implies that the DMP is satisfied, and Item (i) then implies that $(\mathcal{A}^{-1})_{ij} \geq 0$ for all $i, j \in \{1:I\}$. This shows that $\mathcal{A}$ is a nonsingular M-matrix (see Definition 28.16); see also Exercises 33.3 and 33.5. $\qquad\square$

Lemma 33.9 shows that a sufficient condition for the DMP to hold is that the stiffness matrix $\mathcal{A}$ is a Z-matrix. Since ensuring this property on general diffusion-advection-reaction PDEs is delicate, we continue the discussion by focusing on the Poisson equation, i.e., $\mathbb{d} := \mathbb{I}_d$, $\boldsymbol{\beta} := \mathbf{0}$, and $\mu := 0$. For all $i \in \{1:I\}$, let $\mathcal{I}(i) := \{j \in \{1:I\} \mid \varphi_i\varphi_j \not\equiv 0\}$ and $\mathcal{I}^*(i) := \mathcal{I}(i)\backslash\{i\}$.

Definition 33.11 (Weakly acute meshes). *A simplicial mesh is said to be* weakly acute *if the stiffness matrix of the Laplacian is a Z-matrix, i.e.,*

$$\int_D \nabla\varphi_i \cdot \nabla\varphi_j \, \mathrm{d}x \leq 0, \qquad \forall i \in \{1:I\}, \ \forall j \in \mathcal{I}^*(i). \tag{33.14}$$

The condition (33.14) is always satisfied in dimension $d = 1$. In higher dimension, (33.14) boils down to a geometric restriction on the mesh. Notice that for all $i \in \{1{:}I\}$ and all $j \in \mathcal{I}^*(i)$, the collection of the mesh cells having both $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ as vertices, say $\mathcal{T}_{ij}$, is nonempty. Let $K \in \mathcal{T}_{ij}$. Let $F_{K,i}$ (resp. $F_{K,j}$) be the face of K opposite to $\boldsymbol{z}_i$ (resp. $\boldsymbol{z}_j$). Let $\boldsymbol{n}_{K,i}$, $\boldsymbol{n}_{K,j}$ be the two unit normal vectors to $F_{K,i}$ and $F_{K,j}$, respectively, pointing outward. Let $\boldsymbol{z}_{K,i}^*$ (resp. $\boldsymbol{z}_{K,j}^*$) be the ℓ^2-orthogonal projection of $\boldsymbol{z}_i$ onto $F_{K,i}$ (resp. $F_{K,j}$). We have $\varphi_{i|K}(\boldsymbol{x}) = h_{K,i}^{-1}(\boldsymbol{z}_{K,i}^* - \boldsymbol{x}){\cdot}\boldsymbol{n}_{K,i}$ and $\nabla\varphi_{i|K} = -h_{K,i}^{-1}\boldsymbol{n}_{K,i}$ with $h_{K,i} := \|\boldsymbol{z}_i - \boldsymbol{z}_{K,i}^*\|_{\ell^2}$. Recalling that $|K| = \frac{1}{d}|F_{K,i}|h_{K,i}$ and setting $\cos(\alpha_{K,ij}) := -\boldsymbol{n}_{K,i}{\cdot}\boldsymbol{n}_{K,j}$ (i.e., $\alpha_{K,ij} \in (0,\pi)$ is the dihedral angle between $F_{K,i}$ and $F_{K,j}$), we infer that

$$\int_K \nabla\varphi_i{\cdot}\nabla\varphi_j \, \mathrm{d}x = -\frac{|F_{K,i}||F_{K,j}|}{d^2|K|}\cos(\alpha_{K,ij}). \tag{33.15}$$

Thus, a sufficient condition for (33.14) to hold true, that is, for the mesh to be weakly acute, is that for all $K \in \mathcal{T}_h$ and for every pair of distinct faces of K, say F, F', we have $\boldsymbol{n}_{K|F}{\cdot}\boldsymbol{n}_{K|F'} \leq 0$, i.e., the dihedral angle between F and F' is in $(0, \frac{\pi}{2}]$. We say in this case that the mesh is *nonobtuse*; see also Brandts et al. [84]. However, a weakly acute mesh is not necessarily nonobtuse since (33.14) only requires that $\sum_{K \in \mathcal{T}_{ij}} \int_K \nabla\varphi_i{\cdot}\nabla\varphi_j \, \mathrm{d}x \leq 0$, whereas (33.15) requires that *each term* in the sum is nonpositive. This leads us to look for a necessary and sufficient condition so that (33.14) holds true.

Lemma 33.12 (Geometric identity). *The following holds true (with the convention that $|F_{K,i} \cap F_{K,j}| = 1$ for $d = 2$):*

$$\int_K \nabla\varphi_i{\cdot}\nabla\varphi_j \, \mathrm{d}x = -\frac{|F_{K,i} \cap F_{K,j}|}{d(d-1)}\cot(\alpha_{K,ij}). \tag{33.16}$$

Proof. Since K is fixed, we drop the index K in the proof. Since F_j is a simplex in $\mathbb{R}^{d-1}$, we have $|F_j| = \frac{1}{d-1}|F_i \cap F_j|h_{i,j}$ where $h_{i,j} := \|\boldsymbol{z}_{i,j}^* - \boldsymbol{z}_i\|_{\ell^2}$ and $\boldsymbol{z}_{i,j}^*$ is the projection of $\boldsymbol{z}_i$ onto $F_i \cap F_j$. Thus, we have

$$\nabla\varphi_i{\cdot}\nabla\varphi_{j|K} = -\frac{\cos(\alpha_{ij})}{h_i h_j} = -\frac{\cos(\alpha_{ij})|F_j|}{h_i d|K|} = -\frac{|F_i \cap F_j|\cos(\alpha_{ij})}{d(d-1)|K|}\frac{h_{i,j}}{h_i},$$

and it remains to show that $\frac{h_{i,j}}{h_i} = \frac{1}{\sin(\alpha_{ij})}$. Letting $\boldsymbol{m}_j' = \boldsymbol{n}_j - (\boldsymbol{n}_j{\cdot}\boldsymbol{n}_i)\boldsymbol{n}_i$ so that $\|\boldsymbol{m}_j'\|_{\ell^2} = \sin(\alpha_{ij})$, we set $\boldsymbol{n}_j' := \frac{1}{\sin(\alpha_{ij})}(\boldsymbol{n}_j - (\boldsymbol{n}_j{\cdot}\boldsymbol{n}_i)\boldsymbol{n}_i)$. The set $\{\boldsymbol{n}_i, \boldsymbol{n}_j'\}$ is an orthonormal basis of the plane orthogonal to the $(d-2)$-dimensional manifold $F_i \cap F_j$. Let $\boldsymbol{z}_k$ be one of the $(d-1)$ vertices in $F_i \cap F_j$. By definition, $\boldsymbol{z}_i^* - \boldsymbol{z}_k = \boldsymbol{z}_i - \boldsymbol{z}_k - ((\boldsymbol{z}_i - \boldsymbol{z}_k){\cdot}\boldsymbol{n}_i)\boldsymbol{n}_i$ and $\boldsymbol{z}_{i,j}^* - \boldsymbol{z}_k = \boldsymbol{z}_i - \boldsymbol{z}_k - ((\boldsymbol{z}_i - \boldsymbol{z}_k){\cdot}\boldsymbol{n}_i)\boldsymbol{n}_i - ((\boldsymbol{z}_i - \boldsymbol{z}_k){\cdot}\boldsymbol{n}_j')\boldsymbol{n}_j'$. Hence, we have

$$h_{i,j}^2 = \|\boldsymbol{z}_i^* - \boldsymbol{z}_i - ((\boldsymbol{z}_i - \boldsymbol{z}_k){\cdot}\boldsymbol{n}_j')\boldsymbol{n}_j'\|_{\ell^2}^2 = h_i^2 + |(\boldsymbol{z}_i - \boldsymbol{z}_k){\cdot}\boldsymbol{n}_j'|^2.$$

But $|(\boldsymbol{z}_i - \boldsymbol{z}_k){\cdot}\boldsymbol{n}_j'|^2 = \frac{\cos^2(\alpha_{ij})}{\sin^2(\alpha_{ij})}h_i^2$ since $(\boldsymbol{z}_i - \boldsymbol{z}_k){\cdot}\boldsymbol{n}_j = 0$ and $|(\boldsymbol{z}_i - \boldsymbol{z}_k){\cdot}\boldsymbol{n}_i|^2 = h_i^2$. Thus, we have $h_{i,j}^2 = \left(1 + \frac{\cos^2(\alpha_{ij})}{\sin^2(\alpha_{ij})}\right)h_i^2 = \frac{1}{\sin(\alpha_{ij})^2}h_i^2$. $\square$

Corollary 33.13 (Necessary and sufficient condition). *The condition* (33.14) *is fulfilled iff*

$$\sum_{K \in \mathcal{T}_{ij}} |F_{K,i} \cap F_{K,j}|\cot(\alpha_{K,ij}) \geq 0, \qquad \forall i \in \{1{:}I\}, \ \forall j \in \mathcal{I}^*(i). \tag{33.17}$$

If $d = 2$, $\mathcal{T}_{ij}$ consists of only two cells, and the identity $(\cot(\alpha)+\cot(\beta))\sin(\alpha)\sin(\beta) = \sin(\alpha+\beta)$ shows that (33.14) holds true iff the sum of the two angles opposite to any interior face is less than or equal to π; see Xu and Zikatanov [396, Eq. (2.5)].

Remark 33.14 (Obstructions). It is noticed in Brandts et al. [83, §5.2] that (33.14) cannot hold true in dimension $d \geq 4$, and that the strict inequality cannot hold true in dimension three on Cartesian meshes. ◻

Remark 33.15 (Nonlinear stabilization). An alternative approach to enforce the discrete maximum principle that avoids geometric requirements on the mesh is to add a nonlinear viscosity term to the discrete problem; see Burman and Ern [98, 99], Barrenechea et al. [45]. ◻

33.3 Discrete problem with quadratures

In this section, we study the influence of quadratures when approximating a scalar elliptic PDE by means of finite elements.

33.3.1 Continuous and discrete settings

For simplicity, we drop the lower-order terms in the PDE which becomes $-\nabla{\cdot}(\mathrm{d}\nabla u) = f$ in D, and we consider homogeneous Dirichlet boundary conditions, i.e., $u = 0$ on ∂D. Thus, the weak formulation is posed in $V := H_0^1(D)$, which we equip with the norm $\|v\|_V := \|\nabla v\|_{\boldsymbol{L}^2(D)} = |v|_{H^1(D)}$, and the bilinear and linear forms are

$$a(v, w) = \int_D (\mathrm{d}\nabla v){\cdot}\nabla w \, \mathrm{d}x, \qquad \ell(w) = \int_D f w \, \mathrm{d}x. \tag{33.18}$$

We assume that the model problem is well-posed, i.e., the conditions of the BNB theorem (or the Lax–Milgram lemma) are fulfilled. Hence, there is $\alpha > 0$ s.t. $\alpha\|\nabla v\|_{\boldsymbol{L}^2(D)} \leq \sup_{w \in V} \frac{|a(v,w)|}{\|\nabla w\|_{\boldsymbol{L}^2(D)}}$ for all $v \in V$.

We consider the H_0^1-conforming finite element space $V_h := P_{k,0}^{\mathrm{g}}(\mathcal{T}_h) \subset H_0^1(D)$ defined in (32.4c). If the integrals defining the forms a and ℓ are evaluated exactly in the discrete problem, well-posedness follows automatically in the setting of the Lax–Milgram lemma if a is V-coercive, and well-posedness holds in the setting of the BNB theorem if the bilinear form a satisfies the following uniform inf-sup condition on $V_h{\times}V_h$ for all $h \in \mathcal{H}$ (see Chapter 32):

$$\exists \alpha_0 > 0, \quad \forall v_h \in V_h, \quad \alpha_0\|\nabla v_h\|_{\boldsymbol{L}^2(D)} \leq \sup_{w_h \in V_h} \frac{|a(v_h, w_h)|}{\|\nabla w_h\|_{\boldsymbol{L}^2(D)}}. \tag{33.19}$$

Moreover, in both situations one obtains H^1-seminorm error estimates of order h^r if u is in $H^{1+r}(D)$ with $r \in (0, k]$. But in practice the integrals defining the forms a and ℓ have to be evaluated approximately by means of quadratures as described in Chapter 30. Therefore, a natural question is whether the quadratures impact the well-posedness of the discrete problem and its error analysis.

Recalling Definition 30.1, we consider a quadrature in $\widehat{K}$ with nodes $\{\widehat{\boldsymbol{\xi}}_l\}_{l \in \{1:l_{\mathcal{Q}}\}}$ and weights $\{\widehat{\omega}_l\}_{l \in \{1:l_{\mathcal{Q}}\}}$. The largest integer k such that the quadrature is exact for any polynomial in $\mathbb{P}_{k,d}$ is called quadrature order and is denoted by $k_{\mathcal{Q}}$. Recalling Proposition 30.2, we construct a quadrature in every mesh cell $K \in \mathcal{T}_h$ by setting $\{\boldsymbol{\xi}_{lK} := \boldsymbol{T}_K(\widehat{\boldsymbol{\xi}}_l)\}_{l \in \{1:l_{\mathcal{Q}}\}}$ for the nodes and $\{\omega_{lK} := \widehat{\omega}_l |\det(\mathbb{J}_K(\widehat{\boldsymbol{\xi}}_l))|\}_{l \in \{1:l_{\mathcal{Q}}\}}$ for the weights, where $\boldsymbol{T}_K : \widehat{K} \to K$ is the geometric mapping. This quadrature allows us to approximate the integral of any continuous function ϕ in K as $\int_K \phi(\boldsymbol{x}) \, \mathrm{d}x \approx \sum_{l \in \{1:l_{\mathcal{Q}}\}} \omega_{lK} \, \phi(\boldsymbol{\xi}_{lK})$. The quadrature error $E_K : C^0(K) \to \mathbb{R}$ is defined by setting $E_K(\phi) := \int_K \phi(\boldsymbol{x}) \, \mathrm{d}x - \sum_{l \in \{1:l_{\mathcal{Q}}\}} \omega_{lK} \, \phi(\boldsymbol{\xi}_{lK})$.

33.3.2   Well-posedness with quadratures

For simplicity, we assume that the mesh $\mathcal{T}_h$ is affine and that the diffusion coefficients d_{ij} and the source term f are continuous in every mesh cell $K \in \mathcal{T}_h$. The use of a quadrature in every mesh cell to evaluate the exact forms a and ℓ defined in (33.18) leads to the following approximate forms:

$$a_{\mathcal{Q}}(v_h, w_h) := \sum_{K \in \mathcal{T}_h} \sum_{l \in \{1:l_{\mathcal{Q}}\}} \omega_{lK}\big(\mathrm{d}(\boldsymbol{\xi}_{lK})\nabla v_h(\boldsymbol{\xi}_{lK})\big)\cdot\nabla w_h(\boldsymbol{\xi}_{lK}), \tag{33.20a}$$

$$\ell_{\mathcal{Q}}(w_h) := \sum_{K \in \mathcal{T}_h} \sum_{l \in \{1:l_{\mathcal{Q}}\}} \omega_{lK} f(\boldsymbol{\xi}_{lK}) w_h(\boldsymbol{\xi}_{lK}), \tag{33.20b}$$

for all $(v_h, w_h) \in V_h \times V_h$. It is not possible in general to extend $a_{\mathcal{Q}}$ and $\ell_{\mathcal{Q}}$ to $H_0^1(D)$, since functions in $H_0^1(D)$ are not necessarily defined pointwise. The discrete problem with quadratures is formulated as follows:

$$\begin{cases} \text{Find } u_h \in V_h \text{ such that} \\ a_{\mathcal{Q}}(u_h, w_h) = \ell_{\mathcal{Q}}(w_h), \quad \forall w_h \in V_h. \end{cases} \tag{33.21}$$

Lemma 33.16 (Well-posedness). *Assume that $\widehat{P} \subset \mathbb{P}_{l,d}$ for some integer $l \geq 1$. Assume that $k_{\mathcal{Q}} \geq 2l - 2$ and that $\mathrm{d} \in \mathbb{W}^{1,\infty}(\mathcal{T}_h) := W^{1,\infty}(\mathcal{T}_h; \mathbb{R}^{d \times d})$. Assume that the inf-sup condition (33.19) is satisfied. Define the length scale $\ell_0 := \frac{\alpha_0}{|\mathrm{d}|_{\mathbb{W}^{1,\infty}(\mathcal{T}_h)}}$. (i) There is $\varrho > 0$ such that for all $h \in \mathcal{H} \cap (0, \varrho\ell_0]$, the approximate bilinear form $a_{\mathcal{Q}}$ satisfies an inf-sup condition on $V_h \times V_h$ with constant $\alpha_{\mathcal{Q}} := \frac{1}{2}\alpha_0$. (ii) The discrete problem (33.21) is well-posed.*

Proof. We only need to establish the item (i) since the item (ii) follows from (i). Let $v_h \in V_h$. Owing to (33.19), we have

$$\sup_{w_h \in V_h} \frac{|a_{\mathcal{Q}}(v_h, w_h)|}{\|\nabla w_h\|_{\boldsymbol{L}^2(D)}} \geq \sup_{w_h \in V_h} \frac{|a(v_h, w_h)|}{\|\nabla w_h\|_{\boldsymbol{L}^2(D)}} - \sup_{w_h \in V_h} \frac{|(a - a_{\mathcal{Q}})(v_h, w_h)|}{\|\nabla w_h\|_{\boldsymbol{L}^2(D)}}$$

$$\geq \alpha_0 \|\nabla v_h\|_{\boldsymbol{L}^2(D)} - \sup_{w_h \in V_h} \frac{|(a - a_{\mathcal{Q}})(v_h, w_h)|}{\|\nabla w_h\|_{\boldsymbol{L}^2(D)}}.$$

Recalling that $E_K(\cdot)$ denotes the quadrature error, we have

$$(a - a_{\mathcal{Q}})(v_h, w_h) = \sum_{K \in \mathcal{T}_h} E_K\big((\mathrm{d}\nabla v_h)\cdot\nabla w_h\big).$$

For all $i, j \in \{1:d\}$, let us set $p := \partial_i v_h \partial_j w_h$. Since $\partial_{i'}\widehat{v}_h \circ \boldsymbol{T}_K \in \mathbb{P}_{l-1,d}$ and $\mathbb{J}_K$ is constant over $\widehat{K}$ (recall that the mesh is affine) we have $\partial_i v_h \circ \boldsymbol{T}_K = \sum_{i' \in \{1:d\}} \mathbb{J}_{K,ii'}^{-\mathsf{T}}(\partial_{i'}\widehat{v}_h) \in \mathbb{P}_{l-1,d}$. A similar argument shows that $\partial_j w_h \circ \boldsymbol{T}_K \in \mathbb{P}_{l-1,d}$. This proves that $p \circ \boldsymbol{T}_K \in \mathbb{P}_{2l-2,d}$. We now use Lemma 30.10 with $\phi := \mathrm{d}_{ij}$ and $p := \partial_i v_h \partial_j w_h$. The assumptions of the lemma are met with $m := 1$ and $n := 2l - 2$ (so that $n + m - 1 = 2l - 2 \leq k_{\mathcal{Q}}$). Since $\|\partial_i v_h \partial_j w_h\|_{L^1(K)} \leq \|\partial_i v_h\|_{L^2(K)}\|\partial_j w_h\|_{L^2(K)}$, we infer that there is $c_{\mathcal{Q}}$ such that for all $v_h, w_h \in V_h$, all $K \in \mathcal{T}_h$, and all $h \in \mathcal{H}$,

$$|E_K((\mathrm{d}\nabla v_h)\cdot\nabla w_h)| \leq c_{\mathcal{Q}} h_K |\mathrm{d}|_{\mathbb{W}^{1,\infty}(K)}\|\nabla v_h\|_{\boldsymbol{L}^2(K)}\|\nabla w_h\|_{\boldsymbol{L}^2(K)}.$$

(Notice that it is natural that the above estimate depends on $|\mathrm{d}|_{\mathbb{W}^{1,\infty}(K)}$ because $E_K((\mathrm{d}\nabla v_h)\cdot\nabla w_h)$ is zero if d is constant over K.) Owing to the Cauchy–Schwarz inequality, we infer that

$$|(a - a_{\mathcal{Q}})(v_h, w_h)| \leq c_{\mathcal{Q}} h |\mathrm{d}|_{\mathbb{W}^{1,\infty}(\mathcal{T}_h)}\|\nabla v_h\|_{\boldsymbol{L}^2(D)}\|\nabla w_h\|_{\boldsymbol{L}^2(D)}.$$

Taking $\varrho := \frac{1}{2c_{\mathcal{Q}}}$ and assuming that $h \in (0, \varrho\ell_0]$ with $\ell_0 := \frac{\alpha_0}{|\mathrm{d}|_{\mathbb{W}^{1,\infty}(\mathcal{T}_h)}}$ implies that $c_{\mathcal{Q}} h |\mathrm{d}|_{\mathbb{W}^{1,\infty}(\mathcal{T}_h)} \leq \frac{\alpha_0}{2}$. Combining the above estimates then yields $\sup_{w_h \in V_h} \frac{|a_{\mathcal{Q}}(v_h, w_h)|}{\|\nabla w_h\|_{\boldsymbol{L}^2(D)}} \geq \frac{\alpha_0}{2}\|\nabla v_h\|_{\boldsymbol{L}^2(D)}$, which is the expected bound. $\qquad\square$

33.3.3   Error analysis with quadratures

Theorem 33.17 (Error estimate). *Assume* $\mathbb{P}_{k,d} \subset \widehat{P} \subset \mathbb{P}_{l,d}$ *with the integers* $l \geq k \geq 1$. *Assume that* $k_{\mathcal{Q}} \geq l + k - 2$, $\mathbb{d} \in \mathbb{W}^{k,\infty}(\mathcal{T}_h)$, *and* $f \in W^{k,\infty}(\mathcal{T}_h)$. *Assume that* (33.21) *is well-posed and let* $\alpha_{\mathcal{Q}}$ *denote the inf-sup constant of* $a_{\mathcal{Q}}$ *on* $V_h \times V_h$ *(see Lemma 33.16). Assume that* $u \in H^{k+1}(D)$. *There is* c *s.t. for all* $h \in \mathcal{H} \cap (0, \varrho \ell_0]$,

$$|u - u_h|_{H^1(D)} \leq c\,h^k\big(|u|_{H^{k+1}(D)} + \alpha_{\mathcal{Q}}^{-1}(C_{\mathcal{Q}}(\mathbb{d}, u) + C_{\mathcal{Q}}(f))\big), \tag{33.22}$$

where we have set

$$C_{\mathcal{Q}}(\mathbb{d}, u) := \sum_{m \in \{0:k\}} |\mathbb{d}|_{\mathbb{W}^{k-m,\infty}(\mathcal{T}_h)}|u|_{H^{m+1}(D)},$$

$$C_{\mathcal{Q}}(f) := |D|^{\frac{1}{2}} \max(\ell_D|f|_{W^{k,\infty}(\mathcal{T}_h)}, |f|_{W^{k-1,\infty}(\mathcal{T}_h)}),$$

and ℓ_D *is a length scale associated with* D, *e.g.,* $\ell_D := \operatorname{diam}(D)$.

Proof. Since the discrete problem with quadratures is stable, we bound the error using Strang's first lemma (Lemma 27.12). We consider $v_h := \mathcal{I}_{h0}^{\mathrm{g,av}}(u)$, where $\mathcal{I}_{h0}^{\mathrm{g,av}}$ is the quasi-interpolation operator introduced in §22.4.2. Recalling that $\|\cdot\|_V := \|\nabla(\cdot)\|_{\mathbf{L}^2(D)}$, this yields

$$\|\nabla(u - u_h)\|_{\mathbf{L}^2(D)} \leq c\big(\|\nabla(u - \mathcal{I}_{h0}^{\mathrm{g,av}}(u))\|_{\mathbf{L}^2(D)} + \alpha_{\mathcal{Q}}^{-1}\|\delta_h^{\mathrm{St1}}(v_h)\|_{V_h'}\big),$$

where the consistency error $\delta_h^{\mathrm{St1}}(v_h) \in V_h'$ is such that

$$\langle \delta_h^{\mathrm{St1}}(v_h), w_h \rangle_{V_h', V_h} := (a - a_{\mathcal{Q}})(v_h, w_h) - (\ell - \ell_{\mathcal{Q}})(w_h).$$

(1) Bound on $(a - a_{\mathcal{Q}})$. Since $(a_{\mathcal{Q}} - a)(v_h, w_h) = \sum_{K \in \mathcal{T}_h} E_K\big((\mathbb{d}\nabla v_h)\cdot\nabla w_h\big)$, we can use the bound (30.6) from Lemma 30.10 with $\phi := \mathbb{d}_{ij}\partial_i v_h$, $p := \partial_j w_h$, $m := k$, and $n := l - 1$ (so that $n + m - 1 = l + k - 2 \leq k_{\mathcal{Q}}$) for all $i, j \in \{1:d\}$ to infer that

$$|E_K\big((\mathbb{d}\nabla v_h)\cdot\nabla w_h\big)| \leq c \sum_{i,j\in\{1:d\}} h_K^k |\mathbb{d}_{ij}\partial_i v_h|_{W^{k,\infty}(K)}\|\partial_j w_h\|_{L^1(K)}.$$

Combining the Leibniz product rule with the inverse inequality (12.3) (with $p := \infty$ and $r := 2$) leads to

$$|\mathbb{d}_{ij}\partial_i v_h|_{W^{k,\infty}(K)} \leq c \sum_{m\in\{0:k\}} |\mathbb{d}_{ij}|_{W^{k-m,\infty}(K)}|K|^{-\frac{1}{2}}|\partial_i v_h|_{H^m(K)}.$$

Applying again an inverse inequality, we infer that

$$|E_K\big((\mathbb{d}\nabla v_h)\cdot\nabla w_h\big)| \leq c\,h_K^k \sum_{m\in\{0:k\}} |\mathbb{d}|_{\mathbb{W}^{k-m,\infty}(K)}\|\nabla v_h\|_{\mathbf{H}^m(K)}\|\nabla w_h\|_{\mathbf{L}^2(K)}.$$

As a result, we have

$$\|(a - a_{\mathcal{Q}})(v_h, w_h)\|_{V_h'} \leq c\,h^k \sum_{m\in\{0:k\}} |\mathbb{d}|_{\mathbb{W}^{k-m,\infty}(\mathcal{T}_h)}|u|_{H^{m+1}(D)},$$

where we used the estimate $|v_h|_{H^m(\mathcal{T}_h)} = |\mathcal{I}_{h0}^{\mathrm{g,av}}(u)|_{H^m(\mathcal{T}_h)} \leq c|u|_{H^m(D)}$ which follows from Theorem 22.14.

(2) Bound on $(\ell - \ell_{\mathcal{Q}})$. We have $(\ell_{\mathcal{Q}} - \ell)(w_h) = \sum_{K\in\mathcal{T}_h} E_K(f w_h)$. We cannot apply the bound (30.6) from Lemma 30.10 with $\phi := f$, $p := w_h$, $m := k$, and $n := l$ since $n + m - 1 = l + k - 1$

may be larger than $k_{\mathcal{Q}}$. Instead, we use the bound (30.7) with $m := k$ and $n := l$ (since $n + m - 2 = l + k - 2 \leq k_{\mathcal{Q}}$) yielding

$$|(\ell_{\mathcal{Q}} - \ell)(w_h)| \leq c \sum_{K \in \mathcal{T}_h} h_K^k (|f|_{W^{k,\infty}(K)} \|w_h\|_{L^1(K)} + |f|_{W^{k-1,\infty}(K)} \|\nabla w_h\|_{\boldsymbol{L}^1(K)})$$

$$\leq c\, h^k \max(C_{\mathrm{PS}}^{-1} \ell_D |f|_{W^{k,\infty}(\mathcal{T}_h)}, |f|_{W^{k-1,\infty}(\mathcal{T}_h)}) |D|^{\frac{1}{2}} \|\nabla w_h\|_{\boldsymbol{L}^2(D)},$$

where C_{PS} is the global Poincaré–Steklov constant from (3.11). The rest of the proof follows readily. □

Remark 33.18 (Literature). The above analysis is inspired from Ciarlet [124, §4.1], Ciarlet and Raviart [126], Dautray and Lions [154, §XII.5]. It is possible to refine the analysis by assuming $f \in W^{k,q}(\mathcal{T}_h)$ with $q > \frac{d}{k}$ and $q \geq 2$. The analysis with approximate Neumann conditions can be done by assuming that surface quadratures of order at least $k + l - 1$ are used to approximate the boundary integrals; see [154, §XII.5]. □

Exercises

Exercise 33.1 (Regularity assumption). Let u_h solve (33.5). Assume that $u \in H^{1+r}(D)$ with $r \in (0,k]$. Prove that $\|u - u_h\|_{H^1(D)} \leq c(h^r |u|_{H^{1+r}(D)} + (\sum_{F \in \mathcal{F}_h^\partial} h_F^{-1} \|g - g_h\|_{L^2(F)}^2)^{\frac{1}{2}})$. (*Hint*: consider $v_h := \mathcal{I}_{h0}^{\mathrm{g,av}}(u) + \sum_{a \in \mathcal{A}_h^\partial} \sigma_a^\partial(g)\varphi_a$, and follow the proof of Theorem 22.14 to bound $\|u - v_h\|_{H^1(D)}$.)

Exercise 33.2 (Non-homogeneous Dirichlet). Let $\mathcal{A}$ denote the system matrix in (33.10). Let $\mathsf{R} \in \mathbb{R}^I$ and let $k \geq 1$. Consider the Krylov space $S_k := \mathrm{span}\{\mathsf{R}, \mathcal{A}\mathsf{R}, \ldots, \mathcal{A}^{k-1}\mathsf{R}\}$. For all $\mathsf{V} \in \mathbb{R}^I$, write $\mathsf{V} := (\mathsf{V}^\circ, \mathsf{V}^\partial)^\mathsf{T}$. Assume that $\mathsf{R}^\partial = 0$. (i) Prove that $\mathsf{Y}^\partial = 0$ for all $\mathsf{Y} \in S_k$. (ii) Prove that if $\mathcal{A}^{\circ\circ}$ is symmetric, the restriction of $\mathcal{A}$ to S_k is symmetric.

Exercise 33.3 (DMP). Assume that the stiffness matrix is a Z-matrix. Assume the following: (i) $\mathcal{A}_{ii} \geq -\sum_{j \neq i} \mathcal{A}_{ij}$ for all $i \in \{1{:}I\}$; (ii) $\exists i_* \in \{1{:}I\}$ such that $\mathcal{A}_{i_* i_*} > -\sum_{j \neq i_*} \mathcal{A}_{i_* j}$; (iii) For all $i \in \{1{:}I\}$, $i \neq i_*$, there exists a path $[i =: i_1, \ldots, i_J := i_*]$ such that $\mathcal{A}_{i_j i_{j+1}} < 0$ for all $j \in \{1{:}J{-}1\}$. Prove that $\mathcal{A}$ is a nonsingular M-matrix. (*Hint*: let $\mathsf{B} \leq 0$, let $\mathsf{U} := \mathcal{A}^{-1}\mathsf{B}$, and proceeding by contradiction, assume that there is $i \in \{1{:}I\}$ s.t. $U_i = \max_{j \in \{1{:}I\}} U_j > 0$.)

Exercise 33.4 (Obtuse mesh). The mesh shown in Figure 33.1 contains three interior nodes with coordinates $\boldsymbol{z}_1 := (1,1)$, $\boldsymbol{z}_2 := (3,1)$, and $\boldsymbol{z}_3 := (2, \frac{3}{2})$. The sum of the two angles opposite the edge linking $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ is larger than π. (i) Assemble the $3{\times}3$ stiffness matrix $\mathcal{A}$ generated by the three shape functions associated with the three interior nodes $\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3$. Is $\mathcal{A}$ a Z-matrix? (*Hint*: the local stiffness matrix is translation- and scale-invariant, there are four shapes of triangles in the mesh, and one can work on triangles with vertices $((0,0),(1,0),(0,1))$, $((0,0),(1,0),(0,\frac{1}{2}))$, $((-1,0),(1,0),(0,\frac{1}{2}))$, and $((-1,1),(1,1),(0,1))$.) (ii) Compute $\mathcal{A}^{-1}$. Is $\mathcal{A}$ an M-matrix?

Exercise 33.5 (1D DMP). Consider the equation $\mu u + \beta u' - \nu u'' = f$ in $D := (0,1)$. Let $\mathcal{T}_h$ be the uniform mesh composed of the cells $[ih, (i{+}1)h]$, $\forall i \in \{0{:}I\}$, with uniform meshsize $h := \frac{1}{I+1}$. Assume $\mu \in \mathbb{R}_+$, $\beta \in \mathbb{R}$, $\nu \in \mathbb{R}_+$ and $f \in L^1(D)$. Let $u_h := \sum_{i \in \{0{:}I+1\}} U_i \varphi_i \in P_1^{\mathrm{g}}(\mathcal{T}_h)$ be such that $\int_D ((\mu u_h + \beta u_h')\varphi_i + \nu u_h' \varphi_i') \, \mathrm{d}x = \int_D f\varphi_i \, \mathrm{d}x$ for all $i \in \{1{:}I\}$. Let $F_i := \int_D f\varphi_i \, \mathrm{d}x / \int_D \varphi_i \, \mathrm{d}x$. Assume that $\frac{\nu}{h} \geq \frac{|\beta|}{2} + \frac{\mu h}{6}$. (i) Show that $\min(U_{i-1}, U_{i+1}, \frac{F_i}{\mu}) \leq U_i \leq \max(U_{i-1}, U_{i+1}, \frac{F_i}{\mu})$ for all $i \in \{1{:}I\}$. (*Hint*: write the linear system as $\mu U_i + \alpha_{i-1}(\mu, \beta, \nu)(U_i - U_{i-1}) + \alpha_{i+1}(\mu, \beta, \nu)(U_i - $
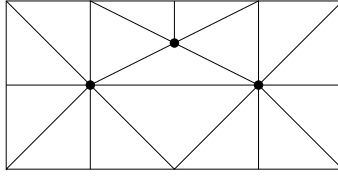
Figure 33.1: Illustration for Exercise 33.4.

$U_{i+1}) = F_i$.) (ii) Show that $\min(U_0, U_{I+1}, \frac{\min_{j\in\{1:I\}} F_j}{\mu}) \leq U_i \leq \max(U_0, U_{I+1}, \frac{\max_{j\in\{1:I\}} F_j}{\mu})$ for all $i \in \{1{:}I\}$.

Exercise 33.6 (1D DMP, pure diffusion). Let $D := (0,1)$, $f \in L^\infty(D)$, and a nonuniform mesh $\mathcal{T}_h$ of D with nodes $\{x_i\}_{i\in\{0:I+1\}}$. Let $u_h \in P_1^g(\mathcal{T}_h)$ be s.t. $u_h(0) = a$, $u_h(1) = b$, and $\int_D u_h' v_h' \, \mathrm{d}x = \int_D f v_h \, \mathrm{d}x$ for all $v_h \in P_{1,0}^g(\mathcal{T}_h)$. (i) Show that $\max_{x\in D} u_h(x) \leq \max(a,b) + \frac{1}{4} \operatorname{ess\,sup}_{x\in D} f(x)$. (*Hint*: test with $\phi_h \in P_{1,0}^g(\mathcal{T}_h)$ s.t. $\phi_{h|[0,x_i]} := \frac{x}{x_i}$ and $\phi_{h|[x_i,1]} := \frac{1-x}{1-x_i}$ for all $i \in \{1{:}I\}$.) (ii) Let ϕ_h be the function defined in the hint. Compute $-\partial_{xx}\phi_h$. Comment on the result.

Exercise 33.7 (Maximum principle). Let D be a bounded Lipschitz domain in $\mathbb{R}^d$. Let $\boldsymbol{x}_0 \in D$ and $R \in \mathbb{R}$ be s.t. $\max_{\boldsymbol{x}\in D} \|\boldsymbol{x} - \boldsymbol{x}_0\|_{\ell^2} \leq R$. (i) Let $\phi(\boldsymbol{x}) := -\frac{1}{2d}\|\boldsymbol{x} - \boldsymbol{x}_0\|_{\ell^2}^2$. Compute $-\Delta\phi$. Give an upper bound on $\max_{\boldsymbol{x}\in D} \phi(\boldsymbol{x})$ and a lower bound on $\min_{\boldsymbol{x}\in\partial D} \phi(\boldsymbol{x})$. (ii) Let $f \in L^\infty(D)$ and let $u \in H^1(D)$ solve $-\Delta u = f$. Let $M := \operatorname{ess\,sup}_{\boldsymbol{x}\in D} f(\boldsymbol{x})$. Give an upper bound on $-\Delta(u - M\phi)$. (iii) Prove that $\max_{\boldsymbol{x}\in D} u(\boldsymbol{x}) \leq \max_{\boldsymbol{x}\in\partial D} u(\boldsymbol{x}) + M_+ \frac{R^2}{2d}$ with $M_+ := \max(M, 0)$. (*Hint*: use (i) from Theorem 33.6.)

Chapter 34

A posteriori error analysis

An a posteriori error estimate is an upper bound on the approximation error that can be computed by using only the discrete solution and the problem data. Such an estimate can serve the twofold purpose of judging the quality of the discrete solution and of guiding an adaptive procedure that modifies the discretization iteratively in order to diminish the approximation error. A posteriori error estimates should involve constants that are all computable, or sharp estimates from above of these constants. For the purpose of mesh adaptation, the error estimate should be a sum of local contributions (usually called indicators) that can be used to mark those cells requiring further refinement at the next iteration of the adaptive procedure. It is then important that the indicators represent a local lower bound on the error. A posteriori and a priori error estimates are conceptually different. A priori error estimates rely on the stability of the discrete problem to provide decay estimates of the error that depend on high-order Sobolev norms of the exact solution which are inaccessible to computation. A posteriori error estimates rely on the stability of the continuous problem and provide computable upper bounds on the error.

34.1 The residual and its dual norm

A key notion in a posteriori error analysis is the residual and its dual norm.

34.1.1 Model problem and residual

For simplicity, we focus on the purely diffusive version of the model problem (32.1) with homogeneous Dirichlet boundary conditions. We denote by u the unique function in $V := H_0^1(D)$ such that $a(u, w) = \ell(w)$ for all $w \in V$, where $a(v, w) := \int_D (\mathbb{d}\nabla v) \cdot \nabla w \, dx$ and $\ell(w) := \int_D fw \, dx$ with $f \in L^2(D)$. As in §32.1, we assume that $\mathbb{d}$ is defined on D with values in $\mathbb{R}^{d \times d}$ and that $\mathbb{d}(\boldsymbol{x})$ is symmetric with all its eigenvalues in the interval $[\lambda_\flat, \lambda_\sharp]$ for a.e. $\boldsymbol{x} \in D$, where $0 < \lambda_\flat \leq \lambda_\sharp < \infty$.

Let $(\mathcal{T}_h)_{h \in \mathcal{H}}$ be a shape-regular family of matching simplicial meshes of D, let $V_h := P_{k,0}^{\mathrm{g}}(\mathcal{T}_h)$ be the $H_0^1(D)$-conforming finite element space of some degree $k \geq 1$, and let $u_h \in V_h$ be the corresponding approximate solution such that $a(u_h, w_h) = \ell(w_h)$ for all $w_h \in V_h$ (see §32.1). Let $V' := \mathcal{L}(V; \mathbb{R})$ be the real space of bounded linear forms acting on V. In the present setting, we have $V' := \mathcal{L}(H_0^1(D); \mathbb{R}) = H^{-1}(D)$; see Definition 4.10. We denote the action of an element of V' on a function in V by using brackets. We equip the space V with the H^1-seminorm, i.e.,

$\|v\|_V := \|\nabla v\|_{\boldsymbol{L}^2(D)} = |v|_{H^1(D)}$. The Poincaré–Steklov inequality implies that this seminorm is indeed a norm on V; see Lemma 3.27 (with $p := 2$).

Definition 34.1 (Residual). *The residual of the discrete solution $u_h \in V_h$ is the element $\rho(u_h) \in V' := H^{-1}(D)$ acting as follows:*

$$\langle \rho(u_h), \varphi \rangle := \ell(\varphi) - a(u_h, \varphi), \qquad \forall \varphi \in V := H_0^1(D). \tag{34.1}$$

The boundedness of the bilinear form a together with the assumption $f \in L^2(D)$ implies that $\rho(u_h)$ is bounded on V. Using the embedding $L^2(D) \hookrightarrow H^{-1}(D)$, (34.1) is equivalent to $\rho(u_h) := f + \nabla\cdot(\mathrm{d}\nabla u_h) \in H^{-1}(D)$. Moreover, since u satisfies $a(u, \varphi) = \ell(\varphi)$ for all $\varphi \in H_0^1(D)$, we infer that

$$\langle \rho(u_h), \varphi \rangle := a(u - u_h, \varphi), \qquad \forall \varphi \in H_0^1(D), \tag{34.2}$$

which is equivalent to saying that $\rho(u_h) := \nabla\cdot(\mathrm{d}\nabla(u_h - u)) \in H^{-1}(D)$.

Remark 34.2 (Extensions). We refer the reader to Verfürth [378, §4.3-4.4] for other boundary conditions and lower-order terms in the PDE, to Verfürth [377] for the analysis of singularly perturbed regimes and a precise tracking of the model parameters in the error constants, to Ciarlet and Vohralík [122] for sign-changing diffusion coefficients, and to Cohen et al. [138] for a source term in $H^{-1}(D)$. □

34.1.2 The residual dual norm and the error

The dual space $V' := H^{-1}(D)$ is equipped with the norm

$$\|\eta\|_{H^{-1}(D)} := \sup_{\varphi \in H_0^1(D)} \frac{|\langle \eta, \varphi \rangle|}{\|\nabla\varphi\|_{\boldsymbol{L}^2(D)}}, \tag{34.3}$$

for all $\eta \in H^{-1}(D)$. Our first important observation is that the dual norm of the residual is closely related to the H^1-seminorm of the error.

Lemma 34.3 (Error and residual). *Let α and M denote, respectively, the stability and boundedness constants of the bilinear form a with respect to the H^1-seminorm. The following holds true:*

$$\frac{1}{M}\|\rho(u_h)\|_{H^{-1}(D)} \le \|\nabla(u - u_h)\|_{\boldsymbol{L}^2(D)} \le \frac{1}{\alpha}\|\rho(u_h)\|_{H^{-1}(D)}. \tag{34.4}$$

Proof. Owing to (34.2), we have $\|\rho(u_h)\|_{H^{-1}(D)} = \sup_{\varphi \in H_0^1(D)} \frac{|a(u-u_h,\varphi)|}{\|\nabla\varphi\|_{\boldsymbol{L}^2(D)}}$, and the stability and boundedness of a imply that

$$\alpha\|\nabla(u - u_h)\|_{\boldsymbol{L}^2(D)} \le \sup_{\varphi \in H_0^1(D)} \frac{|a(u - u_h, \varphi)|}{\|\nabla\varphi\|_{\boldsymbol{L}^2(D)}} \le M\|\nabla u - u_h\|_{\boldsymbol{L}^2(D)}. \qquad \square$$

Remark 34.4 (Stability). The coercivity of a is not needed to prove (34.4), it is just the inf-sup stability that is invoked (see the BNB theorem). □

Lemma 34.3 is fundamental since it provides two-sided bounds on the approximation error in terms of the residual. These two bounds are computable since they only depend on u_h and on the problem data (f and d). Considering the $H^{-1}(D)$-norm is a key ingredient in the argumentation since it allows us to get rid of u by replacing $a(u, \varphi)$ by $\ell(\varphi)$ for all $\varphi \in H_0^1(D)$. The price to pay for this replacement is that the $H^{-1}(D)$-norm is not computable since it invokes the supremum over an infinite-dimensional space. We will circumvent this difficulty in §34.2.

34.1.3 Localization of dual norms

An objection that is often raised with dual norms is that they are not localizable. While this is generally true for arbitrary elements of $H^{-1}(D)$, localization is actually possible when the linear form vanishes on a set of functions with local support forming a partition of unity. Let us first observe that we can define the restriction of $\eta \in H^{-1}(D)$ to an open Lipschitz subset $U \subset D$ to be the bounded linear form $\eta_{|U} \in H^{-1}(U)$ such that $\langle \eta_{|U}, \psi \rangle_U := \langle \eta, E_U(\psi) \rangle$ for all $\psi \in H_0^1(U)$, where $E_U(\psi) \in H_0^1(D)$ denotes the zero-extension of ψ to D and $\langle \cdot, \cdot \rangle_U$ the duality pairing between $H^{-1}(U)$ and $H_0^1(U)$. Note that $\|\eta_{|U}\|_{H^{-1}(U)} \leq \|\eta\|_{H^{-1}(D)}$. We abuse the notation by just writing $\eta \in H^{-1}(U)$ when the context is unambiguous.

Consider the vertices $\boldsymbol{z} \in \mathcal{V}_h$ of the mesh $\mathcal{T}_h$ and the global shape functions $\{\psi_{\boldsymbol{z}}\}_{\boldsymbol{z} \in \mathcal{V}_h}$ associated with the $\mathbb{P}_1$ Lagrange finite elements (called hat or Courant basis functions); see §19.2.1. For all $\boldsymbol{z} \in \mathcal{V}_h$, let $\mathcal{T}_{\boldsymbol{z}}$ be the collection of the mesh cells sharing $\boldsymbol{z}$ and let $D_{\boldsymbol{z}} := \operatorname{int}(\bigcup_{K \in \mathcal{T}_{\boldsymbol{z}}} K)$. The set $D_{\boldsymbol{z}}$ is called *finite element star* and its diameter is denoted by $h_{D_{\boldsymbol{z}}}$. Recall that $\psi_{\boldsymbol{z}}$ is supported $\overline{D_{\boldsymbol{z}}}$ (see the left panel of Figure 21.1). The hat basis functions form a partition of unity since $\sum_{\boldsymbol{z} \in \mathcal{V}_h} \psi_{\boldsymbol{z}} = 1$ in $\overline{D}$. We even have a local partition of unity since $\sum_{\boldsymbol{z} \in \mathcal{V}_K} \psi_{\boldsymbol{z}|K} = 1$ for all $K \in \mathcal{T}_h$, where $\mathcal{V}_K$ is the collection of the $(d+1)$ vertices of K. To handle homogeneous Dirichlet conditions, we define the subset $\mathcal{V}_h^\circ \subset \mathcal{V}_h$ composed of the interior vertices (i.e., not lying on ∂D) and the subset $\mathcal{V}_h^\partial := \mathcal{V}_h \backslash \mathcal{V}_h^\circ$ composed of the boundary vertices.

Definition 34.5 (Poincaré–Steklov constant $C_{\mathrm{PS}, \boldsymbol{z}}$). *For all $\boldsymbol{z} \in \mathcal{V}_h^\circ$, let $H_*^1(D_{\boldsymbol{z}}) := \{v \in H^1(D_{\boldsymbol{z}}) \mid \int_{D_{\boldsymbol{z}}} v \, \mathrm{d}x = 0\}$, and for all $\boldsymbol{z} \in \mathcal{V}_h^\partial$, let $H_*^1(D_{\boldsymbol{z}}) := \{v \in H^1(D_{\boldsymbol{z}}) \mid v_{|\partial D_{\boldsymbol{z}} \cap \partial D} = 0\}$. We define*

$$C_{\mathrm{PS}, \boldsymbol{z}} := h_{D_{\boldsymbol{z}}}^{-1} \sup_{v \in H_*^1(D_{\boldsymbol{z}})} \frac{\|v\|_{L^2(D_{\boldsymbol{z}})}}{\|\nabla v\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})}}. \tag{34.5}$$

Remark 34.6 ($C_{\mathrm{PS}, \boldsymbol{z}}$). The constant $C_{\mathrm{PS}, \boldsymbol{z}}$ in Definition 34.5 is uniformly bounded on shape-regular mesh sequences. For $\boldsymbol{z} \in \mathcal{V}_h^\circ$, one has $C_{\mathrm{PS}, \boldsymbol{z}} \leq \pi^{-1}$ if $D_{\boldsymbol{z}}$ is convex; see (12.13). Sharp estimates in the nonconvex case can be found in Eymard et al. [197], Repin [334], Veeser and Verfürth [372], Šebestová and Vejchodský [346]; see also Exercise 22.3. For $\boldsymbol{z} \in \mathcal{V}_h^\partial$, one has $C_{\mathrm{PS}, \boldsymbol{z}} \leq 1$ if there is a vector $\boldsymbol{t} \in \mathbb{R}^d$ such that for a.e. $\boldsymbol{x} \in D_{\boldsymbol{z}}$, the straight line drawn from $\boldsymbol{x}$ in the direction of $\boldsymbol{t}$ first hits $\partial D_{\boldsymbol{z}}$ at a point in ∂D; see Vohralík [381]. We refer the reader to [381, 334] for the general case. $\square$

Proposition 34.7 (Localization). *Let $\eta \in H^{-1}(D)$. (i) We have*

$$\sum_{\boldsymbol{z} \in \mathcal{V}_h} \|\eta\|_{H^{-1}(D_{\boldsymbol{z}})}^2 \leq (d+1)\|\eta\|_{H^{-1}(D)}^2. \tag{34.6}$$

(ii) If η does not have low-frequency components, i.e., if the following holds:

$$\langle \eta, \psi_{\boldsymbol{z}} \rangle = 0, \qquad \forall \boldsymbol{z} \in \mathcal{V}_h^\circ, \tag{34.7}$$

then letting $\check{C}_{\mathrm{PS}} := \max_{\boldsymbol{z} \in \mathcal{V}_h}(1 + h_{D_{\boldsymbol{z}}} \|\nabla \psi_{\boldsymbol{z}}\|_{\boldsymbol{L}^\infty(D_{\boldsymbol{z}})} C_{\mathrm{PS}, \boldsymbol{z}})$, we have

$$\|\eta\|_{H^{-1}(D)}^2 \leq (d+1)\check{C}_{\mathrm{PS}}^2 \sum_{\boldsymbol{z} \in \mathcal{V}_h} \|\eta\|_{H^{-1}(D_{\boldsymbol{z}})}^2. \tag{34.8}$$

Proof. For all $\boldsymbol{z} \in \mathcal{V}_h$, let $v_{\boldsymbol{z}} \in H_0^1(D_{\boldsymbol{z}})$ be the Riesz–Fréchet representative of $\eta_{|D_{\boldsymbol{z}}}$. Then $\|\eta\|_{H^{-1}(D_{\boldsymbol{z}})}^2 = \langle \eta, v_{\boldsymbol{z}} \rangle_{D_{\boldsymbol{z}}} = \langle \eta, E_{D_{\boldsymbol{z}}}(v_{\boldsymbol{z}}) \rangle = \|\nabla v_{\boldsymbol{z}}\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})}^2$. Let us set $v := \sum_{\boldsymbol{z} \in \mathcal{V}_h} E_{D_{\boldsymbol{z}}}(v_{\boldsymbol{z}})$. Since $v \in H_0^1(D)$, we infer that

$$\sum_{\boldsymbol{z} \in \mathcal{V}_h} \|\eta\|_{H^{-1}(D_{\boldsymbol{z}})}^2 = \sum_{\boldsymbol{z} \in \mathcal{V}_h} \langle \eta, E_{D_{\boldsymbol{z}}}(v_{\boldsymbol{z}}) \rangle = \langle \eta, v \rangle \leq \|\eta\|_{H^{-1}(D)} \|\nabla v\|_{\boldsymbol{L}^2(D)}.$$

Using the Cauchy–Schwarz inequality and rearranging the sums leads to

$$\|\nabla v\|_{\boldsymbol{L}^2(D)}^2 = \sum_{K\in\mathcal{T}_h} \left\|\sum_{\boldsymbol{z}\in\mathcal{V}_K} \nabla v_{\boldsymbol{z}}\right\|_{\boldsymbol{L}^2(K)}^2 \le (d+1) \sum_{K\in\mathcal{T}_h} \sum_{\boldsymbol{z}\in\mathcal{V}_K} \|\nabla v_{\boldsymbol{z}}\|_{\boldsymbol{L}^2(K)}^2$$
$$= (d+1) \sum_{\boldsymbol{z}\in\mathcal{V}_h} \|\nabla v_{\boldsymbol{z}}\|_{L^2(D_{\boldsymbol{z}})}^2 = (d+1) \sum_{\boldsymbol{z}\in\mathcal{V}_h} \|\eta\|_{H^{-1}(D_{\boldsymbol{z}})}^2.$$

Combining the above bounds yields (34.6). Let us prove (34.8), i.e., we assume now that $\langle\eta,\psi_{\boldsymbol{z}}\rangle = 0$ for all $\boldsymbol{z}\in\mathcal{V}_h^\circ$, i.e., that (34.7) holds true. Let $\varphi\in H_0^1(D)$. Since $\psi_{\boldsymbol{z}}\varphi\in H_0^1(D_{\boldsymbol{z}})$, the partition of unity implies that

$$\langle\eta,\varphi\rangle = \sum_{\boldsymbol{z}\in\mathcal{V}_h} \langle\eta,\psi_{\boldsymbol{z}}\varphi\rangle_{D_{\boldsymbol{z}}} = \sum_{\boldsymbol{z}\in\mathcal{V}_h^\circ} \langle\eta,\psi_{\boldsymbol{z}}(\varphi-\underline{\varphi}^{\boldsymbol{z}})\rangle_{D_{\boldsymbol{z}}} + \sum_{\boldsymbol{z}\in\mathcal{V}_h^\partial} \langle\eta,\psi_{\boldsymbol{z}}\varphi\rangle_{D_{\boldsymbol{z}}}, \qquad (34.9)$$

with $\underline{\varphi}^{\boldsymbol{z}} := \frac{1}{|D_{\boldsymbol{z}}|}\int_{D_{\boldsymbol{z}}} \varphi\,\mathrm{d}x$, since $\langle\eta,\psi_{\boldsymbol{z}}\underline{\varphi}^{\boldsymbol{z}}\rangle = \underline{\varphi}^{\boldsymbol{z}}\langle\eta,\psi_{\boldsymbol{z}}\rangle = 0$ for all $\boldsymbol{z}\in\mathcal{V}_h^\circ$ owing to (34.7). We have

$$\|\nabla(\psi_{\boldsymbol{z}}(\varphi-\underline{\varphi}^{\boldsymbol{z}}))\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})} \le (1+h_{D_{\boldsymbol{z}}}\|\nabla\psi_{\boldsymbol{z}}\|_{\boldsymbol{L}^\infty(D_{\boldsymbol{z}})} C_{\mathrm{PS},\boldsymbol{z}})\|\nabla\varphi\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})}$$
$$\le \check{C}_{\mathrm{PS}}\|\nabla\varphi\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})},$$

where we used that $\nabla(\psi_{\boldsymbol{z}}(\varphi-\underline{\varphi}^{\boldsymbol{z}})) = \psi_{\boldsymbol{z}}\nabla\varphi + (\varphi-\underline{\varphi}^{\boldsymbol{z}})\nabla\psi_{\boldsymbol{z}}$, the triangle inequality, $\|\psi_{\boldsymbol{z}}\|_{\boldsymbol{L}^\infty(D_{\boldsymbol{z}})} = 1$, and the definitions of $C_{\mathrm{PS},\boldsymbol{z}}$ and $\check{C}_{\mathrm{PS}}$. Proceeding similarly, we infer the same bound on $\|\nabla(\psi_{\boldsymbol{z}}\varphi)\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})}$ for all $\boldsymbol{z}\in\mathcal{V}_h^\partial$. Using the Cauchy–Schwarz inequality, we infer that

$$|\langle\eta,\varphi\rangle| \le \left(\sum_{\boldsymbol{z}\in\mathcal{V}_h} \|\eta\|_{H^{-1}(D_{\boldsymbol{z}})}^2\right)^{\frac{1}{2}} \check{C}_{\mathrm{PS}} \left(\sum_{\boldsymbol{z}\in\mathcal{V}_h} \|\nabla\varphi\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})}^2\right)^{\frac{1}{2}}.$$

This gives (34.8) since $\sum_{\boldsymbol{z}\in\mathcal{V}_h} \|\nabla\varphi\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})}^2 = (d+1)\|\nabla\varphi\|_{\boldsymbol{L}^2(D)}^2$. □

The bound (34.8) means that we can consider local test functions in $\{H_0^1(D_{\boldsymbol{z}})\}_{\boldsymbol{z}\in\mathcal{V}_h}$ to explore the action on the whole space $H_0^1(D)$ of a linear form $\eta\in H^{-1}(D)$ satisfying (34.7). Note that the residual $\rho(u_h)$ satisfies (34.7): this is the Galerkin orthogonality property for the hat basis functions.

Remark 34.8 (Value of $\check{C}_{\mathrm{PS}}$). Using an inverse inequality to bound $\|\nabla\psi_{\boldsymbol{z}}\|_{L^\infty(D_{\boldsymbol{z}})}$, we can see that the constant $\check{C}_{\mathrm{PS}}$ from Lemma 34.7 is uniformly bounded on shape-regular mesh sequences. See also Remark 34.6. □

Remark 34.9 (Literature). The proof of Proposition 34.7 is inspired by Carstensen and Funken [108]; see also Babuška and Miller [36] (for the idea of working on finite element stars), Cohen et al. [138], Ciarlet and Vohralík [122], Blechta et al. [59]. □

34.2 Global upper bound

We derive a computable upper bound on the error by using Lemma 34.3. Inspired by Nochetto and Veeser [315], Veeser and Verfürth [372], we achieve this by relying on two key ideas: (i) the residual is the sum of an L^2-function and a measure supported in the mesh interfaces; (ii) localization is

achieved by exploiting that the residual vanishes on the hat basis functions in the same spirit as Proposition 34.7. The error upper bound derived herein belongs to the class of residual-based a posteriori estimates pioneered by Babuška and Rheinbolt [39]. Another class of a posteriori error estimates based on local flux equilibration in finite element stars is discussed in Chapter 52.

Let us first observe that the residual $\rho(u_h) \in H^{-1}(D)$ admits the following representation (see Exercise 34.1): For all $\varphi \in H_0^1(D)$,

$$\langle \rho(u_h), \varphi \rangle = \sum_{K \in \mathcal{T}_h} \int_K r^{\mathrm{v}}(u_h) \varphi \, \mathrm{d}x + \sum_{F \in \mathcal{F}_h^\circ} \int_F r^{\mathrm{s}}(u_h) \varphi \, \mathrm{d}s, \qquad (34.10)$$

with densities $r^{\mathrm{v}}(u_h) \in L^2(D)$ and $r^{\mathrm{s}}(u_h) \in L^\infty(\mathcal{F}_h^\circ)$ defined by

$$r^{\mathrm{v}}(u_h)_{|K} := f_{|K} + (\nabla{\cdot}(\mathrm{d}\nabla u_h))_{|K}, \quad \forall K \in \mathcal{T}_h, \qquad (34.11a)$$
$$r^{\mathrm{s}}(u_h)_{|F} := [\![\mathrm{d}\nabla u_h]\!]_F{\cdot}\boldsymbol{n}_F, \qquad \forall F \in \mathcal{F}_h^\circ, \qquad (34.11b)$$

and $[\![\cdot]\!]_F$ denotes the jump across F using the orientation of the unit normal $\boldsymbol{n}_F$ (see Definitions 8.10 and 18.2).

Definition 34.10 (Trace inequality constant). *Let $\boldsymbol{z} \in \mathcal{V}_h$, let $\mathcal{F}_{\boldsymbol{z}}^\circ$ be the collection of the interfaces sharing $\boldsymbol{z}$, and let $H_*^1(D_{\boldsymbol{z}})$ be defined as in Definition 34.5. Then we set*

$$C_{\mathrm{tr},\boldsymbol{z}} := h_{D_{\boldsymbol{z}}}^{-\frac{1}{2}} \sup_{v \in H_*^1(D_{\boldsymbol{z}})} \frac{\|v\|_{L^2(\mathcal{F}_{\boldsymbol{z}}^\circ)}}{\|\nabla v\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})}}, \qquad (34.12)$$

with the notation $\|v\|_{L^2(\mathcal{F}_{\boldsymbol{z}}^\circ)} := \left(\sum_{F \in \mathcal{F}_{\boldsymbol{z}}^\circ} \|v\|_{L^2(F)}^2\right)^{\frac{1}{2}}$.

Remark 34.11 ($C_{\mathrm{tr},\boldsymbol{z}}$). The constant $C_{\mathrm{tr},\boldsymbol{z}}$ in Definition 34.10 is uniformly bounded on shape-regular mesh sequences; see Exercise 34.2. □

Theorem 34.12 (Upper bound). *Define the vertex-based error indicators*

$$\eta_{\boldsymbol{z}}^{\mathrm{v}}(u_h) := h_{D_{\boldsymbol{z}}} \|\psi_{\boldsymbol{z}}^{\frac{1}{2}} r^{\mathrm{v}}(u_h)\|_{L^2(D_{\boldsymbol{z}})}, \quad \eta_{\boldsymbol{z}}^{\mathrm{s}}(u_h) := h_{D_{\boldsymbol{z}}}^{\frac{1}{2}} \|\psi_{\boldsymbol{z}}^{\frac{1}{2}} r^{\mathrm{s}}(u_h)\|_{L^2(\mathcal{F}_{\boldsymbol{z}}^\circ)},$$
$$\eta_{\boldsymbol{z}}(u_h) := (d+1)^{\frac{1}{2}} \big(C_{\mathrm{PS},\boldsymbol{z}} \eta_{\boldsymbol{z}}^{\mathrm{v}}(u_h) + C_{\mathrm{tr},\boldsymbol{z}} \eta_{\boldsymbol{z}}^{\mathrm{s}}(u_h)\big), \qquad (34.13)$$

with $C_{\mathrm{PS},\boldsymbol{z}}$ defined in (34.5) and $C_{\mathrm{tr},\boldsymbol{z}}$ defined in (34.12). The following global a posteriori estimate holds true:

$$\alpha \|\nabla(u - u_h)\|_{\boldsymbol{L}^2(D)} \leq \left(\sum_{\boldsymbol{z} \in \mathcal{V}_h} \eta_{\boldsymbol{z}}(u_h)^2\right)^{\frac{1}{2}}. \qquad (34.14)$$

Proof. Our starting point is the error upper bound from Lemma 34.3, i.e., $\alpha \|\nabla(u - u_h)\|_{\boldsymbol{L}^2(D)} \leq \sup_{\varphi \in H_0^1(D)} \frac{|\langle \rho(u_h), \varphi \rangle|}{\|\nabla \varphi\|_{\boldsymbol{L}^2(D)}}$. Using (34.9), we infer that $\langle \rho(u_h), \varphi \rangle = \sum_{\boldsymbol{z} \in \mathcal{V}_h^\circ} \langle \rho(u_h), \psi_{\boldsymbol{z}}(\varphi - \underline{\varphi}^{\boldsymbol{z}}) \rangle_{D_{\boldsymbol{z}}} + \sum_{\boldsymbol{z} \in \mathcal{V}_h^\partial} \langle \rho(u_h), \psi_{\boldsymbol{z}} \varphi \rangle_{D_{\boldsymbol{z}}}$, where $\underline{\varphi}^{\boldsymbol{z}}$ is the mean value of φ over $D_{\boldsymbol{z}}$. Consider $\boldsymbol{z} \in \mathcal{V}_h^\circ$. Exploiting the representation (34.10) and since $\psi_{\boldsymbol{z}}$ is supported in $D_{\boldsymbol{z}}$, we infer that

$$\langle \rho(u_h), \psi_{\boldsymbol{z}}(\varphi - \underline{\varphi}^{\boldsymbol{z}}) \rangle_{D_{\boldsymbol{z}}} = \sum_{K \in \mathcal{T}_{\boldsymbol{z}}} \int_K r^{\mathrm{v}}(u_h) \psi_{\boldsymbol{z}}(\varphi - \underline{\varphi}^{\boldsymbol{z}}) \, \mathrm{d}x + \sum_{F \in \mathcal{F}_{\boldsymbol{z}}^\circ} \int_F r^{\mathrm{s}}(u_h) \psi_{\boldsymbol{z}}(\varphi - \underline{\varphi}^{\boldsymbol{z}}) \, \mathrm{d}s.$$

Let $\mathfrak{T}_1, \mathfrak{T}_2$ denote the two terms on the right-hand side. Using the Cauchy–Schwarz inequality, $\|\psi_{\boldsymbol{z}}\|_{L^\infty(D_{\boldsymbol{z}})} = 1$, Definition 34.5, and $\nabla \underline{\varphi}^{\boldsymbol{z}} = \boldsymbol{0}$ yields

$$|\mathfrak{T}_1| \leq \|\psi_{\boldsymbol{z}} r^{\mathrm{v}}(u_h)\|_{L^2(D_{\boldsymbol{z}})} \|\varphi - \underline{\varphi}^{\boldsymbol{z}}\|_{L^2(D_{\boldsymbol{z}})} \leq C_{\mathrm{PS},\boldsymbol{z}} h_{D_{\boldsymbol{z}}} \|\psi_{\boldsymbol{z}}^{\frac{1}{2}} r^{\mathrm{v}}(u_h)\|_{L^2(D_{\boldsymbol{z}})} \|\nabla \varphi\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})}.$$

Proceeding similarly and invoking Definition 34.10 leads to

$$|\mathfrak{T}_2| \leq C_{\mathrm{tr},\boldsymbol{z}} h_{D_{\boldsymbol{z}}}^{\frac{1}{2}} \|\psi_{\boldsymbol{z}}^{\frac{1}{2}} r^{\mathrm{s}}(u_h)\|_{L^2(\mathcal{F}_{\boldsymbol{z}}^\circ)} \|\nabla\varphi\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})}.$$

Similar bounds are verified for $\boldsymbol{z} \in \mathcal{V}_h^\partial$. We conclude the proof by using the Cauchy–Schwarz inequality and $\sum_{\boldsymbol{z}\in\mathcal{V}_h} \|\nabla\varphi\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})}^2 = (d+1)\|\nabla\varphi\|_{\boldsymbol{L}^2(D)}^2$. $\qquad\square$

Remark 34.13 (Variants). The factor $(d+1)$ in $\eta_{\boldsymbol{z}}(u_h)$ can be avoided by invoking Poincaré–Steklov inequalities based on norms weighted by the hat basis functions (see Veeser and Verfürth [372]). The weights $\psi_{\boldsymbol{z}}^{\frac{1}{2}}$ in $\eta_{\boldsymbol{z}}^{\mathrm{v}}(u_h)$ and $\eta_{\boldsymbol{z}}^{\mathrm{s}}(u_h)$ are not essential, but they will help to deduce Corollary 34.14. Another route to derive an upper bound similar to (34.14) consists of combining the upper bound from Lemma 34.3 with the localization property (34.8) and the residual representation (34.10). The above proof of Theorem 34.12 is slightly more direct and therefore leads to somewhat sharper values for the constants weighting the error indicators. $\qquad\square$

Since adaptive procedures usually mark cells rather than vertices (see Morin et al. [306] for an example of vertex-based marking), we reformulate the global upper bound (34.14) in terms of cell-based error indicators.

Corollary 34.14 (Cell-based upper bound). *The following holds true:*

$$\alpha\|\nabla(u-u_h)\|_{\boldsymbol{L}^2(D)} \leq C_{\mathrm{GUB}} \bigg(\sum_{K\in\mathcal{T}_h} \Big(\eta_K^{\mathrm{v}}(u_h)^2 + \eta_K^{\mathrm{s}}(u_h)^2 \Big) \bigg)^{\frac{1}{2}}, \tag{34.15}$$

with the constant $C_{\mathrm{GUB}} := (d+1)^{\frac{1}{2}} \max_{\boldsymbol{z}\in\mathcal{V}_h}(2^{\frac{1}{2}} C_{\mathrm{PS},\boldsymbol{z}}\vartheta_{\boldsymbol{z}}, C_{\mathrm{tr},\boldsymbol{z}}\varrho_{\boldsymbol{z}}^{\frac{1}{2}})$, *the geometric factors* $\vartheta_{\boldsymbol{z}} := \max_{K\in\mathcal{T}_{\boldsymbol{z}}} \frac{h_{D_{\boldsymbol{z}}}}{h_K}$ *and* $\varrho_{\boldsymbol{z}} := \max_{F\in\mathcal{F}_{\boldsymbol{z}}^\circ} \frac{h_{D_{\boldsymbol{z}}}}{h_F}$, *and the cell-based error indicators*

$$\eta_K^{\mathrm{v}}(u_h) := h_K \|r^{\mathrm{v}}(u_h)\|_{L^2(K)}, \qquad \eta_K^{\mathrm{s}}(u_h) := h_K^{\frac{1}{2}} \|r^{\mathrm{s}}(u_h)\|_{L^2(\mathcal{F}_K^\circ)}, \tag{34.16}$$

where $\|v\|_{L^2(\mathcal{F}_K^\circ)} := (\sum_{F\in\mathcal{F}_K^\circ} \|v\|_{L^2(F)}^2)^{\frac{1}{2}}$ *and* $\mathcal{F}_K^\circ$ *is the collection of the faces of K that are interfaces.*

Proof. Since $(a+b)^2 \leq 2(a^2+b^2)$ for all $a,b \in \mathbb{R}$, recalling (34.13) we have

$$\sum_{\boldsymbol{z}\in\mathcal{V}_h} \eta_{\boldsymbol{z}}(u_h)^2 \leq \sum_{\boldsymbol{z}\in\mathcal{V}_h} 2(d+1) C_{\mathrm{PS},\boldsymbol{z}}^2 h_{D_{\boldsymbol{z}}}^2 \|\psi_{\boldsymbol{z}}^{\frac{1}{2}} r^{\mathrm{v}}(u_h)\|_{L^2(D_{\boldsymbol{z}})}^2$$
$$+ \sum_{\boldsymbol{z}\in\mathcal{V}_h} 2(d+1) C_{\mathrm{tr},\boldsymbol{z}}^2 h_{D_{\boldsymbol{z}}} \|\psi_{\boldsymbol{z}}^{\frac{1}{2}} r^{\mathrm{s}}(u_h)\|_{L^2(\mathcal{F}_{\boldsymbol{z}}^\circ)}^2.$$

Let $\mathfrak{T}_1, \mathfrak{T}_2$ denote the two terms on the right-hand side. We have

$$\mathfrak{T}_1 \leq 2(d+1) \max_{\boldsymbol{z}\in\mathcal{V}_h}(C_{\mathrm{PS},\boldsymbol{z}}^2 \vartheta_{\boldsymbol{z}}^2) \sum_{\boldsymbol{z}\in\mathcal{V}_h} \sum_{K\in\mathcal{T}_{\boldsymbol{z}}} h_K^2 \|\psi_{\boldsymbol{z}}^{\frac{1}{2}} r^{\mathrm{v}}(u_h)\|_{L^2(K)}^2$$
$$= 2(d+1) \max_{\boldsymbol{z}\in\mathcal{V}_h}(C_{\mathrm{PS},\boldsymbol{z}}^2 \vartheta_{\boldsymbol{z}}^2) \sum_{K\in\mathcal{T}_h} h_K^2 \|r^{\mathrm{v}}(u_h)\|_{L^2(K)}^2,$$

where we used $\vartheta_{\boldsymbol{z}} := \max_{K\in\mathcal{T}_z} \frac{h_{D_{\boldsymbol{z}}}}{h_K}$ and $D_{\boldsymbol{z}} := \mathrm{int}(\bigcup_{K\in\mathcal{T}_{\boldsymbol{z}}} K)$ in the first line and where the identity in the second line follows by exchanging the two summations and using that the restrictions to any mesh cell of the hat basis functions form a partition of unity. The reasoning for $\mathfrak{T}_2$ is similar

since the restrictions to any interface of the hat basis functions also form a partition of unity. This leads to

$$\mathfrak{T}_2 \leq (d+1) \max_{\boldsymbol{z} \in \mathcal{V}_h}(C_{\mathrm{tr},\boldsymbol{z}}^2 \varrho_{\boldsymbol{z}}) \sum_{F \in \mathcal{F}_h^\circ} 2h_F \|r^{\mathrm{s}}(u_h)\|_{L^2(F)}^2.$$

Finally, we remove the factor 2 by introducing a cell-based summation and observing that every interface $F := \partial K_l \cap \partial K_r \in \mathcal{F}_h^\circ$ is shared by two mesh cells and that $h_F \leq h_K$ for all $K \in \{K_l, K_r\}$. □

Remark 34.15 (C_{GUB}). The constant C_{GUB} in Corollary 34.14 is uniformly bounded on shape-regular mesh sequences. □

Remark 34.16 (Dual-weighted residual estimate). Let $\psi \in H^{-1}(D)$ be some linear output functional. Let $z_\psi \in H_0^1(D)$ solve the dual problem $a(\varphi, z_\psi) := \langle \psi, \varphi \rangle$ for all $\varphi \in H_0^1(D)$. Then $\langle \psi, u - u_h \rangle = a(u - u_h, z_\psi) = \langle \rho(u_h), z_\psi \rangle$, i.e., the output error $\langle \psi, u - u_h \rangle$ is equal to the residual tested against the dual solution z_ψ. For instance, we have $\|u - u_h\|_{L^2(D)}^2 = \langle \rho(u_h), z_\psi \rangle$ with $z_\psi \in H_0^1(D)$ s.t. $a(\varphi, z_\psi) := (u - u_h, \varphi)_{L^2(D)}$ for all $\varphi \in H_0^1(D)$. See Becker and Rannacher [48] for further insight on the whole approach. Notice though that one must approximate z_ψ to obtain a computable estimate. □

34.3 Local lower bound

Our goal in this section is to bound the cell-based error indicators $\eta_K^{\mathrm{v}}(u_h)$ and $\eta_K^{\mathrm{s}}(u_h)$ defined in (34.16) by the approximation error in the mesh cell K (and some of its neighbors) for all $K \in \mathcal{T}_h$. This will give lower bounds on the approximation error. These lower bounds differ from the upper bounds on two aspects: they are local (recall that the upper bounds are global), and they involve generic constants whose value may depend on the regularity of the mesh sequence and the polynomial degree k. The symbol c denotes these generic constants (the value can change at each occurrence).

To put the upcoming results into perspective, we observe that for every subset U of D, defining $M_U := \|d\|_{L^\infty(U;\mathbb{R}^{d \times d})}$, the bound

$$\|\rho(u_h)\|_{H^{-1}(U)} \leq M_U \|\nabla(u - u_h)\|_{\boldsymbol{L}^2(U)} \tag{34.17}$$

follows from $\langle \rho(u_h), \varphi \rangle_U = \int_U (d\nabla(u - u_h)) \cdot \nabla\varphi \, dx$ for all $\varphi \in H_0^1(U)$ and the Cauchy–Schwarz inequality. After observing that $r^{\mathrm{v}}(u_h)_{|K} = \rho(u_h)_{|K}$ (where we used the (slightly abusive) notation on the restriction to $H^{-1}(K)$ of the residual), what we now need is a bound of the form

$$\|r^{\mathrm{v}}(u_h)\|_{L^2(K)} \leq c\, h_K^{-1} \|r^{\mathrm{v}}(u_h)\|_{H^{-1}(K)}, \qquad \forall K \in \mathcal{T}_h. \tag{34.18}$$

If (34.18) were indeed true, we would immediately deduce that

$$\eta_K^{\mathrm{v}}(u_h) := h_K \|r^{\mathrm{v}}(u_h)\|_{L^2(K)} \leq c\, \|r^{\mathrm{v}}(u_h)\|_{H^{-1}(K)} \leq c\, M_K \|\nabla(u - u_h)\|_{\boldsymbol{L}^2(K)}.$$

Unfortunately, (34.18) is an inverse-like inequality (where one norm is that of a dual space). Hence, it cannot be valid for every function in $L^2(K)$ (recall that $r^{\mathrm{v}}(u_h)$ depends on the source term f and the tensor d). We refer the reader to [315, Pbm. 32] for a concrete example.

To address this problem, we introduce the notion of oscillation. For all $K \in \mathcal{T}_h$ and every integer $l^{\mathrm{v}} \geq 0$, we use the notation $P_{l^{\mathrm{v}}}(K) := \mathbb{P}_{l^{\mathrm{v}},d} \circ \boldsymbol{T}_K^{-1}$, where $\boldsymbol{T}_K : \widehat{K} \to K$ is the geometric mapping.

Similarly, for all $F \in \mathcal{F}_h^\circ$ and every integer $l^{\mathrm{s}} \geq 0$, we use the notation $P_{l^{\mathrm{s}}}(F) := \mathbb{P}_{l^{\mathrm{s}},d-1} \circ \boldsymbol{T}_{K,F}^{-1}$, where K is a cell having F as face, $\boldsymbol{T}_{K,F} := \boldsymbol{T}_{K|\widehat{F}} \circ \boldsymbol{T}_{\widehat{F}} : \widehat{F}^{d-1} \to F$, $\widehat{F} := \boldsymbol{T}_K^{-1}(F)$, $\widehat{F}^{d-1}$ is the reference simplex in $\mathbb{R}^{d-1}$, and $\boldsymbol{T}_{\widehat{F}} : \widehat{F}^{d-1} \to \widehat{F}$ is an affine bijective mapping (recall that $P_{l^{\mathrm{s}}}(F)$ is independent of the choice of K); see §20.2.

Definition 34.17 (Oscillation). *Let $l^{\mathrm{v}}, l^{\mathrm{s}} \in \mathbb{N}$, $K \in \mathcal{T}_h$, and $F \in \mathcal{F}_h^\circ$. Let $\bar{r}^{\mathrm{v}}(u_h)$ be the L^2-orthogonal projection of $r^{\mathrm{v}}(u_h)$ onto $P_{l^{\mathrm{v}}}(K)$. Let $\bar{r}^{\mathrm{s}}(u_h)$ be the L^2-orthogonal projection of $r^{\mathrm{s}}(u_h)$ onto $P_{l^{\mathrm{s}}}(F)$. The oscillation indicators are defined by*

$$\phi_K^{\mathrm{v}}(u_h, f, \mathrm{d}) := h_K \|r^{\mathrm{v}}(u_h) - \bar{r}^{\mathrm{v}}(u_h)\|_{L^2(K)}, \tag{34.19a}$$

$$\phi_F^{\mathrm{s}}(u_h, f, \mathrm{d}) := h_F^{\frac{1}{2}} \|r^{\mathrm{s}}(u_h) - \bar{r}^{\mathrm{s}}(u_h)\|_{L^2(F)}. \tag{34.19b}$$

Lemma 34.18 (Verfürth's inverse inequalities). (i) *Let $T_K^{\mathrm{v}} : L^2(K) \to H^{-1}(K)$ be defined by $\langle T_K^{\mathrm{v}}(r), \varphi \rangle_K := \int_K r\varphi \, \mathrm{d}x$ for all $\varphi \in H_0^1(K)$ and all $r \in L^2(K)$. There is c, depending on l^{v}, such that for all $K \in \mathcal{T}_h$, all $h \in \mathcal{H}$, and all $q \in P_{l^{\mathrm{v}}}(K)$,*

$$\|q\|_{L^2(K)} \leq c \, h_K^{-1} \|T_K^{\mathrm{v}}(q)\|_{H^{-1}(K)}. \tag{34.20}$$

(ii) *Let $T_F^{\mathrm{s}} : L^2(F) \to H^{-1}(D_F)$ be defined by $\langle T_F^{\mathrm{s}}(r), \varphi \rangle_{D_F} := \int_F r\varphi \, \mathrm{d}s$ for all $\varphi \in H_0^1(D_F)$ and all $r \in L^2(F)$, where $D_F := \mathrm{int}(K_l \cup K_r)$ with $F := \partial K_l \cap \partial K_r$. There is c, depending on l^{s}, such that for all $F \in \mathcal{F}_h^\circ$, all $h \in \mathcal{H}$, and all $g \in P_{l^{\mathrm{s}}}(F)$,*

$$\|g\|_{L^2(F)} \leq c \, h_F^{-\frac{1}{2}} \|T_F^{\mathrm{s}}(g)\|_{H^{-1}(D_F)}. \tag{34.21}$$

Proof. The proof hinges on the use of suitable cell- and face-based bubble functions introduced by Verfürth [378, §3.6]. These functions vanish on the boundary of K and D_F, respectively.
(1) Proof of (34.20). Let $K \in \mathcal{T}_h$. The cell-based bubble function $b_K^{\mathrm{v}} := (d+1)^{d+1} \lambda_0^K \dots \lambda_d^K$, where $\{\lambda_i^K\}_{i \in \{0:d\}}$ are the barycentric coordinates in K, is such that $\|q\|_{L^2(K)}^2 \leq c_1 \|(b_K^{\mathrm{v}})^{\frac{1}{2}} q\|_{L^2(K)}^2$ and $\|\nabla(b_K^{\mathrm{v}} q)\|_{\boldsymbol{L}^2(K)} \leq c_2 h_K^{-1} \|q\|_{L^2(K)}$ for all $q \in P_{l^{\mathrm{v}}}(K)$ (both inequalities are established on the reference element by invoking norm equivalence in polynomial spaces and then transferred back to K by the geometric mapping $\boldsymbol{T}_K$). Noticing that $b_K^{\mathrm{v}} q \in H_0^1(K)$, we infer that

$$
\begin{aligned}
\|q\|_{L^2(K)}^2 &\leq c_1 \|(b_K^{\mathrm{v}})^{\frac{1}{2}} q\|_{L^2(K)}^2 = c_1 \langle T_K^{\mathrm{v}}(q), b_K^{\mathrm{v}} q \rangle_K \\
&\leq c_1 \|T_K^{\mathrm{v}}(q)\|_{H^{-1}(K)} \|\nabla(b_K^{\mathrm{v}} q)\|_{\boldsymbol{L}^2(K)} \\
&\leq c_1 c_2 h_K^{-1} \|T_K^{\mathrm{v}}(q)\|_{H^{-1}(K)} \|q\|_{L^2(K)}.
\end{aligned}
$$

(2) Proof of (34.21). Let $F := \partial K_l \cap \partial K_r \in \mathcal{F}_h^\circ$, and assume without loss of generality that in both cells, the vertex opposite to F is associated with the barycentric coordinate λ_0^K with $K \in \mathcal{T}_F := \{K_l, K_r\}$. The face-based bubble function b_F^{s} such that $b_{F|K}^{\mathrm{s}} := d^d \lambda_1^K \dots \lambda_d^K$, for all $K \in \mathcal{T}_F$, is in $H_0^1(D_F)$ and is such that $\|g\|_{L^2(F)}^2 \leq c_3 \|(b_F^{\mathrm{s}})^{\frac{1}{2}} \widetilde{g}\|_{L^2(F)}^2$ and $\|\nabla(b_F^{\mathrm{s}} \widetilde{g})\|_{\boldsymbol{L}^2(D_F)} \leq c_4 h_F^{-\frac{1}{2}} \|g\|_{L^2(F)}$ for all $g \in P_{l^{\mathrm{s}}}(F)$, where $\widetilde{g}$ is the extension of g to D_F defined by $\widetilde{g}_{|K} := (((g \circ (\boldsymbol{T}_K)_{|\widehat{F}}) \circ \Pi_{\widehat{F}}) \circ \boldsymbol{T}_K^{-1})$ for all $K \in \mathcal{T}_F$, where $\widehat{F} := \boldsymbol{T}_K^{-1}(F)$ and $\Pi_{\widehat{F}}$ is the orthogonal projection onto $\widehat{F}$. Note that $\widetilde{g}_{|F} = g$. The above bounds are again proved on the reference element by using the pullback by $\boldsymbol{T}_K$. Since $b_F^{\mathrm{s}} \widetilde{g} \in H_0^1(D_F)$, we conclude that

$$
\begin{aligned}
\|g\|_{L^2(F)}^2 &\leq c_3 \|(b_F^{\mathrm{s}})^{\frac{1}{2}} g\|_{L^2(F)}^2 = c_3 \langle T_F^{\mathrm{s}}(g), b_F^{\mathrm{s}} \widetilde{g} \rangle_{D_F} \\
&\leq c_3 \|T_F^{\mathrm{s}}(g)\|_{H^{-1}(D_F)} \|\nabla(b_F^{\mathrm{s}} \widetilde{g})\|_{\boldsymbol{L}^2(D_F)} \\
&\leq c_3 c_4 \tilde{h}_F^{-\frac{1}{2}} \|T_F^{\mathrm{s}}(g)\|_{H^{-1}(D_F)} \|g\|_{L^2(F)}. \qquad \square
\end{aligned}
$$

The operator $T_K^{\mathrm{v}} : L^2(K) \to H^{-1}(K)$ is nothing but the natural injection of $L^2(K)$ into $H^{-1}(K)$. Observe also that $T_K^{\mathrm{v}}(r^{\mathrm{v}}(u_h)) = \rho(u_h)_{|K}$ in $H^{-1}(K)$. We now establish a local lower bound on the error using the cell-based indicators $\eta_K^{\mathrm{v}}(u_h)$ and $\eta_K^{\mathrm{s}}(u_h)$ defined in (34.16).

Theorem 34.19 (Local lower bound). *For all $K \in \mathcal{T}_h$, let $\mathcal{T}_K^{\mathrm{f}}$ be the set composed of K and those cells sharing an interface with K, and let $D_K^{\mathrm{f}} := \mathrm{int}(\bigcup_{K' \in \mathcal{T}_K^{\mathrm{f}}} K')$. There is c such that for all $K \in \mathcal{T}_h$ and all $h \in \mathcal{H}$,*

$$\eta_K^{\mathrm{v}}(u_h) + \eta_K^{\mathrm{s}}(u_h) \le c \left(M_{D_K^{\mathrm{f}}} \|\nabla(u - u_h)\|_{\boldsymbol{L}^2(D_K^{\mathrm{f}})} + \phi_{\mathcal{T}_K^{\mathrm{f}}}(u_h, f, \mathrm{d}) \right), \tag{34.22}$$

with $\phi_{\mathcal{T}_K^{\mathrm{f}}}(u_h, f, \mathrm{d}) := \sum_{K' \in \mathcal{T}_K^{\mathrm{f}}} \phi_{K'}^{\mathrm{v}}(u_h, f, \mathrm{d}) + \sum_{F \in \mathcal{F}_K^\circ} \phi_F^{\mathrm{s}}(u_h, f, \mathrm{d})$, where $\phi_{K'}^{\mathrm{v}}$ and ϕ_F^{s} are defined in (34.19), and $M_{D_K^{\mathrm{f}}} := \|\mathrm{d}\|_{L^\infty(D_K^{\mathrm{f}}; \mathbb{R}^{d \times d})}$.

Proof. Let $K \in \mathcal{T}_h$. Owing to (34.20) and the triangle inequality, we infer that

$$
\begin{aligned}
h_K \|\bar{r}^{\mathrm{v}}(u_h)\|_{L^2(K)} &\le c \, \|T_K^{\mathrm{v}}(\bar{r}^{\mathrm{v}}(u_h))\|_{H^{-1}(K)} \\
&\le c \left(\|T_K^{\mathrm{v}}(r^{\mathrm{v}}(u_h))\|_{H^{-1}(K)} + \|T_K^{\mathrm{v}}(r^{\mathrm{v}}(u_h) - \bar{r}^{\mathrm{v}}(u_h))\|_{H^{-1}(K)} \right) \\
&\le c' \left(\|\rho(u_h)\|_{H^{-1}(K)} + \phi_K^{\mathrm{v}}(u_h, f, \mathrm{d}) \right),
\end{aligned}
$$

since $T_K^{\mathrm{v}}(r^{\mathrm{v}}(u_h)) := \rho(u_h)_{|K}$ and $\|T_K^{\mathrm{v}}(r)\|_{H^{-1}(K)} \le c h_K \|r\|_{L^2(K)}$ for all $r \in L^2(K)$ (see Exercise 34.3). Using the triangle inequality and $\|\rho(u_h)\|_{H^{-1}(K)} \le M_K \|\nabla(u - u_h)\|_{\boldsymbol{L}^2(K)}$ owing to (34.17), we infer that

$$\eta_K^{\mathrm{v}}(u_h) := h_K \|r^{\mathrm{v}}(u_h)\|_{L^2(K)} \le c \left(M_K \|\nabla(u - u_h)\|_{\boldsymbol{L}^2(K)} + \phi_K^{\mathrm{v}}(u_h, f, \mathrm{d}) \right).$$

Let $F \in \mathcal{F}_K^\circ$. By using (34.21) and $\|T_F^{\mathrm{s}}(g)\|_{H^{-1}(D_F)} \le c \tilde{h}_F^{\frac{1}{2}} \|g\|_{L^2(F)}$ for all $g \in L^2(F)$ (see Exercise 34.3) and proceeding similarly, we infer that

$$h_F^{\frac{1}{2}} \|r^{\mathrm{s}}(u_h)\|_{L^2(F)} \le c \left(\|T_F^{\mathrm{s}}(r^{\mathrm{s}}(u_h))\|_{H^{-1}(D_F)} + \phi_F^{\mathrm{s}}(u_h, f, \mathrm{d}) \right).$$

Since $\langle T_F^{\mathrm{s}}(r^{\mathrm{s}}(u_h)), \varphi \rangle := -\int_{D_F} r^{\mathrm{v}}(u_h) \varphi \, \mathrm{d}x + \langle \rho(u_h), \varphi \rangle$ and $\|\varphi\|_{L^2(D_F)} \le c h_F \|\nabla\varphi\|_{\boldsymbol{L}^2(D_F)}$ for all $\varphi \in H_0^1(D_F)$ owing to the Poincaré-Steklov inequality in $H_0^1(D_F)$ (the constant c is independent on F and h), we infer that

$$\|T_F^{\mathrm{s}}(r^{\mathrm{s}}(u_h))\|_{H^{-1}(D_F)} \le c \big(h_F \|r^{\mathrm{v}}(u_h)\|_{L^2(D_F)} + \|\rho(u_h)\|_{H^{-1}(D_F)} \big).$$

Since $\eta_K^{\mathrm{s}}(u_h) := h_K^{\frac{1}{2}} \|r^{\mathrm{s}}(u_h)\|_{L^2(\mathcal{F}_K^\circ)}$, we conclude by using the regularity of the mesh sequence, the above bound on $r^{\mathrm{v}}(u_h)$, and (34.17). $\square$

Remark 34.20 (Oscillation). The oscillation term somehow pollutes the local lower bound in Theorem 34.19. As emphasized above, this is the price to pay to have computable error indicators. It is usually recommended in the literature to take $l^{\mathrm{v}} := 2k - 2$ and $l^{\mathrm{s}} := 2k - 1$ for general d, where k is the polynomial degree of the finite elements. When the diffusion tensor d is piecewise constant, taking $l^{\mathrm{s}} := k - 1$ makes the face-based oscillation ϕ_F^{s} to vanish (i.e., it is not necessary to invoke ϕ_F^{s}), and taking $l^{\mathrm{s}} := k - 1$ transforms the cell-based oscillation into a data oscillation since in this case $\phi_K^{\mathrm{v}} = h_K \|f - \overline{f}^{\mathrm{v}}\|_{L^2(K)}$. With the above choices for l^{v} and l^{s}, the oscillation is expected to be of higher-order than the approximation error (see Cascón et al. [113] and Exercise 34.4). On coarse meshes however the oscillation can be the dominant (or even be the only) contribution to the approximation error. $\square$

34.4   Adaptivity

This section outlines important ideas and results on adaptive mesh refinement driven by a posteriori error estimates. We do not consider mesh coarsening, even though this is also a practically important topic. The analysis of adaptive finite element methods (i.e., finite element solvers employing adaptive mesh refinement) has witnessed extensive progress over the years. The convergence of the adaptive procedure and its (quasi-)optimality in terms of error decay rates as a function of the number of the degrees of freedom is now well understood. Seminal contributions include those in Dörfler [171], Morin et al. [305], Binev et al. [56], Stevenson [356, 357], Cascón et al. [113]. Comprehensive surveys can be found in Nochetto et al. [316], Nochetto and Veeser [315], and Verfürth [378, p. 264]. An axiomatic presentation with numerous references is proposed in Carstensen et al. [112].

Algorithm 34.1 Adaptive finite element solver.

    Build an initial grid $\mathcal{T}_0$ and choose a tolerance `TOL`
    for $n = 0, 1, \dots$ until $\eta(u_n, \mathcal{T}_n) \leq$ `TOL` **do**
        $u_n \leftarrow$ `SOLVE`$(\mathcal{T}_n)$
        $\{\eta_K(u_n)\}_{K \in \mathcal{T}_n} \leftarrow$ `ESTIMATE`$(u_n, \mathcal{T}_n)$
        $\mathcal{M}_n \leftarrow$ `MARK`$(\{\eta_K(u_n)\}_{K \in \mathcal{T}_n}, \mathcal{T}_n)$
        $\mathcal{T}_{n+1} \leftarrow$ `REFINE`$(\mathcal{M}_n, \mathcal{T}_n)$
        $n \leftarrow n + 1$
    end for

The core of an adaptive finite element solver is outlined in Algorithm 34.1, which generates a sequence of (matching simplicial) meshes $\mathcal{T}_0, \mathcal{T}_1, \dots$ (we omit the subscript h to simplify the notation). The module `SOLVE` consists of building the finite element space V_n from the current mesh $\mathcal{T}_n$ and solving for the discrete solution $u_n \in V_n$. The module `ESTIMATE` computes the cell-based error estimators $\{\eta_K(u_n)\}_{K \in \mathcal{T}_n}$ defined in (34.16). The module `MARK` uses these estimators to mark some cells in $\mathcal{T}_n$ for refinement. The marked cells are collected in the set $\mathcal{M}_n \subset \mathcal{T}_n$. The fourth module `REFINE` uses the marked cells in $\mathcal{M}_n$ and the current mesh $\mathcal{T}_n$ to build a new mesh $\mathcal{T}_{n+1}$ for the next iteration. The termination criterion of the adaptive loop compares the global upper bound $\eta(u_n, \mathcal{T}_n) := (\sum_{K \in \mathcal{T}_n} \eta_K(u_n)^2)^{\frac{1}{2}}$ to the user-prescribed tolerance `TOL`.

The modules `SOLVE` and `ESTIMATE` have been already discussed. The module `MARK` selects mesh cells for refinement using Dörfler's marking [171] (also called *bulk chasing* criterion) as follows: Given a fixed parameter $\theta \in (0, 1)$, `MARK` determines a set $\mathcal{M}_n \subset \mathcal{T}_n$ of (almost) minimal cardinality such that

$$\eta(u_n, \mathcal{M}_n) \geq \theta \, \eta(u_n, \mathcal{T}_n), \tag{34.23}$$

where $\eta(u_n, \mathcal{M}_n) := (\sum_{K \in \mathcal{M}_n} \eta_K(u_n)^2)^{\frac{1}{2}}$. This marking means that the set $\mathcal{M}_n$ contains a substantial part of the total (or bulk) error. Taking θ small typically means that few mesh cells are marked. A mesh of minimal cardinality $\mathcal{M}_n^*$ is one such that $\text{card}(\mathcal{M}_n) \geq \text{card}(\mathcal{M}_n^*)$ for all $\mathcal{M}_n \subset \mathcal{T}_n$ s.t. $\eta(u_n, \mathcal{M}_n) \geq \theta \, \eta(u_n, \mathcal{T}_n)$. Building a set $\mathcal{M}_n^*$ of minimal cardinality entails sorting all the mesh cells, which is of superlinear complexity (for instance the complexity of the merge-sort algorithm is $\text{card}(\mathcal{T}_n) \ln(\text{card}(\mathcal{T}_n))$). By relaxing the minimality requirement, one can use a sorting algorithm of linear complexity based on binning, thereby producing a set $\mathcal{M}_n$ of cardinality $\text{card}(\mathcal{M}_n) \leq c \, \text{card}(\mathcal{M}_n^*)$ for some uniform constant c.

The module `REFINE` refines all the marked cells in $\mathcal{M}_n$ at least once. Refining mesh cells is usually done by using a double labeling technique indicating how the cells are to be subdivided and giving a rule to label the newly created subcells. An important example in dimension two
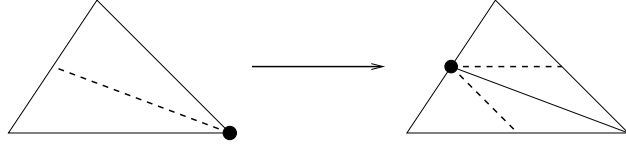
Figure 34.1: Newest vertex bisection: mesh cell with flagged vertex and dashed line indicating the bisecting method (left). If the cell is marked for refinement, two new cells are produced (right) and the newest vertex is flagged in both cells.

is the Newest vertex bisection (NVB), where one vertex of the cell is labeled to indicate that the opposite edge is to be bisected if refinement is required. If two new subcells are indeed created, the midpoint of the bisected edge is in turn labeled; see Figure 34.1. A three-dimensional extension of the NVB exists; see Stevenson [357]. One can verify that a sequence of meshes produced by NVB is shape-regular. However, since the new mesh $\mathcal{T}_{n+1}$ must remain a matching mesh, the module REFINE cannot be completely local. The propagation of refinement beyond the set of marked cells is a rather delicate issue. A crucial result on the cumulative effect of refinement propagation shown in Binev et al. [56] for $d = 2$ and [357] for $d > 2$ is that, provided the initial labeling of $\mathcal{T}_0$ satisfies some suitable requirements, there is a uniform constant c such that any sequence of successively bisected meshes satisfies the following bound:

$$\text{card}(\mathcal{T}_{n+1}) - \text{card}(\mathcal{T}_0) \leq c \sum_{m \in \{0:n\}} \text{card}(\mathcal{M}_m), \tag{34.24}$$

whereas single-step uniform bounds of the form $\text{card}(\mathcal{T}_{m+1}) - \text{card}(\mathcal{T}_m) \leq c\,\text{card}(\mathcal{M}_m)$ may not hold true.

The first important result for the adaptive finite element solver is a contraction property (implying convergence with geometric rate). This property can be stated on the quasi-error defined as a weighted sum of the approximation error plus the estimator. In particular, it is shown in Cascón et al. [113] that using Dörfler's marking and bisecting marked elements at least once, there exist $\gamma > 0$ and $\rho \in (0, 1)$ such that

$$\mathcal{E}_{n+1} \leq \rho \mathcal{E}_n, \tag{34.25}$$

where $\mathcal{E}_n := \|\nabla(u - u_n)\|_{\boldsymbol{L}^2(D)} + \gamma\eta(u_n, \mathcal{T}_n)$. The proof uses the global error upper bound, but not the lower bound. The symmetry and coercivity of the bilinear form a and the nesting of the finite element spaces are also used. Strict error reduction ($\gamma = 0$ in the definition of $\mathcal{E}_n$) is not true in general as shown in Morin et al. [305].

The second important result deals with convergence rates. For simplicity, we first discuss the case without oscillation, and we consider the Laplacian with piecewise polynomial source term on the initial mesh $\mathcal{T}_0$. For a real number $s > 0$ and a function $y \in H_0^1(D)$, we consider the following quantity:

$$|y|_{A_s} := \sup_{N > 0} N^s \inf_{\mathcal{T} \in \mathbb{T}_N} \inf_{v \in V(\mathcal{T})} \|\nabla(y - v)\|_{\boldsymbol{L}^2(D)}, \tag{34.26}$$

where $\mathbb{T}_N$ is the set of matching simplicial meshes that are refinements of the initial mesh $\mathcal{T}_0$ with cardinality such that $\text{card}(\mathcal{T}) - \text{card}(\mathcal{T}_0) \leq N$, and where $V(\mathcal{T})$ is the $H_0^1(D)$-conforming finite element space (of some order k) built using the mesh $\mathcal{T}$. Observe that $\inf_{v \in V(\mathcal{T})} \|\nabla(y - v)\|_{\boldsymbol{L}^2(D)}$ represents the best-approximation error of y in $V(\mathcal{T})$. Moreover, the discrete solution $u_{\mathcal{T}} \in V(\mathcal{T})$ from the Galerkin approximation delivers the quasi-optimal error bound $\|\nabla(u - u_{\mathcal{T}})\|_{\boldsymbol{L}^2(D)} \leq \frac{M}{\alpha} \inf_{v \in V(\mathcal{T})} \|\nabla(u - v)\|_{\boldsymbol{L}^2(D)}$. Using (34.26), we define the approximation class

$$A_s := \{y \in H_0^1(D) \mid |y|_{A_s} < \infty\}. \tag{34.27}$$

Membership in A_s informs on the optimal decay rate one can expect for the approximation error. Specifically, if u is in A_s, then there is c s.t. for every $N > 0$, there is an optimal mesh $\mathcal{T}_N^*$ such that $\|\nabla(u - u_{\mathcal{T}_N^*})\|_{\boldsymbol{L}^2(D)} \leq cN^{-s}|u|_{A_s}$. Finding an optimal mesh $\mathcal{T}_N^*$ is computationally untractable. Fortunately, it turns out that the adaptive finite element procedure from Algorithm 34.1 selects meshes $\{\mathcal{T}_n\}_{n\geq 0}$ delivering optimal decay rates. Indeed, it is shown in Cascón et al. [113] (see also Binev et al. [56], Stevenson [357]) that using Dörfler's marking with a parameter θ small enough together with a sorting algorithm such that $\mathrm{card}(\mathcal{M}_n) \leq c\,\mathrm{card}(\mathcal{M}_n^*)$ for some uniform constant c, and if the complexity estimate (34.24) for REFINE holds true, there is c such that for all $n \geq 1$,

$$\|\nabla(u - u_n)\|_{\boldsymbol{L}^2(D)} \leq c\,|u|_{A_s}(\mathrm{card}(\mathcal{T}_n) - \mathrm{card}(\mathcal{T}_0))^{-\frac{1}{s}}. \tag{34.28}$$

Note that the error lower bound is used in this proof. In the general case with oscillations, the problem data f and d are included in the definition of the approximation class, and the decay rate is established in Cascón et al. [113] for the total error defined as the sum of the approximation error and the data oscillation. An alternative viewpoint (see Carstensen et al. [112]) is to introduce approximation classes and decay rates for the estimator, and then use the error lower bound to infer decay rates for the approximation error.

Exercises

Exercise 34.1 (Residual). Prove (34.10). (*Hint*: integrate by parts.)

Exercise 34.2 (Trace inequality in stars). Let $C_{\mathrm{tr},\boldsymbol{z}}$ be defined in (34.12). Prove that $C_{\mathrm{tr},\boldsymbol{z}} \leq \varpi_{\boldsymbol{z}}^{\frac{1}{2}}(dC_{\mathrm{PS},\boldsymbol{z}}^2 + 2C_{\mathrm{PS},\boldsymbol{z}})^{\frac{1}{2}}$ with $\varpi_{\boldsymbol{z}} := h_{D_{\boldsymbol{z}}} \max_{F \in \mathcal{F}_{\boldsymbol{z}}^\circ} \frac{|F|}{|D_F|}$ and $D_F := \mathrm{int}(K_l \cup K_r)$ with $F := \partial K_l \cap \partial K_r$. (*Hint*: see the proof of Lemma 12.15.)

Exercise 34.3 (Bound on dual norm). (i) Prove that $\|T_K^{\mathrm{v}}(f)\|_{H^{-1}(K)} \leq ch_K\|f\|_{L^2(K)}$ for all $f \in L^2(K)$. (*Hint*: use a scaled Poincaré–Steklov inequality for functions $\varphi \in H_0^1(K)$.) (ii) Prove that $\|T_F^{\mathrm{s}}(g)\|_{H^{-1}(D_F)} \leq ch_F^{\frac{1}{2}}\|g\|_{L^2(F)}$ for all $g \in L^2(F)$. (*Hint*: use the multiplicative trace inequality from Lemma 12.15.)

Exercise 34.4 (Oscillation). (i) Let $P_m^{(p)} : L^p(K) \to \mathbb{P}_m$ be the best-approximation operator in $L^p(K)$ for $p \in [1, \infty]$ and $m \in \mathbb{N}$. Prove that

$$\|(I - P_m^{(2)})(\theta v_h)\|_{L^2(K)} \leq \|(I - P_{m-n}^{(\infty)})(\theta)\|_{L^\infty(K)}\|v_h\|_{L^2(K)},$$

for all $\theta \in L^\infty(K)$ and all $v_h \in \mathbb{P}_n$ with $n \leq m$. (ii) Consider the oscillation indicators defined in (34.19) with $l^{\mathrm{v}} := 2k-2$ and $l^{\mathrm{s}} := 2k-1$. Prove that $\phi_K^{\mathrm{v}}(u_h, f, \mathrm{d}) \leq h_K\|(I - P_{2k-2}^{(2)})(f)\|_{L^2(K)} + c(\|(I - P_{k-1}^{(\infty)})(\nabla\cdot\mathrm{d})\|_{\boldsymbol{L}^\infty(K)} + \|(I - P_k^{(\infty)})(\mathrm{d})\|_{\mathbb{L}^\infty(K)})\|\nabla u_h\|_{\boldsymbol{L}^2(K)}$ with $(\nabla\cdot\mathrm{d})_i := \sum_{j\in\{1:d\}} \frac{\partial}{\partial x_j}\mathrm{d}_{ji}$ for all $i \in \{1{:}d\}$. Prove that $\phi_F^{\mathrm{s}}(u_h, f, \mathrm{d}) \leq c\|(I - P_k^{(\infty)})(\mathrm{d})\|_{\mathbb{L}^\infty(F)}\|\nabla u_h\|_{\boldsymbol{L}^2(D_F)}$ with best-approximation operator $P_k^{(\infty)}$ mapping to $L^\infty(F)$. What are the decay rates of the oscillation terms for smooth f and d? (iii) What happens if $l^{\mathrm{v}} := k$ and $l^{\mathrm{s}} := k-1$ for piecewise constant d?

Exercise 34.5 (Error reduction). Consider two discrete spaces $V_{h_1} \subset V_{h_2} \subset H_0^1(D)$ with corresponding discrete solutions u_{h_1} and u_{h_2}, respectively. Consider the norm $\|v\|_a := a(v,v)^{\frac{1}{2}}$ for all $v \in H_0^1(D)$. Prove that $\|u - u_{h_1}\|_a^2 = \|u - u_{h_2}\|_a^2 + \|u_{h_2} - u_{h_1}\|_a^2$. (*Hint*: use the Galerkin orthogonality property.)

Exercise 34.6 (Approximation class for smooth solution). Let D be a Lipschitz polyhedron in $\mathbb{R}^d$. Prove that $H^{k+1}(D) \subset A_{k/d}$. (*Hint*: consider uniformly refined meshes.)

Exercise 34.7 (Graded mesh). Let $D := (0,1)$ and let $(x_i)_{i\in\{0:I\}}$, $I \geq 2$, be a mesh of D. Let $u \in W^{1,1}(D)$ and consider the piecewise constant function u_I such that $u_I(x) := u(x_{i-1})$ for all $x \in (x_{i-1}, x_i)$ and all $i \in \{1{:}I\}$. (i) Assume $u \in W^{1,\infty}(D)$. Prove that the decay rate $\|u - u_I\|_{L^\infty(D)} \leq \frac{1}{I}\|u'\|_{L^\infty(D)}$ is achieved using a uniform mesh. (ii) Assume now $u \in W^{1,1}(D)$. Prove that the decay rate $\|u - u_I\|_{L^\infty(D)} \leq \frac{1}{I}\|u'\|_{L^1(D)}$ is achieved using a graded mesh such that $x_i := \Phi^{(-1)}(\frac{i}{I})$, where $\Phi(s) := \frac{1}{\|u'\|_{L^1(D)}} \int_0^s |u'(t)|\, \mathrm{d}t$ for all $s \in (0,1)$ and all $i \in \{0{:}I\}$.

Chapter 35

The Helmholtz problem

The objective of this chapter is to give a brief overview of the analysis of the Helmholtz problem and its approximation using H^1-conforming finite elements. The Helmholtz problem arises when modeling electromagnetic or acoustic scattering problems in the frequency domain. One specificity of this elliptic problem is that one cannot apply the Lax-Milgram lemma to establish well-posedness. The correct way to tackle the Helmholtz problem is to invoke the BNB theorem (Theorem 25.9). In the entire chapter, D is a Lipschitz domain in $\mathbb{R}^d$ with $d \geq 1$, i.e., a nonempty open bounded and connected subset of $\mathbb{R}^d$ with a Lipschitz boundary.

35.1 Robin boundary conditions

We investigate in this section the Helmholtz problem with Robin boundary conditions. Given $f \in L^2(D)$, $g \in L^2(\partial D)$, and $\kappa \in \mathbb{R}$, our goal is to find a function $u : D \to \mathbb{C}$ such that

$$-\Delta u - \kappa^2 u = f \quad \text{in } D, \qquad \partial_n u - \mathrm{i}\kappa u = g \quad \text{on } \partial D, \tag{35.1}$$

with $\mathrm{i}^2 = -1$. Notice that the Robin boundary condition couples the real and imaginary parts of u. The sign of the parameter κ is irrelevant in what follows, but to simplify some expressions, we henceforth assume that $\kappa > 0$. All that is said below remains valid when $\kappa < 0$ by replacing κ by $|\kappa|$ in the definitions of the norms and in the upper bounds. Note that κ^{-1} is a length scale. The problem (35.1) can be reformulated as follows in weak form:

$$\begin{cases} \text{Find } u \in V := H^1(D) \text{ such that} \\ a(u, w) = \ell(w), \quad \forall w \in V, \end{cases} \tag{35.2}$$

with the sesquilinear form

$$a(v, w) := \int_D (\nabla v \cdot \nabla \overline{w} - \kappa^2 v \overline{w}) \, \mathrm{d}x - \mathrm{i}\kappa \int_{\partial D} \gamma^{\mathrm{g}}(v) \gamma^{\mathrm{g}}(\overline{w}) \, \mathrm{d}s, \tag{35.3}$$

and the antilinear form $\ell(w) := \int_D f\overline{w} \, \mathrm{d}x + \int_{\partial D} g\gamma^{\mathrm{g}}(\overline{w}) \, \mathrm{d}s$, where $\gamma^{\mathrm{g}} : H^1(D) \to H^{\frac{1}{2}}(\partial D)$ is the trace map.

Remark 35.1 (Sommerfeld radiation condition). The Helmholtz problem is in general posed on unbounded domains, and the proper "boundary condition to set at infinity" is the Sommerfeld

radiation condition $\lim_{r\to\infty} r^{\frac{d-1}{2}}(\boldsymbol{e}{\cdot}\nabla u(r\boldsymbol{e}) - \mathrm{i}\kappa u(r\boldsymbol{e})) = 0$ for every unit vector $\boldsymbol{e} \in \mathbb{R}^d$ and the convergence must be uniform with respect to $\boldsymbol{n}$. One usually simplifies this problem by truncating the domain and replacing the Sommerfeld radiation condition by a Robin boundary condition as in (35.1). □

Remark 35.2 (Wave equation). The Helmholtz problem can be derived by considering the wave equation $\partial_{tt} v - c^2 \Delta v = g(\boldsymbol{x})\cos(\omega t)$ in $D{\times}(0,T)$ with appropriate initial data and boundary conditions; see §46.2.1 and §46.2.2. Here, c is the wave speed and g is some forcing. Assuming that the solution is of the form $v(\boldsymbol{x},t) = \Re(u(\boldsymbol{x})e^{\mathrm{i}\omega t})$, the complex amplitude u solves $\omega^2 u - c^2 \Delta u = g$. We then recover (35.1) by setting $\kappa := \frac{\omega}{c}$. □

35.1.1 Well-posedness

Contrary to what was done in the previous chapters, we cannot apply the Lax–Milgram lemma to establish that the weak formulation (35.2) is well-posed since the sesquilinear form a is not coercive. We are going to invoke instead the BNB theorem (Theorem 25.9), and with this goal in mind, we first establish an abstract result.

Lemma 35.3 (Gårding). *Let $V \hookrightarrow L$ be two Banach spaces with compact embedding. Let $a : V{\times}V \to \mathbb{C}$ be a bounded sesquilinear form. Assume that there exist two real numbers $\beta, \gamma > 0$ such that the following holds true:*

$$|a(v,v)| + \beta\|v\|_L^2 \geq \gamma\|v\|_V^2, \quad \forall v \in V, \tag{35.4a}$$

$$[\, a(v,w) = 0, \ \forall w \in V \,] \implies [\, v = 0 \,]. \tag{35.4b}$$

Then there is $\alpha > 0$ such that $\inf_{v\in V} \sup_{w\in V} \frac{|a(v,w)|}{\|v\|_V \|w\|_V} \geq \alpha$.

Proof. Let us argue by contradiction like in the proof of the Peetre–Tartar lemma (Lemma A.20). Assume that for every integer $n \geq 1$, there is $v_n \in V$ with $\|v_n\|_V = 1$ and $\sup_{w\in V} |a(v_n,w)|/\|w\|_V \leq \frac{1}{n}$. Since the embedding $V \hookrightarrow L$ is compact, there is a subsequence $(v_l)_{l\in\mathcal{S}}$, $\mathcal{S} \subset \mathbb{N}$, such that $(v_l)_{l\in\mathcal{S}}$ converges strongly to some v in L. The assumption (35.4a) implies that

$$\gamma\|v_m - v_n\|_V^2 \leq \beta\|v_m - v_n\|_L^2 + |a(v_m - v_n, v_m - v_n)|$$
$$\leq \beta\|v_m - v_n\|_L^2 + |a(v_m, v_m)| + |a(v_m, v_n)| + |a(v_n, v_m)| + |a(v_n, v_n)|.$$

Since $|a(v_l, v_{l'})| = |a(v_l, v_{l'})|/\|v_{l'}\|_V \leq \frac{1}{l}$, for all $l, l' \in \{m, n\}$, we infer that $\gamma\|v_m - v_n\|_V^2 \leq \beta\|v_m - v_n\|_L^2 + 2(m^{-1} + n^{-1})$, which in turn implies that $(v_l)_{l\in\mathcal{S}}$ is a Cauchy sequence in V. As a result, $v \in V$ and $\sup_{w\in V} |a(v,w)|/\|w\|_V = 0$, which means that $a(v,w) = 0$ for all $w \in V$. The assumption (35.4b) implies that $v = 0$, which contradicts $1 = \lim_{\mathcal{S}\ni l\to\infty} \|v_l\|_V = \|v\|_V$. □

Remark 35.4 (Gårding's inequality). Inequalities like (35.4a) are called *Gårding's inequality* in the literature. □

Theorem 35.5 (BNB, Robin BCs). *Let $V := H^1(D)$ be equipped with the norm $\|v\|_V := \{\|\nabla v\|_{\boldsymbol{L}^2(D)}^2 + \kappa\|v\|_{L^2(\partial D)}^2\}^{\frac{1}{2}}$. The sesquilinear form a defined in (35.3) satisfies the conditions of the BNB theorem.*

Proof. We are going to verify (35.4a) and (35.4b) from Lemma 35.3.
(1) Let $v \in V$. The real and imaginary parts of $a(v,v)$ are

$$\Re(a(v,v)) = \|\nabla v\|_{\boldsymbol{L}^2(D)}^2 - \kappa^2\|v\|_{L^2(D)}^2, \tag{35.5a}$$

$$\Im(a(v,v)) = -\kappa\|v\|_{L^2(\partial D)}^2. \tag{35.5b}$$

Using that $\sqrt{2}(x^2 + y^2)^{\frac{1}{2}} \geq x - y$ for all $x, y \in \mathbb{R}$, this implies that

$$\sqrt{2}|a(v,v)| \geq \|v\|_V^2 - \kappa^2 \|v\|_{L^2(D)}^2.$$

Hence, (35.4a) holds true with $\beta := \frac{1}{\sqrt{2}}\kappa^2$ and $\gamma := \frac{1}{\sqrt{2}}$.

(2) Let us now assume that $a(v, w) = 0$ for all $w \in V$. We are going to prove that $v = 0$ by arguing by contradiction. The inequality $|a(v,v)| \geq -\Im(a(v,v)) = \kappa\|v\|_{L^2(\partial D)}^2$ implies that $\gamma^{\mathrm{g}}(v) = 0$. Hence, $v \in H_0^1(D)$. Let us embed D into a ball of radius R large enough, say $R > R_0 := \mathrm{diam}(D)$, and without loss of generality, we assume that this ball is centered at $\mathbf{0}$. Let B_R be the ball in question and let us set $D_R^{\mathrm{c}} := D^{\mathrm{c}} \cap B_R$, where D^{c} denotes the complement of D in $\mathbb{R}^d$. Since $v_{|\partial D} = 0$, we can extend v by zero over D_R^{c}, and we denote by $\widetilde{v}_R$ the extension in question. We have $\widetilde{v}_R \in H_0^1(B_R)$, $(\nabla \widetilde{v}_R)_{|D} \in \boldsymbol{H}(\mathrm{div}; D)$, and $(\nabla \widetilde{v}_R)_{|D_R^{\mathrm{c}}} \in \boldsymbol{H}(\mathrm{div}; D_R^{\mathrm{c}})$. Since the Robin boundary condition implies that $\partial_n v_{|\partial D} = 0$, we infer that the normal component of $\nabla \widetilde{v}_R$ is continuous across ∂D. Reasoning as in the proof of Theorem 18.10, we conclude that $\nabla \widetilde{v}_R$ is a member of $\boldsymbol{H}(\mathrm{div}; B_R)$. This means that $\Delta \widetilde{v}_R \in L^2(B_R)$. Since $\widetilde{v}_R \in H_0^1(B_R)$ and $\widetilde{v}_R$ vanishes on an open subset of B_R, we can invoke the unique continuation principle (see Theorem 31.4) to infer that $\widetilde{v}_R = 0$ in B_R. Hence, $v = 0$ in D and the property (35.4b) holds true. $\qquad\square$

Remark 35.6 (Alternative proof). Instead of invoking the unique continuation principle in the above proof, one can use the spectral theorem for symmetric compact operators (see Theorem 46.21). The above reasoning shows that $\widetilde{v}_R \in H_0^1(B_R)$ and $-\Delta \widetilde{v}_R = \kappa^2 \widetilde{v}_R$ in B_R. Hence, if $\widetilde{v}_R$ is not zero, then κ^2 is an eigenvalue of the Laplace operator equipped with homogeneous Dirichlet boundary conditions on every ball centered at $\mathbf{0}$ in $\mathbb{R}^d$ with radius larger than R_0. However, Theorem 46.21 says that the eigenvalues of the Laplace operator in $H_0^1(B_R)$ are countable with no accumulation point and are of the form $(R^{-2}\lambda_n)_{n \in \mathbb{N}}$ for every $R > 0$, where $(\lambda_n)_{n \in \mathbb{N}}$ are the eigenvalues of the Laplace operator in $H_0^1(B_1)$. Assuming that the eigenvalues are ordered in increasing order, let $R_0' > R_0$ be large enough so that there is some $n \in \mathbb{N}$ such that $\kappa^2(R_0')^2 = \lambda_n$ with $\lambda_n < \lambda_{n+1}$. Let δ be defined by $\kappa^2(R_0' + \delta)^2 := \frac{1}{2}(\lambda_n + \lambda_{n+1})$. Then $\kappa^2(R_0' + \delta)^2$ cannot be in the set $\{\lambda_n\}_{n \in \mathbb{N}}$, but this is a contradiction since the above reasoning with $R := R_0' + \delta$ shows that $\kappa^2 R^2 = \kappa^2(R_0' + \delta)^2$ is a member of the sequence $(\lambda_n)_{n \in \mathbb{N}}$ if $\widetilde{v}_R$ is not zero. This proves that $\widetilde{v}_R = 0$. $\qquad\square$

35.1.2 A priori estimates on the solution

In this section, we derive a priori estimates on the weak solution of (35.2). We are particularly interested in estimating the possible dependence of the upper bound on the (nondimensional) quantity $\kappa\ell_D$ with $\ell_D := \mathrm{diam}(D)$. The following result, established in Melenk [299, Prop. 8.1.4] and Hetmaniuk [243], delivers a sharp upper bound on the V-norm of the weak solution that relies on the relatively strong assumption that the domain D is star-shaped with respect to some point in D which we take to be $\mathbf{0}$.

Lemma 35.7 (A priori estimate). *Assume that D is a bounded Lipschitz domain and star-shaped w.r.t. $\mathbf{0}$, i.e., there exists $r > 0$ s.t. $\boldsymbol{x}\cdot\boldsymbol{n} > r\ell_D$ for all $\boldsymbol{x} \in \partial D$. Let $V := H^1(D)$ be equipped with the norm $\|v\|_V := \{\|\nabla v\|_{\boldsymbol{L}^2(D)}^2 + \kappa\|v\|_{L^2(\partial D)}^2\}^{\frac{1}{2}}$. There is a constant c that depends only on D (i.e., it is independent of $\kappa\ell_D$) such that the weak solution of (35.2) satisfies*

$$\kappa\|u\|_{L^2(D)} + \|u\|_V \leq c\,(\ell_D\|f\|_{L^2(D)} + \ell_D^{\frac{1}{2}}\|g\|_{L^2(\partial D)}). \tag{35.6}$$

Proof. We only give the proof when κ is bounded away from zero, say $\kappa\ell_D \geq 1$ since the proof in the other case is similar; see [299, 243]. Since we assume that $\mathbf{0} \in D$, we have $\|\boldsymbol{x}\|_{\ell^2} \leq \ell_D$ for all

$\boldsymbol{x} \in D$. We write $C(f,g) := c(\ell_D \|f\|_{L^2(D)} + \ell_D^{\frac{1}{2}} \|g\|_{L^2(\partial D)})$, where as usual the value of the constant c can change at each occurrence as long as it is independent of κ.

(1) In the first step of the proof, we assume that $\nabla u_{|\partial D} \in \boldsymbol{L}^2(\partial D)$ (we establish this smoothness property in the second step). Let us multiply the PDE $-\Delta u - \kappa^2 u = f$ with $\boldsymbol{x}\cdot\nabla\overline{u}$ and integrate over D. The identity (35.11) from Lemma 35.8 with $\boldsymbol{m} := \boldsymbol{x}$ implies that

$$-\Re\left(\int_D \Delta u\, \boldsymbol{x}\cdot\nabla\overline{u}\,\mathrm{d}x\right) = \left(1 - \frac{d}{2}\right)\|\nabla u\|_{\boldsymbol{L}^2(D)}^2$$
$$+ \frac{1}{2}\int_{\partial D}(\boldsymbol{x}\cdot\boldsymbol{n})\|\nabla u\|_{\ell^2}^2\,\mathrm{d}s - \Re\left(\int_{\partial D}(\partial_n u)(\boldsymbol{x}\cdot\nabla\overline{u})\,\mathrm{d}s\right),$$

since $\nabla\boldsymbol{x} = (\nabla\boldsymbol{x})^{\mathsf{T}} = \mathbb{I}_d$ and $\nabla\cdot\boldsymbol{x} = d$ so that $\mathbb{e}(\boldsymbol{x}) = (1 - \frac{d}{2})\mathbb{I}_d$ (see Lemma 35.8). This identity is often called *Rellich's identity* in the literature. Using the PDE $-\Delta u - \kappa^2 u = f$, the Robin boundary condition $\partial_n u = \mathrm{i}\kappa u + g$, and the assumption $\boldsymbol{x}\cdot\boldsymbol{n} > r\ell_D$ on ∂D, we obtain

$$\frac{r\ell_D}{2}\|\nabla u\|_{\boldsymbol{L}^2(\partial D)}^2 \le \left(\frac{d}{2} - 1\right)\|\nabla u\|_{\boldsymbol{L}^2(D)}^2 + \Re\left(\int_D \kappa^2 u(\boldsymbol{x}\cdot\nabla\overline{u})\,\mathrm{d}x\right)$$
$$+ \Re\left(\int_D f(\boldsymbol{x}\cdot\nabla\overline{u})\,\mathrm{d}x\right) + \Re\left(\int_{\partial D}(\mathrm{i}\kappa u + g)(\boldsymbol{x}\cdot\nabla\overline{u})\,\mathrm{d}s\right).$$

Since $\Re(\int_D u(\boldsymbol{x}\cdot\nabla\overline{u})\,\mathrm{d}x) = -\frac{d}{2}\|u\|_{L^2(D)}^2 + \frac{1}{2}\int_{\partial D}(\boldsymbol{x}\cdot\boldsymbol{n})|u|^2\,\mathrm{d}s$, this leads to

$$\frac{r\ell_D}{2}\|\nabla u\|_{\boldsymbol{L}^2(\partial D)}^2 + \frac{d\kappa^2}{2}\|u\|_{L^2(D)}^2 \le \left(\frac{d}{2} - 1\right)\|\nabla u\|_{\boldsymbol{L}^2(D)}^2 + \frac{\kappa^2\ell_D}{2}\|u\|_{L^2(\partial D)}^2$$
$$+ \Re\left(\int_D f(\boldsymbol{x}\cdot\nabla\overline{u})\,\mathrm{d}x\right) + \Re\left(\int_{\partial D}(\mathrm{i}\kappa u + g)(\boldsymbol{x}\cdot\nabla\overline{u})\,\mathrm{d}s\right).$$

We now bound the last two terms on the right-hand side by using Young's inequality, which yields

$$\Re\left(\int_D f(\boldsymbol{x}\cdot\nabla\overline{u})\,\mathrm{d}x\right) + \Re\left(\int_{\partial D}(\mathrm{i}\kappa u + g)(\boldsymbol{x}\cdot\nabla\overline{u})\,\mathrm{d}s\right) \le \gamma_1\|\nabla u\|_{\boldsymbol{L}^2(D)}^2$$
$$+ \frac{1}{4\gamma_1}\ell_D^2\|f\|_{L^2(D)}^2 + \frac{r\ell_D}{4}\|\nabla u\|_{\boldsymbol{L}^2(\partial D)}^2 + \frac{2\ell_D}{r}\left(\kappa^2\|u\|_{L^2(\partial D)}^2 + \|g\|_{L^2(\partial D)}^2\right),$$

where $\gamma_1 > 0$ can be chosen as small as needed. Rearranging the terms gives

$$\frac{r\ell_D}{4}\|\nabla u\|_{\boldsymbol{L}^2(\partial D)}^2 + \frac{d\kappa^2}{2}\|u\|_{L^2(D)}^2 \le \left(\frac{d}{2} - 1 + \gamma_1\right)\|\nabla u\|_{\boldsymbol{L}^2(D)}^2$$
$$+ \frac{r+4}{2r}\kappa^2\ell_D\|u\|_{L^2(\partial D)}^2 + C(f,g)^2. \quad (35.7)$$

Let us now bound the norms $\|\nabla u\|_{\boldsymbol{L}^2(D)}^2$ and $\|u\|_{L^2(\partial D)}^2$ appearing on the right-hand side. Owing to (35.5a) and Young's inequality, we infer that

$$\|\nabla u\|_{\boldsymbol{L}^2(D)}^2 = \kappa^2\|u\|_{L^2(D)}^2 + \Re\left((f,u)_{L^2(D)} + (g, \gamma^{\mathrm{g}}(u))_{L^2(\partial D)}\right)$$
$$\le (1 + \gamma_2)\kappa^2\|u\|_{L^2(D)}^2 + \frac{1}{4\gamma_2\kappa^2}\|f\|_{L^2(D)}^2 + \frac{1}{2\kappa}\|g\|_{L^2(\partial D)}^2 + \frac{1}{2}\kappa\|u\|_{L^2(\partial D)}^2,$$

where $\gamma_2 > 0$ can be chosen as small as needed. Since we assumed above that $\kappa\ell_D \ge 1$, we obtain

$$\|\nabla u\|_{\boldsymbol{L}^2(D)}^2 \le (1 + \gamma_2)\kappa^2\|u\|_{L^2(D)}^2 + \frac{1}{2}\kappa\|u\|_{L^2(\partial D)}^2 + C(f,g)^2. \quad (35.8)$$

Owing to (35.5b), we infer that

$$\kappa\|u\|^2_{L^2(\partial D)} = -\Im\Big((f,u)_{L^2(D)} + (g,\gamma^{\mathrm{g}}(u))_{L^2(\partial D)}\Big),$$

and applying Young's inequality with a positive real number θ gives

$$\frac{1}{2}\kappa\|u\|^2_{L^2(\partial D)} \le \theta\kappa\|u\|^2_{L^2(D)} + \frac{1}{4\theta\kappa}\|f\|^2_{L^2(D)} + \frac{1}{2\kappa}\|g\|^2_{L^2(\partial D)}.$$

Taking $\theta := \gamma_3\kappa$ with $\gamma_3 > 0$ as small as needed leads to (recall that $\kappa\ell_D \ge 1$)

$$\frac{1}{2}\kappa\|u\|^2_{L^2(\partial D)} \le \gamma_3\kappa^2\|u\|^2_{L^2(D)} + C(f,g)^2. \tag{35.9}$$

In addition, taking $\theta := \frac{1}{2\ell_D}\frac{r}{r+4}$ and multiplying by $\frac{r+4}{r}\kappa\ell_D$ yields

$$\frac{r+4}{2r}\kappa^2\ell_D\|u\|^2_{L^2(\partial D)} \le \frac{1}{2}\kappa^2\|u\|^2_{L^2(D)} + C(f,g)^2. \tag{35.10}$$

Inserting (35.9) into (35.8) gives $\|\nabla u\|^2_{\boldsymbol{L}^2(D)} \le (1+\gamma_2+\gamma_3)\kappa^2\|u\|^2_{L^2(D)} + C(f,g)^2$, and inserting this bound into (35.7), we obtain

$$\frac{r\ell_D}{4}\|\nabla u\|^2_{\boldsymbol{L}^2(\partial D)} + \frac{d\kappa^2}{2}\|u\|^2_{L^2(D)} \le \Big(\frac{d}{2}-1+\gamma_1\Big)(1+\gamma_2+\gamma_3)\kappa^2\|u\|^2_{L^2(D)}$$
$$+ \frac{r+4}{2r}\kappa^2\ell_D\|u\|^2_{L^2(\partial D)} + C(f,g)^2.$$

Using now the bound on $\|u\|^2_{L^2(\partial D)}$ from (35.10), we infer that

$$\frac{r\ell_D}{4}\|\nabla u\|^2_{\boldsymbol{L}^2(\partial D)} + \frac{d}{2}\kappa^2\|u\|^2_{L^2(D)} \le \Big(\Big(\frac{d}{2}-1+\gamma_1\Big)(1+\gamma_2+\gamma_3)+\frac{1}{2}\Big)\kappa^2\|u\|^2_{L^2(D)} + C(f,g)^2.$$

Letting $\gamma_1 := \frac{1}{4d}, \gamma_2 = \gamma_3 := \frac{1}{8d}$, we observe that $(\frac{d}{2}-1+\gamma_1)(1+\gamma_2+\gamma_3) = \frac{d}{2}-\frac{7}{8}+\frac{1}{16d^2} \le \frac{d}{2}-\frac{1}{4}$ for all $d \ge 1$. We conclude that

$$\frac{r\ell_D}{4}\|\nabla u\|^2_{\boldsymbol{L}^2(\partial D)} + \frac{\kappa^2}{4}\|u\|^2_{L^2(D)} \le C(f,g)^2.$$

Invoking once again the bounds (35.8) and (35.9), we infer that

$$\kappa^2\|u\|^2_{L^2(D)} + \kappa\|u\|^2_{L^2(\partial D)} + \|\nabla u\|^2_{\boldsymbol{L}^2(D)} + \ell_D\|\nabla u\|^2_{\boldsymbol{L}^2(\partial D)} \le C(f,g)^2,$$

which shows that the a priori estimate (35.6) holds true.

(2) It remains to prove that indeed $\nabla u_{|\partial D} \in \boldsymbol{L}^2(\partial D)$. Recall that u is in the functional space $Y := \{y \in H^1(D) \mid \Delta y \in L^2(D), \partial_n y \in L^2(\partial D)\}$ owing to (35.1) and our assumption that $f \in L^2(D)$ and $g \in L^2(\partial D)$. We are going to show by means of a density argument that any function $y \in Y$ is such that $\nabla y_{|\partial D} \in \boldsymbol{L}^2(\partial D)$. Let $(\varphi_m)_{m\in\mathbb{N}}$ be a sequence in $C^\infty(\overline{D})$ converging to y in Y (such a sequence can be constructed by using mollifying operators, as in §23.1). Let us set $f_m := -\Delta\varphi_m - \varphi_m$ and $g_m := \partial_n\varphi_m - \mathrm{i}\kappa\varphi_m$. Then $(f_m)_{m\in\mathbb{N}}$ and $(g_m)_{m\in\mathbb{N}}$ are Cauchy sequences in $L^2(D)$ and $L^2(\partial D)$, respectively. Moreover, the bound from Step (1) implies that $\|\nabla(\varphi_m - \varphi_p)\|_{\boldsymbol{L}^2(\partial D)} \le c(\ell_D^{\frac{1}{2}}\|f_m - f_p\|_{L^2(D)} + \|g_m - g_p\|_{L^2(\partial D)})$ for all $m, p \in \mathbb{N}$, which shows that $(\nabla\varphi_m)_{m\in\mathbb{N}}$ is a Cauchy sequence in $\boldsymbol{L}^2(\partial D)$. The uniqueness of the limit in the distribution sense finally shows that $\nabla y_{|\partial D} \in \boldsymbol{L}^2(\partial D)$. $\qquad\square$

Lemma 35.8 (Special identity). *For all* $q \in \{v \in H^1(D;\mathbb{C}) \mid \Delta v \in L^2(D;\mathbb{C}), \ \nabla v \in L^2(\partial D;\mathbb{C}^d)\}$ *and all* $\boldsymbol{m} \in W^{1,\infty}(D;\mathbb{R}^d)$*, letting* $\mathbb{e}(\boldsymbol{m}) := \frac{1}{2}(\nabla\boldsymbol{m} + (\nabla\boldsymbol{m})^{\mathsf{T}} - (\nabla\cdot\boldsymbol{m})\mathbb{I}_d)$*, we have*

$$-\Re\left(\int_D \Delta q(\boldsymbol{m}\cdot\nabla\overline{q})\,\mathrm{d}x\right) = \Re\left(\int_D \nabla q\cdot(\mathbb{e}(\boldsymbol{m})\nabla\overline{q})\,\mathrm{d}x\right)$$
$$+ \frac{1}{2}\int_{\partial D}(\boldsymbol{m}\cdot\boldsymbol{n})\|\nabla q\|_{\ell^2}^2\,\mathrm{d}s - \Re\left(\int_{\partial D}(\boldsymbol{n}\cdot\nabla q)(\boldsymbol{m}\cdot\nabla\overline{q})\,\mathrm{d}s\right). \quad (35.11)$$

Proof. See Exercise 35.4 and Hetmaniuk [243, Lem. 3.2]. $\qquad\square$

A detailed analysis of the Helmholtz problem (35.2) using integral representations is done in Esterhazy and Melenk [195, §2]. The following result is established therein.

Theorem 35.9 (BNB, Robin BCs). *Let* D *be a Lipschitz domain in* $\mathbb{R}^d$*,* $d \in \{2,3\}$*. Let* $V := H^1(D)$ *be equipped with the norm* $\|v\|_V := \kappa\|v\|_{L^2(D)} + \|\nabla v\|_{\boldsymbol{L}^2(D)}$*. Let* $k_0 > 0$ *be a fixed number and set* $\kappa_0 := k_0\ell_D^{-1}$*. Then there is* $c > 0$*, depending on* D *and* k_0*, such that the following holds true for all* $\kappa \geq \kappa_0$*:*

$$\inf_{v\in V}\sup_{w\in V}\frac{|a(v,w)|}{\|v\|_V\|w\|_V} \geq c\,(\kappa\ell_D)^{-s}, \quad (35.12)$$

with $s := \frac{7}{2}$ *in general, and* $s := 1$ *if* D *is convex or if* D *is star-shaped or if* ∂D *is smooth.*

This theorem implies, in particular, that for every $f \in V' := (H^1(D))'$ and $g \in H^{-\frac{1}{2}}(\partial D) = (H^{\frac{1}{2}}(\partial D))'$, the problem (35.2) is uniquely solvable in V, and its solution satisfies the a priori bound $\|u\|_V \leq c(\kappa\ell_D)^{\frac{7}{2}}(\|f\|_{V'} + \|g\|_{H^{-\frac{1}{2}}(\partial D)})$. If $f \in L^2(D)$ and $g \in L^2(\partial D)$, this estimate can be improved to $\|u\|_V \leq c(\kappa\ell_D)^{\frac{5}{2}}(\ell_D\|f\|_{L^2(D)} + \kappa^{-\frac{1}{2}}\|g\|_{L^2(\partial D)})$; see [195, Thm. 2.5].

35.2   Mixed boundary conditions

We consider in this section the Helmholtz problem with mixed Dirichlet and Robin boundary conditions. The problem is formulated as follows: For $f \in L^2(D)$, $g \in L^2(\partial D_\mathrm{r})$, and $\kappa \in \mathbb{R}$, find a complex-valued function u such that

$$-\Delta u - \kappa^2 u = f \text{ in } D, \quad u = 0 \text{ on } \partial D_\mathrm{d}, \quad \partial_n u - \mathrm{i}\kappa u = g \text{ on } \partial D_\mathrm{r}, \quad (35.13)$$

where $\{\partial D_\mathrm{d}, \partial D_\mathrm{r}\}$ is a partition of ∂D. We assume that the subsets ∂D_d and ∂D_r have a Lipschitz boundary and have positive (surface) measure. As before, we assume that $\kappa > 0$ for simplicity. The above problem is reformulated as follows:

$$\begin{cases} \text{Find } u \in V := \{v \in H^1(D) \mid \gamma^\mathrm{g}(v)_{|\partial D_\mathrm{d}} = 0\} \text{ such that} \\ a(u,w) = \ell(w), \quad \forall w \in V, \end{cases} \quad (35.14)$$

with the sesquilinear form

$$a(v,w) := \int_D (\nabla v\cdot\nabla\overline{w} - \kappa^2 v\overline{w})\,\mathrm{d}x - \mathrm{i}\kappa\int_{\partial D_\mathrm{r}} v\overline{w}\,\mathrm{d}s, \quad (35.15)$$

and the antilinear form $\ell(w) := \int_D f\overline{w}\,\mathrm{d}x + \int_{\partial D_\mathrm{r}} g\overline{w}\,\mathrm{d}s$. Here again, we cannot apply the Lax–Milgram lemma since a is not coercive on V. We are going to invoke instead the BNB theorem.

Theorem 35.10 (BNB, mixed BCs). *Let the space V defined in (35.14) be equipped with the norm $\|v\|_V := \|\nabla v\|_{\boldsymbol{L}^2(D)}$. The sesquilinear form a defined in (35.15) satisfies the conditions of the BNB theorem.*

Proof. We are going to invoke Lemma 35.3. We can proceed as in the proof of Theorem 35.5 to prove the Gårding inequality (35.4a), but we proceed slightly differently to prove (35.4b). Let us assume that $a(v, w) = 0$ for all $w \in V$. The inequality $|a(v, v)| \geq \kappa \|v\|_{L^2(\partial D_{\mathrm{r}})}^2$ implies that $v_{|\partial D_{\mathrm{r}}} = 0$. Since $|\partial D_{\mathrm{r}}| > 0$, there exists a point $\boldsymbol{x}_0 \in \partial D_{\mathrm{r}}$ and there is $r_0 > 0$ such that $B(\boldsymbol{x}_0, r_0) \cap \partial D \subset \partial D_{\mathrm{r}}$. Let $D_{r_0}^{\mathrm{c}} := D^{\mathrm{c}} \cap B(\boldsymbol{x}_0, r_0)$. We extend v by zero over $D_{r_0}^{\mathrm{c}}$, denote the extension in question by $\widetilde{v}_{r_0}$ and set $\widetilde{D}_{r_0} := \mathrm{int}(\overline{D} \cup \overline{D}_{r_0}^{\mathrm{c}})$. We have $\widetilde{v}_{r_0} \in H_0^1(\widetilde{D}_{r_0})$, $(\nabla \widetilde{v}_{r_0})_{|D} \in \boldsymbol{H}(\mathrm{div}; D)$, and $(\nabla \widetilde{v}_{r_0})_{|D_{r_0}^{\mathrm{c}}} \in \boldsymbol{H}(\mathrm{div}; D_{r_0}^{\mathrm{c}})$. Since the Robin boundary condition implies that $(\partial_n v)_{|\partial D_{\mathrm{r}}} = 0$, we infer that the normal component of $\nabla \widetilde{v}_{r_0}$ is continuous across $\partial D_{\mathrm{r}} \cap B(\boldsymbol{x}_0, r_0)$. Reasoning as in the proof of Theorem 18.10, we conclude that $\nabla \widetilde{v}_{r_0}$ is a member of $\boldsymbol{H}(\mathrm{div}; \widetilde{D}_{r_0})$, i.e., $\Delta \widetilde{v}_{r_0} \in L^2(\widetilde{D}_{r_0})$. In conclusion, we have $\widetilde{v}_{r_0} \in H_0^1(\widetilde{D}_{r_0})$, $-\Delta \widetilde{v}_{r_0} = \kappa^2 \widetilde{v}_{r_0}$ in $\widetilde{D}_{r_0}$, and $\widetilde{v}_{r_0|D_{r_0}^{\mathrm{c}}} = 0$. The unique continuation principle (Theorem 31.4) implies that $\widetilde{v}_{r_0} = 0$. Hence, $v = 0$. $\qquad\square$

Following Ihlenburg and Babuška [251], we now set $D := (0, \ell_D)$ and investigate the one-dimensional version of the problem (35.13). A homogeneous Dirichlet boundary condition is enforced at $\{x = 0\}$, and a homogeneous Robin condition is enforced at $\{x = \ell_D\}$. The space V becomes $V := \{v \in H^1(D) \mid v(0) = 0\}$.

Theorem 35.11 (BNB, mixed BCs, 1D). *Let $D := (0, \ell_D)$. Let the space V be equipped with the norm $\|v\|_V := \|\partial_x v\|_{L^2(D)}$. There are two constants $0 < c_\flat \leq c_\sharp$, both uniform with respect to κ, such that*

$$\frac{c_\flat}{1 + \kappa \ell_D} \leq \inf_{v \in V} \sup_{w \in V} \frac{|a(v, w)|}{\|v\|_V \|w\|_V} \leq \sup_{v \in V} \sup_{w \in V} \frac{|a(v, w)|}{\|v\|_V \|w\|_V} \leq \frac{c_\sharp}{1 + \kappa \ell_D}.$$

Proof. (1) Let us start with the lower bound. Let $v \in V$, $v \neq 0$, and let $z \in V$ solve $a(w, z) = (w, \kappa^2 v)_{L^2(D)}$ for all $w \in V$. It is shown in Exercise 35.1 that this problem has a unique solution in V, and it is shown in Exercise 35.2 that $\|z\|_V \leq 4\kappa \ell_D \|v\|_V$. Then we have

$$|a(v, v + z)| \geq \Re(a(v, v + z)) = \Re(a(v, v)) + \kappa^2 \|v\|_{L^2(D)}^2$$

$$= \|v'\|_{L^2(D)}^2 = \|v\|_V^2 = \frac{1}{4\kappa \ell_D + 1} \|v\|_V (\|v\|_V + 4\kappa \ell_D \|v\|_V)$$

$$\geq \frac{1}{4\kappa \ell_D + 1} \|v\|_V (\|v\|_V + \|z\|_V) \geq \frac{1}{4\kappa \ell_D + 1} \|v\|_V \|v + z\|_V.$$

This shows that the lower bound holds true.

(2) Let us now prove the upper bound. Let $v \in V$.

(2.a) If $\kappa \ell_D \leq 2$, then we can invoke the following Poincaré–Steklov inequality in V: there is a constant $\tilde{C}_{\mathrm{PS}} > 0$ s.t. $\tilde{C}_{\mathrm{PS}}(\ell_D^{-1} \|v\|_{L^2(D)} + \ell_D^{-\frac{1}{2}} |v(\ell_D)|) \leq \|v\|_V$ (see the proof of Proposition 31.21). Using the Cauchy–Schwarz inequality in (35.3) implies that

$$|a(v, w)| \leq \|v\|_V \|w\|_V + \kappa^2 \|v\|_{L^2(D)} \|w\|_{L^2(D)} + \kappa |v(\ell_D)| |w(\ell_D)|$$

$$\leq \max(1, \tilde{C}_{\mathrm{PS}}^{-2})(1 + \kappa \ell_D + (\kappa \ell_D)^2) \|v\|_V \|w\|_V.$$

Since we assumed $\kappa \ell_D \leq 2$, this leads to the bound $|a(v, w)| \leq c(1 + \kappa \ell_D)^{-1} \|v\|_V \|w\|_V$ with $c := \max(1, \tilde{C}_{\mathrm{PS}}^{-2}) \max_{t \in [0,2]} (1 + t + t^2)(1 + t)$.

(2.b) Let us now assume that $\kappa \ell_D \geq 2$. Let φ be a smooth nonnegative function equal to 1 on

$[0, \frac{1}{2}\ell_D]$ and such that $\varphi(\ell_D) = \partial_x\varphi(\ell_D) = 0$. Let us set $w(x) := \varphi(x)\sin(\kappa x)/\kappa$ so that $w \in V$, $w(0) = 0$, $w(\ell_D) = 0$, and $\partial_x w(\ell_D) = 0$. Let us set $\eta(x) := \partial_x w(x) - \partial_x w(0) + \kappa^2 \int_0^x w(s)\,\mathrm{d}s$, and $c_\varphi := \max(2\ell_D\|\partial_x\varphi\|_{L^\infty(D)}, \ell_D^2\|\partial_{xx}\varphi\|_{L^\infty(D)})$. Since w is real-valued and vanishes at $x = \ell_D$ and $v(0) = 0$, we have

$$a(v, w) = \int_0^{\ell_D} \partial_x v \partial_x w\,\mathrm{d}x - \kappa^2 \int_0^{\ell_D} vw\,\mathrm{d}x$$
$$= \int_0^{\ell_D} (\partial_x v)\eta\,\mathrm{d}x + v(\ell_D)\partial_x w(0) - \kappa^2\int_0^{\ell_D}\left(vw + \partial_x v \int_0^x w(s)\,\mathrm{d}s\right)\mathrm{d}x.$$

The last term is equal to $-\kappa^2 v(\ell_D)\int_0^{\ell_D} w(s)\,\mathrm{d}s$ since $v(0) = 0$. Since $\eta(\ell_D) = -\partial_x w(0) + \kappa^2\int_0^{\ell_D} w(s)\,\mathrm{d}s$ and $|v(\ell_D)| \le \ell_D^{\frac{1}{2}}\|v\|_V$, we infer that

$$|a(v, w)| = \int_0^{\ell_D}(\partial_x v)\eta\,\mathrm{d}x - v(\ell_D)\eta(\ell_D)$$
$$\le \|v\|_V(\|\eta\|_{L^2(D)} + \ell_D^{\frac{1}{2}}|\eta(\ell_D)|) \le 2\ell_D^{\frac{1}{2}}\|v\|_V\|\eta\|_{L^\infty(D)}.$$

Since $\eta(0) = 0$, we have $\|\eta\|_{L^\infty(D)} \le \ell_D\|\partial_x\eta\|_{L^\infty(D)}$. After observing that

$$\partial_x\eta(x) = \partial_{xx}\varphi(x)\sin(\kappa x)/\kappa + 2\partial_x\varphi(x)\cos(\kappa x)$$

and recalling the above bounds on the derivatives of φ, we deduce that $\|\eta\|_{L^\infty(D)} \le c_\varphi(1 + (\kappa\ell_D)^{-1})$. Hence, we have $|a(v, w)| \le 2c_\varphi(1 + (\kappa\ell_D)^{-1})\ell_D^{\frac{1}{2}}\|v\|_V$. After observing that

$$\|w\|_V^2 \ge \int_0^{\frac{1}{2}\ell_D}\cos(\kappa x)^2\,\mathrm{d}x \ge \frac{\ell_D}{4} - \frac{1}{4\kappa} \ge \frac{\ell_D}{8},$$

since $\kappa\ell_D \ge 2$, we conclude that $\|w\|_V \ge (\frac{1}{8}\ell_D)^{\frac{1}{2}}$. Hence, $|a(v, w)| \le c(1 + \kappa\ell_D)^{-1}\|v\|_V\|w\|_V$, and the proof is complete. $\qquad\square$

Remark 35.12 (Literature). Theorem 35.11 has been derived in Ihlenburg and Babuška [251, Thm. 1], and we refer the reader to this work for an exhaustive analysis of the continuous problem in one dimension with $g := 0$. Two- and three-dimensional versions of Lemma 35.7 for mixed boundary conditions are established in Hetmaniuk [243]. $\qquad\square$

35.3   Dirichlet boundary conditions

We consider in this section the Helmholtz problem with Dirichlet boundary conditions: For $f \in L^2(D;\mathbb{R})$ and $\kappa \in \mathbb{R}$, find u such that

$$-\Delta u - \kappa^2 u = f \quad \text{in } D, \qquad u = 0 \quad \text{on } \partial D. \tag{35.16}$$

As before, we assume that $\kappa > 0$ for simplicity. Note that the solution is now real-valued. We reformulate the above problem as follows:

$$\begin{cases} \text{Find } u \in V := H_0^1(D) \text{ such that} \\ a(u, w) = \ell(w), \quad \forall w \in V, \end{cases} \tag{35.17}$$

with the bilinear form

$$a(v,w) := \int_D (\nabla v \cdot \nabla w - \kappa^2 vw)\, \mathrm{d}x, \tag{35.18}$$

and the linear form $\ell(v) := \int_D fv\, \mathrm{d}x$. As above, we are going to rely on the BNB theorem to establish the well-posedness (35.17) since a is not coercive. But contrary to the case with Robin or mixed boundary conditions, the enforcement of Dirichlet conditions leads to a conditional stability depending on the value of κ. In other words, resonance phenomena can occur if κ takes values in some discrete subset of $\mathbb{R}_+$ associated with the spectrum of the Laplacian operator in D with Dirichlet conditions.

Since the embedding $H_0^1(D) \hookrightarrow L^2(D)$ is compact and the operator $(-\Delta)^{-1} : L^2(D) \to L^2(D)$ is self-adjoint, there exists a Hilbertian basis of $L^2(D)$ composed of eigenvectors of the Laplace operator (see Theorem 46.21). Let $(\psi_l)_{l \in \mathbb{N}}$ be the basis in question and let $(\lambda_l)_{l \in \mathbb{N}}$ be the corresponding eigenvalues with the normalization $\|\psi_l\|_{L^2(D)} = 1$. Then every function $v \in H_0^1(D)$ admits a unique expansion $v := \sum_{l \in \mathbb{N}} v_l \psi_l$ with $\|\nabla v\|_{\boldsymbol{L}^2(D)}^2 = \sum_{l \in \mathbb{N}} \lambda_l v_l^2$, $\|v\|_{L^2(D)}^2 = \sum_{l \in \mathbb{N}} v_l^2$. Notice that $a(v,w) = \sum_{l \in \mathbb{N}} (\lambda_l - \kappa^2) v_l w_l$ for all $v = \sum_{l \in \mathbb{N}} v_l \psi_l$, $w = \sum_{l \in \mathbb{N}} w_l \psi_l$ in $H_0^1(D)$. Let us denote by $l(\kappa)$ the largest integer such that $\lambda_{l(\kappa)} < \kappa^2$ with the convention that $l(\kappa) = -1$ if $\kappa^2 \leq \lambda_0$. The well-posedness of the problem (35.17) follows from the following result.

Theorem 35.13 (BNB, Dirichlet BCs). *Let $V := H_0^1(D)$ be equipped with the norm $\|v\|_V := \|\nabla v\|_{\boldsymbol{L}^2(D)}$. Assume that $\kappa^2 \notin \{\lambda_l\}_{l \in \mathbb{N}}$. Then the bilinear form a satisfies the conditions of the BNB theorem with the constant $\alpha(\kappa) := \min_{l \in \mathbb{N}} |\lambda_l - \kappa^2|/\lambda_l > 0$.*

Proof. Let $v \in H_0^1(D)$ with $v := \sum_{l \in \mathbb{N}} v_l \psi_l$. Let us set $w := \sum_{l \leq l(\kappa)} -v_l \psi_l + \sum_{l(\kappa) < l} v_l \psi_l$ with the convention that $l \in \mathbb{N}$ in the sums. Then we have

$$a(v,w) = \sum_{l \leq l(\kappa)} (\kappa^2 - \lambda_l) v_l^2 + \sum_{l(\kappa) < l} (\lambda_l - \kappa^2) v_l^2 \geq \alpha(\kappa) \sum_{l \in \mathbb{N}} \lambda_l v_l^2 = \alpha(\kappa) \|v\|_V^2.$$

The assertion follows readily from $\|w\|_V = \|v\|_V$. The reader is referred to Ciarlet [120, §3.1] for more details on this problem. $\square$

In general, $\alpha(\kappa)$ behaves like $\alpha_0 \gamma(\kappa)(\kappa \ell_D)^{-1}$, where $\gamma(\kappa) \in (0,1]$ and α_0 only depends on D. For $D := (0, \ell_D)$, the eigenvalues of the Laplace operator are $\lambda_l := \pi l^2 \ell_D^{-2}$. Let $\beta \in (0,1)$ and $L \in \mathbb{N} \setminus \{0\}$ be s.t. $\kappa^2 := \pi(L+\beta)^2 \ell_D^{-2}$. Then $\alpha(\kappa) = \min(\beta(2L+\beta)/L^2, (1+\beta)(2L+1+\beta)/(L+1)^2)$, and the claim follows readily. Notice that $\gamma(\kappa)$ becomes arbitrarily small as κ approaches an eigenvalue of the Laplace operator, i.e., if β is close to 0.

35.4 H^1-conforming approximation

We now formulate an H^1-conforming approximation of the Helmholtz problem with one of the boundary conditions discussed in the previous sections (Robin, mixed or Dirichlet). At this stage, we do not specify the norm with which we equip the space V: we just assume that it is an H^1-like norm that can contain some lower-order terms depending on κ (see Example 35.18).

Let $(\mathcal{T}_h)_{h \in \mathcal{H}}$ be a shape-regular mesh sequence so that each mesh covers D exactly. In the case of mixed boundary conditions, we also assume that the meshes are compatible with the corresponding partition of the boundary ∂D. Let $k \geq 1$ be the degree of the underlying finite element. Let $P_k^{\mathrm{g}}(\mathcal{T}_h)$ be the H^1-conforming finite element space considered in §18.2.3 and §32.1. For the Robin problem, we set $V_h := P_k^{\mathrm{g}}(\mathcal{T}_h)$, and for the mixed and the Dirichlet problems we set

$$V_h := \{v_h \in P_k^{\mathrm{g}}(\mathcal{T}_h) \mid v_{h|\partial D_{\mathrm{d}}} = 0\}. \tag{35.19}$$

We construct an approximation of the Helmholtz problem as follows:

$$\begin{cases} \text{Find } u_h \in V_h \text{ such that} \\ a(u_h, w_h) = \ell(w_h), \quad \forall w_h \in V_h. \end{cases} \tag{35.20}$$

A first way to investigate the stability of the discrete problem (35.20) consists of reasoning by perturbation using the fact that the continuous problem is well-posed. Such a result can be obtained by invoking a variation of Fortin's lemma (a more abstract version of this variation is discussed in Exercise 35.3). Recall that the elliptic projection $\Pi_h^{\mathrm{E}} : V \to V_h$ is defined for all $v \in V$ s.t. $(\nabla(v - \Pi_h^{\mathrm{E}}(v)), \nabla w_h)_{\boldsymbol{L}^2(D)} = 0$ for all $w_h \in V_h$ (see §32.4).

Lemma 35.14 (Modified Fortin). *Assume that there are positive real numbers* γ_{stb}, c_{app}, s *such that the elliptic projection satisfies for all* $v \in V$,

$$\gamma_{\mathrm{stb}} \|\Pi_h^{\mathrm{E}}(v)\|_V \le \|v\|_V, \qquad \|v - \Pi_h^{\mathrm{E}}(v)\|_{L^2(D)} \le c_{\mathrm{app}} h^s \ell_D^{1-s} \|v\|_V. \tag{35.21}$$

Let α *be the inf-sup constant of* a *on* $V \times V$. *Let* $\iota_{L,V} > 0$ *be such that*

$$\|v\|_{L^2(D)} \le \iota_{L,V} \ell_D \|v\|_V. \tag{35.22}$$

Assume that $h \in \mathcal{H} \cap (0, \ell_0(\kappa)]$ *with* $\ell_0(\kappa) := (\frac{1}{2} c_{\mathrm{app}}^{-1} \iota_{L,V}^{-1} \alpha \ell_D^{s-2} \kappa^{-2})^{\frac{1}{s}}$. *Then the restriction of* a *to* $V_h \times V_h$ *satisfies the following inf-sup condition:*

$$\inf_{v_h \in V_h} \sup_{w_h \in W_h} \frac{|a(v_h, w_h)|}{\|v_h\|_V \|w_h\|_V} \ge \alpha_0 := \frac{1}{2} \gamma_{\mathrm{stb}} \alpha > 0. \tag{35.23}$$

Proof. Using that $\Pi_h^{\mathrm{E}}(V) \subset V_h$ and the assumptions on Π_h^{E}, we have

$$\begin{aligned} \gamma_{\mathrm{stb}}^{-1} \sup_{w_h \in V_h} \frac{|a(v_h, w_h)|}{\|w_h\|_V} &\ge \gamma_{\mathrm{stb}}^{-1} \sup_{w \in V} \frac{|a(v_h, \Pi_h^{\mathrm{E}}(w))|}{\|\Pi_h^{\mathrm{E}}(w)\|_V} \ge \sup_{w \in V} \frac{|a(v_h, \Pi_h^{\mathrm{E}}(w))|}{\|w\|_V} \\ &\ge \sup_{w \in V} \frac{|a(v_h, w) + \kappa^2 (v_h, w - \Pi_h^{\mathrm{E}}(w))_{L^2(D)}|}{\|w\|_V} \\ &\ge \sup_{w \in V} \frac{|a(v_h, w)|}{\|w\|_V} - c_{\mathrm{app}} \iota_{L,V} h^s \ell_D^{2-s} \kappa^2 \|v_h\|_V \ge (\alpha - c_{\mathrm{app}} \iota_{L,V} h^s \ell_D^{2-s} \kappa^2) \|v_h\|_V. \end{aligned}$$

Since $h \le \ell_0(\kappa)$, using the definition of $\ell_0(\kappa)$ yields $\gamma_{\mathrm{stb}}^{-1} \sup_{w_h \in V_h} \frac{|a(v_h, w_h)|}{\|w_h\|_V} \ge \frac{1}{2} \alpha \|v_h\|_V$, i.e., (35.23) holds true with $\alpha_0 := \frac{1}{2} \gamma_{\mathrm{stb}} \alpha$. $\qquad\square$

The above result can be applied with $s := 1$ when full elliptic regularity is available. One always has $s > \frac{1}{2}$ in polyhedra (see Theorem 31.31).

Remark 35.15 (Duality argument). A duality argument is implicitly present in the assumptions of Lemma 35.14 since duality has to be invoked to establish the approximation property $\|v - \Pi_h^{\mathrm{E}}(v)\|_{L^2(D)} \le c_{\mathrm{app}} h^s \ell_D^{1-s} \|v\|_V$ (see Theorem 32.15). $\qquad\square$

A second way to investigate the stability of the discrete problem (35.20) is a technique introduced by Schatz [343] based on the Aubin–Nitsche duality argument.

Lemma 35.16 (Schatz). *Let* V, W *be two Banach spaces,* W *being reflexive. Let* a *be a bounded sesquilinear form on* $V \times W$ *satisfying the conditions of the BNB theorem with inf-sup and boundedness constants* $0 < \alpha \le \|a\|$. *Let* L *be a Hilbert space such that* $\|v\|_L \le \iota_{L,V} \|v\|_V$ *for all* $v \in V$ *(i.e.,* $V \hookrightarrow L$). *Let* $(V_h)_{h \in \mathcal{H}}$, $(W_h)_{h \in \mathcal{H}}$ *be sequences of finite-dimensional subspaces equipped, respectively, with the norm of* V *and the norm of* W. *Assume the following:*

(i) *(Gårding's inequality) There are $c_V > 0$, $c_L \geq 0$ s.t. $c_V\|v_h\|_V - c_L\|v_h\|_L \leq \sup_{w_h \in W_h} \frac{|a(v_h, w_h)|}{\|w_h\|_W}$ for all $v_h \in V_h$.*

(ii) *(Duality argument) There is a subspace $W_s \hookrightarrow W$ and real numbers c_{smo}, c_{app}, and $s \in (0, 1]$ s.t. $\inf_{w_h \in W_h} \|z - w_h\|_W \leq c_{\mathrm{app}} h^s \|z\|_{W_s}$ for all $z \in W_s$ and all $h \in \mathcal{H}$. Moreover, for all $g \in L$, the unique solution $z \in W$ to the adjoint problem $a(v, z) = (v, g)_L$ for all $v \in V$, satisfies $\|z\|_{W_s} \leq c_{\mathrm{smo}} \|g\|_L$.*

Assume that $h \in \mathcal{H} \cap (0, \ell_0(\kappa)]$ with $\ell_0(\kappa) := (\frac{1}{2}c_V c_L^{-1}\|a\|^{-1}c_{\mathrm{app}}^{-1}c_{\mathrm{smo}}^{-1})^{\frac{1}{s}}$. Then the restriction of a to $V_h \times W_h$ satisfies the discrete inf-sup condition (35.23) with $\alpha_0 \geq \frac{c_V}{2(\|a\| + c_L \iota_{L,V} + \frac{1}{2}c_V)}\alpha$.

Proof. Let $v_h \neq 0$ be a member of V_h. Consider the antilinear form $\ell_h \in (W_h)'$ defined by $\ell_h(w_h) := a(v_h, w_h)$ for all $w_h \in W_h$. (Note that $\ell_h := A_h(v_h)$ with $A_h \in \mathcal{L}(V_h; W_h')$ s.t. $\langle A_h(y_h), w_h \rangle_{W_h', W_h} := a(y_h, w_h)$ for all $(y_h, w_h) \in V_h \times W_h$.) Owing to the Hahn–Banach theorem (Theorem C.13), we can extend ℓ_h to W. Let $\widetilde{\ell_h}$ be the extension in question with $\|\widetilde{\ell_h}\|_{W'} = \|\ell_h\|_{W_h'}$. Since a satisfies the conditions of the BNB theorem, there exists $u \in V$ such that $a(u, w) := \widetilde{\ell_h}(w)$ for all $w \in W$. (Notice that $u := A^{-1}(\widetilde{\ell_h})$ with $A \in \mathcal{L}(V; W')$ s.t. $\langle A(y), w \rangle_{W', W} := a(y, w)$ for all $(y, w) \in V \times W$.) Using the inf-sup condition satisfied by a on $V \times W$, we infer that

$$\sup_{w_h \in W_h} \frac{|a(v_h, w_h)|}{\|w_h\|_W} = \sup_{w_h \in W_h} \frac{|\ell_h(w_h)|}{\|w_h\|_W} = \|\ell_h\|_{W_h'} = \|\widetilde{\ell_h}\|_{W'} = \sup_{w \in W} \frac{|a(u, w)|}{\|w\|_W} \geq \alpha\|u\|_V.$$

The rest of the proof consists of showing that there is c s.t. $\|u\|_V \geq c\|v_h\|_V$ for all $h \in \mathcal{H}$. Invoking Gårding's inequality on V_h gives

$$c_V\|v_h\|_V - c_L\|v_h\|_L \leq \sup_{w_h \in W_h} \frac{|a(v_h, w_h)|}{\|w_h\|_W} = \sup_{w_h \in W_h} \frac{|a(u, w_h)|}{\|w_h\|_W} \leq \|a\|\|u\|_V,$$

where we used that $a(u - v_h, w_h) = 0$ for all $w_h \in W_h$ (Galerkin orthogonality property) and the boundedness of the sesquilinear form a on $V \times W$. Since $\|v\|_L \leq \iota_{L,V}\|v\|_V$ for all $v \in V$, we infer that

$$c_V\|v_h\|_V \leq c_L\|v_h - u\|_L + (c_L\iota_{L,V} + \|a\|)\|u\|_V.$$

We now establish an upper bound on $\|v_h - u\|_L$. Let $z \in W$ solve $a(v, z) = (v, u - v_h)_L$ for all v in V. The Galerkin orthogonality property implies that $\|u - v_h\|_L^2 = a(u - v_h, z) = a(u - v_h, z - z_h)$ for all $z_h \in W_h$. Hence, we have

$$\|u - v_h\|_L^2 \leq \|a\|\|u - v_h\|_V c_a h^s \|z\|_{W_s} \leq \|a\|\|u - v_h\|_V c_{\mathrm{app}} c_{\mathrm{smo}} h^s \|u - v_h\|_L,$$

so that $\|u - v_h\|_L \leq \|a\| c_{\mathrm{app}} c_{\mathrm{smo}} h^s \|u - v_h\|_V$. This in turn implies that

$$\begin{aligned}
c_V\|v_h\|_V &\leq c_L\|v_h - u\|_L + (c_L\iota_{L,V} + \|a\|)\|u\|_V \\
&\leq c_L\|a\| c_{\mathrm{app}} c_{\mathrm{smo}} h^s \|u - v_h\|_V + (c_L\iota_{L,V} + \|a\|)\|u\|_V.
\end{aligned}$$

Using the triangle inequality gives

$$(c_V - c_L\|a\| c_{\mathrm{app}} c_{\mathrm{smo}} h^s)\|v_h\|_V \leq \left(\|a\| + c_L\iota_{L,V} + c_L\|a\| c_{\mathrm{app}} c_{\mathrm{smo}} h^s\right)\|u\|_V.$$

Provided $h \leq \ell_0(\kappa)$ we obtain $c_L\|a\| c_{\mathrm{app}} c_{\mathrm{smo}} h^s \leq \frac{1}{2}c_V$, so that

$$\frac{c_V}{2(\|a\| + c_L\iota_{L,V} + \frac{1}{2}c_V)}\|v_h\|_V \leq \|u\|_V.$$

This concludes the proof. $\qquad\square$

Both Lemma 35.14 and Lemma 35.16 imply that there is $\ell_0(\kappa)$ such that, if $h \in \mathcal{H} \cap (0, \ell_0(\kappa)]$, the discrete inf-sup condition (35.23) holds true with a constant that is uniform with respect to the meshsize but may depend on κ. To emphasize this dependency, let us write this constant as $\alpha_0(\kappa)$. We can now invoke Babuška's lemma (Lemma 26.14) to infer a quasi-optimal bound on the approximation error.

Corollary 35.17 (Error estimate). *There is $\ell_0(\kappa)$ s.t. the following quasi-optimal error estimate holds true for all $h \in \mathcal{H} \cap (0, \ell_0(\kappa)]$:*

$$\|u - u_h\|_V \leq \left(1 + \frac{\|a\|}{\alpha_0(\kappa)}\right) \inf_{v_h \in V_h} \|u - v_h\|_V. \tag{35.24}$$

Example 35.18 (Dependence on κ). In order to illustrate the above results, let us assume that we impose Robin boundary conditions with the norm $\|v\|_V := \|\nabla v\|_{\boldsymbol{L}^2(D)} + \kappa \|v\|_{L^2(D)}$. Let us also assume that full elliptic regularity holds true, i.e., the conclusion of Theorem 35.9 is fulfilled with $s := 1$. Then $\alpha(\kappa) \sim (\kappa \ell_D)^{-1}$ for all $\kappa \geq \kappa_0$. Moreover, we have $c_{\text{app}} \sim 1$, $s := 1$, $\iota_{L,V} \sim \kappa \ell_D$ in Lemma 35.14, so that $\ell_0(\kappa) \sim \ell_D^{-1} \kappa^{-2} \kappa \ell_D (\ell_D \kappa)^{-1} = \kappa^{-2} \ell_D^{-1}$, and $\alpha_0(\kappa) \sim (\kappa \ell_D)^{-1}$. The error estimate (35.24) gives $\|u - u_h\|_V \leq (1 + \kappa \ell_D) \inf_{v_h \in V_h} \|u - v_h\|_V$. Let us now use Lemma 35.16 with $\|a\| \sim 1$, $c_V := 1$, $c_L := \kappa$, $\iota_{L,V} := \kappa^{-1}$, $c_{\text{app}} \sim 1$, $s := 1$. In this case, it can be shown that $c_{\text{smo}} \sim \kappa \ell_D$. Then we have again $\ell_0(\kappa) \sim c_V c_L^{-1} \|a\|^{-1} c_{\text{a}}^{-1} c_{\text{smo}}^{-1} \sim \kappa^{-2} \ell_D^{-1}$ and $\alpha_0(\kappa) \sim (\kappa \ell_D)^{-1}$ leading to the same error estimate.   □

Remark 35.19 (Literature). The reader is referred to Ihlenburg and Babuška [251] for an exhaustive analysis of the one-dimensional Helmholtz problem with mixed boundary conditions and its Galerkin approximation in one dimension with $g := 0$. In particular, the following statements are proved therein: (i) For piecewise linear continuous finite elements on a uniform mesh, α_h scales exactly like $(\kappa \ell_D)^{-1}$ uniformly in $h \in \mathcal{H}$, i.e., the discrete problem is well-posed for all $h \in \mathcal{H}$ (see [251, Thm. 4]); (ii) The $\mathbb{P}_1$ Galerkin method delivers a quasi-optimal error estimate in the H^1-seminorm with a constant proportional to $\kappa \ell_D$ if $\kappa h < 1 < \kappa \ell_D$ (see [251, Cor. 2]).   □

Remark 35.20 (Dispersion error). It is shown in [251, Thm. 5] that $\|\nabla(u - u_h)\|_{\boldsymbol{L}^2(D)} \leq \ell_D (h\kappa/\pi)(1 + ch\kappa^2 \ell_D) \|f\|_{L^2}$, where c is independent of $h \in \mathcal{H}$ and $\kappa \geq 0$. The term proportional to $h\kappa^2 \ell_D$ is usually called *pollution error* or *dispersion error*. This term grows unboundedly when κ grows even if $h\kappa < 1$. The question whether the pollution error could be reduced or eliminated by using stabilization techniques (i.e., discontinuous approximation techniques or methods similar to those presented in Chapters 57–60) has been extensively addressed in the literature. We refer the reader to Burman et al. [102], Feng and Wu [200], Melenk and Sauter [300], Peterseim [325], and the literature therein for more details. For instance, it is shown in [300, Thm. 5.8] that under some appropriate assumptions the pollution effect can be suppressed if one assumes that $\kappa h/k$ is sufficiently small and that the polynomial degree k is at least $\mathcal{O}(\ln(\kappa))$. It is shown in [102, Thm. 6] that the pollution error disappears in one dimension for some specific κ-dependent choices of the penalty parameter of the CIP method (see §58.3 for details on CIP). The pollution error is also shown to disappear in [325, Thm. 6.2] for a localized Petrov-Galerkin method where the global shape functions each have a support of size rh with the oversampling condition $r \gtrsim \ln(\kappa \ell_D)$.   □

Exercises

Exercise 35.1 (1D Helmholtz, well-posedness). Let $D := (0, \ell_D)$, $\kappa > 0$, and consider the Helmholtz problem with mixed boundary conditions: $-\partial_{xx} u - \kappa^2 u = f$ in D, $u(0) = 0$, and

$\partial_x u(\ell_D) - i\kappa u(\ell_D) = 0$. (i) Give a weak formulation in $V := \{v \in H^1(D) \mid v(0) = 0\}$. (ii) Show by invoking an ODE argument that if the weak formulation has a solution, then it is unique. (iii) Show that the weak problem is well-posed. (*Hint*: use Lemma 35.3.)

Exercise 35.2 (Green's function, 1D). Let $G : D \times D \to \mathbb{C}$ be the function defined by

$$G(x, s) := \kappa^{-1} \begin{cases} \sin(\kappa x)e^{i\kappa s} & \text{if } x \in [0, s], \\ \sin(\kappa s)e^{i\kappa x} & \text{if } x \in [s, 1]. \end{cases}$$

(i) Prove that for all $x \in D$, the function $D \ni s \mapsto G(x, s) \in \mathbb{C}$ solves the PDE $-\partial_{ss}u - \kappa^2 u = \delta_{s=x}$ in D with the boundary conditions $u(0) = 0$ and $\partial_s u(\ell_D) - i\kappa u(\ell_D) = 0$ (i.e., G is the Green's function of the Helmholtz problem from Exercise 35.1). (ii) Find $H(x, s)$ s.t. $\partial_s H(x, s) = \partial_x G(x, s)$. (iii) Let $u(x) := \int_0^{\ell_D} G(x, s)f(s)\,\mathrm{d}s$. Prove that $\|u\|_{L^2(D)} \leq \kappa^{-1}\|f\|_{L^2(D)}$, $|u|_{H^1(D)} \leq \|f\|_{L^2(D)}$, and $|u|_{H^2(D)} \leq (\kappa + 1)\|f\|_{L^2(D)}$. (iv) Let $v \in L^2(D)$ and let $\widetilde{z}(x) := \kappa^2 \int_0^{\ell_D} G(x, s)v(s)\,\mathrm{d}s$. What is the PDE solved by $\widetilde{z}$? Same question for $z(x) := \kappa^2 \int_0^{\ell_D} \overline{G}(x, s)v(s)\,\mathrm{d}s$. *Note*: The function z is invoked in Step (1) of the proof of Theorem 35.11. (v) Assume now that $v \in H^1(D)$ with $v(0) = 0$, and let z and $\widetilde{z}$ be defined as above. Prove that $\max(|z|_{H^1(D)}, |\widetilde{z}|_{H^1(D)}) \leq 4\kappa\ell_D|v|_{H^1(D)}$. (*Hint*: see Ihlenburg and Babuška [251, p. 14] (up to the factor 4).)

Exercise 35.3 (Variation on Fortin's lemma). Let V, W be two Banach spaces and let a be a bounded sesquilinear form on $V \times W$ like in Fortin's Lemma 26.9. Let $(V_h)_{h \in \mathcal{H}}$, $(W_h)_{h \in \mathcal{H}}$ be sequences of subspaces of V and W equipped with the norm of V and W, respectively. Assume that there exists a map $\Pi_h : W \to W_h$ and constants $\gamma_{\Pi_h} > 0$, $c(h) > 0$ such that $|a(v_h, w - \Pi_h(w))| \leq c(h)\|v_h\|_V\|w\|_W$, $\gamma_{\Pi_h}\|\Pi_h(w)\|_W \leq \|w\|_W$ for all $v_h \in V_h$, all $w \in W$, and all $h \in \mathcal{H}$. Assume that $\lim_{h \to 0} c(h) = 0$. Prove that the discrete inf-sup condition (26.5a) holds true for $h \in \mathcal{H}$ small enough.

Exercise 35.4 (Lemma 35.8). (i) Prove that $\Re((\boldsymbol{m} \cdot \nabla v)\overline{v}) = \frac{1}{2}\boldsymbol{m} \cdot \nabla |v|^2$ for all $v \in H^1(D; \mathbb{C})$ and $\boldsymbol{m} \in \mathbb{R}^d$. (ii) Prove that $\Re(\boldsymbol{m} \cdot ((\nabla \boldsymbol{v})^\mathsf{T}\overline{\boldsymbol{v}})) = \frac{1}{2}\boldsymbol{m} \cdot \nabla\|\boldsymbol{v}\|_{\ell^2(\mathbb{C}^d)}^2$ for all $\boldsymbol{v} \in H^1(D; \mathbb{C}^d)$ and $\boldsymbol{m} \in \mathbb{R}^d$. (iii) Let $q \in H^2(D; \mathbb{C})$ and let $D^2 q$ denote the Hessian matrix of q, i.e., $(D^2 q)_{ij} = \partial_{x_i x_j}^2 q$ for all $i, j \in \{1{:}d\}$. Show that $\Re(\boldsymbol{m} \cdot ((D^2 q)\nabla \overline{q})) = \frac{1}{2}\boldsymbol{m} \cdot \nabla\|\nabla q\|_{\ell^2(\mathbb{C}^d)}^2$. (iv) Prove that (35.11) holds true for all $q \in \{v \in H^1(D; \mathbb{C}) \mid \Delta v \in L^2(D; \mathbb{C}), \ \nabla v \in L^2(\partial D; \mathbb{C}^d)\}$ and all $\boldsymbol{m} \in W^{1,\infty}(D; \mathbb{R}^d)$. (*Hint*: assume first that $q \in H^2(D; \mathbb{C})$.)

Chapter 36

Crouzeix–Raviart approximation

In Part VIII, composed of Chapters 36 to 41, we study various nonconforming approximations of an elliptic model problem. We first study the Poisson equation with a homogeneous Dirichlet condition and then address a diffusion PDE with contrasted coefficients. Nonconformity means that the discrete trial and test spaces are not subspaces of $H^1(D)$. Nonconformity has many sources. It may be that the discrete shape functions have nonzero jumps across the mesh interfaces. It may be that the Dirichlet conditions are enforced weakly. Another possible reason is that the approximation involves discrete unknowns associated with the mesh faces as in hybrid methods. All of these situations are studied in the following chapters. The objective of the present chapter is to study the nonconforming approximation of the Poisson equation by Crouzeix–Raviart finite elements. Another objective is to illustrate the abstract error analysis of Chapter 27.

36.1 Model problem

Let D be a Lipschitz domain in $\mathbb{R}^d$. We assume for simplicity that D is a polyhedron. We focus on the Poisson equation with homogeneous Dirichlet boundary conditions:

$$-\Delta u = f \quad \text{in } D, \qquad u = 0 \quad \text{on } \partial D, \tag{36.1}$$

with source term $f \in L^2(D)$. The weak formulation is as follows:

$$\begin{cases} \text{Find } u \in V := H_0^1(D) \text{ such that} \\ a(u,w) = \ell(w), \quad \forall w \in V, \end{cases} \tag{36.2}$$

with

$$a(v,w) := \int_D \nabla v \cdot \nabla w \, dx, \qquad \ell(w) := \int_D f w \, dx. \tag{36.3}$$

Owing to the Poincaré–Steklov inequality (see (3.11) with $p := 2$), there is $C_{\mathrm{PS}} > 0$ such that $C_{\mathrm{PS}} \|v\|_{L^2(D)} \leq \ell_D \|\nabla v\|_{\boldsymbol{L}^2(D)}$ for all $v \in V$, where ℓ_D is a length scale associated with D, e.g., $\ell_D := \mathrm{diam}(D)$. Hence, V equipped with the norm $\|v\|_V := \|\nabla v\|_{\boldsymbol{L}^2(D)} = |v|_{H^1(D)}$ is a Hilbert space, and the bilinear form a coincides with the inner product in V. Owing to the Lax–Milgram lemma, (36.2) is well-posed. We refer the reader to §41.2 for the more general PDE $-\nabla \cdot (\lambda \nabla u) = f$ with contrasted diffusivity λ.

36.2   Crouzeix–Raviart discretization

In this section, we recall Crouzeix–Raviart finite element, we define the corresponding approxima-
tion space, we formulate the discrete problem, and we establish its well-posedness. We also derive
some important stability estimates for Crouzeix–Raviart finite elements.

36.2.1   Crouzeix–Raviart finite elements

The Crouzeix–Raviart finite element is introduced in §7.5; see [151] for the original work to approx-
imate the Stokes equations. Let $\widehat{K}$ be the unit simplex in $\mathbb{R}^d$ with vertices $\{\widehat{z}_i\}_{i\in\{0:d\}}$. Let $\widehat{F}_i$ be
the face of $\widehat{K}$ opposite to $\widehat{z}_i$. The *Crouzeix–Raviart finite element* is defined by setting $\widehat{P} := \mathbb{P}_{1,d}$
and by using the following degrees of freedom (dofs) on $\widehat{P}$:

$$\widehat{\sigma}_i^{\mathrm{CR}}(\widehat{p}) := \frac{1}{|\widehat{F}_i|} \int_{\widehat{F}_i} \widehat{p}\,\mathrm{d}s, \qquad \forall i \in \{0{:}d\}. \tag{36.4}$$

Let $(\mathcal{T}_h)_{h\in\mathcal{H}}$ be a shape-regular matching mesh sequence composed of affine simplices so that
each mesh covers D exactly. Let $\mathcal{T}_h$ be a mesh and let K be a cell in $\mathcal{T}_h$. Using the Crouzeix–
Raviart element as reference finite element and letting the transformation ψ_K be the pullback by
the geometric mapping, i.e., $\psi_K(v) := v \circ \boldsymbol{T}_K$, Proposition 9.2 allows us to generate a Crouzeix–
Raviart finite element in K. We have $P_K := \psi_K^{-1}(\widehat{P}) = \mathbb{P}_{1,d} \circ \boldsymbol{T}_K^{-1} = \mathbb{P}_{1,d}$ since $\boldsymbol{T}_K$ is affine, and
the local dofs in K are for all $p \in P_K$,

$$\sigma_{K,i}^{\mathrm{CR}}(p) := \widehat{\sigma}_i^{\mathrm{CR}}(\psi_K(p)) = \frac{1}{|\widehat{F}_i|} \int_{\widehat{F}_i} p \circ \boldsymbol{T}_K\,\mathrm{d}\widehat{s} = \frac{1}{|F_{K,i}|} \int_{F_{K,i}} p\,\mathrm{d}s, \tag{36.5}$$

for all $i \in \{1{:}d\}$, where $\{F_{K,i} := \boldsymbol{T}_K(\widehat{F}_i)\}_{i\in\{0:d\}}$ are the faces of K. The local interpolation operator
$\mathcal{I}_K^{\mathrm{CR}} : V(K) := W^{1,1}(K) \to P_K$ is such that $\mathcal{I}_K^{\mathrm{CR}}(v) := \sum_{i\in\{0:d\}} \sigma_{K,i}^{\mathrm{CR}}(v)\theta_{K,i}^{\mathrm{CR}}$ for all $v \in V(K)$, where
$\{\theta_{K,i}\}_{i\in\{0:d\}}$ are the local shape functions in K s.t. $\sigma_{K,i}^{\mathrm{CR}}(\theta_{K,j}^{\mathrm{CR}}) = \delta_{ij}$ for all $i,j \in \{0{:}d\}$. Recall
that $\theta_i^{\mathrm{CR}} := 1 - d\lambda_i$, where $\{\lambda_i\}_{i\in\{0:d\}}$ are the barycentric coordinates in K.

Lemma 36.1 (Local interpolation). *There is c s.t. for all $r \in [0,1]$, all $p \in [1,\infty]$, all $v \in W^{1+r,p}(K)$, all $K \in \mathcal{T}_h$, and all $h \in \mathcal{H}$,*

$$\|v - \mathcal{I}_K^{\mathrm{CR}}(v)\|_{L^p(K)} + h_K|v - \mathcal{I}_K^{\mathrm{CR}}(v)|_{W^{1,p}(K)} \le c\,h_K^{1+r}|v|_{W^{1+r,p}(K)}. \tag{36.6}$$

Proof. Let $v \in W^{1+r,p}(K)$. The error estimates for $r \in \{0,1\}$ follow from Theorem 11.13 with
$k := 1$ and $l := 1$ since $V(K) := W^{1,1}(K)$. For $r \in (0,1)$, we use Corollary 12.13, the $W^{1,p}$-stability
of $\mathcal{I}_K^{\mathrm{CR}}$, and the fact that $P_K := \mathbb{P}_{1,d}$ is pointwise invariant under $\mathcal{I}_K^{\mathrm{CR}}$ to infer that

$$|v - \mathcal{I}_K^{\mathrm{CR}}(v))|_{W^{1,p}(K)} \le \inf_{p\in\mathbb{P}_{1,d}} |v - p - \mathcal{I}_K^{\mathrm{CR}}(v-p))|_{W^{1,p}(K)}$$

$$\le c \inf_{p\in\mathbb{P}_{1,d}} |v - p|_{W^{1,p}(K)} \le c'\,h_K^r|v|_{W^{1+r,p}(K)}.$$

The bound on $\|v - \mathcal{I}_K^{\mathrm{CR}}(v)\|_{L^p(K)}$ follows by proceeding similarly and using that $\|\mathcal{I}_K^{\mathrm{CR}}(v)\|_{L^p(K)} \le \|v\|_{L^p(K)} + ch_K|v|_{W^{1,p}(K)}$. $\qquad\qquad\square$

36.2.2   Crouzeix–Raviart finite element space

Consider the broken finite element space defined in (18.4) with $k := 1$,

$$P_1^{\mathrm{b}}(\mathcal{T}_h) := \{v_h \in L^\infty(D) \mid v_{h|K} \in \mathbb{P}_{1,d}, \ \forall K \in \mathcal{T}_h\}.$$

Recall that the set $\mathcal{F}_h^\circ$ is the collection of the interior faces (interfaces) in the mesh, and the faces are oriented by the unit normal vector $\boldsymbol{n}_F$ (see Chapter 10 on mesh orientation). For all $F \in \mathcal{F}_h^\circ$, there are two cells K_l, K_r s.t. $F := \partial K_l \cap \partial K_r$ and $\boldsymbol{n}_F$ points from K_l to K_r, i.e., $\boldsymbol{n}_F := \boldsymbol{n}_{K_l} = -\boldsymbol{n}_{K_r}$. The notion of jump across F is defined by setting $[\![v]\!]_F := v_{|K_l} - v_{|K_r}$. It is convenient to use a common notation for interfaces and boundary faces by writing $[\![v]\!]_F := v_{|K_l}$ for every boundary face $F := \partial K_l \cap \partial D \in \mathcal{F}_h^\partial$. The Crouzeix–Raviart finite element space is defined as

$$P_1^{\mathrm{CR}}(\mathcal{T}_h) := \{v_h \in P_1^{\mathrm{b}}(\mathcal{T}_h) \mid \int_F [\![v_h]\!]_F \, \mathrm{d}s = 0, \, \forall F \in \mathcal{F}_h^\circ\}. \tag{36.7}$$

The condition $\int_F [\![v_h]\!]_F \, \mathrm{d}s = 0$ is equivalent to the continuity of v_h at the barycenter $\boldsymbol{x}_F$ of F. Note that $P_1^{\mathrm{CR}}(\mathcal{T}_h)$ is not H^1-conforming since membership in $H^1(D)$ requires having zero-jumps pointwise (see Theorem 18.8).

Let $F \in \mathcal{F}_h$ be a mesh face. Let us denote by $\mathcal{T}_F := \{K \in \mathcal{T}_h \mid F \in \mathcal{F}_K\}$ the collection of the mesh cells having F as face ($\mathcal{T}_F$ contains two cells for $F \in \mathcal{F}_h^\circ$ and one cell for $F \in \mathcal{F}_h^\partial$). Let φ_F^{CR} be the function such that $\varphi_{F|K}^{\mathrm{CR}}$ is the local shape function in K associated with F if $K \in \mathcal{T}_F$ and $\varphi_{F|K}^{\mathrm{CR}} := 0$ otherwise; see Figure 36.1 for $d = 2$. Note that $\mathrm{supp}(\varphi_F^{\mathrm{CR}}) = D_F := \mathrm{int}(\bigcup_{K \in \mathcal{T}_F} K)$, i.e., D_F is the collection of all the points in the (one or two) mesh cells containing F. Let γ_F^{CR} be the linear form on $P_1^{\mathrm{CR}}(\mathcal{T}_h)$ such that $\gamma_F^{\mathrm{CR}}(v_h) := |F|^{-1} \int_F v_h \, \mathrm{d}s$ for all $v_h \in P_1^{\mathrm{CR}}(\mathcal{T}_h)$. Although v_h may be multivalued at F, the quantity $\gamma_F^{\mathrm{CR}}(v_h)$ is well defined since $\int_F [\![v_h]\!]_F \, \mathrm{d}s = 0$.
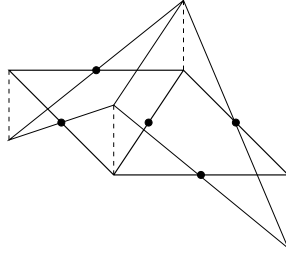
Figure 36.1: Global shape function for the Crouzeix–Raviart finite element. The support is materialized by thick lines and the graph by thin lines. Bullets indicate the barycenter of the edges.

Proposition 36.2 (Global dofs). $\{\varphi_F^{\mathrm{CR}}\}_{F \in \mathcal{F}_h}$ *is a basis of* $P_1^{\mathrm{CR}}(\mathcal{T}_h)$*, and* $\{\gamma_F^{\mathrm{CR}}\}_{F \in \mathcal{F}_h}$ *is a basis of* $\mathcal{L}(P_1^{\mathrm{CR}}(\mathcal{T}_h); \mathbb{R})$.

Proof. φ_F^{CR} is a member of $P_1^{\mathrm{CR}}(\mathcal{T}_h)$ since φ_F^{CR} is piecewise affine by construction and its mean value on a mesh face is 0 or 1. Consider now real numbers $\{\alpha_F\}_{F \in \mathcal{F}_h}$ s.t. the function $w := \sum_{F \in \mathcal{F}_h} \alpha_F \varphi_F^{\mathrm{CR}}$ vanishes identically. Observing that $\gamma_{F'}^{\mathrm{CR}}(\varphi_F^{\mathrm{CR}}) = \delta_{FF'}$ for all $F, F' \in \mathcal{F}_h$, where $\delta_{FF'}$ denotes the Kronecker symbol, we infer that $\alpha_{F'} = \gamma_{F'}^{\mathrm{CR}}(w) = 0$ for all $F' \in \mathcal{F}_h$. Hence, the functions $\{\varphi_F^{\mathrm{CR}}\}_{F \in \mathcal{F}_h}$ are linearly independent. Finally, let $v_h \in P_1^{\mathrm{CR}}(\mathcal{T}_h)$ and set $w_h := \sum_{F \in \mathcal{F}_h} \gamma_F^{\mathrm{CR}}(v_h) \varphi_F^{\mathrm{CR}}$. Then, $v_{h|K}$ and $w_{h|K}$ are in P_K for all $K \in \mathcal{T}_h$, and $\sigma_{K,i}(w_{h|K}) = \sigma_{K,i}(v_{h|K})$ for all $i \in \{0{:}d\}$. Unisolvence implies that $v_{h|K} = w_{h|K}$, so that $v_h = w_h$ since $K \in \mathcal{T}_h$ is arbitrary. This shows that $\{\varphi_F^{\mathrm{CR}}\}_{F \in \mathcal{F}_h}$ is a basis of $P_1^{\mathrm{CR}}(\mathcal{T}_h)$. By using similar arguments, it follows that $\{\gamma_F^{\mathrm{CR}}\}_{F \in \mathcal{F}_h}$ is a basis of $\mathcal{L}(P_1^{\mathrm{CR}}(\mathcal{T}_h); \mathbb{R})$. $\qquad\square$

Proposition 36.2 implies that the dimension of $P_1^{\mathrm{CR}}(\mathcal{T}_h)$ is equal to the number of faces (edges in dimension two) in the mesh. Moreover, the global Crouzeix–Raviart interpolation operator acts

on every function v in $W^{1,1}(D)$ as follows: For all $\boldsymbol{x} \in D$,

$$\mathcal{I}_h^{\mathrm{CR}}(v)(\boldsymbol{x}) := \sum_{F \in \mathcal{F}_h} \gamma_F^{\mathrm{CR}}(v) \varphi_F^{\mathrm{CR}}(\boldsymbol{x}) = \sum_{F \in \mathcal{F}_h} \left(\frac{1}{|F|} \int_F v \, \mathrm{d}s \right) \varphi_F^{\mathrm{CR}}(\boldsymbol{x}).$$

Since $\mathcal{I}_h^{\mathrm{CR}}(v)_{|K} = \mathcal{I}_K^{\mathrm{CR}}(v_{|K})$ for all $K \in \mathcal{T}_h$, the approximation results of Lemma 36.1 can be rephrased in terms of $\mathcal{I}_h^{\mathrm{CR}}$.

36.2.3 Discrete problem and well-posedness

We account for the homogeneous Dirichlet boundary condition by considering the following subspace of $P_1^{\mathrm{CR}}(\mathcal{T}_h)$:

$$P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h) := \left\{ v_h \in P_1^{\mathrm{CR}}(\mathcal{T}_h) \mid \int_F v_h \, \mathrm{d}s = 0, \, \forall F \in \mathcal{F}_h^{\partial} \right\}, \tag{36.8}$$

where $\mathcal{F}_h^{\partial}$ is the collection of the mesh faces located at the boundary. By proceeding as in Proposition 36.2, one can verify that $\{\varphi_F^{\mathrm{CR}}\}_{F \in \mathcal{F}_h^{\circ}}$ is a basis of $P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$, and $\{\gamma_F^{\mathrm{CR}}\}_{F \in \mathcal{F}_h^{\circ}}$ is a basis of $\mathcal{L}(P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h); \mathbb{R})$. The dimension of $P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$ is the number of internal faces (edges if $d = 2$) in the mesh.

The bilinear form a introduced in (36.3) is not well defined on $P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$ since this space is not H^1-conforming. Since functions in $P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$ are piecewise smooth, we can localize their gradient to the mesh cells. To this purpose, we introduce the notion of broken gradient on the broken Sobolev space $W^{1,p}(\mathcal{T}_h)$ with $p \in [1, \infty]$. Recall from Definition 18.1 that a function $v \in W^{1,p}(\mathcal{T}_h)$ is s.t. $\nabla(v_{|K}) \in \boldsymbol{L}^p(K)$ for all $K \in \mathcal{T}_h$.

Definition 36.3 (Broken gradient). *Let* $p \in [1, \infty]$. *The* broken gradient *operator* $\nabla_h : W^{1,p}(\mathcal{T}_h) \to \boldsymbol{L}^p(D)$ *is defined by setting* $(\nabla_h v)_{|K} := \nabla(v_{|K})$ *for all* $K \in \mathcal{T}_h$.

A crucial consequence of Lemma 18.9 is that $\nabla_h v = \nabla v$ whenever $v \in W^{1,p}(D)$. This property will be often used for the solution to the model problem (36.2) since $u \in H_0^1(D)$. We define the following discrete bilinear and linear forms on $V_h \times V_h$ and on V_h, respectively:

$$a_h(v_h, w_h) := \int_D \nabla_h v_h \cdot \nabla_h w_h \, \mathrm{d}x, \qquad \ell_h(w_h) := \int_D f w_h \, \mathrm{d}x, \tag{36.9}$$

and we consider the following discrete problem:

$$\begin{cases} \text{Find } u_h \in V_h := P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h) \text{ such that} \\ a_h(u_h, w_h) = \ell_h(w_h), \quad \forall w_h \in V_h. \end{cases} \tag{36.10}$$

Lemma 36.4 (Coercivity, well-posedness). (i) *The map*

$$v_h \mapsto \|v_h\|_{V_h} := a_h(v_h, v_h)^{\frac{1}{2}} = \|\nabla_h v_h\|_{\boldsymbol{L}^2(D)} \tag{36.11}$$

is a norm on $P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$. (ii) *Equipping* V_h *with this norm, the bilinear form* a_h *is coercive on* V_h *with* $\alpha_h := 1$. (iii) *The discrete problem* (36.10) *is well-posed.*

Proof. (i) The only nontrivial property is to prove that $\|v_h\|_{V_h} = 0$ implies that $v_h = 0$ for all $v_h \in V_h$. If $\|v_h\|_{V_h} = 0$, then v_h is piecewise constant. The additional property $\int_F [\![v_h]\!]_F \, \mathrm{d}s = 0$ for all $F \in \mathcal{F}_h^{\circ}$ implies that v_h is globally constant on D. That $v_h = 0$ follows from $\int_F v_h \, \mathrm{d}s = 0$ for all $F \in \mathcal{F}_h^{\partial}$.
(ii)-(iii) Since $\|\cdot\|_{V_h}$ is a norm on V_h, coercivity follows from the definition of $\|\cdot\|_{V_h}$, and well-posedness follows from the Lax–Milgram lemma. $\qquad \square$

Remark 36.5 (Nonsmooth right-hand side). We observe that it is not clear how one should account for a source term f in $H^{-1}(D)$ in (36.10), since it is not clear how f would act on (discrete) functions that are not in $H_0^1(D)$. One possibility is to consider the discrete linear form $\ell_h(w_h) := \langle f, \mathcal{J}_{h,0}^{\mathrm{av}}(w_h)\rangle_{H^{-1}(D),H_0^1(D)}$ where $\mathcal{J}_{h,0}^{\mathrm{av}} : P_1^{\mathrm{b}}(\mathcal{T}_h) \to P_{1,0}^{\mathrm{g}}(\mathcal{T}_h)$ is the averaging operator with boundary conditions introduced in §22.4.1. A general theory addressing this type of difficulty is developed in Veeser and Zanotti [373]. $\qquad\square$

36.2.4 Discrete Poincaré–Steklov inequality

On the H_0^1-conforming subspace $P_{1,0}^{\mathrm{g}}(\mathcal{T}_h) := P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h) \cap H_0^1(D)$, the norm $\|\cdot\|_{V_h}$ defined in (36.11) coincides with the H^1-seminorm. Owing to the Poincaré–Steklov inequality, we know that there is $C_{\mathrm{PS}} > 0$ s.t. $C_{\mathrm{PS}}\|v_h\|_{L^2(D)} \le \ell_D\|\nabla v_h\|_{\boldsymbol{L}^2(D)} = \ell_D\|v_h\|_{V_h}$ for all $v_h \in P_{1,0}^{\mathrm{g}}(\mathcal{T}_h)$. We now prove that a similar inequality is available on the larger space $P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$.

Lemma 36.6 (Discrete Poincaré–Steklov inequality). *There is $C_{\mathrm{PS}}^{\mathrm{CR}} > 0$ s.t. for all $v_h \in P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$ and all $h \in \mathcal{H}$,*

$$C_{\mathrm{PS}}^{\mathrm{CR}}\|v_h\|_{L^2(D)} \le \ell_D\|\nabla_h v_h\|_{\boldsymbol{L}^2(D)}. \tag{36.12}$$

Proof. Let $v_h \in P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$. Let $\phi \in H_0^1(D)$ solve $\Delta\phi = v_h$ and let $\boldsymbol{\sigma} := \nabla\phi$. Then $\nabla\cdot\boldsymbol{\sigma} = v_h$. Elliptic regularity implies that there is $s > \frac{1}{2}$ such that $\phi \in H^{1+s}(D)$ (see Theorem 31.33) so that $\boldsymbol{\sigma} \in \boldsymbol{H}^s(D)$. Moreover, there is $\gamma_D > 0$ such that $\gamma_D(\|\boldsymbol{\sigma}\|_{\boldsymbol{L}^2(D)} + \ell_D^s|\boldsymbol{\sigma}|_{\boldsymbol{H}^s(D)}) \le \ell_D\|v_h\|_{L^2(D)}$. Integrating by parts cellwise, we infer that

$$\|v_h\|_{L^2(D)}^2 = \int_D v_h \nabla\cdot\boldsymbol{\sigma}\,\mathrm{d}x = \sum_{K\in\mathcal{T}_h}\int_K v_{h|K}\nabla\cdot\boldsymbol{\sigma}\,\mathrm{d}x$$

$$= -\sum_{K\in\mathcal{T}_h}\int_K \boldsymbol{\sigma}\cdot\nabla(v_{h|K})\,\mathrm{d}x + \sum_{K\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_K}\int_F \boldsymbol{\sigma}\cdot\boldsymbol{n}_K v_{h|K}\,\mathrm{d}s$$

$$= -\int_D \boldsymbol{\sigma}\cdot\nabla_h v_h\,\mathrm{d}x + \sum_{K\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_K}\int_F \boldsymbol{\sigma}\cdot\boldsymbol{n}_K v_{h|K}\,\mathrm{d}s =: \mathfrak{T}_1 + \mathfrak{T}_2,$$

where $\mathcal{F}_K$ is the collection of the faces of K and $\boldsymbol{n}_K$ the outward unit normal to K (observe that $\boldsymbol{\sigma}$ is single-valued on F since $\boldsymbol{\sigma} \in \boldsymbol{H}^s(D)$ with $s > \frac{1}{2}$). The Cauchy–Schwarz inequality implies that

$$|\mathfrak{T}_1| \le \|\boldsymbol{\sigma}\|_{\boldsymbol{L}^2(D)}\|\nabla_h v_h\|_{\boldsymbol{L}^2(D)}.$$

Consider now $\mathfrak{T}_2$. If $F := \partial K_l \cap \partial K_r$ is an interface, the integral over F appears twice in the sum. Since $\int_F v_{h|K_l}\,\mathrm{d}s = \int_F v_{h|K_r}\,\mathrm{d}s$ by definition of $P_1^{\mathrm{CR}}(\mathcal{T}_h)$ and since $\boldsymbol{n}_{K_l} = -\boldsymbol{n}_{K_r}$, we can subtract from $\boldsymbol{\sigma}$ a constant function on F that we take equal to $\underline{\boldsymbol{\sigma}}_F := \frac{1}{|F|}\int_F \boldsymbol{\sigma}\,\mathrm{d}s$. The same conclusion is valid for the boundary faces since $\int_F v_h\,\mathrm{d}s = 0$ on such faces by definition of $P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$. This leads to

$$\mathfrak{T}_2 = \sum_{K\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_K}\int_F (\boldsymbol{\sigma} - \underline{\boldsymbol{\sigma}}_F)\cdot\boldsymbol{n}_K v_{h|K}\,\mathrm{d}s$$

$$= \sum_{K\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_K}\int_F (\boldsymbol{\sigma} - \underline{\boldsymbol{\sigma}}_F)\cdot\boldsymbol{n}_K (v_{h|K} - \underline{v}_F)\,\mathrm{d}s,$$

where the subtraction of the single-valued quantity $\underline{v}_F := \frac{1}{|F|}\int_F v_h\,\mathrm{d}s$ is justified as above. Applying Lemma 36.8 below to $\sigma_{|K}$ and to $v_{h|K}$, using $h_K \le \ell_D$ for all $K \in \mathcal{T}_h$, and invoking the

Cauchy–Schwarz inequality yields

$$|\mathfrak{T}_2| \leq c \sum_{K \in \mathcal{T}_h} h_K^{s-\frac{1}{2}} |\boldsymbol{\sigma}|_{\boldsymbol{H}^s(K)} h_K^{\frac{1}{2}} \|\nabla(v_{h|K})\|_{\boldsymbol{L}^2(K)}$$

$$\leq c\,\ell_D^s \sum_{K \in \mathcal{T}_h} |\boldsymbol{\sigma}|_{\boldsymbol{H}^s(K)} \|\nabla(v_{h|K})\|_{\boldsymbol{L}^2(K)} \leq c\,\ell_D^s |\boldsymbol{\sigma}|_{\boldsymbol{H}^s(D)} \|\nabla_h v_h\|_{\boldsymbol{L}^2(D)},$$

since $\sum_{K \in \mathcal{T}_h} |\boldsymbol{\sigma}|_{\boldsymbol{H}^s(K)}^2 \leq |\boldsymbol{\sigma}|_{\boldsymbol{H}^s(D)}^2$. Combining the above bounds on $\mathfrak{T}_1$ and $\mathfrak{T}_2$, we infer that

$$\|v_h\|_{L^2(D)}^2 \leq \left(\|\boldsymbol{\sigma}\|_{\boldsymbol{L}^2(D)} + c\,\ell_D^s |\boldsymbol{\sigma}|_{\boldsymbol{H}^s(D)}\right)\|\nabla_h v_h\|_{\boldsymbol{L}^2(D)},$$

and (36.12) follows from $\gamma_D(\|\boldsymbol{\sigma}\|_{\boldsymbol{L}^2(D)} + \ell_D^s |\boldsymbol{\sigma}|_{\boldsymbol{H}^s(D)}) \leq \ell_D \|v_h\|_{L^2(D)}$.   □

Remark 36.7 (Literature). The above proof is adapted from Temam [363, Prop. 4.13]; see also Croisille and Greff [150].   □

Lemma 36.8 (Poincaré–Steklov on faces). *Let $s \in (\frac{1}{2}, 1]$. There is c s.t.*

$$\|\psi - \underline{\psi}_F\|_{L^2(F)} \leq c\,h_K^{s-\frac{1}{2}} |\psi|_{H^s(K)}, \tag{36.13}$$

for all $\psi \in H^s(K)$ with $\underline{\psi}_F := \frac{1}{|F|} \int_F \psi\,\mathrm{d}s$; all $K \in \mathcal{T}_h$, all $F \in \mathcal{F}_K$, and all $h \in \mathcal{H}$ (the constant c grows unboundedly as $s \downarrow \frac{1}{2}$).

Proof. Let $\tilde{\psi} := \psi - \frac{1}{|K|} \int_K \psi\,\mathrm{d}x$. With obvious notation, we have $\psi - \underline{\psi}_F = \tilde{\psi} - \underline{\tilde{\psi}}_F$. The triangle inequality and the Cauchy–Schwarz inequality imply that $\|\psi - \underline{\psi}_F\|_{L^2(F)} \leq 2\|\tilde{\psi}\|_{L^2(F)}$. Using the trace inequality (12.17) yields

$$\|\psi - \underline{\psi}_F\|_{L^2(F)} \leq c(h_K^{-\frac{1}{2}} \|\tilde{\psi}\|_{L^2(K)} + h_K^{s-\frac{1}{2}} |\tilde{\psi}|_{H^s(K)}).$$

The expected bound follows from $|\tilde{\psi}|_{H^s(K)} = |\psi|_{H^s(K)}$ and the Poincaré–Steklov inequality ((12.13) if $s = 1$ or (12.14) if $s \in (\frac{1}{2}, 1)$) on K, which gives $\|\tilde{\psi}\|_{L^2(K)} \leq ch_K^s |\psi|_{H^s(K)}$.   □

36.2.5 Bound on the jumps

Bounding the jumps of functions in $P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$ is useful in many situations. The following result will be invoked in the next section.

Lemma 36.9 (Bound on the jumps). *There is c s.t. for all $v_h \in P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$ and all $h \in \mathcal{H}$,*

$$c^{-1} \sum_{F \in \mathcal{F}_h} h_F^{-1} \|[\![v_h]\!]\|_{L^2(F)}^2 \leq \inf_{v \in H_0^1(D)} \|\nabla_h(v - v_h)\|_{\boldsymbol{L}^2(D)}^2$$

$$\leq c \sum_{F \in \mathcal{F}_h} h_F^{-1} \|[\![v_h]\!]\|_{L^2(F)}^2. \tag{36.14}$$

Proof. Let $v_h \in P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$. For all $K \in \mathcal{T}_h$, let us set $H_*^1(K) := \{\phi \in H^1(K) \mid \int_K \phi\,\mathrm{d}x = 0\}$ and let $\mathcal{F}_K$ be the collection of the faces of K. For all $F \in \mathcal{F}_K$, let $\psi_{K,F} \in H_*^1(K)$ solve the local Neumann problem:

$$\int_K \nabla\psi_{K,F}\cdot\nabla\phi\,\mathrm{d}x = \epsilon_{K,F} \int_F [\![v_h]\!]_F \phi\,\mathrm{d}s, \quad \forall\phi \in H_*^1(K), \tag{36.15}$$

where $\epsilon_{K,F} := \boldsymbol{n}_K\cdot\boldsymbol{n}_F = \pm 1$. This problem is well-posed since $\int_F [\![v_h]\!]_F\, \mathrm{d}s = 0$ for all $F \in \mathcal{F}_h$. Since $\psi_{K,F} \in H^1_*(K)$, the multiplicative trace inequality (12.17) (with $s := 1$ and $p := 2$) together with the Poincaré–Steklov inequality (12.13) implies that $\|\psi_{K,F}\|_{L^2(F)} \leq c h_K^{\frac{1}{2}}\|\nabla\psi_{K,F}\|_{\boldsymbol{L}^2(K)}$. Taking $\phi := \psi_{K,F}$ as a test function in (36.15), we infer that

$$\|\nabla\psi_{K,F}\|^2_{\boldsymbol{L}^2(K)} = \epsilon_{K,F}\int_F [\![v_h]\!]_F\psi_{K,F}\, \mathrm{d}s \leq \|[\![v_h]\!]\|_{L^2(F)}\|\psi_{K,F}\|_{L^2(F)}$$

$$\leq c\, h_K^{\frac{1}{2}}\|[\![v_h]\!]_F\|_{L^2(F)}\|\nabla\psi_{K,F}\|_{\boldsymbol{L}^2(K)}.$$

Owing to the regularity of the mesh sequence, we infer that

$$\|\nabla\psi_{K,F}\|_{\boldsymbol{L}^2(K)} \leq c\, h_F^{\frac{1}{2}}\|[\![v_h]\!]_F\|_{L^2(F)}.$$

(1) Let us prove the first bound in (36.14). Let $v \in H^1_0(D)$. Let c_K be the mean value of the function $(v_h - v)$ over K. The restriction of $(v_h - v - c_K)$ to K is in $H^1_*(K)$. Let $F \in \mathcal{F}_h$. Taking $\phi_K := (v_h - v)_{|K} - c_K$ as a test function in (36.15) and summing over $K \in \mathcal{T}_F$, we infer that

$$\sum_{K\in\mathcal{T}_F}\int_K \nabla\psi_{K,F}\cdot\nabla(v_h - v)_{|K}\, \mathrm{d}x = \sum_{K\in\mathcal{T}_F}\int_K \nabla\psi_{K,F}\cdot\nabla\phi_K\, \mathrm{d}x$$

$$= \sum_{K\in\mathcal{T}_F}\epsilon_{K,F}\int_F [\![v_h]\!]_F\phi_K\, \mathrm{d}s = \sum_{K\in\mathcal{T}_F}\epsilon_{K,F}\int_F [\![v_h]\!]_F(v_{h|K} - v - c_K)\, \mathrm{d}s$$

$$= \int_F [\![v_h]\!]_F[\![v_h - v - c_K]\!]_F\, \mathrm{d}s = \int_F [\![v_h]\!]_F[\![v_h - v]\!]_F\, \mathrm{d}s = \int_F [\![v_h]\!]^2_F\, \mathrm{d}s,$$

where we used that $\int_F [\![v_h]\!]_F\, \mathrm{d}s = 0$ to eliminate c_K and the fact that $v \in H^1_0(D)$ to eliminate $[\![v]\!]_F$. Using the Cauchy–Schwarz inequality and the above bound on $\|\nabla\psi_{K,F}\|_{\boldsymbol{L}^2(K)}$, we obtain

$$h_F^{-1}\|[\![v_h]\!]_F\|^2_{L^2(F)} \leq c\sum_{K\in\mathcal{T}_F}\|\nabla(v - v_{h|K})\|^2_{\boldsymbol{L}^2(K)}. \tag{36.16}$$

Summing over $F \in \mathcal{F}_h$ leads to the first bound in (36.14).

(2) To prove the second bound in (36.14), we estimate the infimum over $v \in H^1_0(D)$ by taking $v := \mathcal{J}^{\mathrm{g,av}}_{h,0}(v_h)$ where $\mathcal{J}^{\mathrm{g,av}}_{h,0} : P^{\mathrm{b}}_1(\mathcal{T}_h) \to P^{\mathrm{g}}_{1,0}(\mathcal{T}_h) \subset H^1_0(D)$ is the averaging operator with zero trace introduced in §22.4.1. Then the second bound in (36.14) follows from Lemma 22.12 and the regularity of the mesh sequence. $\qquad\square$

The bound (36.14) can be adapted to the case where $v_h \in P^{\mathrm{CR}}_1(\mathcal{T}_h)$, i.e., without any boundary prescription. The summations over the mesh faces are then restricted to the mesh interfaces, and the infimum is taken over the functions v in $H^1(D)$. The idea of introducing the local Neumann problem (36.15) has been considered in Achdou et al. [4].

36.3 Error analysis

In this section, we first establish an error estimate by using the coercivity norm and the abstract error estimate from Lemma 27.5. Then we derive an improved L^2-error estimate by adapting the duality argument from §32.3.

36.3.1   Energy error estimate

We perform the error analysis under the assumption that the solution to the model problem (36.2) is in $H^{1+r}(D)$ with $r > \frac{1}{2}$, i.e., we set

$$V_{\mathrm{s}} := H^{1+r}(D) \cap H_0^1(D), \quad r > \frac{1}{2}. \tag{36.17}$$

The assumption $u \in V_{\mathrm{s}}$ is reasonable in the setting of the Poisson equation with Dirichlet conditions in a Lipschitz polyhedron since it is consistent with the elliptic regularity theory (see Theorem 31.33). The important property of a function $v \in V_{\mathrm{s}}$ that we use here is that its normal derivative $\boldsymbol{n}_K\cdot\nabla v$ is meaningful in $L^2(\partial K)$ for all $K \in \mathcal{T}_h$. Actually, the full trace of ∇v on ∂K is meaningful on $\boldsymbol{L}^2(\partial K)$, and this trace is single-valued on any interface $F \in \mathcal{F}_h^\circ$ (see Remark 18.4). Therefore, we have $[\![\nabla v]\!]_F = \boldsymbol{0}$ for all $v \in V_{\mathrm{s}}$ and all $F \in \mathcal{F}_h^\circ$.

The discrete space $V_h := P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$ is equipped with the norm $\|\cdot\|_{V_h}$ defined in (36.11), and we introduce the space $V_\sharp := V_{\mathrm{s}} + V_h$ equipped with the norm $\|\cdot\|_{V_\sharp}$ defined by

$$\|v\|_{V_\sharp}^2 := \sum_{K \in \mathcal{T}_h} \Big(\|\nabla v\|_{\boldsymbol{L}^2(K)}^2 + h_K \|\boldsymbol{n}_K\cdot\nabla v_{|K}\|_{L^2(\partial K)}^2 \Big). \tag{36.18}$$

A discrete trace inequality shows that there is $c_\sharp$ s.t. $\|v_h\|_{V_\sharp} \leq c_\sharp \|v_h\|_{V_h}$ for all $v_h \in V_h$ and all $h \in \mathcal{H}$, i.e., (27.5) holds true. Using the forms a_h and ℓ_h defined in (36.9), the consistency error is s.t.

$$\langle \delta_h(v_h), w_h \rangle_{V_h', V_h} := \ell_h(w_h) - a_h(v_h, w_h), \qquad \forall v_h, w_h \in V_h. \tag{36.19}$$

Lemma 36.10 (Consistency/boundedness). *Assume* (36.17). *There is* $\omega_\sharp$, *uniform w.r.t.* $u \in V_{\mathrm{s}}$, *s.t. for all* $v_h \in V_h$ *and all* $h \in \mathcal{H}$,

$$\|\delta_h(v_h)\|_{V_h'} \leq \omega_\sharp \|u - v_h\|_{V_\sharp}. \tag{36.20}$$

Proof. Let $v_h, w_h \in V_h$. Since the normal derivative $\boldsymbol{n}_K\cdot\nabla u$ is meaningful in $L^2(\partial K)$ for all $K \in \mathcal{T}_h$, we have

$$\ell_h(w_h) = \sum_{K \in \mathcal{T}_h} \int_K f w_{h|K}\, \mathrm{d}x = \sum_{K \in \mathcal{T}_h} \int_K -(\Delta u) w_{h|K}\, \mathrm{d}x$$

$$= \sum_{K \in \mathcal{T}_h} \int_K \nabla u\cdot\nabla w_{h|K}\, \mathrm{d}x - \sum_{K \in \mathcal{T}_h} \sum_{F \in \mathcal{F}_K} \int_F (\boldsymbol{n}_K\cdot\nabla u) w_{h|K}\, \mathrm{d}s$$

$$= \int_D \nabla u\cdot\nabla_h w_h\, \mathrm{d}x - \sum_{K \in \mathcal{T}_h} \sum_{F \in \mathcal{F}_K} \int_F (\boldsymbol{n}_K\cdot\nabla u) w_{h|K}\, \mathrm{d}s.$$

Note that we write $\boldsymbol{n}_K\cdot\nabla u$ instead of $\boldsymbol{n}_K\cdot\nabla u_{|K}$ since ∇u is single-valued on F because $u \in V_{\mathrm{s}}$. We want to exchange the order of the summations on the right-hand side. Recalling that for every interface $F := \partial K_l \cap \partial K_r \in \mathcal{F}_h^\circ$ with $\boldsymbol{n}_F$ pointing from K_l to K_r, i.e., $\boldsymbol{n}_F := \boldsymbol{n}_{K_l} = -\boldsymbol{n}_{K_r}$, we have

$$(\boldsymbol{n}_{K_l}\cdot\nabla u) w_{h|K_l} + (\boldsymbol{n}_{K_r}\cdot\nabla u) w_{h|K_r} = (\boldsymbol{n}_{K_l}\cdot\nabla u)[\![w_h]\!]_F.$$

For every boundary face $F := \partial K_l \cap \partial D \in \mathcal{F}_h^\partial$, recall that we have conventionally set $[\![w_h]\!]_F := w_{h|K_l}$. Thus, we infer that

$$\sum_{K \in \mathcal{T}_h} \sum_{F \in \mathcal{F}_K} \int_F (\boldsymbol{n}_K\cdot\nabla u) w_{h|K}\, \mathrm{d}s = \sum_{F \in \mathcal{F}_h} \int_F (\boldsymbol{n}_{K_l}\cdot\nabla u)[\![w_h]\!]_F\, \mathrm{d}s.$$

Setting $\eta := u - v_h$, we can write the consistency error as follows:

$$\langle \delta_h(v_h), w_h \rangle_{V_h', V_h} = \int_D \nabla_h \eta \cdot \nabla_h w_h \, \mathrm{d}x - \sum_{F \in \mathcal{F}_h} \int_F (\boldsymbol{n}_{K_l} \cdot \nabla u) [\![w_h]\!]_F \, \mathrm{d}s$$

$$= \int_D \nabla_h \eta \cdot \nabla_h w_h \, \mathrm{d}x - \sum_{F \in \mathcal{F}_h} \int_F (\boldsymbol{n}_{K_l} \cdot \nabla \eta_{|K_l}) [\![w_h]\!]_F \, \mathrm{d}s,$$

where we used that $\int_F (\boldsymbol{n}_{K_l} \cdot \nabla v_{h|K_l}) [\![w_h]\!]_F \, \mathrm{d}s = 0$ for all $F \in \mathcal{F}_h$ by definition of the Crouzeix–Raviart space $V_h = P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$. We conclude by invoking the Cauchy–Schwarz inequality, the first bound on the jumps in (36.14) which implies that $\sum_{F \in \mathcal{F}_h} h_F^{-1} \| [\![w_h]\!]_F \|_{L^2(F)}^2 \le c \| w_h \|_{V_h}^2$ (bound the infimum by taking $v := 0$), and the regularity of the mesh sequence. $\qquad \square$

Theorem 36.11 (Convergence). *Let u solve (36.2) and let u_h solve (36.10). Assume (36.17).*
(i) There is c s.t. the following quasi-optimal error estimate holds true for all $h \in \mathcal{H}$,

$$\| u - u_h \|_{V_\sharp} \le c \inf_{v_h \in V_h} \| u - v_h \|_{V_\sharp}. \tag{36.21}$$

(ii) Letting $t := \min(1, r)$, we have

$$\| u - u_h \|_{V_\sharp} \le c \left(\sum_{K \in \mathcal{T}_h} h_K^{2t} |u|_{H^{1+t}(K)}^2 \right)^{\frac{1}{2}}. \tag{36.22}$$

Proof. (i) The estimate (36.21) follows from Lemma 27.5 combined with stability (Lemma 36.4) and consistency/boundedness (Lemma 36.10).
(ii) The bound (36.22) follows from (36.21) by taking $v_h := \mathcal{I}_h^{\mathrm{CR}}(u)$. Letting $\eta := u - \mathcal{I}_h^{\mathrm{CR}}(u)$, we indeed have $\| \nabla \eta_{|K} \|_{\boldsymbol{L}^2(K)} \le c h_K^t |u|_{H^{1+t}(K)}$ for all $K \in \mathcal{T}_h$ owing to Lemma 36.1. Moreover, invoking the multiplicative trace inequality (12.17), we obtain

$$h_K^{\frac{1}{2}} \| \boldsymbol{n}_K \cdot \nabla \eta_{|K} \|_{L^2(\partial K)} \le \| \nabla \eta_{|K} \|_{\boldsymbol{L}^2(K)} + h_K^t |\eta_{|K}|_{H^{1+t}(K)},$$

and we have $|\eta_{|K}|_{H^{1+t}(K)} = |u|_{H^{1+r}(K)}$ since $\mathcal{I}_h^{\mathrm{CR}}(u)$ is affine in K. $\qquad \square$

Remark 36.12 (Strang 2). The analysis can also be done by invoking Strang's second lemma (Lemma 27.15). Let us set $V_\sharp := H_0^1(D) + P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$ and let us equip this space with the norm $\| \cdot \|_{V_\sharp}$ defined in (36.18). The discrete bilinear form a_h can be extended to a bilinear form $a_\sharp$ having boundedness constant equal to 1 on $V_\sharp \times V_h$. Lemma 27.15 leads to the error bound

$$\| u - v_h \|_{V_\sharp} \le c \left(\inf_{v_h \in V_h} \| u - v_h \|_{V_\sharp} + \| \delta_h^{\mathrm{St2}}(u) \|_{V_h'} \right),$$

with the consistency error s.t. for all $w_h \in P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$,

$$\langle \delta_h^{\mathrm{St2}}(u), w_h \rangle_{V_h', V_h} := \ell_h(w_h) - a_h(u, w_h) = \sum_{K \in \mathcal{T}_h} \int_K (f w_h - \nabla u \cdot \nabla w_{h|K}) \, \mathrm{d}x$$

$$= - \sum_{K \in \mathcal{T}_h} \int_{\partial K} (\boldsymbol{n}_K \cdot \nabla u) w_{h|K} \, \mathrm{d}s.$$

Thus, the consistency error does not vanish identically, i.e., the Crouzeix–Raviart finite element method is not strongly consistent in the sense defined in Remark 27.16. Since we have

$$\sum_{K \in \mathcal{T}_h} \int_{\partial K} (\boldsymbol{n}_K \cdot \nabla u) w_{h|K} \, \mathrm{d}s = \sum_{K \in \mathcal{T}_h} \int_{\partial K} (\boldsymbol{n}_K \cdot \nabla(u - v_h)) w_{h|K} \, \mathrm{d}s,$$

for all $v_h \in P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$, by proceeding as in the proof of Theorem 36.11, we infer again that the quasi-optimal error estimate (36.21) holds true. $\square$

36.3.2 L^2-error estimate

The goal of this section is to derive an improved L^2-error estimate of the form $\|u - u_h\|_{L^2(D)} \le c h^\gamma \ell_D^{1-\gamma} \|u - u_h\|_{V_\sharp}$ for some real number $\gamma > 0$, where ℓ_D is a length scale associated with D, e.g., $\ell_D := \mathrm{diam}(D)$.

   Proceeding as in §32.3, we invoke a *duality argument.* We consider for all $g \in L^2(D)$ the adjoint solution $\zeta_g \in V := H_0^1(D)$ such that

$$a(v, \zeta_g) = (v, g)_{L^2(D)}, \qquad \forall v \in V. \tag{36.23}$$

Notice that $-\Delta \zeta_g = g$ in D and $\gamma^{\mathrm{g}}(\zeta_g) = 0$. Owing to the elliptic regularity theory (see §31.4), there is $s \in (0, 1]$ and a constant c_{smo} such that $\|\zeta_g\|_{H^{1+s}(D)} \le c_{\mathrm{smo}} \ell_D^2 \|g\|_{L^2(D)}$ for all $g \in L^2(D)$. In the present setting of the Poisson equation with Dirichlet conditions in a Lipschitz polyhedron, it is reasonable to assume that $s \in (\frac{1}{2}, 1]$.

Theorem 36.13 (L^2-estimate). *Let u solve (36.2) and let u_h solve (36.10). Assume that the elliptic regularity index satisfies $s \in (\frac{1}{2}, 1]$. There is c s.t. for all $h \in \mathcal{H}$,*

$$\|u - u_h\|_{L^2(D)} \le c \, h^s \ell_D^{1-s} \|u - u_h\|_{V_\sharp}. \tag{36.24}$$

Proof. Let $e := u - u_h$ and set $Y_h := P_{1,0}^{\mathrm{g}}(\mathcal{T}_h) := P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h) \cap H_0^1(D)$. Then $(\nabla_h e, \nabla y_h)_{\boldsymbol{L}^2(D)} = (\nabla u, \nabla y_h)_{\boldsymbol{L}^2(D)} - (\nabla_h u_h, \nabla y_h)_{\boldsymbol{L}^2(D)} = 0$ for all $y_h \in Y_h$. Since $\|e\|_{L^2(D)}^2 = -(e, \Delta \zeta_e)_{\boldsymbol{L}^2(D)}$, we have

$$\|e\|_{L^2(D)}^2 = (\nabla_h e, \nabla \zeta_e)_{\boldsymbol{L}^2(D)} - \big((e, \Delta \zeta_e)_{L^2(D)} + (\nabla_h e, \nabla \zeta_e)_{\boldsymbol{L}^2(D)}\big)$$
$$= (\nabla_h e, \nabla(\zeta_e - y_h))_{\boldsymbol{L}^2(D)} - \langle \delta^{\mathrm{adj}}(\zeta_e), e \rangle_{V_\sharp', V_\sharp},$$

where we introduced $\delta^{\mathrm{adj}}(\zeta_e) \in V_\sharp'$ s.t. $\langle \delta^{\mathrm{adj}}(\zeta_e), v \rangle_{V_\sharp', V_\sharp} := (v, \Delta \zeta_e)_{L^2(D)} + (\nabla_h v, \nabla \zeta_e)_{\boldsymbol{L}^2(D)}$ and used that $(\nabla_h e, \nabla y_h)_{\boldsymbol{L}^2(D)} = 0$ for all $y_h \in V_h$. Let us set $\|\delta^{\mathrm{adj}}(\zeta_e)\|_{V_\sharp'} := \sup_{v \in V_\sharp} \frac{|\langle \delta^{\mathrm{adj}}(\zeta_e), v \rangle_{V_\sharp', V_\sharp}|}{\|v\|_{V_\sharp}}$. The Cauchy–Schwarz inequality and the definition of the $\|\cdot\|_{V_\sharp}$- and $\|\cdot\|_{V_\sharp'}$-norms imply that

$$\|e\|_{L^2(D)}^2 \le \left(\inf_{y_h \in Y_h} \|\nabla(\zeta_e - y_h)\|_{\boldsymbol{L}^2(D)} + \|\delta^{\mathrm{adj}}(\zeta_e)\|_{V_\sharp'} \right) \|e\|_{V_\sharp}.$$

It remains to bound the two terms between parentheses on the right-hand side. Using the quasi-interpolation operator $\mathcal{I}_{h0}^{\mathrm{g,av}}$ from §22.4, we infer that

$$\inf_{y_h \in Y_h} \|\nabla(\zeta_e - y_h)\|_{\boldsymbol{L}^2(D)} \le \|\nabla(\zeta_e - \mathcal{I}_{h0}^{\mathrm{g,av}}(\zeta_e))\|_{\boldsymbol{L}^2(D)}$$
$$\le c \, h^s |\zeta_e|_{H^{1+s}(D)} \le c \, h^s \ell_D^{-1-s} \|\zeta_e\|_{H^{1+s}(D)} \le c \, c_{\mathrm{smo}} \, h^s \ell_D^{1-s} \|e\|_{L^2(D)},$$

where we used the approximation properties of $\mathcal{I}_{h0}^{\mathrm{g,av}}$ from Theorem 22.14 and the elliptic regularity theory to bound $\|\zeta_e\|_{H^{1+s}(D)}$ by $\|e\|_{L^2(D)}$. Let us now estimate $\|\delta^{\mathrm{adj}}(\zeta_e)\|_{V_\sharp'}$. By proceeding as in

the proof of Lemma 36.10 (observe that $[\![\nabla\zeta_e]\!]_F = \mathbf{0}$ for all $F \in \mathcal{F}_h^\circ$), we infer that we have, for all $v := v_\mathrm{s} + v_h \in V_\sharp := V_\mathrm{s} + V_h$ with $v_\mathrm{s} \in V_\mathrm{s}$ and $v_h \in V_h$, and all $z_h \in V_h$,

$$
\begin{aligned}
\langle \delta^{\mathrm{adj}}(\zeta_e), v \rangle_{V_\sharp', V_\sharp} &= \sum_{F \in \mathcal{F}_h} \int_F \boldsymbol{n}_{K_l} \cdot \nabla\zeta_e [\![v_h]\!]_F \, \mathrm{d}s \\
&= \sum_{F \in \mathcal{F}_h} \int_F \boldsymbol{n}_{K_l} \cdot \nabla(\zeta_e - z_h)_{|K_l} [\![v_h]\!]_F \, \mathrm{d}s \\
&\leq c \, \|\zeta_e - z_h\|_{V_\sharp} \bigg(\sum_{F \in \mathcal{F}_h} h_F^{-1} \|[\![v_h]\!]_F\|_{L^2(F)}^2 \bigg)^{\frac{1}{2}},
\end{aligned}
$$

where we used that $\boldsymbol{n}_{K_l} \cdot \nabla z_{h|K_l}$ is constant on F. Using the leftmost inequality in (36.14) with $\inf_{w \in H_0^1(D)} \|\nabla_h(w - v_h)\|_{\boldsymbol{L}^2(D)}^2 \leq \|\nabla_h(v_\mathrm{s} + v_h)\|_{\boldsymbol{L}^2(D)}^2$, we infer that $\sum_{F \in \mathcal{F}_h} h_F^{-1} \|[\![v_h]\!]_F\|_{L^2(F)}^2 \leq c\|v_\mathrm{s} + v_h\|_{V_\sharp}^2 = c\|v\|_{V_\sharp}^2$. Thus, $\|\delta^{\mathrm{adj}}(\zeta_e)\|_{V_\sharp'} \leq c' \inf_{z_h \in V_h} \|\zeta_e - z_h\|_{V_\sharp}$. Using the approximation properties of V_h, we conclude that $\|\delta^{\mathrm{adj}}(\zeta_e)\|_{V_\sharp'} \leq ch^s |\zeta_e|_{H^{1+s}(D)}$, and reasoning as above yields $\|\delta^{\mathrm{adj}}(\zeta_e)\|_{V_\sharp'} \leq ch^s \ell_D^{1-s} \|e\|_{L^2(D)}$. $\qquad\square$

36.3.3 Abstract nonconforming duality argument

Let us finish with an abstract formulation of the above duality argument that can be applied in the context of nonconforming approximation techniques. Let V a Banach space, L be a Hilbert space, and assume that V embeds continuously into L (i.e., $V \hookrightarrow L$) and V is dense in L. Identifying L with L', we are in the situation where

$$
V \hookrightarrow L \equiv L' \hookrightarrow V', \tag{36.25}
$$

with continuous and dense embeddings. Let $a : V \times V \to \mathbb{C}$ be a bounded sesquilinear form satisfying the assumptions of the BNB theorem (Theorem 25.9). For all $f \in L$ we denote by ξ_f the unique solution to the problem

$$
a(\xi_f, v) = (f, v)_L, \qquad \forall v \in V. \tag{36.26}
$$

Similarly, for all $g \in L$ we denote by $\zeta_g \in V$ the unique solution to the adjoint problem

$$
a(v, \zeta_g) = (v, g)_L, \qquad \forall v \in V. \tag{36.27}
$$

These two problems are well-posed since a satisfies the assumptions of the BNB theorem. Let $A^{\mathrm{adj}} \in \mathcal{L}(V; V')$ be s.t. $\langle A^{\mathrm{adj}}(w), v \rangle_{V', V} = \overline{a(v, w)}$ for all $(v, w) \in V \times V$. Owing to (36.25) and (36.27), we have $A^{\mathrm{adj}}(\zeta_g) = g$ in L.

We assume that we have at hand two subspaces $V_\mathrm{s} \subset V$ and $Z_\mathrm{s} \subset V$ s.t. the maps $V' \ni f \mapsto \xi_f \in V_\mathrm{s}$ and $V' \ni g \mapsto \zeta_g \in Z_\mathrm{s}$ are bounded. Let $V_h \subset L$ be a finite-dimensional subspace of L (but not necessarily of V). Let $Y_h \subseteq V_h$. We set $V_\sharp := V_\mathrm{s} + V_h$ and $Z_\sharp := Z_\mathrm{s} + Y_h$, and we equip these spaces with norms denoted by $\|\cdot\|_{V_\sharp}$ and $\|\cdot\|_{Z_\sharp}$.

Lemma 36.14 (L-norm estimate). *Let $a_\sharp$ be a bounded sesquilinear form on $V_\sharp \times Z_\sharp$. Let $\|a_\sharp\|$ be the norm of $a_\sharp$ on $V_\sharp \times Z_\sharp$. Let $u \in V_\mathrm{s}$ and $u_h \in V_h$. Assume that the following Galerkin orthogonality property holds true:*

$$
a_\sharp(u - u_h, y_h) = 0, \qquad \forall y_h \in Y_h. \tag{36.28}
$$

Let $e := u - u_h$ and let $\delta^{\mathrm{adj}}(\zeta_e) \in V_\sharp'$ be the adjoint consistency error:

$$\langle \delta^{\mathrm{adj}}(\zeta_e), v \rangle_{V_\sharp', V_\sharp} := (v, A^{\mathrm{adj}}(\zeta_e))_L - a_\sharp(v, \zeta_e), \qquad \forall v \in V_\sharp. \qquad (36.29)$$

Then the following estimate holds true:

$$\|e\|_L \leq \left(\frac{\|\delta^{\mathrm{adj}}(\zeta_e)\|_{V_\sharp'}}{\|e\|_L} + \|a_\sharp\| \inf_{y_h \in Y_h} \frac{\|\zeta_e - y_h\|_{Z_\sharp}}{\|e\|_L} \right) \|e\|_{V_\sharp}, \qquad (36.30)$$

Proof. Using the identity $A^{\mathrm{adj}}(\zeta_e) = e$ and the Galerkin orthogonality property (36.28), we infer that

$$\|e\|_L^2 = (e, A^{\mathrm{adj}}(\zeta_e))_L = (e, A^{\mathrm{adj}}(\zeta_e))_L - a_\sharp(e, \zeta_e) + a_\sharp(e, \zeta_e)$$
$$= \langle \delta^{\mathrm{adj}}(\zeta_e), e \rangle_{V_\sharp', V_\sharp} + a_\sharp(e, \zeta_e - y_h).$$

The boundedness of $a_\sharp$ on $V_\sharp \times Z_\sharp$ and the definition of the dual norm $\|\delta^{\mathrm{adj}}(\zeta_e)\|_{V_\sharp'}$ imply that (36.30) holds true. $\square$

Example 36.15 (Crouzeix–Raviart). Lemma 36.14 can be applied to the Crouzeix–Raviart approximation with $V_{\mathrm{s}} := H^{1+r}(D) \cap H_0^1(D)$, $Z_{\mathrm{s}} := H^{1+s}(D) \cap H_0^1(D)$, $a_\sharp(v, w) := (\nabla_h v, \nabla_h w)_{\boldsymbol{L}^2(D)}$, and equipping the spaces $V_\sharp := V_{\mathrm{s}} + V_h$, $Z_\sharp := Z_{\mathrm{s}} + Y_h$, $Y_h := V_h \cap H_0^1(D)$, with the broken energy norm. Note that the adjoint consistency error is nonzero, and that the proof of Theorem 36.13 shows that both terms on the right-hand side of (36.30) converge with the same rate w.r.t. $h \in \mathcal{H}$. $\square$

Exercises

Exercise 36.1 (Commuting properties). Let K be a simplex in $\mathbb{R}^d$ and let Π_K^0 denote the L^2-orthogonal projection onto constants. Prove that $\nabla(\mathcal{I}_K^{\mathrm{CR}}(p)) = \Pi_K^0(\nabla p)$ and $\nabla \cdot (\boldsymbol{\mathcal{I}}_K^{\mathrm{CR}}(\boldsymbol{\sigma})) = \Pi_K^0(\nabla \cdot \boldsymbol{\sigma})$ for all $p \in H^1(K)$ and all $\boldsymbol{\sigma} \in \boldsymbol{L}^2(K)$ with $\nabla \cdot \boldsymbol{\sigma} \in L^1(K)$ and $\boldsymbol{\mathcal{I}}_K^{\mathrm{CR}}$ defined componentwise using $\mathcal{I}_h^{\mathrm{CR}}$.

Exercise 36.2 (Best approximation). Let $v \in H^1(D)$. A global best-approximation of v in $P_1^{\mathrm{CR}}(\mathcal{T}_h)$ in the broken H^1-seminorm is a function $v_h^{\mathrm{CR}} \in P_1^{\mathrm{CR}}(\mathcal{T}_h)$ s.t.

$$\sum_{K \in \mathcal{T}_h} \|\nabla(v - v_h^{\mathrm{CR}})\|_{\boldsymbol{L}^2(K)}^2 = \min_{v_h \in P_1^{\mathrm{CR}}(\mathcal{T}_h)} \sum_{K \in \mathcal{T}_h} \|\nabla(v - v_h)\|_{\boldsymbol{L}^2(K)}^2.$$

(i) Write a characterization of v_h^{CR} in weak form and show that v_h^{CR} is unique up to an additive constant. (*Hint*: adapt Proposition 25.8.) (ii) Let v_h^{b} be a global best-approximation of v in the broken finite element space $P_1^{\mathrm{b}}(\mathcal{T}_h)$; see §32.2. Prove that $\sum_{K \in \mathcal{T}_h} \|\nabla(v - v_h^{\mathrm{CR}})\|_{\boldsymbol{L}^2(K)}^2 = \sum_{K \in \mathcal{T}_h} \|\nabla(v - v_h^{\mathrm{b}})\|_{\boldsymbol{L}^2(K)}^2$. (*Hint*: using Exercise 36.1, show that $v_h^{\mathrm{CR}} = \mathcal{I}_h^{\mathrm{CR}}(v)$ up to an additive constant.)

Exercise 36.3 ($\boldsymbol{H}(\mathrm{div})$-flux recovery). Let u_h solve (36.10). Assume that f is piecewise constant on $\mathcal{T}_h$. Set $\boldsymbol{\sigma}_{h|K} := -\nabla u_{h|K} + \frac{1}{d} f_{|K}(\boldsymbol{x} - \boldsymbol{x}_K)$, where $\boldsymbol{x}_K$ is the barycenter of K for all $K \in \mathcal{T}_h$. Prove that $\boldsymbol{\sigma}_h$ is in the lowest-order Raviart–Thomas finite element space $\boldsymbol{P}_0^{\mathrm{d}}(\mathcal{T}_h)$ and that $\nabla \cdot \boldsymbol{\sigma} = f$; see Marini [295] (*Hint*: evaluate $\int_F [\![\boldsymbol{\sigma}_h]\!] \cdot \boldsymbol{n}_F \varphi_F^{\mathrm{CR}} \, \mathrm{d}s$ for all $F \in \mathcal{F}_h^\circ$.)

Exercise 36.4 (Discrete Helmholtz). Let $D \subset \mathbb{R}^2$ be a simply connected polygon. Prove that $\boldsymbol{P}_0^{\mathrm{b}}(\mathcal{T}_h) = \nabla P_1^{\mathrm{g}}(\mathcal{T}_h) \oplus \nabla_h^{\perp} P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$, where

$$\nabla_h^{\perp} P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h) := \{\boldsymbol{v}_h \in \boldsymbol{P}_0^{\mathrm{b}}(\mathcal{T}_h) \mid \exists q_h \in P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h) \mid \boldsymbol{v}_{h|K} = \nabla^{\perp}(q_{h|K}), \forall K \in \mathcal{T}_h\},$$

and $\nabla^{\perp}$ is the two-dimensional curl operator defined in Remark 16.17. (*Hint*: prove that the decomposition is L^2-orthogonal and use a dimension argument based on Euler's relations.)

Exercise 36.5 (Rannacher–Turek). Let $K := [-1,1]^d$. For all $i \in \{1{:}d\}$ and $\alpha \in \{l, r\}$, let $F_{i,\alpha}$ be the face of K corresponding to $\{x_i = -1\}$ when $\alpha = l$ and to $\{x_i = 1\}$ when $\alpha = r$. Observe that there are $2d$ such faces, each of measure 2^{d-1}. Let P be spanned by the $2d$ functions $\{1, x_1, \ldots, x_d, x_1^2 - x_2^2, \ldots, x_{d-1}^2 - x_d^2\}$. Consider the linear forms $\sigma_{i,\alpha}(p) := 2^{1-d} \int_{F_{i,\alpha}} p \, \mathrm{d}s$ for all $i \in \{1{:}d\}$ and $\alpha \in \{l, r\}$. Setting $\Sigma := \{\sigma_{i,\alpha}\}_{i \in \{1{:}d\}, \alpha \in \{l,r\}}$, prove that (K, P, Σ) is a finite element. *Note*: this element has been introduced by [330] for the mixed discretization of the Stokes equations on Cartesian grids.

Exercise 36.6 (Quadratic space). Let $\mathcal{T}_h$ be a triangulation of a simply connected domain $D \subset \mathbb{R}^2$ and let

$$P_2^{\mathrm{CR}}(\mathcal{T}_h) := \{v_h \in P_2^{\mathrm{b}}(\mathcal{T}_h) \mid \int_F [\![v_h]\!]_F (q \circ \boldsymbol{T}_F^{-1}) \, \mathrm{d}s = 0, \forall F \in \mathcal{F}_h^{\circ}, \forall q \in \mathbb{P}_{1,1}\},$$

where $\boldsymbol{T}_F$ is an affine bijective mapping from the unit segment $\widehat{S}^1 = [-1,1]$ to F. Orient all the faces $F \in \mathcal{F}_h$ and define the two Gauss points $\boldsymbol{g}_F^{\pm}$ on F that are the image by $\boldsymbol{T}_F$ of $\widehat{g}^{\pm} := \pm\frac{\sqrt{3}}{3}$, in such a way that the orientation of F goes from $\boldsymbol{g}_F^-$ to $\boldsymbol{g}_F^+$. For all $K \in \mathcal{T}_h$, let $\{\lambda_{0,K}, \lambda_{1,K}, \lambda_{2,K}\}$ be the barycentric coordinates in K and set $b_K := 2 - 3(\lambda_{0,K}^2 + \lambda_{1,K}^2 + \lambda_{2,K}^2)$ (this function is usually called Fortin–Soulié bubble [204]). One can verify that a polynomial $p \in \mathbb{P}_{2,2}$ vanishes at the six points $\{\boldsymbol{g}_F^{\pm}\}_{F \in \mathcal{F}_K}$ if and only if $p = \alpha b_K$ for some $\alpha \in \mathbb{R}$. *Note*: this shows that these six points, which lie on an ellipse, cannot be taken as nodes of a $\mathbb{P}_{2,2}$ Lagrange element. (i) Extending b_K by zero outside K, verify that $b_K \in P_2^{\mathrm{CR}}(\mathcal{T}_h)$. (ii) Set $B := \mathrm{span}_{K \in \mathcal{T}_h}\{b_K\}$ and $B_* := \{v_h \in B \mid \int_D v_h \, \mathrm{d}x = 0\}$. Prove that $P_2^{\mathrm{g}}(\mathcal{T}_h) + B_* \subset P_2^{\mathrm{CR}}(\mathcal{T}_h)$ and that $P_2^{\mathrm{g}}(\mathcal{T}_h) \cap B_* = \{0\}$. (iii) Define $J : P_2^{\mathrm{CR}}(\mathcal{T}_h) \to \mathbb{R}^{2N_{\mathrm{f}}}$ s.t. $J(v_h) := (v_h(\boldsymbol{g}_F^-), v_h(\boldsymbol{g}_F^+))_{F \in \mathcal{F}_h}$ for all $v_h \in P_2^{\mathrm{CR}}(\mathcal{T}_h)$. Prove that $\dim(\ker(J)) = N_{\mathrm{c}}$ and $\dim(\mathrm{im}(J)) \leq 2N_{\mathrm{f}} - N_{\mathrm{c}}$. (*Hint*: any polynomial $p \in \mathbb{P}_{2,2}$ satisfies $\sum_{F \in \mathcal{F}_K}(p(\boldsymbol{g}_F^+) - p(\boldsymbol{g}_F^-)) = 0$ for all $K \in \mathcal{T}_h$.) (iv) Prove that $P_2^{\mathrm{CR}}(\mathcal{T}_h) = P_2^{\mathrm{g}}(\mathcal{T}_h) \oplus B_*$; see Greff [222]. (*Hint*: use a dimensional argument and Euler's relation from Remark 8.13.)

Chapter 37

Nitsche's boundary penalty method

The main objective of this chapter is to present a technique to treat Dirichlet boundary conditions in a natural way using a penalty method. This technique is powerful and has many extensions. In particular, the idea is reused in the next chapter for discontinuous Galerkin methods. Another objective of this chapter is to illustrate the abstract error analysis of Chapter 27.

37.1   Main ideas and discrete problem

Let D be a Lipschitz domain in $\mathbb{R}^d$. We assume for simplicity that D is a polyhedron. Let $f \in L^2(D)$ be the source term, and let $g \in H^{\frac{1}{2}}(\partial D)$ be the Dirichlet boundary data. We consider the Poisson equation with Dirichlet conditions

$$-\Delta u = f \quad \text{in } D, \qquad \gamma^{\mathrm{g}}(u) = g \quad \text{on } \partial D, \tag{37.1}$$

where $\gamma^{\mathrm{g}} : H^1(D) \to H^{\frac{1}{2}}(\partial D)$ is the trace map. Let $u_g \in H^1(D)$ be a lifting of g, i.e., $\gamma^{\mathrm{g}}(u_g) = g$ (recall that $\gamma^{\mathrm{g}} : H^1(D) \to H^{\frac{1}{2}}(\partial D)$ is the trace map). We seek $u_0 \in H_0^1(D)$ s.t. $a(u_0, w) = \ell(w) - a(u_g, w)$ for all $w \in H_0^1(D)$, with

$$a(v, w) := \int_D \nabla v \cdot \nabla w \, \mathrm{d}x, \qquad \ell(w) := \int_D f w \, \mathrm{d}x. \tag{37.2}$$

This problem is well-posed in $H_0^1(D)$ owing to the Lax–Milgram lemma and the Poincaré–Steklov inequality in $H_0^1(D)$. Then the unique weak solution to (37.1) is $u := u_0 + u_g$ (see §31.2.2).

   In this chapter, we take a route that is different from the above approach to construct an approximation of the solution. Instead of enforcing the Dirichlet boundary condition strongly, we are going to construct an H^1-conforming discretization of (37.1) that enforces this condition naturally. This means that we no longer require that the discrete test functions vanish at the boundary. The discrete counterpart of the bilinear form a must then be modified accordingly. To motivate the modification in question, let us proceed informally by assuming that all the functions we manipulate are sufficiently smooth. Testing (37.1) with a function w which we do not require to vanish at the boundary, the integration by parts formula (4.8b) gives

$$a(u, w) - \int_{\partial D} (\boldsymbol{n} \cdot \nabla u) w \, \mathrm{d}s = \ell(w). \tag{37.3}$$

The idea of Nitsche is to modify (37.3) by adding a term proportional to $\int_{\partial D} uw \, \mathrm{d}s$ on both sides of the above identity. This leads to

$$a(u, w) - \int_{\partial D} (\boldsymbol{n} \cdot \nabla u) w \, \mathrm{d}s + \varpi \int_{\partial D} uw \, \mathrm{d}s = \ell(w) + \varpi \int_{\partial D} gw \, \mathrm{d}s, \qquad (37.4)$$

where the boundary value of u has been replaced by g in the boundary integral on the right-hand side. The yet unspecified parameter ϖ is assumed to be positive. Heuristically, if u satisfies (37.4) and if ϖ is large, one expects u to be close to g at the boundary. For this reason, ϖ is called *penalty parameter*.

The above ideas lead to an approximation method employing discrete trial and test spaces composed of functions that are not required to vanish at the boundary. Let $\mathcal{T}_h$ be a mesh from a shape-regular sequence of meshes so that each mesh covers D exactly. Let $\mathcal{F}_h^{\partial}$ be the collection of the boundary faces. Let $P_k^{\mathrm{g}}(\mathcal{T}_h)$ be the H^1-conforming finite element space of degree $k \geq 1$ based on $\mathcal{T}_h$; see (19.10). We consider the following discrete problem:

$$\begin{cases} \text{Find } u_h \in V_h := P_k^{\mathrm{g}}(\mathcal{T}_h) \text{ such that} \\ a_h(u_h, w_h) = \ell_h(w_h), \quad \forall w_h \in V_h, \end{cases} \qquad (37.5)$$

where the discrete forms a_h and ℓ_h are inspired from (37.4):

$$a_h(v_h, w_h) := a(v_h, w_h) - \int_{\partial D} (\boldsymbol{n} \cdot \nabla v_h) w_h \, \mathrm{d}s + \sum_{F \in \mathcal{F}_h^{\partial}} \varpi(h_F) \int_F v_h w_h \, \mathrm{d}s,$$

$$\ell_h(w_h) := \ell(w_h) + \sum_{F \in \mathcal{F}_h^{\partial}} \varpi(h_F) \int_F gw_h \, \mathrm{d}s.$$

The second term in the definition of a_h is called *consistency term* (this term plays a key role when estimating the consistency error) and the third one is called *penalty term*. The penalty parameter $\varpi(h_F) > 0$, yet to be defined, depends on the diameter of the face F (or a uniformly equivalent local length scale). The stability analysis will reveal that $\varpi(h_F)$ should scale like h_F^{-1}. The approximation setting associated with Nitsche's boundary penalty method is nonconforming, i.e., $V_h \not\subset V := H_0^1(D)$, since functions in V_h may not vanish at the boundary, whereas functions in V do.

Remark 37.1 (Literature, extensions). The boundary penalty method has been introduced by Nitsche [314] to treat Dirichlet boundary conditions. It was extended in Juntunen and Stenberg [262] to Robin boundary conditions. We refer the reader to §41.3 where the more general PDE $-\nabla \cdot (\lambda \nabla u) = f$ with contrasted diffusivity λ is treated. $\qquad \square$

37.2   Stability and well-posedness

The main objective of this section is to prove that the discrete bilinear form a_h is coercive on V_h if the penalty parameter is large enough. This is done by showing that the consistency term can be appropriately bounded. For all $F \in \mathcal{F}_h^{\partial}$, let us denote by K_l the unique mesh cell having F as a face, i.e., $F := \partial K_l \cap \partial D$. Let $\mathcal{T}_h^{\partial D}$ be the collection of the mesh cells having at least one boundary face, i.e., $\mathcal{T}_h^{\partial D} := \bigcup_{F \in \mathcal{F}_h^{\partial}} \{K_l\}$. (The set $\mathcal{T}_h^{\partial D}$ should not be confused with the larger set $\mathcal{T}_h^{\partial}$ defined in (22.28), which is the collection of the mesh cells touching the boundary.)

Let n_∂ denote the maximum number of boundary faces that a mesh cell in $\mathcal{T}_h^{\partial D}$ can have, i.e., $n_\partial := \max_{K \in \mathcal{T}_h^{\partial D}} \mathrm{card}(\mathcal{F}_K \cap \mathcal{F}_h^\partial)$ ($n_\partial \leq d$ for simplicial meshes). Owing to the regularity of the mesh sequence, the discrete trace inequality from Lemma 12.8 (with $p = q := 2$) implies that there is c_{dt} such that for all $v_h \in V_h$, all $F \in \mathcal{F}_h^\partial$, and all $h \in \mathcal{H}$,

$$\|\boldsymbol{n}{\cdot}\nabla v_h\|_{L^2(F)} \leq c_{\mathrm{dt}} h_F^{-\frac{1}{2}} \|\nabla v_h\|_{\boldsymbol{L}^2(K_l)}. \tag{37.6}$$

Lemma 37.2 (Bound on consistency term). *The following holds true for all $v_h \in V_h$:*

$$\left| \int_{\partial D} (\boldsymbol{n}{\cdot}\nabla v_h) v_h \, \mathrm{d}s \right| \leq n_\partial^{\frac{1}{2}} c_{\mathrm{dt}} \left(\sum_{K \in \mathcal{T}_h^{\partial D}} \|\nabla v_h\|_{\boldsymbol{L}^2(K)}^2 \right)^{\frac{1}{2}} \left(\sum_{F \in \mathcal{F}_h^\partial} \frac{1}{h_F} \|v_h\|_{L^2(F)}^2 \right)^{\frac{1}{2}}.$$

Proof. Let $v_h \in V_h$. Let $F \in \mathcal{F}_h^\partial$. Using the Cauchy–Schwarz inequality, bounding the normal component of the gradient by its Euclidean norm, and using the discrete trace inequality (37.6) componentwise, we infer that

$$\left| \int_{\partial D} (\boldsymbol{n}{\cdot}\nabla v_h) v_h \, \mathrm{d}s \right| \leq \left(\sum_{F \in \mathcal{F}_h^\partial} h_F \|\boldsymbol{n}{\cdot}\nabla v_h\|_{\boldsymbol{L}^2(F)}^2 \right)^{\frac{1}{2}} \left(\sum_{F \in \mathcal{F}_h^\partial} \frac{1}{h_F} \|v_h\|_{L^2(F)}^2 \right)^{\frac{1}{2}}$$

$$\leq c_{\mathrm{dt}} \left(\sum_{F \in \mathcal{F}_h^\partial} \|\nabla v_h\|_{\boldsymbol{L}^2(K_l)}^2 \right)^{\frac{1}{2}} \left(\sum_{F \in \mathcal{F}_h^\partial} \frac{1}{h_F} \|v_h\|_{L^2(F)}^2 \right)^{\frac{1}{2}}.$$

Finally, we have $\sum_{F \in \mathcal{F}_h^\partial} \|\cdot\|_{\boldsymbol{L}^2(K_l)}^2 = \sum_{K \in \mathcal{T}_h^{\partial D}} \mathrm{card}(\mathcal{F}_K \cap \mathcal{F}_h^\partial) \|\cdot\|_{\boldsymbol{L}^2(K)}^2 \leq n_\partial \sum_{K \in \mathcal{T}_h^{\partial D}} \|\cdot\|_{\boldsymbol{L}^2(K)}^2$. $\square$

We equip the space V_h with the following norm:

$$\|v_h\|_{V_h}^2 := \|\nabla v_h\|_{\boldsymbol{L}^2(D)}^2 + |v_h|_\partial^2, \qquad |v_h|_\partial^2 := \sum_{F \in \mathcal{F}_h^\partial} \frac{1}{h_F} \|v_h\|_{L^2(F)}^2. \tag{37.7}$$

Note that $\|v_h\|_{V_h} = 0$ implies that v_h is constant on D and vanishes on ∂D, so that $v_h = 0$. Hence, $\|\cdot\|_{V_h}$ is a norm on V_h. Note also that the two terms composing the norm $\|\cdot\|_{V_h}$ are dimensionally consistent.

Lemma 37.3 (Coercivity, well-posedness). *Assume that the penalty parameter $\varpi(h_F)$ is defined s.t.*

$$\varpi(h_F) := \varpi_0 \frac{1}{h_F}, \qquad \forall F \in \mathcal{F}_h^\partial, \tag{37.8}$$

with $\varpi_0 > \frac{1}{4} n_\partial c_{\mathrm{dt}}^2$. (i) The following coercivity property holds true:

$$a_h(v_h, v_h) \geq \alpha \|v_h\|_{V_h}^2, \qquad \forall v_h \in V_h, \tag{37.9}$$

with $\alpha := \frac{\varpi_0 - \frac{1}{4} n_\partial c_{\mathrm{dt}}^2}{1 + \varpi_0} > 0$, (ii) The discrete problem (37.5) is well-posed.

Proof. Let $v_h \in V_h$. We have

$$a_h(v_h, v_h) = \|\nabla v_h\|_{\boldsymbol{L}^2(D)}^2 - \int_{\partial D} (\boldsymbol{n}{\cdot}\nabla v_h) v_h \, \mathrm{d}s + \varpi_0 |v_h|_\partial^2.$$

Setting $z := (\sum_{K \in \mathcal{T}_h \setminus \mathcal{T}_h^{\partial D}} \|\nabla v_h\|_{\boldsymbol{L}^2(K)}^2)^{\frac{1}{2}}$, $x := (\sum_{K \in \mathcal{T}_h^{\partial D}} \|\nabla v_h\|_{\boldsymbol{L}^2(K)}^2)^{\frac{1}{2}}$, and $y := |v_h|_\partial$, and using Lemma 37.2, we infer that

$$a_h(v_h, v_h) \geq z^2 + (x^2 - n_\partial^{\frac{1}{2}} c_{\mathrm{dt}} xy + \varpi_0 y^2).$$

Coercivity follows from the inequality $x^2 - 2\beta xy + \varpi_0 y^2 \geq \frac{\varpi_0 - \beta^2}{1+\varpi_0}(x^2 + y^2)$ applied with $\beta := \frac{1}{2} n_\partial^{\frac{1}{2}} c_{\mathrm{dt}}$ (see Exercise 37.2) and since $\frac{\varpi_0 - \beta^2}{1+\varpi_0} \leq \frac{\varpi_0}{1+\varpi_0} \leq 1$. Finally, well-posedness follows from the Lax–Milgram lemma.   □

Remark 37.4 (Choice of penalty parameter). Ensuring the stability condition $\varpi_0 > \frac{1}{4} n_\partial c_{\mathrm{dt}}^2$ requires in practice to know a reasonable upper bound on the constant c_{dt}. The results of §12.2 show that c_{dt} scales like the polynomial degree k. More precisely, Lemma 12.10 shows that for simplices one can take $c_{\mathrm{dt}} := ((k+1)(k+d)/d)^{\frac{1}{2}}$ with $h_F := |K_l|/|F|$.   □

37.3 Error analysis

In this section, we derive an energy error estimate, that is, we bound the error by using the coercivity norm and the abstract error estimate from Lemma 27.5. We also derive an improved L^2-error estimate by means of a duality argument.

37.3.1 Energy error estimate

We perform the error analysis under the assumption that the solution to (37.1) is in $H^{1+r}(D)$ with $r > \frac{1}{2}$, i.e., we set

$$V_{\mathrm{s}} := H^{1+r}(D), \quad r > \frac{1}{2}. \tag{37.10}$$

The assumption $u \in V_{\mathrm{s}}$ is reasonable in the setting of the Poisson equation with Dirichlet conditions in a Lipschitz polyhedron since it is consistent with the elliptic regularity theory (see Theorem 31.33). The important property that we use is that for any function $v \in V_{\mathrm{s}}$, the normal derivative $\boldsymbol{n} \cdot \nabla v$ at the boundary is meaningful in $L^2(\partial D)$. We consider the space $V_\sharp := V_{\mathrm{s}} + V_h$ equipped with the norm

$$\|v\|_{V_\sharp}^2 := \|\nabla v\|_{\boldsymbol{L}^2(D)}^2 + |v|_\partial^2 + \sum_{F \in \mathcal{F}_h^\partial} h_F \|\boldsymbol{n} \cdot \nabla v\|_{L^2(F)}^2, \tag{37.11}$$

with $|v|_\partial^2 := \sum_{F \in \mathcal{F}_h^\partial} \frac{1}{h_F} \|v\|_{L^2(F)}^2$. A discrete trace inequality shows that there is $c_\sharp$ s.t. $\|v_h\|_{V_\sharp} \leq c_\sharp \|v_h\|_{V_h}$ for all $v_h \in V_h$ and all $h \in \mathcal{H}$, i.e., (27.5) holds true. Recall from Definition 27.3 that the consistency error is defined by setting $\langle \delta_h(v_h), w_h \rangle_{V_h', V_h} := \ell_h(w_h) - a_h(v_h, w_h)$ for all $v_h, w_h \in V_h$.

Lemma 37.5 (Consistency/boundedness). *Assume (37.10). There is $\omega_\sharp$, uniform w.r.t. $u \in V_{\mathrm{s}}$, s.t. for all $v_h \in V_h$ and all $h \in \mathcal{H}$,*

$$\|\delta_h(v_h)\|_{V_h'} \leq \omega_\sharp \|u - v_h\|_{V_\sharp}. \tag{37.12}$$

Proof. Let $v_h, w_h \in V_h$. Since the normal derivative $\boldsymbol{n}\cdot\nabla u$ is meaningful at the boundary, using the PDE and the boundary condition in (37.1), we infer that

$$
\ell_h(w_h) = \int_D -(\Delta u)w_h \, \mathrm{d}x + \sum_{F \in \mathcal{F}_h^\partial} \varpi_0 \frac{1}{h_F} \int_F g w_h \, \mathrm{d}s
$$

$$
= \int_D \nabla u \cdot \nabla w_h \, \mathrm{d}x - \int_{\partial D} (\boldsymbol{n}\cdot\nabla u)w_h \, \mathrm{d}s + \sum_{F \in \mathcal{F}_h^\partial} \varpi_0 \frac{1}{h_F} \int_F u w_h \, \mathrm{d}s.
$$

Letting $\eta := u - v_h$, this implies that

$$
\langle \delta_h(v_h), w_h \rangle_{V_h', V_h} = \int_D \nabla \eta \cdot \nabla w_h \, \mathrm{d}x - \int_{\partial D} (\boldsymbol{n}\cdot\nabla \eta)w_h \, \mathrm{d}s + \sum_{F \in \mathcal{F}_h^\partial} \varpi_0 \frac{1}{h_F} \int_F \eta w_h \, \mathrm{d}s.
$$

Using the Cauchy–Schwarz inequality, we obtain the estimate (37.12) with $\omega_\sharp := \max(1, \varpi_0)$. $\quad\square$

Theorem 37.6 (Convergence). *Let u solve (37.1) and let u_h solve (37.5) with the penalty parameter $\varpi_0 > \frac{1}{4} n_\partial c_{\mathrm{dt}}^2$. Assume (37.10). (i) There is c s.t. the following quasi-optimal error estimate holds true for all $h \in \mathcal{H}$:*

$$
\|u - u_h\|_{V_\sharp} \le c \inf_{v_h \in V_h} \|u - v_h\|_{V_\sharp}. \tag{37.13}
$$

(ii) *Letting $t := \min(k, r)$, we have*

$$
\|u - u_h\|_{V_\sharp} \le c \left(\sum_{K \in \mathcal{T}_h} h_K^{2t} |u|_{H^{1+t}(K)}^2 \right)^{\frac{1}{2}}. \tag{37.14}
$$

Proof. (i) The estimate (37.13) follows from Lemma 27.5 combined with stability (Lemma 37.3) and consistency/boundedness (Lemma 37.5).
(ii) The proof of (37.14) is left as an exercise. $\quad\square$

37.3.2 L^2-norm estimate

We derive an improved error estimate of the form $\|u - u_h\|_{L^2(D)} \le ch^\gamma \ell_D^{1-\gamma} \|u - u_h\|_{V_\sharp}$ for some real number $\gamma > 0$, where ℓ_D is a length scale associated with D, e.g., $\ell_D := \mathrm{diam}(D)$. Proceeding as in §36.3.2, we invoke a *duality argument*. We consider the adjoint solution $\zeta_r \in V := H_0^1(D)$ for all $r \in L^2(D)$ such that

$$
a(v, \zeta_r) = (v, r)_{L^2(D)}, \qquad \forall v \in V, \tag{37.15}
$$

i.e., ζ_r solves $-\Delta \zeta_r = r$ in D and $\gamma^{\mathrm{g}}(\zeta_r) = 0$. (Note that we enforce a homogeneous Dirichlet condition on the adjoint solution.) Owing to the elliptic regularity theory (see §31.4), there is $s \in (0, 1]$ and a constant c_{smo} such that

$$
\|\zeta_r\|_{H^{1+s}(D)} \le c_{\mathrm{smo}} \, \ell_D^2 \|r\|_{L^2(D)}, \qquad \forall r \in L^2(D). \tag{37.16}
$$

In the present setting of the Poisson equation with Dirichlet conditions in a Lipschitz polyhedron, it is reasonable to assume that $s \in (\frac{1}{2}, 1]$.

Theorem 37.7 (L^2-estimate). *Let u solve (37.1) and let u_h solve (37.5). Assume that the elliptic regularity index satisfies $s \in (\frac{1}{2}, 1]$. There is c s.t. for all $h \in \mathcal{H}$,*

$$
\|u - u_h\|_{L^2(D)} \le c \, h^{\frac{1}{2}} \ell_D^{\frac{1}{2}} \|u - u_h\|_{V_\sharp}. \tag{37.17}
$$

Proof. Set $e := u - u_h$. We apply the abstract error estimate of Lemma 36.14 with $V_\sharp := V_s + V_h$ as above, $Z_s := H^{1+s}(D) \cap H_0^1(D)$, $Y_h := V_h \cap H_0^1(D)$, and $Z_\sharp := Z_s + Y_h$ equipped with the H^1-seminorm. We consider the bilinear form $a_\sharp(v, w) := (\nabla v, \nabla w)_{L^2(D)}$. Notice that $a_\sharp$ is bounded on $V_\sharp \times Z_\sharp$. Moreover, $a_\sharp(e, y_h) = 0$ for all $y_h \in Y_h$ since $Y_h \subset H_0^1(D)$, i.e., the Galerkin orthogonality property (36.28) holds true. Lemma 36.14 implies that

$$\|e\|_{L^2(D)} \leq \left(\frac{\|\delta^{\mathrm{adj}}(\zeta_e)\|_{V_\sharp'}}{\|e\|_{L^2(D)}} + \inf_{y_h \in Y_h} \frac{\|\nabla(\zeta_e - y_h)\|_{\boldsymbol{L}^2(D)}}{\|e\|_{L^2(D)}} \right) \|e\|_{V_\sharp},$$

where the first and the second term between parentheses are the *adjoint consistency error* and the interpolation error on the adjoint solution, respectively. Let us first bound the adjoint consistency error. Recall that $\delta^{\mathrm{adj}}(\zeta_e)$ is defined in such a way that the following identity holds true: For all $v \in V_\sharp$,

$$\langle \delta^{\mathrm{adj}}(\zeta_e), v \rangle_{V_\sharp', V_\sharp} = -(v, \Delta\zeta_e)_{L^2(D)} - a_\sharp(v, \zeta_e) = -(v, \boldsymbol{n} \cdot \nabla\zeta_e)_{L^2(\partial D)}.$$

The Cauchy–Schwarz inequality implies that

$$|\langle \delta^{\mathrm{adj}}(\zeta_e), v \rangle_{V_\sharp', V_\sharp}| \leq h^{\frac{1}{2}} \|\nabla\zeta_e\|_{L^2(\partial D)} |v|_\partial \leq h^{\frac{1}{2}} \|\nabla\zeta_e\|_{L^2(\partial D)} \|v\|_{V_\sharp}$$

$$\leq c\, h^{\frac{1}{2}} \ell_D^{-\frac{3}{2}} \|\zeta_e\|_{H^{1+s}(D)} \|v\|_{V_\sharp},$$

since $s > \frac{1}{2}$. Using (37.16), we infer that $\|\delta^{\mathrm{adj}}(\zeta_e)\|_{V_\sharp'} \leq ch^{\frac{1}{2}} \ell_D^{\frac{1}{2}} \|e\|_{L^2(D)}$. To bound the interpolation error on the adjoint solution, we consider the quasi-interpolation operator $\mathcal{I}_{h0}^{\mathrm{g,av}}$ from §22.4. Since $\mathcal{I}_{h0}^{\mathrm{g,av}}(\zeta_e) \in Y_h$, we deduce that

$$\inf_{y_h \in Y_h} \|\nabla(\zeta_e - y_h)\|_{\boldsymbol{L}^2(D)} \leq \|\nabla(\zeta_e - \mathcal{I}_{h0}^{\mathrm{g,av}}(\zeta_e))\|_{\boldsymbol{L}^2(D)}$$

$$\leq c\, h^s |\zeta_e|_{H^{1+s}(D)} \leq c\, h^s \ell_D^{-1-s} \|\zeta_e\|_{H^{1+s}(D)} \leq c\, c_{\mathrm{smo}}\, h^s \ell_D^{1-s} \|e\|_{L^2(D)},$$

where we used the approximation properties of $\mathcal{I}_{h0}^{\mathrm{g,av}}$ from Theorem 22.14 and the estimate (37.16). Since $s > \frac{1}{2}$ and $h \leq \ell_D$, we have $h^s \ell_D^{1-s} \leq h^{\frac{1}{2}} \ell_D^{\frac{1}{2}}$, and this concludes the proof. $\qquad\square$

37.3.3 Symmetrization

The estimate (37.17) is suboptimal by a factor $h^{s-\frac{1}{2}}$, and this loss of optimality is caused by the adjoint consistency error which is only of order $h^{\frac{1}{2}}$. This shortcoming can be avoided by symmetrizing a_h and modifying ℓ_h consistently. More precisely, we define

$$a_h^{\mathrm{sym}}(v_h, w_h) := a(v_h, w_h) - \int_{\partial D} (\boldsymbol{n} \cdot \nabla v_h) w_h \, \mathrm{d}s - \int_{\partial D} v_h(\boldsymbol{n} \cdot \nabla w_h) \, \mathrm{d}s$$

$$+ \sum_{F \in \mathcal{F}_h^\partial} \varpi_0 \frac{1}{h_F} \int_F v_h w_h \, \mathrm{d}s,$$

$$\ell_h^{\mathrm{sym}}(w_h) := \ell(w_h) - \int_{\partial D} g(\boldsymbol{n} \cdot \nabla w_h) \, \mathrm{d}s + \sum_{F \in \mathcal{F}_h^\partial} \varpi_0 \frac{1}{h_F} \int_F g w_h \, \mathrm{d}s.$$

Consider the following discrete problem:

$$\begin{cases} \text{Find } u_h \in V_h \text{ such that} \\ a_h^{\mathrm{sym}}(u_h, w_h) = \ell_h^{\mathrm{sym}}(w_h), \quad \forall w_h \in V_h. \end{cases} \tag{37.18}$$

Adapting the proof of Lemma 37.3, one can show that the problem (37.18) is well-posed if one chooses the stabilization parameter s.t. $\varpi_0 > n_\partial c_{\mathrm{dt}}^2$.

Theorem 37.8 (L^2-estimate). *Let u solve (37.1) and let u_h solve (37.18). Assume $\varpi_0 > n_\partial c_{\mathrm{dt}}^2$ and that there is $s \in (\frac{1}{2}, 1]$ s.t. the adjoint solution satisfies the a priori estimate (37.16). There is c s.t. for all $h \in \mathcal{H}$,*

$$\|u - u_h\|_{L^2(D)} \leq c\, h^s \ell_D^{1-s} \|u - u_h\|_{V_\sharp}. \tag{37.19}$$

Proof. We proceed as in the proof of Theorem 37.7 with the same spaces $V_\sharp$, $Z_\sharp$, and Y_h, but now we set $a_\sharp(v, w) := (\nabla v, \nabla w)_{\boldsymbol{L}^2(D)} - (v, \boldsymbol{n}{\cdot}\nabla w)_{L^2(\partial D)}$. We equip $Z_\sharp$ with the same norm as $V_\sharp$, so that $a_\sharp$ is bounded on $V_\sharp{\times}Z_\sharp$. The Galerkin orthogonality property still holds true for $a_\sharp$. Indeed, we have

$$a_\sharp(u, y_h) = (f, y_h)_{L^2(D)} - (g, \boldsymbol{n}{\cdot}\nabla y_h)_{L^2(\partial D)}$$
$$= \ell_h^{\mathrm{sym}}(y_h) = a_h^{\mathrm{sym}}(u_h, y_h) = a_\sharp(u_h, y_h), \qquad \forall y_h \in Y_h,$$

since $\gamma^{\mathrm{g}}(u) = g$ and y_h vanishes on ∂D. Now the adjoint consistency error vanishes, and we still have $\|\zeta_e - \mathcal{I}_{h0}^{\mathrm{g,av}}(\zeta_e)\|_{Z_\sharp} \leq c\, h^s |\zeta_e|_{H^{1+s}(D)}$. $\qquad\blacksquare$

Exercises

Exercise 37.1 (Poincaré–Steklov). Let $\check{C}_{\mathrm{PS}}$ be defined in (31.23). Prove that $\check{C}_{\mathrm{PS}} \ell_D^{-1} \|v\|_{L^2(D)} \leq (\|\nabla v\|_{\boldsymbol{L}^2(D)}^2 + |v|_\partial^2)^{\frac{1}{2}}$ for all $v \in H^1(D)$. (*Hint*: use $h \leq \ell_D$ and (31.23).)

Exercise 37.2 (Quadratic inequality). Prove that $x^2 - 2\beta xy + \varpi_0 y^2 \geq \frac{\varpi_0 - \beta^2}{1+\varpi_0}(x^2 + y^2)$ for all real numbers x, y, $\varpi_0 \geq 0$ and $\beta \geq 0$.

Exercise 37.3 (Error estimate). Prove (37.14). (*Hint*: consider the quasi-interpolation operator from §22.3.)

Exercise 37.4 (Gradient). Let U be an open bounded set in $\mathbb{R}^d$, let $s \in (0, 1)$, and set $\boldsymbol{H}_{00}^s(U) := [\boldsymbol{L}^2(U), \boldsymbol{H}_0^1(U)]_{s,2}$. (i) Show that $\nabla : \boldsymbol{H}^{1-s}(U) \to (\boldsymbol{H}_{00}^s(U))'$ is bounded for all $s \in (0, 1)$. (*Hint*: use Theorems A.27 and A.30.) (ii) Assume that U is Lipschitz. Show that $\nabla : \boldsymbol{H}^{1-s}(U) \to \boldsymbol{H}^{-s}(U)$ is bounded for all $s \in (0, 1)$, $s \neq \frac{1}{2}$. (*Hint*: see (3.7), Theorem 3.19; see also Grisvard [223, Lem. 1.4.4.6].)

Exercise 37.5 (L^2-estimate). (i) Modify the proof of Theorem 37.7 by measuring the interpolation error on the adjoint solution with the operator $\mathcal{I}_h^{\mathrm{g,av}}$ instead of $\mathcal{I}_{h0}^{\mathrm{g,av}}$, i.e., use $Y_h := V_h$ instead of $Y_h := V_h \cap H_0^1(D)$. (*Hint*: set $a_\sharp(v, w) := (\nabla v, \nabla w)_{\boldsymbol{L}^2(D)} - (\boldsymbol{n}{\cdot}\nabla v, w)_{L^2(\partial D)} + \sum_{F \in \mathcal{F}_h^\partial} \varpi_0 \frac{1}{h_F}(v, w)_{L^2(F)}$.) (ii) Do the same for the proof of Theorem 37.8.

Chapter 38

Discontinuous Galerkin

The goal of this chapter is to study the approximation of an elliptic model problem by the discontinuous Galerkin (dG) method. The distinctive feature of dG methods is that the trial and the test spaces are broken finite element spaces (see §18.1.2). Inspired by the boundary penalty method from Chapter 37, dG formulations are obtained by adding a consistency term at all the mesh interfaces and boundary faces, boundary conditions are weakly enforced à la Nitsche, and continuity across the mesh interfaces is weakly enforced by penalizing the jumps. The dG method we study here is called symmetric interior penalty (SIP) because the consistency term is symmetrized to maintain the symmetry of the discrete bilinear form. Incidentally, the symmetry property is important to derive optimal L^2-error estimates assuming full elliptic regularity pickup. We also discuss a useful reformulation of the dG method by lifting the jumps, leading to the important notion of discrete gradient reconstruction.

38.1   Model problem

For simplicity, we focus on the Poisson equation with homogeneous Dirichlet boundary conditions:

$$\begin{cases} \text{Find } u \in V := H_0^1(D) \text{ such that} \\ a(u, w) = \ell(w), \quad \forall w \in V, \end{cases} \tag{38.1}$$

with $a(v, w) := \int_D \nabla v \cdot \nabla w \, dx$, $\ell(w) := \int_D fw \, dx$, $f \in L^2(D)$, and D is a Lipschitz polyhedron in $\mathbb{R}^d$. This problem is well-posed owing to the Lax–Milgram lemma and the Poincaré–Steklov inequality in $H_0^1(D)$. We refer the reader to §41.4 for the more general PDE $-\nabla \cdot (\lambda \nabla u) = f$ with contrasted diffusivity λ.

38.2   Symmetric interior penalty

In this section, we derive the dG approximation of the model problem (38.1) using the SIP method and show that the discrete problem is well-posed.

38.2.1   Discrete problem

Although dG methods can be used on general meshes composed of polyhedral cells, we consider for simplicity a shape-regular sequence $(\mathcal{T}_h)_{h\in\mathcal{H}}$ of affine matching meshes so that each mesh covers D exactly. Let $W^{1,1}(\mathcal{T}_h;\mathbb{R}^q)$, $q \geq 1$, be the broken Sobolev space introduced in Definition 18.1. Recall that every interface $F := \partial K_l \cap \partial K_r \in \mathcal{F}_h^\circ$ is oriented by the fixed unit normal vector $\boldsymbol{n}_F$ pointing from K_l to K_r, i.e., $\boldsymbol{n}_F := \boldsymbol{n}_{K_l} = -\boldsymbol{n}_{K_r}$, and that the jump across F of a function $v \in W^{1,1}(\mathcal{T}_h;\mathbb{R}^q)$ is defined by setting $[\![v]\!]_F := v_{|K_l} - v_{|K_r}$ a.e. on F. We also need the following notion of face average.

Definition 38.1 (Average). *For all* $F := \partial K_l \cap \partial K_r \in \mathcal{F}_h^\circ$, *the* average *of a function* $v \in$ $W^{1,1}(\mathcal{T}_h;\mathbb{R}^q)$ *on* F *is defined as*

$$\{v\}_F := \frac{1}{2}\big(v_{|K_l} + v_{|K_r}\big) \qquad \textit{a.e. on } F. \tag{38.2}$$

As for jumps, the subscript F *is dropped when the context is unambiguous.*

To be more concise, it is customary in the dG literature dedicated to elliptic PDEs to define the jump and the average of a function at the boundary faces by setting $[\![v]\!]_F := \{v\}_F := v_{|K_l}$ a.e. on $F := \partial K_l \cap \partial D \in \mathcal{F}_h^\partial$ (i.e., K_l is the unique mesh cell having the boundary face F among its faces).

Let $(\widehat{K}, \widehat{P}, \widehat{\Sigma})$ be the reference finite element which we assume to be of degree $k \geq 1$. Let us consider the broken finite element space (see (18.4)) s.t.

$$V_h := P_k^{\mathrm{b}}(\mathcal{T}_h) := \{v_h \in L^\infty(D) \mid \psi_K(v_{h|K}) \in \widehat{P}, \forall K \in \mathcal{T}_h\}, \tag{38.3}$$

where $\psi_K(v) := v \circ \boldsymbol{T}_K$ is the pullback by the geometric mapping $\boldsymbol{T}_K$. The approximation setting in dG methods is nonconforming since functions in V_h can jump across the mesh interfaces and can have nonzero boundary values, whereas membership in $V := H_0^1(D)$ requires continuity across the interfaces (see Theorem 18.8) and zero boundary values. Nonconformity implies that we cannot work with the bilinear form a. The construction of the discrete bilinear form a_h on $V_h \times V_h$ is a bit more involved than for the Crouzeix–Raviart finite element method from Chapter 36, where it was sufficient to replace the weak gradient ∇ by the broken gradient ∇_h (see Definition 36.3) to build a_h from a. Instead, the SIP method hinges on the following discrete bilinear form:

$$a_h(v_h, w_h) := \int_D \nabla_h v_h \cdot \nabla_h w_h \, \mathrm{d}x - \sum_{F \in \mathcal{F}_h} \int_F \{\nabla_h v_h\} \cdot \boldsymbol{n}_F [\![w_h]\!] \, \mathrm{d}s$$

$$- \sum_{F \in \mathcal{F}_h} \int_F [\![v_h]\!]\{\nabla_h w_h\} \cdot \boldsymbol{n}_F \, \mathrm{d}s + \sum_{F \in \mathcal{F}_h} \varpi(h_F) \int_F [\![v_h]\!][\![w_h]\!] \, \mathrm{d}s, \tag{38.4}$$

where the second and the fourth terms on the right-hand side are reminiscent of Nitsche's boundary penalty method. The second term is called *consistency term* since it is important to establish consistency/boundedness (see Lemma 38.9). The third term, which is called *adjoint consistency term*, makes the discrete bilinear form a_h symmetric and it is important to establish an improved L^2-error estimate (see Theorem 38.12). The fourth term is important to establish coercivity (see Lemma 38.6). It penalizes jumps across interfaces and values at boundary faces and is, therefore, called *penalty term*. Coercivity requires that the penalty parameter be s.t. $\varpi(h_F) := \varpi_0 h_F^{-1}$, where $\varpi_0 > 0$ has to be chosen large enough, and on shape-regular mesh sequences, the local length scale h_F can be taken to be the diameter of F.

We consider the following discrete problem:

$$\begin{cases} \text{Find } u_h \in V_h \text{ such that} \\ a_h(u_h, w_h) = \ell_h(w_h), \quad \forall w_h \in V_h, \end{cases} \tag{38.5}$$

where the discrete linear form is given by

$$\ell_h(w_h) := \int_D f w_h \, \mathrm{d}x, \qquad \forall w_h \in V_h. \tag{38.6}$$

This choice for ℓ_h is possible since the source term in the model problem (38.1) is assumed to be in $L^2(D)$. A more general setting, e.g., $f \in H^{-1}(D)$, is discussed in Remark 36.5. Furthermore, it is legitimate to extend a_h to $(H^{1+r}(D) + V_h) \times V_h$, $r > \frac{1}{2}$, since $\nabla u \in \boldsymbol{H}^r(D)$ implies that $(\nabla u)_{|F}$ is well defined as an integrable function for all $F \in \mathcal{F}_h$. To motivate the appearance of the consistency term in the definition of a_h, let us prove the following important result.

Lemma 38.2 (Consistency term). *Assume that $u \in H^{1+r}(D)$, $r > \frac{1}{2}$. Then we have $a_h(u, w_h) = \ell_h(w_h)$ for all $w_h \in V_h$.*

Proof. We have $[\![u]\!]_F = 0$ a.e. on all $F \in \mathcal{F}_h$ (use Theorem 18.8 for $F \in \mathcal{F}_h^\circ$ and $\gamma^{\mathrm{g}}(u) = 0$ for $F \in \mathcal{F}_h^\partial$) and $\nabla_h u = \nabla u$ (see Lemma 18.9). Since $\nabla u \in \boldsymbol{H}^r(D)$, $r > \frac{1}{2}$, we also have $[\![\nabla u]\!] \cdot \boldsymbol{n}_F = 0$ a.e. on all $F \in \mathcal{F}_h^\circ$ (see Remark 18.4). We infer that

$$a_h(u, w_h) = \int_D \nabla u \cdot \nabla_h w_h \, \mathrm{d}x - \sum_{F \in \mathcal{F}_h} \int_F (\nabla u \cdot \boldsymbol{n}_F) [\![w_h]\!] \, \mathrm{d}s.$$

We conclude by performing elementwise integration by parts as follows:

$$\int_D \nabla u \cdot \nabla_h w_h \, \mathrm{d}x = \int_D -(\Delta u) w_h \, \mathrm{d}x + \sum_{K \in \mathcal{T}_h} \sum_{F \in \mathcal{F}_K} \int_F (\nabla u \cdot \boldsymbol{n}_K) w_{h|K} \, \mathrm{d}s$$

$$= \ell_h(w_h) + \sum_{F \in \mathcal{F}_h} \int_F (\nabla u \cdot \boldsymbol{n}_F) [\![w_h]\!] \, \mathrm{d}s. \qquad \square$$

Remark 38.3 (Literature). The SIP approximation has been analyzed in Arnold [15] (see also Baker [44], Wheeler [394]). $\qquad \square$

Remark 38.4 (Nonmatching meshes). It is possible to consider nonmatching meshes if the diameter of each interface $F \in \mathcal{F}_h^\circ$ is uniformly equivalent to the diameter of the two cells sharing F. $\qquad \square$

38.2.2 Coercivity and well-posedness

We equip the space V_h with the following norm:

$$\|v_h\|_{V_h}^2 := \|\nabla_h v_h\|_{\boldsymbol{L}^2(D)}^2 + |v_h|_{\mathrm{J}}^2, \qquad |v_h|_{\mathrm{J}}^2 := \sum_{F \in \mathcal{F}_h} \frac{1}{h_F} \|[\![v_h]\!]\|_{L^2(F)}^2. \tag{38.7}$$

That $\|\cdot\|_{V_h}$ is a norm on V_h (and not just a seminorm) can be verified directly: If $\|v_h\|_{V_h} = 0$, then v_h is piecewise constant and $[\![v_h]\!]_F = 0$ for all $F \in \mathcal{F}_h$. This means that v_h is constant on D and vanishes at ∂D, so that $v_h = 0$. Our first step in the analysis is to bound from above the consistency term. Recall that $\mathcal{T}_F := \{K \in \mathcal{T}_h \mid F \in \mathcal{F}_K\}$ is the collection of the mesh cells having F as face. Let $|\mathcal{T}_F|$ denote the cardinality of the set $\mathcal{T}_F$ ($|\mathcal{T}_F| = 2$ for all $F \in \mathcal{F}_h^\circ$ and $|\mathcal{T}_F| = 1$ for all $F \in \mathcal{F}_h^\partial$).

Lemma 38.5 (Consistency term). *Let us set for all* $(v_h, w_h) \in V_h \times V_h$,

$$n_h(v_h, w_h) := - \sum_{F \in \mathcal{F}_h} \int_F \{\nabla_h v_h\} \cdot \boldsymbol{n}_F [\![w_h]\!] \, ds. \tag{38.8}$$

Then the following holds true for all $v_h \in V_h$:

$$\sup_{w_h \in V_h} \frac{|n_h(v_h, w_h)|}{|w_h|_\mathrm{J}} \leq \left(\sum_{F \in \mathcal{F}_h} \frac{1}{|\mathcal{T}_F|} \sum_{K \in \mathcal{T}_F} h_F \|\boldsymbol{n}_F \cdot \nabla(v_{h|K})\|^2_{L^2(F)} \right)^{\frac{1}{2}}. \tag{38.9}$$

Proof. The Cauchy–Schwarz inequality leads to

$$|n_h(v_h, w_h)| \leq \sum_{F \in \mathcal{F}_h} h_F^{\frac{1}{2}} \|\boldsymbol{n}_F \cdot \{\nabla_h v_h\}\|_{L^2(F)} \times h_F^{-\frac{1}{2}} \|[\![w_h]\!]\|_{L^2(F)}$$

$$\leq \left(\sum_{F \in \mathcal{F}_h} h_F \|\boldsymbol{n}_F \cdot \{\nabla_h v_h\}\|^2_{L^2(F)} \right)^{\frac{1}{2}} |w_h|_\mathrm{J},$$

Letting $\boldsymbol{g}_h := \nabla_h v_h$, (38.9) follows from $\{\boldsymbol{g}_h\}_F = \frac{1}{|\mathcal{T}_F|} \sum_{K \in \mathcal{T}_F} \boldsymbol{g}_{h|K}$ and

$$\|\boldsymbol{n}_F \cdot \{\boldsymbol{g}_h\}\|^2_{L^2(F)} = \frac{1}{|\mathcal{T}_F|^2} \left\| \sum_{K \in \mathcal{T}_F} \boldsymbol{n}_F \cdot \boldsymbol{g}_{h|K} \right\|^2_{L^2(F)} \leq \frac{1}{|\mathcal{T}_F|} \sum_{K \in \mathcal{T}_F} \|\boldsymbol{n}_F \cdot \boldsymbol{g}_{h|K}\|^2_{L^2(F)}. \qquad \square$$

We shall use the same discrete trace inequality as in Chapter 37 to prove a coercivity property. Let c_dt be the smallest constant such that

$$\|\boldsymbol{n}_F \cdot \nabla_h w_{h|K}\|_{L^2(F)} \leq c_\mathrm{dt} h_F^{-\frac{1}{2}} \|\nabla_h w_h\|_{\boldsymbol{L}^2(K)}, \tag{38.10}$$

for all $w_h \in V_h$, all $K \in \mathcal{T}_h$, all $F \in \mathcal{F}_K$, and all $h \in \mathcal{H}$. Let $n_\partial := \max_{K \in \mathcal{T}_h} |\mathcal{F}_K|$ be the largest number of faces per mesh cell, i.e., $n_\partial \leq d + 1$ for simplicial meshes (the definition of n_∂ differs from that of Chapter 37).

Lemma 38.6 (Coercivity, well-posedness). *Let the penalty parameter be s.t.* $\varpi(h_F) := \varpi_0 h_F^{-1}$ *with* $\varpi_0 > n_\partial c_\mathrm{dt}^2$. (i) *We have*

$$a_h(v_h, v_h) \geq \alpha \|v_h\|^2_{V_h}, \qquad \forall v_h \in V_h, \tag{38.11}$$

with $\alpha := \frac{\varpi_0 - n_\partial c_\mathrm{dt}^2}{1 + \varpi_0} > 0$. (ii) *The discrete problem* (38.5) *is well-posed.*

Proof. Let $v_h \in V_h$. Our starting observation is that

$$a_h(v_h, v_h) = \|\nabla_h v_h\|^2_{\boldsymbol{L}^2(D)} + 2n_h(v_h, v_h) + \varpi_0 |v_h|^2_\mathrm{J}.$$

Using (38.9) and (38.10), we infer that

$$|n_h(v_h, v_h)| \leq \left(\sup_{w_h \in V_h} \frac{|n_h(v_h, w_h)|}{|w_h|_\mathrm{J}} \right) |v_h|_\mathrm{J} \leq n_\partial^{\frac{1}{2}} c_\mathrm{dt} \|\nabla_h v_h\|_{\boldsymbol{L}^2(D)} |v_h|_\mathrm{J},$$

since $|\mathcal{T}_F| \geq 1$, $\sum_{F \in \mathcal{F}_h} \sum_{K \in \mathcal{T}_F}(\cdot) = \sum_{K \in \mathcal{T}_h} \sum_{F \in \mathcal{F}_K}(\cdot)$, and $|\mathcal{F}_K| \leq n_\partial$, so that

$$\sum_{F \in \mathcal{F}_h} \sum_{K \in \mathcal{T}_F} \|\boldsymbol{g}_{h|K}\|_{\boldsymbol{L}^2(K)}^2 \leq n_\partial \sum_{K \in \mathcal{T}_h} \|\boldsymbol{g}_{h|K}\|_{\boldsymbol{L}^2(K)}^2 = n_\partial \|\boldsymbol{g}_h\|_{\boldsymbol{L}^2(D)}^2$$

with $\boldsymbol{g}_h := \nabla_h v_h$. This leads to the lower bound

$$a_h(v_h, v_h) \geq \|\nabla_h v_h\|_{\boldsymbol{L}^2(D)}^2 - 2 n_\partial^{\frac{1}{2}} c_{\mathrm{dt}} \|\nabla_h v_h\|_{\boldsymbol{L}^2(D)} |v_h|_{\mathrm{J}} + \varpi_0 |v_h|_{\mathrm{J}}^2,$$

whence we infer the coercivity property (38.11) by using the quadratic inequality from Exercise 37.2. Finally, the well-posedness of (38.5) follows from the Lax–Milgram lemma. □

Remark 38.7 (Penalty parameter). As in the boundary penalty method from Chapter 37, one needs a (reasonable) upper bound on the constant c_{dt} to choose a value of ϖ_0 that guarantees coercivity. The results of §12.2 show that c_{dt} scales essentially as k^2. An alternative penalty strategy allowing for an easy-to-compute value of ϖ_0 is discussed in Remark 38.17, but this technique requires local inversions of small mass matrices. □

Remark 38.8 (Discrete Sobolev inequality). Let ℓ_D be a length scale associated with D, e.g., $\ell_D := \mathrm{diam}(D)$. One can show that there is $C_{\mathrm{SOB}} > 0$ such that $C_{\mathrm{SOB}} \|v_h\|_{L^q(D)} \leq \ell_D \|v_h\|_{V_h}$ for all $v_h \in V_h$, all $h \in \mathcal{H}$, and all $q \in [1, \infty)$ if $d = 2$ and $q \in [1, \frac{2d}{d-2}]$ if $d \geq 3$; see Buffa and Ortner [95], Di Pietro and Ern [164]. The reader is referred to Arnold [15], Brenner [86] for similar estimates in broken Hilbert Sobolev spaces ($q = 2$). □

38.2.3 Variations on boundary conditions

The non-homogeneous Dirichlet boundary condition $u = g$ on ∂D with $g \in H^{\frac{1}{2}}(\partial D)$ is discretized by modifying the right-hand side in (38.5) as follows:

$$\ell_h^{\mathrm{nD}}(w_h) := \ell(w_h) - \sum_{F \in \mathcal{F}_h^\partial} \int_F g(\boldsymbol{n}_F \cdot \nabla_h w_h - \varpi(h_F) w_h) \, \mathrm{d}s. \tag{38.12}$$

For the Robin boundary condition $\gamma u + \boldsymbol{n} \cdot \nabla u = g$ on ∂D with $g \in L^2(\partial D)$ and $\gamma \in L^\infty(\partial D)$ taking nonnegative values on ∂D ($\gamma := 0$ corresponds to the Neumann problem), the discrete bilinear form and the right-hand side become

$$a_h^{\mathrm{Rb}}(v_h, w_h) := \int_D \nabla_h v_h \cdot \nabla_h w_h \, \mathrm{d}x - \sum_{F \in \mathcal{F}_h^\circ} \int_F \{\nabla_h v_h\} \cdot \boldsymbol{n}_F [\![w_h]\!] \, \mathrm{d}s \tag{38.13a}$$

$$- \sum_{F \in \mathcal{F}_h^\circ} \int_F [\![v_h]\!] \{\nabla_h w_h\} \cdot \boldsymbol{n}_F \, \mathrm{d}s + \sum_{F \in \mathcal{F}_h^\circ} \varpi(h_F) \int_F [\![v_h]\!] [\![w_h]\!] \, \mathrm{d}s + \sum_{F \in \mathcal{F}_h^\partial} \int_F \gamma v_h w_h \, \mathrm{d}s,$$

$$\ell_h^{\mathrm{Rb}}(w_h) := \ell(w_h) + \sum_{F \in \mathcal{F}_h^\partial} \int_F g w_h \, \mathrm{d}s. \tag{38.13b}$$

One can verify that Lemma 38.2 still holds true in both cases.

38.3   Error analysis

In this section, we derive an energy error estimate, that is, we bound the error by using the coercivity norm and the abstract error estimate from Lemma 27.5. We also derive an improved L^2-error estimate by means of a duality argument. We assume that $u \in V_{\mathrm{s}}$ with

$$V_{\mathrm{s}} := H^{1+r}(D) \cap H_0^1(D), \qquad r > \frac{1}{2}. \tag{38.14}$$

The assumption $u \in V_{\mathrm{s}}$ is reasonable in the setting of the Poisson equation with Dirichlet conditions in a Lipschitz polyhedron since it is consistent with the elliptic regularity theory (see Theorem 31.33). The important property that we use is that for any function $v \in V_{\mathrm{s}}$ the normal derivative $\boldsymbol{n}_K \cdot \nabla v$ is meaningful in $L^2(\partial K)$ for all $K \in \mathcal{T}_h$. Recall that the discrete space is $V_h := P_k^{\mathrm{b}}(\mathcal{T}_h)$ equipped with the $\|\cdot\|_{V_h}$-norm defined in (38.7). We set $V_\sharp := V_{\mathrm{s}} + V_h$ and we equip this space with the norm

$$\|v\|_{V_\sharp}^2 := \|\nabla_h v\|_{\boldsymbol{L}^2(D)}^2 + |v|_{\mathrm{J}}^2 + \sum_{K \in \mathcal{T}_h} h_K \|\boldsymbol{n}_K \cdot \nabla v_{|K}\|_{L^2(\partial K)}^2, \tag{38.15}$$

with $|v|_{\mathrm{J}}^2 := \sum_{F \in \mathcal{F}_h} \frac{1}{h_F} \|[\![v]\!]\|_{L^2(F)}^2$. A discrete trace inequality shows that there is $c_\sharp$ s.t. $\|v_h\|_{V_\sharp} \leq c_\sharp \|v_h\|_{V_h}$ for all $v_h \in V_h$ and all $h \in \mathcal{H}$, i.e., (27.5) holds true. Using the discrete bilinear forms a_h and ℓ_h defined in (38.4) and (38.6), respectively, the consistency error is s.t. $\langle \delta_h(v_h), w_h \rangle_{V_h', V_h} := \ell_h(w_h) - a_h(v_h, w_h)$ for all $v_h, w_h \in V_h$.

Lemma 38.9 (Consistency/boundedness). *Assume* (38.14). *There is $\omega_\sharp$, uniform w.r.t. $u \in V_{\mathrm{s}}$, s.t. for all $v_h \in V_h$ and all $h \in \mathcal{H}$,*

$$\|\delta_h(v_h)\|_{V_h'} \leq \omega_\sharp \|u - v_h\|_{V_\sharp}. \tag{38.16}$$

Proof. Let $v_h \in V_h$ and let us set $\eta := u - v_h$. Owing to Lemma 38.2 and since $[\![u]\!]_F = 0$ for all $F \in \mathcal{F}_h$, we infer that for all $w_h \in W_h$,

$$\langle \delta_h(v_h), w_h \rangle_{V_h', V_h} = \int_D \nabla_h \eta \cdot \nabla_h w_h \, \mathrm{d}x + n_\sharp(\eta, w_h)$$

$$- \sum_{F \in \mathcal{F}_h} \int_F [\![\eta]\!] \{\nabla_h w_h\} \cdot \boldsymbol{n}_F \, \mathrm{d}s + \sum_{F \in \mathcal{F}_h} \frac{\varpi_0}{h_F} \int_F [\![\eta]\!] [\![w_h]\!] \, \mathrm{d}s,$$

where $n_\sharp(v, w_h) := -\sum_{F \in \mathcal{F}_h} \int_F \{\nabla_h v\} \cdot \boldsymbol{n}_F [\![w_h]\!] \, \mathrm{d}s$ is understood as an extension to $V_\sharp \times V_h$ of the discrete bilinear form n_h originally defined on $V_h \times V_h$ by (38.8). (Note that the assumption $r > \frac{1}{2}$ in the definition of V_{s} is crucial for this extension to make sense.) The Cauchy–Schwarz inequality implies that

$$\left| \int_D \nabla_h \eta \cdot \nabla_h w_h \, \mathrm{d}x + \sum_{F \in \mathcal{F}_h} \frac{\varpi_0}{h_F} \int_F [\![\eta]\!] [\![w_h]\!] \, \mathrm{d}s \right|$$

$$\leq \|\nabla_h \eta\|_{\boldsymbol{L}^2(D)} \|\nabla_h w_h\|_{\boldsymbol{L}^2(D)} + \varpi_0 |\eta|_{\mathrm{J}} |w_h|_{\mathrm{J}} \leq \max(1, \varpi_0) \|\eta\|_{V_\sharp} \|w_h\|_{V_h}.$$

Since the bound (38.9) is still valid for $n_\sharp(\eta, w_h)$, we also have

$$|n_\sharp(\eta, w_h)| \leq \left(\sum_{F \in \mathcal{F}_h} \frac{1}{|\mathcal{T}_F|} \sum_{K \in \mathcal{T}_F} h_F \|\boldsymbol{n}_F \cdot \nabla(\eta_{|K})\|_{L^2(F)}^2 \right)^{\frac{1}{2}} |w_h|_{\mathrm{J}}$$

$$\leq c \|\eta\|_{V_\sharp} |w_h|_{\mathrm{J}} \leq c \|\eta\|_{V_\sharp} \|w_h\|_{V_h}.$$

(This is where we use the contribution of the normal derivative to the $\|\cdot\|_{V_\sharp}$-norm.) Proceeding as in the proof of Lemma 38.5, we finally infer that

$$\left| \sum_{F \in \mathcal{F}_h} \int_F [\![\eta]\!] \{\nabla_h w_h\} \cdot \boldsymbol{n}_F \, \mathrm{d}s \right| \leq |\eta|_{\mathrm{J}} \left(\sum_{F \in \mathcal{F}_h} \frac{1}{|\mathcal{T}_F|} \sum_{K \in \mathcal{T}_F} h_F \|\nabla(w_{h|K})\|^2_{\boldsymbol{L}^2(F)} \right)^{\frac{1}{2}}$$

$$\leq n_\partial^{\frac{1}{2}} c_{\mathrm{dt}} |\eta|_{\mathrm{J}} \|\nabla_h w_h\|_{\boldsymbol{L}^2(D)} \leq n_\partial^{\frac{1}{2}} c_{\mathrm{dt}} \|\eta\|_{V_\sharp} \|w_h\|_{V_h},$$

where we used the discrete trace inequality (38.10) as in the proof of Lemma 38.6. Collecting the above bounds shows that $|\langle \delta_h(v_h), w_h \rangle_{V_h', V_h}| \leq c\|\eta\|_{V_\sharp} \|w_h\|_{V_h}$, i.e., (38.16) holds true. $\square$

Theorem 38.10 (Convergence). *Let u solve (38.1) and let u_h solve (38.5) with the penalty parameter $\varpi_0 > c_{\mathrm{dt}}^2 n_\partial$. Assume (38.14). (i) There is c s.t. the following holds true for all $h \in \mathcal{H}$:*

$$\|u - u_h\|_{V_\sharp} \leq c \inf_{v_h \in V_h} \|u - v_h\|_{V_\sharp}. \tag{38.17}$$

(ii) *Letting $t := \min(k, r)$, we have*

$$\|u - u_h\|_{V_\sharp} \leq c \left(\sum_{K \in \mathcal{T}_h} h_K^{2t} |u|^2_{H^{1+t}(K)} \right)^{\frac{1}{2}}. \tag{38.18}$$

Proof. (i) The estimate (38.17) follows from Lemma 27.5 combined with stability (Lemma 38.6) and consistency/boundedness (Lemma 38.9).

(ii) We bound the infimum in (38.17) by taking $v_h := \mathcal{I}_h^\sharp(u)$, where $\mathcal{I}_h^\sharp : L^1(D) \to P_k^{\mathrm{b}}(\mathcal{T}_h)$ is the L^1-stable interpolation operator from §18.3. We need to bound $\|\nabla(\eta_{|K})\|_{\boldsymbol{L}^2(K)} + h_K^{\frac{1}{2}} \|\nabla(\eta_{|K})\|_{L^2(\partial K)}$ for all $K \in \mathcal{T}_h$ and $h_F^{-\frac{1}{2}} \|[\![\eta]\!]_F\|_{L^2(F)}$ for all $F \in \mathcal{F}_h$, with $\eta := u - \mathcal{I}_h^\sharp(u)$. Theorem 18.14 implies that $\|\nabla(\eta_{|K})\|_{\boldsymbol{L}^2(K)} \leq ch_K^t |u|_{H^{1+t}(K)}$. Moreover, Corollary 18.15 implies that $h_K^{\frac{1}{2}} \|\nabla(\eta_{|K})\|_{L^2(\partial K)} \leq ch_K^t |u|_{H^{1+t}(K)}$ and that $\|\eta_{|K}\|_{L^2(F)} \leq ch_K^{t+\frac{1}{2}} |u|_{H^{1+t}(K)}$ for any face $F \in \mathcal{F}_K$. Since $[\![\eta]\!]_F := \eta_{|K_l}$ for all $F := \partial K_l \cap \partial D \in \mathcal{F}_h^\partial$ and $[\![\eta]\!]_F := \eta_{|K_l} - \eta_{|K_r}$ for all $F := \partial K_l \cap \partial K_r \in \mathcal{F}_h^\circ$, we can use the shape-regularity of the mesh sequence and the triangle inequality for the jump to infer that $h_F^{-\frac{1}{2}} \|[\![\eta]\!]_F\|_{L^2(F)} \leq c \sum_{K \in \mathcal{T}_F} h_K^t |u|_{H^{1+t}(K)}$ for all $F \in \mathcal{F}_h$. This leads to (38.18). $\square$

Remark 38.11 (L^2-orthogonal projection). Note that, as shown in Remark 18.18, $\mathcal{I}_h^\sharp$ is the L^2-orthogonal projection onto $P_k^{\mathrm{b}}(\mathcal{T}_h)$ since ψ_K is the pullback by the geometric mapping $\boldsymbol{T}_K$. $\square$

We now derive an L^2-error estimate by invoking a *duality argument* as in §36.3.3. For all $g \in L^2(D)$, we consider the adjoint solution $\zeta_g \in V := H_0^1(D)$ s.t. $a(v, \zeta_g) = (v, g)_{L^2(D)}$ for all $v \in V$, i.e., $-\Delta \zeta_g = g$ in D and $\gamma^{\mathrm{g}}(\zeta_g) = 0$. Owing to the elliptic regularity theory (see §31.4), there is $s \in (0, 1]$ and a constant c_{smo} such that $\|\zeta_g\|_{H^{1+s}(D)} \leq c_{\mathrm{smo}} \ell_D^2 \|g\|_{L^2(D)}$ for all $g \in L^2(D)$. In the present setting of the Poisson equation with Dirichlet conditions in a Lipschitz polyhedron, it is reasonable to assume that $s \in (\frac{1}{2}, 1]$.

Theorem 38.12 (L^2-estimate). *Under the assumptions of Theorem 38.10 and assuming that the elliptic regularity index satisfies $s \in (\frac{1}{2}, 1]$, there is c such that for all $h \in \mathcal{H}$,*

$$\|u - u_h\|_{L^2(D)} \leq c\, h^s \ell_D^{1-s} \|u - u_h\|_{V_\sharp}. \tag{38.19}$$

Proof. Apply Lemma 36.14 and use exact adjoint consistency; see Exercise 38.3. $\square$

Remark 38.13 (Variations on symmetry). Let us set

$$a_h(v_h, w_h) := \int_D \nabla_h v_h \cdot \nabla_h w_h \, \mathrm{d}x - \sum_{F \in \mathcal{F}_h} \int_F \{\nabla_h v_h\} \cdot \boldsymbol{n}_F [\![w_h]\!] \, \mathrm{d}s$$

$$- \theta \sum_{F \in \mathcal{F}_h} \int_F \{\nabla_h w_h\} \cdot \boldsymbol{n}_F [\![v_h]\!] \, \mathrm{d}s + \sum_{F \in \mathcal{F}_h} \varpi(h_F) \int_F [\![v_h]\!] [\![w_h]\!] \, \mathrm{d}s, \qquad (38.20)$$

where θ is a real number ($\theta := 1$ corresponds to the SIP formulation). The choice $\theta := -1$ gives the method usually called nonsymmetric interior penalty (NIP). This choice is interesting since it simplifies the analysis of the coercivity in that the consistency term cancels with the added nonsymmetric term. The original idea can be traced back to the method in Oden et al. [318], where the nonsymmetric method is introduced without the penalty term. The convergence analysis when the penalty term is included can be found in Rivière et al. [335, 336], where it is shown that coercivity only requires $\varpi_0 > 0$; see also Larson and Niklasson [274] for the inf-sup stability analysis. The incomplete interior penalty (IIP) method corresponds to the choice $\theta := 0$. Similarly to SIP, a minimal threshold on the penalty parameter ϖ_0 is required for the coercivity; see Dawson et al. [157]. Whenever $\theta \neq 1$, the analysis of the L^2-error estimate proceeds as in §37.3.2 (accounting for an adjoint consistency error), and one only obtains $\|u - u_h\|_{L^2(D)} \leq ch^{\frac{1}{2}} \|u - u_h\|_{V_\sharp}$ even if full elliptic regularity holds true ($s = 1$). $\qquad \square$

Remark 38.14 (L^∞-estimates). Pointwise dG error estimates are found in Kanschat and Rannacher [264], Chen and Chen [117], Guzmán [233]. $\qquad \square$

38.4   Discrete gradient and fluxes

In this section, we introduce the notion of discrete gradient and use it to derive an alternative viewpoint on the SIP bilinear form. One interesting outcome is a reformulation of the discrete problem (38.5) in terms of local problems with numerical fluxes.

38.4.1   Liftings

Loosely speaking the discrete gradient consists of the broken gradient plus a correction associated with the jumps. This correction is formulated in terms of local liftings introduced in Bassi and Rebay [46] and analyzed in Brezzi et al. [93] (see also Perugia and Schötzau [323] for the hp-analysis). Let $F \in \mathcal{F}_h$ and an integer $l \geq 0$. Consider the local *lifting operator* $\boldsymbol{\mathcal{L}}_F^l : L^2(F) \to \boldsymbol{P}_l^{\mathrm{b}}(\mathcal{T}_h) := P_l^{\mathrm{b}}(\mathcal{T}_h; \mathbb{R}^d)$ s.t. for all $\varphi \in L^2(F)$, the discrete function $\boldsymbol{\mathcal{L}}_F^l(\varphi)$ is defined as

$$\int_D \boldsymbol{\mathcal{L}}_F^l(\varphi) \cdot \boldsymbol{\tau}_h \, \mathrm{d}x := \int_F \{\boldsymbol{\tau}_h\} \cdot \boldsymbol{n}_F \varphi \, \mathrm{d}s, \qquad \forall \boldsymbol{\tau}_h \in \boldsymbol{P}_l^{\mathrm{b}}(\mathcal{T}_h). \qquad (38.21)$$

By localizing the support of $\boldsymbol{\tau}_h$ to a single mesh cell, we infer that $\boldsymbol{\mathcal{L}}_F^l(\varphi)$ is collinear to $\boldsymbol{n}_F$ and is supported in the set $D_F := \mathrm{int}(\bigcup_{K \in \mathcal{T}_F} K)$. In practice, the Cartesian components of the polynomial function $\boldsymbol{\mathcal{L}}_F^l(\varphi)$ can be computed in each $K \in \mathcal{T}_F$ by inverting the local mass matrix with entries $\mathcal{M}_{K,ij} := \int_K \theta_{K,i} \theta_{K,j} \, \mathrm{d}x$, where the functions $\theta_{K,i}$ are the local shape functions in K.

Consider now a function $v \in H^1(\mathcal{T}_h)$. We define the global lifting of the jumps of v as follows:

$$\boldsymbol{\mathcal{L}}_h^l([\![v]\!]) := \sum_{F \in \mathcal{F}_h} \boldsymbol{\mathcal{L}}_F^l([\![v]\!]).$$

This makes sense since $[\![v]\!]_F \in L^2(F)$ for all $F \in \mathcal{F}_h$. A consequence of $\mathrm{supp}(\mathcal{L}_F^l([\![v]\!])) = D_F$ is that $\mathcal{L}_h^l([\![v]\!])_{|K} := \sum_{F \in \mathcal{F}_K} \mathcal{L}_F^l([\![v]\!])$ for all $K \in \mathcal{T}_h$, i.e., only the jumps across the faces of K contribute to the restriction to K of the global lifting $\mathcal{L}_h^l([\![v]\!])$.

Lemma 38.15 (Stability). *The following holds true for all $l \geq 0$:*

$$\|\mathcal{L}_F^l(\varphi)\|_{\boldsymbol{L}^2(D_F)} \leq c_{\mathrm{dt}} h_F^{-\frac{1}{2}} \|\varphi\|_{L^2(F)}, \qquad \forall \varphi \in L^2(F),\ \forall F \in \mathcal{F}_h, \tag{38.22a}$$

$$\|\mathcal{L}_h^l([\![v]\!])\|_{\boldsymbol{L}^2(D)} \leq n_\partial^{\frac{1}{2}} c_{\mathrm{dt}} |v|_{\mathrm{J}}, \qquad \forall v \in H^1(\mathcal{T}_h), \tag{38.22b}$$

where c_{dt} is the constant from the discrete trace inequality (38.10).

Proof. The proof of (38.22a) is proposed in Exercise 38.4. To prove (38.22b), we use the Cauchy–Schwarz inequality and the definition of n_∂ to infer that

$$\|\mathcal{L}_h^l([\![v]\!])\|_{\boldsymbol{L}^2(K)}^2 = \int_K \left| \sum_{F \in \mathcal{F}_K} \mathcal{L}_F^l([\![v]\!]) \right|^2 \mathrm{d}x \leq n_\partial \sum_{F \in \mathcal{F}_K} \|\mathcal{L}_F^l([\![v]\!])\|_{\boldsymbol{L}^2(K)}^2,$$

for all $K \in \mathcal{T}_h$. Summing over the mesh cells, recalling that the support of $\mathcal{L}_F^l([\![v]\!])$ is D_F, and using (38.22a) yields (38.22b). $\quad\square$

Definition 38.16 (Discrete gradient). *Let $l \geq 0$. The* discrete gradient operator $\mathfrak{G}_h^l : H^1(\mathcal{T}_h) \to \boldsymbol{L}^2(D)$ *is defined as follows:*

$$\mathfrak{G}_h^l(v) := \nabla_h v - \mathcal{L}_h^l([\![v]\!]), \qquad \forall v \in H^1(\mathcal{T}_h). \tag{38.23}$$

We can now use Definition 38.16 to derive alternative expressions for the SIP bilinear form a_h defined in (38.4). Recalling that $V_h := P_k^{\mathrm{b}}(\mathcal{T}_h)$, $k \geq 1$, we choose the polynomial degree of the liftings such that $l \in \{k-1, k\}$. Since $\nabla_h v_h, \nabla_h w_h \in \mathbb{P}_{k-1}^{\mathrm{b}}(\mathcal{T}_h) \subset \mathbb{P}_l^{\mathrm{b}}(\mathcal{T}_h)$ for all $v_h, w_h \in V_h$, we infer that

$$\int_D \nabla_h v_h \cdot \nabla_h w_h \, \mathrm{d}x - \sum_{F \in \mathcal{F}_h} \int_F \left(\{\nabla_h v_h\} \cdot \boldsymbol{n}_F [\![w_h]\!] + [\![v_h]\!] \{\nabla_h w_h\} \cdot \boldsymbol{n}_F \right) \mathrm{d}s$$
$$= \int_D \mathfrak{G}_h^l(v_h) \cdot \mathfrak{G}_h^l(w_h) \, \mathrm{d}x - \int_D \mathcal{L}_h^l([\![v_h]\!]) \cdot \mathcal{L}_h^l([\![w_h]\!]) \, \mathrm{d}x. \tag{38.24}$$

Recalling the expression (38.4) of a_h, we obtain

$$a_h(v_h, w_h) := \int_D \mathfrak{G}_h^l(v_h) \cdot \mathfrak{G}_h^l(w_h) \, \mathrm{d}x + \tilde{s}_h(v_h, w_h), \tag{38.25}$$

with $\tilde{s}_h(v_h, w_h) := \sum_{F \in \mathcal{F}_h} \varpi(h_F) \int_F [\![v_h]\!] [\![w_h]\!] \, \mathrm{d}s - \int_D \mathcal{L}_h^l([\![v_h]\!]) \cdot \mathcal{L}_h^l([\![w_h]\!]) \, \mathrm{d}x$. The estimate (38.22b) from Lemma 38.15 implies that

$$a_h(v_h, v_h) \geq \|\mathfrak{G}_h^l(v_h)\|_{\boldsymbol{L}^2(D)}^2 + (\varpi_0 - n_\partial c_{\mathrm{dt}}^2) |v_h|_{\mathrm{J}}^2, \tag{38.26}$$

for all $v_h \in V_h$, showing again the relevance of the condition $\varpi_0 > n_\partial c_{\mathrm{dt}}^2$ for coercivity (see Lemma 38.6).

Remark 38.17 (Alternative penalty strategy). It is possible to penalize the liftings of the jumps instead of the jumps, leading to the following modification of the SIP bilinear form:

$$\check{a}_h(v_h, w_h) := \int_D \nabla_h v_h \cdot \nabla_h w_h \, dx - \sum_{F \in \mathcal{F}_h} \int_F \{\nabla_h v_h\} \cdot \boldsymbol{n}_F [\![w_h]\!] \, ds$$

$$- \sum_{F \in \mathcal{F}_h} \int_F [\![v_h]\!] \{\nabla_h w_h\} \cdot \boldsymbol{n}_F \, ds + \sum_{F \in \mathcal{F}_h} \varpi_0 \int_D \boldsymbol{\mathcal{L}}_F^l([\![v_h]\!]) \cdot \boldsymbol{\mathcal{L}}_F^l([\![w_h]\!]) \, dx.$$

The main advantage of this formulation is that coercivity holds true as soon as $\varpi_0 > n_\partial$, thereby avoiding the constant c_{dt} from (38.10). However, the discretization stencil is larger since the dofs in two cells $K, K' \in \mathcal{T}_h$ are coupled if there is $K'' \in \mathcal{T}_h$ s.t. $\partial K \cap \partial K'' \in \mathcal{F}_h^\circ$ and $\partial K' \cap \partial K'' \in \mathcal{F}_h^\circ$ (for the usual penalty strategy the coupling condition is $\partial K \cap \partial K' \in \mathcal{F}_h^\circ$). □

Remark 38.18 (Choosing l). The computation of the discrete gradient can be done with any $l \geq k - 1$. The minimal choice is $l = k - 1$, but choosing $l = k$ may be more interesting from the implementation point of view since it does not require the user to construct the finite element space $P_{k-1}^{\mathrm{b}}(\mathcal{T}_h)$. □

Remark 38.19 (Literature). The discrete gradient is an important notion in the design and analysis of dG methods for nonlinear problems. We refer the reader to Ten Eyck and Lew [364] for nonlinear mechanics, to Burman and Ern [100], Buffa and Ortner [95] for Leray–Lions operators, and to Di Pietro and Ern [164] for the incompressible Navier–Stokes equations. Moreover, an important stability result established in John et al. [260] is that if $l = k + 1$, then there is c s.t. $\|v_h\|_{V_h} \leq c \|\mathfrak{G}_h^l(v_h)\|_{L^2(D)}$ for all $v_h \in V_h$ and all $h \in \mathcal{H}$. Since the proof of this result invokes Raviart–Thomas functions, simplicial meshes are required, but hanging nodes are still allowed under some assumptions. An interesting consequence of this stability result is that for $l = k + 1$, replacing a_h defined in (38.4) by $\tilde{a}_h(v_h, w_h) := \int_D \mathfrak{G}_h^l(v_h) \cdot \mathfrak{G}_h^l(w_h) \, dx$ gives a stable and optimally convergent dG discretization without any penalty parameters. Notice that $\tilde{a}_h$ does not deliver exact consistency because liftings are discrete objects; see Exercise 38.6. For the same reason, the bilinear form a_h defined in (38.4) coincides with the right-hand side of (38.25) on $V_h \times V_h$, but the two sides of the equality produce different results on $V_\sharp \times V_h$. □

38.4.2 Local formulation with fluxes

Let $K \in \mathcal{T}_h$ and consider a smooth function $\xi \in C^1(K)$. Integration by parts shows that the solution to (38.1), if it is smooth enough, satisfies

$$\int_K f\xi \, dx = \int_K -(\Delta u)\xi \, dx = \int_K \nabla u \cdot \nabla \xi \, dx - \int_{\partial K} (\nabla u \cdot \boldsymbol{n}_K)\xi \, ds.$$

Splitting the boundary integral over the faces $F \in \mathcal{F}_K$ yields

$$\int_K \nabla u \cdot \nabla \xi \, dx + \sum_{F \in \mathcal{F}_K} \epsilon_{K,F} \int_F \Phi_F(u)\xi \, ds = \int_K f\xi \, dx, \tag{38.27}$$

where $\Phi_F(u) := -\nabla u \cdot \boldsymbol{n}_F$, $\epsilon_{K,F} = \boldsymbol{n}_K \cdot \boldsymbol{n}_F$, and $\boldsymbol{n}_K$ is the outward normal to K ($\boldsymbol{n}_K \cdot \boldsymbol{n}_F = \pm 1$, for all $F \in \mathcal{F}_K$, depending on the orientation of F). The function Φ_F is called exact flux since (38.27) expresses a balance between the source term in K, the diffusion processes in K, and the fluxes across all the faces in $\mathcal{F}_K$. An interesting feature of dG methods is that one obtains a discrete counterpart of (38.27) when the test function is supported only in the mesh cell K.

Lemma 38.20 (Local formulation). *Let u_h solve (38.5). Let the numerical flux on a mesh face $F \in \mathcal{F}_h$ be defined by*

$$\widehat{\Phi}_F(u_h) := -\{\nabla_h u_h\} \cdot \boldsymbol{n}_F + \varpi(h_F)[\![u_h]\!]. \tag{38.28}$$

Then the following holds true for all $q \in P_K$ and all $K \in \mathcal{T}_h$:

$$\int_K \mathfrak{G}_h^l(u_h) \cdot \nabla q \, \mathrm{d}x + \sum_{F \in \mathcal{F}_K} \epsilon_{K,F} \int_F \widehat{\Phi}_F(u_h) q \, \mathrm{d}s = \int_K f q \, \mathrm{d}x. \tag{38.29}$$

Proof. Let $\mathbb{1}_K$ be the indicator function of K and let q be arbitrary in P_K. Using the test function $w_h := q\mathbb{1}_K$ in (38.5), we obtain $a_h(u_h, q\mathbb{1}_K) = \int_K f q \, \mathrm{d}x$. Then, (38.29) follows by invoking (38.24) and by making use of the identity $[\![q\mathbb{1}_K]\!]_F = \epsilon_{K,F} q$ if $F \in \mathcal{F}_K$ and $[\![q\mathbb{1}_K]\!]_F = 0$ otherwise. $\qquad\square$

The numerical flux $\widehat{\Phi}_F(u_h)$ consists of a centered flux, $-\{\nabla_h u_h\} \cdot \boldsymbol{n}_F$, originating from the consistency term, plus a stabilization term, $\varpi(h_F)[\![u_h]\!]$, originating from the penalty term. A unified presentation of dG methods for the Poisson equation based on fluxes can be found in Arnold et al. [21].

38.4.3 Equilibrated $\boldsymbol{H}(\mathrm{div})$ flux recovery

The vector-valued function $\boldsymbol{\sigma} := -\nabla u$ is called *diffusive flux.* This function is important in many applications where the underlying PDE expresses a conservation principle in the form $\nabla \cdot \boldsymbol{\sigma} = f$ in D. Since $\boldsymbol{\sigma} \in \boldsymbol{H}(\mathrm{div}; D)$, Theorem 18.10 implies that $[\![\boldsymbol{\sigma}]\!] \cdot \boldsymbol{n}_F = 0$ for all $F \in \mathcal{F}_h^\circ$ (possibly in a weak sense if $\boldsymbol{\sigma}$ is not smooth enough). From a physical viewpoint, this zero-jump condition expresses the fact that what flows out of a mesh cell through one of its faces flows into the neighboring mesh cell.

The local formulation (38.29) provides a natural way of reconstructing a discrete diffusive flux $\boldsymbol{\sigma}_h$ in $\boldsymbol{H}(\mathrm{div}; D)$ that closely approximates $\boldsymbol{\sigma}$. Assuming that the mesh is matching and simplicial, we now describe a way to reconstruct $\boldsymbol{\sigma}_h$ in the Raviart–Thomas finite element space $\boldsymbol{P}_l^{\mathrm{d}}(\mathcal{T}_h)$ defined in (19.16) with $l \in \{k-1, k\}$. The reconstruction is explicit and amounts to prescribing the global degrees of freedom of $\boldsymbol{\sigma}_h$ in $\boldsymbol{P}_h^{\mathrm{d}}(\mathcal{T}_h)$; see Ern et al. [193], Kim [268].

Lemma 38.21 (Flux recovery). *Let $\boldsymbol{\sigma}_h \in \boldsymbol{P}_h^{\mathrm{d}}(\mathcal{T}_h)$ be such that*

$$\int_F (\boldsymbol{\sigma}_h \cdot \boldsymbol{n}_F)(q \circ \boldsymbol{T}_F^{-1}) \, \mathrm{d}s = \int_F \widehat{\Phi}_F(u_h)(q \circ \boldsymbol{T}_F^{-1}) \, \mathrm{d}s, \qquad \forall F \in \mathcal{F}_h, \forall q \in \mathbb{P}_{l,d-1},$$

$$\text{and if } l \geq 1, \quad \int_K \boldsymbol{\sigma}_h \cdot \boldsymbol{r} \, \mathrm{d}x = -\int_K \mathfrak{G}_h^l(u_h) \cdot \boldsymbol{r} \, \mathrm{d}x, \qquad \forall K \in \mathcal{T}_h, \forall \boldsymbol{r} \in \mathbb{P}_{l-1,d},$$

where $\boldsymbol{T}_F$ is an affine bijective mapping from the unit simplex of $\mathbb{R}^{d-1}$ to F for all $F \in \mathcal{F}_h$. Let $\mathcal{I}_h^{\mathrm{b}}$ denote the L^2-orthogonal projection onto $P_l^{\mathrm{b}}(\mathcal{T}_h)$. Then we have

$$\nabla \cdot \boldsymbol{\sigma}_h = \mathcal{I}_h^{\mathrm{b}}(f). \tag{38.30}$$

Proof. Integrating by parts on a cell $K \in \mathcal{T}_h$ and using (38.29), we infer that

$$\int_K (\nabla \cdot \boldsymbol{\sigma}_h) q \, \mathrm{d}x = -\int_K \boldsymbol{\sigma}_h \cdot \nabla q \, \mathrm{d}x + \sum_{F \in \mathcal{F}_K} \int_F (\boldsymbol{\sigma}_h \cdot \boldsymbol{n}_K)(q_{|F} \circ \boldsymbol{T}_F) \circ \boldsymbol{T}_F^{-1} \, \mathrm{d}s = \int_K f q \, \mathrm{d}x,$$

for all $q \in \mathbb{P}_{l,d}$ since $\nabla q \in \mathbb{P}_{l-1,d}$ and $q_{|F} \circ \boldsymbol{T}_F \in \mathbb{P}_{l,d-1}$ (see Lemma 7.10). Then (38.30) is a consequence of $\nabla \cdot \boldsymbol{\sigma}_h \in P_l^{\mathrm{b}}(\mathcal{T}_h)$. $\qquad\square$

Equation (38.30) shows that $\nabla \cdot \boldsymbol{\sigma}_h$ optimally approximates the source term. By proceeding as in Di Pietro and Ern [165, §5.5.3], it is possible to show that $\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)} \leq c(\|u - u_h\|_{V_\sharp} + h\|f - \mathcal{I}_h^{\mathrm{b}}(f)\|_{L^2(D)})$.

Exercises

Exercise 38.1 (Elementary dG identities). (i) Let $F := \partial K_l \cap \partial K_r \in \mathcal{F}_h^\circ$. Prove that $2\{\boldsymbol{\sigma}\cdot\boldsymbol{n}_K q\} = (\{\boldsymbol{\sigma}\}[\![q]\!] + [\![\boldsymbol{\sigma}]\!]\{q\})\cdot\boldsymbol{n}_F$. (ii) Let $\theta_l, \theta_r \in [0,1]$ such that $\theta_l + \theta_r = 1$. Let $[\![a]\!]_\theta := 2(\theta_r a_l - \theta_l a_r)$ and $\{a\}_\theta := \theta_l a_l + \theta_r a_r$. Show that $\{ab\} = \{a\}\{b\}_\theta + \frac{1}{4}[\![a]\!]_\theta[\![b]\!]$.

Exercise 38.2 (Boundary conditions). (i) Assume that u solves the Poisson problem (38.1) with the non-homogeneous Dirichlet condition $u = g$ on ∂D. Let a_h^θ be defined in (38.20). Devise $\ell_h^{\theta,\mathrm{nD}}$ so that exact consistency holds for the following formulation: Find $u_h \in V_h$ such that $a_h^\theta(u_h, w_h) = \ell_h^{\theta,\mathrm{nD}}(w_h)$ for all $w_h \in V_h$. (ii) Assume that u solves the Poisson problem with the Robin condition $\gamma u + \boldsymbol{n}\cdot\nabla u = g$ on ∂D. Let ℓ_h^{Rb} be defined in (38.13b). Devise a_h^{Rb} so that exact consistency holds for the following formulation: Find $u_h \in V_h$ such that $a_h^{\theta,\mathrm{Rb}}(u_h, w_h) = \ell_h^{\mathrm{Rb}}(w_h)$ for all $w_h \in V_h$.

Exercise 38.3 (L^2-estimate). Prove Theorem 38.12. (*Hint*: see the proof of Theorem 37.8.)

Exercise 38.4 (Local lifting). Prove (38.22a). (*Hint*: use (38.10).)

Exercise 38.5 (Local formulation). Write the local formulation of the OBB, NIP, and IIP dG methods discussed in Remark 38.13.

Exercise 38.6 (Extending (38.25)). Let $\tilde{a}_h$ (resp., a_h) be defined by extending (38.25) (resp., (38.4)) to $V_\sharp \times V_h$. Show that $\tilde{a}_h(v, w_h) = a_h(v, w_h) + \sum_{F\in\mathcal{F}_h} \int_F \{\nabla_h v - \mathcal{I}_h^{\mathrm{b}}(\nabla_h v)\}\cdot\boldsymbol{n}_F[\![w_h]\!]\,\mathrm{d}s$ for all $(v, w_h) \in V_\sharp \times V_h$.

Exercise 38.7 (Discrete gradient). Let $(v_h)_{h\in\mathcal{H}}$ be a sequence in $(V_h)_{h\in\mathcal{H}}$ (meaning that $v_h \in V_h$ for all $h \in \mathcal{H}$). Assume that there is C s.t. $\|v_h\|_{V_h} \leq C$ for all $h \in \mathcal{H}$. One can show that there is $v \in L^2(D)$ such that, up to a subsequence, $v_h \to v$ in $L^2(D)$ as $h \to 0$; see [165, Thm. 5.6]. (i) Show that, up to a subsequence, $\mathfrak{G}_h^l(v_h)$ weakly converges to some $\boldsymbol{G}$ in $\boldsymbol{L}^2(D)$ as $h \to 0$. (*Hint*: bound $\|\mathfrak{G}_h^l(v_h)\|_{\boldsymbol{L}^2(D)}$.) (ii) Show that $\boldsymbol{G} = \nabla v$ and that $v \in H_0^1(D)$. (*Hint*: extend functions by zero outside D and prove first that $\int_{\mathbb{R}^d} \mathfrak{G}_h^l(v_h)\cdot\boldsymbol{\Phi}\,\mathrm{d}x = -\int_{\mathbb{R}^d} v_h\nabla\cdot\boldsymbol{\Phi}\,\mathrm{d}x + \sum_{F\in\mathcal{F}_h} \int_F \{\boldsymbol{\Phi}-\mathcal{I}_h^{\mathrm{b}}\boldsymbol{\Phi}\}\cdot\boldsymbol{n}_F[\![v_h]\!]\,\mathrm{d}s$ for all $\boldsymbol{\Phi} \in \boldsymbol{C}_0^\infty(\mathbb{R}^d)$.)

Chapter 39

Hybrid high-order method

As in Chapter 38, we want to approximate the Poisson equation with homogeneous Dirichlet conditions, but this time we use the hybrid high-order (HHO) method. In this method, the discrete solution is composed of a pair: a face component that approximates the trace of the solution on the mesh faces and a cell component that approximates the solution in the mesh cells. The cell unknowns can be eliminated locally by static condensation. The two key ideas behind the HHO method are a local reconstruction operator and a local stabilization operator. Altogether the approximation setting is nonconforming since the solution is approximated by piecewise polynomials that can jump across the mesh interfaces. The error analysis leads to $\mathcal{O}(h^{k+1})$ convergence rates in H^1 for smooth solutions if polynomials of degree $k \geq 0$ are used for the face and the cell unknowns. Moreover, we show that the HHO method is closely related to the hybridizable discontinuous Galerkin (HDG) method.

39.1 Local operators

Local reconstruction and stabilization operators associated with each mesh cell lie at the heart of the HHO method. Although these operators can be defined on general meshes composed of cells having a polyhedral shape, for simplicity, we are going to restrict our attention to simplicial meshes.

39.1.1 Discrete setting

Let $K \in \mathcal{T}_h$ be a mesh cell, where $\mathcal{T}_h$ is a member of a shape-regular sequence of affine simplicial meshes. Let $k \geq 0$ be the polynomial degree. We consider a pair $\hat{v}_K := (v_K, v_{\partial K})$, where v_K is defined on K and $v_{\partial K}$ is defined on the faces $F \in \mathcal{F}_K$ composing the boundary ∂K of K. We write $\hat{v}_K := (v_K, v_{\partial K}) \in \hat{V}_K^k := V_K^k \times V_{\partial K}^k$ with

$$V_K^k := \mathbb{P}_{k,d} \circ \boldsymbol{T}_K^{-1}, \quad V_{\partial K}^k := \prod_{F \in \mathcal{F}_K} \mathbb{P}_{k,d-1} \circ \boldsymbol{T}_F^{-1}, \tag{39.1}$$

where $\boldsymbol{T}_K : \widehat{S}^d \to K$, $\boldsymbol{T}_F : \widehat{S}^{d-1} \to F$ are affine geometric mappings defined on the reference simplices of $\mathbb{R}^d$ and $\mathbb{R}^{d-1}$, respectively (see Figure 39.1). We have $\dim(\hat{V}_K^k) = \binom{k+d}{d} + (d+1)\binom{k+d-1}{d-1}$. Functions in $V_{\partial K}^k$ are defined independently on each face composing the boundary ∂K of K. More

precisely, if $v_{\partial K} \in V_{\partial K}^k$ and $F_1, F_2 \in \mathcal{F}_K$ are two distinct faces of K, then $v_{\partial K|F_1}$ and $v_{\partial K|F_2}$ may not have the same restriction on $F_1 \cap F_2$.

$$k = 0 \qquad\qquad k = 1 \qquad\qquad k = 2$$
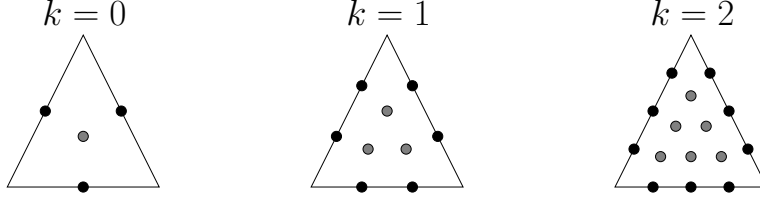
Figure 39.1: Local unknowns for the HHO method ($d = 2$). Each bullet on the faces and in the cell conventionally represents one basis function, which can be of modal or nodal type. The face basis functions are not necessarily continuous at the cell vertices.

39.1.2 Local reconstruction and stabilization

Let $V_K^{k+1} := \mathbb{P}_{k+1,d} \circ \boldsymbol{T}_K^{-1}$. We define a reconstruction operator $\mathsf{R} : \hat{V}_K^k \to V_K^{k+1}$ s.t., for every pair $\hat{v}_K = (v_K, v_{\partial K}) \in \hat{V}_K^k$, the function $\mathsf{R}(\hat{v}_K) \in V_K^{k+1}$ is s.t. for all $q \in V_K^{k+1}$,

$$\begin{aligned}(\nabla\mathsf{R}(\hat{v}_K), \nabla q)_{\boldsymbol{L}^2(K)} &:= -(v_K, \Delta q)_{L^2(K)} + (v_{\partial K}, \boldsymbol{n}_K \cdot \nabla q)_{L^2(\partial K)} \\ &= (\nabla v_K, \nabla q)_{\boldsymbol{L}^2(K)} - (v_K - v_{\partial K}, \boldsymbol{n}_K \cdot \nabla q)_{L^2(\partial K)},\end{aligned} \tag{39.2}$$

and $(\mathsf{R}(\hat{v}_K) - v_K, 1)_{L^2(K)} := 0$. This local Neumann problem (where the unknown is $\mathsf{R}(\hat{v}_K)$) makes sense since the right-hand side of (39.2) vanishes when the function q is constant. The second equality in (39.2) is obtained by integration by parts. The reconstruction operator depends on K and k, but for simplicity we just write R. Notice that $\mathsf{R}(\hat{v}_K) = v_K$ if $v_{\partial K} = v_{K|\partial K}$. In practice, the computation of $\mathsf{R}(\hat{v}_K)$ requires inverting the local stiffness matrix in K of order $\binom{k+d+1}{d} - 1$.

Let $\hat{\mathcal{I}}_K^k : H^1(K) \to \hat{V}_K^k$ be the local interpolation operator s.t.

$$\hat{\mathcal{I}}_K^k(v) := (\Pi_K^k(v), \Pi_{\partial K}^k(v_{|\partial K})) \in \hat{V}_K^k, \qquad \forall v \in H^1(K), \tag{39.3}$$

where $\Pi_K^k : L^2(K) \to V_K^k$ is the L^2-orthogonal projection onto V_K^k and $\Pi_{\partial K}^k : L^2(\partial K) \to V_{\partial K}^k$ is the L^2-orthogonal projection onto $V_{\partial K}^k$.

Lemma 39.1 (Elliptic projection). $\mathcal{E}_K := \mathsf{R} \circ \hat{\mathcal{I}}_K^k : H^1(K) \to V_K^{k+1}$ *is the elliptic projection onto* V_K^{k+1}, *i.e.,* $(\nabla(\mathcal{E}_K(v) - v), \nabla q)_{\boldsymbol{L}^2(K)} = 0$ *and* $(\mathcal{E}_K(v) - v, 1)_{L^2(K)} = 0$ *for all* $q \in V_K^{k+1}$ *and all* $v \in H^1(K)$.

Proof. Let $v \in H^1(K)$ and $\phi := \mathsf{R}(\hat{\mathcal{I}}_K^k(v)) = \mathsf{R}(\Pi_K^k(v), \Pi_{\partial K}^k(v_{|\partial K}))$. Using the definition (39.2) of the reconstruction operator, we infer that

$$\begin{aligned}(\nabla\phi, \nabla q)_{\boldsymbol{L}^2(K)} &= -(\Pi_K^k(v), \Delta q)_{L^2(K)} + (\Pi_{\partial K}^k(v_{|\partial K}), \boldsymbol{n}_K \cdot \nabla q)_{L^2(\partial K)} \\ &= -(v, \Delta q)_{L^2(K)} + (v, \boldsymbol{n}_K \cdot \nabla q)_{L^2(\partial K)} = (\nabla v, \nabla q)_{\boldsymbol{L}^2(K)},\end{aligned}$$

for all $q \in V_K^{k+1}$, since $\Delta q \in V_K^k$ and $\boldsymbol{n}_K \cdot \nabla q \in V_{\partial K}^k$ (recall that all the faces are planar so that $\boldsymbol{n}_K$ is piecewise constant). Moreover, $(\mathsf{R}(\hat{\mathcal{I}}_K^k(v)), 1)_{L^2(K)} = (\Pi_K^k(v), 1)_{L^2(K)} = (v, 1)_{L^2(K)}$ owing to the definition of R and $\hat{\mathcal{I}}_K^k$. $\qquad\square$

The main issue with the reconstruction operator is that $\nabla \mathsf{R}(\hat{v}_K) = \mathbf{0}$ does not imply that v_K and $v_{\partial K}$ are constant functions taking the same value. This can be seen from a dimension argument: We have $\ker(\mathsf{R}) \subset \{\hat{v}_K \in \hat{V}_K^k \mid \nabla \mathsf{R}(\hat{v}_K) = \mathbf{0}\}$ and $\dim(\ker(\mathsf{R})) = \dim(\hat{V}_K^k) - \dim(\mathrm{im}(\mathsf{R})) \geq \dim(\hat{V}_K^k) - \dim(V_K^{k+1}) = \binom{k+d-1}{d-1}\frac{kd+1}{k+1} > 1$ (unless $k=0$ and $d=1$ where the difference is equal to 1). To fix this issue, a local stabilization operator is introduced. Among various possibilities, we focus on an operator that maps $\hat{V}_K^k$ to face-based functions $\mathsf{S} : \hat{V}_K^k \to V_{\partial K}^k$ s.t. for all $\hat{v}_K \in \hat{V}_K^k$,

$$\mathsf{S}(\hat{v}_K) := \Pi_{\partial K}^k \big(v_{K|\partial K} - v_{\partial K} + ((I - \Pi_K^k)\mathsf{R}(\hat{v}_K))_{|\partial K} \big), \tag{39.4}$$

where I is the identity. The stabilization operator depends on K and k, but for simplicity we just write S. Letting $\delta_{\partial K} := v_{K|\partial K} - v_{\partial K}$ be the difference between the trace of the cell component and the face component on ∂K, the operator S in (39.4) can be rewritten as

$$\mathsf{S}(\hat{v}_K) = \Pi_{\partial K}^k \big(\delta_{\partial K} - ((I - \Pi_K^k)\mathsf{R}(0, \delta_{\partial K}))_{|\partial K} \big), \tag{39.5}$$

where we used $\mathsf{R}(\hat{v}_K) = \mathsf{R}(v_K, v_{K|\partial K}) - \mathsf{R}(0, \delta_{\partial K})$, by linearity, and that $(I - \Pi_K^k)\mathsf{R}(v_K, v_{K|\partial K}) = 0$ since $\mathsf{R}(v_K, v_{K|\partial K}) = v_K$ and $v_K \in V_K^k$. The identity (39.5) shows that $\mathsf{S}(\hat{v}_K)$ only depends (linearly) on the difference $(v_{K|\partial K} - v_{\partial K})$. The role of S is to help enforce the matching between the trace of the cell component and the face component. In the discrete problem, this matching is enforced in a least-squares manner (see §39.2.1). In practice, computing $\mathsf{S}(\hat{v}_K)$ requires to evaluate L^2-orthogonal projections in the cell and on its faces, which entails inverting the mass matrix in K, which is of size $\binom{k+d}{d}$, and inverting the mass matrix in each face $F \in \mathcal{F}_K$, which is of size $\binom{k+d-1}{d-1}$.

We now show that the operator S leads to an important stability result. We equip the space $\hat{V}_K^k$ with the following H^1-like seminorm: For all $\hat{v}_K \in \hat{V}_K^k$,

$$|\hat{v}_K|_{\hat{V}_K^k}^2 := \|\nabla v_K\|_{\boldsymbol{L}^2(K)}^2 + h_K^{-1}\|v_K - v_{\partial K}\|_{L^2(\partial K)}^2. \tag{39.6}$$

Lemma 39.2 (Stability). *There are $0 < \alpha \leq \omega$ s.t. for all $\hat{v}_K \in \hat{V}_K^k$, all $K \in \mathcal{T}_h$, and all $h \in \mathcal{H}$,*

$$\alpha \, |\hat{v}_K|_{\hat{V}_K^k}^2 \leq \|\nabla \mathsf{R}(\hat{v}_K)\|_{\boldsymbol{L}^2(K)}^2 + h_K^{-1}\|\mathsf{S}(\hat{v}_K)\|_{L^2(\partial K)}^2 \leq \omega \, |\hat{v}_K|_{\hat{V}_K^k}^2. \tag{39.7}$$

Proof. Let $\hat{v}_K = (v_K, v_{\partial K}) \in \hat{V}_K^k$ and set $r_K := \mathsf{R}(\hat{v}_K)$.
(1) Lower bound. Let us first bound $\|\nabla v_K\|_{\boldsymbol{L}^2(K)}$. Taking $q := v_K$ in (39.2) and using the Cauchy–Schwarz inequality yields

$$\|\nabla v_K\|_{\boldsymbol{L}^2(K)}^2 = (\nabla r_K, \nabla v_K)_{\boldsymbol{L}^2(K)} + (v_K - v_{\partial K}, \boldsymbol{n}_K \cdot \nabla v_K)_{L^2(\partial K)}$$

$$\leq \|\nabla r_K\|_{\boldsymbol{L}^2(K)}\|\nabla v_K\|_{\boldsymbol{L}^2(K)} + h_K^{-\frac{1}{2}}\|v_K - v_{\partial K}\|_{L^2(\partial K)}h_K^{\frac{1}{2}}\|\boldsymbol{n}_K \cdot \nabla v_K\|_{L^2(\partial K)}.$$

A discrete trace inequality yields $h_K^{\frac{1}{2}}\|\boldsymbol{n}_K \cdot \nabla v_K\|_{L^2(\partial K)} \leq c\|\nabla v_K\|_{\boldsymbol{L}^2(K)}$. These bounds imply that

$$\|\nabla v_K\|_{\boldsymbol{L}^2(K)} \leq c \, (\|\nabla r_K\|_{\boldsymbol{L}^2(K)} + h_K^{-\frac{1}{2}}\|v_K - v_{\partial K}\|_{L^2(\partial K)}). \tag{39.8}$$

Let us now bound $h_K^{-\frac{1}{2}}\|v_K - v_{\partial K}\|_{L^2(\partial K)}$. We have

$$\|\Pi_{\partial K}^k \big(((I - \Pi_K^k)r_K)_{|\partial K} \big)\|_{L^2(\partial K)} \leq \|(I - \Pi_K^k)r_K\|_{L^2(\partial K)}$$

$$\leq c \, h_K^{-\frac{1}{2}}\|(I - \Pi_K^k)r_K\|_{L^2(K)} \leq c' h_K^{\frac{1}{2}}\|\nabla r_K\|_{\boldsymbol{L}^2(K)},$$

owing to a discrete trace inequality and the local Poincaré–Steklov inequality (12.13) since $(I - \Pi_K^k)r_K$ has zero mean value in K. Using the definition of S and the fact that $v_{K|\partial K} - v_{\partial K}$ is in $V_{\partial K}^k$, we infer that $v_{K|\partial K} - v_{\partial K} = \mathsf{S}(\hat{v}_K) - \Pi_{\partial K}^k(((I - \Pi_K^k)r_K)_{|\partial K})$. The triangle inequality and the above bound imply that

$$h_K^{-\frac{1}{2}}\|v_K - v_{\partial K}\|_{L^2(\partial K)} \leq h_K^{-\frac{1}{2}}\|\mathsf{S}(\hat{v}_K)\|_{L^2(\partial K)} + c\,\|\nabla r_K\|_{\boldsymbol{L}^2(K)}.$$

Combining this estimate with (39.8) proves the lower bound in (39.7).
(2) Upper bound. Using the definition (39.2) of R, we have

$$\begin{aligned}
\|\nabla r_K\|_{\boldsymbol{L}^2(K)} &= \sup_{q \in V_K^{k+1}} \frac{(\nabla r_K, \nabla q)_{\boldsymbol{L}^2(K)}}{\|\nabla q\|_{\boldsymbol{L}^2(K)}} \\
&= \sup_{q \in V_K^{k+1}} \frac{(\nabla v_K, \nabla q)_{\boldsymbol{L}^2(K)} - (v_K - v_{\partial K}, \boldsymbol{n}_K\!\cdot\!\nabla q)_{L^2(\partial K)}}{\|\nabla q\|_{\boldsymbol{L}^2(K)}} \\
&\leq \|\nabla v_K\|_{\boldsymbol{L}^2(K)} + c\,h_K^{-\frac{1}{2}}\|v_K - v_{\partial K}\|_{L^2(\partial K)},
\end{aligned}$$

where the last bound follows from the Cauchy–Schwarz inequality and a discrete trace inequality. Moreover, the triangle inequality and the L^2-stability of $\Pi_{\partial K}^k$ imply that

$$\|\mathsf{S}(\hat{v}_K)\|_{L^2(\partial K)} \leq \|v_K - v_{\partial K}\|_{L^2(\partial K)} + \|\Pi_{\partial K}^k((I - \Pi_K^k)r_K)_{|\partial K}\|_{L^2(\partial K)}.$$

Invoking the above bound on $\|\Pi_{\partial K}^k(((I - \Pi_K^k)r_K)_{|\partial K})\|_{L^2(\partial K)}$ yields

$$h_K^{-\frac{1}{2}}\|\mathsf{S}(\hat{v}_K)\|_{L^2(\partial K)} \leq h_K^{-\frac{1}{2}}\|v_K - v_{\partial K}\|_{L^2(\partial K)} + c'\,\|\nabla r_K\|_{\boldsymbol{L}^2(K)}.$$

Combining the above bounds proves the upper bound in (39.7). $\qquad\square$

Another important property of the stabilization operator is that it leads to optimal approximation properties when combined with the interpolation operator $\hat{\mathcal{I}}_K^k$.

Lemma 39.3 (Approximation property of $\mathsf{S} \circ \hat{\mathcal{I}}_K^k$). *There is c s.t. for all $v \in H^1(K)$, all $K \in \mathcal{T}_h$, and all $h \in \mathcal{H}$,*

$$h_K^{-\frac{1}{2}}\|\mathsf{S}(\hat{\mathcal{I}}_K^k(v))\|_{L^2(\partial K)} \leq c\,\|\nabla(v - \mathcal{E}_K(v))\|_{\boldsymbol{L}^2(K)}. \tag{39.9}$$

Proof. Let $v \in H^1(K)$ and set $\eta := v - \mathcal{E}_K(v)$. Owing to the definitions of S and $\hat{\mathcal{I}}_K^k$ and the fact that $\mathsf{R} \circ \hat{\mathcal{I}}_K^k = \mathcal{E}_K$, we infer that

$$\begin{aligned}
\mathsf{S}(\hat{\mathcal{I}}_K^k(v)) &= \Pi_{\partial K}^k\left(\Pi_K^k(v)_{|\partial K} - \Pi_{\partial K}^k(v_{|\partial K}) + ((I - \Pi_K^k)\mathcal{E}_K(v))_{|\partial K}\right) \\
&= \Pi_K^k(\eta)_{|\partial K} - \Pi_{\partial K}^k(\eta_{|\partial K}),
\end{aligned}$$

where we used that $\Pi_{\partial K}^k(\Pi_K^k(\eta)_{|\partial K}) = \Pi_K^k(\eta)_{|\partial K}$ and $\Pi_{\partial K}^k \circ \Pi_{\partial K}^k = \Pi_{\partial K}^k$. Invoking the triangle inequality, the L^2-stability of the projections Π_K^k and $\Pi_{\partial K}^k$, and a discrete trace inequality leads to

$$\|\mathsf{S}(\hat{\mathcal{I}}_K^k(v))\|_{L^2(\partial K)} \leq c\,(h_K^{-\frac{1}{2}}\|\eta\|_{L^2(K)} + \|\eta\|_{L^2(K)}) \leq c'h_K^{\frac{1}{2}}\|\nabla\eta\|_{\boldsymbol{L}^2(K)},$$

where the last bound follows from the multiplicative trace inequality (12.16) and the Poincaré–Steklov inequality (12.13) (since $(\eta, 1)_{L^2(K)} = 0$). $\qquad\square$

Remark 39.4 (Literature). The HHO method was introduced in Di Pietro and Ern [166], Di Pietro et al. [168]. Its algebraic realization and implementation are discussed in Cicuttin et al. [128], and an open-source library is available at `https://github.com/wareHHOuse/diskpp`. □

Remark 39.5 (Variants). Let $k \geq 0$ be the face polynomial degree. As observed in Cockburn et al. [137], the cell polynomial degree can be taken equal to $(k-1)$, if $k \geq 1$, or to k (as considered above), or to $(k+1)$. In the first case, the HHO method is related, up to an equivalent form of the stabilization, to the nonconforming Virtual Element method of Ayuso de Dios et al. [32] (which adopts a viewpoint close in spirit to §39.1.3). In the third case, the stabilization operator can be simplified to $\mathsf{S}(\hat{v}_K) := \Pi_{\partial K}^k(v_{K|\partial K} - v_{\partial K})$, as considered by Lehrenfeld and Schöberl for the HDG method (see [280, 281] and Oikawa [319]). All these methods have also close connections with the weak Galerkin method of Wang and Ye [386]. □

39.1.3 Finite element viewpoint

In this section, we briefly outline how the above setting can be understood within the finite element viewpoint by identifying a triple (K, P, Σ) (see Definition 5.2). Recall that $k \geq 0$ is the polynomial degree. Consider the space

$$\mathcal{V}_K^k := \left\{ v \in H^1(K) \mid \Delta v \in \mathbb{P}_{k,d} \circ \boldsymbol{T}_K^{-1}, \ \boldsymbol{n}_K \cdot (\nabla v)_{|\partial K} \in V_{\partial K}^k \right\}, \tag{39.10}$$

with $V_{\partial K}^k$ defined in (39.1). We observe that $V_K^{k+1} := \mathbb{P}_{k+1,d} \circ \boldsymbol{T}_K^{-1} \subset \mathcal{V}_K^k$, but there are functions in $\mathcal{V}_K^k$ that are not in V_K^{k+1} and these functions are not accessible to direct computation (they can be approximated by solving a subgrid problem in K). A key observation is the following.

Lemma 39.6 ($\mathcal{V}_K^k \leftrightarrow \hat{V}_K^k$). *The functional space $\mathcal{V}_K^k$ is finite-dimensional and the restriction of $\hat{\mathcal{I}}_K^k$ to $\mathcal{V}_K^k$, i.e., $\hat{\mathcal{I}}_{K|\mathcal{V}_K^k}^k : \mathcal{V}_K^k \to \hat{V}_K^k$, is an isomorphism.*

Proof. (1) Let us first prove that $\hat{\mathcal{I}}_{K|\mathcal{V}_K^k}^k : \mathcal{V}_K^k \to \hat{V}_K^k$ is injective. Let $v \in \mathcal{V}_K^k$ be s.t. $\hat{\mathcal{I}}_K^k(v) = (0,0) \in \hat{V}_K^k$. Integrating by parts, we infer that

$$\|\nabla v\|_{\boldsymbol{L}^2(K)}^2 = -(v, \Delta v)_{L^2(K)} + (v, \boldsymbol{n}_K \cdot \nabla v)_{L^2(\partial K)}$$
$$= -(\hat{\mathcal{I}}_K^k(v)_K, \Delta v)_{L^2(K)} + (\hat{\mathcal{I}}_K^k(v)_{\partial K}, \boldsymbol{n}_K \cdot \nabla v)_{L^2(\partial K)} = 0,$$

where we used the definitions of $\hat{\mathcal{I}}_K^k$ and of $\mathcal{V}_K^k$. Hence, v is constant on K and since $\Pi_K^k(v) = 0$, the mean value of v in K vanishes. Thus, $v = 0$.
(2) Consider the map $\Phi_K : \hat{V}_K^k \to \mathcal{V}_K^k$ s.t. for all $\hat{v}_K = (v_K, v_{\partial K}) \in \hat{V}_K^k$, the function $\Phi_K(\hat{v}_K)$ is the unique solution in $H^1(K)$ of the well-posed Neumann problem $-\Delta(\Phi_K(\hat{v}_K)) = v_K - \bar{v}_K + \frac{|\partial K|}{|K|}\bar{v}_{\partial K}$ in K, $-\boldsymbol{n}_K \cdot \nabla(\Phi_K(\hat{v}_K)) = v_{\partial K}$ on ∂K, and $(\Phi_K(\hat{v}_K) - v_K, 1)_{L^2(K)} = 0$, where $\bar{v}_K$ and $\bar{v}_{\partial K}$ denote the mean value of v_K and $v_{\partial K}$ on K and ∂K, respectively. By definition, $\Phi_K(\hat{v}_K) \in \mathcal{V}_K^k$. Moreover, $\Phi_K(\hat{v}_K)$ is clearly injective.
(3) Combining Steps (1) and (2), the rank nullity theorem implies $\dim(\mathcal{V}_K^k) = \dim(\hat{V}_K^k)$ and $\hat{\mathcal{I}}_{K|\mathcal{V}_K^k}^k : \mathcal{V}_K^k \to \hat{V}_K^k$ is an isomorphism. □

Let Σ be the collection of the following linear forms acting on $\mathcal{V}_K^k$:

$$\sigma_{F,m}^{\mathrm{f}}(v) := \frac{1}{|F|} \int_F v\,(\zeta_m \circ \boldsymbol{T}_F^{-1})\,\mathrm{d}s, \quad \forall F \in \mathcal{F}_K, \tag{39.11a}$$

$$\sigma_m^{\mathrm{c}}(v) := \frac{1}{|K|} \int_K v\,(\psi_m \circ \boldsymbol{T}_K^{-1})\,\mathrm{d}x, \tag{39.11b}$$

where $\{\zeta_m\}_{m\in\{1:n_{\mathrm{sh}}^{\mathrm{f}}\}}$ is a basis of $\mathbb{P}_{k,d-1}$ with $n_{\mathrm{sh}}^{\mathrm{f}} := \dim(\mathbb{P}_{k,d-1}) = \binom{k+d-1}{d-1}$ and $\{\psi_m\}_{m\in\{1:n_{\mathrm{sh}}^{\mathrm{c}}\}}$ is a basis of $\mathbb{P}_{k,d}$ with $n_{\mathrm{sh}}^{\mathrm{c}} := \dim(\mathbb{P}_{k,d}) = \binom{d+k}{d}$.

Lemma 39.7 (Finite element). *The triple* $(K, \mathcal{V}_K^k, \Sigma)$ *is a finite element.*

Proof. Direct consequence of Lemma 39.6. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

39.2   Discrete problem

We now show how to assemble the discrete problem, how to reduce its size by static condensation, and how the HHO and HDG methods are connected. Let D be a Lipschitz polyhedron in $\mathbb{R}^d$ and $f \in L^2(D)$. The model problem is

$$\begin{cases} \text{Find } u \in V := H_0^1(D) \text{ such that} \\ a(u,w) = \ell(w), \quad \forall w \in V, \end{cases} \tag{39.12}$$

with $a(v,w) := (\nabla v, \nabla w)_{\boldsymbol{L}^2(D)}$ and $\ell(w) := (f,w)_{L^2(D)}$ for all $v, w \in V$.

39.2.1   Assembling and well-posedness

Let $k \geq 0$. Let $\mathcal{T}_h$ be a member of a shape-regular family of affine simplicial meshes. Let $\mathcal{F}_h = \mathcal{F}_h^{\circ} \cup \mathcal{F}_h^{\partial}$ be the collection of the faces of $\mathcal{T}_h$, where $\mathcal{F}_h^{\circ}$ is the collection of the mesh interfaces and $\mathcal{F}_h^{\partial}$ the collection of the mesh boundary faces. Let us set $\hat{V}_h^k := V_{\mathcal{T}_h}^k \times V_{\mathcal{F}_h}^k$, where

$$V_{\mathcal{T}_h}^k := \{ v_{\mathcal{T}_h} \in L^2(D) \mid v_{\mathcal{T}_h|K} \in V_K^k, \, \forall K \in \mathcal{T}_h \}, \tag{39.13a}$$

$$V_{\mathcal{F}_h}^k := \{ v_{\mathcal{F}_h} \in L^2(\mathcal{F}_h) \mid v_{\mathcal{F}_h|\partial K} \in V_{\partial K}^k, \, \forall K \in \mathcal{T}_h \}. \tag{39.13b}$$

Note that functions in $V_{\mathcal{F}_h}^k$ are single-valued on the mesh interfaces. For every pair $\hat{v}_h := (v_{\mathcal{T}_h}, v_{\mathcal{F}_h}) \in \hat{V}_h^k$ and all $K \in \mathcal{T}_h$, it is convenient to define $v_K := v_{\mathcal{T}_h|K}$ and $v_{\partial K} := v_{\mathcal{F}_h|\partial K}$, so that $\hat{v}_K := (v_K, v_{\partial K}) \in \hat{V}_K^k$. For all $K \in \mathcal{T}_h$, we define the local forms $\hat{a}_K$ and ℓ_K s.t.

$$\hat{a}_K(\hat{v}_K, \hat{w}_K) := (\nabla \mathsf{R}(\hat{v}_K), \nabla \mathsf{R}(\hat{w}_K))_{\boldsymbol{L}^2(K)} + h_K^{-1}(\mathsf{S}(\hat{v}_K), \mathsf{S}(\hat{w}_K))_{L^2(\partial K)},$$

$$\ell_K(w_K) := (f, w_K)_{L^2(K)},$$

for all $\hat{v}_K, \hat{w}_K \in \hat{V}_K^k$. We define the global forms $\hat{a}_h$ and ℓ_h s.t.

$$\hat{a}_h(\hat{v}_h, \hat{w}_h) := \sum_{K\in\mathcal{T}_h} \hat{a}_K(\hat{v}_K, \hat{w}_K), \quad \ell_h(w_{\mathcal{T}_h}) := \sum_{K\in\mathcal{T}_h} \ell_K(w_K), \tag{39.14}$$

for all $\hat{v}_h, \hat{w}_h \in \hat{V}_h^k$. Notice that only the cell component of the test function is used to evaluate ℓ_h. We enforce strongly the homogeneous Dirichlet boundary condition by zeroing out the discrete unknowns associated with the boundary faces, i.e., we consider

$$\hat{V}_{h,0}^k := V_{\mathcal{T}_h}^k \times V_{\mathcal{F}_h,0}^k, \quad V_{\mathcal{F}_h,0}^k := \{ v_{\mathcal{F}_h} \in V_{\mathcal{F}_h}^k \mid v_{\mathcal{F}_h|F} := 0, \, \forall F \in \mathcal{F}_h^{\partial} \}. \tag{39.15}$$

The discrete problem is as follows:

$$\begin{cases} \text{Find } \hat{u}_h \in \hat{V}_{h,0}^k \text{ such that} \\ \hat{a}_h(\hat{u}_h, \hat{w}_h) = \ell_h(w_{\mathcal{T}_h}), \quad \forall \hat{w}_h \in \hat{V}_{h,0}^k. \end{cases} \tag{39.16}$$

In other words, the HHO method produces a discrete solution having two components: a piecewise polynomial function in the mesh cells and a piecewise polynomial function on the mesh faces. The second component is single-valued at the mesh interfaces, vanishes at the boundary faces, and its value can jump from one interface to a neighboring one.

To establish the well-posedness of (39.16), we prove that the bilinear form $\hat{a}_h$ is coercive on $\hat{V}_{h,0}^k$. We equip this space with the norm

$$\|\hat{v}_h\|_{\hat{V}_{h,0}^k}^2 := \sum_{K \in \mathcal{T}_h} |\hat{v}_K|_{\hat{V}_K^k}^2, \qquad \forall \hat{v}_h \in \hat{V}_{h,0}^k. \tag{39.17}$$

The only nontrivial property to verify that we have indeed defined a norm is that $\|\hat{v}_h\|_{\hat{V}_{h,0}^k} = 0$ implies $\hat{v}_h = (0,0)$. Let $v_h \in \hat{V}_{h,0}^k$ be s.t. $\|\hat{v}_h\|_{\hat{V}_{h,0}^k} = 0$, i.e., $|\hat{v}_K|_{\hat{V}_K^k} = 0$ for all $K \in \mathcal{T}_h$. Then recalling (39.6) we infer that v_K and $v_{\partial K}$ are constant functions taking the same value in each mesh cell. On cells having a boundary face, this value must be zero since $v_{\mathcal{F}_h}$ vanishes on the boundary faces. We can repeat the argument for the cells sharing an interface with those cells, and we can move inward and reach all the cells in $\mathcal{T}_h$ by repeating this process a finite number of times. Thus, $\hat{v}_h = (0,0)$.

Lemma 39.8 (Coercivity, well-posedness). (i) *The bilinear form $\hat{a}_h$ is coercive on $\hat{V}_{h,0}^k$.* (ii) *The discrete problem* (39.16) *is well-posed.*

Proof. The coercivity of $\hat{a}_h$ follows by summing the lower bound from Lemma 39.2 over the mesh cells, which yields

$$\hat{a}_h(\hat{v}_h, \hat{v}_h) \geq \alpha \|\hat{v}_h\|_{\hat{V}_{h,0}^k}^2, \qquad \forall \hat{v}_h \in \hat{V}_{h,0}^k. \tag{39.18}$$

Well-posedness is a consequence of the Lax–Milgram lemma. $\qquad\qquad\qquad\qquad\qquad\qquad$ □

Remark 39.9 (Finite element viewpoint). The role of the stabilization in the HHO method can also be understood by taking inspiration from the ideas at the heart of the virtual element method (Beirão da Veiga et al. [50]). Since manipulating functions $v \in \mathcal{V}_K^k$ (see (39.10)) is unpractical because these functions are not known explicitly, one would like to manipulate only the projection $\mathcal{E}_K(v) \in V_K^{k+1}$ which is computable from the dofs $\{\sigma(v)\}_{\sigma \in \Sigma}$ of v, that is, from the polynomial pair $\hat{\mathcal{I}}_K^k(v) \in \hat{V}_K^k$. The local bilinear form on $\mathcal{V}_K^k \times \mathcal{V}_K^k$ is $\mathfrak{a}_K(v,w) := (\nabla \mathcal{E}_K(v), \nabla \mathcal{E}_K(w))_{L^2(K)} + h_K^{-1}(\mathsf{S}(\hat{\mathcal{I}}_K^k(v)), \mathsf{S}(\hat{\mathcal{I}}_K^k(v)))_{L^2(\partial K)}$. To prove that $\mathfrak{a}_K(v,v)$ controls $\|\nabla v\|_{\boldsymbol{L}^2(K)}^2$, one needs to control $\|\nabla(v - \mathcal{E}_K(v))\|_{\boldsymbol{L}^2(K)}^2$, and this is where the stabilization comes into play (see Exercise 39.2). $\qquad$ □

39.2.2 Static condensation and global transmission problem

The problem (39.16) can be solved by using a Schur complement technique consisting of locally eliminating all the cell unknowns (this technique is also known as static condensation; see §28.1.2). In other words, (39.16) can be reformulated in the form of local problems patched together by a global transmission problem. To see this, we define $U_\mu \in V_K^k$ for all $\mu \in V_{\partial K}^k$, and we define $U_r \in V_K^k$ for all $r \in L^2(K)$, s.t.

$$\hat{a}_K((U_\mu, 0), (q, 0)) := -\hat{a}_K((0, \mu), (q, 0)), \quad \forall q \in V_K^k, \tag{39.19a}$$

$$\hat{a}_K((U_r, 0), (q, 0)) := (r, q)_{L^2(K)}, \qquad \forall q \in V_K^k. \tag{39.19b}$$

These problems are well-posed since $\hat{a}_K$ is coercive on $V_K^k \times \{0\}$.

Proposition 39.10 (Transmission problem). *The pair* $\hat{u}_h \in \hat{V}_{h,0}^k$ *solves* (39.16) *iff* $u_K = U_{u_{\partial K}} + U_{f_{|K}}$ *for all* $K \in \mathcal{T}_h$*, and* $u_{\mathcal{F}_h} \in V_{\mathcal{F}_h,0}^k$ *solves the following global transmission problem: For all* $w_{\mathcal{F}_h} \in V_{\mathcal{F}_h,0}^k$*,*

$$\sum_{K \in \mathcal{T}_h} \hat{a}_K((U_{u_{\partial K}}, u_{\partial K}),(U_{w_{\partial K}}, w_{\partial K})) = \sum_{K \in \mathcal{T}_h} \ell_K(U_{w_{\partial K}}). \tag{39.20}$$

Proof. Assume that $\hat{u}_h$ solves (39.16). Let $K \in \mathcal{T}_h$ and $w_K \in V_K^k$. Since $\hat{a}_K((u_K, u_{\partial K}),(w_K,0)) = \ell_K(w_K) = \hat{a}_K((U_{f_{|K}},0),(w_K,0))$, we infer that

$$\hat{a}_K((u_K - U_{f_{|K}}, u_{\partial K}),(w_K,0)) = \hat{a}_K((u_K, u_{\partial K}) - (U_{f_{|K}},0),(w_K,0))$$
$$= 0 = \hat{a}_K((U_{u_{\partial K}}, u_{\partial K}),(w_K,0)),$$

showing that $u_K - U_{f_{|K}} = U_{u_{\partial K}}$. This implies that for all $w_{\partial K} \in V_{\partial K}^k$,

$$\hat{a}_K((U_{u_{\partial K}}, u_{\partial K}),(U_{w_{\partial K}}, w_{\partial K})) = \hat{a}_K((u_K, u_{\partial K}),(U_{w_{\partial K}}, w_{\partial K})) - \hat{a}_K((U_{f_{|K}},0),(U_{w_{\partial K}}, w_{\partial K}))$$
$$= \hat{a}_K((u_K, u_{\partial K}),(U_{w_{\partial K}}, w_{\partial K})),$$

where we used the symmetry of $\hat{a}_K$ and that $\hat{a}_K(U_{w_{\partial K}}, w_{\partial K}),(q,0)) = 0$ for all $q \in V_K^k$. Summing over $K \in \mathcal{T}_h$ shows that $u_{\mathcal{F}_h}$ solves the transmission problem (39.20). The converse statement is proved in Exercise 39.6. $\square$

Remark 39.11 (Transmission problem). Following Cockburn [130], one can show that the problem (39.12) can be reformulated as a transmission problem. Let $a_K(\phi, \psi) := (\nabla\phi, \nabla\psi)_{\boldsymbol{L}^2(K)}$ and $\ell_K(\psi) := (f, \psi)_{L^2(K)}$ be the restrictions to K of the exact forms a and ℓ. There is a unique lifting $U_\mu \in H^1(K)$ for all $\mu \in H^{\frac{1}{2}}(\partial K)$ s.t. $U_{\mu|\partial K} := \mu$ and $a_K(U_\mu, \psi) := 0$ for all $\psi \in H_0^1(K)$. Similarly, there is a unique lifting $U_r \in H_0^1(K)$ for all $r \in L^2(K)$ s.t. $U_{r|\partial K} := 0$ and $a_K(U_r, \psi) := (r, \psi)_{L^2(K)}$ for all $\psi \in H_0^1(K)$. For all $v \in H^1(D)$, we slightly abuse the notation by writing $v_{|\mathcal{F}_h}$ for the restriction of v to the mesh faces, and for every function λ defined on the mesh faces, we write $\lambda_{\partial K}$ for its restriction to the boundary of any mesh cell $K \in \mathcal{T}_h$. Since the weak solution u is in $H_0^1(D)$, $u_{|\mathcal{F}_h}$ is in the trace space Λ defined as

$$\Lambda := \{\lambda \in L^2(\mathcal{F}_h) \mid \lambda_{\partial K} \in H^{\frac{1}{2}}(\partial K), \forall K \in \mathcal{T}_h; \ \lambda_{|\mathcal{F}_h^\partial} = 0\}. \tag{39.21}$$

By definition, functions in Λ are single-valued on the mesh interfaces. Then the function u is the weak solution iff there exists $\lambda \in \Lambda$ s.t. $u_{|K} = U_{\lambda_{\partial K}} + U_{f_{|K}}$ for all $K \in \mathcal{T}_h$ and λ solves the transmission problem

$$\sum_{K \in \mathcal{T}_h} a_K(U_{\lambda_{\partial K}}, U_{\mu_{\partial K}}) = \sum_{K \in \mathcal{T}_h} \ell_K(U_{\mu_{\partial K}}), \quad \forall\mu \in \Lambda. \tag{39.22}$$

Moreover, assuming that $\boldsymbol{n}_F \cdot (\nabla u)_{|F}$ is in $L^2(F)$ for all $F \in \mathcal{F}_h$ (this is the case if $u \in H^{1+r}(D)$, $r > \frac{1}{2}$), we have for all $\mu \in \Lambda$,

$$\sum_{F \in \mathcal{F}_h^\circ} ([\![\nabla u]\!]_F \cdot \boldsymbol{n}_F, \mu)_{L^2(F)} = \sum_{K \in \mathcal{T}_h} \big(a_K(U_{\lambda_{\partial K}}, U_{\mu_{\partial K}}) - \ell_K(U_{\mu_{\partial K}})\big) = 0. \tag{39.23}$$

The identity (39.23) shows that the global transmission problem (39.22) expresses the continuity of the normal component of ∇u across the mesh interfaces. In conclusion, the problem (39.22) consists of seeking $\lambda \in \Lambda$ such that $[\![U_f + U_\lambda]\!]_F = 0$ and $\boldsymbol{n}_F \cdot [\![\nabla(U_f + U_\lambda)]\!]_F = 0$ for all $F \in \mathcal{F}_h^\circ$, $(U_f + U_\lambda)_{|\partial D} = 0$, and $-\Delta(U_f + U_\lambda)_{|K} = f_{|K}$ for all $K \in \mathcal{T}_h$. We refer the reader to Exercise 39.6 for more details. $\square$

Remark 39.12 (Finite element viewpoint). Recalling §39.1.3, we define the high-order Crouzeix–Raviart-type finite element space

$$\mathcal{V}_h^k := \{v \in L^2(D) \mid v_{|K} \in \mathcal{V}_K^k, \ \forall K \in \mathcal{T}_h, \ (\llbracket v \rrbracket, q \circ \boldsymbol{T}_F^{-1})_{L^2(F)} = 0, \ \forall q \in \mathbb{P}_{k,d-1}, \ \forall F \in \mathcal{F}_h^\circ\},$$

and

$$\mathcal{V}_{h,0}^k := \{v \in \mathcal{V}_h^k \mid (v, q \circ \boldsymbol{T}_F^{-1})_{L^2(F)} = 0, \ \forall q \in \mathbb{P}_{k,d-1}, \ \forall F \in \mathcal{F}_h^\partial\}.$$

For all $v_h, w_h \in \mathcal{V}_h^k$, let $\mathfrak{a}_h(v_h, w_h) := \sum_{K \in \mathcal{T}_h} \mathfrak{a}_K(v_h, w_h)$, with $\mathfrak{a}_K$ defined in Remark 39.9, and $\ell_h(w_h) := \sum_{K \in \mathcal{T}_h} (f, w_h)_{L^2(K)}$. Then the problem consisting of seeking $u_h \in \mathcal{V}_{h,0}^k$ s.t. $\mathfrak{a}_h(u_h, w_h) = \ell_h(w_h)$ for all $w_h \in \mathcal{V}_{h,0}^k$ is well-posed, and $\hat{\mathcal{I}}_k^k(u_{h|K})$ is the HHO solution in K. $\qquad\square$

39.2.3 Comparison with HDG and flux recovery

In this section, we compare the HHO method with the *hybridizable discontinuous Galerkin (HDG)* method. In the HDG method, one approximates a triple, whereas one approximates a pair in the HHO method. Let us consider the dual variable $\boldsymbol{\sigma} := -\nabla u$ (sometimes called flux), the primal variable u, and its trace $\lambda := u_{|\mathcal{F}_h}$ on the mesh faces. HDG methods approximate the triple $(\boldsymbol{\sigma}, u, \lambda)$ by introducing local spaces $\boldsymbol{S}_K$, V_K, and V_F for all $K \in \mathcal{T}_h$ and all $F \in \mathcal{F}_h^\circ$, and by defining a numerical flux trace that includes a stabilization operator. Let us define the broken spaces

$$\boldsymbol{S}_{\mathcal{T}_h} := \{\boldsymbol{\tau}_{\mathcal{T}_h} \in \boldsymbol{L}^2(D) \mid \boldsymbol{\tau}_K \in \boldsymbol{S}_K, \forall K \in \mathcal{T}_h\}, \tag{39.24a}$$

$$V_{\mathcal{T}_h} := \{v_{\mathcal{T}_h} \in L^2(D) \mid v_K \in V_K, \forall K \in \mathcal{T}_h\}, \tag{39.24b}$$

$$V_{\mathcal{F}_h} := \{\mu_{\mathcal{F}_h} \in L^2(\mathcal{F}_h) \mid \mu_F \in V_F, \forall F \in \mathcal{F}_h\}, \tag{39.24c}$$

with $\boldsymbol{\tau}_K := \boldsymbol{\tau}_{\mathcal{T}_h|K}$, $v_K := v_{\mathcal{T}_h|K}$, and $\mu_F := \mu_{\mathcal{F}_h|F}$, and let us set $V_{\mathcal{F}_h,0} := \{\mu_{\mathcal{F}_h} \in V_{\mathcal{F}_h} \mid \mu_{\mathcal{F}_h|F} = 0, \forall F \in \mathcal{F}_h^\partial\}$. The HDG method consists of seeking $(\boldsymbol{\sigma}_{\mathcal{T}_h}, u_{\mathcal{T}_h}, \lambda_{\mathcal{F}_h}) \in \boldsymbol{S}_{\mathcal{T}_h} \times V_{\mathcal{T}_h} \times V_{\mathcal{F}_h,0}$ s.t. the following holds true for all $(\boldsymbol{\tau}_K, w_K, \mu_F) \in \boldsymbol{S}_K \times V_K \times V_F$, all $K \in \mathcal{T}_h$, and all $F \in \mathcal{F}_h^\circ$:

$$(\boldsymbol{\sigma}_K, \boldsymbol{\tau}_K)_{\boldsymbol{L}^2(K)} - (u_K, \nabla \cdot \boldsymbol{\tau}_K)_{L^2(K)} + (\lambda_{\partial K}, \boldsymbol{\tau}_K \cdot \boldsymbol{n}_K)_{L^2(\partial K)} = 0, \tag{39.25a}$$

$$-(\boldsymbol{\sigma}_K, \nabla w_K)_{\boldsymbol{L}^2(K)} + (\boldsymbol{\phi}_{\partial K} \cdot \boldsymbol{n}_K, w_K)_{L^2(\partial K)} = (f, w_K)_{L^2(K)}, \tag{39.25b}$$

$$(\llbracket \boldsymbol{\phi}_h \rrbracket \cdot \boldsymbol{n}_F, \mu_F)_{L^2(F)} = 0, \tag{39.25c}$$

where the numerical flux trace $\boldsymbol{\phi}_{\partial K}$ is defined by

$$\boldsymbol{\phi}_{\partial K} := \boldsymbol{\sigma}_{K|\partial K} + \tau_{\partial K}(u_{K|\partial K} - \lambda_{\partial K})\boldsymbol{n}_K \quad \text{on } \partial K, \tag{39.26}$$

for all $K \in \mathcal{T}_h$, with $\lambda_{\partial K} := (\lambda_F)_{F \in \mathcal{F}_K}$ and $\tau_{\partial K}$ is a linear stabilization operator. The equation (39.25a) is the discrete counterpart of $\boldsymbol{\sigma} = -\nabla u$, the equation (39.25b) that of $\nabla \cdot \boldsymbol{\sigma} = f$, and the equation (39.25c) weakly enforces the continuity of the normal component of the numerical flux trace. Various HDG methods are realized by choosing the local spaces $\boldsymbol{S}_K$, V_K, V_F, and the stabilization operator $\tau_{\partial K}$. In general, the stabilization operator $\tau_{\partial K}$ in the HDG method acts pointwise on ∂K. We will see in Proposition 39.13 that in the HHO method $\tau_{\partial K}$ acts collectively on ∂K.

Let $\tilde{\mathsf{S}} : V_{\partial K}^k \to V_{\partial K}^k$ be s.t. $\tilde{\mathsf{S}}(\mu) := \Pi_{\partial K}^k(\mu - ((I - \Pi_K^k)\mathsf{R}(0, \mu))_{|\partial K})$. The HHO stabilization operator satisfies $\mathsf{S}(\hat{v}_K) = \tilde{\mathsf{S}}(v_{K|\partial K} - v_{\partial K})$; see (39.5). By definition, the adjoint of $\tilde{\mathsf{S}}$, say $\tilde{\mathsf{S}}^* : V_{\partial K}^k \to V_{\partial K}^k$, is s.t. $(\tilde{\mathsf{S}}^*(\lambda), \mu)_{L^2(\partial K)} := (\lambda, \tilde{\mathsf{S}}(\mu))_{L^2(\partial K)}$ for all $\lambda, \mu \in V_{\partial K}^k$.

Proposition 39.13 (HHO vs. HDG). *Let* $\boldsymbol{S}_K := \nabla V_K^{k+1}$, $V_K := \mathbb{P}_{k,d} \circ \boldsymbol{T}_K^{-1}$, $V_F := \mathbb{P}_{k,d-1} \circ \boldsymbol{T}_F^{-1}$, *and* $\tau_{\partial K} := h_K^{-1} \tilde{\mathsf{S}}^* \circ \tilde{\mathsf{S}}$ *for all* $K \in \mathcal{T}_h$ *and all* $F \in \mathcal{F}_h^\circ$. *(i) If* $\hat{u}_h := (u_{\mathcal{T}_h}, u_{\mathcal{F}_h})$ *solves the HHO problem* (39.16), *then* $(\boldsymbol{\sigma}_{\mathcal{T}_h}, u_{\mathcal{T}_h}, u_{\mathcal{F}_h})$ *solves the HDG problem* (39.25) *with* $\boldsymbol{\sigma}_K := -\nabla \mathsf{R}(\hat{u}_K)$ *for all* $K \in \mathcal{T}_h$. *(ii) Conversely, if* $(\boldsymbol{\sigma}_{\mathcal{T}_h}, u_{\mathcal{T}_h}, \lambda_{\mathcal{F}_h})$ *solves the HDG problem* (39.25), *then* $\boldsymbol{\sigma}_K = -\nabla \mathsf{R}(u_K, \lambda_{\partial K})$ *for all* $K \in \mathcal{T}_h$, *and* $(u_{\mathcal{T}_h}, \lambda_{\mathcal{F}_h})$ *solves the HHO problem* (39.16).

Proof. We only prove the forward statement since the proof of the converse statement follows by similar arguments. Let $\hat{u}_h \in \hat{V}_{h,0}^k$ solve (39.16). Let $\boldsymbol{\sigma}_{\mathcal{T}_h}$ be s.t. $\boldsymbol{\sigma}_K := -\nabla \mathsf{R}(\hat{u}_K)$ for all $K \in \mathcal{T}_h$. Note that $\boldsymbol{\sigma}_K \in \boldsymbol{S}_K$. For all $\boldsymbol{\tau}_K := \nabla q \in \boldsymbol{S}_K = \nabla V_K^{k+1}$ with $q \in V_K^{k+1}$, using the definition (39.2) of R shows that $\boldsymbol{\sigma}_K$ solves (39.25a). Let now w_K be an arbitrary function in V_K^k. Since (39.2) implies that $(\boldsymbol{\sigma}_K, \nabla \mathsf{R}(w_K, 0))_{\boldsymbol{L}^2(K)} = (\boldsymbol{\sigma}_K, \nabla w_K)_{\boldsymbol{L}^2(K)} - (\boldsymbol{\sigma}_K \cdot \boldsymbol{n}_K, w_K)_{L^2(\partial K)}$, we have

$$
\begin{aligned}
(f, w_K)_{L^2(K)} &= \hat{a}_K(\hat{u}_K, (w_K, 0)) \\
&= -(\boldsymbol{\sigma}_K, \nabla \mathsf{R}(w_K, 0))_{\boldsymbol{L}^2(K)} + h_K^{-1}(\tilde{\mathsf{S}}(u_{K|\partial K} - u_{\partial K}), \tilde{\mathsf{S}}(w_{K|\partial K}))_{L^2(\partial K)} \\
&= -(\boldsymbol{\sigma}_K, \nabla w_K)_{\boldsymbol{L}^2(K)} + (\boldsymbol{\sigma}_K \cdot \boldsymbol{n}_K + \tau_{\partial K}(u_{K|\partial K} - u_{\partial K}), w_K)_{L^2(\partial K)}.
\end{aligned}
$$

This shows that (39.25b) holds true with $\boldsymbol{\phi}_{\partial K}$ defined in (39.26). Finally, let μ_F be an arbitrary function in V_F and let us denote by $\tilde{\mu}_F$ the extension by zero of μ_F to all the faces in $\mathcal{F}_h$ except F. By definition, we have $(\nabla \mathsf{R}(0, \tilde{\mu}_F), \nabla w_K)_{\boldsymbol{L}^2(K)} = (\mu_F, \boldsymbol{n}_K \cdot \nabla w_K)_{L^2(F)}$ for all $w_K \in V_K$. Hence, letting $\mathcal{T}_F := \{ K \in \mathcal{T}_h \mid F \in \mathcal{F}_K \}$, the proof of (39.25c) follows from

$$
\begin{aligned}
0 &= -\sum_{K \in \mathcal{T}_F} \hat{a}_K(\hat{u}_K, (0, \tilde{\mu}_F)) \\
&= \sum_{K \in \mathcal{T}_F} (\boldsymbol{\sigma}_K, \nabla \mathsf{R}(0, \tilde{\mu}_F))_{\boldsymbol{L}^2(K)} + h_K^{-1}(\tilde{\mathsf{S}}(u_{K|\partial K} - u_{\partial K}), \tilde{\mathsf{S}}(\tilde{\mu}_F))_{L^2(\partial K)} \\
&= \sum_{K \in \mathcal{T}_F} (\boldsymbol{\phi}_{h|\partial K} \cdot \boldsymbol{n}_K, \tilde{\mu}_F)_{L^2(\partial K)} = (\llbracket \boldsymbol{\phi}_h \rrbracket \cdot \boldsymbol{n}_F, \mu_F)_{L^2(F)}. \qquad \square
\end{aligned}
$$

Remark 39.14 (Flux recovery). Proposition 39.13 shows that one can post-process the HHO method by computing the numerical flux traces

$$
\boldsymbol{\phi}_{\partial K}(\hat{u}_K) := -\nabla \mathsf{R}(\hat{u}_K)_{|\partial K} + h_K^{-1}(\tilde{\mathsf{S}}^* \tilde{\mathsf{S}}(u_{K|\partial K} - u_{\partial K})) \boldsymbol{n}_K. \tag{39.27}
$$

Defining the global flux trace $\boldsymbol{\phi}_h(\hat{u}_h)_{|\partial K} := \boldsymbol{\phi}_{\partial K}(\hat{u}_K)$ for all $K \in \mathcal{T}_h$ gives $(\llbracket \boldsymbol{\phi}_h(\hat{u}_h) \rrbracket \cdot \boldsymbol{n}_F, \mu_F)_{L^2(F)} = 0$ for all $\mu_F \in V_F$. Since both factors are polynomials of degree at most k, we infer that $\llbracket \boldsymbol{\phi}_h(\hat{u}_h) \rrbracket \cdot \boldsymbol{n}_F = 0$. Finally, the above flux traces can be lifted as Raviart–Thomas vector-valued functions defined in the mesh cells as was done for HDG methods in Cockburn et al. [135]. $\square$

Remark 39.15 (Literature). HDG methods were introduced in Cockburn et al. [134]; see also [130] for a review and [135] for the convergence analysis. The link between the HHO and HDG methods is explored in Cockburn et al. [137]. With the simple polynomial spaces defined in (39.1), the HHO stabilization operator yields optimal error estimates for all $k \geq 0$ even on polyhedral meshes. Achieving this result with the HDG method with the simpler operator $\mathsf{S}^k := \Pi_{\partial K}^k(v_{K|\partial K} - v_{\partial K})$ on polyhedral meshes requires a subtle design of the local spaces; see Cockburn et al. [136]. $\square$

39.3 Error analysis

This section is devoted to the error analysis of the HHO method. We adopt a point of view similar to that of Lemma 27.5, where the notions of stability and consistency/boundedness were

essential. Since stability has been established in Lemma 39.8, we now turn our attention to consistency/boundedness. We slightly adapt the notion of the consistency error since the solution to (39.12) is a function defined on D, whereas the discrete solution is a pair composed of a function defined on D and a function defined on the mesh faces. Let $\hat{\mathcal{I}}_h^k : V := H_0^1(D) \to \hat{V}_{h,0}^k$ be the global interpolation operator s.t. for all $v \in V$, $\hat{\mathcal{I}}_h^k(v) \in \hat{V}_{h,0}^k$ is specified as follows: For all $K \in \mathcal{T}_h$,

$$((\hat{\mathcal{I}}_h^k(v))_K, (\hat{\mathcal{I}}_h^k(v))_{\partial K}) := \hat{\mathcal{I}}_K^k(v_{|K}) = (\Pi_K^k(v_{|K}), \Pi_{\partial K}^k(v_{|\partial K})) \in \hat{V}_K^k. \tag{39.28}$$

Notice that $\hat{\mathcal{I}}_h^k(v)$ is well defined in $\hat{V}_{h,0}^k$ since v has zero jumps across the mesh interfaces (see Theorem 18.8) and zero traces at the boundary faces. We define the consistency error $\delta_h(\hat{v}_h) \in (\hat{V}_{h,0}^k)'$ s.t. for all $\hat{v}_h, \hat{w}_h \in \hat{V}_{h,0}^k$,

$$\langle \delta_h(\hat{v}_h), \hat{w}_h \rangle_{(\hat{V}_{h,0}^k)', \hat{V}_{h,0}^k} := \ell_h(w_{\mathcal{T}_h}) - \hat{a}_h(\hat{v}_h, \hat{w}_h). \tag{39.29}$$

To avoid distracting technicalities, we assume in the error analysis that $u \in H^{1+r}(D)$, $r > \frac{1}{2}$. This assumption can be removed as discussed in §41.5. Recall from Lemma 39.1 the local elliptic projection $\mathcal{E}_K : H^1(K) \to V_K^{k+1}$.

Lemma 39.16 (Consistency/boundedness). *Assume that the solution to (39.12) satisfies $u \in H^{1+r}(D) \cap H_0^1(D)$, $r > \frac{1}{2}$. There is c s.t. for all $h \in \mathcal{H}$,*

$$\|\delta_h(\hat{\mathcal{I}}_h^k(u))\|_{(\hat{V}_{h,0}^k)'}^2 \le c \sum_{K \in \mathcal{T}_h} \|u - \mathcal{E}_K(u)\|_{\sharp,K}^2, \tag{39.30}$$

where we defined for all $K \in \mathcal{T}_h$ and all $v \in H^{1+r}(K)$, $r > \frac{1}{2}$,

$$\|v\|_{\sharp,K} := \|\nabla v\|_{\mathbf{L}^2(K)} + h_K^{\frac{1}{2}} \|\nabla v\|_{\mathbf{L}^2(\partial K)}. \tag{39.31}$$

Proof. Let $\hat{w}_h \in \hat{V}_{h,0}^k$. Integrating by parts cellwise, we observe that

$$\begin{aligned}
\ell_h(w_{\mathcal{T}_h}) &= \sum_{K \in \mathcal{T}_h} \ell_K(w_K) = \sum_{K \in \mathcal{T}_h} -(\Delta u, w_K)_{L^2(K)} \\
&= \sum_{K \in \mathcal{T}_h} \left((\nabla u, \nabla w_K)_{\mathbf{L}^2(K)} - (\mathbf{n}_K \cdot \nabla u, w_K)_{L^2(\partial K)} \right) \\
&= \sum_{K \in \mathcal{T}_h} \left((\nabla u, \nabla w_K)_{\mathbf{L}^2(K)} - (\mathbf{n}_K \cdot \nabla u, w_K - w_{\partial K})_{L^2(\partial K)} \right),
\end{aligned}$$

where we used that $\sum_{K \in \mathcal{T}_h} (\mathbf{n}_K \cdot \nabla u, w_{\partial K})_{L^2(\partial K)} = 0$ since ∇u and $w_{\mathcal{F}_h}$ are single-valued on the mesh interfaces and $w_{\mathcal{F}_h}$ vanishes at the boundary faces. Moreover, since $\mathcal{E}_K = \mathsf{R} \circ \hat{\mathcal{I}}_K^k$, using the definition of $\mathsf{R}(\hat{w}_K)$ leads to

$$\begin{aligned}
(\nabla \mathsf{R}(\hat{\mathcal{I}}_K^k(u)), \nabla \mathsf{R}(\hat{w}_K))_{\mathbf{L}^2(K)} &= (\nabla \mathcal{E}_K(u), \nabla \mathsf{R}(\hat{w}_K))_{\mathbf{L}^2(K)} \\
&= (\nabla \mathcal{E}_K(u), \nabla w_K)_{\mathbf{L}^2(K)} - (\mathbf{n}_K \cdot \nabla \mathcal{E}_K(u), w_K - w_{\partial K})_{L^2(\partial K)}.
\end{aligned}$$

Using the definition of $\hat{a}_K$ and since $(\nabla(u - \mathcal{E}_K(u)), \nabla w_K)_{\mathbf{L}^2(K)} = 0$, we have

$$\langle \delta_h(\hat{\mathcal{I}}_h^k(u)), \hat{w}_h \rangle_{(\hat{V}_{h,0}^k)', \hat{V}_{h,0}^k} = -\sum_{K \in \mathcal{T}_h} (\mathfrak{T}_{1,K} + \mathfrak{T}_{2,K})$$

with

$$\mathfrak{T}_{1,K} := (\boldsymbol{n}_K \cdot \nabla \eta_K, w_K - w_{\partial K})_{L^2(\partial K)},$$
$$\mathfrak{T}_{2,K} := h_K^{-1}(\mathsf{S}(\hat{\mathcal{I}}_K^k(u)), \mathsf{S}(\hat{w}_K))_{L^2(\partial K)},$$

and $\eta_K := u_{|K} - \mathcal{E}_K(u)$ for all $K \in \mathcal{T}_h$. The Cauchy–Schwarz inequality and the definition of $|\hat{w}_K|_{\hat{V}_K^k}$ imply that

$$|\mathfrak{T}_{1,K}| \leq \|\nabla \eta_K\|_{\boldsymbol{L}^2(\partial K)} \|w_K - w_{\partial K}\|_{L^2(\partial K)} \leq h_K^{\frac{1}{2}} \|\nabla \eta_K\|_{\boldsymbol{L}^2(\partial K)} |\hat{w}_K|_{\hat{V}_K^k}.$$

Moreover, $|\mathfrak{T}_{2,K}| \leq h_K^{-\frac{1}{2}} \|\mathsf{S}(\hat{\mathcal{I}}_K^k(u))\|_{L^2(\partial K)} h_K^{-\frac{1}{2}} \|\mathsf{S}(\hat{w}_K)\|_{L^2(\partial K)}$ owing to the Cauchy–Schwarz inequality. The first factor is bounded in Lemma 39.3, and the second one in Lemma 39.2. Hence, $|\mathfrak{T}_{2,K}| \leq c \|\nabla \eta_K\|_{\boldsymbol{L}^2(K)} |\hat{w}_K|_{\hat{V}_K^k}$. Collecting these bounds and summing over the mesh cells leads to (39.30). $\square$

Theorem 39.17 (Error estimate). *Let u be the solution to (39.12) and let $\hat{u}_h := (u_{\mathcal{T}_h}, u_{\mathcal{F}_h}) \in \hat{V}_{h,0}^k$ solve (39.16). Assume that $u \in H^{1+r}(D) \cap H_0^1(D)$, $r > \frac{1}{2}$. Recall the notation $u_K := u_{\mathcal{T}_h|K}$, $u_{\partial K} := u_{\mathcal{F}_h|\partial K}$, and $\hat{u}_K := (u_K, u_{\partial K})$ for all $K \in \mathcal{T}_h$. (i) There is c s.t. for all $h \in \mathcal{H}$,*

$$\sum_{K \in \mathcal{T}_h} \|\nabla(u - \mathsf{R}(\hat{u}_K))\|_{\boldsymbol{L}^2(K)}^2 \leq c \sum_{K \in \mathcal{T}_h} \|u - \mathcal{E}_K(u)\|_{\sharp,K}^2. \tag{39.32}$$

(ii) Letting $t := \min(k+1, r)$, we have

$$\sum_{K \in \mathcal{T}_h} \|\nabla(u - \mathsf{R}(\hat{u}_K))\|_{\boldsymbol{L}^2(K)}^2 \leq c \sum_{K \in \mathcal{T}_h} h_K^{2t} |u|_{H^{1+t}(K)}^2. \tag{39.33}$$

Proof. (i) We adapt the proof of Lemma 27.5 to account for the use of the reconstruction operator. Set $\hat{\zeta}_h^k := \hat{\mathcal{I}}_h^k(u) - \hat{u}_h \in \hat{V}_{h,0}^k$ so that $\hat{\zeta}_K^k = \hat{\mathcal{I}}_K^k(u_{|K}) - \hat{u}_K$ for all $K \in \mathcal{T}_h$. Notice that $a_h(\hat{\zeta}_h^k, \hat{\zeta}_h^k) = -\delta_h(\hat{\mathcal{I}}_h^k(u)), \hat{\zeta}_h^k\rangle_{(\hat{V}_{h,0}^k)', \hat{V}_{h,0}^k}$. Then the coercivity property (39.18) implies that

$$\alpha \sum_{K \in \mathcal{T}_h} \|\nabla \mathsf{R}(\hat{\zeta}_K^k)\|_{\boldsymbol{L}^2(K)}^2 \leq \frac{\hat{a}_h(\hat{\zeta}_h^k, \hat{\zeta}_h^k)}{\|\hat{\zeta}_h^k\|_{\hat{V}_{h,0}^k}^2} \sum_{K \in \mathcal{T}_h} \|\nabla \mathsf{R}(\hat{\zeta}_K^k)\|_{\boldsymbol{L}^2(K)}^2$$

$$\leq \frac{\hat{a}_h(\hat{\zeta}_h^k, \hat{\zeta}_h^k)^2}{\|\hat{\zeta}_h^k\|_{\hat{V}_{h,0}^k}^2} = \frac{\langle \delta_h(\hat{\mathcal{I}}_h^k(u)), \hat{\zeta}_h^k\rangle_{(\hat{V}_{h,0}^k)', \hat{V}_{h,0}^k}^2}{\|\hat{\zeta}_h^k\|_{\hat{V}_{h,0}^k}^2} \leq \|\delta_h(\hat{\mathcal{I}}_h^k(u))\|_{(\hat{V}_{h,0}^k)'}^2.$$

Lemma 39.16 yields $\sum_{K \in \mathcal{T}_h} \|\nabla \mathsf{R}(\hat{\zeta}_K^k)\|_{\boldsymbol{L}^2(K)}^2 \leq c \sum_{K \in \mathcal{T}_h} \|u - \mathcal{E}_K(u)\|_{\sharp,K}^2$. Since $\mathsf{R}(\hat{\mathcal{I}}_K^k(u)) = \mathcal{E}_K(u)$ for all $K \in \mathcal{T}_h$, we infer that

$$u_{|K} - \mathsf{R}(\hat{u}_K) = u_{|K} - \mathcal{E}_K(u) + \mathsf{R}(\hat{\zeta}_K^k).$$

Then the estimate (39.32) follows from the triangle inequality.
(ii) The estimate (39.33) results from the approximation properties of the local elliptic projection; see Exercise 39.3. $\square$

Remark 39.18 (L^2-estimate). Improved L^2-error estimates can be established if elliptic regularity pickup can be invoked; see [168]. $\square$

Exercises

Exercise 39.1 (Stabilization). Prove that $\hat{a}_K(\hat{v}_K, \hat{v}_K)$ is equivalent to $\|\nabla r_K\|^2_{\boldsymbol{L}^2(K)} + \hat{\theta}_K(\hat{v}_K, \hat{v}_K)$ for all $\hat{v}_K \in \hat{V}_K^k$, with $r_K := \mathsf{R}(\hat{v}_K)$ and

$$\hat{\theta}_K(\hat{v}_K, \hat{v}_K) := h_K^{-2}\|v_K - \Pi_K^k(r_K)\|^2_{L^2(K)} + h_K^{-1}\|v_{\partial K} - \Pi_{\partial K}^k(r_K)\|^2_{L^2(\partial K)}.$$

(*Hint*: note that $\mathsf{S}(\hat{v}_K) = \Pi_{\partial K}^k(v_K - \Pi_K^k(r_K))_{|\partial K} - (v_{\partial K} - \Pi_{\partial K}^k(r_K))$, and to bound $\hat{a}_K(\hat{v}_K, \hat{v}_K)$ from below, prove that $\hat{\theta}_K(\hat{v}_K, \hat{v}_K)^{\frac{1}{2}} \leq c\, h_K^{-1}\|v_K - r_K\|_{L^2(K)} + h_K^{-\frac{1}{2}}\|\mathsf{S}(\hat{v}_K)\|_{L^2(\partial K)}$, then invoke the Poincaré–Steklov inequality, the triangle inequality, and the lower bound from Lemma 39.2.)

Exercise 39.2 (Finite element viewpoint). Let $\mathcal{V}_K^k$ be defined in (39.10). Let $\mathcal{E}_K : H^1(K) \to V_K^{k+1}$ be the elliptic projection and set $\delta := v - \mathcal{E}_K(v)$ for all $v \in \mathcal{V}_K^k$. (i) Prove that

$$h_K^{-1}\|\Pi_K^k(\delta)\|_{L^2(K)} \leq c\left(\|\nabla\mathcal{E}_K(v)\|_{\boldsymbol{L}^2(K)} + h_K^{-\frac{1}{2}}\|\mathsf{S}(\hat{\mathcal{I}}_K^k(v))\|_{L^2(\partial K)}\right).$$

(*Hint*: use the Poincaré–Steklov inequality in K and the lower bound from Lemma 39.2.) (ii) Prove that

$$\|\nabla\delta\|_{\boldsymbol{L}^2(K)} \leq c\left(\|\nabla\mathcal{E}_K(v)\|_{\boldsymbol{L}^2(K)} + h_K^{-\frac{1}{2}}\|\mathsf{S}(\hat{\mathcal{I}}_K^k(v))\|_{L^2(\partial K)}\right).$$

(*Hint*: integrate by parts $\|\nabla\delta\|^2_{\boldsymbol{L}^2(K)}$ and accept as a fact that a discrete trace inequality and an inverse inequality are valid on $\mathcal{V}_K^k$, then use that $\mathsf{S}(\hat{\mathcal{I}}_K^k(v)) = \Pi_{\partial K}^k(\Pi_K^k(\delta)_{|\partial K}) - \Pi_{\partial K}^k(\delta_{|\partial K})$.) (iii) Let $\mathfrak{a}_K(v, w) := (\nabla\mathcal{E}_K(v), \nabla\mathcal{E}_K(w))_{\boldsymbol{L}^2(K)} + h_K^{-1}(\mathsf{S}(\hat{\mathcal{I}}_K^k(v)), \mathsf{S}(\hat{\mathcal{I}}_K^k(v)))_{L^2(\partial K)}$ on $\mathcal{V}_K^k \times \mathcal{V}_K^k$. Prove that $\mathfrak{a}_K(v, v) \geq c\|\nabla v\|^2_{\boldsymbol{L}^2(K)}$ with $c > 0$.

Exercise 39.3 (Elliptic projection). Prove the second bound in Theorem 39.17. (*Hint*: introduce the L^2-orthogonal projection Π_K^{k+1}.)

Exercise 39.4 (Reconstruction). (i) Let $\mathsf{G} : \hat{V}_K^k \to \boldsymbol{V}_K^k := \mathbb{P}_{k,d} \circ \boldsymbol{T}_K^{-1}$ be s.t. $(\mathsf{G}(\hat{v}_K), \boldsymbol{q})_{\boldsymbol{L}^2(K)} = -(v_K, \nabla\cdot\boldsymbol{q})_{L^2(K)} + (v_{\partial K}, \boldsymbol{n}_K\cdot\boldsymbol{q})_{L^2(\partial K)}$ for all $\boldsymbol{q} \in \boldsymbol{V}_K^k$. Prove that $\Pi_{\nabla V_K^{k+1}}\mathsf{G} = \nabla\mathsf{R}$, where $\Pi_{\nabla V_K^{k+1}}$ is the $\boldsymbol{L}^2$-orthogonal projection onto ∇V_K^{k+1}. (ii) Let $\mathsf{G}_{\mathrm{RT}} : \hat{V}_K^k \to \boldsymbol{V}_K^k := (\boldsymbol{\psi}_K^{\mathrm{d}})^{-1}(\mathbb{R}\mathbb{T}_{k,d})$ be s.t. $(\mathsf{G}_{\mathrm{RT}}(\hat{v}_K), \boldsymbol{q})_{\boldsymbol{L}^2(K)} = -(v_K, \nabla\cdot\boldsymbol{q})_{L^2(K)} + (v_{\partial K}, \boldsymbol{n}_K\cdot\boldsymbol{q})_{L^2(\partial K)}$ for all $\boldsymbol{q} \in \boldsymbol{V}_K^k$, where $\boldsymbol{\psi}_K^{\mathrm{d}}$ is the contravariant Piola transformation defined in (9.9c), and $\mathbb{R}\mathbb{T}_{k,d}$ is the Raviart–Thomas polynomial space. Prove that $\|\mathsf{G}_{\mathrm{RT}}(\hat{v}_K)\|_{\boldsymbol{L}^2(K)} \geq c|\hat{v}_K|_{\hat{V}_K^k}$ with $c > 0$. (*Hint*: use the dofs of the Raviart–Thomas element; see John et al. [260] for the seminal idea in the context of dG methods.)

Exercise 39.5 ($k = 0$). (i) Derive the HHO method in 1D for $k = 0$, as well as the global transmission problem. (ii) Prove that, in dimension $d \geq 2$ for $k = 0$, $\mathsf{R}(\hat{v}_K)(\boldsymbol{x}) = v_K + \sum_{F\in\mathcal{F}_K}\frac{|F|}{|K|}(v_F - v_K)\boldsymbol{n}_{K|F}\cdot(\boldsymbol{x} - \boldsymbol{x}_K)$ for all $\boldsymbol{x} \in K$, with $v_F := v_{\partial K|F}$ for all $F \in \mathcal{F}_K$, and $\boldsymbol{x}_K$ is the barycenter of K, and $\mathsf{S}(\hat{v}_K)_{|F} = v_K - v_F - \nabla\mathsf{R}(\hat{v}_K)\cdot(\boldsymbol{x}_K - \boldsymbol{x}_F)$, where $\boldsymbol{x}_F$ is the barycenter of F for all $F \in \mathcal{F}_K$ (*Hint*: any function $q \in \mathbb{P}_{1,d} \circ \boldsymbol{T}_K^{-1}$ is of the form $q(\boldsymbol{x}) = q_K + \boldsymbol{G}_q\cdot(\boldsymbol{x} - \boldsymbol{x}_K)$, where $q_K := q(\boldsymbol{x}_K)$ is the mean value of q over K and $\boldsymbol{G}_q := \nabla q$, and use also (7.1).)

Exercise 39.6 (Transmission problem). (i) Prove the converse statement in Proposition 39.10. (*Hint*: write $\hat{w}_K = (w_K - U_{w_{\partial K}}, 0) + (U_{w_{\partial K}}, w_{\partial K})$.) (ii) Justify Remark 39.11. (*Hint*: for the converse statement show that $a_K(u, w) - \ell_K(w) = a_K(U_{\lambda_{\partial K}}, U_\mu) - \ell_K(U_\mu)$ with $\mu := w_{\partial K}$.) (iii) Adapt the statement if a_K is nonsymmetric. (*Hint*: consider $U_\lambda^* \in H^1(K)$ s.t. $U_{\lambda|\partial K}^* = \lambda$ and $a_K(\psi, U_\lambda^*) = 0$ for all $\psi \in H_0^1(K)$.) (iv) Prove (39.23).

Exercise 39.7 (HDG). Consider the HDG method. Assume the following: if $(v_K, \mu_{\partial K}) \in V_K \times V_{\partial K}$ with $V_{\partial K} := \prod_{F \in \mathcal{F}_K} V_F$ is s.t. $(\tau_{\partial K}(v_{K|\partial K} - \mu_{\partial K}), v_{K|\partial K} - \mu_{\partial K})_{L^2(\partial K)} = 0$ and $(v_K, \nabla \cdot \boldsymbol{\tau}_K)_{L^2(K)} - (\mu_{\partial K}, \boldsymbol{\tau}_K \cdot \boldsymbol{n}_K)_{L^2(\partial K)} = 0$ for all $\boldsymbol{\tau}_K \in \boldsymbol{S}_K$, then v_K and $\mu_{\partial K}$ are constant functions taking the same value. Prove that the discrete problem (39.25) is well-posed. (*Hint*: derive an energy identity.)

Exercise 39.8 (Space Λ). Let Λ be defined in (39.21). Recall that the trace map $\gamma^{\mathrm{g}}_{\partial K} : H^1(K) \to H^{\frac{1}{2}}(\partial K)$ is surjective. (i) Prove that there are constants $0 < c_1 \le c_2$ s.t. $c_1 \|\nabla U_\mu\|_{\boldsymbol{L}^2(K)} \le |\mu|_{H^{\frac{1}{2}}(\partial K)} \le c_2 \|\nabla U_\mu\|_{\boldsymbol{L}^2(K)}$ for all $\mu \in H^{\frac{1}{2}}(\partial K)$, all $K \in \mathcal{T}_h$, and all $h \in \mathcal{H}$. (*Hint*: prove first the bounds on the reference cell $\widehat{K}$.) (ii) Set $\|\lambda\|^2_\Lambda := \sum_{K \in \mathcal{T}_h} |\lambda_{\partial K}|^2_{H^{\frac{1}{2}}(\partial K)}$. Verify that $\|\cdot\|_\Lambda$ indeed defines a norm on Λ, and that Λ is a Hilbert space. (*Hint*: for all $\lambda \in \Lambda$, consider the function $U_\lambda : D \to \mathbb{R}$ s.t. $U_{\lambda|K} := U_{\lambda_{\partial K}}$ for all $K \in \mathcal{T}_h$, and prove that $U_\lambda \in H^1_0(D)$.)

Exercise 39.9 (Liftings, 1D). Consider a uniform mesh of $D := (0, 1)$ with nodes $x_i := ih$, $i := \frac{1}{I+1}$ for all $i \in \{0{:}(I+1)\}$. Consider the PDE $-u'' = f$ in D with $u(0) = u(1) = 0$. (i) Prove that (39.22) amounts to $\mathcal{A}X = B$ with $\mathcal{A} = h^{-1}\mathrm{tridiag}(-1, 2, -1)$, $X_i = \lambda_i$, and $B_i = \int_{x_{i-1}}^{x_{i+1}} \varphi_i f \, \mathrm{d}s$ for all $i \in \{1{:}I\}$. (*Hint*: prove that U_λ is affine on every cell $K_i = [x_{i-1}, x_i]$.) Prove that $\lambda_i = u(x_i)$. (*Hint*: write $f = -u''$ and integrate by parts. This remarkable fact only happens in 1D.) (ii) Let $k \ge 2$. For all $m \ge 1$, set $\phi_m := (2(2m+1))^{-\frac{1}{2}}(L_{m+1} - L_{m-1})$, where L_m is the Legendre polynomial of degree m (see §6.1). Verify that $\{\phi_m\}_{m \in \{1:k-1\}}$ is a basis of $\mathbb{P}^\circ_k := \{p \in \mathbb{P}_k \mid p(\pm 1) = 0\}$. Prove that $U_{f_{|\widehat{K}}}(x) = \int_{\widehat{K}} G(x, s) f(s) \, \mathrm{d}s$ on $\widehat{K} := [-1, 1]$ with the discrete Green's function $G(x, s) := \sum_{m \in \{1:k-1\}} \phi_m(x) \phi_m(s)$. (*Hint*: observe that $\phi'_m = L_m$.) Infer the expression of $U_{f_{|K_i}}$ for every cell K_i.

Chapter 40

Contrasted diffusivity (I)

The goal of Chapters 40 and 41 is to investigate the approximation of a diffusion model problem with contrasted diffusivity and revisit the error analysis of the various nonconforming approximation methods presented in the previous chapters. The essential difficulty is that the elliptic regularity theory (see §31.4) tells us that the Sobolev smoothness index of the solution u may be just barely larger than one. For this reason, we are going to perform the error analysis under the assumption that $u \in H^{1+r}(D)$, $r > 0$ (recall that we assumed $r > \frac{1}{2}$ in the previous chapters). This will be done by using again the abstract error analysis from §27.2, but the lack of smoothness will require that we invoke the tools devised in Chapter 17 to give a proper meaning to the normal derivative of the solution at the mesh faces. We assume that $d \geq 2$ since the analysis is much simpler if $d = 1$.

40.1 Model problem

Let D be a Lipschitz domain in $\mathbb{R}^d$, which we assume for simplicity to be a polyhedron. We consider the following scalar model problem:

$$-\nabla{\cdot}(\lambda\nabla u) = f \quad \text{in } D, \qquad \gamma^{\mathrm{g}}(u) = g \quad \text{on } \partial D, \tag{40.1}$$

with the trace map $\gamma^{\mathrm{g}} : H^1(D) \to H^{\frac{1}{2}}(\partial D)$, the Dirichlet boundary data $g \in H^{\frac{1}{2}}(\partial D)$, and the scalar-valued diffusion coefficient $\lambda \in L^\infty(D)$ which we assume to be uniformly bounded from below away from zero. For simplicity, we also assume that λ is piecewise constant in D, i.e., there is a partition of D into M disjoint Lipschitz polyhedra $\{D_i\}_{i\in\{1:M\}}$ s.t. $\lambda_{|D_i}$ is a positive real number for all $i \in \{1{:}M\}$. A central notion in this chapter is the diffusive flux which is defined as follows:

$$\boldsymbol{\sigma}(v) := -\lambda\nabla v \in \boldsymbol{L}^2(D), \qquad \forall v \in H^1(D). \tag{40.2}$$

In the previous chapters, we considered elliptic PDEs with a source term $f \in L^2(D)$. We are now going to relax a bit this hypothesis by only assuming that $f \in L^q(D)$ with $q > \frac{2d}{2+d}$. Note that $q > 1$ since $d \geq 2$. Since $\frac{2d}{2+d} < 2$, we are going to assume without loss of generality that $q \leq 2$ in the entire chapter. Readers who wish to simplify some arguments can think that $q = 2$ in what follows.

In the case of the homogeneous Dirichlet condition ($g := 0$), the weak formulation of the model problem (40.1) is as follows:

$$\begin{cases} \text{Find } u \in V := H_0^1(D) \text{ such that} \\ a(u, w) = \ell(w), \quad \forall w \in V, \end{cases} \tag{40.3}$$

with the bilinear and linear forms

$$a(v, w) := \int_D \lambda \nabla v \cdot \nabla w \, dx, \qquad \ell(w) := \int_D f w \, dx. \tag{40.4}$$

The bilinear form a is coercive on V owing to the Poincaré–Steklov inequality, and it is also bounded on $V \times V$ owing to the Cauchy–Schwarz inequality. The linear form ℓ is bounded on V since the Sobolev embedding theorem (Theorem 2.31) and Hölder's inequality imply that $|\ell(w)| \leq \|f\|_{L^q(D)} \|w\|_{L^{q'}(D)} \leq c\|f\|_{L^q(D)} \|w\|_{H^1(D)}$ with $\frac{1}{q} + \frac{1}{q'} = 1$. Note that $q \geq \frac{2d}{2+d}$ is the minimal integrability requirement on f for this boundedness property to hold true. The above coercivity and boundedness properties combined with the Lax–Milgram lemma imply that (40.3) is well-posed. For the non-homogeneous Dirichlet boundary condition, one invokes the surjectivity of the trace map γ^{g} to infer the existence of a lifting of g, say $u_g \in H^1(D)$, and one decomposes the solution to (40.1) as $u := u_g + u_0$, where $u_0 \in H_0^1(D)$ solves the weak problem (40.3) with $\ell(w)$ replaced by $\ell_g(w) := \ell(w) - \widetilde{a}(u_g, w)$ with $\widetilde{a}(u, v) := \int_D \lambda \nabla v \cdot \nabla w \, dx$. The weak formulation thus modified is well-posed since ℓ_g is bounded on $H_0^1(D)$.

Lemma 40.1 (A priori regularity). *If the solution to (40.3) is s.t. $u \in H^{1+r}(D)$, $r > 0$, and if the source term f is in $L^q(D)$, $\frac{2d}{2+d} < q \leq 2$, then*

$$u \in V_{\mathrm{s}} := \{v \in H_0^1(D) \mid \boldsymbol{\sigma}(v) \in \boldsymbol{L}^p(D), \, \nabla \cdot \boldsymbol{\sigma}(v) \in L^q(D)\}, \tag{40.5}$$

where the real numbers p, q are such that

$$2 < p, \qquad \frac{2d}{2+d} < q \leq 2. \tag{40.6}$$

Proof. The Sobolev embedding theorem implies that there is $p > 2$ s.t. $\boldsymbol{H}^r(D) \hookrightarrow \boldsymbol{L}^p(D)$. Indeed, if $2r < d$, we have $\boldsymbol{H}^r(D) \hookrightarrow \boldsymbol{L}^s(D)$ for all $s \in [2, \frac{2d}{d-2r}]$ and we can take $p := \frac{2d}{d-2r}$, whereas if $2r \geq d$, we have $\boldsymbol{H}^r(D) \hookrightarrow \boldsymbol{H}^{\frac{d}{2}}(D) \hookrightarrow \boldsymbol{L}^s(D)$ for all $s \in [2, \infty)$, and we can take any $p > 2$. The above argument implies that $\nabla u \in \boldsymbol{L}^p(D)$, and since λ is piecewise constant and $\boldsymbol{\sigma}(u) = -\lambda \nabla u$, we have $\boldsymbol{\sigma}(u) \in \boldsymbol{L}^p(D)$. Since $\nabla \cdot \boldsymbol{\sigma}(u) = f$ and $f \in L^q(D)$ with $q > \frac{2d}{2+d}$ by assumption, we have $\nabla \cdot \boldsymbol{\sigma}(u) \in L^q(D)$. $\qquad \square$

The smoothness assumption $u \in H^{1+r}(D)$, $r > 0$, is reasonable owing to the elliptic regularity theory (see Theorem 31.36). In general, one expects that $r \leq \frac{1}{2}$ whenever u is supported in at least two contiguous subdomains where λ takes different values since otherwise the normal derivative of u would be continuous across the interface separating the two subdomains in question, and owing to the discontinuity of λ, the normal component of the diffusive flux $\boldsymbol{\sigma}(u)$ would be discontinuous across the interface, which would contradict the fact that $\boldsymbol{\sigma}(u)$ has a weak divergence. It is possible that $r > \frac{1}{2}$ when u is supported in one subdomain only. If $r \geq 1$, we notice that we must have $f \in L^2(D)$ (since $f_{|D_i} = -\lambda_{|D_i}(\Delta u)_{|D_i}$ for all $i \in \{1:M\}$), i.e., we can assume that $q = 2$ if $r \geq 1$.

Remark 40.2 (Extensions). One can also consider lower-order terms in (40.1), e.g., $-\nabla \cdot (\lambda \nabla u) + \boldsymbol{\beta} \cdot \nabla u + \mu u = f$ with $\boldsymbol{\beta} \in \boldsymbol{W}^{1,\infty}(D)$ and $\mu \in L^\infty(D)$ s.t. $\mu - \frac{1}{2} \nabla \cdot \boldsymbol{\beta} \geq 0$ a.e. in D (for simplicity). The present error analysis still applies provided the lower-order terms are not too large, e.g.,

$\lambda \geq \max(h\|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(D)}, h^2\|\mu\|_{L^\infty(D)})$, where $h \in \mathcal{H}$ denotes the meshsize. Stabilization techniques like those discussed in Chapter 61 have to be invoked when the lower-order terms are large. Furthermore, the error analysis can be extended to account for a piecewise constant tensor-valued diffusivity $\mathbb{d}$. Then the various constants in the error estimate depend on the square-root of the anisotropy ratios measuring the contrast between the largest and the smallest eigenvalue of $\mathbb{d}$ in each subdomain D_i. Finally, one can consider that the diffusion tensor $\mathbb{d}$ is piecewise smooth instead of being piecewise constant, and a reasonable requirement is that $\mathbb{d}_{|D_i}$ is Lipschitz for all $i \in \{1:M\}$. Notice though that this last extension entails some subtleties in the analysis because the discrete diffusive flux is no longer a piecewise polynomial function. $\square$

40.2   Discrete setting

We introduce in this section the discrete setting that we are going to use to approximate the solution to (40.3). Let $\mathcal{T}_h$ be a mesh from a shape-regular sequence. We assume that $\mathcal{T}_h$ is oriented in a generation-compatible way and that $\mathcal{T}_h$ covers each of the subdomains $\{D_i\}_{i\in\{1:M\}}$ exactly, so that $\lambda_K := \lambda_{|K}$ is constant for all $K \in \mathcal{T}_h$. Let $(\widehat{K}, \widehat{P}, \widehat{\Sigma})$ be the reference finite element. We assume that $\mathbb{P}_{k,d} \subset \widehat{P} \subset W^{k+1,\infty}(\widehat{K})$ for some $k \geq 1$. For all $K \in \mathcal{T}_h$, let $\boldsymbol{T}_K : \widehat{K} \to K$ be the geometric mapping and let $\psi_K^{\mathrm{g}}(v) := v \circ \boldsymbol{T}_K$ be the pullback by the geometric mapping. Recall the broken finite element space defined as

$$P_k^{\mathrm{b}}(\mathcal{T}_h) := \{v_h \in L^\infty(D) \mid v_{h|K} \in P_K, \ \forall K \in \mathcal{T}_h\}, \tag{40.7}$$

with the local space $P_K := (\psi_K^{\mathrm{g}})^{-1}(\widehat{P}) \subset W^{k+1,\infty}(K)$. For all $v_h \in P_k^{\mathrm{b}}(\mathcal{T}_h)$, we define the broken diffusive flux $\boldsymbol{\sigma}(v_h) \in \boldsymbol{L}^2(D)$ by setting $\boldsymbol{\sigma}(v_h)_{|K} := -\lambda_K \nabla(v_{h|K})$ for all $K \in \mathcal{T}_h$. Recalling the notion of broken gradient (see Definition 36.3), we have $\boldsymbol{\sigma}(v_h) := -\lambda \nabla_h v_h$.

Recall that the set $\mathcal{F}_h^\circ$ is the collection of the mesh interfaces and the set $\mathcal{F}_h^\partial$ is the collection of the mesh faces at the boundary. For all $F \in \mathcal{F}_h^\circ$, there are two cells $K_l, K_r \in \mathcal{T}_h$ s.t. $F := \partial K_l \cap \partial K_r$, and F is oriented by the unit normal vector $\boldsymbol{n}_F$ pointing from K_l to K_r, i.e., $\boldsymbol{n}_F := \boldsymbol{n}_{K_l} = -\boldsymbol{n}_{K_r}$. For all $F \in \mathcal{F}_h^\partial$, we write $F := \partial K_l \cap \partial D$, and F is oriented by the unit normal vector pointing toward the outside of D, i.e., $\boldsymbol{n}_F := \boldsymbol{n}_{K_l} = \boldsymbol{n}$. For all $F \in \mathcal{F}_h$, let $\mathcal{T}_F$ be the collection of the one or two mesh cells sharing F, i.e., $\mathcal{T}_F := \{K_l, K_r\}$ for all $F \in \mathcal{F}_h^\circ$ and $\mathcal{T}_F := \{K_l\}$ for all $F \in \mathcal{F}_h^\partial$. For all $K \in \mathcal{T}_h$, let $\mathcal{F}_K$ be the collection of the faces of K and let $\epsilon_{K,F} := \boldsymbol{n}_K \cdot \boldsymbol{n}_F = \pm 1$. The jump of a function $v \in W^{1,1}(\mathcal{T}_h)$ across the mesh face $F \in \mathcal{F}_h$ is defined a.e. on F by setting $[\![v]\!]_F := v_{|K_l} - v_{|K_r}$ if $F := \partial K_l \cap \partial K_r \in \mathcal{F}_h^\circ$ (see §18.1.1) and $[\![v]\!]_F := v_{|K_l}$ if $F := \partial K_l \cap \partial D \in \mathcal{F}_h^\partial$. It is also useful to consider weighted averages at the mesh faces. For all $F \in \mathcal{F}_h^\circ$, we assume that we have at hand two real numbers such that

$$\theta_{K_l,F}, \theta_{K_r,F} \in [0,1] \quad \text{and} \quad \theta_{K_l,F} + \theta_{K_r,F} = 1. \tag{40.8}$$

We then set

$$\{v\}_{F,\theta} := \theta_{K_l,F} v_{|K_l} + \theta_{K_r,F} v_{|K_r}, \tag{40.9a}$$

$$\{v\}_{F,\bar{\theta}} := \theta_{K_r,F} v_{|K_l} + \theta_{K_l,F} v_{|K_r}. \tag{40.9b}$$

Whenever $\theta_{K_l,F} = \theta_{K_r,F} := \frac{1}{2}$, these two definitions coincide with the usual arithmetic average (see Definition 38.1). In order to use a common notation for interfaces and boundary faces, we write for all $F := \partial K_l \cap \partial D \in \mathcal{F}_h^\partial$, $\theta_{K_l,F} := 1$ and $\{v\}_\theta = v_{|K_l}$. We omit the subscript F in the jump

and the average whenever the context is unambiguous. The following identity (see Exercise 40.3) will be useful:

$$\llbracket vw \rrbracket = \{v\}_\theta \llbracket w \rrbracket + \llbracket v \rrbracket \{w\}_{\bar{\theta}}. \tag{40.10}$$

40.3   The bilinear form $n_\sharp$

In this section, we give a proper meaning to the normal trace of the diffusive flux of the solution to (40.3) over each mesh face. To this purpose, we are going to rely on the face-to-cell lifting operator introduced in §17.1.

40.3.1   Face localization of the normal diffusive flux

Let p, q be two real numbers satisfying the requirement (40.6). Since $z \mapsto \frac{zd}{z+d}$ is an increasing function, there is $\widetilde{p} \in (2, p]$ such that $q \geq \frac{\widetilde{p}d}{p+d}$. With the three numbers $p, q, \widetilde{p}$ in hand, we now invoke the existence of a face-to-cell lifting operator that has been established in Lemma 17.1. Let us recall this result for completeness.

Lemma 40.3 (Face-to-cell lifting). *For every mesh cell $K \in \mathcal{T}_h$ and every face $F \in \mathcal{F}_K$, there exists a lifting operator $L_F^K : W^{\frac{1}{p}, \widetilde{p}'}(F) \to W^{1, \widetilde{p}'}(K)$ s.t.*

$$\gamma_{\partial K}^{\mathrm{g}}(L_F^K(\phi))_{|\partial K \setminus F} = 0, \qquad \gamma_{\partial K}^{\mathrm{g}}(L_F^K(\phi))_{|F} = \phi, \tag{40.11}$$

for all $\phi \in W^{\frac{1}{p}, \widetilde{p}'}(F)$. Moreover, there is c s.t.

$$h_K^{\frac{d}{p}} |L_F^K(\phi)|_{W^{1,p'}(K)} + h_K^{-1+\frac{d}{q}} \|L_F^K(\phi)\|_{L^{q'}(K)} \leq c\, h_K^{-\frac{1}{p}+\frac{d}{p}} \|\phi\|_{W^{\frac{1}{p}, \widetilde{p}'}(F)}, \tag{40.12}$$

for all $\phi \in W^{\frac{1}{p}, \widetilde{p}'}(F)$ with $\|\phi\|_{W^{\frac{1}{p}, \widetilde{p}'}(F)} := \|\phi\|_{L^{\widetilde{p}'}(F)} + h_F^{\frac{1}{p}} |\phi|_{W^{\frac{1}{p}, \widetilde{p}'}(F)}$, all $K \in \mathcal{T}_h$, all $F \in \mathcal{F}_K$, and all $h \in \mathcal{H}$.

Let $K \in \mathcal{T}_h$ be a mesh cell and consider the functional space

$$\boldsymbol{S}^{\mathrm{d}}(K) := \{\boldsymbol{\tau} \in \boldsymbol{L}^p(K) \mid \nabla{\cdot}\boldsymbol{\tau} \in L^q(K)\}, \tag{40.13}$$

where the superscript $^{\mathrm{d}}$ refers to the divergence operator. We equip $\boldsymbol{S}^{\mathrm{d}}(K)$ with the following dimensionally consistent norm:

$$\|\boldsymbol{\tau}\|_{\boldsymbol{S}^{\mathrm{d}}(K)} := \|\boldsymbol{\tau}\|_{\boldsymbol{L}^p(K)} + h_K^{1+d(\frac{1}{p}-\frac{1}{q})} \|\nabla{\cdot}\boldsymbol{\tau}\|_{L^q(K)}. \tag{40.14}$$

With the lifting operator L_F^K in hand, the normal trace on any face F of K of any field $\boldsymbol{\tau} \in \boldsymbol{S}^{\mathrm{d}}(K)$, denoted by $(\boldsymbol{\tau}{\cdot}\boldsymbol{n}_K)_{|F}$, is defined to be the linear form in $(W^{\frac{1}{p}, \widetilde{p}'}(F))'$ whose action on any function $\phi \in W^{\frac{1}{p}, \widetilde{p}'}(F)$ is

$$\langle (\boldsymbol{\tau}{\cdot}\boldsymbol{n}_K)_{|F}, \phi \rangle_F := \int_K \left(\boldsymbol{\tau}{\cdot}\nabla L_F^K(\phi) + (\nabla{\cdot}\boldsymbol{\tau}) L_F^K(\phi) \right) \mathrm{d}x. \tag{40.15}$$

Here, $\langle \cdot, \cdot \rangle_F$ denotes the duality pairing between $(W^{\frac{1}{p}, \widetilde{p}'}(F))'$ and $W^{\frac{1}{p}, \widetilde{p}'}(F)$. Notice that the right-hand side of (40.15) is well defined owing to Hölder's inequality and (40.12). Owing to (40.11), we readily verify that we have indeed defined an extension of the normal trace since we have $\langle (\boldsymbol{\tau}{\cdot}\boldsymbol{n}_K)_{|F}, \phi \rangle_F = \int_F (\boldsymbol{\tau}{\cdot}\boldsymbol{n}_K)\phi \, \mathrm{d}s$ whenever the field $\boldsymbol{\tau}$ is smooth. Let us now derive an important bound on the linear form $(\boldsymbol{\tau}{\cdot}\boldsymbol{n}_K)_{|F}$ when it acts on a function from the space P_F which is composed of the restrictions to F of the functions in P_K. Notice that $P_F \subset W^{\frac{1}{p}, \widetilde{p}'}(F)$.

Lemma 40.4 (Bound on normal component). *There is c s.t.*

$$|\langle(\boldsymbol{\tau}\cdot\boldsymbol{n}_K)_{|F},\phi_h\rangle_F| \le c\, h_K^{d(\frac{1}{2}-\frac{1}{p})}\|\boldsymbol{\tau}\|_{\boldsymbol{S}^{\mathrm{d}}(K)}h_F^{-\frac{1}{2}}\|\phi_h\|_{L^2(F)}, \tag{40.16}$$

for all $\boldsymbol{\tau}\in\boldsymbol{S}^{\mathrm{d}}(K)$, all $\phi_h\in P_F$, all $K\in\mathcal{T}_h$, all $F\in\mathcal{F}_K$, and all $h\in\mathcal{H}$.

Proof. A direct consequence of (40.15), Hölder's inequality, and Lemma 40.3 is that the following holds true for all $\phi\in W^{\frac{1}{p},\tilde{p}'}(F)$:

$$|\langle(\boldsymbol{\tau}\cdot\boldsymbol{n}_K)_{|F},\phi\rangle_F| \le c\, h_K^{-\frac{1}{p}+d(\frac{1}{p}-\frac{1}{p})}\|\boldsymbol{\tau}\|_{\boldsymbol{S}^{\mathrm{d}}(K)}\|\phi\|_{W^{\frac{1}{p},\tilde{p}'}(F)}.$$

Since $\|\phi\|_{W^{\frac{1}{p},\tilde{p}'}(F)} := \|\phi\|_{L^{\tilde{p}'}(F)} + h_F^{\frac{1}{p}}|\phi|_{W^{\frac{1}{p},\tilde{p}'}(F)}$, the assertion (40.16) follows from the inverse inequality $\|\phi_h\|_{W^{\frac{1}{p},\tilde{p}'}(F)} \le c h_F^{(d-1)(\frac{1}{2}-\frac{1}{p})}\|\phi_h\|_{L^2(F)}$, which is valid for all $\phi_h\in P_F$, and the regularity of the mesh sequence. $\qquad\square$

40.3.2 Definition of $n_\sharp$ and key identities

Let us consider the functional space V_{s} defined in (40.5). For all $v\in V_{\mathrm{s}}$, Lemma 40.1 shows that $\boldsymbol{\sigma}(v)_{|K}\in\boldsymbol{S}^{\mathrm{d}}(K)$ for all $K\in\mathcal{T}_h$, and Lemma 40.4 implies that it is possible to give a meaning by duality to the normal component of $\boldsymbol{\sigma}(v)_{|K}$ on all the faces of K separately. Since we have set $\boldsymbol{\sigma}(v_h)_{|K}=-\lambda_K\nabla(v_{h|K})$ for all $v_h\in P_k^{\mathrm{b}}(\mathcal{T}_h)$, and since we have $P_K\subset W^{k+1,\infty}(K)$ with $k\ge 1$, we infer that $\boldsymbol{\sigma}(v_h)_{|K}\in\boldsymbol{S}^{\mathrm{d}}(K)$ as well. Thus, $\boldsymbol{\sigma}(v)_{|K}\in\boldsymbol{S}^{\mathrm{d}}(K)$ for all $v\in V_\sharp^{\mathrm{b}}:=V_{\mathrm{s}}+P_k^{\mathrm{b}}(\mathcal{T}_h)$. Let us now introduce the bilinear form $n_\sharp:V_\sharp^{\mathrm{b}}\times P_k^{\mathrm{b}}(\mathcal{T}_h)\to\mathbb{R}$ defined as follows:

$$n_\sharp(v,w_h) := \sum_{F\in\mathcal{F}_h}\sum_{K\in\mathcal{T}_F}\epsilon_{K,F}\theta_{K,F}\langle(\boldsymbol{\sigma}(v)_{|K}\cdot\boldsymbol{n}_K)_{|F},[\![w_h]\!]\rangle_F, \tag{40.17}$$

where the (yet to be defined) weights $\{\theta_{K,F}\}_{F\in\mathcal{F}_h,K\in\mathcal{T}_F}$ are assumed to satisfy (40.8). The definition (40.17) is meaningful since $[\![w_h]\!]_F\in P_F$ for all $w_h\in P_k^{\mathrm{b}}(\mathcal{T}_h)$. The factor $\epsilon_{K,F}$ in (40.17) handles the relative orientation of $\boldsymbol{n}_K$ and $\boldsymbol{n}_F$, whereas the weights $\theta_{K,F}$ will help achieve robustness w.r.t. the diffusivity contrast. We will see in the next section how these weights must depend on the diffusion coefficient.

The following lemma is fundamental to understand the role that the bilinear form $n_\sharp$ will play in the next section in the analysis of various nonconforming approximation methods. Recall the definition (40.9) of the weighted average $\{\cdot\}_\theta$.

Lemma 40.5 (Identities for $n_\sharp$). *For every choice of weights $\{\theta_{K,F}\}_{F\in\mathcal{F}_h,K\in\mathcal{T}_F}$, we have*

$$n_\sharp(v_h,w_h) = \sum_{F\in\mathcal{F}_h}\int_F\{\boldsymbol{\sigma}(v_h)\}_\theta\cdot\boldsymbol{n}_F[\![w_h]\!]\,\mathrm{d}s, \tag{40.18a}$$

$$n_\sharp(v,w_h) = \sum_{K\in\mathcal{T}_h}\int_K\left(\boldsymbol{\sigma}(v)\cdot\nabla w_{h|K}+(\nabla\cdot\boldsymbol{\sigma}(v))w_{h|K}\right)\mathrm{d}x, \tag{40.18b}$$

for all $v_h,w_h\in P_k^{\mathrm{b}}(\mathcal{T}_h)$ and all $v\in V_{\mathrm{s}}$.

Proof. (1) Proof of (40.18a). Let $v_h,w_h\in P_k^{\mathrm{b}}(\mathcal{T}_h)$. Since the restriction of $\boldsymbol{\sigma}(v_h)$ to each mesh cell is smooth, and since the restriction of $L_F^K([\![w_h]\!])$ to ∂K is nonzero only on the face $F\in\mathcal{F}_K$ where

it coincides with $[\![w_h]\!]$, we have

$$\langle(\boldsymbol{\sigma}(v_h)_{|K}\cdot\boldsymbol{n}_K)_{|F}, [\![w_h]\!]\rangle_F = \int_K \Big(\boldsymbol{\sigma}(v_h)_{|K}\cdot\nabla L_F^K([\![w_h]\!]) + (\nabla\cdot\boldsymbol{\sigma}(v_h)_{|K})L_F^K([\![w_h]\!])\Big)\,\mathrm{d}x$$

$$= \int_{\partial K} \boldsymbol{\sigma}(v_h)_{|K}\cdot\boldsymbol{n}_K L_F^K([\![w_h]\!])\,\mathrm{d}s = \int_F \boldsymbol{\sigma}(v_h)_{|K}\cdot\boldsymbol{n}_K[\![w_h]\!]\,\mathrm{d}s,$$

where we used the divergence formula in K. After using the definitions of $\epsilon_{K,F}$ and of $\theta_{K,F}$, we obtain

$$n_\sharp(v_h, w_h) = \sum_{F\in\mathcal{F}_h}\sum_{K\in\mathcal{T}_F} \epsilon_{K,F}\theta_{K,F}\int_F \boldsymbol{\sigma}(v_h)_{|K}\cdot\boldsymbol{n}_K[\![w_h]\!]\,\mathrm{d}s$$

$$= \sum_{F\in\mathcal{F}_h}\int_F \{\boldsymbol{\sigma}(v_h)\}_\theta\cdot\boldsymbol{n}_F[\![w_h]\!]\,\mathrm{d}s.$$

(2) Proof of (40.18b). Let $v \in V_\mathrm{s}$ and $w_h \in P_k^\mathrm{b}(\mathcal{T}_h)$. Let $\mathcal{K}_\delta^\mathrm{d} : \boldsymbol{L}^1(D) \to \boldsymbol{C}^\infty(\overline{D})$ and $\mathcal{K}_\delta^\mathrm{b} : L^1(D) \to C^\infty(\overline{D})$ be the mollification operators introduced in §23.1. Recall the following key commuting property:

$$\nabla\cdot(\mathcal{K}_\delta^\mathrm{d}(\boldsymbol{\tau})) = \mathcal{K}_\delta^\mathrm{b}(\nabla\cdot\boldsymbol{\tau}), \qquad (40.19)$$

for all $\boldsymbol{\tau} \in \boldsymbol{L}^1(D)$ s.t. $\nabla\cdot\boldsymbol{\tau} \in L^1(D)$. It is important to realize that this property can be applied to $\boldsymbol{\sigma}(v)$ for all $v \in V_\mathrm{s}$ since $\nabla\cdot\boldsymbol{\sigma}(v) \in L^1(D)$ by definition of V_s. (Note that this property cannot be applied to $\boldsymbol{\sigma}(v_h)$ with $v_h \in P_k^\mathrm{b}(\mathcal{T}_h)$, since the normal component of $\boldsymbol{\sigma}(v_h)$ is in general discontinuous across the mesh interfaces, i.e., $\boldsymbol{\sigma}(v_h)$ does not have a weak divergence; see Theorem 18.10.) Let us consider the mollified bilinear form

$$n_{\sharp\delta}(v, w_h) := \sum_{F\in\mathcal{F}_h}\sum_{K\in\mathcal{T}_F} \epsilon_{K,F}\theta_{K,F}\langle(\mathcal{K}_\delta^\mathrm{d}(\boldsymbol{\sigma}(v))_{|K}\cdot\boldsymbol{n}_K)_{|F}, [\![w_h]\!]\rangle_F.$$

Owing to the commuting property (40.19), we infer that

$$\langle(\mathcal{K}_\delta^\mathrm{d}(\boldsymbol{\sigma}(v))_{|K}\cdot\boldsymbol{n}_K)_{|F}, [\![w_h]\!]\rangle_F = \int_K \Big(\mathcal{K}_\delta^\mathrm{d}(\boldsymbol{\sigma}(v))\cdot L_F^K([\![w_h]\!]) + \mathcal{K}_\delta^\mathrm{b}(\nabla\cdot\boldsymbol{\sigma}(v))L_F^K([\![w_h]\!])\Big)\,\mathrm{d}x.$$

Theorem 23.4 implies that

$$\lim_{\delta\to 0}\int_K \Big(\mathcal{K}_\delta^\mathrm{d}(\boldsymbol{\sigma}(v))\cdot L_F^K([\![w_h]\!]) + \mathcal{K}_\delta^\mathrm{b}((\nabla\cdot\boldsymbol{\sigma}(v)))L_F^K([\![w_h]\!])\Big)\,\mathrm{d}x =$$

$$\int_K \Big(\boldsymbol{\sigma}(v)\cdot L_F^K([\![w_h]\!]) + (\nabla\cdot\boldsymbol{\sigma}(v))L_F^K([\![w_h]\!])\Big)\,\mathrm{d}x = \langle(\boldsymbol{\sigma}(v)_{|K}\cdot\boldsymbol{n}_K)_{|F}, [\![w_h]\!]\rangle_F.$$

Summing over the mesh faces and the associated mesh cells, we infer that

$$\lim_{\delta\to 0} n_{\sharp\delta}(v, w_h) = n_\sharp(v, w_h).$$

Since the mollified function $\mathcal{K}_\delta^\mathrm{d}(\boldsymbol{\sigma}(v))$ is smooth, by repeating the calculation done in Step (1), we also have

$$n_{\sharp\delta}(v, w_h) = \sum_{F\in\mathcal{F}_h}\int_F \{\mathcal{K}_\delta^\mathrm{d}(\boldsymbol{\sigma}(v))\}_\theta\cdot\boldsymbol{n}_F[\![w_h]\!]\,\mathrm{d}s.$$

Using the identity (40.10), $[\![\mathcal{K}_\delta^{\mathrm{d}}(\boldsymbol{\sigma}(v))]\!]\cdot\boldsymbol{n}_F = 0$ for all $F \in \mathcal{F}_h^\circ$, the divergence formula in K, and the commuting property (40.19), we obtain

$$
\begin{aligned}
n_{\sharp\delta}(v, w_h) &= \sum_{F \in \mathcal{F}_h} \int_F \{\mathcal{K}_\delta^{\mathrm{d}}(\boldsymbol{\sigma}(v))\}_\theta \cdot \boldsymbol{n}_F [\![w_h]\!]\, \mathrm{d}s + \sum_{F \in \mathcal{F}_h^\circ} \int_F [\![\mathcal{K}_\delta^{\mathrm{d}}(\boldsymbol{\sigma}(v))]\!]\cdot\boldsymbol{n}_F \{w_h\}_{\bar\theta}\, \mathrm{d}s \\
&= \sum_{F \in \mathcal{F}_h} \int_F [\![w_h \mathcal{K}_\delta^{\mathrm{d}}(\boldsymbol{\sigma}(v))]\!]\cdot\boldsymbol{n}_F\, \mathrm{d}s = \sum_{K \in \mathcal{T}_h} \int_{\partial K} \mathcal{K}_\delta^{\mathrm{d}}(\boldsymbol{\sigma}(v))\cdot\boldsymbol{n}_K w_{h|K}\, \mathrm{d}s \\
&= \sum_{K \in \mathcal{T}_h} \int_K \left(\mathcal{K}_\delta^{\mathrm{d}}(\boldsymbol{\sigma}(v))\cdot\nabla w_{h|K} + \mathcal{K}_\delta^{\mathrm{b}}(\nabla\cdot\boldsymbol{\sigma}(v))w_{h|K} \right) \mathrm{d}x.
\end{aligned}
$$

Invoking again Theorem 23.4 leads to the assertion since

$$
\lim_{\delta \to 0} n_{\sharp\delta}(v, w_h) = \sum_{K \in \mathcal{T}_h} \int_K \left(\boldsymbol{\sigma}(v)\cdot\nabla w_{h|K} + (\nabla\cdot\boldsymbol{\sigma}(v))w_{h|K} \right) \mathrm{d}x. \qquad \square
$$

Remark 40.6 (Identity (40.18b)). We are going to use the identity (40.18b) to assert that $\boldsymbol{\sigma}(v)\cdot\boldsymbol{n}$ is continuous across the mesh interfaces without assuming that v is smooth, say $v \in H^{1+r}(D)$ with $r > \frac{1}{2}$. $\qquad \square$

We now establish an important boundedness estimate on the bilinear form $n_\sharp$. Since $\boldsymbol{\sigma}(v)_{|K} \in \boldsymbol{S}^{\mathrm{d}}(K)$ for all $K \in \mathcal{T}_h$ and all $v \in V_\sharp^{\mathrm{b}}$, we can equip the space $V_\sharp^{\mathrm{b}}$ with the seminorm

$$
|v|_{n_\sharp}^2 := \sum_{K \in \mathcal{T}_h} \lambda_K^{-1} \left(h_K^{2d(\frac{1}{2} - \frac{1}{p})} \|\boldsymbol{\sigma}(v)_{|K}\|_{\boldsymbol{L}^p(K)}^2 + h_K^{2d(\frac{2+d}{2d} - \frac{1}{q})} \|\nabla\cdot\boldsymbol{\sigma}(v)_{|K}\|_{L^q(K)}^2 \right). \tag{40.20}
$$

We notice that this seminorm is dimensionally consistent with the classical energy-norm defined as $\sum_{K \in \mathcal{T}_h} \lambda_K \|\nabla v_{|K}\|_{\boldsymbol{L}^2(K)}^2$. The reader is invited to verify that $|v|_\sharp \leq c\lambda_\flat^{-\frac{1}{2}}(\ell_D^{d(\frac{1}{2} - \frac{1}{p})}\|\boldsymbol{\sigma}(v)\|_{\boldsymbol{L}^p(D)} + \ell_D^{d(\frac{2+d}{2d} - \frac{1}{q})}\|\nabla\cdot\boldsymbol{\sigma}(v)\|_{L^q(D)})$, for all $v \in V_{\mathrm{s}}$; see Exercise 40.2.

In order to get robust error estimates with respect to λ, we need to avoid any dependency on the ratio of the values taken by λ in two adjacent subdomains since otherwise the error estimates become meaningless when the diffusion coefficient λ is highly contrasted. To avoid such dependencies, we introduce the following diffusion-dependent weights for all $F := \partial K_l \cap \partial K_r \in \mathcal{F}_h^\circ$:

$$
\theta_{K_l,F} := \frac{\lambda_{K_r}}{\lambda_{K_l} + \lambda_{K_r}}, \qquad \theta_{K_r,F} := \frac{\lambda_{K_l}}{\lambda_{K_l} + \lambda_{K_r}}. \tag{40.21}
$$

We also define

$$
\lambda_F := \frac{2\lambda_{K_l}\lambda_{K_r}}{\lambda_{K_l} + \lambda_{K_r}} \text{ if } F \in \mathcal{F}_h^\circ \quad \text{and} \quad \lambda_F := \lambda_{K_l} \text{ if } F \in \mathcal{F}_h^\partial. \tag{40.22}
$$

The two key properties we are going to use are that, for all $F \in \mathcal{F}_h$ and all $K \in \mathcal{T}_F$, $|\mathcal{T}_F|\lambda_K\theta_{K,F} = \lambda_F$ and $\lambda_F \leq |\mathcal{T}_F| \min_{K \in \mathcal{T}_F} \lambda_K$ (recall that $|\mathcal{T}_F|$ is the cardinality of the set $\mathcal{T}_F$).

Lemma 40.7 (Boundedness of $n_\sharp$). *With the weights defined in (40.21) and λ_F defined in (40.22) for all $F \in \mathcal{F}_h$, there is c, uniform w.r.t. λ, s.t. for all $v \in V_\sharp^{\mathrm{b}}$, all $w_h \in P_k^{\mathrm{b}}(\mathcal{T}_h)$, and all $h \in \mathcal{H}$,*

$$
|n_\sharp(v, w_h)| \leq c\,|v|_{n_\sharp} \left(\sum_{F \in \mathcal{F}_h} \lambda_F h_F^{-1} \|[\![w_h]\!]\|_{L^2(F)}^2 \right)^{\frac{1}{2}}. \tag{40.23}
$$

Proof. Let $v \in V_s + P_k^b(\mathcal{T}_h)$ and $w_h \in P_k^b(\mathcal{T}_h)$. Owing to the definition (40.17) of $n_\sharp$ and the estimate (40.16) from Lemma 40.4, we infer that

$$
|n_\sharp(v, w_h)| \le c \sum_{F \in \mathcal{F}_h} \sum_{K \in \mathcal{T}_F} \theta_{K,F} h_K^{d(\frac{1}{2} - \frac{1}{p})} \|\boldsymbol{\sigma}(v)_{|K}\|_{\boldsymbol{S}^d(K)} h_F^{-\frac{1}{2}} \|[\![w_h]\!]\|_{L^2(F)}
$$

$$
\le c \Bigg(\sum_{F \in \mathcal{F}_h} \sum_{K \in \mathcal{T}_F} \lambda_K^{-\frac{1}{2}} h_K^{d(\frac{1}{2} - \frac{1}{p})} \|\boldsymbol{\sigma}(v)_{|K}\|_{\boldsymbol{L}^p(K)} |\mathcal{T}_F|^{-\frac{1}{2}} \lambda_F^{\frac{1}{2}} h_F^{-\frac{1}{2}} \|[\![w_h]\!]\|_{L^2(F)}
$$

$$
+ \sum_{F \in \mathcal{F}_h} \sum_{K \in \mathcal{T}_F} \lambda_K^{-\frac{1}{2}} h_K^{d(\frac{2+d}{2d} - \frac{1}{q})} \|\nabla{\cdot}\boldsymbol{\sigma}(v)_{|K}\|_{L^q(K)} |\mathcal{T}_F|^{-\frac{1}{2}} \lambda_F^{\frac{1}{2}} h_F^{-\frac{1}{2}} \|[\![w_h]\!]\|_{L^2(F)} \Bigg),
$$

where we used that $\theta_{K,F} \le \theta_{K,F}^{\frac{1}{2}}$ (since $\theta_{K,F} \le 1$), $|\mathcal{T}_F| \lambda_K \theta_{K,F} = \lambda_F$, the definition of $\|{\cdot}\|_{\boldsymbol{S}^d(K)}$, and $1 + d(\frac{1}{2} - \frac{1}{q}) = d(\frac{2+d}{2d} - \frac{1}{q})$. Owing to the Cauchy–Schwarz inequality, we infer that

$$
\sum_{F \in \mathcal{F}_h} \sum_{K \in \mathcal{T}_F} a_K |\mathcal{T}_F|^{-\frac{1}{2}} b_F \le \Big(\sum_{K \in \mathcal{T}_h} |\mathcal{F}_K| a_K^2 \Big)^{\frac{1}{2}} \Big(\sum_{F \in \mathcal{F}_h} b_F^2 \Big)^{\frac{1}{2}},
$$

for all real numbers $\{a_K\}_{K \in \mathcal{T}_h}$, $\{b_F\}_{F \in \mathcal{F}_h}$, where we used that

$$
\sum_{F \in \mathcal{F}_h} \sum_{K \in \mathcal{T}_F} (\cdot) = \sum_{K \in \mathcal{T}_h} \sum_{F \in \mathcal{F}_K} (\cdot)
$$

for the term involving the a_K's. Since $|\mathcal{F}_K|$ is uniformly bounded ($|\mathcal{F}_K| = d + 1$ for simplicial meshes), applying this bound to the two terms composing the estimate on $|n_\sharp(v, w_h)|$ leads to the expected boundedness property. $\square$

Remark 40.8 (Literature). Diffusion-dependent averages have been introduced in Dryja [173] for discontinuous Galerkin methods and have been analyzed in various contexts in Burman and Zunino [101], Dryja et al. [174], Di Pietro et al. [167], Ern et al. [194]. $\square$

Exercises

Exercise 40.1 (Normal flux). Let $\boldsymbol{\sigma} \in \{\boldsymbol{\tau} \in \boldsymbol{L}^p(K) \mid \nabla{\cdot}\boldsymbol{\tau} \in L^2(K)\}$, $p > 2$. Let $\gamma_{\partial K}^d(\boldsymbol{\sigma}) \in H^{-\frac{1}{2}}(\partial K)$ be s.t. $\langle \gamma_{\partial K}^d(\boldsymbol{\sigma}), \phi \rangle_{\partial K} := \int_K \boldsymbol{\sigma}{\cdot}\nabla v(\phi)\, \mathrm{d}x + \int_K (\nabla{\cdot}\boldsymbol{\sigma}) v(\phi)\, \mathrm{d}x$ for all $\phi \in H^{\frac{1}{2}}(\partial K)$, where $v(\phi) \in H^1(K)$ is a lifting of ϕ, i.e., $\gamma_{\partial K}^g(v(\phi)) = \phi$ (see (4.12)). Prove that $\langle \gamma_{\partial K}^d(\boldsymbol{\sigma}), \phi \rangle_{\partial K} = \sum_{F \in \mathcal{F}_K} \langle (\boldsymbol{\sigma}{\cdot}\boldsymbol{n}_K)_{|F}, \phi_{|F} \rangle_F$. (*Hint*: reason as in the proof of (40.18b).)

Exercise 40.2 (Bound on $|v|_\sharp$). Prove that for all $v \in V_s$, $|v|_{n_\sharp} \le c\lambda_\flat^{-\frac{1}{2}} (\ell_D^{d(\frac{1}{2} - \frac{1}{p})} \|\boldsymbol{\sigma}(v)\|_{\boldsymbol{L}^p(D)} + \ell_D^{d(\frac{2+d}{2d} - \frac{1}{q})} \|\nabla{\cdot}\boldsymbol{\sigma}(v)\|_{L^q(D)})$. (*Hint*: for the sum with $\boldsymbol{L}^p$-norms, use Hölder's inequality after observing that $h_K^d \le c|K|$, and for the sum with L^q norms, use that $(\sum_{K \in \mathcal{T}_h} a_K^t)^{\frac{1}{t}} \le (\sum_{K \in \mathcal{T}_h} a_K^s)^{\frac{1}{s}}$ for real numbers $t \ge s$.)

Exercise 40.3 (Jump identity). Let $F := \partial K_l \cap \partial K_r \in \mathcal{F}_h^\circ$. Let $\theta_l, \theta_r \in [0, 1]$ be s.t. $\theta_l + \theta_r = 1$. Set $\{a\}_\theta := \theta_l a_l + \theta_r a_r$ and $\{a\}_{\bar\theta} := \theta_r a_l + \theta_l a_r$. (i) Show that $[\![ab]\!] = \{a\}_{\bar\theta} [\![b]\!] + [\![a]\!] \{b\}_\theta$. (ii) Show that $[\![ab]\!] = \{a\}_\theta [\![b]\!] + [\![a]\!] \{b\}_{\bar\theta}$.

Chapter 41

Contrasted diffusivity (II)

In this chapter, we continue the study of the model elliptic problem (40.3) with contrasted diffusivity. Now that we have in hand our key tool, that is, the bilinear form $n_\sharp$ introduced in §40.3, we perform the error analysis when the model problem (40.3) is approximated by one of the nonconforming methods introduced previously, i.e., Crouzeix–Raviart finite elements, Nitsche's boundary penalty method, the discontinuous Galerkin (dG) method, and the hybrid high-order (HHO) method.

41.1 Continuous and discrete settings

Recall that the model problem (40.3) consists of seeking $u \in V := H_0^1(D)$ s.t. $a(u, w) = \ell(w)$ for all $w \in V$, with $a(v, w) := \int_D \lambda \nabla v \cdot \nabla w \, dx$ and $\ell(w) := \int_D f w \, dx$. We assume that the solution to (40.3) is in the functional space V_s defined in (40.5) with the real numbers p, q satisfying (40.6), i.e.,

$$u \in V_s := \{v \in H_0^1(D) \mid \boldsymbol{\sigma}(v) \in \boldsymbol{L}^p(D), \ \nabla \cdot \boldsymbol{\sigma}(v) \in L^q(D)\}, \tag{41.1}$$

where $\boldsymbol{\sigma}(v) := -\lambda \nabla v$ and the real numbers p, q are such that

$$2 < p, \qquad \frac{2d}{2+d} < q \le 2. \tag{41.2}$$

Let $(\mathcal{T}_h)_{h \in \mathcal{H}}$ be a shape-regular sequence of simplicial affine meshes so that each mesh covers D exactly. Let $k \ge 1$ and consider the broken polynomial space $P_k^b(\mathcal{T}_h)$ defined in (40.7). The discrete problem takes the generic form

$$\begin{cases} \text{Find } u_h \in V_h \text{ such that} \\ a_h(u_h, w_h) = \ell_h(w_h), \quad \forall w_h \in V_h, \end{cases} \tag{41.3}$$

where the subspace $V_h \subset P_k^b(\mathcal{T}_h)$ and the forms a_h and ℓ_h depend on the approximation method.

For all the approximation methods, the error analysis relies on Lemma 27.5, and the main issue is to prove consistency/boundedness. Recall from Definition 27.3 that the consistency error is defined by setting

$$\langle \delta_h(v_h), w_h \rangle_{V_h', V_h} := \ell_h(w_h) - a_h(v_h, w_h), \qquad \forall v_h, w_h \in V_h. \tag{41.4}$$

The key tool to prove consistency/boundedness is the bilinear form $n_\sharp : V_\sharp^{\mathrm{b}} \times P_k^{\mathrm{b}}(\mathcal{T}_h) \to \mathbb{R}$ with $V_\sharp^{\mathrm{b}} := V_{\mathrm{s}} + P_k^{\mathrm{b}}(\mathcal{T}_h)$ s.t.

$$n_\sharp(v, w_h) := \sum_{F \in \mathcal{F}_h} \sum_{K \in \mathcal{T}_F} \epsilon_{K,F} \theta_{K,F} \langle (\boldsymbol{\sigma}(v)_{|K} \cdot \boldsymbol{n}_K)_{|F}, [\![w_h]\!] \rangle_F, \tag{41.5}$$

with the orientation factor $\epsilon_{K,F} := \boldsymbol{n}_K \cdot \boldsymbol{n}_F = \pm 1$ and the diffusion-dependent weights s.t. for all $F := \partial K_l \cap \partial K_r \in \mathcal{F}_h^\circ$,

$$\theta_{K_l,F} := \frac{\lambda_{K_r}}{\lambda_{K_l} + \lambda_{K_r}}, \qquad \theta_{K_r,F} := \frac{\lambda_{K_l}}{\lambda_{K_l} + \lambda_{K_r}}, \tag{41.6}$$

with the convention $\theta_{K_l,F} := 1$ for all $F := \partial K_l \cap \partial D \in \mathcal{F}_h^\partial$. Notice that $\theta_{K_l,F}, \theta_{K_r,F} \in [0,1]$ and $\theta_{K_l,F} + \theta_{K_r,F} = 1$. For all $v \in W^{1,1}(\mathcal{T}_h)$, we define weighted averages a.e. on every face $F \in \mathcal{F}_h$ as follows: If $F \in \mathcal{F}_h^\circ$,

$$\{v\}_{F,\theta} := \theta_{K_l,F} v_{|K_l} + \theta_{K_r,F} v_{|K_r}, \tag{41.7a}$$

$$\{v\}_{F,\bar\theta} := \theta_{K_r,F} v_{|K_l} + \theta_{K_l,F} v_{|K_r}, \tag{41.7b}$$

and $\{v\}_\theta := v_{|K_l}$ if $F := \partial K_l \cap \partial D \in \mathcal{F}_h^\partial$. We omit the subscript F in the jump and the average whenever the context is unambiguous.

The key properties of the bilinear form $n_\sharp$ we are going to invoke are the following: For all $v_h, w_h \in P_k^{\mathrm{b}}(\mathcal{T}_h)$ and all $v \in V_{\mathrm{s}}$, we have (see Lemma 40.5)

$$n_\sharp(v_h, w_h) = \sum_{F \in \mathcal{F}_h} \int_F \{\boldsymbol{\sigma}(v_h)\}_\theta \cdot \boldsymbol{n}_F [\![w_h]\!] \, \mathrm{d}s, \tag{41.8a}$$

$$n_\sharp(v, w_h) = \sum_{K \in \mathcal{T}_h} \int_K \left(\boldsymbol{\sigma}(v) \cdot \nabla w_{h|K} + (\nabla \cdot \boldsymbol{\sigma}(v)) w_{h|K} \right) \mathrm{d}x, \tag{41.8b}$$

and (see Lemma 40.7) there is c, uniform w.r.t. λ, s.t. for all $v \in V_\sharp^{\mathrm{b}}$, all $w_h \in P_k^{\mathrm{b}}(\mathcal{T}_h)$, and all $h \in \mathcal{H}$,

$$|n_\sharp(v, w_h)| \le c \, |v|_{n_\sharp} \left(\sum_{F \in \mathcal{F}_h} \lambda_F \frac{1}{h_F} \|[\![w_h]\!]\|_{L^2(F)}^2 \right)^{\frac{1}{2}}, \tag{41.9}$$

where $|\cdot|_{n_\sharp}$ is defined in (40.20) and

$$\lambda_F := \frac{2\lambda_{K_l} \lambda_{K_r}}{\lambda_{K_l} + \lambda_{K_r}} \text{ if } F \in \mathcal{F}_h^\circ \quad \text{and} \quad \lambda_F := \lambda_{K_l} \text{ if } F \in \mathcal{F}_h^\partial. \tag{41.10}$$

We consider the dimensionally consistent seminorm $|v|_{\lambda,p,q}^2 := \|\lambda^{\frac{1}{2}} \nabla_h v\|_{\boldsymbol{L}^2(D)}^2 + |v|_{n_\sharp}^2$ for all $v \in V_\sharp^{\mathrm{b}}$. Since λ is piecewise constant, we have

$$|v|_{\lambda,p,q}^2 := \sum_{K \in \mathcal{T}_h} \lambda_K \left(\|\nabla(v_{|K})\|_{\boldsymbol{L}^2(K)}^2 + h_K^{2d(\frac{1}{2}-\frac{1}{p})} \|\nabla(v_{|K})\|_{\boldsymbol{L}^p(K)}^2 + h_K^{2d(\frac{d+2}{2d}-\frac{1}{q})} \|\Delta(v_{|K})\|_{L^q(K)}^2 \right). \tag{41.11}$$

41.2   Crouzeix–Raviart approximation

We consider in this section the Crouzeix–Raviart finite element space introduced in Chapter 36, that is,

$$P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h) := \left\{ v_h \in P_1^{\mathrm{b}}(\mathcal{T}_h) \mid \int_F [\![v_h]\!]_F \, \mathrm{d}s = 0, \, \forall F \in \mathcal{F}_h \right\}. \tag{41.12}$$

The discrete problem takes the form (41.3) with $V_h := P_{1,0}^{\text{CR}}(\mathcal{T}_h)$ and

$$a_h(v_h, w_h) := \int_D \lambda \nabla_h v_h \cdot \nabla_h w_h \, dx, \qquad \ell_h(w_h) := \int_D f w_h \, dx. \qquad (41.13)$$

We equip V_h with the norm $\|v_h\|_{V_h} := \|\lambda^{\frac{1}{2}} \nabla_h v_h\|_{\boldsymbol{L}^2(D)}$. Adapting Lemma 36.4 to the present setting leads to the following result.

Lemma 41.1 (Coercivity, well-posedness). (i) *The bilinear form a_h is coercive on V_h with constant $\alpha := 1$.* (ii) *The discrete problem (41.3) is well-posed.*

Let $V_\sharp := V_{\text{s}} + V_h$ be equipped with the norm $\|v\|_{V_\sharp} := |v|_{\lambda,p,q}$ with $|v|_{\lambda,p,q}$ defined in (41.11) (this is indeed a norm on $V_\sharp$ since $|v|_{\lambda,p,q} = 0$ implies that v is piecewise constant and hence vanishes identically owing to the definition of V_h). Invoking inverse inequalities shows that there is $c_\sharp$ s.t. $\|v_h\|_{V_\sharp} \leq c_\sharp \|v_h\|_{V_h}$ for all $v_h \in V_h$ and all $h \in \mathcal{H}$, i.e., (27.5) holds true.

Lemma 41.2 (Consistency/boundedness). *There is $\omega_\sharp$, uniform w.r.t. $u \in V_{\text{s}}$ and λ, s.t. $\|\delta_h(v_h)\|_{V_h'} \leq \omega_\sharp \|u - v_h\|_{V_\sharp}$ for all $v_h \in V_h$ and all $h \in \mathcal{H}$.*

Proof. Let $v_h, w_h \in V_h$. Since $V_h \subset P_k^{\text{b}}(\mathcal{T}_h)$, the identity (41.8a) implies that

$$n_\sharp(v_h, w_h) = \sum_{F \in \mathcal{F}_h} \int_F \{\boldsymbol{\sigma}(v_h)\}_\theta \cdot \boldsymbol{n}_F [\![w_h]\!] \, ds = 0,$$

because $\{\boldsymbol{\sigma}(v_h)\}_\theta \cdot \boldsymbol{n}_F$ is constant on F. Invoking the identity (41.8b) with $v := u$ and since $f = \nabla \cdot \boldsymbol{\sigma}(u)$, we also have

$$\ell_h(w_h) = -\int_D \boldsymbol{\sigma}(u) \cdot \nabla_h w_h \, dx + n_\sharp(u, w_h).$$

Combining the above two identities and letting $\eta := u - v_h$, we obtain

$$\langle \delta_h(v_h), w_h \rangle_{V_h', V_h} = \int_D \lambda \nabla_h \eta \cdot \nabla_h w_h \, dx + n_\sharp(u, w_h)$$
$$= \int_D \lambda \nabla_h \eta \cdot \nabla_h w_h \, dx + n_\sharp(\eta, w_h).$$

The assertion follows by invoking the Cauchy–Schwarz inequality, the boundedness of $n_\sharp$ (see (41.9)), and the bound $\sum_{F \in \mathcal{F}_h} \lambda_F h_F^{-1} \|[\![w_h]\!]\|_{L^2(F)}^2 \leq c \|w_h\|_{V_h}^2$ which follows by adapting the proof of Lemma 36.9 with $v := 0$ and using that $\lambda_F \leq |\mathcal{T}_F| \min_{K \in \mathcal{T}_F} \lambda_K$. $\qquad \square$

Theorem 41.3 (Error estimate). *Let u solve (40.3) and u_h solve (41.3). Assume that $u \in H^{1+r}(D)$, $r > 0$. (i) There is c, uniform w.r.t. λ, s.t. for all $h \in \mathcal{H}$,*

$$\|u - u_h\|_{V_\sharp} \leq c \inf_{v_h \in V_h} \|u - v_h\|_{V_\sharp}. \qquad (41.14)$$

(ii) *Letting $t := \min(1, r)$, we have*

$$\|u - u_h\|_{V_\sharp} \leq c \left(\sum_{K \in \mathcal{T}_h} \lambda_K h_K^{2t} |u|_{H^{1+t}(K)}^2 + \lambda_K^{-1} h_K^{2d(\frac{d+2}{2d} - \frac{1}{q})} \|f\|_{L^q(K)}^2 \right)^{\frac{1}{2}}. \qquad (41.15)$$

Proof. (i) The estimate (41.14) follows from Lemma 27.5 combined with stability (Lemma 41.1) and consistency/boundedness (Lemma 41.2).

(ii) We bound the infimum in (41.14) by considering $\eta := u - \mathcal{I}_h^{\mathrm{CR}}(u)$. For all $K \in \mathcal{T}_h$, Lemma 36.1 implies that $\|\nabla(\eta_{|K})\|_{\boldsymbol{L}^2(K)} \leq ch_K^t|u|_{H^{1+t}(K)}$. Moreover, invoking the embedding $\boldsymbol{H}^t(\widehat{K}) \hookrightarrow \boldsymbol{L}^p(\widehat{K})$ we obtain the bound (see (17.19))

$$h_K^{d(\frac{1}{2}-\frac{1}{p})}\|\nabla(\eta_{|K})\|_{\boldsymbol{L}^p(K)} \leq c\left(\|\nabla(\eta_{|K})\|_{\boldsymbol{L}^2(K)} + h_K^t|\nabla(\eta_{|K})|_{\boldsymbol{H}^t(K)}\right). \tag{41.16}$$

Observing that $|\nabla(\eta_{|K})|_{H^t(K)} = |u|_{H^{1+t}(K)}$ since $\mathcal{I}_h^{\mathrm{CR}}(u)$ is affine on K and using again Lemma 36.1 gives $h_K^{d(\frac{1}{2}-\frac{1}{p})}\|\nabla(\eta_{|K})\|_{\boldsymbol{L}^p(K)} \leq ch_K^t|u|_{H^{1+t}(K)}$. Finally, we have $\Delta(\eta_{|K}) = -\lambda_K^{-1}f$ in K. $\qquad\square$

Remark 41.4 (Convergence). Note that the rightmost term in the estimate (41.15) converges as $\mathcal{O}(h)$ when $q = 2$. Note also that convergence is lost when $q \leq \frac{2d}{d+2}$, which is somewhat natural since in this case the linear form $w \mapsto \int_D fw \,\mathrm{d}x$ is no longer bounded on $H^1(D)$. $\qquad\square$

Remark 41.5 (Weights). Although the weights introduced in (40.21) are not used in the Crouzeix–Raviart discretization, they play a role in the error analysis. More precisely, we used the boundedness of the bilinear form $n_\sharp$ together with $\lambda_F \leq |\mathcal{T}_F|\min_{K\in\mathcal{T}_F}\lambda_K$ in the proof of Lemma 41.2. $\qquad\square$

41.3 Nitsche's boundary penalty method

We consider in this section the boundary penalty method introduced in Chapter 37. Recall that $V_h := P_k^{\mathrm{g}}(\mathcal{T}_h)$, $k \geq 1$, i.e., V_h is H^1-conforming with

$$P_k^{\mathrm{g}}(\mathcal{T}_h) := \{v_h \in P_k^{\mathrm{b}}(\mathcal{T}_h) \mid [\![v_h]\!]_F = 0, \ \forall F \in \mathcal{F}_h^\circ\}. \tag{41.17}$$

The discrete problem is (41.3) with $V_h := P_k^{\mathrm{g}}(\mathcal{T}_h)$,

$$a_h(v_h, w_h) := a(v_h, w_h) + \sum_{F \in \mathcal{F}_h^\partial} \int_F \left(\boldsymbol{\sigma}(v_h)\cdot\boldsymbol{n} + \varpi_0\frac{\lambda_{K_l}}{h_F}v_h\right)w_h \,\mathrm{d}s, \tag{41.18}$$

and $\ell_h(w_h) := \ell(w_h) + \sum_{F \in \mathcal{F}_h^\partial} \varpi_0\frac{\lambda_{K_l}}{h_F}\int_F gw_h \,\mathrm{d}s$, where the exact forms a and ℓ are defined in (40.4), $F := \partial K_l \cap \partial D$, and the user-dependent penalty parameter ϖ_0 is yet to be chosen large enough.

We equip V_h with the norm $\|v_h\|_{V_h}^2 := \|\lambda^{\frac{1}{2}}\nabla v_h\|_{\boldsymbol{L}^2(D)}^2 + |v_h|_\partial^2$ with $|v_h|_\partial^2 := \sum_{F \in \mathcal{F}_h^\partial} \frac{\lambda_{K_l}}{h_F}\|v_h\|_{L^2(F)}^2$. Recall the discrete trace inequality stating that there is c_{dt} s.t. $\|\boldsymbol{n}\cdot\nabla v_h\|_{L^2(F)} \leq c_{\mathrm{dt}}h_F^{-\frac{1}{2}}\|\nabla v_h\|_{\boldsymbol{L}^2(K_l)}$ for all $v_h \in V_h$, all $F \in \mathcal{F}_h^\partial$, and all $h \in \mathcal{H}$. Let n_∂ denote the maximum number of boundary faces that a mesh cell can have ($n_\partial \leq d$ for simplicial meshes), and let $\mathcal{T}_h^{\partial D}$ be the collection of the mesh cells having at least one boundary face.

Lemma 41.6 (Coercivity, well-posedness). *Assume that the penalty parameter satisfies $\varpi_0 > \frac{1}{4}n_\partial c_{\mathrm{dt}}^2$. (i) a_h is coercive on V_h with constant $\alpha := \frac{\varpi_0 - \frac{1}{4}n_\partial c_{\mathrm{dt}}^2}{1+\varpi_0} > 0$. (ii) The discrete problem (41.3) is well-posed.*

Proof. Adapting the proof of Lemma 37.2, we infer that for all $v_h \in V_h$,

$$\left|\sum_{F \in \mathcal{F}_h^\partial} \int_F \lambda_{K_l}(\boldsymbol{n}\cdot\nabla v_h)v_h \,\mathrm{d}s\right| \leq n_\partial^{\frac{1}{2}}c_{\mathrm{dt}}\left(\sum_{K\in\mathcal{T}_h^{\partial D}} \lambda_K\|\nabla v_h\|_{\boldsymbol{L}^2(K)}^2\right)^{\frac{1}{2}}|v_h|_\partial.$$

The rest of the proof is similar to that of Lemma 37.3. □

Let $V_\sharp := V_s + V_h$ be equipped with the norm $\|v\|_{V_\sharp}^2 := |v|_{\lambda,p,q}^2 + |v|_\partial^2$ with $|v|_{\lambda,p,q}$ defined in (41.11) (the summations involving the terms $\|\nabla(v_{|K})\|_{\boldsymbol{L}^p(K)}$ and $\|\Delta(v_{|K})\|_{L^q(K)}$ in the definition of $|\cdot|_{\lambda,p,q}$ in (41.11) can be restricted here to $K \in \mathcal{T}_h^{\partial D}$) and $|v|_\partial^2 := \sum_{F \in \mathcal{F}_h^\partial} \frac{\lambda_{K_l}}{h_F} \|v\|_{L^2(F)}^2$. Invoking inverse inequalities shows that there is $c_\sharp$ s.t. $\|v_h\|_{V_\sharp} \le c_\sharp \|v_h\|_{V_h}$ for all $v_h \in V_h$ and all $h \in \mathcal{H}$, i.e., (27.5) holds true.

Lemma 41.7 (Consistency/boundedness). *There is $\omega_\sharp$, uniform w.r.t. $u \in V_s$ and λ, but depending on p and q, s.t. $\|\delta_h(v_h)\|_{V_h'} \le \omega_\sharp \|u - v_h\|_{V_\sharp}$ for all $v_h \in V_h$ and all $h \in \mathcal{H}$.*

Proof. Let $v_h, w_h \in V_h$. Using the identity (41.8a) for $n_\sharp$, $[\![w_h]\!]_F = 0$ for all $F \in \mathcal{F}_h^\circ$ (since V_h is H^1-conforming), and the definition of the weights at the boundary faces, we infer that $n_\sharp(v_h, w_h) = \sum_{F \in \mathcal{F}_h^\partial} \int_F \boldsymbol{\sigma}(v_h) \cdot \boldsymbol{n} w_h \, \mathrm{d}s$. Hence, $a_h(v_h, w_h) = a(v_h, w_h) + n_\sharp(v_h, w_h) + \sum_{F \in \mathcal{F}_h^\partial} \varpi_0 \frac{\lambda_{K_l}}{h_F} \int_F v_h w_h \, \mathrm{d}s$. Invoking the identity (41.8b) for the exact solution u and observing that $f = \nabla \cdot \boldsymbol{\sigma}(u)$, we infer that $\int_D f w_h \, \mathrm{d}x = a(u, w_h) + n_\sharp(u, w_h)$. Recalling that $\gamma^g(u) = g$, and letting $\eta := u - v_h$, we obtain

$$\langle \delta_h(v_h), w_h \rangle_{V_h', V_h} = a(\eta, w_h) + n_\sharp(\eta, w_h) + \sum_{F \in \mathcal{F}_h^\partial} \varpi_0 \frac{\lambda_{K_l}}{h_F} \int_F \eta w_h \, \mathrm{d}s.$$

We conclude using the Cauchy–Schwarz inequality and the boundedness of $n_\sharp$ from (41.9), where the summation in $|v|_{n_\sharp}$ can be restricted to the mesh cells in $\mathcal{T}_h^{\partial D}$ since $[\![w_h]\!]_F = 0$ across all the mesh interfaces. □

Theorem 41.8 (Error estimate). *Let u solve (40.3) and u_h solve (41.3) with the penalty parameter $\varpi_0 > \frac{1}{4} n_\partial c_{\mathrm{dt}}^2$. Assume that $u \in H^{1+r}(D)$, $r > 0$. (i) There is c, uniform w.r.t. λ, s.t. for all $h \in \mathcal{H}$,*

$$\|u - u_h\|_{V_\sharp} \le c \inf_{v_h \in V_h} \|u - v_h\|_{V_\sharp}. \tag{41.19}$$

(ii) Letting $t := \min(r, k)$ and $\chi_t := 1$ if $t \le 1$ and $\chi_t := 0$ if $t > 1$, we have

$$\|u - u_h\|_{V_\sharp} \le c \left(\sum_{K \in \mathcal{T}_h} \lambda_K h_K^{2t} |u|_{H^{1+t}(\check{\mathcal{T}}_K)}^2 + \frac{\chi_t}{\lambda_K} h_K^{2d(\frac{d+2}{2d} - \frac{1}{q})} \|f\|_{L^q(K)}^2 \right)^{\frac{1}{2}}, \tag{41.20}$$

where $\check{\mathcal{T}}_K$ is the collection of the mesh cells having at least a common vertex with K, and $|u|_{H^{1+t}(\check{\mathcal{T}}_K)}$ can be replaced by $|u|_{H^{1+t}(K)}$ if $1 + t > \frac{d}{2}$.

Proof. (i) The estimate (41.19) follows from Lemma 27.5 combined with stability (Lemma 41.6) and consistency/boundedness (Lemma 41.7).
(ii) We bound the infimum in (41.19) by considering $\eta := u - \mathcal{I}_h^{g,\mathrm{av}}(u)$, where $\mathcal{I}_h^{g,\mathrm{av}}$ is the quasi-interpolation operator introduced in §22.3. We take the polynomial degree of $\mathcal{I}_h^{g,\mathrm{av}}$ to be $\ell := \lceil t \rceil$ (recall that $\lceil t \rceil$ is the smallest integer $n \in \mathbb{N}$ s.t. $n \ge t$). Notice that $\ell \ge 1$ because $r > 0$ and $k \ge 1$, and $\ell \le k$ because $t \le k$. Hence, $\mathcal{I}_h^{g,\mathrm{av}}(u) \in V_h$. We need to bound all the terms composing the norm $\|\eta\|_{V_\sharp}$. Owing to Theorem 22.6 (with $m := 1$), we have $\|\nabla(\eta_{|K})\|_{\boldsymbol{L}^2(K)} \le c h_K^t |u|_{H^{1+t}(\check{\mathcal{T}}_K)}$ for all $K \in \mathcal{T}_h$. Moreover, using Exercise 22.5, we have $h_F^{-\frac{1}{2}} \|\eta\|_{L^2(F)} \le c h_{K_l}^t |u|_{H^{1+t}(\check{\mathcal{T}}_{K_l})}$ for all $F \in \mathcal{F}_h^\partial$. It remains to estimate $h_K^{d(\frac{1}{2} - \frac{1}{p})} \|\nabla(\eta_{|K})\|_{\boldsymbol{L}^p(K)}$ and $h_K^{d(\frac{d+2}{2d} - \frac{1}{q})} \|\Delta(\eta_{|K})\|_{L^q(K)}$ for all $K \in \mathcal{T}_h^{\partial D}$. Using (41.16), the above bound on $\|\nabla(\eta_{|K})\|_{\boldsymbol{L}^2(K)}$, and $|\nabla(\eta_{|K})|_{\boldsymbol{H}^t(K)} = |\nabla u|_{\boldsymbol{H}^t(K)} = |u|_{H^{1+t}(K)}$ since $\ell < 1 + t$, we infer that $h_K^{d(\frac{1}{2} - \frac{1}{p})} \|\nabla(\eta_{|K})\|_{\boldsymbol{L}^p(K)} \le c h_K^t |u|_{H^{1+t}(\check{\mathcal{T}}_K)}$. Moreover, if $t \le 1$, we

have $\ell = 1$ so that $\|\Delta(\eta_{|K})\|_{L^q(K)} = \|\Delta u\|_{L^q(K)} = \lambda_K^{-1}\|f\|_{L^q(K)}$. But, if $t > 1$, we infer that $r > 1$ so that we can set $q := 2$ (recall that $f_{|D_i} = -\lambda_{|D_i}(\Delta u)_{|D_i}$ for all $i \in \{1{:}M\}$, and $u \in H^2(D)$ if $r \geq 1$), and we estimate $\|\Delta(\eta_{|K})\|_{L^2(K)}$ by using Theorem 22.6 (with $m := 2$). Finally, if $1+t > \frac{d}{2}$, we can use the canonical interpolation operator $\mathcal{I}_h^{\mathrm{g}}$ instead of $\mathcal{I}_h^{\mathrm{g,av}}$, and this allows us to replace $|u|_{H^{1+t}(\check{\mathcal{T}}_K)}$ by $|u|_{H^{1+t}(K)}$ in (41.20). $\qquad\square$

Remark 41.9 (Localization). One obtains the same upper bound as in (41.20) when using conforming finite elements for the approximation, i.e., $V_h \subset H_0^1(D)$; see Exercise 41.1. Notice also that $\sum_{K \in \mathcal{T}_h} \lambda_K h_K^{2t}|u|_{H^{1+t}(\check{\mathcal{T}}_K)}^2 \leq c\|\lambda\|_{L^\infty(D)} \sum_{K \in \mathcal{T}_h} h_K^{2t}|u|_{H^{1+t}(K)}^2$. $\qquad\square$

Remark 41.10 (Literature). An alternative analysis based on the approach of Gudi [226] is developed in Lüthen et al. [291]. $\qquad\square$

41.4 Discontinuous Galerkin

We consider in this section the symmetric interior penalty (SIP) discontinuous Galerkin method introduced in Chapter 38 (i.e., $\theta := 1$ in (38.20)). The discrete problem is (41.3) with $V_h := P_k^{\mathrm{b}}(\mathcal{T}_h)$, $k \geq 1$, the bilinear form

$$a_h(v_h, w_h) := \int_D \lambda \nabla_h v_h \cdot \nabla_h w_h \, \mathrm{d}x + \sum_{F \in \mathcal{F}_h} \int_F \{\boldsymbol{\sigma}(v_h)\}_\theta \cdot \boldsymbol{n}_F \llbracket w_h \rrbracket \, \mathrm{d}s$$

$$+ \sum_{F \in \mathcal{F}_h} \int_F \llbracket v_h \rrbracket \{\boldsymbol{\sigma}(w_h)\}_\theta \cdot \boldsymbol{n}_F \, \mathrm{d}s + \sum_{F \in \mathcal{F}_h} \varpi_0 \frac{\lambda_F}{h_F} \int_F \llbracket v_h \rrbracket \llbracket w_h \rrbracket \, \mathrm{d}s,$$

and the linear form $\ell_h(w_h) := \ell(w_h) + \sum_{F \in \mathcal{F}_h^\partial} \varpi_0 \frac{\lambda_{K_l}}{h_F} \int_F g w_h \, \mathrm{d}s$, where ℓ is defined in (40.4), and the user-specified penalty parameter ϖ_0 is yet to be chosen large enough. We equip V_h with the norm $\|v_h\|_{V_h}^2 := \|\lambda^{\frac{1}{2}}\nabla_h v_h\|_{\boldsymbol{L}^2(D)}^2 + |v_h|_{\mathrm{J}}^2$ with $|v_h|_{\mathrm{J}}^2 := \sum_{F \in \mathcal{F}_h} \frac{\lambda_F}{h_F}\|\llbracket v_h \rrbracket\|_{L^2(F)}^2$. Recall the discrete trace inequality stating that there is c_{dt} s.t. $\|\boldsymbol{n}_F \cdot \nabla v_h\|_{L^2(F)} \leq c_{\mathrm{dt}} h_F^{-\frac{1}{2}}\|\nabla v_h\|_{\boldsymbol{L}^2(K)}$ for all $v_h \in V_h$, all $K \in \mathcal{T}_h$, all $F \in \mathcal{F}_K$, and all $h \in \mathcal{H}$. Let n_∂ denote the maximum number of faces that a mesh cell can have ($n_\partial \leq d + 1$ for simplicial meshes).

Lemma 41.11 (Coercivity, well-posedness). *Assume that the penalty parameter satisfies* $\varpi_0 > n_\partial c_{\mathrm{dt}}^2$. *(i)* a_h *is coercive on* V_h *with constant* $\alpha := \frac{\varpi_0 - n_\partial c_{\mathrm{dt}}^2}{1 + \varpi_0} > 0$. *(ii) The discrete problem (41.3) is well-posed.*

Proof. Let $v_h \in V_h$. Our starting observation is that

$$a_h(v_h, v_h) = \|\lambda^{\frac{1}{2}}\nabla_h v_h\|_{\boldsymbol{L}^2(D)}^2 + 2n_\sharp(v_h, v_h) + \varpi_0|v_h|_{\mathrm{J}}^2.$$

Proceeding as in the proof of Lemma 38.5 and Lemma 38.6, and using that $|\mathcal{T}_F|\theta_{K,F}\lambda_K\lambda_F^{-1} = 1$ for all $F \in \mathcal{F}_h$ and all $K \in \mathcal{T}_F$, the reader is invited to verify that

$$|n_\sharp(v_h, w_h)| \leq n_\partial^{\frac{1}{2}} c_{\mathrm{dt}} \|\lambda^{\frac{1}{2}}\nabla_h v_h\|_{\boldsymbol{L}^2(D)}|w_h|_{\mathrm{J}}. \tag{41.21}$$

We can then conclude as in the proof of Lemma 38.6. $\qquad\square$

Let $V_\sharp := V_{\mathrm{s}} + V_h$ be equipped with the norm $\|v\|_{V_\sharp}^2 := |v|_{\lambda,p,q}^2 + |v|_{\mathrm{J}}^2$ with $|v|_{\lambda,p,q}$ defined in (41.11) and $|v|_{\mathrm{J}}^2 := \sum_{F \in \mathcal{F}_h} \frac{\lambda_F}{h_F}\|\llbracket v \rrbracket\|_{L^2(F)}^2$. Invoking inverse inequalities shows that there is $c_\sharp$ s.t. $\|v_h\|_{V_\sharp} \leq c_\sharp\|v_h\|_{V_h}$ for all $v_h \in V_h$ and all $h \in \mathcal{H}$, i.e., (27.5) holds true.

Lemma 41.12 (Consistency/boundedness). *There is $\omega_\sharp$, uniform w.r.t. $u \in V_S$ and λ, s.t. $\|\delta_h(v_h)\|_{V'_h} \leq \omega_\sharp \|u - v_h\|_{V_\sharp}$ for all $v_h \in V_h$ and all $h \in \mathcal{H}$.*

Proof. Let $v_h, w_h \in V_h$. Owing to the identity (41.8b) and since $f = \nabla \cdot \boldsymbol{\sigma}(u)$, we infer that $\int_D f w_h \, dx = \sum_{K \in \mathcal{T}_h} -(\boldsymbol{\sigma}(u), \nabla_h w_h)_{\boldsymbol{L}^2(K)} + n_\sharp(u, w_h)$. Hence, we have

$$\ell_h(w_h) = n_\sharp(u, w_h) - \int_D \boldsymbol{\sigma}(u) \cdot \nabla_h w_h \, dx + \sum_{F \in \mathcal{F}_h^\partial} \varpi_0 \int_F \frac{\lambda_F}{h_F} g w_h \, ds.$$

Using the identity (41.8a), we obtain

$$a_h(v_h, w_h) = \int_D -\boldsymbol{\sigma}(v_h) \cdot \nabla_h w_h \, dx + n_\sharp(v_h, w_h)$$
$$- \sum_{F \in \mathcal{F}_h} \int_F [\![v_h]\!] \{\boldsymbol{\sigma}(w_h)\}_\theta \cdot \boldsymbol{n}_F \, ds + \sum_{F \in \mathcal{F}_h} \varpi_0 \int_F \frac{\lambda_F}{h_F} [\![v_h]\!] [\![w_h]\!] \, ds.$$

Setting $\eta := u - v_h$ and using that $[\![u]\!]_F = 0$ for all $F \in \mathcal{F}_h^\circ$ and $[\![u]\!]_F = g$ for all $F \in \mathcal{F}_h^\partial$, we obtain

$$\langle \delta_h(v_h), w_h \rangle_{V'_h, V_h} = \int_D \lambda \nabla \eta \cdot \nabla_h w_h \, dx + n_\sharp(\eta, w_h)$$
$$- \sum_{F \in \mathcal{F}_h} \int_F [\![\eta]\!] \{\boldsymbol{\sigma}(w_h)\}_\theta \cdot \boldsymbol{n}_F \, ds + \sum_{F \in \mathcal{F}_h} \varpi_0 \int_F \frac{\lambda_F}{h_F} [\![\eta]\!] [\![w_h]\!] \, ds.$$

We conclude by using the Cauchy–Schwarz inequality for the first and the fourth terms on the right-hand side, using the boundedness estimate on $n_\sharp$ from (41.9) for the second term, and by proceeding as in the proof of (41.21) to bound the third term. $\qquad\square$

Theorem 41.13 (Error estimate). *Let u solve (40.3) and u_h solve (41.3) with the penalty parameter $\varpi_0 > n_\partial c_{\mathrm{dt}}^2$. Assume that $u \in H^{1+r}(D)$, $r > 0$. (i) There is c, uniform w.r.t. λ, s.t. for all $h \in \mathcal{H}$,*

$$\|u - u_h\|_{V_\sharp} \leq c \inf_{v_h \in V_h} \|u - v_h\|_{V_\sharp}. \tag{41.22}$$

(ii) Letting $t := \min(r, k)$ and $\chi_t := 1$ if $t \leq 1$ and $\chi_t := 0$ if $t > 1$, we have

$$\|u - u_h\|_{V_\sharp} \leq c \left(\sum_{K \in \mathcal{T}_h} \lambda_K h_K^{2t} |u|_{H^{1+t}(K)}^2 + \frac{\chi_t}{\lambda_K} h_K^{2d(\frac{d+2}{2d} - \frac{1}{q})} \|f\|_{L^q(K)}^2 \right)^{\frac{1}{2}}. \tag{41.23}$$

Proof. Proceed as in the proof of Theorem 41.8, where we now use the L^1-stable interpolation operator $\mathcal{I}_h^\sharp : L^1(D) \to P_k^{\mathrm{b}}(\mathcal{T}_h)$ from §18.3 to estimate the best-approximation error. $\qquad\square$

41.5 The hybrid high-order method

We consider in this section the hybrid high-order (HHO) method from Chapter 39. The discrete space $\hat{V}_{h,0}^k := V_{\mathcal{T}_h}^k \times V_{\mathcal{F}_h,0}^k$ is defined in (39.15) with $k \geq 0$. Recall that

$$V_{\mathcal{T}_h}^k := \{v_{\mathcal{T}_h} \in L^2(D) \mid v_{h|K} \in V_K^k, \forall K \in \mathcal{T}_h\}, \tag{41.24a}$$
$$V_{\mathcal{F}_h,0}^k := \{v_{\mathcal{F}_h} \in L^2(\mathcal{F}_h) \mid v_{\mathcal{F}_h|\partial K} \in V_{\partial K}^k, \forall K \in \mathcal{T}_h; \ v_{\mathcal{F}_h|\mathcal{F}_h^\partial} = 0\}, \tag{41.24b}$$

with $V_K^k := \mathbb{P}_{k,d} \circ \boldsymbol{T}_K^{-1}$ and $V_{\partial K}^k := \prod_{F \in \mathcal{F}_K} \mathbb{P}_{k,d-1} \circ \boldsymbol{T}_F^{-1}$, where $\boldsymbol{T}_K$ and $\boldsymbol{T}_F$ are affine geometric mappings. For every pair $\hat{v}_h := (v_{\mathcal{T}_h}, v_{\mathcal{F}_h}) \in \hat{V}_{h,0}^k$, $v_{\mathcal{T}_h}$ is a collection of cell polynomials of degree at most k, and $v_{\mathcal{F}_h}$ is a collection of face polynomials of degree at most k which are single-valued at the mesh interfaces and vanish at the boundary faces (so as to enforce strongly the homogeneous Dirichlet condition). Recall the notation $\hat{v}_K := (v_K, v_{\partial K}) \in \hat{V}_K^k := V_K^k \times V_{\partial K}^k$ with $v_K := v_{\mathcal{T}_h|K}$ and $v_{\partial K} := v_{\mathcal{F}_h|\partial K}$ for all $K \in \mathcal{T}_h$. The (λ-independent) local bilinear form $\hat{a}_K$ on $\hat{V}_K^k \times \hat{V}_K^k$ is defined as follows:

$$\hat{a}_K(\hat{v}_K, \hat{w}_K) := (\nabla \mathsf{R}(\hat{v}_K), \nabla \mathsf{R}(\hat{w}_K))_{\boldsymbol{L}^2(K)} + h_K^{-1}(\mathsf{S}(\hat{v}_K), \mathsf{S}(\hat{w}_K))_{L^2(\partial K)},$$

with the local reconstruction and stabilization operators R and S defined in (39.2) and in (39.4), respectively.

The discrete problem is as follows: Find $\hat{u}_h \in \hat{V}_{h,0}^k$ s.t.

$$\hat{a}_h(\hat{u}_h, \hat{w}_h) = \ell_h(w_{\mathcal{T}_h}), \quad \forall \hat{w}_h \in \hat{V}_{h,0}^k, \tag{41.25}$$

with the forms

$$\hat{a}_h(\hat{v}_h, \hat{w}_h) := \sum_{K \in \mathcal{T}_h} \lambda_K \hat{a}_K(\hat{v}_K, \hat{w}_K), \qquad \ell_h(w_{\mathcal{T}_h}) := \sum_{K \in \mathcal{T}_h} (f, w_K)_{L^2(K)}.$$

Recalling (39.6), we equip the discrete space $\hat{V}_{h,0}^k$ with the norm $\|\hat{v}_h\|_{\hat{V}_{h,0}^k}^2 := \sum_{K \in \mathcal{T}_h} \lambda_K |\hat{v}_K|_{\hat{V}_K^k}^2$, i.e., we set

$$\|\hat{v}_h\|_{\hat{V}_{h,0}^k}^2 := \sum_{K \in \mathcal{T}_h} \left(\lambda_K \|\nabla v_K\|_{\boldsymbol{L}^2(K)}^2 + \lambda_K h_K^{-1} \|v_K - v_{\partial K}\|_{L^2(\partial K)}^2 \right). \tag{41.26}$$

A straightforward consequence of Lemma 39.2 is that the bilinear form $\hat{a}_h$ is coercive on $\hat{V}_{h,0}^k$. Owing to the Lax–Milgram lemma, the discrete problem (41.25) is, therefore, well-posed.

As in Chapter 39, the local interpolation operator $\hat{\mathcal{I}}_K^k : H^1(K) \to \hat{V}_K^k$ for all $K \in \mathcal{T}_h$ is s.t. $\hat{\mathcal{I}}_K^k(v) := (\Pi_K^k(v), \Pi_{\partial K}^k(v_{|\partial K})) \in \hat{V}_K^k$ for all $v \in H^1(K)$, where Π_K^k and $\Pi_{\partial K}^k$ are the L^2-orthogonal projections onto V_K^k and $V_{\partial K}^k$, respectively. The local elliptic projection $\mathcal{E}_K : H^1(K) \to V_K^{k+1} := \mathbb{P}_{k+1,d} \circ \boldsymbol{T}_K^{-1}$ is s.t. $(\nabla(\mathcal{E}_K(v) - v), \nabla w)_{\boldsymbol{L}^2(K)} = 0$ for all $w \in V_K^{k+1}$, and $(\mathcal{E}_K(v) - v, 1)_{L^2(K)} = 0$. We define global counterparts of these operators, $\hat{\mathcal{I}}_h^k : H^1(D) \to V_{\mathcal{T}_h}^k \times V_{\mathcal{F}_h}^k$ and $\mathcal{E}_h : H^1(D) \to P_{k+1}^{\mathrm{b}}(\mathcal{T}_h)$, that are simply defined locally by setting $\hat{\mathcal{I}}_h^k(v)_{|K} := \hat{\mathcal{I}}_K^k(v_{|K})$ and $\mathcal{E}_h(v)_{|K} := \mathcal{E}_K(v_{|K})$.

Recalling the duality pairing $\langle \cdot, \cdot \rangle_F$ defined in (40.15), the generalization to the HHO method of the bilinear form $n_\sharp$ defined in (41.5) is the bilinear form defined on $(V_{\mathrm{s}} + P_{k+1}^{\mathrm{b}}(\mathcal{T}_h)) \times \hat{V}_{h,0}^k$ that acts as follows:

$$n_\sharp(v, \hat{w}_h) := \sum_{K \in \mathcal{T}_h} \sum_{F \in \mathcal{F}_K} \langle (\boldsymbol{\sigma}(v) \cdot \boldsymbol{n}_K)_{|F}, (w_K - w_{\partial K})_{|F} \rangle_F. \tag{41.27}$$

We now establish the counterparts of the identities (41.8a)-(41.8b) and the boundedness estimate (41.9).

Lemma 41.14 (Identities and boundedness for $n_\sharp$). *The following holds true for all $\hat{w}_h \in \hat{V}_{h,0}^k$, all $v_h \in P_{k+1}^{\mathrm{b}}(\mathcal{T}_h)$, and all $v \in V_{\mathrm{s}}$:*

$$n_\sharp(v_h, \hat{w}_h) = \sum_{K \in \mathcal{T}_h} \int_K \lambda_K \nabla v_{h|K} \cdot \nabla (\mathsf{R}(\hat{w}_K) - w_K) \, \mathrm{d}x, \tag{41.28a}$$

$$n_\sharp(v, \hat{w}_h) = \sum_{K \in \mathcal{T}_h} \int_K \left(\boldsymbol{\sigma}(v) \cdot \nabla w_K + (\nabla \cdot \boldsymbol{\sigma}(v)) w_K \right) \mathrm{d}x. \tag{41.28b}$$

Moreover, there is c, uniform w.r.t. λ, s.t. for all $v \in V_{\mathrm{s}} + P_{k+1}^{\mathrm{b}}(\mathcal{T}_h)$, all $\hat{w}_h \in \hat{V}_{h,0}^k$, and all $h \in \mathcal{H}$,

$$|n_\sharp(v, \hat{w}_h)| \le c \, |v|_{n_\sharp} \left(\sum_{K \in \mathcal{T}_h} \lambda_K h_K^{-1} \|w_K - w_{\partial K}\|_{L^2(\partial K)}^2 \right)^{\frac{1}{2}}, \tag{41.29}$$

with the seminorm $|\cdot|_{n_\sharp}$ defined in (40.20).

Proof. See Exercise 41.3 for the proof of (41.28). The proof of (41.29) uses the same arguments as the proof of Lemma 40.7. $\square$

Remark 41.15 ((41.28b)). The right-hand side of (41.28b) does not depend on the face-based function $w_{\mathcal{F}_h}$. This identity replaces the argument in the proof of Lemma 39.16 invoking the continuity of the normal component of $\boldsymbol{\sigma}(u)$ across the mesh interfaces, which makes sense only when the solution to (40.3) is smooth enough, say $\boldsymbol{\sigma}(u) \in \boldsymbol{H}^r(D)$ with $r > \frac{1}{2}$. $\square$

Let $V_\sharp := V_{\mathrm{s}} + P_{k+1}^{\mathrm{b}}(\mathcal{T}_h)$ be equipped with the seminorm $\|v\|_{V_\sharp} := |v|_{\lambda,p,q}$ with $|v|_{\lambda,p,q}$ defined in (41.11). Note that $\|v\|_{V_\sharp} = 0$ implies that $v = 0$ if v has zero mean value in each mesh cell $K \in \mathcal{T}_h$. This is the case for instance if one takes $v := u - \mathcal{E}_h(u)$. Recalling (39.29), the consistency error is defined s.t. $\langle \delta_h(\hat{\mathcal{I}}_h(u)), \hat{w}_h \rangle_{(\hat{V}_{h,0}^k)', \hat{V}_{h,0}^k} := \ell_h(w_{\mathcal{T}_h}) - \hat{a}_h(\hat{\mathcal{I}}_h^k(u), \hat{w}_h)$ for all $\hat{w}_h \in V_{h,0}^k$.

Lemma 41.16 (Consistency/boundedness). *There is $\omega_\sharp$, uniform w.r.t. $u \in V_{\mathrm{s}}$ and λ, s.t. for all $h \in \mathcal{H}$,*

$$\|\delta_h(\hat{\mathcal{I}}_h(u))\|_{(\hat{V}_{h,0}^k)'} \le \omega_\sharp \|u - \mathcal{E}_h(u)\|_{V_\sharp}. \tag{41.30}$$

Proof. Since $\boldsymbol{\sigma}(u) = -\lambda \nabla u$, $\nabla \cdot \boldsymbol{\sigma}(u) = f$, and $u \in V_{\mathrm{s}}$, (41.28b) implies that $\int_D f w_h \, \mathrm{d}x = \sum_{K \in \mathcal{T}_h} a_K(u, w_K) + n_\sharp(u, \hat{w}_h)$, where $a_K(u, w_K) := (-\boldsymbol{\sigma}(u), \nabla w_K)_{\boldsymbol{L}^2(K)}$. This implies that

$$\hat{\ell}_h(\hat{w}_h) = \sum_{K \in \mathcal{T}_h} a_K(u, w_K) + n_\sharp(u, \hat{w}_h).$$

Using first the definition of $\hat{a}_h$, then the identity $\mathsf{R} \circ \hat{\mathcal{I}}_K^k = \mathcal{E}_K$, and finally (41.28a) with $v_h := \mathcal{E}_h(u)$, we obtain

$$\hat{a}_h(\hat{\mathcal{I}}_h^k(u), \hat{w}_h) = \sum_{K \in \mathcal{T}_h} a_K(\mathcal{E}_K(u), w_K) + n_\sharp(\mathcal{E}_h(u), \hat{w}_h) + \sum_{K \in \mathcal{T}_h} \lambda_K h_K^{-1}(\mathsf{S}(\hat{\mathcal{I}}_K^k(u)), \mathsf{S}(\hat{w}_K))_{L^2(\partial K)}.$$

Subtracting these two identities and using that $a_K(u - \mathcal{E}_K(u), w_K) = 0$ for all $K \in \mathcal{T}_h$ leads to $\langle \delta_h(\hat{\mathcal{I}}_h^k(u)), \hat{w}_h \rangle_{(\hat{V}_{h,0}^k)', \hat{V}_{h,0}^k} = \mathfrak{T}_1 + \mathfrak{T}_2$ with

$$\mathfrak{T}_1 := n_\sharp(u - \mathcal{E}_K(u), \hat{w}_h),$$
$$\mathfrak{T}_2 := -\sum_{K \in \mathcal{T}_h} \lambda_K h_K^{-1}(\mathsf{S}(\hat{\mathcal{I}}_K^k(u)), \mathsf{S}(\hat{w}_K)_{L^2(\partial K)}.$$

To bound $\mathfrak{T}_1$, we invoke (41.29) and use $\sum_{K \in \mathcal{T}_h} \lambda_K h_K^{-1} \|w_K - w_{\partial K}\|_{L^2(\partial K)}^2 \le \|\hat{w}_h\|_{\hat{V}_{h,0}^k}^2$ owing to (41.26). We bound $\mathfrak{T}_2$ as in the proof of Lemma 39.16. $\square$

Theorem 41.17 (Error estimate). *Let u solve (40.3) and $\hat{u}_h \in \hat{V}_{h,0}^k$ solve (41.25). Assume that $u \in H^{1+r}(D)$, $r > 0$. (i) There is c, uniform w.r.t. λ, s.t. for all $h \in \mathcal{H}$,*

$$\sum_{K \in \mathcal{T}_h} \lambda_K \|\nabla(u - \mathsf{R}(\hat{u}_K))\|_{\boldsymbol{L}^2(K)}^2 \le c \, \|u - \mathcal{E}_h(u)\|_{V_\sharp}^2. \tag{41.31}$$

(ii) *Let $t := \min(r, k+1)$ and $\chi_t := 1$ if $t \leq 1$ and $\chi_t := 0$ if $t > 1$. We have*

$$\sum_{K \in \mathcal{T}_h} \lambda_K \|\nabla(u - \mathsf{R}(\hat{u}_K))\|^2_{\boldsymbol{L}^2(K)} \leq c \sum_{K \in \mathcal{T}_h} \lambda_K h_K^{2t} |u|^2_{H^{1+t}(K)} + \frac{\chi_t}{\lambda_K} h_K^{2d(\frac{d+2}{2d} - \frac{1}{q})} \|f\|^2_{L^q(K)}. \qquad (41.32)$$

Proof. See Exercise 41.3.   □

Remark 41.18 (HHO vs. dG). The HHO method is somewhat simpler than the dG method when it comes to solving problems with contrasted coefficients. For the HHO method, one assembles cellwise the local bilinear forms $\hat{a}_K$ weighted by the local diffusion coefficient λ_K, whereas for the dG method one has to invoke interface-based values of the diffusion coefficient to construct the penalty term.   □

Exercises

Exercise 41.1 (Conforming finite elements). Consider the approximation of (40.3) by conforming finite elements. Let $V := H_0^1(D)$, $V_h := P_{k,0}^{\mathrm{g}}(\mathcal{T}_h) \subset V$, $k \geq 1$, and consider the norm $\|v\|_V := \|\lambda^{\frac{1}{2}} \nabla v\|_{\boldsymbol{L}^2(D)}$. Assume $u \in H^{1+r}(D)$, $r > 0$, and set $t := \min(r, k)$. Prove that there is c, uniform w.r.t. λ, s.t. $\|u - u_h\|_V \leq c(\sum_{K \in \mathcal{T}_h} \lambda_K h_K^{2t} |u|^2_{H^{1+t}(\check{\mathcal{T}}_K)})^{\frac{1}{2}}$ for all $h \in \mathcal{H}$, where $\check{\mathcal{T}}_K$ is the collection of the mesh cells sharing at least a vertex with K, and that $|u|_{H^{1+t}(\check{\mathcal{T}}_K)}$ can be replaced by $|u|_{H^{1+t}(K)}$ if $1 + t > \frac{d}{2}$.

Exercise 41.2 (dG). Prove the estimate (41.21).

Exercise 41.3 (HHO). (i) Prove (41.28a) (*Hint*: adapt the proof of (40.18a), i.e., use the definition of the pairing $\langle \cdot, \cdot \rangle_F$ together with the definition (39.2) for R). (ii) Prove (41.28b). (*Hint*: adapt the proof of (40.18b). (iii) Prove the error bound (41.31). (*Hint*: see the proof of (39.32) in Theorem 39.17.) (iv) Prove (41.32). (*Hint*: set $\ell = \lceil t \rceil$ and consider the elliptic projection of degree ℓ, say $\mathcal{E}_K^\ell$, for all $K \in \mathcal{T}_h$.)

Chapter 42

Linear elasticity

The four chapters composing Part IX deal with the approximation of vector-valued elliptic PDEs endowed with a multicomponent coercivity property either in $\boldsymbol{H}^1$ (linear elasticity) or in $\boldsymbol{H}(\mathrm{curl})$ (Maxwell's equations in some specific regimes). The present chapter is concerned with the linear elasticity equations where the main tool to establish coercivity is Korn's inequality. We consider $\boldsymbol{H}^1$-conforming and nonconforming approximations, and we address the robustness of the approximation in the incompressible limit.

42.1 Continuum mechanics

Let D be a Lipschitz domain in $\mathbb{R}^d$, $d = 3$. We assume that D represents a deformable medium, initially at equilibrium, that is subjected to an external load $\boldsymbol{f} : D \to \mathbb{R}^3$. Our goal is to determine the displacement field $\boldsymbol{u} : D \to \mathbb{R}^3$ induced by $\boldsymbol{f}$ once the system has reached equilibrium again. Let $\mathbb{s} : D \to \mathbb{R}^{3\times 3}$ be the *stress tensor* in the medium. We write $\mathbb{s}(\boldsymbol{u})$ since this tensor depends on the displacement field. The equilibrium conditions under the external load $\boldsymbol{f}$ can be expressed as

$$\nabla{\cdot}\mathbb{s}(\boldsymbol{u}) + \boldsymbol{f} = \boldsymbol{0} \quad \text{in } D, \tag{42.1}$$

and the balance of the angular momentum requires that $\mathbb{s}(\boldsymbol{u})$ be symmetric, i.e., $\mathbb{s}(\boldsymbol{u}) = \mathbb{s}(\boldsymbol{u})^\mathsf{T}$. We assume that the deformations are small enough so that the linear elasticity theory applies. Let $\mathbb{e}(\boldsymbol{u}) : D \to \mathbb{R}^{3\times 3}$ be the (linearized) *strain rate tensor* defined as

$$\mathbb{e}(\boldsymbol{u}) := \frac{1}{2}(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^\mathsf{T}). \tag{42.2}$$

In the framework of linear isotropic elasticity, the stress tensor is related to the strain rate tensor by the relation

$$\mathbb{s}(\boldsymbol{u}) = 2\mu\mathbb{e}(\boldsymbol{u}) + \lambda\,\mathrm{tr}(\mathbb{e}(\boldsymbol{u}))\mathbb{I}_d, \tag{42.3}$$

where λ and μ are the phenomenological parameters called *Lamé coefficients*, and $\mathbb{I}_d$ is the identity tensor in $\mathbb{R}^{d\times d}$. Using (42.2), we also have

$$\mathbb{s}(\boldsymbol{u}) = \mu(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^\mathsf{T}) + \lambda(\nabla{\cdot}\boldsymbol{u})\mathbb{I}_d.$$

Owing to thermodynamic stability, the Lamé coefficients are such that $\mu > 0$ and $\lambda + \frac{2}{3}\mu > 0$. We henceforth assume that there are $\mu_{\min}, \kappa_{\min} > 0$ s.t.

$$\mu, \lambda \in L^\infty(D), \quad \mu(\boldsymbol{x}) \geq \mu_{\min}, \quad \lambda(\boldsymbol{x}) + \tfrac{2}{3}\mu(\boldsymbol{x}) \geq \kappa_{\min}, \quad \text{a.e. } \boldsymbol{x} \text{ in } D. \tag{42.4}$$

Remark 42.1 (Cauchy–Navier). If μ and λ are constant in D, the identity $\nabla{\cdot}\mathbb{e}(\boldsymbol{u}) = \frac{1}{2}\big(\Delta\boldsymbol{u} + \nabla(\nabla{\cdot}\boldsymbol{u})\big)$ implies that (42.1) can be rewritten $-\mu\Delta\boldsymbol{u} - (\mu + \lambda)\nabla(\nabla{\cdot}\boldsymbol{u}) = \boldsymbol{f}$ in D. This PDE is called *Cauchy–Navier formulation* in linear elasticity. □

Remark 42.2 (Incompressibility). The coefficient $\kappa := \lambda + \frac{2}{3}\mu$, called bulk modulus, describes the compressibility of the material. Very large values w.r.t. μ, i.e., $\lambda \gg \mu$, correspond to almost incompressible materials. □

Remark 42.3 (Material parameters). Instead of using λ and μ, it is sometimes more convenient to consider the *Young modulus*, E, and the *Poisson coefficient*, ν, defined as follows:

$$E := \mu\frac{3\lambda + 2\mu}{\lambda + \mu} \qquad \nu := \frac{\lambda}{2(\lambda + \mu)}. \tag{42.5}$$

The Poisson coefficient is such that $-1 < \nu < \frac{1}{2}$. An almost incompressible material corresponds to a Poisson coefficient very close to $\frac{1}{2}$. □

Remark 42.4 (Linearity). The linear isotropic elasticity model is in general valid for problems involving infinitesimal strains. In this case, the medium responds linearly to externally applied loads so that one can normalize the problem and consider arbitrary loads. □

Remark 42.5 (A bit of history). The finite element method was originally developed in the 1950s by aeronautical engineers to solve problems of continuum mechanics that could not be easily handled by classical finite difference techniques since they involved complex geometries; see, e.g., Levy [282], Argyris and Kelsey [14], and the references cited in Oden [317]. At the same time, theoretical researches on the approximation of the linear elasticity equations were carried out by Turner et al. [367], and eventually in 1960, Clough [129] coined the terminology "finite elements". □

Definition 42.6 (Rigid displacement). *A rigid displacement $\boldsymbol{r} : D \to \mathbb{R}^3$ is a global motion of the medium D consisting of a translation and a rotation, i.e., $\boldsymbol{r}$ is a member of the following six-dimensional vector space:*

$$\boldsymbol{R} := \mathbb{P}_{0,3} + \boldsymbol{x}{\times}\mathbb{P}_{0,3} = \mathbb{N}_{0,3}, \tag{42.6}$$

where $\mathbb{N}_{0,3}$ is the lowest-order Nédélec polynomial space defined in §15.1.

Lemma 42.7 (Kernel of strain rate). *For all $\boldsymbol{r} \in \boldsymbol{L}^1_{\mathrm{loc}}(D)$, $\boldsymbol{r} \in \boldsymbol{R}$ iff $\mathbb{e}(\boldsymbol{r}) = 0$.*

Proof. Let $\boldsymbol{r} \in \boldsymbol{R}$. Then $\nabla\boldsymbol{r}$ is skew-symmetric so that $\mathbb{e}(\boldsymbol{r}) = 0$. Conversely, let $\boldsymbol{r} \in \boldsymbol{L}^1_{\mathrm{loc}}(D)$ be such that $\mathbb{e} := \mathbb{e}(\boldsymbol{r}) = 0$. Since $\partial_{ij}r_k = \partial_{ji}r_k$ in the distribution sense for all $i, j, k \in \{1{:}3\}$, we have

$$\begin{aligned}
\partial_k(\partial_j r_i) &= \partial_k e_{ij} + \tfrac{1}{2}\partial_k\partial_j r_i - \tfrac{1}{2}\partial_k\partial_i r_j \\
&= \partial_k e_{ij} + \tfrac{1}{2}\partial_j(\partial_k r_i + \partial_i r_k) - \tfrac{1}{2}\partial_i(\partial_k r_j + \partial_j r_k) \\
&= \partial_k e_{ij} + \partial_j e_{ik} - \partial_i e_{jk} = 0.
\end{aligned}$$

This implies that all the Cartesian components of $\boldsymbol{r}$ are first-order polynomials, i.e., $\boldsymbol{r}(\boldsymbol{x}) = \boldsymbol{\alpha} + \mathbb{B}\boldsymbol{x}$, with $\boldsymbol{\alpha} \in \mathbb{R}^3$ and $\mathbb{B} \in \mathbb{R}^{3\times 3}$. Moreover, $\mathbb{e}(\boldsymbol{r}) = 0$ implies that $\mathbb{B} + \mathbb{B}^{\mathsf{T}} = 0$, i.e., the matrix $\mathbb{B}$ is skew-symmetric. Therefore, there exists a vector $\boldsymbol{\beta} \in \mathbb{R}^3$ such that $\mathbb{B}\boldsymbol{x} = \boldsymbol{\beta}{\times}\boldsymbol{x}$. Thus, $\boldsymbol{r} \in \boldsymbol{R}$. □
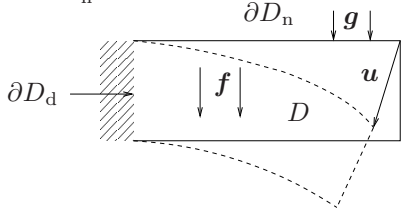
Lemma 42.7 implies that if the displacement field $\boldsymbol{u}$ satisfies the equilibrium condition (42.1), then $\boldsymbol{u} + \boldsymbol{r}$, where $\boldsymbol{r}$ is a rigid displacement, also satisfies this equation. We will see below that the rigid displacements can be controlled by the boundary conditions.

42.2 Weak formulation and well-posedness

In this section, we present a weak formulation for the linear elasticity problem and establish its well-posedness in the framework of the Lax–Milgram lemma. The key tool to prove coercivity are Korn's inequalities.

42.2.1 Weak formulation

The model problem (42.1)-(42.3) must be supplemented with boundary conditions. We consider a boundary partition $\partial D = \partial D_{\mathrm{d}} \cup \partial D_{\mathrm{n}}$ such that $|\partial D_{\mathrm{d}}| > 0$. The displacement is imposed to vanish on ∂D_{d} (we say that the medium is clamped at ∂D_{d}), and a normal load $\boldsymbol{g} : \partial D_{\mathrm{n}} \to \mathbb{R}^3$ is imposed on ∂D_{n}. This leads to the following problem:

$$\nabla{\cdot}\mathbb{s}(\boldsymbol{u}) + \boldsymbol{f} = \boldsymbol{0} \qquad \text{in } D, \tag{42.7a}$$

$$\mathbb{s}(\boldsymbol{u}) = 2\mu\mathbb{e}(\boldsymbol{u}) + \lambda\,\mathrm{tr}(\mathbb{e}(\boldsymbol{u}))\mathbb{I}_d \ \text{ in } D, \tag{42.7b}$$

$$\boldsymbol{u} = \boldsymbol{0} \qquad \text{on } \partial D_{\mathrm{d}}, \tag{42.7c}$$

$$\mathbb{s}(\boldsymbol{u})\boldsymbol{n} = \boldsymbol{g} \qquad \text{on } \partial D_{\mathrm{n}}. \tag{42.7d}$$

Clamping the medium at ∂D_{d} allows one to control the rigid displacements since the only field $\boldsymbol{r} \in \boldsymbol{R}$ such that $\boldsymbol{r}_{|\partial D_{\mathrm{d}}} = \boldsymbol{0}$ is $\boldsymbol{r} = \boldsymbol{0}$ provided $|\partial D_{\mathrm{d}}| > 0$.

To derive a weak formulation for (42.7), we take the scalar product of the equilibrium equation with a smooth test function $\boldsymbol{v} : D \to \mathbb{R}^3$. Since $\int_D \big(\nabla{\cdot}\mathbb{s}(\boldsymbol{u})\big){\cdot}\boldsymbol{v}\,\mathrm{d}x = -\int_D \mathbb{s}(\boldsymbol{u}){:}\nabla\boldsymbol{v}\,\mathrm{d}x + \int_{\partial D} \boldsymbol{v}{\cdot}\mathbb{s}(\boldsymbol{u})\boldsymbol{n}\,\mathrm{d}s$ and $\mathbb{s}(\boldsymbol{u}){:}\nabla\boldsymbol{v} = \mathbb{s}(\boldsymbol{u}){:}\mathbb{e}(\boldsymbol{v})$ owing to the symmetry of $\mathbb{s}(u)$ (here the double dot product is defined as $\mathbb{e}{:}\mathbb{s} := \sum_{j,k\in\{1:d\}} e_{jk}s_{jk} = \mathrm{tr}(\mathbb{e}\,\mathbb{s}^{\mathsf{T}})$), we have

$$\int_D \mathbb{s}(\boldsymbol{u}){:}\mathbb{e}(\boldsymbol{v})\,\mathrm{d}x - \int_{\partial D} \boldsymbol{v}{\cdot}\mathbb{s}(\boldsymbol{u})\boldsymbol{n}\,\mathrm{d}s = \int_D \boldsymbol{f}{\cdot}\boldsymbol{v}\,\mathrm{d}x.$$

The displacement $\boldsymbol{u}$ and the test function $\boldsymbol{v}$ are taken in the functional space

$$\boldsymbol{V}_{\mathrm{d}} := \{\boldsymbol{v} \in \boldsymbol{H}^1(D) \mid \gamma^{\mathrm{g}}(\boldsymbol{v})_{|\partial D_{\mathrm{d}}} = \boldsymbol{0}\}, \tag{42.8}$$

where $\gamma^{\mathrm{g}} : \boldsymbol{H}^1(D) \to \boldsymbol{H}^{\frac{1}{2}}(\partial D)$ acts componentwise as the trace map from Theorem 3.10, i.e., $\gamma^{\mathrm{g}}(\boldsymbol{v}) = \boldsymbol{v}_{|\partial D}$ if $\boldsymbol{v}$ is a smooth function. Since the measure of ∂D_{d} is positive, the following Poincaré–Steklov inequality holds true on $\boldsymbol{V}_{\mathrm{d}}$: There is $\tilde{C}_{\mathrm{PS}} > 0$ such that

$$\tilde{C}_{\mathrm{PS}}\|\boldsymbol{v}\|_{\boldsymbol{L}^2(D)} \leq \ell_D\|\nabla\boldsymbol{v}\|_{\mathbb{L}^2(D)}, \qquad \forall\boldsymbol{v} \in \boldsymbol{V}_{\mathrm{d}}, \tag{42.9}$$

where ℓ_D is a length scale associated with D, e.g., $\ell_D := \mathrm{diam}(D)$. Therefore, $\boldsymbol{V}_{\mathrm{d}} \ni \boldsymbol{v} \mapsto \|\boldsymbol{v}\|_{\boldsymbol{V}_{\mathrm{d}}} := \|\nabla\boldsymbol{v}\|_{\mathbb{L}^2(D)} = |\boldsymbol{v}|_{\boldsymbol{H}^1(D)}$ is a norm on $\boldsymbol{V}_{\mathrm{d}}$. This norm is equivalent to the $\boldsymbol{H}^1$-norm in $\boldsymbol{V}_{\mathrm{d}}$ since $\|\boldsymbol{v}\|_{\boldsymbol{V}_{\mathrm{d}}} \leq \ell_D^{-1}\|\boldsymbol{v}\|_{\boldsymbol{H}^1(D)} \leq (1+\tilde{C}_{\mathrm{PS}}^{-2})^{\frac{1}{2}}\|\boldsymbol{v}\|_{\boldsymbol{V}_{\mathrm{d}}}$, where $\|\boldsymbol{v}\|_{\boldsymbol{H}^1(D)} := (\|\boldsymbol{v}\|_{\boldsymbol{L}^2(D)}^2 + \ell_D^2\|\nabla\boldsymbol{v}\|_{\mathbb{L}^2(D)}^2)^{\frac{1}{2}}$ for all $\boldsymbol{v} \in \boldsymbol{V}_{\mathrm{d}}$. A possible weak formulation of (42.7) is as follows:

$$\begin{cases} \text{Find } \boldsymbol{u} \in \boldsymbol{V}_{\mathrm{d}} \text{ such that} \\ a(\boldsymbol{u}, \boldsymbol{w}) = \ell(\boldsymbol{w}), \quad \forall\boldsymbol{w} \in \boldsymbol{V}_{\mathrm{d}}, \end{cases} \tag{42.10}$$

with the following bilinear and linear forms:

$$a(\boldsymbol{v}, \boldsymbol{w}) := \int_D \mathbb{s}(\boldsymbol{v}){:}\mathbb{e}(\boldsymbol{w})\,\mathrm{d}x = \int_D \big(2\mu\mathbb{e}(\boldsymbol{v}){:}\mathbb{e}(\boldsymbol{w}) + \lambda(\nabla{\cdot}\boldsymbol{v})(\nabla{\cdot}\boldsymbol{w})\big)\,\mathrm{d}x,$$

$$\ell(\boldsymbol{w}) := \int_D \boldsymbol{f}{\cdot}\boldsymbol{w}\,\mathrm{d}x + \int_{\partial D_{\mathrm{n}}} \boldsymbol{g}{\cdot}\boldsymbol{w}\,\mathrm{d}s.$$

In the language of continuum mechanics, the test function $\boldsymbol{w}$ plays the role of a virtual displacement, and the weak formulation (42.10) expresses the *principle of virtual work*. Moreover, recalling that $\|\mathbb{e}(\boldsymbol{v})\|_{\ell^2}^2 = \mathbb{e}(\boldsymbol{v}):\mathbb{e}(\boldsymbol{v}) = \sum_{i,j\in\{1:d\}}|e(\boldsymbol{v})_{ij}|^2$, the quantity $\mathfrak{E}(\boldsymbol{v}) := \frac{1}{2}a(\boldsymbol{v},\boldsymbol{v}) - \ell(\boldsymbol{v})$, i.e.,

$$\mathfrak{E}(\boldsymbol{v}) := \frac{1}{2}\int_D \left(2\mu\|\mathbb{e}(\boldsymbol{v})\|_{\ell^2}^2 + \lambda|\nabla\cdot\boldsymbol{v}|^2\right)\mathrm{d}x - \int_D \boldsymbol{f}\cdot\boldsymbol{v}\,\mathrm{d}x - \int_{\partial D_\mathrm{n}}\boldsymbol{g}\cdot\boldsymbol{v}\,\mathrm{d}s, \qquad (42.11)$$

represents the total energy of the deformed medium at equilibrium. The quadratic terms correspond to the energy of the elastic deformation, and the linear terms represent the potential energy associated with the volume and boundary loads. Note that $\mathfrak{E}(\boldsymbol{v})$ is not bounded from below over the whole space $\boldsymbol{H}^1(D)$ since $a(\boldsymbol{r},\boldsymbol{r}) = 0$ for all $\boldsymbol{r}\in\boldsymbol{R}$ and $\ell(\boldsymbol{r})$ may be arbitrarily large for some rigid displacements. Proceeding as in §31.3.3 for scalar elliptic PDEs leads to the following result.

Proposition 42.8 (Weak solution). *Let D be a Lipschitz domain in $\mathbb{R}^3$ with $\partial D = \partial D_\mathrm{d} \cup \partial D_\mathrm{n}$. Let $\boldsymbol{f}\in\boldsymbol{L}^2(D)$ and $\boldsymbol{g}\in\boldsymbol{L}^2(\partial D_\mathrm{n})$. If the function $\boldsymbol{u}\in\boldsymbol{V}_\mathrm{d}$ solves (42.10), then it satisfies (42.7a)-(42.7b) a.e. in D, (42.7c) a.e. on ∂D_d, and (42.7d) a.e. on ∂D_n in the sense that $\langle \mathbb{s}(\boldsymbol{u})\boldsymbol{n}, \widetilde{\boldsymbol{v}}\rangle_{\boldsymbol{H}^{-\frac{1}{2}},\boldsymbol{H}^{\frac{1}{2}}} = \int_{\partial D_\mathrm{n}}\boldsymbol{g}\cdot\boldsymbol{v}\,\mathrm{d}s$ for all $\boldsymbol{v}\in\widetilde{\boldsymbol{H}}^{\frac{1}{2}}(\partial D_\mathrm{n}) := \{\boldsymbol{v}\in\boldsymbol{H}^{\frac{1}{2}}(\partial D_\mathrm{n}) \mid \widetilde{\boldsymbol{v}}\in\boldsymbol{H}^{\frac{1}{2}}(\partial D)\}$, where $\widetilde{\boldsymbol{v}}$ is the zero extension of $\boldsymbol{v}$ to ∂D.*

42.2.2 Korn's inequalities and well-posedness

There are two Korn's inequalities. These inequalities will be invoked to establish the coercivity of the bilinear form a. The first one deals with the simpler situation where the displacement field vanishes on the whole boundary. The second one does not say anything on the boundary values. To unify the notation, we use the same symbol C_K to denote the constant associated with the first and the second Korn inequality.

Theorem 42.9 (Korn's first inequality). *Let D be a Lipschitz domain in $\mathbb{R}^d$. Setting $C_\mathrm{K} := \frac{1}{\sqrt{2}}$, the following holds true:*

$$C_\mathrm{K}\|\nabla\boldsymbol{v}\|_{\mathbb{L}^2(D)} \le \|\mathbb{e}(\boldsymbol{v})\|_{\mathbb{L}^2(D)}, \qquad \forall\boldsymbol{v}\in\boldsymbol{H}_0^1(D). \qquad (42.12)$$

Proof. Let $\boldsymbol{v}\in\boldsymbol{C}_0^\infty(D)$. Since $\boldsymbol{v}$ vanishes at the boundary, we have

$$\int_D \nabla\boldsymbol{v}:\nabla\boldsymbol{v}^\mathsf{T}\,\mathrm{d}x = \sum_{i,j\in\{1:d\}}\int_D (\partial_i v_j)(\partial_j v_i)\,\mathrm{d}x = -\sum_{i,j\in\{1:d\}}\int_D (\partial_{ij}^2 v_j)v_i\,\mathrm{d}x$$

$$= \sum_{i,j\in\{1:d\}}\int_D (\partial_i v_i)(\partial_j v_j)\,\mathrm{d}x = \int_D (\nabla\cdot\boldsymbol{v})^2\,\mathrm{d}x \ge 0.$$

A density argument then shows that the above inequality holds true for all $\boldsymbol{v}\in\boldsymbol{H}_0^1(D)$. As a result, we infer that for all $\boldsymbol{v}\in\boldsymbol{H}_0^1(D)$,

$$\int_D \mathbb{e}(\boldsymbol{v}):\mathbb{e}(\boldsymbol{v})\,\mathrm{d}x = \frac{1}{4}\int_D (\nabla\boldsymbol{v}+\nabla\boldsymbol{v}^\mathsf{T}):(\nabla\boldsymbol{v}+\nabla\boldsymbol{v}^\mathsf{T})\,\mathrm{d}x$$

$$= \frac{1}{2}\int_D \nabla\boldsymbol{v}:\nabla\boldsymbol{v}\,\mathrm{d}x + \frac{1}{2}\int_D \nabla\boldsymbol{v}:\nabla\boldsymbol{v}^\mathsf{T}\,\mathrm{d}x$$

$$\ge \frac{1}{2}\int_D \nabla\boldsymbol{v}:\nabla\boldsymbol{v}\,\mathrm{d}x = \frac{1}{2}\|\nabla\boldsymbol{v}\|_{\mathbb{L}^2(D)}^2. \qquad \square$$

Theorem 42.10 (Korn's second inequality). *Let D be a Lipschitz domain in $\mathbb{R}^d$. There is $C_{\mathrm{K}} > 0$ s.t.*

$$C_{\mathrm{K}}\|\nabla \boldsymbol{v}\|_{\mathbb{L}^2(D)} \le \|\mathbb{e}(\boldsymbol{v})\|_{\mathbb{L}^2(D)} + \ell_D^{-1}\|\boldsymbol{v}\|_{\boldsymbol{L}^2(D)}, \qquad \forall \boldsymbol{v} \in \boldsymbol{H}^1(D). \qquad (42.13)$$

Moreover, for every closed subspace $\boldsymbol{V}$ of $\boldsymbol{H}^1(D)$ s.t. $\boldsymbol{V} \cap \boldsymbol{R} = \{\boldsymbol{0}\}$, there is $C_{\mathrm{K}} > 0$ s.t.

$$C_{\mathrm{K}}\|\nabla \boldsymbol{v}\|_{\mathbb{L}^2(D)} \le \|\mathbb{e}(\boldsymbol{v})\|_{\mathbb{L}^2(D)}, \qquad \forall \boldsymbol{v} \in \boldsymbol{V}. \qquad (42.14)$$

Proof. For (42.13), see Ciarlet [123, p. 11], Duvaut and Lions [183, p. 110], McLean [298, Thm. 10.2]. The inequality (42.14) results from (42.13) and the Peetre–Tartar lemma (Lemma A.20). Let us define $\boldsymbol{X} := \boldsymbol{V}$, $\mathbb{Y} := \mathbb{L}^2(D)$, $A : \boldsymbol{X} \to \mathbb{Y}$ with $A(\boldsymbol{v}) := \mathbb{e}(\boldsymbol{v})$, $\boldsymbol{Z} := \boldsymbol{L}^2(D)$, and let T be the compact injection from $\boldsymbol{X}$ into $\boldsymbol{Z}$. Lemma 42.7 implies that $\ker(A) \subset \boldsymbol{R}$. But $\boldsymbol{V} \cap \boldsymbol{R} = \{\boldsymbol{0}\}$, so that $\ker(A) = \{\boldsymbol{0}\}$, i.e., A is injective. Moreover, (42.13) implies that there is $c > 0$ s.t. $\|\boldsymbol{v}\|_{\boldsymbol{H}^1(D)} \le c(\ell_D\|A(\boldsymbol{v})\|_{\mathbb{L}^2(D)} + \|T(\boldsymbol{v})\|_{\boldsymbol{L}^2(D)})$. Then the Peetre–Tartar lemma asserts that there is $c > 0$ s.t. $\|\boldsymbol{v}\|_{\boldsymbol{H}^1(D)} \le c\ell_D\|A(\boldsymbol{v})\|_{\mathbb{L}^2(D)}$ for all $\boldsymbol{v} \in \boldsymbol{X}$. Since $\ell_D\|\nabla \boldsymbol{v}\|_{\mathbb{L}^2(D)} \le \|\boldsymbol{v}\|_{\boldsymbol{H}^1(D)}$, this proves (42.14). $\square$

Theorem 42.11 (Well-posedness). *Let D be a Lipschitz domain in $\mathbb{R}^3$ with $\partial D = \partial D_{\mathrm{d}} \cup \partial D_{\mathrm{n}}$. Assume that $|\partial D_{\mathrm{d}}| > 0$. Let $\boldsymbol{f} \in \boldsymbol{L}^2(D)$ and, if $|\partial D_{\mathrm{n}}| > 0$, let $\boldsymbol{g} \in \boldsymbol{L}^2(\partial D_{\mathrm{n}})$. Let λ, μ satisfy (42.4). (i) The problem (42.10) is well-posed. (ii) (42.10) is equivalent to the variational formulation $\boldsymbol{u} = \arg\min_{\boldsymbol{v} \in \boldsymbol{V}_{\mathrm{d}}} \mathfrak{E}(\boldsymbol{v})$ with the energy functional $\mathfrak{E}$ defined in (42.11).*

Proof. (i) We apply the Lax–Milgram lemma. The linear form ℓ is bounded on $\boldsymbol{H}^1(D)$, and the boundedness of the bilinear form a on $\boldsymbol{H}^1(D) \times \boldsymbol{H}^1(D)$ is a consequence of the assumption $\lambda, \mu \in L^\infty(D)$. Let us verify the coercivity of a on $\boldsymbol{V}_{\mathrm{d}}$. If $\lambda \ge 0$, we have $2\mu\|\mathbb{e}(\boldsymbol{v})\|_{\ell^2}^2 + \lambda|\nabla\cdot\boldsymbol{v}|^2 \ge 2\mu\|\mathbb{e}(\boldsymbol{v})\|_{\ell^2}^2$. If $\lambda < 0$, we use the inequality $\|\mathbb{e}(\boldsymbol{v})\|_{\ell^2}^2 \ge \frac{1}{3}(\nabla\cdot\boldsymbol{v})^2$ (which follows from $\|\mathbb{e}(\boldsymbol{v})\|_{\ell^2}^2 = \sum_{i,j}|e(\boldsymbol{v})_{ij}|^2 \ge \sum_i |e(\boldsymbol{v})_{ii}|^2 = \sum_i |\partial_i v_i|^2 \ge \frac{1}{3}(\nabla\cdot\boldsymbol{v})^2$) to infer that $2\mu\|\mathbb{e}(\boldsymbol{v})\|_{\ell^2}^2 + \lambda|\nabla\cdot\boldsymbol{v}|^2 \ge 3\kappa\|\mathbb{e}(\boldsymbol{v})\|_{\ell^2}^2$. Recalling the assumption (42.4), this shows that in all the cases we obtain

$$2\mu\|\mathbb{e}(\boldsymbol{v})\|_{\ell^2}^2 + \lambda|\nabla\cdot\boldsymbol{v}|^2 \ge \rho_{\min}\|\mathbb{e}(\boldsymbol{v})\|_{\ell^2}^2,$$

with $\rho_{\min} := 2\mu_{\min}$ if $\lambda \ge 0$ a.e. in D and $\rho_{\min} := \min(2\mu_{\min}, 3\kappa_{\min})$ otherwise. Notice that we have $\rho_{\min} > 0$. The above bounds imply that

$$a(\boldsymbol{v}, \boldsymbol{v}) = \int_D (2\mu\|\mathbb{e}(\boldsymbol{v})\|_{\ell^2}^2 + \lambda|\nabla\cdot\boldsymbol{v}|^2)\,\mathrm{d}x$$

$$\ge \rho_{\min} \int_D \|\mathbb{e}(\boldsymbol{v})\|_{\ell^2}^2\,\mathrm{d}x = \rho_{\min}\|\mathbb{e}(\boldsymbol{v})\|_{\mathbb{L}^2(D)}^2, \qquad \forall \boldsymbol{v} \in \boldsymbol{V}_{\mathrm{d}}.$$

If $\partial D_{\mathrm{d}} = \partial D$, then we have $\boldsymbol{V}_{\mathrm{d}} := \boldsymbol{H}_0^1(D)$, and we invoke Korn's first inequality (see (42.12)). Otherwise, we invoke Korn's second inequality (see (42.14)) since $|\partial D_{\mathrm{d}}| > 0$ implies that $\boldsymbol{V}_{\mathrm{d}} \cap \boldsymbol{R} = \{\boldsymbol{0}\}$. In both cases, there is $C_{\mathrm{K}} > 0$ s.t.

$$a(\boldsymbol{v}, \boldsymbol{v}) \ge \rho_{\min}C_{\mathrm{K}}^2\|\nabla \boldsymbol{v}\|_{\mathbb{L}^2(D)}^2 = \rho_{\min}C_{\mathrm{K}}^2\|\boldsymbol{v}\|_{\boldsymbol{V}_{\mathrm{d}}}^2, \qquad \forall \boldsymbol{v} \in \boldsymbol{V}_{\mathrm{d}}, \qquad (42.15)$$

i.e., a is $\boldsymbol{V}_{\mathrm{d}}$-coercive. This shows that the problem (42.10) is well-posed.
(ii) The equivalence of the problem (42.10) and the variational formulation minimizing the energy functional $\mathfrak{E}$ follows from Proposition 25.8, since the bilinear form a is symmetric. $\square$

Remark 42.12 (Regularity pickup). The elliptic regularity theory applies to the linear elasticity equations with smooth Lamé parameters. In particular, if ∂D is smooth and either the homogeneous Dirichlet or Neumann boundary condition is applied, there exist $s \in (\frac{1}{2}, 1]$ and c_{smo}

s.t. $\|\boldsymbol{u}\|_{\boldsymbol{H}^{1+s}(D)} \leq c_{\text{smo}}\mu_{\min}^{-1}\ell_D^2\|\boldsymbol{f}\|_{\boldsymbol{L}^2(D)}$ for all $\boldsymbol{f} \in \boldsymbol{L}^2(D)$; see e.g., Mazzucato and Nistor [297, §10]. This property is not valid for mixed Dirichlet–Neumann boundary conditions, as already illustrated in Exercise 31.5 for scalar elliptic PDEs. For nonsmooth domains and mixed Dirichlet–Neumann boundary conditions, it is nevertheless possible to show that there exists s (usually in $(0, \frac{1}{2})$) such that $\|\boldsymbol{u}\|_{\boldsymbol{H}^{1+s}(D)} \leq c_{\text{smo}}\mu_{\min}^{-1}\ell_D^2\|\boldsymbol{f}\|_{\boldsymbol{L}^2(D)}$. $\square$

Remark 42.13 (Pure traction). The pure traction (or Neumann) problem is obtained when $\partial D_{\text{n}} = \partial D$ and $\partial D_{\text{d}} = \emptyset$ in (42.7). In this case, the data must satisfy the compatibility condition $\int_D \boldsymbol{f}\cdot\boldsymbol{r}\,dx + \int_{\partial D} \boldsymbol{g}\cdot\boldsymbol{r}\,ds = 0$ for all $\boldsymbol{r} \in \boldsymbol{R}$. Uniqueness of the weak solution is obtained by additionally prescribing, e.g., that $\int_D \boldsymbol{v}\,dx := \boldsymbol{0}$ and $\int_D \nabla\times\boldsymbol{v}\,dx := \boldsymbol{0}$; see Exercise 42.3. $\square$

Remark 42.14 (Elasticity functionals). The weak formulation (42.10) is posed in terms of one dependent variable: the displacement field. The strain and the stress fields are evaluated from the displacement field by means of (42.2) and (42.3). As shown in Theorem 42.11, the weak solution is the minimizer of the energy functional $\mathfrak{E}$ defined in (42.11). It is possible to characterize the solution of the linear elasticity equations by an optimality condition using other functionals. One important example consists of using both the stress and the displacement fields as independent variables and to look for a critical point of the *Hellinger–Reissner functional* (see [241, 333] and Exercise 42.1). This approach constitutes the basis of the mixed stress-displacement finite element methods in elasticity (see §42.4.2). Among other possible approaches are the three-field formulation that consists of treating the displacement, the strain tensor, and the stress tensor as independent variables (see Fraejis de Veubeke [205], Hu [248], Washizu [388]) and the intrinsic formulation where the only dependent variable is the strain tensor (see Ciarlet and Ciarlet [125]). $\square$

42.3 H^1-conforming approximation

Let D be a polyhedron in $\mathbb{R}^3$. Let $(\mathcal{T}_h)_{h\in\mathcal{H}}$ be a shape-regular sequence of affine matching meshes so that each mesh covers D exactly. For simplicity, we assume that the material is clamped at ∂D, i.e., $\partial D_{\text{d}} = \partial D$ so that $\boldsymbol{V}_{\text{d}} := \boldsymbol{H}_0^1(D)$. Let $P_k^{\text{g}}(\mathcal{T}_h) := \{v_h \in C^0(\overline{D}) \mid v_{h|K} \circ \boldsymbol{T}_K \in \widehat{P}, \forall K \in \mathcal{T}_h\}$ be the scalar-valued H^1-conforming finite element space constructed in §19.2.1, where $k \geq 1$ is the degree of the reference finite element $(\widehat{K}, \widehat{P}, \widehat{\Sigma})$. Let $P_{k,0}^{\text{g}}(\mathcal{T}_h) := P_k^{\text{g}}(\mathcal{T}_h) \cap H_0^1(D)$ be the corresponding $H_0^1(D)$-conforming subspace. We define the $\boldsymbol{H}_0^1$-conforming approximation space

$$\boldsymbol{V}_{h0} := P_{k,0}^{\text{g}}(\mathcal{T}_h) \times P_{k,0}^{\text{g}}(\mathcal{T}_h) \times P_{k,0}^{\text{g}}(\mathcal{T}_h), \tag{42.16}$$

and consider the following discrete problem:

$$\begin{cases} \text{Find } \boldsymbol{u}_h \in \boldsymbol{V}_{h0} \text{ such that} \\ a(\boldsymbol{u}_h, \boldsymbol{w}_h) = \ell(\boldsymbol{w}_h), \quad \forall \boldsymbol{w}_h \in \boldsymbol{V}_{h0}. \end{cases} \tag{42.17}$$

Since a is coercive and the approximation setting is conforming, i.e., $\boldsymbol{V}_{h0} \subset \boldsymbol{V}_{\text{d}}$, the discrete problem (42.17) is well-posed. Recalling the energy functional $\mathfrak{E}$ defined in (42.11), we have $\boldsymbol{u}_h = \arg\min_{\boldsymbol{v}\in\boldsymbol{V}_{h0}} \mathfrak{E}(\boldsymbol{v}_h)$ and $\mathfrak{E}(\boldsymbol{u}_h) \geq \mathfrak{E}(\boldsymbol{u})$ owing to the conformity of the approximation setting.

Remark 42.15 (Collocation). Let $\{\varphi_a\}_{a\in\mathcal{A}_h}$ be the global shape functions of $P_{k,0}^{\text{g}}(\mathcal{T}_h)$, and let $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\}$ be the canonical basis of $\mathbb{R}^3$. Then one can use $\{\boldsymbol{e}_i\varphi_a\}_{a\in\mathcal{A}_h, i\in\{1:3\}}$ as the global shape functions of $\boldsymbol{V}_{h0}$. With this choice, (42.17) leads to a collocalized scheme since the three components of the discrete displacement field $\boldsymbol{u}_h$ are associated with the same scalar-valued global shape function. $\square$

Theorem 42.16 (Error estimate). *Let the assumptions of Theorem* 42.11 *hold true. Let* $\boldsymbol{u}$ *solve* (42.10) *and let* $\boldsymbol{u}_h$ *solve* (42.17). (i) *There is c s.t.*

$$|\boldsymbol{u} - \boldsymbol{u}_h|_{\boldsymbol{H}^1(D)} \leq c \inf_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}} |\boldsymbol{u} - \boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}, \tag{42.18}$$

for all $h \in \mathcal{H}$ *and* $\lim_{h \to 0} |\boldsymbol{u} - \boldsymbol{u}_h|_{\boldsymbol{H}^1(D)} = 0$. (ii) *If* $\boldsymbol{u} \in \boldsymbol{H}^{1+r}(D)$ *for some* $r \in (0, k]$, *the following holds true:*

$$|\boldsymbol{u} - \boldsymbol{u}_h|_{\boldsymbol{H}^1(D)} \leq c \left(\sum_{K \in \mathcal{T}_h} h_K^{2r} |\boldsymbol{u}|_{\boldsymbol{H}^{1+r}(K)}^2 \right)^{\frac{1}{2}} \leq c \, h^r |\boldsymbol{u}|_{\boldsymbol{H}^{1+r}(D)}. \tag{42.19}$$

(iii) *Letting s be the index of the elliptic regularity pickup, we have*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{L}^2(D)} \leq c \, h^s \ell_D^{1-s} |\boldsymbol{u} - \boldsymbol{u}_h|_{\boldsymbol{H}^1(D)}. \tag{42.20}$$

Proof. The proof goes along the same lines as those presented in Chapter 32 for scalar elliptic PDEs, i.e., (42.18) follows from Céa's lemma (Lemma 26.13), (42.19) from the approximation properties of the quasi-interpolation operator with zero trace from §22.4 (and the regularity of the mesh sequence), and (42.20) from the Aubin–Nitsche lemma (Lemma 32.11). □

A shortcoming of low-order H^1-conforming finite elements is their poor performance when approximating nearly-incompressible materials. This phenomenon is known in the literature as *volume* or *dilatation locking*. Other types of locking can occur in linear elasticity problems, such as *shear locking* in plate models when the plate thickness is very small. For simplicity, we focus on volume locking and on how to avoid it. Nearly-incompressible materials are characterized by the fact that the ratio $\frac{\lambda}{\mu}$ of the Lamé parameters is very large (or equivalently the Poisson coefficient ν is very close to $\frac{1}{2}$, see (42.5)). In this situation, the displacement field is nearly divergence-free.

It has long been known that the H^1-conforming approximation of nearly-incompressible materials on triangular meshes may not behave properly on meshes that are not fine enough if $k = 1$. Moreover, the method converges sub-optimally for $k \in \{2, 3\}$ and delivers optimal-order convergence for $k \geq 4$. On quadrilateral meshes, volume locking cannot be avoided for all $k \geq 1$. We refer the reader, e.g., to Vogelius [380], Scott and Vogelius [345], Babuška and Suri [40]. To understand why H^1-conforming finite elements may fail, let us inspect how the error estimate (42.18) depends on the Lamé parameters μ and λ. To simplify the discussion, we assume that these parameters are constant, and since we are concerned with the case $\frac{\lambda}{\mu} \gg 1$, we assume that λ is nonnegative. We first observe that the bilinear form a is $\boldsymbol{H}_0^1$-coercive with coercivity constant being proportional to μ since Korn's first inequality (see (42.12)) and $\lambda \geq 0$ imply that

$$a(\boldsymbol{v}, \boldsymbol{v}) \geq 2\mu \|\mathbb{e}(\boldsymbol{v})\|_{\mathbb{L}^2(D)}^2 \geq \mu |\boldsymbol{v}|_{\boldsymbol{H}^1(D)}^2.$$

Moreover, since $\|\mathbb{e}(\boldsymbol{v})\|_{\mathbb{L}^2(D)} \leq |\boldsymbol{v}|_{\boldsymbol{H}^1(D)}$, the Cauchy–Schwarz inequality implies the following boundedness property:

$$a(\boldsymbol{v}, \boldsymbol{w}) \leq \mu |\boldsymbol{v}|_{\boldsymbol{H}^1(D)} |\boldsymbol{w}|_{\boldsymbol{H}^1(D)} + \lambda \|\nabla \cdot \boldsymbol{v}\|_{L^2(D)} \|\nabla \cdot \boldsymbol{w}\|_{L^2(D)}.$$

Following the proof of Céa's lemma, we infer that

$$|\boldsymbol{u} - \boldsymbol{u}_h|_{\boldsymbol{H}^1(D)} \leq c \inf_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}} \left(|\boldsymbol{u} - \boldsymbol{v}_h|_{\boldsymbol{H}^1(D)} + \frac{\lambda}{\mu} \|\nabla \cdot (\boldsymbol{u} - \boldsymbol{v}_h)\|_{L^2(D)} \right), \tag{42.21}$$

where the punchline is that, in contrast to (42.18), this error estimate features a constant c that is uniform w.r.t. μ and λ. The first term on the right-hand side decays as the best-approximation error

of $\boldsymbol{u}$ in $\boldsymbol{V}_{h0}$. This is also the case for the second term (since $\|\nabla{\cdot}(\boldsymbol{u}-\boldsymbol{v}_h)\|_{L^2(D)} \leq \sqrt{3}|\boldsymbol{u}-\boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}$), but the scaling by the multiplicative factor $\frac{\lambda}{\mu} \gg 1$ causes this term to be very large on practically feasible meshes. In other words, the second term tends to zero as $h \to 0$, but this asymptotic range is only visible on meshes that are refined enough to beat the large constant $\frac{\lambda}{\mu}$.

There are several possibilities to circumvent this bottleneck and to devise approximation methods that are *robust* w.r.t. volume locking. The first route consists of introducing the auxiliary variable $p := \lambda\nabla{\cdot}\boldsymbol{u}$ which plays the role of the pressure in the incompressible limit. The idea is then to devise a mixed finite element approximation for the pair $(\boldsymbol{u}, p)$. This approach requires some care in choosing the finite element spaces to approximate the displacement and the pressure, and is analyzed in Chapters 53 and 54 in the context of the Stokes equations. One can also consider mixed finite element methods that approximate both the stress and the displacement fields; see §42.4.2. Another route consists of using a nonconforming approximation for the displacement in such a way that the error on the approximation of $\nabla{\cdot}\boldsymbol{u}$ only depends on the smoothness of $\nabla{\cdot}\boldsymbol{u}$. Examples include the nonconforming finite element methods in Fortin and Soulié [204], Fortin [202], Falk [199], Brenner and Sung [88], the discontinuous Galerkin methods in Hansbo and Larson [239, 240], Wihler [395], Cockburn et al. [133], the hybridizable discontinuous Galerkin methods in Soon et al. [351], Fu et al. [209], the discontinuous Petrov–Galerkin method in Carstensen and Hellwig [110], and the hybrid high-order method in Di Pietro and Ern [166] that we briefly present in §42.4.3.

42.4   Further topics

This section briefly reviews some other discretization techniques to approximate the model problem (42.10): Crouzeix–Raviart elements, mixed finite elements, and hybrid high-order (HHO) methods.

42.4.1   Crouzeix–Raviart approximation

Let $\boldsymbol{P}_1^{\mathrm{CR}}(\mathcal{T}_h) := P_1^{\mathrm{CR}}(\mathcal{T}_h; \mathbb{R}^3)$ be the vector-valued Crouzeix–Raviart finite element space (see Chapter 36 for the scalar-valued space $P_1^{\mathrm{CR}}(\mathcal{T}_h) := P_1^{\mathrm{CR}}(\mathcal{T}_h; \mathbb{R})$). Using $\boldsymbol{P}_1^{\mathrm{CR}}(\mathcal{T}_h)$ to approximate the components of the displacement field leads to the desirable property on the approximation of the divergence since $\nabla{\cdot}(\boldsymbol{\mathcal{I}}_K^{\mathrm{CR}}(\boldsymbol{u})) = \Pi_K^0(\nabla{\cdot}\boldsymbol{u})$ for all $K \in \mathcal{T}_h$, where Π_K^0 is the L^2-orthogonal projection onto constants in K, i.e., $\nabla{\cdot}(\boldsymbol{\mathcal{I}}_K^{\mathrm{CR}}(\boldsymbol{u}))$ is equal to the mean value of $\nabla{\cdot}\boldsymbol{u}$ in K (see Exercise 36.1). Unfortunately, the Crouzeix–Raviart finite element fails to satisfy the broken version of Korn's inequality, since it is possible to find nonzero discrete fields $\boldsymbol{v}_h \in \boldsymbol{P}_1^{\mathrm{CR}}(\mathcal{T}_h)$ such that locally in each mesh cell $K \in \mathcal{T}_h$, $\mathbb{e}(\boldsymbol{v}_{h|K})$ vanishes identically on K. This is a striking difference with the scalar-valued case where a broken version of the Poincaré–Steklov inequality holds true (see Lemma 36.6). For pure traction (Neumann) boundary conditions, the failure to satisfy a discrete Korn inequality can be shown by the following dimension argument; see [199]. Let $\boldsymbol{P}_{1,*}^{\mathrm{CR}}(\mathcal{T}_h) := \{\boldsymbol{v}_h \in \boldsymbol{P}_1^{\mathrm{CR}}(\mathcal{T}_h) \mid \int_D \boldsymbol{v}_h \,\mathrm{d}x = \boldsymbol{0}, \sum_{K\in\mathcal{T}_h} \int_K \nabla{\times}\boldsymbol{v}_h \,\mathrm{d}x = \boldsymbol{0}\}$, where the two integral conditions (altogether, six scalar conditions) are meant to remove global rigid-body motions from the space (see Remark 42.13). Let N_{c}, N_{f}, and $N_{\mathrm{f}}^{\partial}$ denote the number of cells, faces, and boundary faces in the mesh. Observe that $4N_{\mathrm{c}} = 2N_{\mathrm{f}} - N_{\mathrm{f}}^{\partial}$ (indeed, separating all the mesh cells, we obtain $4N_{\mathrm{c}}$ faces, and this number is equal to $2N_{\mathrm{f}} - N_{\mathrm{f}}^{\partial}$ since there are two faces contributing to each interface but one face contributing to each boundary face). Let $\mathbb{e}_h : \boldsymbol{P}_{1,*}^{\mathrm{CR}}(\mathcal{T}_h) \to \mathbb{L}^2(D)$ be s.t. $\mathbb{e}_h(\boldsymbol{v}_h)_{|K} = \mathbb{e}(\boldsymbol{v}_{h|K})$. Since $\mathbb{e}_h(\boldsymbol{v}_h)$ is piecewise constant on $\mathcal{T}_h$ and takes symmetric values in

$\mathbb{R}^{3\times3}$, we have $\dim(\mathrm{im}(\mathbb{e}_h)) \leq 6N_\mathrm{c}$. We infer that

$$\dim(\ker(\mathbb{e}_h)) = \dim(\boldsymbol{P}_{1,*}^{\mathrm{CR}}(\mathcal{T}_h)) - \dim(\mathrm{im}(\mathbb{e}_h))$$

$$= 3N_\mathrm{f} - 6 - \dim(\mathrm{im}(\mathbb{e}_h)) \geq 3N_\mathrm{f} - 6 - 6N_\mathrm{c} = \frac{3}{2}(N_\mathrm{f}^\partial - 4),$$

which is positive as soon as the mesh is composed of more than one cell. The discrete Korn inequality can also fail on some meshes when enforcing pure displacement (Dirichlet) boundary conditions as shown in Figure 42.1. Here, $D := (-1,1)^2$. Let $\boldsymbol{z}_0 := \boldsymbol{0}$, $\boldsymbol{z}_5 = \boldsymbol{z}_1 := (1,1)$, $\boldsymbol{z}_2 := (-1,1)$, $\boldsymbol{z}_3 := (-1,-1)$, $\boldsymbol{z}_4 := (1,-1)$, and let K_i be the triangle with vertices $\boldsymbol{z}_0, \boldsymbol{z}_i, \boldsymbol{z}_{i+1}$ for all $i \in \{1{:}4\}$. Consider the mesh $\mathcal{T}_h := \bigcup_{i\in\{1:4\}} K_i$. The vector field shown in Figure 42.1 is piecewise linear and defined by $\boldsymbol{v}_{h|K_1} := -2(y,-x) + (0,-2)$, $\boldsymbol{v}_{h|K_2} := 2(y,-x) + (-2,0)$, $\boldsymbol{v}_{h|K_3} := -2(y,-x) + (0,2)$, and $\boldsymbol{v}_{h|K_4} := 2(y,-x) + (2,0)$. One readily verifies that $\boldsymbol{v}_h$ is in $\boldsymbol{P}_{1,*}^{\mathrm{CR}}(\mathcal{T}_h)$ and is such that $\int_F \boldsymbol{v}_h \,\mathrm{d}s = \boldsymbol{0}$ for all $F \in \mathcal{F}_h$, but $\mathbb{e}(\boldsymbol{v}_h)_{|K_i} = \mathbb{O}$ for all $i \in \{1{:}4\}$.
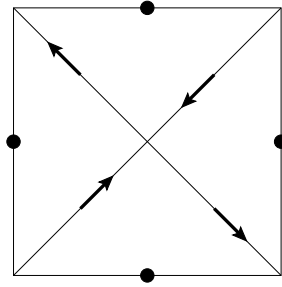
Figure 42.1: Failure of the discrete Korn inequality using Crouzeix–Raviart displacements on a mesh composed of four cells. The bullets symbolize zero displacement at the midpoint of the four boundary edges, and the arrows show the displacement at the midpoint of the four internal edges.

42.4.2 Mixed finite elements

The idea in mixed finite element methods for linear elasticity is to approximate both the stress and the displacement fields. Besides robustness w.r.t. volume locking, mixed finite element methods ensure a direct approximation of the equilibrium condition (42.1), and the discrete strain can be recovered locally from the discrete stress by inverting the constitutive relation (42.3). However, the relation between displacement and stress, i.e., (42.2), is less direct (i.e., it is only obtained in a weak form). In contrast, using the displacement-based formulation (42.17) ensures that (42.2) is satisfied locally (i.e., one can define the discrete strain as $\mathbb{e}(\boldsymbol{u}_h)$), but the equilibrium condition (42.1) and the constitutive relation (42.3) are only satisfied in a weak sense. One difficulty with mixed finite element methods for elasticity is the devising of discrete spaces with symmetric stresses. The idea of relaxing this symmetry constraint by means of an auxiliary variable (that can be interpreted as a rotation) was originally proposed in Fraejis de Veubeke [206] and was further developed and analyzed in Amara and Thomas [8], Arnold et al. [19], Stenberg [353, 355], Morley [307]; see also the more recent and comprehensive presentation in Arnold et al. [24], Boffi et al. [64]. Mixed finite elements with symmetric stresses have been proposed in Arnold and Winther [17] in dimension two and extended to dimension three in Arnold et al. [25], but the number of local degrees of freedom is fairly substantial.

42.4.3 Hybrid high-order (HHO) approximation

We refer the reader to Chapter 39 for a detailed presentation and analysis of the HHO method when approximating a scalar-valued elliptic PDE. For simplicity, we consider here homogeneous Dirichlet conditions on the displacement, and we suppose that the Lamé coefficients take constant values.

Let D be a polyhedron in $\mathbb{R}^d$ and $(\mathcal{T}_h)_{h\in\mathcal{H}}$ be a shape-regular sequence of affine meshes so that each mesh covers D exactly. To fix the ideas, we assume that $d = 3$. Let $k \geq 1$ be the polynomial degree. The local unknowns are $\mathbb{R}^d$-valued polynomials of degree at most k on the mesh cells and the mesh faces. For all $K \in \mathcal{T}_h$, we let $\hat{\boldsymbol{V}}_K^k := \boldsymbol{V}_K^k \times \boldsymbol{V}_{\partial K}^k$ with

$$\boldsymbol{V}_K^k := \mathbb{P}_{k,d} \circ \boldsymbol{T}_K^{-1}, \quad \boldsymbol{V}_{\partial K}^k := \prod_{F\in\mathcal{F}_K} \mathbb{P}_{k,d-1} \circ \boldsymbol{T}_F^{-1}, \tag{42.22}$$

where $\mathcal{F}_K$ is the collection of the faces of K and $\boldsymbol{T}_K : \widehat{S}^d \to K$ and $\boldsymbol{T}_F : \boldsymbol{S}^{d-1} \to F$ are affine geometric mappings defined on the reference simplices of $\mathbb{R}^d$ and $\mathbb{R}^{d-1}$, respectively. Pairs in $\hat{\boldsymbol{V}}_K^k$ are denoted by $\hat{\boldsymbol{v}}_K := (\boldsymbol{v}_K, \boldsymbol{v}_{\partial K})$.

There are three key ingredients to devise the HHO method for linear elasticity (the first and the third ones are similar to those introduced in §39.1): (i) a displacement reconstruction operator, (ii) a divergence reconstruction operator, and (iii) a stabilization operator. The displacement reconstruction operator $\mathsf{R} : \hat{\boldsymbol{V}}_K^k \to \boldsymbol{V}_K^{k+1} := \mathbb{P}_{k+1,d} \circ \boldsymbol{T}_K^{-1}$ is defined by solving the following local Neumann problem: For all $\hat{\boldsymbol{v}}_K \in \hat{\boldsymbol{V}}_K^k$, the $\mathbb{R}^d$-valued polynomial function $\boldsymbol{d} := \mathsf{R}(\hat{\boldsymbol{v}}_K) \in \boldsymbol{V}_K^{k+1}$ is s.t.

$$(\mathbb{e}(\boldsymbol{d}), \mathbb{e}(\boldsymbol{w}))_{\mathbb{L}^2(K)} := -(\boldsymbol{v}_K, \nabla\cdot\mathbb{e}(\boldsymbol{w}))_{\boldsymbol{L}^2(K)} + (\boldsymbol{v}_{\partial K}, \mathbb{e}(\boldsymbol{w})\boldsymbol{n}_K)_{\boldsymbol{L}^2(\partial K)}, \tag{42.23}$$

for all $\boldsymbol{w} \in \boldsymbol{V}_K^{k+1}$. To obtain a well-posed problem, we recall the space of rigid displacements $\boldsymbol{R} := \mathbb{N}_{0,d}$ and Lemma 42.7. Let $\boldsymbol{R}_K := (\boldsymbol{\psi}_K^{\mathrm{c}})^{-1}(\boldsymbol{R})$, where $\boldsymbol{\psi}_K^{\mathrm{c}}$ is the covariant Piola transformation (see (9.9b)), and observe that $\boldsymbol{R}_K = \boldsymbol{R}$ since the geometric mapping is affine. Then $\boldsymbol{d} \in \boldsymbol{V}_K^{k+1}$ is uniquely defined by prescribing $\int_K \boldsymbol{d}\,\mathrm{d}x := \int_K \boldsymbol{v}_K\,\mathrm{d}x$ and $\int_K \nabla\times\boldsymbol{d}\,\mathrm{d}x := \int_{\partial K} \boldsymbol{n}_K\times\boldsymbol{v}_{\partial K}\,\mathrm{d}s$ (indeed, if $\boldsymbol{r} \in \boldsymbol{R}_K$ is s.t. $\int_K \boldsymbol{r}\,\mathrm{d}x = \boldsymbol{0}$ and $\int_K \nabla\times\boldsymbol{r}\,\mathrm{d}x = \boldsymbol{0}$, then $\boldsymbol{r} = \boldsymbol{0}$; see Exercise 42.3). Furthermore, the divergence reconstruction operator $\mathsf{D} : \hat{\boldsymbol{V}}_K^k \to V_K^k := \mathbb{P}_{k,d} \circ \boldsymbol{T}_K^{-1}$ is defined by solving the following well-posed problem:

$$(\mathsf{D}(\hat{\boldsymbol{v}}_K), q)_{L^2(K)} := -(\boldsymbol{v}_K, \nabla q)_{\boldsymbol{L}^2(K)} + (\boldsymbol{v}_{\partial K}, q\boldsymbol{n}_K)_{\boldsymbol{L}^2(K)}, \tag{42.24}$$

for all $\hat{\boldsymbol{v}}_K \in \hat{\boldsymbol{V}}_K^k$ and all $q \in V_K^k$. Recalling the definition (39.3), we observe that this operator satisfies the following important commuting property:

$$\mathsf{D}(\hat{\mathcal{I}}_K^k(\boldsymbol{v})) = \Pi_K^k(\nabla\cdot\boldsymbol{v}), \qquad \forall\boldsymbol{v} \in \boldsymbol{H}^1(K). \tag{42.25}$$

This property is the key argument to ensure robustness w.r.t. volume locking. Finally, the stabilization operator $\mathsf{S} : \hat{\boldsymbol{V}}_K^k \to \boldsymbol{V}_{\partial K}^k$ is defined as follows: For all $\hat{\boldsymbol{v}}_K \in \hat{\boldsymbol{V}}_K^k$, letting $\boldsymbol{\delta}_{\partial K} := \boldsymbol{v}_{K|\partial K} - \boldsymbol{v}_{\partial K}$,

$$\begin{aligned}\mathsf{S}(\hat{\boldsymbol{v}}_K) &:= \Pi_{\partial K}^k \left(\boldsymbol{v}_{K|\partial K} - \boldsymbol{v}_{\partial K} + ((I - \Pi_K^k)\mathsf{R}(\hat{\boldsymbol{v}}_K))_{|\partial K}\right) \\ &= \Pi_{\partial K}^k \left(\boldsymbol{\delta}_{\partial K} - ((I - \Pi_K^k)\mathsf{R}(\boldsymbol{0}, \boldsymbol{\delta}_{\partial K}))_{|\partial K}\right),\end{aligned} \tag{42.26}$$

where I is the identity, $\Pi_{\partial K}^k : \boldsymbol{L}^2(\partial K) \to \boldsymbol{V}_{\partial K}^k$ is the L^2-orthogonal projection onto $\boldsymbol{V}_{\partial K}^k$, and $\Pi_K^k : \boldsymbol{L}^2(K) \to \boldsymbol{V}_K^k$ is the L^2-orthogonal projection onto V_K^k. Let $\hat{\mathcal{I}}_K^k : \boldsymbol{H}^1(K) \to \hat{\boldsymbol{V}}_K^k$ be the local interpolation operator s.t. $\hat{\mathcal{I}}_K^k(\boldsymbol{v}) := (\Pi_K^k(\boldsymbol{v}), \Pi_{\partial K}^k(\boldsymbol{v}_{|\partial K}))$. Let $\boldsymbol{\mathcal{E}}_K : \boldsymbol{H}^1(K) \to \boldsymbol{V}_K^{k+1}$ be the local elliptic projection s.t. $(\mathbb{e}(\boldsymbol{\mathcal{E}}_K(\boldsymbol{v}) - \boldsymbol{v}), \mathbb{e}(\boldsymbol{w}))_{\mathbb{L}^2(K)} = 0$ for all $\boldsymbol{w} \in \boldsymbol{V}_K^{k+1}$, and $\int_K(\boldsymbol{\mathcal{E}}_K(\boldsymbol{v}) - \boldsymbol{v})\,\mathrm{d}x = \boldsymbol{0}$,

$\int_K \nabla \times (\mathcal{E}_K(\boldsymbol{v}) - \boldsymbol{v}) \, dx = \boldsymbol{0}$. As in Lemma 39.1 and Lemma 39.3, we have $\mathsf{R} \circ \hat{\mathcal{I}}_K^k = \mathcal{E}_K$ and there is c s.t.

$$h_K^{-\frac{1}{2}} \|\mathsf{S}(\hat{\mathcal{I}}_K^k(\boldsymbol{v}))\|_{\boldsymbol{L}^2(\partial K)} \le c \, |\boldsymbol{v} - \mathcal{E}_K(\boldsymbol{v})|_{\boldsymbol{H}^1(K)}, \tag{42.27}$$

for all $\boldsymbol{v} \in \boldsymbol{H}^1(K)$, all $K \in \mathcal{T}_h$, and all $h \in \mathcal{H}$.

For all $K \in \mathcal{T}_h$, we define the bilinear form on $\hat{\boldsymbol{V}}_K^k \times \hat{\boldsymbol{V}}_K^k$ such that

$$\hat{a}_K(\hat{\boldsymbol{v}}_K, \boldsymbol{w}_K) := 2\mu(\mathbb{e}(\mathsf{R}(\hat{\boldsymbol{v}}_K)), \mathbb{e}(\mathsf{R}(\hat{\boldsymbol{w}}_K)))_{\mathbb{L}^2(K)}$$
$$+ \lambda(\mathsf{D}(\hat{\boldsymbol{v}}_K), \mathsf{D}(\hat{\boldsymbol{w}}_K))_{L^2(K)} + 2\mu h_K^{-1}(\mathsf{S}(\hat{\boldsymbol{v}}_K), \mathsf{S}(\hat{\boldsymbol{w}}_K))_{\boldsymbol{L}^2(\partial K)}.$$

We introduce the global discrete spaces

$$\boldsymbol{V}_{\mathcal{T}_h}^k := \{\boldsymbol{v}_{\mathcal{T}_h} \in \boldsymbol{L}^2(D) \mid \boldsymbol{v}_K := \boldsymbol{v}_{\mathcal{T}_h|K} \in \boldsymbol{V}_K^k, \, \forall K \in \mathcal{T}_h\},$$
$$\boldsymbol{V}_{\mathcal{F}_h,0}^k := \{\boldsymbol{v}_{\mathcal{F}_h} \in \boldsymbol{L}^2(\mathcal{F}_h) \mid \boldsymbol{v}_{\partial K} := \boldsymbol{v}_{\mathcal{F}_h|\partial K} \in \boldsymbol{V}_{\partial K}^k, \, \forall K \in \mathcal{T}_h; \, \boldsymbol{v}_{\mathcal{F}_h|\mathcal{F}_h^\partial} = \boldsymbol{0}\},$$

and the product space $\hat{\boldsymbol{V}}_{h,0}^k := \boldsymbol{V}_{\mathcal{T}_h}^k \times \boldsymbol{V}_{\mathcal{F}_h,0}^k$. For every pair $\hat{\boldsymbol{v}}_h := (\boldsymbol{v}_{\mathcal{T}_h}, \boldsymbol{v}_{\mathcal{F}_h}) \in \hat{\boldsymbol{V}}_{h,0}^k$, we denote by $\hat{\boldsymbol{v}}_K := (\boldsymbol{v}_K, \boldsymbol{v}_{\partial K}) \in \hat{\boldsymbol{V}}_K^k$ its local components in the mesh cell $K \in \mathcal{T}_h$. The discrete problem is as follows:

$$\begin{cases} \text{Find } \hat{\boldsymbol{u}}_h \in \hat{\boldsymbol{V}}_{h,0}^k \text{ such that} \\ \hat{a}_h(\hat{\boldsymbol{u}}_h, \hat{\boldsymbol{w}}_h) = \ell_h(\boldsymbol{w}_{\mathcal{T}_h}), \quad \forall \hat{\boldsymbol{w}}_h \in \hat{\boldsymbol{V}}_{h,0}^k, \end{cases} \tag{42.28}$$

where the forms $\hat{a}_h$ and ℓ_h are assembled cellwise by setting $\hat{a}_h(\hat{\boldsymbol{v}}_h, \hat{\boldsymbol{w}}_h) := \sum_{K \in \mathcal{T}_h} \hat{a}_K(\hat{\boldsymbol{v}}_K, \hat{\boldsymbol{w}}_K)$ and $\ell_h(\boldsymbol{w}_{\mathcal{T}_h}) := \sum_{K \in \mathcal{T}_h}(\boldsymbol{f}, \boldsymbol{w}_K)_{\boldsymbol{L}^2(K)}$.

Remark 42.17 (Elimination of cell unknowns). As in §39.1, the cell unknowns can be eliminated locally in the discrete problem (42.28) by using a Schur complement technique, i.e., static condensation. The global transmission problem coupling the face unknowns is of size $3\binom{k+2}{2}N_\mathrm{f}$ (for $d = 3$), where N_f is the number of mesh interfaces (that is, $9N_\mathrm{f}$ in the lowest-order case $k = 1$). Moreover, one can define as in §39.2.3 face-based tractions in each cell that are in equilibrium with the applied load and the internal efforts and that comply at the interfaces with the law of action and reaction. □

Remark 42.18 (Polynomial degree). The minimal value $k \ge 1$ is needed to control the rigid displacements since $\mathbb{P}_{0,d} \subsetneq \boldsymbol{R} \subsetneq \mathbb{P}_{1,d}$ (recall that the minimal value of the polynomial degree for scalar elliptic PDEs is $k \ge 0$). □

Remark 42.19 (Literature). HHO methods for linear elasticity were introduced in Di Pietro and Ern [166]. Applications to nonlinear mechanics are developed in Botti et al. [75], Abbas et al. [1, 2]. □

We equip the local HHO space $\hat{\boldsymbol{V}}_K^k$ with the strain-seminorm

$$|\hat{\boldsymbol{v}}_K|_{\mathbb{e},K}^2 := 2\|\mathbb{e}(\boldsymbol{v}_K)\|_{\mathbb{L}^2(K)}^2 + h_K^{-1}\|\boldsymbol{v}_K - \boldsymbol{v}_{\partial K}\|_{\boldsymbol{L}^2(\partial K)}^2, \tag{42.29}$$

and the global HHO space $\hat{\boldsymbol{V}}_{h,0}^k$ with the norm

$$\|\hat{\boldsymbol{v}}_h\|_{\hat{\boldsymbol{V}}_{h,0}^k}^2 := \sum_{K \in \mathcal{T}_h} \left(\mu|\hat{\boldsymbol{v}}_K|_{\mathbb{e},K}^2 + \lambda\|\mathsf{D}(\hat{\boldsymbol{v}}_K)\|_{L^2(K)}^2 \right). \tag{42.30}$$

This indeed defines a norm on $\hat{\boldsymbol{V}}_{h,0}^k$ since $\|\hat{\boldsymbol{v}}_h\|_{\hat{\boldsymbol{V}}_{h,0}^k} = 0$ implies that for all $K \in \mathcal{T}_h$, $\boldsymbol{v}_K$ is a rigid displacement whose trace on ∂K is $\boldsymbol{v}_{\partial K}$. Since two rigid displacements that coincide on a face are identical, we infer that $\boldsymbol{w}_{\mathcal{T}_h}$ is a global rigid displacement, and the Dirichlet condition enforced on $\boldsymbol{v}_{\mathcal{F}_h} \in \boldsymbol{V}_{\mathcal{F}_h,0}^k$ at the boundary faces implies that $\boldsymbol{v}_{\mathcal{T}_h}$ and $\boldsymbol{v}_{\mathcal{F}_h}$ are zero.

Lemma 42.20 (Stability, well-posedness). (i) *There are $0 < \eta \leq \omega$ s.t.*

$$\eta\,\|\hat{\boldsymbol{v}}_K\|_{\oplus,K}^2 \leq 2\|\oplus(\mathsf{R}(\hat{\boldsymbol{v}}_K))\|_{\mathbb{L}^2(K)}^2 + h_K^{-1}\|\mathsf{S}(\hat{\boldsymbol{v}}_K)\|_{\boldsymbol{L}^2(\partial K)}^2 \leq \omega\,\|\hat{\boldsymbol{v}}_K\|_{\oplus,K}^2,$$

for all $\hat{\boldsymbol{v}}_K \in \hat{\boldsymbol{V}}_K^k$, all $K \in \mathcal{T}_h$, and all $h \in \mathcal{H}$, and we have

$$\hat{a}_h(\hat{\boldsymbol{v}}_h, \hat{\boldsymbol{v}}_h) \geq \min(1,\eta)\,\|\hat{\boldsymbol{v}}_h\|_{\hat{\boldsymbol{V}}_{h,0}^k}^2, \qquad (42.31)$$

for all $\hat{\boldsymbol{v}}_h \in \boldsymbol{V}_{h,0}^k$. (ii) *The discrete problem (42.28) is well-posed.*

Proof. See Exercise 42.5. $\qquad\qquad\Box$

To derive an error estimate, we introduce the consistency error $\delta_{\mathcal{I}}(\boldsymbol{u}) \in (\hat{\boldsymbol{V}}_{h,0}^k)'$ s.t.

$$\langle \delta_{\mathcal{I}}(\boldsymbol{u}), \hat{\boldsymbol{w}}_h \rangle_{(\hat{\boldsymbol{V}}_{h,0}^k)', \hat{\boldsymbol{V}}_{h,0}^k} := \hat{\ell}_h(\hat{\boldsymbol{w}}_h) - \hat{a}_h(\hat{\mathcal{I}}_h^k(\boldsymbol{u}), \hat{\boldsymbol{w}}_h), \qquad \forall \hat{\boldsymbol{w}}_h \in \hat{\boldsymbol{V}}_{h,0}^k,$$

with $\hat{\mathcal{I}}_h^k : \boldsymbol{H}_0^1(D) \to \hat{\boldsymbol{V}}_{h,0}^k$ s.t. $(\hat{\mathcal{I}}_h^k(\boldsymbol{v}))_K := \hat{\mathcal{I}}_K^k(\boldsymbol{v}_{|K})$ for all $\boldsymbol{v} \in \boldsymbol{H}_0^1(D)$ and all $K \in \mathcal{T}_h$. Note that $\hat{\mathcal{I}}_h^k(\boldsymbol{H}_0^1(D)) \subset \hat{\boldsymbol{V}}_{h,0}^k$ since functions in $\boldsymbol{H}_0^1(D)$ have zero jumps across the mesh interfaces and zero traces at the boundary faces.

Lemma 42.21 (Consistency). *Assume that $\boldsymbol{u} \in \boldsymbol{H}^{1+r}(D)$, $r > \frac{1}{2}$. There is c, uniform w.r.t. μ and λ, such that for all $h \in \mathcal{H}$,*

$$\|\delta_{\mathcal{I}}(u)\|_{(\mathcal{V}_h^k)'} \leq c \sum_{K \in \mathcal{T}_h} \left(\mu\|\boldsymbol{u} - \boldsymbol{\mathcal{E}}_K(\boldsymbol{u})\|_{\sharp,K}^2 + \lambda\|\nabla\!\cdot\!\boldsymbol{u} - \Pi_K^k(\nabla\!\cdot\!\boldsymbol{u})\|_{L^2(K)}^2 \right),$$

where $\|\boldsymbol{v}\|_{\sharp,K} := \|\oplus(\boldsymbol{v})\|_{\mathbb{L}^2(K)} + h_K^{\frac{1}{2}}\|\oplus(\boldsymbol{v})\|_{\mathbb{L}^2(\partial K)}$ for all $\boldsymbol{v} \in \boldsymbol{H}^{1+r}(K)$.

Proof. See Exercise 42.5 $\qquad\qquad\Box$

Theorem 42.22 (Error estimate). *Let $\boldsymbol{u}$ solve (42.10) and let $\hat{\boldsymbol{u}}_h \in \hat{\boldsymbol{V}}_{h,0}^k$ solve (42.28). Assume that $\boldsymbol{u} \in \boldsymbol{H}^{1+r}(D)$, $r > \frac{1}{2}$.* (i) *Letting $\|\phi\|_{\dagger,K} := \|\phi\|_{L^2(K)} + h_K^{\frac{1}{2}}\|\phi\|_{L^2(\partial K)}$, there is c s.t. for all $h \in \mathcal{H}$,*

$$\sum_{K \in \mathcal{T}_h} \mu\|\oplus(\boldsymbol{u} - \boldsymbol{r}_K)\|_{\mathbb{L}^2(K)}^2 \leq c \sum_{K \in \mathcal{T}_h} \left(\mu\|\boldsymbol{u} - \boldsymbol{\mathcal{E}}_K(\boldsymbol{u})\|_{\sharp,K}^2 + \lambda\|\nabla\!\cdot\!\boldsymbol{u} - \Pi_K^k(\nabla\!\cdot\!\boldsymbol{u})\|_{\dagger,K}^2 \right),$$

with $\boldsymbol{r}_K := \mathsf{R}(\hat{\boldsymbol{u}}_K)$. (ii) *If $\boldsymbol{u} \in \boldsymbol{H}^{k+2}(D)$ and $\nabla\!\cdot\!\boldsymbol{u} \in H^{k+1}(D)$, then*

$$\sum_{K \in \mathcal{T}_h} \mu\|\oplus(\boldsymbol{u} - \boldsymbol{r}_K)\|_{\mathbb{L}^2(K)}^2 \leq c \sum_{K \in \mathcal{T}_h} h_K^{2(k+1)} \left(\mu|\boldsymbol{u}|_{\boldsymbol{H}^{k+2}(K)}^2 + \lambda|\nabla\!\cdot\!\boldsymbol{u}|_{H^{k+1}(K)}^2 \right).$$

Proof. Use the approximation properties of the local elliptic projection $\boldsymbol{\mathcal{E}}_K$ and the L^2-orthogonal projection Π_K^k (see the proof of Theorem 39.17). $\qquad\qquad\Box$

Exercises

Exercise 42.1 (Compliance). (i) Let $\mathbb{s}(\mathbb{e})$ be defined in (42.3) (i.e., $\mathbb{s}(\mathbb{e}) := 2\mu\mathbb{e} + \lambda\mathrm{tr}(\mathbb{e})\mathbb{I}_d$) and let $\mathbb{A}$ be the fourth order tensor s.t. $\mathbb{s}(\mathbb{e}) = \mathbb{A}\mathbb{e}$. Verify that $\mathbb{A}$ is symmetric positive definite. (*Hint*: compute the quadratic form $\mathbb{A}\mathbb{e}{:}\mathbb{f}$.) Compute $\mathbb{A}^{\frac{1}{2}}\mathbb{e}$. (*Hint*: find $\alpha, \beta \in \mathbb{R}$ s.t. $\mathbb{A}^{\frac{1}{2}}\mathbb{e} = \alpha\mathbb{e} + \beta\,\mathrm{tr}(\mathbb{e})\mathbb{I}$.) (ii) Invert (42.3), i.e., express $\mathbb{e}$ as a function of $\mathbb{s}$ (the fourth-order tensor $\mathbb{C}$ s.t. $\mathbb{e} = \mathbb{C}\mathbb{s}$ is called *compliance tensor*). (*Hint*: compute first $\mathrm{tr}(\mathbb{s})$.) Compute $\mathbb{e}{:}\mathbb{s}$ in terms of $\mathbb{s}'$ and $\mathrm{tr}(\mathbb{s})$ where $\mathbb{t}' := \mathbb{t} - \frac{1}{3}\mathrm{tr}(\mathbb{t})\mathbb{I}$ is the deviatoric (i.e., trace-free) part of the tensor $\mathbb{t}$. (iii) Consider the Hellinger–Reissner functional $\mathfrak{L}_{\mathrm{HR}}(\mathbb{t}, \boldsymbol{v}) := \int_D(\frac{1}{4\mu}\mathbb{t}'{:}\mathbb{t}' + \frac{1}{18\kappa}\mathrm{tr}(\mathbb{t})^2 + (\nabla\cdot\mathbb{t})\cdot\boldsymbol{v} - \boldsymbol{f}\cdot\boldsymbol{v})\,\mathrm{d}x$ on $\mathbb{H} \times \boldsymbol{V}$ where $\mathbb{H} := \{\mathbb{t} \in \mathbb{L}^2(D) \mid \mathbb{t} = \mathbb{t}^\mathsf{T}, \nabla\cdot\mathbb{t} \in \boldsymbol{L}^2(D)\}$ and $\boldsymbol{V} := \boldsymbol{L}^2(D)$. Find the equations (in weak form) satisfied by a critical point $(\mathbb{s}, \boldsymbol{u})$ of $\mathfrak{L}_{\mathrm{HR}}$. Verify that $(\mathbb{s}, \boldsymbol{u})$ satisfies (42.1) and (42.3) a.e. in D. (*Hint*: use a density argument.)

Exercise 42.2 (Second-order system). (i) Find matrices $\mathbb{A}^{jk} \in \mathbb{R}^{d\times d}$ for all $j, k \in \{1{:}d\}$ s.t. $\nabla\cdot\mathbb{s}(\boldsymbol{u}) = \sum_{j,k} \partial_j(\mathbb{A}^{jk}\partial_k\boldsymbol{u})$. (*Hint*: verify that $\sum_{j,k} \partial_j(\lambda(\boldsymbol{e}_j \otimes \boldsymbol{e}_k)\partial_k\boldsymbol{u}) = \nabla(\lambda\nabla\cdot\boldsymbol{u})$ and $\sum_{j,k} \partial_j(\mu(\boldsymbol{e}_k \otimes \boldsymbol{e}_j)\partial_k\boldsymbol{u}) = \nabla\cdot(\mu\nabla\boldsymbol{u}^\mathsf{T})$ where $(\boldsymbol{e}_j)_{j\in\{1:d\}}$ is the canonical basis of $\mathbb{R}^d$.) (ii) Verify that $(\mathbb{A}^{jk})^\mathsf{T} = \mathbb{A}^{kj}$. What is the consequence on the bilinear form $a(\boldsymbol{v}, \boldsymbol{w}) := \int_D \partial_j\boldsymbol{w}^\mathsf{T}\mathbb{A}^{jk}\partial_k\boldsymbol{v}\,\mathrm{d}x$?

Exercise 42.3 (Pure traction). The pure traction problem is $\nabla\cdot\mathbb{s}(\boldsymbol{u}) + \boldsymbol{f} = \boldsymbol{0}$ in D and $\mathbb{s}(\boldsymbol{u})\cdot\boldsymbol{n} = \boldsymbol{g}$ on ∂D. (i) Write a weak formulation in $\boldsymbol{H}^1(D)$. (ii) Show that it is necessary that $\int_D \boldsymbol{f}\cdot\boldsymbol{r}\,\mathrm{d}x + \int_{\partial D} \boldsymbol{g}\cdot\boldsymbol{r}\,\mathrm{d}s = 0$ for a weak solution to exist. (iii) Assume that $\boldsymbol{r} \in \boldsymbol{R}$ satisfies $\int_D \boldsymbol{r}\,\mathrm{d}x = \boldsymbol{0}$ and $\int_D \nabla\times\boldsymbol{r}\,\mathrm{d}x = \boldsymbol{0}$. Show that $\boldsymbol{r} = \boldsymbol{0}$. (iv) Let $\boldsymbol{V} := \{\boldsymbol{v} \in \boldsymbol{H}^1(D) \mid \int_D \boldsymbol{v}\,\mathrm{d}x = \boldsymbol{0}, \int_D \nabla\times\boldsymbol{v}\,\mathrm{d}x = \boldsymbol{0}\}$. Show that the weak formulation is well-posed in $\boldsymbol{V}$.

Exercise 42.4 (Timoshenko beam). Consider a horizontal beam $D := (0, L)$ clamped at $x = 0$ and subjected to a (vertical) force distribution f and to a bending moment distribution m. A (vertical) shear force F and a bending moment M are applied at $x = L$. The unknowns are the vertical displacement u and the rotation angle of the transverse section θ s.t. $-(u'' - \theta') = \frac{\gamma}{EI}f$ and $-\gamma\theta'' - (u' - \theta) = \frac{\gamma}{EI}m$ in D, where E is the Young modulus, I is the area moment of inertia, and $\gamma := \frac{2(1+\nu)I}{S\kappa}$ (S is the cross section area and κ is an empirical correction factor usually set to $\frac{5}{6}$). The boundary conditions are $u(0) = 0$, $\theta(0) = 0$, $(u' - \theta)(L) = \frac{\gamma}{EI}F$, and $\theta'(L) = \frac{1}{EI}M$. (i) Assuming $f, m \in L^2(D)$, write a weak formulation for the pair (u, θ) in $Y := X \times X$ with $X := \{v \in H^1(D) \mid v(0) = 0\}$. (ii) Prove the well-posedness of the weak formulation. (*Hint*: use that $2\int_D \theta u'\,\mathrm{d}x \le \mu\|\theta\|_{L^2(D)}^2 + \frac{1}{\mu}|u|_{H^1(D)}^2$ with μ sufficiently close to 1 and the Poincaré–Steklov inequality.) (iii) Write an H^1-conforming finite element approximation and derive H^1- and L^2-error estimates for u and θ.

Exercise 42.5 (HHO). (i) Prove (42.25). (ii) Prove Lemma 42.20. (*Hint*: see Lemma 39.2 and use the local Korn inequality $\|\boldsymbol{v}\|_{\boldsymbol{L}^2(K)} \le ch_K\|\mathbb{e}(\boldsymbol{v})\|_{\mathbb{L}^2(K)}$ for all $\boldsymbol{v} \in \boldsymbol{H}^1(K)$ s.t. $(\boldsymbol{v}, \boldsymbol{r})_{\boldsymbol{L}^2(K)} = 0$ for all $\boldsymbol{r} \in \boldsymbol{R}_K$; see Horgan [246], Kim [269].) (iii) Prove Lemma 42.21. (*Hint*: adapt the proof of Lemma 39.16.)

Chapter 43

Maxwell's equations: $\boldsymbol{H}(\text{curl})$-approximation

The objective of this chapter is to introduce some model problems derived from Maxwell's equations that all fit the Lax-Milgram formalism in $\boldsymbol{H}(\text{curl})$. The approximation is performed using $\boldsymbol{H}(\text{curl})$-conforming edge (Nédélec) finite elements. The analysis relies on a coercivity argument in $\boldsymbol{H}(\text{curl})$ that exploits the presence of a uniformly positive zero-order term in the formulation. A more robust technique controlling the divergence of the approximated field is presented in Chapter 44. The space dimension is 3 in the entire chapter $(d = 3)$, and D is a Lipschitz domain in $\mathbb{R}^3$.

43.1 Maxwell's equations

We start by recalling some basic facts about Maxwell's equations. The reader is referred to Bossavit [74, Chap. 1], Monk [303, Chap. 1], Assous et al. [27, Chap. 1] for a detailed discussion on this model. Maxwell's equations are partial differential equations providing a macroscopic description of electromagnetic phenomena. These equations describe how the electric field $\boldsymbol{E}$, the magnetic field $\boldsymbol{H}$, the electric displacement field $\boldsymbol{D}$, and the magnetic induction $\boldsymbol{B}$ (sometimes called magnetic flux density) interact through the action of currents $\boldsymbol{j}$ and charges ρ:

$$\partial_t \boldsymbol{D} - \nabla{\times}\boldsymbol{H} = -\boldsymbol{j} \quad \text{(Ampère's law)}, \tag{43.1a}$$

$$\partial_t \boldsymbol{B} + \nabla{\times}\boldsymbol{E} = \boldsymbol{0} \quad \text{(Faraday's law of induction)}, \tag{43.1b}$$

$$\nabla{\cdot}\boldsymbol{D} = \rho \quad \text{(Gauss's law for electricity)}, \tag{43.1c}$$

$$\nabla{\cdot}\boldsymbol{B} = 0 \quad \text{(Gauss's law for magnetism)}. \tag{43.1d}$$

Notice that if $(\nabla{\cdot}\boldsymbol{B})_{|t=0} = 0$, taking the divergence of (43.1b) implies that (43.1d) is satisfied at all times. Similarly, assuming $(\nabla{\cdot}\boldsymbol{D})_{|t=0} = \rho_{|t=0}$ and that the charge conservation equation $\partial_t \rho + \nabla{\cdot}\boldsymbol{j} = 0$ is satisfied at all times implies that (43.1c) is satisfied at all times. This shows that if the data ρ, $\boldsymbol{j}$, $\boldsymbol{B}_{|t=0}$, and $\boldsymbol{D}_{|t=0}$ satisfy the proper constraints, Gauss's laws are just consequences of Ampère's law and Faraday's law.

The system (43.1) is closed by relating the fields through constitutive laws describing microscopic mechanisms of polarization and magnetization:

$$\boldsymbol{D} - \varepsilon_0 \boldsymbol{E} = \boldsymbol{P}, \qquad \boldsymbol{B} = \mu_0(\boldsymbol{H} + \boldsymbol{M}), \tag{43.2}$$

where ε_0 and μ_0 are the electric permittivity and the magnetic permeability of vacuum, and $\boldsymbol{P}$ and $\boldsymbol{M}$ are the polarization and the magnetization fields, respectively. These quantities are the average representatives at the macroscopic scale of complex microscopic interactions that must be modeled. The models in question always involve parameters that need to be identified by measurements or other techniques like homogenization or multiscale models. We have $\boldsymbol{P} := \boldsymbol{0}$ and $\boldsymbol{M} := \boldsymbol{0}$ in vacuum, and it is common to use $\boldsymbol{P} := \varepsilon_0 \varepsilon_r \boldsymbol{E}$ and $\boldsymbol{M} := \mu_r \boldsymbol{H}$ to model isotropic homogeneous dielectric and magnetic materials, where ε_r is the electric susceptibility and μ_r is the magnetic susceptibility. In the rest of the chapter, we assume that

$$\boldsymbol{D} := \epsilon \boldsymbol{E} \quad \text{and} \quad \boldsymbol{B} := \mu \boldsymbol{H}, \tag{43.3}$$

where ϵ and μ are given coefficients that may be space-dependent. The current $\boldsymbol{j}$ and charge density ρ are a priori given, but it is also possible to make these quantities depend on the other fields through phenomenological mechanisms. For instance, it is possible to further decompose the current into one component that depends on the material and another one that is a source. The simplest model doing that is Ohm's law, $\boldsymbol{j} = \boldsymbol{j}_\text{s} + \sigma \boldsymbol{E}$, where σ is the electrical conductivity and $\boldsymbol{j}_\text{s}$ an imposed current.

We now formulate Maxwell's equations in two different regimes: the time-harmonic regime and the eddy current limit.

43.1.1 The time-harmonic regime

We first consider Maxwell's equations in the *time-harmonic regime* where the time-dependence is assumed to be of the form $e^{\text{i}\omega t}$ with $\text{i}^2 = -1$ and ω is a given angular frequency. The time-harmonic version of (43.1a)-(43.1b) is

$$\text{i}\omega\epsilon\boldsymbol{E} + \sigma\boldsymbol{E} - \nabla{\times}\boldsymbol{H} = -\boldsymbol{j}_\text{s}, \qquad \text{in } D, \tag{43.4a}$$

$$\text{i}\omega\mu\boldsymbol{H} + \nabla{\times}\boldsymbol{E} = \boldsymbol{0}, \qquad \text{in } D, \tag{43.4b}$$

$$\boldsymbol{H}_{|\partial D_\text{d}}{\times}\boldsymbol{n} = \boldsymbol{a}_\text{d}, \quad \boldsymbol{E}_{|\partial D_\text{n}}{\times}\boldsymbol{n} = \boldsymbol{a}_\text{n}, \quad \text{on } \partial D, \tag{43.4c}$$

where $\{\partial D_\text{d}, \partial D_\text{n}\}$ forms a partition of the boundary ∂D of D. The dependent variables are the electric field $\boldsymbol{E}$ and the magnetic field $\boldsymbol{H}$. The data are the conductivity σ, the permittivity ϵ, the permeability μ, the current $\boldsymbol{j}_\text{s}$, and the boundary data $\boldsymbol{a}_\text{d}$ and $\boldsymbol{a}_\text{n}$. The material coefficients ϵ and μ can be complex-valued. The system (43.4) models for instance a microwave oven; see e.g., [74, Chap. 9]. The conditions $\boldsymbol{H}_{|\partial D_\text{d}}{\times}\boldsymbol{n} = \boldsymbol{0}$ and $\boldsymbol{E}_{|\partial D_\text{n}}{\times}\boldsymbol{n} = \boldsymbol{0}$ are usually called perfect magnetic conductor and perfect electric conductor boundary conditions, respectively.

Let us assume that the modulus of the magnetic permeability μ is bounded away from zero uniformly in D. It is then possible to eliminate $\boldsymbol{H}$ by using $\boldsymbol{H} = \text{i}(\omega\mu)^{-1}\nabla{\times}\boldsymbol{E}$. The system then takes the following form:

$$(-\omega^2\epsilon + \text{i}\omega\sigma)\boldsymbol{E} + \nabla{\times}(\mu^{-1}\nabla{\times}\boldsymbol{E}) = -\text{i}\omega\boldsymbol{j}_\text{s}, \qquad \text{in } D, \tag{43.5a}$$

$$(\nabla{\times}\boldsymbol{E})_{|\partial D_\text{d}}{\times}\boldsymbol{n} = -\text{i}\omega\mu\boldsymbol{a}_\text{d}, \quad \boldsymbol{E}_{|\partial D_\text{n}}{\times}\boldsymbol{n} = \boldsymbol{a}_\text{n}, \quad \text{on } \partial D. \tag{43.5b}$$

Notice that Gauss's law for electricity is contained in (43.5a) since taking the divergence of the equation yields $\nabla{\cdot}((-\omega^2\epsilon + \text{i}\omega\sigma)\boldsymbol{E}) = \nabla{\cdot}(-\text{i}\omega\boldsymbol{j}_\text{s})$, which is the time-harmonic counterpart of (43.1c) combined with (43.1a). The system (43.5) is often used to model the propagation of electromagnetic waves through various media.

43.1.2 The eddy current problem

When the time scale of interest, say τ, is such that the ratio $\epsilon/(\tau\sigma) \ll 1$, it is legitimate to neglect the displacement current in Ampère's law (i.e., Maxwell's correction $\partial_t \boldsymbol{D}$). This situation occurs in

particular in systems with moving parts (either solid or fluids) whose characteristic speed is much slower than the speed of light. The resulting system, called *eddy current problem*, is as follows:

$$\sigma \boldsymbol{E} - \nabla \times \boldsymbol{H} = -\boldsymbol{j}_{\mathrm{s}}, \qquad\qquad \text{in } D, \qquad\qquad (43.6\text{a})$$

$$\partial_t(\mu \boldsymbol{H}) + \nabla \times \boldsymbol{E} = \boldsymbol{0}, \qquad\qquad \text{in } D, \qquad\qquad (43.6\text{b})$$

$$\boldsymbol{H}_{|\partial D_{\mathrm{d}}} \times \boldsymbol{n} = \boldsymbol{a}_{\mathrm{d}}, \quad \boldsymbol{E}_{|\partial D_{\mathrm{n}}} \times \boldsymbol{n} = \boldsymbol{a}_{\mathrm{n}}, \quad \text{on } \partial D, \qquad\qquad (43.6\text{c})$$

where $\{\partial D_{\mathrm{d}}, \partial D_{\mathrm{n}}\}$ forms a partition of the boundary ∂D of D. The system (43.6) arises in magneto-hydrodynamics (MHD). In this case, $\boldsymbol{j}_{\mathrm{s}}$ is further decomposed into $\boldsymbol{j}_{\mathrm{s}} = \boldsymbol{j}_{\mathrm{s}}' + \sigma \boldsymbol{u} \times \boldsymbol{B}$, where $\boldsymbol{u}$ is the velocity of the fluid occupying the domain D, i.e., the actual current is decomposed into $\boldsymbol{j} = \boldsymbol{j}_{\mathrm{s}}' + \sigma(\boldsymbol{E} + \boldsymbol{u} \times \boldsymbol{B})$.

Let us assume that σ is bounded from below away from zero uniformly in D. It is then possible to eliminate the electric field from (43.6) by using $\boldsymbol{E} = \sigma^{-1}(\nabla \times \boldsymbol{H} - \boldsymbol{j}_{\mathrm{s}})$. The new system to be solved is rewritten as follows:

$$\partial_t(\mu \boldsymbol{H}) + \nabla \times (\sigma^{-1} \nabla \times \boldsymbol{H} - \boldsymbol{u} \times (\mu \boldsymbol{H})) = \nabla \times (\sigma^{-1} \boldsymbol{j}_{\mathrm{s}}'), \qquad \text{in } D, \qquad\qquad (43.7\text{a})$$

$$\boldsymbol{H}_{|\partial D_{\mathrm{d}}} \times \boldsymbol{n} = \boldsymbol{a}_{\mathrm{d}}, \quad (\sigma^{-1} \nabla \times \boldsymbol{H} - \boldsymbol{u} \times (\mu \boldsymbol{H}))_{|\partial D_{\mathrm{n}}} \times \boldsymbol{n} = \boldsymbol{c}_{\mathrm{n}}, \quad \text{on } \partial D, \qquad\qquad (43.7\text{b})$$

where $\boldsymbol{c}_{\mathrm{n}} := \boldsymbol{a}_{\mathrm{n}} + (\sigma^{-1} \boldsymbol{j}_{\mathrm{s}}')_{|\partial D_{\mathrm{n}}} \times \boldsymbol{n}$. At this point, it is possible to further simplify the problem by assuming that either the time evolution is harmonic, i.e., $\boldsymbol{H}(\boldsymbol{x}, t) := \boldsymbol{H}_{\mathrm{sp}}(\boldsymbol{x}) e^{\mathrm{i}\omega t}$, or the time derivative is approximated as $\partial_t \boldsymbol{H}(\boldsymbol{x}, t) \approx \tau^{-1}(\boldsymbol{H}(\boldsymbol{x}, t) - \boldsymbol{H}(\boldsymbol{x}, t - \tau))$, where τ is the time step of the time discretization. After appropriately renaming the dependent variable and the data, say either $\widetilde{\mu} := \mathrm{i}\omega\mu$ and $\boldsymbol{f} := \nabla \times (\sigma^{-1} \boldsymbol{j}_{\mathrm{s}}')$, or $\widetilde{\mu} := \mu\tau^{-1}$ and $\boldsymbol{f} := \nabla \times (\sigma^{-1} \boldsymbol{j}_{\mathrm{s}}') + \widetilde{\mu} \boldsymbol{H}(\boldsymbol{x}, t - \tau)$, the above system reduces to solving the following problem:

$$\widetilde{\mu} \boldsymbol{H} + \nabla \times (\sigma^{-1} \nabla \times \boldsymbol{H} - \boldsymbol{u} \times (\mu \boldsymbol{H})) = \boldsymbol{f}, \qquad\qquad \text{in } D, \qquad\qquad (43.8\text{a})$$

$$\boldsymbol{H}_{|\partial D_{\mathrm{d}}} \times \boldsymbol{n} = \boldsymbol{a}_{\mathrm{d}}, \quad (\sigma^{-1} \nabla \times \boldsymbol{H} - \boldsymbol{u} \times (\mu \boldsymbol{H}))_{|\partial D_{\mathrm{n}}} \times \boldsymbol{n} = \boldsymbol{c}_{\mathrm{n}}, \quad \text{on } \partial D. \qquad\qquad (43.8\text{b})$$

Notice that $\nabla \cdot \boldsymbol{f} = 0$ in both cases. Hence, Gauss's law for magnetism is contained in (43.8a) since taking the divergence of the equation yields $\nabla \cdot (\mu \boldsymbol{H}) = 0$ whether $\widetilde{\mu} := \mathrm{i}\omega\mu$ or $\widetilde{\mu} := \mu\tau^{-1}$.

43.2   Weak formulation

The time-harmonic problem and the eddy current problem have a very similar structure. After lifting the boundary condition (either on ∂D_{n} for the time-harmonic problem or on ∂D_{d} for the eddy current problem) and making appropriate changes of notation, the above two problems (43.5) and (43.8) can be reformulated as follows: Find $\boldsymbol{A} : D \to \mathbb{C}^3$ such that

$$\nu \boldsymbol{A} + \nabla \times (\kappa \nabla \times \boldsymbol{A}) = \boldsymbol{f}, \quad \boldsymbol{A}_{|\partial D_{\mathrm{d}}} \times \boldsymbol{n} = \boldsymbol{0}, \quad (\kappa \nabla \times \boldsymbol{A})_{|\partial D_{\mathrm{n}}} \times \boldsymbol{n} = \boldsymbol{0}, \qquad\qquad (43.9)$$

where ν, κ, and $\boldsymbol{f}$ are complex-valued. We have taken $\boldsymbol{u} := \boldsymbol{0}$ in the MHD problem for simplicity. We have also assumed that the Neumann data is zero to avoid unnecessary technicalities. We have $\nu := -\omega^2\epsilon + \mathrm{i}\omega\sigma$ and $\kappa := \mu^{-1}$ for the time-harmonic problem, and $\nu := \mathrm{i}\omega\mu$ or $\nu := \mu\tau^{-1}$ and $\kappa := \sigma^{-1}$ for the eddy current problem.

43.2.1   Functional setting

Let us assume that $\boldsymbol{f} \in \boldsymbol{L}^2(D) := L^2(D; \mathbb{C}^3)$ and $\nu, \kappa \in L^{\infty}(D; \mathbb{C})$. A weak formulation of (43.9) is obtained by multiplying the PDE by the complex conjugate of a smooth test function $\boldsymbol{b}$ with

zero tangential component over ∂D_d and integrating by parts. Recalling (4.11), we obtain

$$\int_D (\nu\boldsymbol{A}{\cdot}\overline{\boldsymbol{b}} + \kappa\nabla{\times}\boldsymbol{A}{\cdot}\nabla{\times}\overline{\boldsymbol{b}})\,\mathrm{d}x = \int_D \boldsymbol{f}{\cdot}\overline{\boldsymbol{b}}\,\mathrm{d}x.$$

The integral on the left-hand side makes sense if $\boldsymbol{A}, \boldsymbol{b} \in \boldsymbol{H}(\mathrm{curl}; D)$. To be dimensionally coherent, we equip $\boldsymbol{H}(\mathrm{curl}; D)$ with the norm $\|\boldsymbol{b}\|_{\boldsymbol{H}(\mathrm{curl};D)} := (\|\boldsymbol{b}\|^2_{\boldsymbol{L}^2(D)} + \ell_D^2\|\nabla{\times}\boldsymbol{b}\|^2_{\boldsymbol{L}^2(D)})^{\frac{1}{2}}$, where ℓ_D is some characteristic length of D, e.g., $\ell_D := \mathrm{diam}(D)$.

Let $\gamma^c : \boldsymbol{H}(\mathrm{curl}; D) \to \boldsymbol{H}^{-\frac{1}{2}}(\partial D)$ denote the tangential trace operator introduced in (4.11) and let $\langle\cdot,\cdot\rangle_{\partial D}$ denote the duality pairing between $\boldsymbol{H}^{-\frac{1}{2}}(\partial D)$ and $\boldsymbol{H}^{\frac{1}{2}}(\partial D)$. Since the Dirichlet condition $\gamma^c(\boldsymbol{A}) = \boldsymbol{0}$ is enforced on ∂D_d only, we must consider the restriction of the linear forms in $\boldsymbol{H}^{-\frac{1}{2}}(\partial D)$ to functions that are only defined on ∂D_d. Let $\widetilde{\boldsymbol{H}^{\frac{1}{2}}}(\partial D_\mathrm{d})$ be composed of the functions $\boldsymbol{\theta}$ defined on ∂D_d whose zero-extension to ∂D, say $\tilde{\boldsymbol{\theta}}$, is in $\boldsymbol{H}^{\frac{1}{2}}(\partial D)$. Then for all $\boldsymbol{b} \in \boldsymbol{H}(\mathrm{curl}; D)$, the restriction $\gamma^c(\boldsymbol{b})_{|\partial D_\mathrm{d}}$ is defined in $\widetilde{\boldsymbol{H}^{\frac{1}{2}}}(\partial D_\mathrm{d})'$ by using the duality product $\langle\gamma^c(\boldsymbol{b})_{|\partial D_\mathrm{d}}, \boldsymbol{\theta}\rangle_{\partial D_\mathrm{d}} := \langle\gamma^c(\boldsymbol{b}), \tilde{\boldsymbol{\theta}}\rangle_{\partial D}$ for all $\boldsymbol{\theta} \in \widetilde{\boldsymbol{H}^{\frac{1}{2}}}(\partial D_\mathrm{d})$. A weak formulation of (43.9) is the following:

$$\begin{cases} \text{Find } \boldsymbol{A} \in \boldsymbol{V}_\mathrm{d} := \{\boldsymbol{b} \in \boldsymbol{H}(\mathrm{curl}; D) \mid \gamma^c(\boldsymbol{b})_{|\partial D_\mathrm{d}} = \boldsymbol{0}\} \text{ such that} \\ a_{\nu,\kappa}(\boldsymbol{A}, \boldsymbol{b}) = \ell(\boldsymbol{b}), \quad \forall \boldsymbol{b} \in \boldsymbol{V}_\mathrm{d}, \end{cases} \tag{43.10}$$

with the following sesquilinear and antilinear forms:

$$a_{\nu,\kappa}(\boldsymbol{a}, \boldsymbol{b}) := \int_D (\nu\boldsymbol{a}{\cdot}\overline{\boldsymbol{b}} + \kappa\nabla{\times}\boldsymbol{a}{\cdot}\nabla{\times}\overline{\boldsymbol{b}})\,\mathrm{d}x, \qquad \ell(\boldsymbol{b}) := \int_D \boldsymbol{f}{\cdot}\overline{\boldsymbol{b}}\,\mathrm{d}x. \tag{43.11}$$

43.2.2   Well-posedness

We assume that there are real numbers θ, $\nu_\flat > 0$, and $\kappa_\flat > 0$ s.t.

$$\operatorname*{ess\,inf}_{\boldsymbol{x}\in D} \Re\big(e^{i\theta}\nu(\boldsymbol{x})\big) \geq \nu_\flat \quad \text{and} \quad \operatorname*{ess\,inf}_{\boldsymbol{x}\in D} \Re\big(e^{i\theta}\kappa(\boldsymbol{x})\big) \geq \kappa_\flat. \tag{43.12}$$

Let us set $\nu_\sharp := \|\nu\|_{L^\infty(D;\mathbb{C})}$ and $\kappa_\sharp := \|\kappa\|_{L^\infty(D;\mathbb{C})}$.

Theorem 43.1 (Coercivity, well-posedness). (i) *Assume* $\boldsymbol{f} \in \boldsymbol{L}^2(D)$, $\nu, \kappa \in L^\infty(D;\mathbb{C})$, *and* (43.12). *Then the sesquilinear form* $a_{\nu,\kappa}$ *is coercive and bounded:*

$$\Re\big(e^{i\theta}a_{\nu,\kappa}(\boldsymbol{b}, \boldsymbol{b})\big) \geq \min(\nu_\flat, \ell_D^{-2}\kappa_\flat)\|\boldsymbol{b}\|^2_{\boldsymbol{H}(\mathrm{curl};D)}, \tag{43.13a}$$

$$|a_{\nu,\kappa}(\boldsymbol{a}, \boldsymbol{b})| \leq \max(\nu_\sharp, \ell_D^{-2}\kappa_\sharp)\|\boldsymbol{a}\|_{\boldsymbol{H}(\mathrm{curl};D)}\|\boldsymbol{b}\|_{\boldsymbol{H}(\mathrm{curl};D)}, \tag{43.13b}$$

for all $\boldsymbol{a}, \boldsymbol{b} \in \boldsymbol{H}(\mathrm{curl}; D)$. (ii) *The problem* (43.10) *is well-posed.*

Proof. Let us first verify that $\boldsymbol{V}_\mathrm{d}$ is a closed subspace of $\boldsymbol{H}(\mathrm{curl}; D)$. Let $(\boldsymbol{b}_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in $\boldsymbol{V}_\mathrm{d}$. Then $\boldsymbol{b}_n \to \boldsymbol{b}$ in $\boldsymbol{H}(\mathrm{curl}; D)$, and for all $\boldsymbol{\theta} \in \widetilde{\boldsymbol{H}^{\frac{1}{2}}}(\partial D_\mathrm{d})$, we have

$$0 = \langle\gamma^c(\boldsymbol{b}_n)_{|\partial D_\mathrm{d}}, \boldsymbol{\theta}\rangle_{\partial D_\mathrm{d}} := \langle\gamma^c(\boldsymbol{b}_n), \tilde{\boldsymbol{\theta}}\rangle_{\partial D} \to \langle\gamma^c(\boldsymbol{b}), \tilde{\boldsymbol{\theta}}\rangle_{\partial D} =: \langle\gamma^c(\boldsymbol{b}), \boldsymbol{\theta}\rangle_{\partial D_\mathrm{d}},$$

so that $\boldsymbol{b} \in \boldsymbol{V}_\mathrm{d}$. (Recall that (4.11) implies that $\gamma^c : \boldsymbol{H}(\mathrm{curl}; D) \to \boldsymbol{H}^{-\frac{1}{2}}(\partial D)$ is continuous.) Moreover, coercivity follows from (43.12) since we have

$$\begin{aligned} \Re(e^{i\theta}a_{\nu,\kappa}(\boldsymbol{b}, \boldsymbol{b})) &= \int_D \left(\Re(e^{i\theta}\nu)|\boldsymbol{b}|^2 + \Re(e^{i\theta}\kappa)|\nabla{\times}\boldsymbol{b}|^2\right)\mathrm{d}x \\ &\geq \int_D \left(\nu_\flat|\boldsymbol{b}|^2 + \kappa_\flat|\nabla{\times}\boldsymbol{b}|^2\right)\mathrm{d}x \geq \min(\nu_\flat, \ell_D^{-2}\kappa_\flat)\|\boldsymbol{b}\|^2_{\boldsymbol{H}(\mathrm{curl};D)}. \end{aligned}$$

Similarly, the boundedness of $a_{\nu,\kappa}$ follows from $\nu, \kappa \in L^\infty(D;\mathbb{C})$, and the boundedness of ℓ follows from $\boldsymbol{f} \in \boldsymbol{L}^2(D)$. Finally, well-posedness follows from the complex version of the Lax–Milgram lemma. $\qquad\square$

Example 43.2 (Property (43.12)**).** Assume to fix the ideas that κ is real and uniformly positive. If ν is also real and uniformly positive, (43.12) is satisfied with $\theta := 0$, $\kappa_\flat := \operatorname{ess\,inf}_{\boldsymbol{x}\in D} \kappa(\boldsymbol{x})$, and $\nu_\flat := \operatorname{ess\,inf}_{\boldsymbol{x}\in D} \nu(\boldsymbol{x})$. If instead ν is purely imaginary with a uniformly positive imaginary part, (43.12) is satisfied with $\theta := -\frac{\pi}{4}$, $\kappa_\flat := \frac{\sqrt{2}}{2} \operatorname{ess\,inf}_{\boldsymbol{x}\in D} \kappa(\boldsymbol{x})$, and $\nu_\flat := \frac{\sqrt{2}}{2} \operatorname{ess\,inf}_{\boldsymbol{x}\in D} \Im(\nu(\boldsymbol{x}))$. More generally, if $\nu := \rho_\nu e^{i\theta_\nu}$ with $\operatorname{ess\,inf}_{\boldsymbol{x}\in D} \rho_\nu(\boldsymbol{x}) =: \rho_\flat > 0$ and $\theta_\nu(\boldsymbol{x}) \in [\theta_{\min}, \theta_{\max}] \subset (-\pi, \pi)$ a.e. in D, then setting $\delta := \theta_{\max} - \theta_{\min}$ and assuming that $\delta < \pi$, (43.12) is satisfied with $\theta := -\frac{1}{2}(\theta_{\min}+\theta_{\max})\frac{\pi}{2\pi-\delta}$, $\nu_\flat := \min(\cos(\theta_{\min}+\theta), \cos(\theta_{\max}+\theta))\rho_\flat$ and $\kappa_\flat := \cos(\theta) \operatorname{ess\,inf}_{\boldsymbol{x}\in D} \kappa(\boldsymbol{x})$ (see Exercise 43.3). An important example where the condition (43.12) fails is when the two complex numbers ν and κ are collinear and point in opposite directions. In this case, resonances may occur and (43.10) has to be replaced by an eigenvalue problem. $\qquad\square$

43.2.3 Regularity

In the case of constant or smooth coefficients, a smoothness property on the solution to (43.10) can be inferred from the following important result.

Lemma 43.3 (Regularity). *Let D be a Lipschitz domain in $\mathbb{R}^3$. (i) There is $c > 0$ s.t. the following holds true:*

$$c\,\ell_D^s |\boldsymbol{v}|_{\boldsymbol{H}^s(D)} \le \|\boldsymbol{v}\|_{\boldsymbol{L}^2(D)} + \ell_D \|\nabla\times\boldsymbol{v}\|_{\boldsymbol{L}^2(D)} + \ell_D \|\nabla\cdot\boldsymbol{v}\|_{L^2(D)}, \qquad (43.14)$$

with $s := \frac{1}{2}$, for all vector fields $\boldsymbol{v} \in \boldsymbol{H}(\mathrm{curl}; D) \cap \boldsymbol{H}(\mathrm{div}; D)$ with either zero normal trace or zero tangential trace over ∂D. (ii) The estimate remains valid with $s \in (\frac{1}{2}, 1]$ if D is a Lipschitz polyhedron, and with $s := 1$ if D is convex.

Proof. (i) For the proof of (43.14), see Birman and Solomyak [57, Thm. 3.1] and Costabel [142, Thm. 2]. (ii) See Amrouche et al. [10, Prop. 3.7] when D is a Lipschitz polyhedron and [10, Thm. 2.17] when D is convex. $\qquad\square$

Let us consider the problem (43.10) and assume that $\boldsymbol{f} \in \boldsymbol{H}(\mathrm{div}; D)$ and ν is constant (or smooth) over D. Then the unique solution $\boldsymbol{A}$ is such that $\nabla\cdot\boldsymbol{A} = \nu^{-1}\nabla\cdot\boldsymbol{f} \in L^2(D)$. Hence, $\boldsymbol{A} \in \boldsymbol{H}(\mathrm{curl}; D) \cap \boldsymbol{H}(\mathrm{div}; D)$. Moreover, (43.9) implies that $\nabla\times(\kappa\nabla\times\boldsymbol{A}) \in \boldsymbol{L}^2(D)$ so that, assuming that κ is constant (or smooth) over D, we infer that $\nabla\times\boldsymbol{A} \in \boldsymbol{H}(\mathrm{curl}; D) \cap \boldsymbol{H}(\mathrm{div}; D)$. In addition to the above assumptions on ν and κ, let us also assume that $\partial D_\mathrm{n} = \emptyset$ (i.e., $\boldsymbol{A}$ has a zero tangential trace, which implies that $\nabla\times\boldsymbol{A}$ has a zero normal trace a.e. on ∂D). Lemma 43.3 implies that there exists $r > 0$ so that

$$\boldsymbol{A} \in \boldsymbol{H}^r(D), \qquad \nabla\times\boldsymbol{A} \in \boldsymbol{H}^r(D), \qquad (43.15)$$

with $r := \frac{1}{2}$ in general, $r \in (\frac{1}{2}, 1]$ if D is a Lipschitz polyhedron, and $r := 1$ if D is convex. In the more general case of heterogeneous coefficients, we will see in the next chapter (see Lemma 44.2) that the smoothness assumption (43.15) is still valid with a smoothness index $r > 0$ under appropriate assumptions on ν. In the rest of this chapter, we are going to assume that (43.15) holds true with $r > 0$.

43.3   Approximation using edge elements

We assume that the hypotheses of Theorem 43.1 are satisfied so that the boundary-value problem (43.10) is well-posed.

43.3.1   Discrete setting

We consider a shape-regular sequence of affine meshes $(\mathcal{T}_h)_{h \in \mathcal{H}}$ of D. We assume that D is a Lipschitz polyhedron so that each mesh covers D exactly. We also assume that the meshes are compatible with the partition of the boundary into $\{\partial D_\mathrm{d}, \partial D_\mathrm{n}\}$. We consider the Nédélec (or edge) finite elements of some order $k \geq 0$ from Chapter 15 and the corresponding $\boldsymbol{H}(\mathrm{curl})$-conforming finite element space $\boldsymbol{P}_k^\mathrm{c}(\mathcal{T}_h)$ built in Chapter 19. Let $\boldsymbol{V}_{h\mathrm{d}}$ be the subspace of $\boldsymbol{P}_k^\mathrm{c}(\mathcal{T}_h)$ defined by

$$\boldsymbol{V}_{h\mathrm{d}} := \{\boldsymbol{b}_h \in \boldsymbol{P}_k^\mathrm{c}(\mathcal{T}_h) \mid \boldsymbol{b}_{h|\partial D_\mathrm{d}} \times \boldsymbol{n} = \boldsymbol{0}\}. \tag{43.16}$$

Since the Dirichlet boundary condition is strongly enforced in $\boldsymbol{V}_{h\mathrm{d}}$, the approximation setting is conforming, i.e., $\boldsymbol{V}_{h\mathrm{d}} \subset \boldsymbol{V}_\mathrm{d}$. The discrete formulation of (43.10) is

$$\begin{cases} \text{Find } \boldsymbol{A}_h \in \boldsymbol{V}_{h\mathrm{d}} \text{ such that} \\ a_{\nu,\kappa}(\boldsymbol{A}_h, \boldsymbol{b}_h) = \ell(\boldsymbol{b}_h), \quad \forall \boldsymbol{b}_h \in \boldsymbol{V}_{h\mathrm{d}}. \end{cases} \tag{43.17}$$

The Lax–Milgram lemma together with the conformity of the approximation setting implies that (43.17) has a unique solution.

43.3.2 $\boldsymbol{H}(\mathrm{curl})$-error estimate

Theorem 43.4 ($\boldsymbol{H}(\mathrm{curl})$-error estimate). (i) *Under the assumptions of Theorem* 43.1, *there is* c *s.t. for all* $h \in \mathcal{H}$,

$$\|\boldsymbol{A} - \boldsymbol{A}_h\|_{\boldsymbol{H}(\mathrm{curl};D)} \leq c \inf_{\boldsymbol{b}_h \in \boldsymbol{V}_{h\mathrm{d}}} \|\boldsymbol{A} - \boldsymbol{b}_h\|_{\boldsymbol{H}(\mathrm{curl};D)}. \tag{43.18}$$

(ii) *Assuming that either* $\partial D_\mathrm{d} = \partial D$ *or* $\partial D_\mathrm{n} = \partial D$ *and that there is* $r \in (0, k+1]$ *s.t.* $\boldsymbol{A} \in \boldsymbol{H}^r(D)$ *and* $\nabla \times \boldsymbol{A} \in \boldsymbol{H}^r(D)$, *where* $k \geq 0$ *is the degree of the finite element used to build* $\boldsymbol{V}_{h\mathrm{d}}$, *we have*

$$\|\boldsymbol{A} - \boldsymbol{A}_h\|_{\boldsymbol{H}(\mathrm{curl};D)} \leq c\, h^r(|\boldsymbol{A}|_{\boldsymbol{H}^r(D)} + \ell_D |\nabla \times \boldsymbol{A}|_{\boldsymbol{H}^r(D)}). \tag{43.19}$$

Proof. (i) The estimate (43.18) is a direct consequence of Céa's lemma.
(ii) We prove the estimate (43.19) when $\partial D_\mathrm{d} = \partial D$, that is, when $\boldsymbol{V}_{h\mathrm{d}} := \boldsymbol{P}_{k,0}^\mathrm{c}(\mathcal{T}_h) := \{\boldsymbol{b}_h \in \boldsymbol{P}_k^\mathrm{c}(\mathcal{T}_h) \mid \boldsymbol{b}_{h|\partial D} \times \boldsymbol{n} = \boldsymbol{0}\}$. We estimate the infimum in (43.18) by taking $\boldsymbol{b}_h := \mathcal{J}_{h0}^\mathrm{c}(\boldsymbol{A})$, where $\mathcal{J}_{h0}^\mathrm{c} : \boldsymbol{L}^1(D) \to \boldsymbol{P}_{k,0}^\mathrm{c}(\mathcal{T}_h)$ is the commuting quasi-interpolation operator with zero tangential trace introduced in §23.3.3. Owing to the items (ii) and (iii) in Theorem 23.12, we infer that

$$\begin{aligned}
\|\boldsymbol{A} - \mathcal{J}_{h0}^\mathrm{c}(\boldsymbol{A})\|_{\boldsymbol{H}(\mathrm{curl};D)} &\leq \|\boldsymbol{A} - \mathcal{J}_{h0}^\mathrm{c}(\boldsymbol{A})\|_{\boldsymbol{L}^2(D)} + \ell_D \|\nabla \times (\boldsymbol{A} - \mathcal{J}_{h0}^\mathrm{c}(\boldsymbol{A}))\|_{\boldsymbol{L}^2(D)} \\
&= \|\boldsymbol{A} - \mathcal{J}_{h0}^\mathrm{c}(\boldsymbol{A})\|_{\boldsymbol{L}^2(D)} + \ell_D \|\nabla \times \boldsymbol{A} - \mathcal{J}_{h0}^\mathrm{d}(\nabla \times \boldsymbol{A})\|_{\boldsymbol{L}^2(D)} \\
&\leq c \inf_{\boldsymbol{b}_h \in \boldsymbol{P}_{k,0}^\mathrm{c}(\mathcal{T}_h)} \|\boldsymbol{A} - \boldsymbol{b}_h\|_{\boldsymbol{L}^2(D)} + c' \ell_D \inf_{\boldsymbol{d}_h \in \boldsymbol{P}_{k,0}^\mathrm{d}(\mathcal{T}_h)} \|\nabla \times \boldsymbol{A} - \boldsymbol{d}_h\|_{\boldsymbol{L}^2(D)} \\
&\leq c'' h^r(|\boldsymbol{A}|_{\boldsymbol{H}^r(D)} + \ell_D |\nabla \times \boldsymbol{A}|_{\boldsymbol{H}^r(D)}),
\end{aligned}$$

where the last step follows from Corollary 22.16. The proof for $\partial D_\mathrm{n} = \partial D$ is similar if one uses $\mathcal{J}_h^\mathrm{c}, \mathcal{J}_h^\mathrm{d}$ instead of $\mathcal{J}_{h0}^\mathrm{c}, \mathcal{J}_{h0}^\mathrm{d}$. $\qquad\square$

Remark 43.5 ($\nu_\flat$-dependency). The coercivity and boundedness properties in (43.13) show that the constant in the error estimate (43.18) is $c = \frac{\max(\nu_\sharp, \ell_D^{-2}\kappa_\sharp)}{\min(\nu_\flat, \ell_D^{-2}\kappa_\flat)}$, which becomes unbounded when $\nu_\flat$ is very small. This difficulty is addressed in Chapter 44. □

Remark 43.6 (Variants). It is possible to localize (43.19) by using Theorem 22.14 instead of Corollary 22.16 when $\partial D_{\mathrm{d}} = \partial D$, and using Theorem 22.6 instead of Corollary 22.9 when $\partial D_{\mathrm{n}} = \partial D$. Using that $\boldsymbol{A} \in \boldsymbol{H}_0(\mathrm{curl}; D)$, $\nabla \times \boldsymbol{A} \in \boldsymbol{H}_0(\mathrm{div}; D)$, and the regularity of the mesh sequence, Theorem 22.14 and Theorem 22.6 imply that

$$
\|\boldsymbol{A} - \boldsymbol{A}_h\|_{\boldsymbol{H}(\mathrm{curl};D)} \le c \left(\sum_{K \in \mathcal{T}_h} h_K^{2r} (|\boldsymbol{A}|_{\boldsymbol{H}^r(K)} + \ell_D |\nabla \times \boldsymbol{A}|_{\boldsymbol{H}^r(K)})^2 \right)^{\frac{1}{2}},
$$

when $r > \frac{1}{2}$. The seminorm $|\cdot|_{\boldsymbol{H}^r(K)}$ has to be replaced by $|\cdot|_{\boldsymbol{H}^r(D_K)}$ whenever $r \le \frac{1}{2}$, where D_K is the set of the points composing the mesh cells sharing a degree of freedom with K. One can also extend the estimate (43.19) to the case of mixed boundary conditions by adapting the construction of the quasi-interpolation operator and of the commuting projection from Chapters 22 and 23. Finally, we refer the reader to Ciarlet [121, Prop. 4] for an alternative proof of (43.19). □

43.3.3 The duality argument

Recalling the material from §32.3, we would like to apply the Aubin–Nitsche duality argument to deduce an improved error estimate on $\|\boldsymbol{A} - \boldsymbol{A}_h\|_{\boldsymbol{L}^2(D)}$. It is at this point that we realize that the approach we have taken so far is too simplistic. To better understand the problem, let us consider the case $\partial D_{\mathrm{d}} = \partial D$. In this context, we have $\boldsymbol{V}_{\mathrm{d}} := \boldsymbol{H}_0(\mathrm{curl}; D)$ and $\boldsymbol{L} := \boldsymbol{L}^2(D)$, and Theorem 32.8 tells us that the Aubin–Nitsche argument provides a better rate of convergence in the $\boldsymbol{L}^2$-norm if and only if the embedding $\boldsymbol{H}_0(\mathrm{curl}; D) \hookrightarrow \boldsymbol{L}^2(D)$ is compact, which is not the case as shown in Exercise 43.1. The conclusion of this argumentation is that the estimates we have derived so far cannot yield an improved error estimate on $\|\boldsymbol{A} - \boldsymbol{A}_h\|_{\boldsymbol{L}^2(D)}$. A way around this obstacle is to find a space smaller than $\boldsymbol{H}_0(\mathrm{curl}; D)$, where the weak solution $\boldsymbol{A}$ lives and that embeds compactly into $\boldsymbol{L}^2(D)$, and to show that $\boldsymbol{A}_h$ is a convergent nonconforming approximation of $\boldsymbol{A}$ in that space. We are going to see in Chapter 44 that a good candidate is $\boldsymbol{H}_0(\mathrm{curl}; D) \cap \boldsymbol{H}(\mathrm{div}; D)$, as pointed out in Weber [391, Thm. 2.1-2.3]. Recall that the unknown field $\boldsymbol{A}$ stands for $\boldsymbol{E}$ or $\boldsymbol{H}$, and that the Gauss laws (43.1c)-(43.1d) combined with (43.3) imply that $\nabla \cdot (\epsilon \boldsymbol{E}) = \nabla \cdot \boldsymbol{D} = \rho$ and that $\nabla \cdot (\mu \boldsymbol{H}) = \nabla \cdot \boldsymbol{B} = 0$. Thus, it is reasonable to expect some control on the divergence of $\boldsymbol{A}$ and, therefore, to hope for an improved estimate on $\|\boldsymbol{A} - \boldsymbol{A}_h\|_{\boldsymbol{L}^2(D)}$ provided $\nabla \cdot \boldsymbol{A}_h$ is controlled in some sense. This question is addressed in Chapter 44.

Exercises

Exercise 43.1 (Compactness). Let $D := (0, 1)^3$ be the unit cube in $\mathbb{R}^3$. Show that the embedding $\boldsymbol{H}_0(\mathrm{curl}; D) \hookrightarrow \boldsymbol{L}^2(D)$ is not compact. (*Hint*: consider $\boldsymbol{v}_n := \nabla \phi_n$ with $\phi_n(x_1, x_2, x_3) := \frac{1}{n\pi} \sin(n\pi x_1) \sin(n\pi x_2) \sin(n\pi x_3)$, $n \ge 1$, and prove first that $(\boldsymbol{v}_n)_{n \ge 1}$ weakly converges to zero in $\boldsymbol{L}^2(D)$ (see Definition C.28), then compute $\|\boldsymbol{v}_n\|_{\boldsymbol{L}^2(D)}$ and argue by contradiction.)

Exercise 43.2 (Curl). (i) Let $\boldsymbol{v}$ be a smooth field. Show that $\|\nabla \times \boldsymbol{v}\|_{\ell^2}^2 \le 2\nabla \boldsymbol{v} : \nabla \boldsymbol{v}$. (*Hint*: relate $\nabla \times \boldsymbol{v}$ to the components of $(\nabla \boldsymbol{v} - \nabla \boldsymbol{v}^\mathsf{T})$.) (ii) Show that $\|\nabla \times \boldsymbol{v}\|_{\boldsymbol{L}^2(D)} \le |\boldsymbol{v}|_{\boldsymbol{H}^1(D)}$ for all $\boldsymbol{v} \in \boldsymbol{H}_0^1(D)$. (*Hint*: use an integration by parts.)

Exercise 43.3 (Property (43.12)**).** Prove the claim in Example 43.2, i.e., for $[\theta_{\min}, \theta_{\max}] \subset (-\pi, \pi)$ with $\delta := \theta_{\max} - \theta_{\min} < \pi$, letting $\theta := -\frac{1}{2}(\theta_{\min} + \theta_{\max})\frac{\pi}{2\pi - \delta}$, prove that $\theta \in (-\frac{\pi}{2}, \frac{\pi}{2})$ and $[\theta_{\min} + \theta, \theta_{\max} + \theta] \subset (-\frac{\pi}{2}, \frac{\pi}{2})$.

Exercise 43.4 (Dirichlet/Neumann). Let $\boldsymbol{v}$ be a smooth vector field in D such that $\boldsymbol{v}_{|\partial D_{\mathrm{d}}} \times \boldsymbol{n} = \boldsymbol{0}$. Prove that $(\nabla \times \boldsymbol{v})_{|\partial D_{\mathrm{d}}} \cdot \boldsymbol{n} = \boldsymbol{0}$. (*Hint*: compute $\int_D (\nabla \times \boldsymbol{v}) \cdot \nabla q \, \mathrm{d}x$ with q well chosen.)

Chapter 44

Maxwell's equations: control on the divergence

The analysis of Chapter 43 requires a coercivity property in $\boldsymbol{H}(\text{curl})$. There is, however, a loss of coercivity when the lower bound on the model parameter ν becomes very small. This situation occurs in the following two situations: (i) in the low frequency limit ($\omega \to 0$) when $\nu := \mathrm{i}\omega\mu$ as in the eddy current problem; (ii) if $\kappa \in \mathbb{R}$ and $\sigma \ll \omega\epsilon$ when $\nu := -\omega^2\epsilon + \mathrm{i}\omega\sigma$ as in the time-harmonic problem. We have also seen in Chapter 43 that a compactness property needs to be established to deduce an improved $\boldsymbol{L}^2$-error estimate by the duality argument. We show in this chapter that robust coercivity and compactness can be achieved by a weak control on the divergence of the discrete solution. The material of this chapter is based on [188].

44.1 Functional setting

In this section, we present the assumptions on the model problem and introduce a functional setting leading to a key smoothness result on the curl operator.

44.1.1 Model problem

We consider the model problem (43.9) on a Lipschitz domain D in $\mathbb{R}^3$. For simplicity, we restrict the scope to the homogeneous Dirichlet boundary condition $\boldsymbol{A}_{|\partial D}\times\boldsymbol{n} = \boldsymbol{0}$ (so that $\partial D_\mathrm{d} = \partial D$). The weak formulation is

$$
\begin{cases}
\text{Find } \boldsymbol{A} \in \boldsymbol{V}_0 := \boldsymbol{H}_0(\text{curl}; D) \text{ such that} \\
a_{\nu,\kappa}(\boldsymbol{A}, \boldsymbol{b}) = \ell(\boldsymbol{b}), \quad \forall \boldsymbol{b} \in \boldsymbol{V}_0,
\end{cases}
\tag{44.1}
$$

with $a_{\nu,\kappa}(\boldsymbol{a}, \boldsymbol{b}) := \int_D (\nu\boldsymbol{a}\cdot\overline{\boldsymbol{b}} + \kappa\nabla\times\boldsymbol{a}\cdot\nabla\times\overline{\boldsymbol{b}})\,\mathrm{d}x$ and $\ell(\boldsymbol{b}) := \int_D \boldsymbol{f}\cdot\overline{\boldsymbol{b}}\,\mathrm{d}x$. We assume that $\boldsymbol{f} \in \boldsymbol{L}^2(D)$ and that $\nabla\cdot\boldsymbol{f} = 0$. The divergence-free condition on $\boldsymbol{f}$ implies the following important property on the solution $\boldsymbol{A}$:

$$
\nabla\cdot(\nu\boldsymbol{A}) = 0.
\tag{44.2}
$$

Concerning the material properties ν and κ, we make the following assumptions: (i) Boundedness: $\nu, \kappa \in L^\infty(D; \mathbb{C})$ and we set $\nu_\sharp := \|\nu\|_{L^\infty(D;\mathbb{C})}$ and $\kappa_\sharp := \|\kappa\|_{L^\infty(D;\mathbb{C})}$. (ii) Rotated positivity:

there are real numbers θ, $\nu_\flat > 0$, and $\kappa_\flat > 0$ s.t. (43.12) is satisfied, i.e.,

$$\underset{\boldsymbol{x} \in D}{\mathrm{ess\,inf}}\, \Re\big(e^{i\theta}\nu(\boldsymbol{x})\big) \geq \nu_\flat, \qquad \underset{\boldsymbol{x} \in D}{\mathrm{ess\,inf}}\, \Re\big(e^{i\theta}\kappa(\boldsymbol{x})\big) \geq \kappa_\flat. \tag{44.3}$$

We define the contrast factors $\nu_{\sharp/\flat} := \frac{\nu_\sharp}{\nu_\flat}$ and $\kappa_{\sharp/\flat} := \frac{\kappa_\sharp}{\kappa_\flat}$. We also define the magnetic Reynolds number $\gamma_{\nu,\kappa} := \nu_\sharp \ell_D^2 \kappa_\sharp^{-1}$. Several magnetic Reynolds numbers can be defined if the material is highly contrasted, but we will not explore this situation further. (iii) Piecewise smoothness: there is a partition of D into M disjoint Lipschitz polyhedra $\{D_m\}_{m \in \{1:M\}}$ s.t. $\nu_{|D_m}, \kappa_{|D_m} \in W^{1,\infty}(D_m)$ for all $m \in \{1:M\}$. The reader who is not comfortable with this assumption may think of ν, κ being constant without missing anything essential in the analysis.

44.1.2 A key smoothness result on the curl operator

Let us define the (complex-valued) functional spaces

$$M_0 := H_0^1(D), \qquad M_* := \{q \in H^1(D) \mid (q,1)_{L^2(D)} = 0\}, \tag{44.4}$$

as well as the following subspaces of $\boldsymbol{H}(\mathrm{curl}; D)$:

$$\boldsymbol{X}_{0\nu} := \{\boldsymbol{b} \in \boldsymbol{H}_0(\mathrm{curl}; D) \mid (\nu\boldsymbol{b}, \nabla m)_{\boldsymbol{L}^2(D)} = 0,\ \forall m \in M_0\}, \tag{44.5a}$$

$$\boldsymbol{X}_{*\kappa^{-1}} := \{\boldsymbol{b} \in \boldsymbol{H}(\mathrm{curl}; D) \mid (\kappa^{-1}\boldsymbol{b}, \nabla m)_{\boldsymbol{L}^2(D)} = 0,\ \forall m \in M_*\}, \tag{44.5b}$$

where $(\cdot,\cdot)_{\boldsymbol{L}^2(D)}$ denotes the inner product in $\boldsymbol{L}^2(D)$. The main motivation for introducing the above subspaces is that $\boldsymbol{A} \in \boldsymbol{X}_{0\nu}$ owing to (44.2). Moreover, we will see below that $\kappa\nabla{\times}\boldsymbol{A} \in \boldsymbol{X}_{*\kappa^{-1}}$. Taking $m \in C_0^\infty(D)$ in (44.5a) shows that for all $\boldsymbol{b} \in \boldsymbol{X}_{0\nu}$, the field $\nu\boldsymbol{b}$ has a weak divergence in $L^2(D)$ and $\nabla{\cdot}(\nu\boldsymbol{b}) = 0$. Similarly, the definition (44.5b) implies that for all $\boldsymbol{b} \in \boldsymbol{X}_{*\kappa^{-1}}$, the field $\kappa^{-1}\boldsymbol{b}$ has a weak divergence in $L^2(D)$ and $\nabla{\cdot}(\kappa^{-1}\boldsymbol{b}) = 0$. Invoking the integration by parts formula (4.12) and the surjectivity of the trace map $\gamma^{\mathrm{g}} : H^1(D) \to H^{\frac{1}{2}}(\partial D)$ then shows that $\gamma^{\mathrm{d}}(\kappa^{-1}\boldsymbol{b}) = 0$ for all $\boldsymbol{b} \in \boldsymbol{X}_{*\kappa^{-1}}$, where γ^{d} is the normal trace operator (recall that $\gamma^{\mathrm{d}}(\boldsymbol{v}) = \boldsymbol{v}_{|\partial D}{\cdot}\boldsymbol{n}$ if the field $\boldsymbol{v}$ is smooth).

Let us first state a simple result related to the Helmholtz decomposition of vector fields in $\boldsymbol{V}_0 := \boldsymbol{H}_0(\mathrm{curl}; D)$ using the subspace $\boldsymbol{X}_{0\nu}$ (a similar result is available on $\boldsymbol{H}(\mathrm{curl}; D)$ using the subspace $\boldsymbol{X}_{*\kappa^{-1}}$).

Lemma 44.1 (Helmholtz decomposition). *The following holds true:*

$$\boldsymbol{V}_0 = \boldsymbol{X}_{0\nu} \oplus \nabla M_0. \tag{44.6}$$

Proof. Let $\boldsymbol{b} \in \boldsymbol{V}_0$ and let $p \in M_0$ solve $(\nu\nabla p, \nabla q)_{\boldsymbol{L}^2(D)} = (\nu\boldsymbol{b}, \nabla q)_{\boldsymbol{L}^2(D)}$ for all $q \in M_0$. Our assumptions on ν imply that there is a unique solution to this problem. Then we set $\boldsymbol{v} := \boldsymbol{b} - \nabla p$ and observe that $\boldsymbol{v} \in \boldsymbol{X}_{0\nu}$. The sum is direct because if $\boldsymbol{0} = \boldsymbol{v} + \nabla p$, then the identity $\int_D \nu\nabla p{\cdot}\overline{\boldsymbol{v}}\,\mathrm{d}x = 0$, which holds true for all $p \in M_0$ and all $\boldsymbol{v} \in \boldsymbol{X}_{0\nu}$, implies that $\nabla p = \boldsymbol{0} = \boldsymbol{v}$. $\qquad\square$

We can now state the main result of this section. This result extends Lemma 43.3 to heterogeneous domains. Given a smoothness index $s > 0$, we set $\|\boldsymbol{b}\|_{\boldsymbol{H}^s(D)} := (\|\boldsymbol{b}\|_{\boldsymbol{L}^2(D)}^2 + \ell_D^{2s}|\boldsymbol{b}|_{\boldsymbol{H}^s(D)}^2)^{\frac{1}{2}}$, where ℓ_D is some characteristic length of D, e.g., $\ell_D := \mathrm{diam}(D)$.

Lemma 44.2 (Regularity pickup). *Let D be a Lipschitz domain in $\mathbb{R}^3$. (i) Assume that the boundary ∂D is connected and that ν is piecewise smooth. There exist $s > 0$ and $\check{C} > 0$ (depending on D and the contrast factor $\nu_{\sharp/\flat}$ but not on $\nu_\flat$ alone) such that*

$$\check{C}\ell_D^{-1}\|\boldsymbol{b}\|_{\boldsymbol{H}^s(D)} \leq \|\nabla{\times}\boldsymbol{b}\|_{\boldsymbol{L}^2(D)}, \qquad \forall \boldsymbol{b} \in \boldsymbol{X}_{0\nu}. \tag{44.7}$$

(ii) *Assume that D is simply connected and that κ is piecewise smooth. There exist $s' > 0$ and $\check{C}' > 0$ (depending on D and the contrast factor $\kappa_{\sharp/\flat}$ but not on $\kappa_\flat$ alone) such that*

$$\check{C}'\ell_D^{-1}\|\boldsymbol{b}\|_{\boldsymbol{H}^{s'}(D)} \leq \|\nabla\times\boldsymbol{b}\|_{\boldsymbol{L}^2(D)}, \qquad \forall\boldsymbol{b} \in \boldsymbol{X}_{*\kappa^{-1}}. \tag{44.8}$$

Proof. See Jochmann [259], Bonito et al. [70]. □

Remark 44.3 (Smoothness index). There are some situations where the smoothness indices s, s' can be larger than $\frac{1}{2}$. One example is that of isolated inclusions in an otherwise homogeneous material. We refer the reader to Ciarlet [121, §5.2] for further insight and examples. □

Lemma 44.2 has two important consequences. First, by restricting the smoothness index s to zero in (44.7), we obtain the following important stability result on the curl operator.

Lemma 44.4 (Poincaré–Steklov). *Assume that the boundary ∂D is connected and that ν is piecewise smooth. There is $\hat{C}_{\mathrm{PS}} > 0$ (depending on D and the contrast factor $\nu_{\sharp/\flat}$) such that the following Poincaré–Steklov inequality holds true:*

$$\hat{C}_{\mathrm{PS}}\ell_D^{-1}\|\boldsymbol{b}\|_{\boldsymbol{L}^2(D)} \leq \|\nabla\times\boldsymbol{b}\|_{\boldsymbol{L}^2(D)}, \qquad \forall\boldsymbol{b} \in \boldsymbol{X}_{0\nu}. \tag{44.9}$$

The bound (44.9) is what we need to establish a coercivity property on $\boldsymbol{X}_{0\nu}$ that is robust w.r.t. $\nu_\flat$. Indeed, we have

$$\begin{aligned}
\Re\big(e^{\mathrm{i}\theta}a_{\nu,\kappa}(\boldsymbol{b},\boldsymbol{b})\big) &\geq \nu_\flat\|\boldsymbol{b}\|_{\boldsymbol{L}^2(D)}^2 + \kappa_\flat\|\nabla\times\boldsymbol{b}\|_{\boldsymbol{L}^2(D)}^2 \geq \kappa_\flat\|\nabla\times\boldsymbol{b}\|_{\boldsymbol{L}^2(D)}^2 \\
&\geq \frac{1}{2}\kappa_\flat\big(\|\nabla\times\boldsymbol{b}\|_{\boldsymbol{L}^2(D)}^2 + \hat{C}_{\mathrm{PS}}^2\ell_D^{-2}\|\boldsymbol{b}\|_{\boldsymbol{L}^2(D)}^2\big) \\
&\geq \frac{1}{2}\kappa_\flat\ell_D^{-2}\min(1,\hat{C}_{\mathrm{PS}}^2)\|\boldsymbol{b}\|_{\boldsymbol{H}(\mathrm{curl};D)}^2, \tag{44.10}
\end{aligned}$$

for all $\boldsymbol{b} \in \boldsymbol{X}_{0\nu}$, where we recall that $\boldsymbol{H}(\mathrm{curl};D)$ is equipped with the norm $\|\boldsymbol{b}\|_{\boldsymbol{H}(\mathrm{curl};D)} := (\|\boldsymbol{b}\|_{\boldsymbol{L}^2(D)}^2 + \ell_D^2\|\nabla\times\boldsymbol{b}\|_{\boldsymbol{L}^2(D)}^2)^{\frac{1}{2}}$. This shows that the sesquilinear form $a_{\nu,\kappa}$ is coercive on $\boldsymbol{X}_{0\nu}$ with a coercivity constant depending on the contrast factor $\nu_{\sharp/\flat}$ but not on $\nu_\flat$ alone (whereas the coercivity constant on the larger space $\boldsymbol{V}_0$ is $\min(\nu_\flat, \ell_D^{-2}\kappa_\flat)$ (see (43.13a))).

Let us now examine the consequences of Lemma 44.2 on the Sobolev smoothness index of $\boldsymbol{A}$ and $\nabla\times\boldsymbol{A}$. Owing to (44.7), there is $s > 0$ s.t. $\boldsymbol{A} \in \boldsymbol{H}^s(D)$. We will see in §44.3 that the embedding $\boldsymbol{H}^s(D) \hookrightarrow \boldsymbol{L}^2(D)$ is the compactness property that we need to apply the duality argument and derive an improved $\boldsymbol{L}^2$-error estimate. Furthermore, the field $\boldsymbol{R} := \kappa\nabla\times\boldsymbol{A}$ is in $\boldsymbol{X}_{*\kappa^{-1}}$ (notice in particular that $\nabla\times\boldsymbol{R} = \boldsymbol{f} - \nu\boldsymbol{A} \in \boldsymbol{L}^2(D)$), so that we deduce from (44.8) that there is $s' > 0$ s.t. $\boldsymbol{R} \in \boldsymbol{H}^{s'}(D)$. In addition, the material property κ being piecewise smooth, we infer that the following multiplier property holds true (see [259, Lem. 2] and [70, Prop. 2.1]): There exists $\tau > 0$ and $C_{\kappa^{-1}}$ s.t.

$$|\kappa^{-1}\boldsymbol{\xi}|_{\boldsymbol{H}^{\tau'}(D)} \leq C_{\kappa^{-1}}|\boldsymbol{\xi}|_{\boldsymbol{H}^{\tau'}(D)}, \qquad \forall\boldsymbol{\xi} \in \boldsymbol{H}^\tau(D), \quad \forall\tau' \in [0,\tau]. \tag{44.11}$$

Letting $s'' := \min(s',\tau) > 0$, we conclude that $\nabla\times\boldsymbol{A} \in \boldsymbol{H}^{s''}(D)$.

44.2 Coercivity revisited for edge elements

In this section, we revisit the $\boldsymbol{H}(\mathrm{curl})$-error analysis for the approximation of the weak problem (44.1) using Nédélec (or edge) elements (see Chapters 15 and 19). The key tool we are going

to use is a discrete counterpart of the Poincaré–Steklov inequality (44.9). We consider a shape-regular sequence of affine meshes $(\mathcal{T}_h)_{h\in\mathcal{H}}$ of D. We assume that D is a Lipschitz polyhedron and that each mesh covers D exactly.

44.2.1   Discrete Poincaré–Steklov inequality

Let $\boldsymbol{V}_{h0}$ be the $\boldsymbol{H}_0(\mathrm{curl})$-conforming space using Nédélec elements of order $k \geq 0$ defined by

$$\boldsymbol{V}_{h0} := \boldsymbol{P}^{\mathrm{c}}_{k,0}(\mathcal{T}_h) := \{\boldsymbol{b}_h \in \boldsymbol{P}^{\mathrm{c}}_k(\mathcal{T}_h) \mid \boldsymbol{b}_{h|\partial D}\times\boldsymbol{n} = \boldsymbol{0}\}. \tag{44.12}$$

Observe that the Dirichlet condition is enforced strongly in $\boldsymbol{V}_{h0}$. The discrete problem is formulated as follows:

$$\begin{cases} \text{Find } \boldsymbol{A}_h \in \boldsymbol{V}_{h0} \text{ such that} \\ a_{\nu,\kappa}(\boldsymbol{A}_h, \boldsymbol{b}_h) = \ell(\boldsymbol{b}_h), \quad \forall\boldsymbol{b}_h \in \boldsymbol{V}_{h0}. \end{cases} \tag{44.13}$$

Since it is not reasonable to consider the space $\{\boldsymbol{b}_h \in \boldsymbol{V}_{h0} \mid \nabla{\cdot}(\nu\boldsymbol{b}_h) = 0\}$, because the normal component of $\nu\boldsymbol{b}_h$ may jump across the mesh interfaces, we are going to consider instead the subspace

$$\boldsymbol{X}_{h0\nu} := \{\boldsymbol{b}_h \in \boldsymbol{V}_{h0} \mid (\nu\boldsymbol{b}_h, \nabla m_h)_{\boldsymbol{L}^2(D)} = 0, \; \forall m_h \in M_{h0}\}, \tag{44.14}$$

where $M_{h0} := P^{\mathrm{g}}_{k+1,0}(\mathcal{T}_h; \mathbb{C})$ is conforming in $H^1_0(D; \mathbb{C})$. Note that the polynomial degrees of the finite element spaces M_{h0} and $\boldsymbol{V}_{h0}$ are compatible in the sense that $\nabla M_{h0} \subset \boldsymbol{V}_{h0}$. Using this property and proceeding as in Lemma 44.1 proves the following discrete Helmholtz decomposition:

$$\boldsymbol{V}_{h0} = \boldsymbol{X}_{h0\nu} \oplus \nabla M_{h0}. \tag{44.15}$$

Lemma 44.5 (Discrete solution). *Let $\boldsymbol{A}_h \in \boldsymbol{V}_{h0}$ be the unique solution to* (44.13). *Then $\boldsymbol{A}_h \in \boldsymbol{X}_{h0\nu}$.*

Proof. We must show that $(\nu\boldsymbol{A}_h, \nabla m_h)_{\boldsymbol{L}^2(D)} = 0$ for all $m_h \in M_{h0}$. Since $\nabla m_h \in \nabla M_{h0} \subset \boldsymbol{V}_{h0}$, ∇m_h is an admissible test function in (44.13). Recalling that $\nabla{\cdot}\boldsymbol{f} = 0$, we infer that

$$0 = \ell(\nabla m_h) = a_{\nu,\kappa}(\boldsymbol{A}_h, \nabla m_h) = (\nu\boldsymbol{A}_h, \nabla m_h)_{\boldsymbol{L}^2(D)},$$

since $\nabla{\times}(\nabla m_h) = \boldsymbol{0}$. This completes the proof.   □

We now establish a discrete counterpart to the Poincaré–Steklov inequality (44.9). This result is not straightforward since $\boldsymbol{X}_{h0\nu}$ is not a subspace of $\boldsymbol{X}_{0\nu}$. The key tool that we are going to invoke is the stable commuting quasi-interpolation projections from §23.3.3.

Theorem 44.6 (Discrete Poincaré–Steklov). *Under the assumptions of Lemma* 44.4, *there is a constant $\hat{C}'_{\mathrm{PS}} > 0$ (depending on $\hat{C}_{\mathrm{PS}}$, the polynomial degree k, the regularity of the mesh sequence, and the contrast factor $\nu_{\sharp/\flat}$, but not on $\nu_\flat$ alone) s.t. for all $\boldsymbol{x}_h \in \boldsymbol{X}_{h0\nu}$ and all $h \in \mathcal{H}$,*

$$\hat{C}'_{\mathrm{PS}}\ell_D^{-1}\|\boldsymbol{x}_h\|_{\boldsymbol{L}^2(D)} \leq \|\nabla{\times}\boldsymbol{x}_h\|_{\boldsymbol{L}^2(D)}. \tag{44.16}$$

Proof. Let $\boldsymbol{x}_h \in \boldsymbol{X}_{h0\nu}$ be a nonzero discrete field. Let $\phi(\boldsymbol{x}_h) \in M_0 := H^1_0(D)$ be the solution to the following well-posed Poisson problem:

$$(\nu\nabla\phi(\boldsymbol{x}_h), \nabla m)_{\boldsymbol{L}^2(D)} = (\nu\boldsymbol{x}_h, \nabla m)_{\boldsymbol{L}^2(D)}, \qquad \forall m \in M_0.$$

Let us define the *curl-preserving lifting* of $\boldsymbol{x}_h$ s.t. $\boldsymbol{\xi}(\boldsymbol{x}_h) := \boldsymbol{x}_h - \nabla\phi(\boldsymbol{x}_h)$, and let us notice that $\boldsymbol{\xi}(\boldsymbol{x}_h) \in \boldsymbol{X}_{0\nu}$. Upon invoking the quasi-interpolation operators $\mathcal{J}^{\mathrm{c}}_{h0}$ and $\mathcal{J}^{\mathrm{d}}_{h0}$ introduced in §23.3.3, we observe that

$$\boldsymbol{x}_h - \mathcal{J}^{\mathrm{c}}_{h0}(\boldsymbol{\xi}(\boldsymbol{x}_h)) = \mathcal{J}^{\mathrm{c}}_{h0}(\boldsymbol{x}_h - \boldsymbol{\xi}(\boldsymbol{x}_h)) = \mathcal{J}^{\mathrm{c}}_{h0}(\nabla(\phi(\boldsymbol{x}_h))) = \nabla(\mathcal{J}^{\mathrm{g}}_{h0}(\phi(\boldsymbol{x}_h))),$$

where we used that $\mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{x}_h) = \boldsymbol{x}_h$ and the commuting properties of $\mathcal{J}_{h0}^{\mathrm{g}}$ and $\mathcal{J}_{h0}^{\mathrm{c}}$. Since $\boldsymbol{x}_h \in \boldsymbol{X}_{h0\nu}$, we infer that $(\nu\boldsymbol{x}_h, \nabla(\mathcal{J}_{h0}^{\mathrm{g}}(\phi(\boldsymbol{x}_h))))_{\boldsymbol{L}^2(D)} = 0$, so that

$$(\nu\boldsymbol{x}_h, \boldsymbol{x}_h)_{\boldsymbol{L}^2(D)} = (\nu\boldsymbol{x}_h, \boldsymbol{x}_h - \mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{\xi}(\boldsymbol{x}_h)))_{\boldsymbol{L}^2(D)} + (\nu\boldsymbol{x}_h, \mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{\xi}(\boldsymbol{x}_h)))_{\boldsymbol{L}^2(D)}$$
$$= (\nu\boldsymbol{x}_h, \mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{\xi}(\boldsymbol{x}_h)))_{\boldsymbol{L}^2(D)}.$$

Multiplying by $e^{\mathrm{i}\theta}$, taking the real part, and using the Cauchy–Schwarz inequality, we infer that

$$\nu_\flat\|\boldsymbol{x}_h\|_{\boldsymbol{L}^2(D)}^2 \leq \nu_\sharp\|\boldsymbol{x}_h\|_{\boldsymbol{L}^2(D)}\|\mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{\xi}(\boldsymbol{x}_h))\|_{\boldsymbol{L}^2(D)}.$$

The uniform boundedness of $\mathcal{J}_{h0}^{\mathrm{c}}$ on $\boldsymbol{L}^2(D)$, together with the Poincaré–Steklov inequality (44.9) and the identity $\nabla{\times}\boldsymbol{\xi}(\boldsymbol{x}_h) = \nabla{\times}\boldsymbol{x}_h$, implies that

$$\|\mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{\xi}(\boldsymbol{x}_h))\|_{\boldsymbol{L}^2(D)} \leq \|\mathcal{J}_{h0}^{\mathrm{c}}\|_{\mathcal{L}(\boldsymbol{L}^2;\boldsymbol{L}^2)}\|\boldsymbol{\xi}(\boldsymbol{x}_h)\|_{\boldsymbol{L}^2(D)}$$
$$\leq \|\mathcal{J}_{h0}^{\mathrm{c}}\|_{\mathcal{L}(\boldsymbol{L}^2;\boldsymbol{L}^2)}\hat{C}_{\mathrm{PS}}^{-1}\ell_D\|\nabla{\times}\boldsymbol{x}_h\|_{\boldsymbol{L}^2(D)},$$

so that (44.16) holds true with $\hat{C}_{\mathrm{PS}}' := \nu_{\sharp/\flat}^{-1}\|\mathcal{J}_{h0}^{\mathrm{c}}\|_{\mathcal{L}(\boldsymbol{L}^2;\boldsymbol{L}^2)}^{-1}\hat{C}_{\mathrm{PS}}$. □

Remark 44.7 (Literature). There are many ways to prove the discrete Poincaré–Steklov inequality (44.16). One route described in Hiptmair [244, §4.2] consists of invoking subtle regularity estimates from Amrouche et al. [10, Lem. 4.7]. Another one, which avoids invoking regularity estimates, is based on an argument by Kikuchi [267] which is often called *discrete compactness*; see also Monk and Demkowicz [304], Caorsi et al. [106]. The proof is not constructive and is based on an argument by contradiction. The technique used in the proof of Theorem 44.6, inspired from Arnold et al. [23, Thm. 5.11] and Arnold et al. [26, Thm. 3.6], is more recent, and uses the stable commuting quasi-interpolation projections $\mathcal{J}_h^{\mathrm{c}}$ and $\mathcal{J}_{h0}^{\mathrm{c}}$. It was already observed in Boffi [61] that the existence of stable commuting quasi-interpolation operators would imply the discrete compactness property. □

44.2.2 $\boldsymbol{H}(\mathrm{curl})$-error analysis

We are now in a position to revisit the error analysis of §43.3. Let us first show that $\boldsymbol{X}_{h0\nu}$ has the same approximation properties as $\boldsymbol{V}_{h0}$ in $\boldsymbol{X}_{0\nu}$.

Lemma 44.8 (Approximation in $\boldsymbol{X}_{h0\nu}$). *There is c, uniform w.r.t. the model parameters, s.t. for all $\boldsymbol{A} \in \boldsymbol{X}_{0\nu}$ and all $h \in \mathcal{H}$,*

$$\inf_{\boldsymbol{x}_h \in \boldsymbol{X}_{h0\nu}} \|\boldsymbol{A} - \boldsymbol{x}_h\|_{\boldsymbol{H}(\mathrm{curl};D)} \leq c\,\nu_{\sharp/\flat} \inf_{\boldsymbol{b}_h \in \boldsymbol{V}_{h0}} \|\boldsymbol{A} - \boldsymbol{b}_h\|_{\boldsymbol{H}(\mathrm{curl};D)}. \tag{44.17}$$

Proof. Let $\boldsymbol{A} \in \boldsymbol{X}_{0\nu}$. We start by computing the Helmholtz decomposition of $\mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{A})$ in $\boldsymbol{V}_{h0}$ as stated in (44.15). Let $p_h \in M_{h0}$ be the unique solution to the discrete Poisson problem $(\nu\nabla p_h, \nabla q_h)_{\boldsymbol{L}^2(D)} = (\nu\mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{A}), \nabla q_h)_{\boldsymbol{L}^2(D)}$ for all $q_h \in M_{h0}$. Let us define $\boldsymbol{y}_h := \mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{A}) - \nabla p_h$. By construction, $\boldsymbol{y}_h \in \boldsymbol{X}_{h0\nu}$ and $\nabla{\times}\boldsymbol{y}_h = \nabla{\times}\mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{A})$. Hence, $\|\nabla{\times}(\boldsymbol{A} - \boldsymbol{y}_h)\|_{\boldsymbol{L}^2(D)} = \|\nabla{\times}(\boldsymbol{A} - \mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{A}))\|_{\boldsymbol{L}^2(D)}$. Since $\nabla{\cdot}(\nu\boldsymbol{A}) = 0$, we also infer that

$$(\nu\nabla p_h, \nabla p_h)_{\boldsymbol{L}^2(D)} = (\nu\mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{A}), \nabla p_h)_{\boldsymbol{L}^2(D)} = (\nu(\mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{A}) - \boldsymbol{A}), \nabla p_h)_{\boldsymbol{L}^2(D)},$$

which in turn implies that $\|\nabla p_h\|_{\boldsymbol{L}^2(D)} \leq \nu_{\sharp/\flat}\|\mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{A}) - \boldsymbol{A}\|_{\boldsymbol{L}^2(D)}$. The above argument shows that

$$\|\boldsymbol{A} - \boldsymbol{y}_h\|_{\boldsymbol{L}^2(D)} \leq \|\boldsymbol{A} - \mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{A})\|_{\boldsymbol{L}^2(D)} + \|\mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{A}) - \boldsymbol{y}_h\|_{\boldsymbol{L}^2(D)}$$
$$\leq \|\boldsymbol{A} - \mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{A})\|_{\boldsymbol{L}^2(D)} + \|\nabla p_h\|_{\boldsymbol{L}^2(D)}$$
$$\leq (1 + \nu_{\sharp/\flat})\|\boldsymbol{A} - \mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{A})\|_{\boldsymbol{L}^2(D)}.$$

In conclusion, we have proved that

$$\inf_{\boldsymbol{x}_h \in \boldsymbol{X}_{h0\nu}} \|\boldsymbol{A} - \boldsymbol{x}_h\|_{\boldsymbol{H}(\mathrm{curl};D)} \leq \|\boldsymbol{A} - \boldsymbol{y}_h\|_{\boldsymbol{H}(\mathrm{curl};D)} \leq (1 + \nu_{\sharp/\flat})\|\boldsymbol{A} - \mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{A})\|_{\boldsymbol{H}(\mathrm{curl};D)}.$$

Invoking the commutation and approximation properties of the quasi-interpolation operators, we infer that

$$\begin{aligned}
\|\boldsymbol{A} - \mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{A})\|_{\boldsymbol{H}(\mathrm{curl};D)}^2 &= \|\boldsymbol{A} - \mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{A})\|_{\boldsymbol{L}^2(D)}^2 + \ell_D^2 \|\nabla{\times}(\boldsymbol{A} - \mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{A}))\|_{\boldsymbol{L}^2(D)}^2 \\
&= \|\boldsymbol{A} - \mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{A})\|_{\boldsymbol{L}^2(D)}^2 + \ell_D^2 \|\nabla{\times}\boldsymbol{A} - \mathcal{J}_{h0}^{\mathrm{d}}(\nabla{\times}\boldsymbol{A})\|_{\boldsymbol{L}^2(D)}^2 \\
&\leq c \inf_{\boldsymbol{b}_h \in \boldsymbol{P}_0^{\mathrm{c}}(\mathcal{T}_h)} \|\boldsymbol{A} - \boldsymbol{b}_h\|_{\boldsymbol{L}^2(D)}^2 + c'\ell_D^2 \inf_{\boldsymbol{d}_h \in \boldsymbol{P}_0^{\mathrm{d}}(\mathcal{T}_h)} \|\nabla{\times}\boldsymbol{A} - \boldsymbol{d}_h\|_{\boldsymbol{L}^2(D)}^2 \\
&\leq c \inf_{\boldsymbol{b}_h \in \boldsymbol{P}_0^{\mathrm{c}}(\mathcal{T}_h)} \|\boldsymbol{A} - \boldsymbol{b}_h\|_{\boldsymbol{L}^2(D)}^2 + c'\ell_D^2 \inf_{\boldsymbol{b}_h \in \boldsymbol{P}_0^{\mathrm{c}}(\mathcal{T}_h)} \|\nabla{\times}(\boldsymbol{A} - \boldsymbol{b}_h)\|_{\boldsymbol{L}^2(D)}^2,
\end{aligned}$$

where the last bound follows by restricting the minimization set to $\nabla{\times}\boldsymbol{P}_0^{\mathrm{c}}(\mathcal{T}_h)$ since $\nabla{\times}\boldsymbol{P}_0^{\mathrm{c}}(\mathcal{T}_h) \subset \boldsymbol{P}_0^{\mathrm{d}}(\mathcal{T}_h)$. The conclusion follows readily. $\qquad\square$

Theorem 44.9 ($\boldsymbol{H}$(curl)-error estimate). *Let $\boldsymbol{A}$ solve (44.1) and let $\boldsymbol{A}_h$ solve (44.13). Assume that ∂D is connected and that ν is piecewise smooth. There is c, which depends on the discrete Poincaré–Steklov constant $\hat{C}_{\mathrm{PS}}'$ and the contrast factors $\nu_{\sharp/\flat}$ and $\kappa_{\sharp/\flat}$, s.t. for all $h \in \mathcal{H}$,*

$$\|\boldsymbol{A} - \boldsymbol{A}_h\|_{\boldsymbol{H}(\mathrm{curl};D)} \leq c\,\hat{\gamma}_{\nu,\kappa} \inf_{\boldsymbol{b}_h \in \boldsymbol{V}_{h0}} \|\boldsymbol{A} - \boldsymbol{b}_h\|_{\boldsymbol{H}(\mathrm{curl};D)}, \tag{44.18}$$

with $\hat{\gamma}_{\nu,\kappa} := \max(1, \gamma_{\nu,\kappa})$ and the magnetic Reynolds number $\gamma_{\nu,\kappa} := \nu_\sharp \ell_D^2 \kappa_\sharp^{-1}$.

Proof. Owing to Lemma 44.5, $\boldsymbol{A}_h$ also solves the following problem: Find $\boldsymbol{A}_h \in \boldsymbol{X}_{h0\nu}$ s.t.

$$a_{\nu,\kappa}(\boldsymbol{A}_h, \boldsymbol{x}_h) = \ell(\boldsymbol{x}_h), \forall \boldsymbol{x}_h \in \boldsymbol{X}_{h0\nu}.$$

Using the discrete Poincaré–Steklov inequality (44.16) and proceeding as in (44.10), we infer that

$$\Re\big(e^{\mathrm{i}\theta} a_{\nu,\kappa}(\boldsymbol{x}_h, \boldsymbol{x}_h)\big) \geq \frac{1}{2}\kappa_\flat \ell_D^{-2} \min(1, (\hat{C}_{\mathrm{PS}}')^2)\|\boldsymbol{x}_h\|_{\boldsymbol{H}(\mathrm{curl};D)}^2,$$

for all $\boldsymbol{x}_h \in \boldsymbol{X}_{h0\nu}$. Hence, the above problem is well-posed. Recalling the boundedness property (43.13b) of the sesquilinear form $a_{\nu,\kappa}$ and invoking the abstract error estimate (26.18) leads to

$$\|\boldsymbol{A} - \boldsymbol{A}_h\|_{\boldsymbol{H}(\mathrm{curl};D)} \leq \frac{2\max(\nu_\sharp, \ell_D^{-2}\kappa_\sharp)}{\kappa_\flat \ell_D^{-2} \min(1, (\hat{C}_{\mathrm{PS}}')^2)} \inf_{\boldsymbol{x}_h \in \boldsymbol{X}_{h0\nu}} \|\boldsymbol{A} - \boldsymbol{x}_h\|_{\boldsymbol{H}(\mathrm{curl};D)}.$$

We conclude the proof by invoking Lemma 44.8. $\qquad\square$

Remark 44.10 (Neumann boundary condition). The above analysis can be adapted to handle the Neumann condition $(\kappa\nabla{\times}\boldsymbol{A})_{|\partial D}{\times}\boldsymbol{n} = \boldsymbol{0}$; see Exercise 44.3. This condition implies that $(\nabla{\times}(\kappa\nabla{\times}\boldsymbol{A}))_{|\partial D}{\cdot}\boldsymbol{n} = 0$. Moreover, assuming $\boldsymbol{f}_{|\partial D}{\cdot}\boldsymbol{n} = 0$ and taking the normal component of the equation $\nu\boldsymbol{A} + \nabla{\times}(\kappa\nabla{\times}\boldsymbol{A}) = \boldsymbol{f}$ at the boundary gives $\boldsymbol{A}_{|\partial D}{\cdot}\boldsymbol{n} = 0$. Since $\nabla{\cdot}\boldsymbol{f} = 0$, we also have $\nabla{\cdot}(\nu\boldsymbol{A}) = 0$. In other words, we have

$$\boldsymbol{A} \in \boldsymbol{X}_{*\nu} := \{\boldsymbol{b} \in \boldsymbol{H}(\mathrm{curl}; D) \mid (\nu\boldsymbol{b}, \nabla m)_{\boldsymbol{L}^2(D)} = 0, \ \forall m \in M_*\}.$$

The discrete spaces are now $\boldsymbol{V}_h := \boldsymbol{P}_k^{\mathrm{c}}(\mathcal{T}_h)$ and $M_{h*} := P_{k+1}^{\mathrm{g}}(\mathcal{T}_h; \mathbb{C}) \cap M_*$. Using $\boldsymbol{V}_h$ for the discrete trial and test spaces, we infer that

$$\boldsymbol{A}_h \in \boldsymbol{X}_{h*\nu} := \{\boldsymbol{b}_h \in \boldsymbol{V}_h \mid (\nu\boldsymbol{b}_h, \nabla m_h)_{\boldsymbol{L}^2(D)} = 0, \ \forall m_h \in M_{h*}\}.$$

The Poincaré–Steklov inequality (44.16) still holds true provided the assumption that ∂D is connected is replaced by the assumption that D is simply connected. The error analysis from Theorem 44.9 can be readily adapted. $\qquad\square$

44.3   The duality argument for edge elements

Our goal is to derive an improved error estimate in the $\boldsymbol{L}^2$-norm using a duality argument that invokes a weak control on the divergence. The subtlety is that, as already mentioned, the setting is nonconforming since $\boldsymbol{X}_{h0\nu}$ is not a subspace of $\boldsymbol{X}_{0\nu}$. We assume in the section that the boundary ∂D is connected and that the domain D is simply connected. Recalling the smoothness indices $s, s' > 0$ from Lemma 44.2 together with the index $\tau > 0$ from the multiplier property (44.11) and letting $s'' := \min(s', \tau)$, we have $\boldsymbol{A} \in \boldsymbol{H}^s(D)$ and $\nabla{\times}\boldsymbol{A} \in \boldsymbol{H}^{s''}(D)$ with $s, s'' > 0$. In what follows, we set

$$\sigma := \min(s, s''). \tag{44.19}$$

Let us first start with an approximation result on the curl-preserving lifting operator $\boldsymbol{\xi} : \boldsymbol{X}_{h0\nu} \to \boldsymbol{X}_{0\nu}$ defined in the proof of Theorem 44.6. Recall that for all $\boldsymbol{x}_h \in \boldsymbol{X}_{h0\nu}$, the field $\boldsymbol{\xi}(\boldsymbol{x}_h) \in \boldsymbol{X}_{0\nu}$ is s.t. $\boldsymbol{\xi}(\boldsymbol{x}_h) := \boldsymbol{x}_h - \nabla\phi(\boldsymbol{x}_h)$ with $\phi(\boldsymbol{x}_h) \in H^1_0(D)$, implying that $\nabla{\times}\boldsymbol{\xi}(\boldsymbol{x}_h) = \nabla{\times}\boldsymbol{x}_h$.

Lemma 44.11 (Curl-preserving lifting). *Let $s > 0$ be the smoothness index introduced in (44.7). There is c, depending on the constant $\check{C}_D$ from (44.7) and the contrast factor $\nu_{\sharp/\flat}$, s.t. for all $\boldsymbol{x}_h \in \boldsymbol{X}_{h0\nu}$ and all $h \in \mathcal{H}$,*

$$\|\boldsymbol{\xi}(\boldsymbol{x}_h) - \boldsymbol{x}_h\|_{\boldsymbol{L}^2(D)} \leq c\, h^s \ell_D^{1-s} \|\nabla{\times}\boldsymbol{x}_h\|_{\boldsymbol{L}^2(D)}. \tag{44.20}$$

Proof. Let us set $\boldsymbol{e}_h := \boldsymbol{\xi}(\boldsymbol{x}_h) - \boldsymbol{x}_h$. We have seen in the proof of Theorem 44.6 that $\mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{\xi}(\boldsymbol{x}_h)) - \boldsymbol{x}_h \in \nabla M_{h0}$, so that $(\nu\boldsymbol{e}_h, \mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{\xi}(\boldsymbol{x}_h)) - \boldsymbol{x}_h)_{\boldsymbol{L}^2(D)} = 0$ since $\boldsymbol{\xi}(\boldsymbol{x}_h) \in \boldsymbol{X}_{0\nu}$, $M_{h0} \subset M_0$, and $\boldsymbol{x}_h \in \boldsymbol{X}_{h0\nu}$. Since $\boldsymbol{e}_h = (I - \mathcal{J}_{h0}^{\mathrm{c}})(\boldsymbol{\xi}(\boldsymbol{x}_h)) + (\mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{\xi}(\boldsymbol{x}_h)) - \boldsymbol{x}_h)$, we infer that

$$(\nu\boldsymbol{e}_h, \boldsymbol{e}_h)_{\boldsymbol{L}^2(D)} = (\nu\boldsymbol{e}_h, (I - \mathcal{J}_{h0}^{\mathrm{c}})(\boldsymbol{\xi}(\boldsymbol{x}_h)))_{\boldsymbol{L}^2(D)},$$

thereby implying that $\|\boldsymbol{e}_h\|_{\boldsymbol{L}^2(D)} \leq \nu_{\sharp/\flat}\|(I - \mathcal{J}_{h0}^{\mathrm{c}})(\boldsymbol{\xi}(\boldsymbol{x}_h))\|_{\boldsymbol{L}^2(D)}$. Using the approximation properties of $\mathcal{J}_{h0}^{\mathrm{c}}$ yields

$$\|\boldsymbol{e}_h\|_{\boldsymbol{L}^2(D)} \leq c\,\nu_{\sharp/\flat} h^s |\boldsymbol{\xi}(\boldsymbol{x}_h)|_{\boldsymbol{H}^s(D)},$$

and we conclude using the bound $|\boldsymbol{\xi}(\boldsymbol{x}_h)|_{\boldsymbol{H}^s(D)} \leq \check{C}_D \ell_D^{1-s}\|\nabla{\times}\boldsymbol{x}_h\|_{\boldsymbol{L}^2(D)}$ which follows from (44.7) since $\boldsymbol{\xi}(\boldsymbol{x}_h) \in \boldsymbol{X}_{0,\nu}$ and $\nabla{\times}\boldsymbol{\xi}(\boldsymbol{x}_h) = \nabla{\times}\boldsymbol{x}_h$. $\qquad\square$

Lemma 44.12 (Adjoint solution). *Let $\boldsymbol{y} \in \boldsymbol{X}_{0\nu}$ and let $\boldsymbol{\zeta} \in \boldsymbol{X}_{0\nu}$ solve the (adjoint) problem $\nu\boldsymbol{\zeta} + \nabla{\times}(\kappa\nabla{\times}\boldsymbol{\zeta}) := \nu_\flat^{-1}\nu\boldsymbol{y}$. There is c, depending on the constants $\hat{C}_{\mathrm{PS}}$ from (44.9), $\check{C}$, $\check{C}'$ from (44.7)-(44.8), and the contrast factors $\nu_{\sharp/\flat}$, $\kappa_{\sharp/\flat}$, and $\kappa_\sharp C_{\kappa^{-1}}$, s.t. for all $h \in \mathcal{H}$,*

$$|\boldsymbol{\zeta}|_{\boldsymbol{H}^\sigma(D)} \leq c\,\nu_\sharp^{-1}\gamma_{\nu,\kappa}\ell_D^{-\sigma}\|\boldsymbol{y}\|_{L^2(D)}, \tag{44.21a}$$

$$|\nabla{\times}\boldsymbol{\zeta}|_{\boldsymbol{H}^\sigma(D)} \leq c\,\nu_\sharp^{-1}\gamma_{\nu,\kappa}\hat{\gamma}_{\nu,\kappa}\ell_D^{-1-\sigma}\|\boldsymbol{y}\|_{L^2(D)}. \tag{44.21b}$$

Proof. Proof of (44.21a). Testing the adjoint problem with $e^{-i\theta}\boldsymbol{\zeta}$ leads to $\kappa_\flat\|\nabla{\times}\boldsymbol{\zeta}\|_{\boldsymbol{L}^2(D)}^2 \leq \nu_{\sharp/\flat}\|\boldsymbol{y}\|_{L^2(D)}\|\boldsymbol{\zeta}\|_{L^2(D)}$. Using the Poincaré–Steklov inequality (44.9), we can bound $\|\boldsymbol{\zeta}\|_{L^2(D)}$ by $\|\nabla{\times}\boldsymbol{\zeta}\|_{\boldsymbol{L}^2(D)}$, and altogether this gives

$$\|\nabla{\times}\boldsymbol{\zeta}\|_{\boldsymbol{L}^2(D)} \leq \kappa_\flat^{-1}\nu_{\sharp/\flat}\hat{C}_{\mathrm{PS}}^{-1}\ell_D\|\boldsymbol{y}\|_{\boldsymbol{L}^2(D)}. \tag{44.22}$$

Invoking (44.7) with $\sigma \leq s$ yields

$$|\boldsymbol{\zeta}|_{\boldsymbol{H}^\sigma(D)} \leq \check{C}_D^{-1}\ell_D^{1-\sigma}\|\nabla{\times}\boldsymbol{\zeta}\|_{\boldsymbol{L}^2(D)} \leq \kappa_\flat^{-1}\nu_{\sharp/\flat}\check{C}_D^{-1}\hat{C}_{\mathrm{PS}}^{-1}\ell_D^{2-\sigma}\|\boldsymbol{y}\|_{\boldsymbol{L}^2(D)},$$

which proves (44.21a) since $\kappa_\flat^{-1}\ell_D^2 = \kappa_{\sharp/\flat}\nu_\sharp^{-1}\gamma_{\nu,\kappa}$.

Proof of (44.21b). Invoking (44.8) with $\sigma \leq s'$ for $\boldsymbol{b} := \kappa\nabla{\times}\boldsymbol{\zeta}$, which is a member of $\boldsymbol{X}_{*\kappa^{-1}}$, we infer that

$$\check{C}'_D\ell_D^{-1+\sigma}|\boldsymbol{b}|_{\boldsymbol{H}^\sigma(D)} \leq \|\nabla{\times}\boldsymbol{b}\|_{\boldsymbol{L}^2(D)} = \|\nabla{\times}(\kappa\nabla{\times}\boldsymbol{\zeta})\|_{\boldsymbol{L}^2(D)}$$
$$\leq \nu_{\sharp/\flat}\|\boldsymbol{y}\|_{\boldsymbol{L}^2(D)} + \nu_\sharp\|\boldsymbol{\zeta}\|_{\boldsymbol{L}^2(D)},$$

by definition of the adjoint solution $\boldsymbol{\zeta}$ and the triangle inequality. Invoking again the Poincaré–Steklov inequality (44.9) to bound $\|\boldsymbol{\zeta}\|_{\boldsymbol{L}^2(D)}$ by $\|\nabla{\times}\boldsymbol{\zeta}\|_{\boldsymbol{L}^2(D)}$ and using (44.22) yields $\|\boldsymbol{\zeta}\|_{\boldsymbol{L}^2(D)} \leq \kappa_\flat^{-1}\nu_{\sharp/\flat}\hat{C}_{\mathrm{PS}}^{-2}\ell_D^2\|\boldsymbol{y}\|_{\boldsymbol{L}^2(D)}$. As a result, we obtain

$$\check{C}'_D\ell_D^{-1+\sigma}|\boldsymbol{b}|_{\boldsymbol{H}^\sigma(D)} \leq \nu_{\sharp/\flat}(1 + \nu_\sharp\kappa_\flat^{-1}\hat{C}_{\mathrm{PS}}^{-2}\ell_D^2)\|\boldsymbol{y}\|_{\boldsymbol{L}^2(D)},$$

and this concludes the proof of (44.21b) since $|\nabla{\times}\boldsymbol{\zeta}|_{\boldsymbol{H}^\sigma(D)} \leq C_{\kappa^{-1}}|\boldsymbol{b}|_{\boldsymbol{H}^\sigma(D)}$ owing to the multiplier property (44.11) and $\sigma \leq \tau$. $\qquad\square$

We can now state the main result of this section.

Theorem 44.13 (Improved L^2-error estimate). *Let $\boldsymbol{A}$ solve (44.1) and let $\boldsymbol{A}_h$ solve (44.13). There is c, depending on the constants $\hat{C}_{\mathrm{PS}}$ from (44.9), $\check{C}$, $\check{C}'$ from (44.7)-(44.8), and the contrast factors $\nu_{\sharp/\flat}$, $\kappa_{\sharp/\flat}$, and $\kappa_\sharp C_{\kappa^{-1}}$, s.t. for all $h \in \mathcal{H}$,*

$$\|\boldsymbol{A} - \boldsymbol{A}_h\|_{\boldsymbol{L}^2(D)} \leq c \inf_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}} (\|\boldsymbol{A} - \boldsymbol{v}_h\|_{\boldsymbol{L}^2(D)} + \hat{\gamma}_{\nu,\kappa}^3 h^\sigma\ell_D^{-\sigma}\|\boldsymbol{A} - \boldsymbol{v}_h\|_{\boldsymbol{H}(\mathrm{curl})}).$$

Proof. In this proof, we use the symbol c to denote a generic positive constant that can have the same parametric dependencies as in the above statement. Let $\boldsymbol{v}_h \in \boldsymbol{X}_{h0\nu}$ and let us set $\boldsymbol{x}_h := \boldsymbol{A}_h - \boldsymbol{v}_h$. We observe that $\boldsymbol{x}_h \in \boldsymbol{X}_{h0\nu}$. Let $\boldsymbol{\xi}(\boldsymbol{x}_h)$ be the image of $\boldsymbol{x}_h$ by the curl-preserving lifting operator and let $\boldsymbol{\zeta} \in \boldsymbol{X}_{0\nu}$ be the solution to the following adjoint problem:

$$\nu\boldsymbol{\zeta} + \nabla{\times}(\kappa\nabla{\times}\boldsymbol{\zeta}) := \nu_\flat^{-1}\nu\boldsymbol{\xi}(\boldsymbol{x}_h).$$

(1) Let us first bound $\|\boldsymbol{\xi}(\boldsymbol{x}_h)\|_{\boldsymbol{L}^2(D)}$ from above. Recalling that $\boldsymbol{\xi}(\boldsymbol{x}_h) - \boldsymbol{x}_h = -\nabla\phi(\boldsymbol{x}_h)$ and that $(\nu\boldsymbol{\xi}(\boldsymbol{x}_h), \boldsymbol{\xi}(\boldsymbol{x}_h) - \boldsymbol{x}_h)_{\boldsymbol{L}^2(D)} = -(\nu\boldsymbol{\xi}(\boldsymbol{x}_h), \nabla\phi(\boldsymbol{x}_h))_{\boldsymbol{L}^2(D)} = 0$, we infer that

$$(\boldsymbol{\xi}(\boldsymbol{x}_h), \nu\boldsymbol{\xi}(\boldsymbol{x}_h))_{\boldsymbol{L}^2(D)} = (\boldsymbol{x}_h, \nu\boldsymbol{\xi}(\boldsymbol{x}_h))_{\boldsymbol{L}^2(D)}$$
$$= (\boldsymbol{A} - \boldsymbol{v}_h, \nu\boldsymbol{\xi}(\boldsymbol{x}_h))_{\boldsymbol{L}^2(D)} + (\boldsymbol{A}_h - \boldsymbol{A}, \nu\boldsymbol{\xi}(\boldsymbol{x}_h))_{\boldsymbol{L}^2(D)}$$
$$= (\boldsymbol{A} - \boldsymbol{v}_h, \nu\boldsymbol{\xi}(\boldsymbol{x}_h))_{\boldsymbol{L}^2(D)} + \nu_\flat a_{\nu,\kappa}(\boldsymbol{A}_h - \boldsymbol{A}, \boldsymbol{\zeta})$$
$$= (\boldsymbol{A} - \boldsymbol{v}_h, \nu\boldsymbol{\xi}(\boldsymbol{x}_h))_{\boldsymbol{L}^2(D)} + \nu_\flat a_{\nu,\kappa}(\boldsymbol{A}_h - \boldsymbol{A}, \boldsymbol{\zeta} - \mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{\zeta})),$$

where we used the Galerkin orthogonality property on the fourth line. Since we have $|a_{\nu,\kappa}(\boldsymbol{a}, \boldsymbol{b})| \leq \kappa_\sharp\ell_D^{-2}\hat{\gamma}_{\nu,\kappa}\|\boldsymbol{a}\|_{\boldsymbol{H}(\mathrm{curl};D)}\|\boldsymbol{b}\|_{\boldsymbol{H}(\mathrm{curl};D)}$ by (43.13b), we infer from the commutation and approximation properties of the quasi-interpolation operators that

$$\|\boldsymbol{\xi}(\boldsymbol{x}_h)\|_{\boldsymbol{L}^2(D)}^2 \leq \nu_{\sharp/\flat}\|\boldsymbol{A} - \boldsymbol{v}_h\|_{\boldsymbol{L}^2(D)}\|\boldsymbol{\xi}(\boldsymbol{x}_h)\|_{\boldsymbol{L}^2(D)}$$
$$+ c\,\kappa_\sharp\ell_D^{-2}\hat{\gamma}_{\nu,\kappa}h^\sigma\|\boldsymbol{A} - \boldsymbol{A}_h\|_{\boldsymbol{H}(\mathrm{curl};D)}(|\boldsymbol{\zeta}|_{\boldsymbol{H}^\sigma(D)} + \ell_D|\nabla{\times}\boldsymbol{\zeta}|_{\boldsymbol{H}^\sigma(D)}).$$

Owing to the bounds from Lemma 44.12 on the adjoint solution with $\boldsymbol{y} := \boldsymbol{\xi}(\boldsymbol{x}_h)$, we conclude that

$$\|\boldsymbol{\xi}(\boldsymbol{x}_h)\|_{\boldsymbol{L}^2(D)} \leq \nu_{\sharp/\flat}(\|\boldsymbol{A} - \boldsymbol{v}_h\|_{\boldsymbol{L}^2(D)} + c\hat{\gamma}_{\nu,\kappa}^2 h^\sigma\ell_D^{-\sigma}\|\boldsymbol{A} - \boldsymbol{A}_h\|_{\boldsymbol{H}(\mathrm{curl};D)}).$$

(2) The triangle inequality and the identity $\boldsymbol{A} - \boldsymbol{A}_h = \boldsymbol{A} - \boldsymbol{v}_h - \boldsymbol{x}_h$ imply that

$$\|\boldsymbol{A} - \boldsymbol{A}_h\|_{\boldsymbol{L}^2(D)} \leq \|\boldsymbol{A} - \boldsymbol{v}_h\|_{\boldsymbol{L}^2(D)} + \|\boldsymbol{\xi}(\boldsymbol{x}_h) - \boldsymbol{x}_h\|_{\boldsymbol{L}^2(D)} + \|\boldsymbol{\xi}(\boldsymbol{x}_h)\|_{\boldsymbol{L}^2(D)}.$$

We use Lemma 44.11 to bound the second term on the right-hand side as

$$\begin{aligned}
\|\boldsymbol{\xi}(\boldsymbol{x}_h) - \boldsymbol{x}_h\|_{\boldsymbol{L}^2(D)} &\leq c\, h^\sigma \ell_D^{1-\sigma} \|\nabla{\times}\boldsymbol{x}_h\|_{\boldsymbol{L}^2(D)} \\
&\leq c\, h^\sigma \ell_D^{1-\sigma} (\|\nabla{\times}(\boldsymbol{A} - \boldsymbol{v}_h)\|_{\boldsymbol{L}^2(D)} + \|\nabla{\times}(\boldsymbol{A} - \boldsymbol{A}_h)\|_{\boldsymbol{L}^2(D)}),
\end{aligned}$$

and we use (44.18) to infer that $\|\boldsymbol{A} - \boldsymbol{A}_h\|_{\boldsymbol{H}(\mathrm{curl};D)} \leq c\hat{\gamma}_{\nu,\kappa}\|\boldsymbol{A} - \boldsymbol{v}_h\|_{\boldsymbol{H}(\mathrm{curl};D)}$. For the third term on the right-hand side, we use the bound on $\|\boldsymbol{\xi}(\boldsymbol{x}_h)\|_{\boldsymbol{L}^2(D)}$ from Step (1). We conclude by taking the infimum over $\boldsymbol{v}_h \in \boldsymbol{X}_{h0\nu}$, and we use Lemma 44.8 to extend the infimum over $\boldsymbol{V}_{h0}$. □

Remark 44.14 (Literature). The construction of the curl-preserving lifting operator invoked in the proof of Theorem 44.6 and Theorem 44.13 is done in Monk [302, pp. 249-250]. The statement in Lemma 44.11 is similar to that in Monk [303, Lem. 7.6], but the present proof is simplified by the use of the commuting quasi-interpolation operators. The curl-preserving lifting of $\boldsymbol{A} - \boldsymbol{A}_h$ is invoked in Arnold et al. [23, Eq. (9.9)] and denoted therein by $\boldsymbol{\psi}$. The estimate of $\|\boldsymbol{\psi}\|_{\boldsymbol{L}^2(D)}$ given one line above [23, Eq. (9.11)] is similar to (44.3) and is obtained by invoking the commuting quasi-interpolation operators constructed in [23, §5.4] for natural boundary conditions. Note that contrary to the above reference, we invoke the curl-preserving lifting of $\boldsymbol{A}_h - \boldsymbol{v}_h$ instead of $\boldsymbol{A} - \boldsymbol{A}_h$ and make use of Lemma 44.11, which simplifies the argument. Furthermore, the statement of Theorem 44.13 is similar to that of Zhong et al. [405, Thm. 4.1], but the present proof is simpler and does not require the smoothness index σ to be larger than $\frac{1}{2}$. □

Exercises

Exercise 44.1 (Gradient). Let $\phi \in H_0^1(D)$. Prove that $\nabla\phi \in \boldsymbol{H}_0(\mathrm{curl};D)$

Exercise 44.2 (Vector potential). Let $\boldsymbol{v} \in \boldsymbol{L}^2(D)$ with $(\nu\boldsymbol{v}, \nabla m_h)_{\boldsymbol{L}^2(D)} = 0$ for all $m_h \in M_{h0}$. Prove that $(\nu\boldsymbol{v}, \boldsymbol{w}_h)_{\boldsymbol{L}^2(D)} = (\nabla{\times}\boldsymbol{z}_h, \nabla{\times}\boldsymbol{w}_h)_{\boldsymbol{L}^2(D)}$ for all $\boldsymbol{w}_h \in \boldsymbol{V}_{h0}$, where $\boldsymbol{z}_h$ solves a curl-curl problem on $\boldsymbol{X}_{h0\nu}$.

Exercise 44.3 (Neumann condition). Recall Remark 44.10. Assume that D is simply connected so that there is $\hat{C}_{\mathrm{PS}} > 0$ such that $\hat{C}_{\mathrm{PS}}\ell_D^{-1}\|\boldsymbol{b}\|_{\boldsymbol{L}^2(D)} \leq \|\nabla{\times}\boldsymbol{b}\|_{\boldsymbol{L}^2(D)}$ for all $\boldsymbol{b} \in \boldsymbol{X}_{*\nu}$. Prove that there is $\hat{C}'_{\mathrm{PS}} > 0$ such that $\hat{C}'_{\mathrm{PS}}\ell_D^{-1}|\boldsymbol{b}_h\|_{\boldsymbol{L}^2(D)} \leq \|\nabla{\times}\boldsymbol{b}_h\|_{\boldsymbol{L}^2(D)}$ for all $\boldsymbol{b}_h \in \boldsymbol{X}_{h*\nu}$. (*Hint*: adapt the proof of Theorem 44.6 using $\mathcal{J}_h^{\mathrm{c}}$.)

Exercise 44.4 (Discrete Poincaré–Steklov for $\nabla\cdot$). Let ν be as in §44.1.1. Let $\boldsymbol{Y}_{0\nu} := \{\boldsymbol{v} \in \boldsymbol{H}_0(\mathrm{div};D) \mid (\nu\boldsymbol{v}, \nabla{\times}\boldsymbol{\phi})_{\boldsymbol{L}^2(D)} = 0, \ \forall\boldsymbol{\phi} \in \boldsymbol{H}_0(\mathrm{curl};D)\}$ and accept as a fact that there is $\hat{C}_{\mathrm{PS}} > 0$ such that $\hat{C}_{\mathrm{PS}}\ell_D^{-1}\|\boldsymbol{v}\|_{\boldsymbol{L}^2(D)} \leq \|\nabla\cdot\boldsymbol{v}\|_{\boldsymbol{L}^2(D)}$ for all $\boldsymbol{v} \in \boldsymbol{Y}_{0\nu}$. Let $k \geq 0$ and consider the discrete space $\boldsymbol{Y}_{h0\nu} := \{\boldsymbol{v}_h \in \boldsymbol{P}_{k,0}^{\mathrm{d}}(\mathcal{T}_h) \mid (\nu\boldsymbol{v}_h, \nabla{\times}\boldsymbol{\phi}_h)_{\boldsymbol{L}^2(D)} = 0, \ \forall\boldsymbol{\phi}_h \in \boldsymbol{P}_{k,0}^{\mathrm{c}}(\mathcal{T}_h;\mathbb{C})\}$. Prove that there is $\hat{C}'_{\mathrm{PS}} > 0$ such that $\hat{C}'_{\mathrm{PS}}\|\boldsymbol{v}_h\|_{\boldsymbol{L}^2(D)} \leq \ell_D\|\nabla\cdot\boldsymbol{v}_h\|_{\boldsymbol{L}^2(D)}$ for all $\boldsymbol{v}_h \in \boldsymbol{Y}_{h0\nu}$. (*Hint*: adapt the proof of Theorem 44.6 using $\mathcal{J}_{h0}^{\mathrm{d}}$.)

Chapter 45

Maxwell's equations: further topics

In this chapter, we investigate two additional topics on the approximation of Maxwell's equations. First, we study the use of a boundary penalty method inspired by Nitsche's method for elliptic PDEs (see Chapter 37 and §41.3) to enforce the boundary condition on the tangential component. We combine this method with $\boldsymbol{H}(\mathrm{curl})$-conforming elements and with the all-purpose $\boldsymbol{H}^1$-conforming elements. The use of a boundary penalty method is motivated for $\boldsymbol{H}^1$-conforming elements whenever some faces of the domain D are not parallel to the canonical Cartesian planes in $\mathbb{R}^3$ since in this case the boundary condition couples the Cartesian components of the discrete solution. For simplicity, we study the boundary penalty method under the assumption that there is a uniformly positive zero-order term in the model problem. The second topic we explore in this chapter is the use of a least-squares penalty technique to control the divergence in the context of $\boldsymbol{H}^1$-conforming elements. We will see that this technique works well for smooth solutions, but there is an approximability obstruction for nonsmooth solutions.

45.1 Model problem

We consider the weak formulation (44.1) in a Lipschitz domain D in $\mathbb{R}^3$:

$$
\begin{cases}
\text{Find } \boldsymbol{A} \in \boldsymbol{V}_0 := \boldsymbol{H}_0(\mathrm{curl}; D) \text{ such that} \\
a_{\nu,\kappa}(\boldsymbol{A}, \boldsymbol{b}) = \ell(\boldsymbol{b}), \quad \forall \boldsymbol{b} \in \boldsymbol{V}_0,
\end{cases}
\tag{45.1}
$$

with $a_{\nu,\kappa}(\boldsymbol{a}, \boldsymbol{b}) := \int_D (\nu \boldsymbol{a} \cdot \overline{\boldsymbol{b}} + \kappa \nabla \times \boldsymbol{a} \cdot \nabla \times \overline{\boldsymbol{b}}) \, \mathrm{d}x$, $\ell(\boldsymbol{b}) := \int_D \boldsymbol{f} \cdot \overline{\boldsymbol{b}} \, \mathrm{d}x$, and $\boldsymbol{f} \in \boldsymbol{L}^2(D)$. As in §44.1.1, we assume that (i) $\nu, \kappa \in L^\infty(D; \mathbb{C})$ and we set $\nu_\sharp := \|\nu\|_{L^\infty(D;\mathbb{C})}$, $\kappa_\sharp := \|\kappa\|_{L^\infty(D;\mathbb{C})}$; (ii) There are real numbers θ, $\nu_\flat > 0$, and $\kappa_\flat > 0$ s.t. i.e., $\operatorname{ess\,inf}_{\boldsymbol{x} \in D} \Re(e^{\mathrm{i}\theta} \nu(\boldsymbol{x})) \geq \nu_\flat$, $\operatorname{ess\,inf}_{\boldsymbol{x} \in D} \Re(e^{\mathrm{i}\theta} \kappa(\boldsymbol{x})) \geq \kappa_\flat$; (iii) There is a partition of D into M disjoint Lipschitz polyhedra $\{D_m\}_{m \in \{1:M\}}$ s.t. $\nu_{|D_m}, \kappa_{|D_m}$ are constant for all $m \in \{1:M\}$. Recall that $\ell_D := \mathrm{diam}(D)$.

45.2 Boundary penalty method in $H(\mathrm{curl})$

In this section, we apply Nitsche's boundary penalty method (see Chapter 37 and §41.3) to the approximation of the model problem (43.10) using edge (Nédélec) finite elements.

45.2.1 Discrete problem

Let $(\mathcal{T}_h)_{h\in\mathcal{H}}$ be a shape-regular sequence of affine meshes so that each mesh covers D exactly. We assume that each mesh is compatible with the partition $\{D_m\}_{m\in\{1:M\}}$ so that ν and κ are piecewise constant on $\mathcal{T}_h$. We set $\kappa_K := \kappa_{|K}$, $\nu_K := \nu_{|K}$, $\kappa_{r,K} := \Re(e^{i\theta}\kappa_K)$, and $\nu_{r,K} := \Re(e^{i\theta}\nu_K)$ for all $K \in \mathcal{T}_h$ (notice that $\kappa_{r,K} \geq \kappa_\flat > 0$ and $\nu_{r,K} \geq \nu_\flat > 0$). For every boundary face $F \in \mathcal{F}_h^\partial$, we denote by K_l the unique mesh cell having F as a face, i.e., $F := \partial K_l \cap \partial D$. To simplify the notation, we set $\lambda_F := \frac{|\kappa_{K_l}|^2}{\kappa_{r,K_l}}$.

In Nitsche's boundary penalty method, the degrees of freedom associated with the tangential component at the boundary of the trial functions and of the test functions are kept in the trial and in the test spaces. Hence, we set $\boldsymbol{V}_h := \boldsymbol{P}_k^c(\mathcal{T}_h) \subset \boldsymbol{H}(\mathrm{curl};D)$, $k \geq 0$. Since $\boldsymbol{V}_h$ is not a subspace of $\boldsymbol{H}_0(\mathrm{curl};D)$, the approximation setting is nonconforming. The discrete formulation is

$$\begin{cases} \text{Find } \boldsymbol{A}_h \in \boldsymbol{V}_h \text{ such that} \\ a_{\nu,\kappa,h}(\boldsymbol{A}_h, \boldsymbol{b}_h) = \ell(\boldsymbol{b}_h), \quad \forall \boldsymbol{b}_h \in \boldsymbol{V}_h. \end{cases} \tag{45.2}$$

The sesquilinear form $a_{\nu,\kappa,h} : \boldsymbol{V}_h \times \boldsymbol{V}_h \to \mathbb{C}$ is such that

$$a_{\nu,\kappa,h}(\boldsymbol{a}_h, \boldsymbol{b}_h) := a_{\nu,\kappa}(\boldsymbol{a}_h, \boldsymbol{b}_h) - n_h(\boldsymbol{a}_h, \boldsymbol{b}_h) + \sum_{F\in\mathcal{F}_h^\partial} \eta_0 e^{-i\theta} \frac{\lambda_F}{h_F} \int_F (\boldsymbol{a}_h \times \boldsymbol{n}) \cdot (\overline{\boldsymbol{b}}_h \times \boldsymbol{n}) \, \mathrm{d}s, \tag{45.3}$$

where η_0 is a user-dependent parameter to be chosen large enough (see Lemma 45.1), and using the notation $\boldsymbol{\sigma}(\boldsymbol{a}) := \kappa\nabla\times\boldsymbol{a}$ for all $\boldsymbol{a} \in \boldsymbol{H}(\mathrm{curl};D)$,

$$n_h(\boldsymbol{a}_h, \boldsymbol{b}_h) := \int_{\partial D} (\boldsymbol{\sigma}(\boldsymbol{a}_h)\times\boldsymbol{n}) \cdot \overline{\boldsymbol{b}}_h \, \mathrm{d}s. \tag{45.4}$$

45.2.2 Stability and well-posedness

We equip the space $\boldsymbol{V}_h$ with the following stability norm: For all $\boldsymbol{b}_h \in \boldsymbol{V}_h$,

$$\|\boldsymbol{b}_h\|_{\boldsymbol{V}_h}^2 := \sum_{K\in\mathcal{T}_h} \left(\nu_{r,K}\|\boldsymbol{b}_h\|_{\boldsymbol{L}^2(K)}^2 + \kappa_{r,K}\|\nabla\times\boldsymbol{b}_h\|_{\boldsymbol{L}^2(K)}^2 \right) + |\boldsymbol{b}_h|_\partial^2, \tag{45.5a}$$

$$|\boldsymbol{b}_h|_\partial^2 := \sum_{F\in\mathcal{F}_h^\partial} \frac{\lambda_F}{h_F}\|\boldsymbol{b}_h\times\boldsymbol{n}\|_{\boldsymbol{L}^2(F)}^2. \tag{45.5b}$$

Let $\mathcal{T}_h^{\partial D}$ be the collection of all the mesh cells having a boundary face. Let $n_\partial := \max_{K\in\mathcal{T}_h^{\partial D}} |\mathcal{F}_K \cap \mathcal{F}_h^\partial|$ denote the maximum number of boundary faces that a mesh cell in $\mathcal{T}_h^{\partial D}$ can have ($n_\partial \leq d$ for simplicial meshes). As in (37.6), let c_{dt} be the smallest constant such that $\|\boldsymbol{n}\times(\nabla\times\boldsymbol{v}_h)\|_{\boldsymbol{L}^2(F)} \leq c_{\mathrm{dt}}h_F^{-\frac{1}{2}}\|\nabla\times\boldsymbol{v}_h\|_{\boldsymbol{L}^2(F)}$ for all $F \in \mathcal{F}_h^\partial$ and all $\boldsymbol{v}_h \in \boldsymbol{V}_h$.

Lemma 45.1 (Coercivity, well-posedness). *Assume that $\eta_0 > \frac{1}{4}n_\partial c_{\mathrm{dt}}^2$. (i) The following coercivity property holds true:*

$$\Re\big(e^{i\theta}a_{\nu,\kappa,h}(\boldsymbol{b}_h, \boldsymbol{b}_h)\big) \geq \alpha\|\boldsymbol{b}_h\|_{\boldsymbol{V}_h}^2, \qquad \forall \boldsymbol{b}_h \in \boldsymbol{V}_h, \tag{45.6}$$

with $\alpha := \frac{\eta_0 - \frac{1}{4} n_\partial c_{\mathrm{dt}}^2}{1 + \eta_0} > 0$. (ii) *The discrete problem* (45.2) *is well-posed.*

Proof. We only sketch the proof since it is similar to that of Lemma 37.3. For all $\boldsymbol{b}_h \in \boldsymbol{V}_h$, we have

$$\Re\big(e^{\mathrm{i}\theta} a_{\nu,\kappa,h}(\boldsymbol{b}_h, \boldsymbol{b}_h)\big) \geq \|\boldsymbol{b}_h\|_{\boldsymbol{V}_h}^2 - \Re\big(e^{\mathrm{i}\theta} n_h(\boldsymbol{b}_h, \boldsymbol{b}_h)\big).$$

The last term on the right-hand side is bounded by proceeding as in the proof of Lemma 37.2. Using that $(\boldsymbol{\sigma}(\boldsymbol{b}_h) \times \boldsymbol{n}) \cdot \overline{\boldsymbol{b}}_h = -\boldsymbol{\sigma}(\boldsymbol{b}_h) \cdot (\overline{\boldsymbol{b}}_h \times \boldsymbol{n})$ and $|\kappa_{r,K_l}| \leq |\kappa_{K_l}|$, we infer that

$$|n_h(\boldsymbol{b}_h, \boldsymbol{b}_h)| \leq n_\partial^{\frac{1}{2}} c_{\mathrm{dt}} \bigg(\sum_{K \in \mathcal{T}_h^{\partial D}} \kappa_{r,K_l} \|\nabla \times \boldsymbol{b}_h\|_{\boldsymbol{L}^2(K)}^2 \bigg)^{\frac{1}{2}} |\boldsymbol{b}_h|_\partial.$$

Then we use the same quadratic inequality as in the proof of Lemma 37.3 to conclude that (45.6) holds true. Finally, the well-posedness of (45.2) follows from the Lax–Milgram lemma. $\square$

Remark 45.2 (Sesquilinear form $a_{\nu,\kappa,h}$). If in the penalty term in (45.3) one takes the κ-dependent factor equal to $\frac{|\kappa_{K_l}|}{h_F}$ instead of $\frac{\lambda_F}{h_F}$, the minimal value for the parameter η_0 in Lemma 45.1 depends on $\max_{K \in \mathcal{T}_h} \frac{|\kappa_{K_l}|}{\kappa_{r,K_l}}$, which is not convenient in general. Furthermore, one can add the term $-\overline{n_h(\boldsymbol{b}_h, \boldsymbol{a}_h)}$ to the right-hand side of (45.3) to make $a_{\nu,\kappa,h}$ Hermitian. Then the coercivity property (45.6) is valid if $\eta_0 > n_\partial c_{\mathrm{dt}}^2$ with $\alpha := \frac{\eta_0 - n_\partial c_{\mathrm{dt}}^2}{1 + \eta_0} > 0$. $\square$

45.2.3 Error analysis

We perform the error analysis by making only a minimal regularity assumption on $\boldsymbol{A}$, i.e., we assume that (43.15) holds true for some $r > 0$. Our first step consists of extending the tangential trace of $\boldsymbol{\sigma}(\boldsymbol{a}) := \kappa \nabla \times \boldsymbol{a}$. Just like in §40.3.1, we introduce two real numbers p, q such that

$$2 < p, \qquad \frac{2d}{2+d} < q \leq 2, \tag{45.7}$$

and consider $\widetilde{p} \in (2, p]$ such that $q \geq \frac{\widetilde{p}d}{\widetilde{p}+d}$. Notice that $\widetilde{p}$ always exists since $x \mapsto \frac{xd}{x+d}$ is an increasing function. Let $K \in \mathcal{T}_h$ be a mesh cell with outward unit normal $\boldsymbol{n}_K$, and let $F \in \mathcal{F}_K$ be a face of K. We consider the functional space $\boldsymbol{V}^{\mathrm{c}}(K) := \{\boldsymbol{v} \in \boldsymbol{L}^p(K) \mid \nabla \times \boldsymbol{v} \in \boldsymbol{L}^q(K)\}$ equipped with the norm

$$\|\boldsymbol{v}\|_{\boldsymbol{V}^{\mathrm{c}}(K)} := \|\boldsymbol{v}\|_{\boldsymbol{L}^p(K)} + h_K^{1+d(\frac{1}{p}-\frac{1}{q})} \|\nabla \times \boldsymbol{v}\|_{\boldsymbol{L}^q(K)}. \tag{45.8}$$

Recalling that L_F^K is the face-to-cell lifting operator defined in Lemma 17.1, the tangential trace on F of any field $\boldsymbol{v} \in \boldsymbol{V}^{\mathrm{c}}(K)$ is denoted by $(\boldsymbol{v} \times \boldsymbol{n}_K)_{|F}$, and is defined as the antilinear form in $(\boldsymbol{W}^{\frac{1}{p}, \widetilde{p}'}(F))'$ s.t. for all $\boldsymbol{\phi} \in \boldsymbol{W}^{\frac{1}{p}, \widetilde{p}'}(F)$,

$$\langle (\boldsymbol{v} \times \boldsymbol{n}_K)_{|F}, \boldsymbol{\phi} \rangle_F := \int_K \Big(\boldsymbol{v} \cdot \nabla \times L_F^K(\overline{\boldsymbol{\phi}}) - (\nabla \times \boldsymbol{v}) \cdot L_F^K(\overline{\boldsymbol{\phi}}) \Big) \, \mathrm{d}x, \tag{45.9}$$

where $\langle \cdot, \cdot \rangle_F$ denotes the duality pairing between $(\boldsymbol{W}^{\frac{1}{p}, \widetilde{p}'}(F))'$ and $\boldsymbol{W}^{\frac{1}{p}, \widetilde{p}'}(F)$. The right-hand side of (45.9) is well defined since $\boldsymbol{v} \in \boldsymbol{L}^p(K)$, $\nabla \times \boldsymbol{v} \in \boldsymbol{L}^q(K)$, and $L_F^K(\boldsymbol{\phi}) \in \boldsymbol{W}^{1, \widetilde{p}'}(K) \hookrightarrow \boldsymbol{L}^{q'}(K) \cap \boldsymbol{W}^{1,p'}(K)$.

We now properly extend the sesquilinear form n_h defined in (45.4). For this purpose, we set

$$\boldsymbol{V}_{\mathrm{s}} := \{\boldsymbol{a} \in \boldsymbol{H}_0(\mathrm{curl}; D) \mid \boldsymbol{\sigma}(\boldsymbol{a}) \in \boldsymbol{L}^p(D), \nabla \times \boldsymbol{\sigma}(\boldsymbol{a}) \in \boldsymbol{L}^q(D)\}, \tag{45.10}$$

and observe that $\boldsymbol{\sigma}(\boldsymbol{a})_{|K} \in \boldsymbol{V}^{\mathrm{c}}(K)$ for all $\boldsymbol{a} \in \boldsymbol{V}_{\mathrm{s}}$ and all $K \in \mathcal{T}_h$. It turns out that the assumption $\boldsymbol{f} \in \boldsymbol{L}^2(D)$ implies that one can always take $q := 2$.

Lemma 45.3 (Smoothness). *Let $\boldsymbol{A}$ solve* (45.1). *Assume that there is $r > 0$ s.t.*

$$\boldsymbol{A} \in \boldsymbol{H}^r(D), \qquad \nabla{\times}\boldsymbol{A} \in \boldsymbol{H}^r(D). \tag{45.11}$$

Then $\boldsymbol{A} \in \boldsymbol{V}_s$ with $q := 2$ and with either any $p \in (2, \frac{6}{3-2r}]$ if $r < \frac{3}{2}$ or all $p \in (2, \infty)$ otherwise.

Proof. The property $\boldsymbol{\sigma}(\boldsymbol{A}) \in \boldsymbol{L}^p(D)$ with p as in the assertion follows from $\nabla{\times}\boldsymbol{A} \in \boldsymbol{H}^r(D) \hookrightarrow \boldsymbol{L}^p(D)$ owing to the Sobolev embedding theorem and the assumption $\kappa \in L^\infty(D; \mathbb{C})$. Since $\nabla{\times}\boldsymbol{\sigma}(\boldsymbol{A}) = \boldsymbol{f} - \nu\boldsymbol{A}$, $\boldsymbol{f} \in \boldsymbol{L}^2(D)$, and $\nu \in L^\infty(D; \mathbb{C})$, we conclude that $\nabla{\times}\boldsymbol{\sigma}(\boldsymbol{A}) \in \boldsymbol{L}^2(D)$. $\qquad\square$

We define $\boldsymbol{V}_\sharp := \boldsymbol{V}_s + \boldsymbol{V}_h$ and the sesquilinear form on $\boldsymbol{V}_\sharp \times \boldsymbol{V}_h$ such that

$$n_\sharp(\boldsymbol{a}, \boldsymbol{b}_h) := \sum_{F \in \mathcal{F}_h^\partial} \langle (\boldsymbol{\sigma}(\boldsymbol{a})_{|K_l}{\times}\boldsymbol{n})_{|F}, \Pi_F(\boldsymbol{b}_h) \rangle_F, \tag{45.12}$$

where Π_F is the ℓ^2-orthogonal projection onto the hyperplane tangent to F, i.e., $\Pi_F(\boldsymbol{b}_h) := \boldsymbol{n}{\times}(\boldsymbol{b}_h{\times}\boldsymbol{n})$ (note that $\boldsymbol{n} = \boldsymbol{n}_F$ for boundary faces). We observe that (45.12) is meaningful since $\Pi_F(\boldsymbol{b}_h)$ is in $\boldsymbol{W}^{\frac{1}{p}, \widetilde{p}'}(F)$.

Lemma 45.4 (Properties of $n_\sharp$). *The following holds true for all $\boldsymbol{a}_h, \boldsymbol{b}_h \in \boldsymbol{V}_h$ and all $\boldsymbol{a} \in \boldsymbol{V}_s$:*

$$n_\sharp(\boldsymbol{a}_h, \boldsymbol{b}_h) = n_h(\boldsymbol{a}_h, \boldsymbol{b}_h), \tag{45.13a}$$

$$n_\sharp(\boldsymbol{a}, \boldsymbol{b}_h) = \int_D \left(\boldsymbol{\sigma}(\boldsymbol{a}){\cdot}\nabla{\times}\overline{\boldsymbol{b}}_h - (\nabla{\times}\boldsymbol{\sigma}(\boldsymbol{a})){\cdot}\overline{\boldsymbol{b}}_h \right) \mathrm{d}x. \tag{45.13b}$$

Moreover, there is c, uniform w.r.t. κ, such that the following boundedness property holds true for all $\boldsymbol{a} \in \boldsymbol{V}_\sharp$, all $\boldsymbol{b}_h \in \boldsymbol{V}_h$, and all $h \in \mathcal{H}$:

$$|n_\sharp(\boldsymbol{a}, \boldsymbol{b}_h)| \leq c\, |\boldsymbol{a}|_{n_\sharp} |\boldsymbol{b}_h|_\partial, \tag{45.14}$$

with $|\boldsymbol{b}_h|_\partial$ defined in (45.5b) and

$$|\boldsymbol{a}|^2_{n_\sharp} := \sum_{F \in \mathcal{F}_h^\partial} \lambda_F^{-1} \left(h_{K_l}^{2d(\frac{1}{2} - \frac{1}{p})} \|\boldsymbol{\sigma}(\boldsymbol{a})\|^2_{\boldsymbol{L}^p(K_l)} + h_{K_l}^{2d(\frac{d+2}{2d} - \frac{1}{q})} \|\nabla{\times}\boldsymbol{\sigma}(\boldsymbol{a})\|^2_{\boldsymbol{L}^q(K_l)} \right).$$

Proof. (1) Proof of (45.13a). Let $\boldsymbol{a}_h, \boldsymbol{b}_h \in \boldsymbol{V}_h$. We have for all $F \in \mathcal{F}_h^\partial$,

$$\begin{aligned}
\langle (\boldsymbol{\sigma}(\boldsymbol{a}_h)_{|K_l}{\times}\boldsymbol{n})_{|F}, \Pi_F(\boldsymbol{b}_h) \rangle_F &:= \int_{K_l} \left(\boldsymbol{\sigma}(\boldsymbol{a}_h){\cdot}\nabla{\times}L_F^{K_l}(\Pi_F(\overline{\boldsymbol{b}}_h)) - (\nabla{\times}\boldsymbol{\sigma}(\boldsymbol{a}_h)){\cdot}L_F^{K_l}(\Pi_F(\overline{\boldsymbol{b}}_h)) \right) \mathrm{d}x \\
&= \int_{\partial K_l} (\boldsymbol{\sigma}(\boldsymbol{a}_h)_{|K_l}{\times}\boldsymbol{n}){\cdot}L_F^{K_l}(\Pi_F(\overline{\boldsymbol{b}}_h))\, \mathrm{d}s \\
&= \int_F (\boldsymbol{\sigma}(\boldsymbol{a}_h)_{|K_l}{\times}\boldsymbol{n}){\cdot}\Pi_F(\overline{\boldsymbol{b}}_h)\, \mathrm{d}s = \int_F (\boldsymbol{\sigma}(\boldsymbol{a}_h)_{|K_l}{\times}\boldsymbol{n}){\cdot}\overline{\boldsymbol{b}}_h\, \mathrm{d}s,
\end{aligned}$$

since $L_F^{K_l}(\Pi_F(\overline{\boldsymbol{b}}_h))_{|\partial K_l \setminus F} = \boldsymbol{0}$, $L_F^{K_l}(\Pi_F(\overline{\boldsymbol{b}}_h))_{|F} = \Pi_F(\overline{\boldsymbol{b}}_h)$ by definition of $L_F^{K_l}$, and owing to the identity $(\boldsymbol{v}{\times}\boldsymbol{n}){\cdot}\Pi_F(\boldsymbol{w}) = (\boldsymbol{v}{\times}\boldsymbol{n}){\cdot}\boldsymbol{w}$ for all $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^d$. Summing the above identity over $F \in \mathcal{F}_h^\partial$ proves (45.13a).
(2) See Exercise 45.1 for the proof of (45.13b).

(3) We prove (45.14) by proceeding as in the proof of Lemma 40.4. We have

$$|\langle (\boldsymbol{\sigma}(\boldsymbol{a})_{|K_l}\times\boldsymbol{n})_{|F}, \Pi_F(\boldsymbol{b}_h)\rangle_F| \leq c\, h_{K_l}^{d(\frac{1}{2}-\frac{1}{p})} \|\boldsymbol{\sigma}(\boldsymbol{a})\|_{\boldsymbol{V}^c(K_l)} h_F^{-\frac{1}{2}} \|\Pi_F(\boldsymbol{b}_h)\|_{\boldsymbol{L}^2(F)}$$

$$\leq c\left(\lambda_F^{-\frac{1}{2}} h_{K_l}^{d(\frac{1}{2}-\frac{1}{p})} \|\boldsymbol{\sigma}(\boldsymbol{a})\|_{\boldsymbol{L}^p(K_l)} \lambda_F^{\frac{1}{2}} h_F^{-\frac{1}{2}} \|\Pi_F(\boldsymbol{b}_h)\|_{\boldsymbol{L}^2(F)}\right.$$

$$\left.+\lambda_F^{-\frac{1}{2}} h_{K_l}^{d(\frac{d+2}{2d}-\frac{1}{q})} \|\nabla\times\boldsymbol{\sigma}(\boldsymbol{a})\|_{\boldsymbol{L}^q(K_l)} \lambda_F^{\frac{1}{2}} h_F^{-\frac{1}{2}} \|\Pi_F(\boldsymbol{b}_h)\|_{\boldsymbol{L}^2(F)}\right),$$

for all $F \in \mathcal{F}_h^\partial$, where we used that $\|\Pi_F(\boldsymbol{b}_h)\|_{\ell^2} = \|\boldsymbol{b}_h\times\boldsymbol{n}\|_{\ell^2}$ and the definition (45.8) of the norm $\|\boldsymbol{\sigma}(\boldsymbol{a})\|_{\boldsymbol{V}^c(K_l)}$. The rest of the proof is identical to that of Lemma 40.7 by invoking the Cauchy–Schwarz inequality. $\qquad\square$

Recalling that $q := 2$, we equip the space $\boldsymbol{V}_\sharp$ with the norm

$$\|\boldsymbol{b}\|_{\boldsymbol{V}_\sharp}^2 := \sum_{K\in\mathcal{T}_h}\left(\frac{|\nu_K|^2}{\nu_{r,K}}\|\boldsymbol{b}\|_{\boldsymbol{L}^2(K)}^2 + \frac{|\kappa_K|^2}{\kappa_{r,K}}\|\nabla\times\boldsymbol{b}\|_{\boldsymbol{L}^2(K)}^2\right) + |\boldsymbol{b}|_\partial^2 \tag{45.15}$$

$$+ \sum_{F\in\mathcal{F}_h^\partial}\kappa_{r,K_l}\left(h_{K_l}^{2d(\frac{1}{2}-\frac{1}{p})}\|\nabla\times\boldsymbol{b}\|_{\boldsymbol{L}^p(K_l)}^2 + h_{K_l}^2\|\nabla\times(\nabla\times\boldsymbol{b})\|_{\boldsymbol{L}^q(K_l)}^2\right),$$

for all $\boldsymbol{b} \in \boldsymbol{V}_\sharp$, where $|\boldsymbol{b}|_\partial^2 := |\boldsymbol{b}_h|_\partial^2$ with $\boldsymbol{b} := \boldsymbol{b}_s + \boldsymbol{b}_h$, $\boldsymbol{b}_s \in \boldsymbol{V}_s$, $\boldsymbol{b}_h \in \boldsymbol{V}_h$ (this definition is meaningful since functions in $\boldsymbol{V}_s$ have a zero tangential trace at the boundary, so that $\boldsymbol{a}_s + \boldsymbol{a}_h = \boldsymbol{b}_s + \boldsymbol{b}_h$ implies $\boldsymbol{a}_{h|\partial D} = \boldsymbol{b}_{h|\partial D}$). Invoking inverse inequalities shows that there is $c_\sharp$ s.t. $\|\boldsymbol{b}_h\|_{\boldsymbol{V}_\sharp} \leq c_\sharp\|\boldsymbol{b}_h\|_{\boldsymbol{V}_h}$ for all $\boldsymbol{b}_h \in \boldsymbol{V}_h$ and all $h \in \mathcal{H}$, i.e., (27.5) holds true. Note that $c_\sharp$ depends on the factor $\max_{K\in\mathcal{T}_h}(\frac{|\kappa_K|}{\kappa_{r,K}}, \frac{|\nu_K|}{\nu_{r,K}})$.

Lemma 45.5 (Consistency/boundedness). *Let the consistency error be defined by*

$$\langle\delta_h(\boldsymbol{a}_h), \boldsymbol{b}_h\rangle_{\boldsymbol{V}_h', \boldsymbol{V}_h} := \ell(\boldsymbol{b}_h) - a_{\nu,\kappa,h}(\boldsymbol{a}_h, \boldsymbol{b}_h), \qquad \forall\boldsymbol{a}_h, \boldsymbol{b}_h \in \boldsymbol{V}_h.$$

There is $\omega_\sharp$, uniform w.r.t. $\boldsymbol{A}$ and κ, s.t. $\|\delta_h(\boldsymbol{a}_h)\|_{\boldsymbol{V}_h'} \leq \omega_\sharp\|\boldsymbol{A}-\boldsymbol{a}_h\|_{\boldsymbol{V}_\sharp}$ for all $\boldsymbol{a}_h \in \boldsymbol{V}_h$ and all $h \in \mathcal{H}$.

Proof. See Exercise 45.2 $\qquad\square$

Theorem 45.6 (Error estimate). *Let $\boldsymbol{A}$ solve (45.1). Let $\boldsymbol{A}_h$ solve (45.2) with the penalty parameter η_0 as in Lemma 45.1. Assume that the smoothness property (43.15) holds true with $r > 0$. (i) There is c, which can depend on $\max_{K\in\mathcal{T}_h}(\frac{|\kappa_K|}{\kappa_{r,K}}, \frac{|\nu_K|}{\nu_{r,K}})$, such that for all $h \in \mathcal{H}$,*

$$\|\boldsymbol{A}-\boldsymbol{A}_h\|_{\boldsymbol{V}_\sharp} \leq c \inf_{\boldsymbol{a}_h\in\boldsymbol{V}_h}\|\boldsymbol{A}-\boldsymbol{a}_h\|_{\boldsymbol{V}_\sharp}. \tag{45.16}$$

(ii) Letting $t = \min(r, k+1)$, $\chi_t := 1$ if $t \leq 1$ and $\chi_t := 0$ if $t > 1$, we have

$$\|\boldsymbol{A}-\boldsymbol{A}_h\|_{\boldsymbol{V}_\sharp} \leq c\left((h/\ell_D)^t(\nu_\sharp^{\frac{1}{2}}\|\boldsymbol{A}\|_{\boldsymbol{H}^t(D)} + \kappa_\sharp^{\frac{1}{2}}\|\nabla\times\boldsymbol{A}\|_{\boldsymbol{H}^t(D)})\right.$$

$$\left.+ \chi_t\kappa_\sharp^{\frac{1}{2}}h\|\kappa^{-1}(\boldsymbol{f}-\nu\boldsymbol{A})\|_{\boldsymbol{L}^2(D)}\right). \tag{45.17}$$

Proof. (i) Lemma 45.3 implies that $\boldsymbol{A} \in \boldsymbol{V}_s$ with $q := 2$. Then the error estimate (45.16) follows from Lemma 27.5 combined with stability (Lemma 45.1) and consistency/boundedness (Lemma 45.5).

(ii) To prove (45.17), we bound the infimum in (45.16) by taking $\boldsymbol{a}_h := \mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{A})$, where $\mathcal{J}_{h0}^{\mathrm{c}}$ is the commuting quasi-interpolation operator with boundary prescription from Chapter 23, which we take of degree $\ell := \lceil t \rceil - 1$. Note that contrary to the proof of Theorem 41.8, here the best-approximation error is estimated by using an interpolant with boundary prescription so as to facilitate the estimate on the boundary penalty seminorm. Since $\ell < t \leq k+1$, we have $\ell \leq k$, i.e., $\boldsymbol{a}_h \in \boldsymbol{P}_{\ell,0}^{\mathrm{c}}(\mathcal{T}_h) \subset \boldsymbol{V}_h$. Let us set $\boldsymbol{\eta} := \boldsymbol{A} - \boldsymbol{a}_h$, so that we need to bound the five terms composing $\|\boldsymbol{\eta}\|_{\boldsymbol{V}_\sharp}$ (see (45.15)). For the first term, we have

$$\left(\sum_{K \in \mathcal{T}_h} \frac{|\nu_K|^2}{\nu_{r,K}} \|\boldsymbol{\eta}\|_{\boldsymbol{L}^2(K)}^2 \right)^{\frac{1}{2}} \leq c\, \nu_\sharp^{\frac{1}{2}} \|\boldsymbol{\eta}\|_{\boldsymbol{L}^2(D)} \leq c'\, \nu_\sharp^{\frac{1}{2}} \inf_{\boldsymbol{b}_h \in \boldsymbol{P}_{\ell,0}^{\mathrm{c}}(\mathcal{T}_h)} \|\boldsymbol{A} - \boldsymbol{b}_h\|_{\boldsymbol{L}^2(D)}$$

$$\leq c''\, \nu_\sharp^{\frac{1}{2}} (h/\ell_D)^t \|\boldsymbol{A}\|_{\boldsymbol{H}^t(D)},$$

where we used the approximation properties of $\mathcal{J}_h^{\mathrm{c}}$ from Theorem 23.12 in the first line and Corollary 22.16 together with $\ell < t$ in the second line. Considering the second term and using the commuting property $\nabla \times (\mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{a}_h)) = \mathcal{J}_{h0}^{\mathrm{d}}(\nabla \times \boldsymbol{a}_h)$, we infer that

$$\left(\sum_{K \in \mathcal{T}_h} \frac{|\kappa_K|^2}{\kappa_{r,K}} \|\nabla \times \boldsymbol{\eta}\|_{\boldsymbol{L}^2(K)}^2 \right)^{\frac{1}{2}} \leq c\, \kappa_\sharp^{\frac{1}{2}} (h/\ell_D)^t \|\nabla \times \boldsymbol{A}\|_{\boldsymbol{H}^t(D)}.$$

The estimate on the third term is straightforward since $\boldsymbol{\eta}_{|\partial D} \times \boldsymbol{n} = \boldsymbol{0}$. For the fourth term, we invoke the embedding inequality (41.16), and we infer that for all $K \in \mathcal{T}_h$,

$$h_K^{d(\frac{1}{2} - \frac{1}{p})} \|\nabla \times \boldsymbol{\eta}\|_{\boldsymbol{L}^p(K)} \leq c\, (\|\nabla \times \boldsymbol{\eta}\|_{\boldsymbol{L}^2(K)} + h_K^t |\nabla \times \boldsymbol{\eta}|_{\boldsymbol{H}^t(K)})$$

$$= c\, (\|\nabla \times \boldsymbol{\eta}\|_{\boldsymbol{L}^2(K)} + h_K^t |\nabla \times \boldsymbol{A}|_{\boldsymbol{H}^t(K)}),$$

since $\ell < t$ implies that $|\nabla \times \boldsymbol{\eta}|_{\boldsymbol{H}^t(K)} = |\nabla \times \boldsymbol{A}|_{\boldsymbol{H}^t(K)}$. Using the above bound on $\|\nabla \times \boldsymbol{\eta}\|_{\boldsymbol{L}^2(K)}$, together with $\kappa_{r,K} \leq |\kappa_K| \leq \kappa_\sharp$ and $|\nabla \times \boldsymbol{A}|_{\boldsymbol{H}^t(K)} \leq \ell_D^{-t} \|\nabla \times \boldsymbol{A}\|_{\boldsymbol{H}^t(K)}$ for all $K \in \mathcal{T}_h$, we infer that

$$\left(\sum_{F \in \mathcal{F}_h^\partial} \kappa_{r,K_l} h_{K_l}^{2d(\frac{1}{2} - \frac{1}{p})} \|\nabla \times \boldsymbol{\eta}\|_{\boldsymbol{L}^p(K_l)}^2 \right)^{\frac{1}{2}} \leq c\, \kappa_\sharp^{\frac{1}{2}} (h/\ell_D)^t \|\nabla \times \boldsymbol{A}\|_{\boldsymbol{H}^t(D)}.$$

Consider the fifth term, $(\sum_{F \in \mathcal{F}_h^\partial} \kappa_{r,K_l} h_{K_l}^2 \|\nabla \times (\nabla \times \boldsymbol{\eta})\|_{\boldsymbol{L}^2(K_l)}^2)^{\frac{1}{2}}$. If $t \leq 1$, we have $\ell = 0$, so that $\nabla \times (\nabla \times \boldsymbol{\eta}) = \nabla \times (\nabla \times \boldsymbol{A}) = \kappa^{-1}(\boldsymbol{f} - \nu \boldsymbol{A})$ in each cell K_l since κ is piecewise constant. If $t > 1$, using $\|\nabla \times (\nabla \times \boldsymbol{\eta})\|_{\boldsymbol{L}^2(K_l)} \leq 2|\nabla \times \boldsymbol{\eta}|_{\boldsymbol{H}^1(K_l)}$ (see Exercise 43.2), and owing to the commuting property $\nabla \times (\mathcal{J}_{h0}^{\mathrm{c}}(\boldsymbol{a}_h)) = \mathcal{J}_{h0}^{\mathrm{d}}(\nabla \times \boldsymbol{a}_h)$, we obtain $\|\nabla \times (\nabla \times \boldsymbol{\eta})\|_{\boldsymbol{L}^2(K_l)} \leq 2|\nabla \times \boldsymbol{A} - \mathcal{J}_{h0}^{\mathrm{d}}(\nabla \times \boldsymbol{A})|_{\boldsymbol{H}^1(K_l)}$. Since $t > 1$, we infer that

$$\left(\sum_{F \in \mathcal{F}_h^\partial} \kappa_{r,K_l} h_{K_l}^2 \|\nabla \times (\nabla \times \boldsymbol{\eta})\|_{\boldsymbol{L}^2(K_l)}^2 \right)^{\frac{1}{2}} \leq c\kappa_\sharp^{\frac{1}{2}} (h/\ell_D)^t \|\nabla \times \boldsymbol{A}\|_{\boldsymbol{H}^t(D)}.$$

Collecting the above estimates leads to the assertion. $\qquad\square$

Remark 45.7 (Estimate (45.17)**).** If $r > \frac{1}{2}$, then $t > \frac{1}{2}$ and the terms $\ell_D^{-t} \|\cdot\|_{\boldsymbol{H}^t(D)}$ can be replaced by $|\cdot|_{\boldsymbol{H}^t(D)}$ in (45.17). These terms can also be localized as a sum over the mesh cells. $\square$

45.3   Boundary penalty method in H^1

One can also combine Nitsche's boundary penalty method with the use of $\boldsymbol{H}^1$-conforming finite elements. The discrete problem is still (45.2), but the discrete trial and test space is now $\boldsymbol{V}_h :=$ $\boldsymbol{P}_k^{\mathrm{g}}(\mathcal{T}_h) := [P_k^{\mathrm{g}}(\mathcal{T}_h)]^3$, $k \geq 1$, where $P_k^{\mathrm{g}}(\mathcal{T}_h)$ is the scalar-valued H^1-conforming finite element space built in Chapter 19. Letting $(\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3)$ be the canonical basis of $\mathbb{R}^3$ and $\{\varphi_i\}_{i \in \mathcal{A}_h}$ be the global shape functions in $P_k^{\mathrm{g}}(\mathcal{T}_h)$, the basis that we use for $\boldsymbol{P}_k^{\mathrm{g}}(\mathcal{T}_h)$ is $\{\varphi_i \boldsymbol{e}_k\}_{i \in \mathcal{A}_h, k \in \{1:3\}}$. Notice that working with the $\boldsymbol{H}^1$-conforming space $\boldsymbol{V}_h$ leads to a collocalized scheme, i.e., the three components of the discrete field $\boldsymbol{A}_h$ are associated with the same global shape function.

Invoking stability and consistency/boundedness arguments as in §45.2 leads to a quasi-optimal error estimate that is identical to (45.16), except that the best-approximation error is measured with respect to a smaller discrete space since $\boldsymbol{H}^1$-conformity is required. To bound this error, we follow the arguments from Bonito and Guermond [69], Bonito et al. [71], where a mollification operator is considered. For simplicity, we focus on the case with mild smoothness where $r \in (0, \frac{1}{2})$ in (45.11), so that the optimal choice for the polynomial degree is $k := 1$.

Corollary 45.8 (Convergence). *Let $\boldsymbol{A}$ solve (45.1) and assume that (45.11) holds true with $r \in (0, \frac{1}{2})$. Assume that the mesh sequence is quasi-uniform and that the polynomial degree is $k := 1$. There are $h_0 > 0$ and c s.t. for all $h \in \mathcal{H} \cap (0, h_0]$,*

$$\|\boldsymbol{A} - \boldsymbol{A}_h\|_{\boldsymbol{V}_\sharp} \leq c \left((h/\ell_D)^{\frac{r}{2}} \big(\kappa_\sharp^{\frac{1}{2}} \|\nabla \times \boldsymbol{A}\|_{\boldsymbol{H}^r(D)} + (\kappa_\sharp^{\frac{1}{2}} \ell_D^{-1} + \nu_\sharp^{\frac{1}{2}}) \|\boldsymbol{A}\|_{\boldsymbol{H}^r(D)} \big) \right.$$

$$\left. + \kappa_\sharp^{\frac{1}{2}} h \|\kappa^{-1}(\boldsymbol{f} - \nu \boldsymbol{A})\|_{\boldsymbol{L}^2(D)} \right). \tag{45.18}$$

Proof. To prove (45.18), we bound the infimum in (45.16) by taking $\boldsymbol{a}_h := \mathcal{I}_{h,0}^{\mathrm{g}}(\mathcal{K}_{\delta,0}^{\mathrm{c}}(\boldsymbol{A}))$, where $\mathcal{I}_{h,0}^{\mathrm{g}}$ is the canonical interpolation operator associated with $\boldsymbol{P}_{1,0}^{\mathrm{g}}(\mathcal{T}_h) \subset \boldsymbol{P}_1^{\mathrm{g}}(\mathcal{T}_h)$ and $\mathcal{K}_{\delta,0}^{\mathrm{c}} : \boldsymbol{L}^1(D) \to \boldsymbol{C}_0^\infty(D)$ is the mollification operator from §23.4. To simplify the notation, we use here the parameter δ as a length scale (that is, it corresponds to the length scale $2\zeta\delta$ from Lemma 23.15). We take $\delta := (\ell_D h)^{\frac{1}{2}}$, and we assume that h is small enough so that $\mathcal{K}_{\delta,0}^{\mathrm{c}}$ is well defined and Lemma 23.15 can be applied. Using the inverse inequality from the hint of Exercise 23.8 and the approximation properties of $\mathcal{K}_{\delta,0}^{\mathrm{c}}$ from Corollary 23.5, we have $\delta^2 |\mathcal{K}_{\delta,0}^{\mathrm{c}}(\boldsymbol{A})|_{\boldsymbol{H}^2(D)} \leq c\delta^r \ell_D^{-r} \|\boldsymbol{A}\|_{\boldsymbol{H}^r(D)}$. Letting $\boldsymbol{\theta}_h := \boldsymbol{A} - \boldsymbol{a}_h$, we need to bound the five terms composing $\|\boldsymbol{\theta}_h\|_{\boldsymbol{V}_\sharp}$ (see (45.15)). Considering the second term, we have

$$\|\nabla \times \boldsymbol{\theta}_h\|_{\boldsymbol{L}^2(D)} \leq \|\nabla \times (\boldsymbol{A} - \mathcal{K}_{\delta,0}^{\mathrm{c}}(\boldsymbol{A}))\|_{\boldsymbol{L}^2(D)} + \|\nabla \times (\mathcal{K}_{\delta,0}^{\mathrm{c}}(\boldsymbol{A}) - \boldsymbol{a}_h)\|_{\boldsymbol{L}^2(D)}$$

$$\leq \|\nabla \times \boldsymbol{A} - \mathcal{K}_{\delta,0}^{\mathrm{d}}(\nabla \times \boldsymbol{A})\|_{\boldsymbol{L}^2(D)} + 2|\mathcal{K}_{\delta,0}^{\mathrm{c}}(\boldsymbol{A}) - \mathcal{I}_{h,0}^{\mathrm{g}}(\mathcal{K}_{\delta,0}^{\mathrm{c}}(\boldsymbol{A}))|_{\boldsymbol{H}^1(D)}$$

$$\leq c\big(\delta^r \ell_D^{-r} \|\nabla \times \boldsymbol{A}\|_{\boldsymbol{H}^r(D)} + h|\mathcal{K}_{\delta,0}^{\mathrm{c}}(\boldsymbol{A})|_{\boldsymbol{H}^2(D)}\big)$$

$$\leq c\big(\delta^r \ell_D^{-r} \|\nabla \times \boldsymbol{A}\|_{\boldsymbol{H}^r(D)} + h\delta^{r-2} \ell_D^{-r} \|\boldsymbol{A}\|_{\boldsymbol{H}^r(D)}\big)$$

$$= c\,(h/\ell_D)^{\frac{r}{2}} \big(\|\nabla \times \boldsymbol{A}\|_{\boldsymbol{H}^r(D)} + \ell_D^{-1} \|\boldsymbol{A}\|_{\boldsymbol{H}^r(D)}\big),$$

where we used Exercise 43.2 in the second line. The bound on the first term concerning $\nu_\sharp^{\frac{1}{2}} \|\boldsymbol{\theta}_h\|_{\boldsymbol{L}^2(D)}$ is similar. The estimate on the third term related to the boundary penalty term is zero. To bound the fourth and fifth terms, we proceed as in the proof of Theorem 45.6 using $k = 1$. $\square$

Remark 45.9 (Estimate (45.18)). We refer the reader to [69, 71] for further developments. The error estimate (45.18) is only of order $h^{\frac{r}{2}}$. This is the price to pay to invoke only the smoothness of $\nabla \times \boldsymbol{A}$ when using $\boldsymbol{H}^1$-conforming elements. $\square$

45.4 H^1-approximation with divergence control

We now consider the model problem (45.1) when ν is significantly smaller than $\kappa\ell_D^{-2}$ and $\nabla\cdot\boldsymbol{f} = 0$. Recall that this situation was successfully treated in the previous chapter in the context of $\boldsymbol{H}(\mathrm{curl})$-conforming edge elements. In the present section, we consider instead an approximation using $\boldsymbol{H}^1$-conforming elements. For simplicity, we assume that ν is constant, so that the solution to (45.1) satisfies $\nabla\cdot\boldsymbol{A} = 0$. Moreover, we are going to enforce the boundary condition strongly by considering the discrete trial and test space $\boldsymbol{V}_{h0} := \boldsymbol{P}_k^{\mathrm{g}}(\mathcal{T}_h) \cap \boldsymbol{H}_0(\mathrm{curl}; D)$, $k \geq 1$, with $\boldsymbol{P}_k^{\mathrm{g}}(\mathcal{T}_h) := [P_k^{\mathrm{g}}(\mathcal{T}_h)]^3$. Working with the discrete space $\boldsymbol{V}_{h0}$ is viable provided the faces of D are parallel to the canonical Cartesian planes in $\mathbb{R}^3$, so that the boundary condition does not couple the Cartesian components. Recall also that working with $\boldsymbol{P}_k^{\mathrm{g}}(\mathcal{T}_h)$ leads to a collocalized scheme, i.e., the three components of the discrete field $\boldsymbol{A}_h$ are associated with the same global shape function.

45.4.1   A least-squares technique

In the context of edge elements, a weak discrete control on the divergence of $\boldsymbol{A}_h \in \boldsymbol{P}_{k,0}^{\mathrm{c}}(\mathcal{T}_h)$ was achieved by using that $(\nu\boldsymbol{A}_h, \nabla q_h)_{\boldsymbol{L}^2(D)} = 0$ for all $q_h \in P_{k+1,0}^{\mathrm{g}}(\mathcal{T}_h)$. When using $\boldsymbol{H}^1$-conforming elements, it is no longer legitimate to use ∇q_h as a test function. This difficulty can be circumvented by employing a least-squares technique to control the divergence of $\boldsymbol{A}_h$ since $\nabla\cdot\boldsymbol{A}_h$ is an integrable function when $\boldsymbol{A}_h$ is discretized with $\boldsymbol{H}^1$-conforming elements.

The functional setting we have in mind hinges on the space

$$\boldsymbol{Z} := \boldsymbol{H}(\mathrm{curl}; D) \cap \boldsymbol{H}(\mathrm{div}; D), \tag{45.19}$$

with the norm $\|\boldsymbol{b}\|_{\boldsymbol{Z}} := (\|\boldsymbol{b}\|_{\boldsymbol{L}^2(D)}^2 + \ell_D^2\|\nabla\times\boldsymbol{b}\|_{\boldsymbol{L}^2(D)}^2 + \ell_D^2\|\nabla\cdot\boldsymbol{b}\|_{L^2(D)}^2)^{\frac{1}{2}}$. Consider the closed subspace $\boldsymbol{Z}_0 := \boldsymbol{H}_0(\mathrm{curl}; D) \cap \boldsymbol{H}(\mathrm{div}; D)$ and the following weak formulation:

$$\begin{cases} \text{Find } \boldsymbol{A} \in \boldsymbol{Z}_0 \text{ such that} \\ a_{\nu,\kappa,\eta}(\boldsymbol{A}, \boldsymbol{b}) = \ell(\boldsymbol{b}), \quad \forall \boldsymbol{b} \in \boldsymbol{Z}_0, \end{cases} \tag{45.20}$$

with $a_{\nu,\kappa,\eta} := a_{\nu,\kappa} + a_\eta$ and $a_\eta(\boldsymbol{a}, \boldsymbol{b}) := \eta e^{-\mathrm{i}\theta}\kappa_\flat(\nabla\cdot\boldsymbol{a}, \nabla\cdot\boldsymbol{b})_{L^2(D)}$, where $\eta > 0$ is a user-defined penalty parameter.

Proposition 45.10 (Equivalence). $\boldsymbol{A}$ *solves* (45.1) *iff* $\boldsymbol{A}$ *solves* (45.20).

Proof. See Exercise 45.3(i).   □

Owing to the Cauchy–Schwarz inequality, the sesquilinear form $a_{\nu,\kappa,\eta}$ satisfies the following boundedness property:

$$|a_{\nu,\kappa,\eta}(\boldsymbol{a}, \boldsymbol{b})| \leq \max(\nu_\sharp, \kappa_\sharp\ell_D^{-2}, \eta\kappa_\flat\ell_D^{-2})\|\boldsymbol{a}\|_{\boldsymbol{Z}}\|\boldsymbol{b}\|_{\boldsymbol{Z}}, \tag{45.21}$$

for all $\boldsymbol{a}, \boldsymbol{b} \in \boldsymbol{Z}$. Moreover, owing to the following Poincaré–Steklov inequality (see Exercise 45.3(ii)), we have

$$\hat{C}_{\mathrm{PS}}''\ell_D^{-1}\|\boldsymbol{b}\|_{\boldsymbol{L}^2(D)} \leq \left(\|\nabla\times\boldsymbol{b}\|_{\boldsymbol{L}^2(D)}^2 + \|\nabla\cdot\boldsymbol{b}\|_{L^2(D)}^2\right)^{\frac{1}{2}}, \qquad \forall\boldsymbol{b} \in \boldsymbol{Z}_0, \tag{45.22}$$

where $\hat{C}_{\mathrm{PS}}'' > 0$ only depends on D, we infer that

$$\begin{aligned} \Re(e^{\mathrm{i}\theta}a_{\nu,\kappa,\eta}(\boldsymbol{b}, \boldsymbol{b})) &\geq \Re(e^{\mathrm{i}\theta}a_{\nu,\kappa}(\boldsymbol{b}, \boldsymbol{b})) + \eta\kappa_\flat\|\nabla\cdot\boldsymbol{b}\|_{L^2(D)}^2 \\ &\geq \min(1, \eta)\kappa_\flat(\|\nabla\times\boldsymbol{b}\|_{\boldsymbol{L}^2(D)}^2 + \|\nabla\cdot\boldsymbol{b}\|_{L^2(D)}^2) \\ &\geq \frac{1}{2}\min(1, \eta)\kappa_\flat\ell_D^{-2}\min(1, (\hat{C}_{\mathrm{PS}}'')^2)\|\boldsymbol{b}\|_{\boldsymbol{Z}}^2, \end{aligned} \tag{45.23}$$

for all $\boldsymbol{b} \in \boldsymbol{Z}_0$, thus proving the ν-robust coercivity of $a_{\nu,\kappa,\eta}$ on $\boldsymbol{Z}_0$.

A conforming approximation of the model problem (45.20) is realized using $\boldsymbol{H}^1$-conforming elements as follows:

$$\begin{cases} \text{Find } \boldsymbol{A}_h \in \boldsymbol{V}_{h0} := \boldsymbol{P}_k^{\mathrm{g}}(\mathcal{T}_h) \cap \boldsymbol{H}_0(\mathrm{curl}; D) \text{ such that} \\ a_{\nu,\kappa,\eta}(\boldsymbol{A}_h, \boldsymbol{b}_h) = \ell(\boldsymbol{b}_h), \quad \forall \boldsymbol{b}_h \in \boldsymbol{V}_{h0}. \end{cases} \tag{45.24}$$

Since the approximation setting is conforming and the Galerkin orthogonality property holds true, the basic error estimate (26.18) combined with the above boundedness and coercivity properties of $a_{\nu,\kappa,\eta}$ yields

$$\|\boldsymbol{A} - \boldsymbol{A}_h\|_{\boldsymbol{Z}} \leq c \inf_{\boldsymbol{b}_h \in \boldsymbol{V}_{h0}} \|\boldsymbol{A} - \boldsymbol{b}_h\|_{\boldsymbol{Z}}, \tag{45.25}$$

where c is uniform w.r.t. $\nu_\flat^{-1}$. Assuming that $\boldsymbol{A} \in \boldsymbol{H}^{1+r}(D) \cap \boldsymbol{H}_0(\mathrm{curl}; D)$, $0 \leq r \leq k$, and using the approximation results from §19.3 and §19.4, we infer the optimal error estimate $\|\boldsymbol{A} - \boldsymbol{A}_h\|_{\boldsymbol{Z}} \leq c\ell_D h^r |\boldsymbol{A}|_{\boldsymbol{H}^{1+r}(D)}$.

45.4.2 The approximability obstruction

The $\boldsymbol{H}(\mathrm{curl})$-conforming approximation method (based on edge elements) converges optimally when $\boldsymbol{A} \in \boldsymbol{H}^r(D)$, $\nabla \times \boldsymbol{A} \in \boldsymbol{H}^r(D)$ and $r \geq 0$. In contrast, the $\boldsymbol{H}^1$-conforming least-squares technique converges optimally when $\boldsymbol{A} \in \boldsymbol{Z}_0 \cap \boldsymbol{H}^r(D)$ with $r \geq 1$, but it may fail to converge if $\boldsymbol{A}$ is just in $\boldsymbol{Z}_0 \cap \boldsymbol{H}^r(D)$ with $r < 1$, which can be the case when D is a nonconvex polyhedron. The bottleneck comes from the approximability property, i.e., $\inf_{\boldsymbol{b}_h \in \boldsymbol{V}_{h0}} \|\boldsymbol{A} - \boldsymbol{b}_h\|_{\boldsymbol{Z}}$ may not go to zero as $h \to 0$.

Lemma 45.11 (Costabel). *Let D be a nonconvex polyhedron. Then the space $\boldsymbol{H}_0(\mathrm{curl}; D) \cap \boldsymbol{H}^1(D)$ is a closed proper subspace of $\boldsymbol{Z}_0 := \boldsymbol{H}_0(\mathrm{curl}; D) \cap \boldsymbol{H}(\mathrm{div}; D)$, and the space $\boldsymbol{H}_0(\mathrm{div}; D) \cap \boldsymbol{H}^1(D)$ is a closed proper subspace of $\boldsymbol{H}_0(\mathrm{div}; D) \cap \boldsymbol{H}(\mathrm{curl}; D)$ (also equipped with the $\|\cdot\|_{\boldsymbol{Z}}$-norm).*

Proof. See [143, p. 541] and [144, Cor. 2.5]. □

Corollary 45.12 (Approximability obstruction). *Let D be a nonconvex polyhedron. Then $(\boldsymbol{V}_{h0})_{h \in \mathcal{H}}$ cannot have the approximability property in $\boldsymbol{Z}_0$.*

Proof. Since $\boldsymbol{V}_{h0} \subset \boldsymbol{H}_0(\mathrm{curl}; D) \cap \boldsymbol{H}^1(D) =: \boldsymbol{Z}_{0,1}$ and $\boldsymbol{Z}_{0,1}$ is closed in $\boldsymbol{Z}_0$, the limit of all the Cauchy sequences in $\boldsymbol{V}_{h0}$ are in $\boldsymbol{Z}_{0,1}$. Moreover, Lemma 45.11 implies that there are functions of $\boldsymbol{Z}_0$ that lie at a positive distance from $\boldsymbol{Z}_{0,1}$. Thus, Cauchy sequences in $\boldsymbol{V}_{h0} \subset \boldsymbol{Z}_{0,1}$ cannot reach these functions. □

The striking consequence of the above arguments is that using $\boldsymbol{H}^1$-conforming finite elements together with the formulation (45.24) produces a method that is convergent if the solution to (45.1) is at least in $\boldsymbol{H}^1(D)$, as it happens if D is a convex polyhedron (see Lemma 43.3(ii)), but the method may fail to converge if it is used to approximate fields that are not in $\boldsymbol{H}^1(D)$. This example shows that there are situations where the approximability property should not be treated too lightly.

Remark 45.13 (Beyond the approximability obstruction). The source of the problem is that the L^2-based least-squares penalty on $\nabla \cdot \boldsymbol{A}_h$ is too strong. Convergence can be recovered by weakening this control. For instance, one can consider the sesquilinear form $(\boldsymbol{a}, \boldsymbol{b}) \mapsto \eta e^{-i\theta} \int_D d^\gamma(\boldsymbol{x}) \kappa_\flat \nabla \cdot \boldsymbol{a} \nabla \cdot \overline{\boldsymbol{b}} \, dx$, where d is the distance to the set of the reentrant edges of D (assumed to be a three-dimensional polyhedron) and $\gamma > 0$ depends on the strength of the singularities induced by the reentrant edges; see Costabel and Dauge [145], Buffa et al. [96]. An alternative

method developed in Bramble and Pasciak [79], Bramble et al. [82] involves a least-squares approximation of a discrete problem with different test and trial spaces. The numerical method uses piecewise constant functions for the trial space and piecewise linear functions enriched with face bubbles for the test space. Furthermore, a technique based on a local L^2-stabilization of the divergence is introduced in Duan et al. [175, 176]. A method based on the stabilization of the divergence in $H^{-\alpha}(D)$ with $\alpha \in (\frac{1}{2}, 1)$ has been proposed in Bonito and Guermond [69], Bonito et al. [71]. All these methods have been proved to be quasi-optimal for solving the boundary value problem (45.1) and for solving the corresponding eigenvalue problem (see Chapter 46). A similar method, where the stabilization is done in $H^{-1}(D)$, has been proposed in Badia and Codina [43]. However, as reported in [71, §6.4], it seems that controlling the divergence in $H^{-1}(D)$ may not be sufficient to guarantee that the spectrum of the Maxwell operator is well approximated. $\square$

Exercises

Exercise 45.1 (Identity for $n_\sharp$). Prove (45.13b). (*Hint*: use the mollification operators $\mathcal{K}_\delta^c : \boldsymbol{L}^1(D) \to \boldsymbol{C}^\infty(\overline{D})$ and $\mathcal{K}_\delta^d : \boldsymbol{L}^1(D) \to \boldsymbol{C}^\infty(\overline{D})$ from §23.1, and adapt the proof of Lemma 40.5.)

Exercise 45.2 (Consistency/boundedness). Prove Lemma 45.5. (*Hint*: adapt the proof of Lemma 41.7 and use Lemma 45.4.)

Exercise 45.3 (Least-squares penalty on divergence). (i) Prove Proposition 45.10. (*Hint*: use Lemma 44.1 to write $\boldsymbol{A} := \boldsymbol{A}_0 + \nabla p$, where $\boldsymbol{A}_0$ is divergence-free and $p \in H_0^1(D)$, and prove that $p = 0$.) (ii) Prove (45.22). (*Hint*: use Lemma 44.4 for $\boldsymbol{A} - \nabla p$.)

Chapter 46

Symmetric elliptic eigenvalue problems

The three chapters composing Part X deal with the finite element approximation of the spectrum of elliptic differential operators. Ellipticity is crucial here to provide a compactness property that guarantees that the spectrum of the operators in question is well structured. We start by recalling fundamental results on compact operators and symmetric operators in Hilbert spaces. Then, we study the finite element approximation of the spectrum of compact operators. We first focus on the H^1-conforming approximation of symmetric operators, then we treat the (possibly nonconforming) approximation of nonsymmetric operators.

The present chapter contains a brief introduction to the spectral theory of compact operators together with illustrative examples. Eigenvalue problems occur when analyzing the response of devices, buildings, or vehicles to vibrations, or when performing the linear stability analysis of dynamical systems.

46.1   Spectral theory

We briefly recall in this section some essential facts regarding the spectral theory of linear operators. Most of the proofs are omitted since the material is classical and can be found in Brezis [89, Chap. 6], Chatelin [116, pp. 95-120], Dunford and Schwartz [179, Part I, pp. 577-580], Lax [278, Chap. 21&32], Kreyszig [271, pp. 365-521]. In the entire section, L is a complex Banach space, we use the shorthand notation $\mathcal{L}(L) := \mathcal{L}(L; L)$, and I_L denotes the identity operator in L.

46.1.1   Basic notions and examples

Definition 46.1 (Resolvent, spectrum, eigenvalues, eigenvectors). *Let $T \in \mathcal{L}(L)$. The resolvent set of T, $\rho(T)$, and the spectrum of T, $\sigma(T)$, are subsets of $\mathbb{C}$ defined as follows:*

$$\rho(T) := \{\mu \in \mathbb{C} \mid \mu I_L - T \text{ is bijective}\}, \tag{46.1a}$$

$$\sigma(T) := \mathbb{C}\backslash\rho(T) = \{\mu \in \mathbb{C} \mid \mu I_L - T \text{ is not bijective}\}. \tag{46.1b}$$

(Since L is a Banach space, $\mu \in \rho(T)$ iff $(\mu I_L - T)^{-1} \in \mathcal{L}(L)$.) The spectrum of T is decomposed into the following disjoint union:

$$\sigma(T) = \sigma_{\mathrm{p}}(T) \cup \sigma_{\mathrm{c}}(T) \cup \sigma_{\mathrm{r}}(T), \tag{46.2}$$

where the point spectrum*, $\sigma_{\mathrm{p}}(T)$, the* continuous spectrum*, $\sigma_{\mathrm{c}}(T)$, and the* residual spectrum*, $\sigma_{\mathrm{r}}(T)$, are defined as follows:*

$$\sigma_{\mathrm{p}}(T) := \{\mu \in \mathbb{C} \mid \mu I_L - T \text{ is not injective}\},$$
$$\sigma_{\mathrm{c}}(T) := \{\mu \in \mathbb{C} \mid \mu I_L - T \text{ is injective, not surjective, } \overline{\mathrm{im}(\mu I_L - T)} = L\},$$
$$\sigma_{\mathrm{r}}(T) := \{\mu \in \mathbb{C} \mid \mu I_L - T \text{ is injective, not surjective, } \overline{\mathrm{im}(\mu I_L - T)} \neq L\}.$$

Whenever $\sigma_{\mathrm{p}}(T)$ is nonempty, members of $\sigma_{\mathrm{p}}(T)$ are called eigenvalues*, and the nonzero vectors in $\ker(\mu I_L - T)$ are called* eigenvectors *associated with μ, i.e., $0 \neq z \in L$ is an eigenvector associated with μ iff $T(z) = \mu z$.*

Example 46.2 (Finite dimension). *If L is finite-dimensional, $\ker(\mu I_L - T) \neq \{0\}$ iff $(\mu I_L - T)$ is not invertible. In this case, the spectrum of T only consists of eigenvalues, i.e., $\sigma(T) = \sigma_{\mathrm{p}}(T)$ and $\sigma_{\mathrm{c}}(T) = \sigma_{\mathrm{r}}(T) = \emptyset$.* $\qquad\square$

Example 46.3 (Volterra operator). *Let $L := L^2((0,1); \mathbb{C})$ and let us identify L and L' by setting $\langle l', l \rangle_{L',L} := \int_0^1 l'(x)\overline{l}(x)\,\mathrm{d}x$. Let $T : L \to L$ be s.t. $T(f)(x) := \int_0^x f(t)\,\mathrm{d}t$ for a.e. $x \in (0,1)$. We have $\rho(T) = \mathbb{C}\backslash\{0\}$, $\sigma_{\mathrm{p}}(T) = \emptyset$, $\sigma_{\mathrm{c}}(T) = \{0\}$, and $\sigma_{\mathrm{r}}(T) = \emptyset$; see Exercise 46.4.* $\qquad\square$

Theorem 46.4 (Spectral radius). *Let $T \in \mathcal{L}(L)$. (i) The subsets $\rho(T)$ and $\sigma(T)$ are both nonempty. (ii) $\sigma(T)$ is a compact subset of $\mathbb{C}$. (iii) Let*

$$r(T) := \max_{\mu \in \sigma(T)} |\mu| \tag{46.3}$$

be the spectral radius *of T. Then*

$$r(T) = \lim_{n \to \infty} \|T^n\|_{\mathcal{L}(L)}^{\frac{1}{n}} \leq \|T\|_{\mathcal{L}(L)}. \tag{46.4}$$

Proof. See Kreyszig [271], Thm. 7.5.4 for (i), Thm. 7.3.4 for (ii), and Thm. 7.5.5 for (iii). $\qquad\square$

Remark 46.5 ((46.4)). The identity $r(T) = \lim_{n \to \infty} \|T^n\|_{\mathcal{L}(L)}^{\frac{1}{n}}$ is often called Gelfand's formula (see [213, p. 11]). The inequality $\lim_{n \to \infty} \|T^n\|_{\mathcal{L}(L)}^{\frac{1}{n}} \leq \|T\|_{\mathcal{L}(L)}$ may sometimes be strict. For instance, $r(T) = 0$ if $\sigma(T) = \{0\}$, but it can happen in that case that $\|T\|_{\mathcal{L}(L)} > 0$. A simple example is the operator $T : \mathbb{R}^2 \to \mathbb{R}^2$ s.t. $T(X) := \mathbb{A}X$ with $\mathbb{A} := \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$. $\qquad\square$

Let us consider more specifically the eigenvalues of T. Assume that $\sigma_{\mathrm{p}}(T) \neq \emptyset$ and let $\mu \in \sigma_{\mathrm{p}}(T)$. Let us set $K_i := \ker(\mu I_L - T)^i$ for all $i \in \mathbb{N}\backslash\{0\}$. One readily verifies that the spaces K_i are invariant under T. Moreover, $K_1 \subset K_2 \ldots$, and if there is an integer $j \geq 1$ such that $K_j = K_{j+1}$, then $K_j = K_{j'}$ for all $j' > j$.

Definition 46.6 (Ascent, algebraic and geometric multiplicity). *Assume that $\sigma_{\mathrm{p}}(T) \neq \emptyset$ and let $\mu \in \sigma_{\mathrm{p}}(T)$. We say that μ has finite ascent if there is $j \in \mathbb{N} \setminus \{0\}$ such that $K_j = K_{j+1}$, and the smallest integer satisfying this property is called* ascent *of μ and is denoted by $\alpha(\mu)$ (or simply α). Moreover, if K_α is finite-dimensional, then the* algebraic multiplicity *of μ, say m, and the* geometric multiplicity *of μ, say g, are defined as follows:*

$$m := \dim(K_\alpha) \geq \dim(K_1) =: g. \tag{46.5}$$

Whenever $\alpha \geq 2$, nonzero vectors in K_α are called generalized eigenvectors*.*

If the eigenvalue μ has finite ascent α and if K_α is finite-dimensional, then elementary arguments from linear algebra show that $\alpha + g - 1 \le m \le \alpha g$ (note that $\alpha = 1$ iff $m = g$). This inequalities are proved by showing that $g_1 + i - 1 \le g_i \le g_{i-1} + g_1$ for all $i \in \{1{:}\alpha\}$ with $g_i := \dim(K_i)$; see Exercise 46.2. All the eigenvalues have a finite ascent and a finite multiplicity if L is finite-dimensional, or if the operator T is compact (see Theorem 46.14(iv)), but this may not be the case in general.

Example 46.7 (Ascent, algebraic and geometric multiplicity). To illustrate Definition 46.6 in a finite-dimensional setting, we consider the operator $T : \mathbb{R}^4 \to \mathbb{R}^4$ defined by $T(X) := \mathbb{A}X$ for all $X \in L := \mathbb{R}^4$, where

$$
\mathbb{A} := \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.
$$

Then $\mu = 1$ is the only eigenvalue of T. Since

$$
\mathbb{I}_4 - \mathbb{A} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad (\mathbb{I}_4 - \mathbb{A})^2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad (\mathbb{I}_4 - \mathbb{A})^3 = \mathbb{O}_4,
$$

we have $\ker(I_L - T) = \operatorname{span}\{e_1, e_4\}$, $\ker(I_L - T)^2 = \operatorname{span}\{e_1, e_2, e_4\}$, and $\ker(I_L - T)^3 = \ker(I_L - T)^4 = \operatorname{span}\{e_1, e_2, e_3, e_4\}$, where $\{e_1, e_2, e_3, e_4\}$ is the canonical Cartesian basis of $\mathbb{R}^4$. Thus, the ascent of $\mu = 1$ is $\alpha = 3$, its algebraic multiplicity is $m = \dim(\ker(I_L - T)^3) = 4$, and its geometric multiplicity is $g = \dim(\ker(I_L - T)) = 2$. Notice that $\alpha + g - 1 = 4 = m \le 6 = \alpha g$. $\square$

Let us finally explore the relation between the spectrum of T and that of its adjoint $T^* : L' \to L'$ s.t. $\langle T^*(l'), l \rangle_{L', L} := \langle l', T(l) \rangle_{L', L}$ for all $l \in L$ and all $l' \in L'$ (see Definition C.29). Recall that we have adopted the convention that dual spaces are composed of antilinear forms (see Definition A.11 and §C.4), so that $(\lambda T)^* = \overline{\lambda} T^*$ for all $\lambda \in \mathbb{C}$. (The reader should be aware that a usual convention in the mathematical physics literature is that dual spaces are composed of linear forms, in which case $(\lambda T)^* = \lambda T^*$.) Moreover, for any subset $A \subset \mathbb{C}$, we denote $\operatorname{conj}(A) := \{\mu \in \mathbb{C} \mid \overline{\mu} \in A\}$.

Lemma 46.8 (Spectrum of T^*). *Let $T \in (\mathcal{L})$. The following holds true:*

$$
\sigma(T^*) = \operatorname{conj}(\sigma(T)), \qquad \sigma_{\mathrm{r}}(T) \subset \operatorname{conj}(\sigma_{\mathrm{p}}(T^*)) \subset \sigma_{\mathrm{r}}(T) \cup \sigma_{\mathrm{p}}(T). \tag{46.6}
$$

Proof. Corollary C.52 implies that $\mu I_L - T$ is not bijective iff $(\mu I_L - T)^* = \overline{\mu} I_{L'} - T^*$ is not bijective. This proves the first equality. See Exercise 46.1 for the proof of the other two inclusions. $\square$

Example 46.9 (Left and right shifts). Let $p \in (1, \infty)$ and let ℓ^p be the Banach space composed of the complex-valued sequences $x := (x_n)_{n \in \mathbb{N}}$ s.t. $\sum_{n \in \mathbb{N}} |x_n|^p < \infty$. We can identify the dual space of ℓ^p with $\ell^{p'}$, where $\frac{1}{p} + \frac{1}{p'} = 1$, by setting $\langle x, y \rangle_{\ell^{p'}, \ell^p} := \sum_{n \in \mathbb{N}} x_n \overline{y_n}$ with $x := (x_n)_{n \in \mathbb{N}}$ and $y := (y_n)_{n \in \mathbb{N}}$. Consider the left shift operator $\mathsf{L} : \ell^{p'} \to \ell^{p'}$ defined by $\mathsf{L}(x) := (x_1, x_2, \ldots)$ and the right shift operator $\mathsf{R} : \ell^p \to \ell^p$ defined by $\mathsf{R}(x) := (0, x_0, x_1, \ldots)$. Then $\langle x, \mathsf{R}(y) \rangle_{\ell^{p'}, \ell^p} := \sum_{n \ge 1} x_n \overline{y_{n-1}} = \sum_{n \ge 0} x_{n+1} \overline{y_n} = \langle \mathsf{L}(x), y \rangle_{\ell^{p'}, \ell^p}$. This shows that $\mathsf{L} = \mathsf{R}^*$. Similarly, $\mathsf{R} = \mathsf{L}^*$ once the dual of $\ell^{p'}$ is identified with ℓ^p. Observe that $\|\mathsf{R}\|_{\mathcal{L}(\ell^p; \ell^p)} = \|\mathsf{L}\|_{\mathcal{L}(\ell^{p'}; \ell^{p'})} = 1$, so that both $\sigma(\mathsf{R})$ and $\sigma(\mathsf{L})$ are contained in the unit disk $\{\lambda \in \mathbb{C} \mid |\lambda| \le 1\}$ owing to Theorem 46.4(iii). Notice that $0 \not\in \sigma_{\mathrm{p}}(\mathsf{R})$ since R is injective. Assume that there exists $\mu \in \sigma_{\mathrm{p}}(\mathsf{R})$, i.e., there is a nonzero $x \in \ell^p$ s.t. $(\mu x_0, \mu x_1 - x_0, \mu x_2 - x_1, \ldots) = 0$. Then $x_n = 0$ for all $n \in \mathbb{N}$, i.e., $x = 0$, which is absurd (recall that

$\mu \neq 0$). Hence, $\sigma_{\mathrm{p}}(\mathsf{R}) = \emptyset$. Lemma 46.8 in turn implies that $\sigma_{\mathrm{r}}(\mathsf{L}) = \emptyset$ because $\mathsf{L}^* = \mathsf{R}$. Similarly, Lemma 46.8 implies that $\sigma_{\mathrm{r}}(\mathsf{R}) \subset \mathrm{conj}(\sigma_{\mathrm{p}}(\mathsf{L})) \subset \sigma_{\mathrm{r}}(\mathsf{R})$, i.e., $\sigma_{\mathrm{r}}(\mathsf{R}) = \mathrm{conj}(\sigma_{\mathrm{p}}(\mathsf{L}))$. Assuming that $\mu \in \sigma_{\mathrm{p}}(\mathsf{L})$, there is a nonzero vector $x \in \ell^{p'}$ s.t. $\mathsf{L}(x) = \mu x$, which means that $x = x_0(1, \mu, \mu^2, \ldots)$. This vector is in $\ell^{p'}$ iff $|\mu| < 1$. Hence, $\sigma_{\mathrm{p}}(\mathsf{L}) = \{\mu \in \mathbb{C} \mid |\mu| < 1\}$. Since $\sigma_{\mathrm{p}}(\mathsf{L})$ is invariant under complex conjugation, we conclude that $\sigma_{\mathrm{r}}(\mathsf{R}) = \sigma_{\mathrm{p}}(\mathsf{L})$. Finally, since $\sigma(\mathsf{L})$ is closed (see Theorem 46.4(ii)) and $\|\mathsf{L}\|_{\mathcal{L}(\ell^{p'};\ell^{p'})} = 1$, we have $\sigma(\mathsf{L}) \subset \{\mu \in \mathbb{C} \mid |\mu| \leq 1\}$. But $\sigma(\mathsf{L})$ must also contain the closure in $\mathbb{C}$ of $\sigma_{\mathrm{p}}(\mathsf{L}) = \{\mu \in \mathbb{C} \mid |\mu| < 1\}$. Hence, $\sigma(\mathsf{L}) = \{\mu \in \mathbb{C} \mid |\mu| \leq 1\}$. This, in turn, implies that $\sigma_{\mathrm{c}}(\mathsf{L}) = \{\mu \in \mathbb{C} \mid |\mu| = 1\}$. In conclusion, we have established that

$$
\begin{aligned}
\sigma_{\mathrm{p}}(\mathsf{L}) &= \{\mu \in \mathbb{C} \mid |\mu| < 1\} = \sigma_{\mathrm{r}}(\mathsf{R}), \\
\sigma_{\mathrm{c}}(\mathsf{L}) &= \{\mu \in \mathbb{C} \mid |\mu| = 1\} = \sigma_{\mathrm{c}}(\mathsf{R}), \\
\sigma_{\mathrm{r}}(\mathsf{L}) &= \emptyset = \sigma_{\mathrm{p}}(\mathsf{R}). \qquad \square
\end{aligned}
$$

46.1.2 Compact operators in Banach spaces

Since we are going to focus later our attention on the approximation of the eigenvalues and eigenspaces of compact operators, we now recall important facts about such operators. Given two Banach spaces V, W, we say that $T \in \mathcal{L}(V; W)$ is compact if T maps the unit ball of V into a relatively compact set in W (see Definition A.17). Let us also recall (see Theorem A.21) that if there exists a sequence $(T_n)_{n \in \mathbb{N}}$ of operators in $\mathcal{L}(V; W)$ of finite rank s.t. $\lim_{n \to \infty} \|T - T_n\|_{\mathcal{L}(V;W)} = 0$, then T is compact. Conversely, if W is a Hilbert space and $T \in \mathcal{L}(V; W)$ is a compact operator, then there exists a sequence of operators in $\mathcal{L}(V; W)$ of finite rank, $(T_n)_{n \in \mathbb{N}}$, such that $\lim_{n \to \infty} \|T - T_n\|_{\mathcal{L}(V;W)} = 0$.

Example 46.10 (Rellich–Kondrachov). For every bounded Lipschitz domain D, the Rellich–Kondrachov theorem states that the injection $W^{s,p}(D) \hookrightarrow L^q(D)$ is compact for all $q \in [1, \frac{pd}{d-sp})$ if $sp \leq d$ (see Theorem 2.35). $\qquad\square$

Example 46.11 (Hilbert–Schmidt operators). Let $K \in L^2(D \times D; \mathbb{C})$, where D is a bounded set in $\mathbb{R}^d$. Then the Hilbert–Schmidt operator $T : L^2(D; \mathbb{C}) \to L^2(D; \mathbb{C})$ defined s.t. $T(f)(x) := \int_D f(y) K(x, y) \, \mathrm{d}y$ a.e. in D is compact (see Brezis [89, Thm. 6.12]). Note that $T^*(f)(x) := \int_D f(y) \overline{K(y, x)} \, \mathrm{d}y$. $\qquad\square$

Example 46.12 (Identity). The identity $I_{\ell^p} : \ell^p \to \ell^p$, $p \in [1, \infty]$, is not compact. Indeed, consider the sequence $e_n := (\delta_{mn})_{m \in \mathbb{N}}$. For all $N \geq 0$ and $n, m \geq N$, $n \neq m$, we have $\|e_n - e_m\|_{\ell^p} = 2^{\frac{1}{p}}$ for all $p \in [1, \infty)$, and $\|e_n - e_m\|_{\ell^\infty} = 1$. Hence, one cannot extract any Cauchy subsequence in ℓ^p from $(e_n)_{n \in \mathbb{N}}$. $\qquad\square$

Let us now state some important results on compact operators.

Theorem 46.13 (Fredholm alternative). *Let $T \in \mathcal{L}(L)$ be a compact operator. The following properties hold true for all $\mu \in \mathbb{C} \backslash \{0\}$:*

  (i) *$\mu I_L - T$ is injective iff $\mu I_L - T$ is surjective.*

  (ii) *$\ker(\mu I_L - T)$ is finite-dimensional.*

  (iii) *$\mathrm{im}(\mu I_L - T)$ is closed, i.e., $\mathrm{im}(\mu I_L - T) = \ker(\overline{\mu} I_{L'} - T^*)^\perp$.*

  (iv) *$\dim \ker(\mu I_L - T) = \dim \ker(\overline{\mu} I_{L'} - T^*)$.*

Proof. See Brezis [89, Thm. 6.6]. $\qquad\square$

The Fredholm alternative usually refers to Item (i), which implies that every nonzero member of the spectrum of T is an eigenvalue when T is compact. The key result for compact operators is the following theorem.

Theorem 46.14 (Spectrum of compact operators). *Let $T \in \mathcal{L}(L)$ be a compact operator with* $\dim(L) = \infty$. *The following holds true:*

(i) $0 \in \sigma(T)$.

(ii) $\sigma(T)\backslash\{0\} = \sigma_\mathrm{p}(T)\backslash\{0\}$.

(iii) *One of the following three cases holds:* (1) $\sigma(T) = \{0\}$; (2) $\sigma(T)\backslash\{0\}$ *is a finite set;* (3) $\sigma(T)\backslash\{0\}$ *is a sequence converging to* 0.

(iv) *Any $\mu \in \sigma(T)\backslash\{0\}$ has a finite ascent α, and the space $\ker(\mu I_L - T)^\alpha$ is finite-dimensional, i.e., μ has finite algebraic and geometric multiplicity.*

(v) *$\mu \in \sigma(T)$ iff $\overline{\mu} \in \sigma(T^*)$, i.e., $\sigma(T^*) = \mathrm{conj}(\sigma(T))$. The ascent, algebraic and geometric multiplicities of $\mu \in \sigma(T)\backslash\{0\}$ and of $\overline{\mu}$ are equal.*

Proof. See Brezis [89, Thm. 6.8], Lax [278, p. 238], or Kreyszig [271, Thm. 8.3.1 & 8.4.4] for (i)-(iii) and [271, Thm. 8.4.3] for (iv)-(v). $\qquad\square$

The first two items in Theorem 46.14 imply that either T is not injective (i.e., $0 \in \sigma_\mathrm{p}(T)$) and then $\sigma(T) = \sigma_\mathrm{p}(T)$ (and $\sigma_\mathrm{c}(T) = \sigma_\mathrm{r}(T) = \emptyset$), or T is injective (i.e., $0 \notin \sigma_\mathrm{p}(T)$) and then $\sigma(T) = \sigma_\mathrm{p}(T) \cup \{0\}$ (and $\sigma_\mathrm{c}(T) = \{0\}$, $\sigma_\mathrm{r}(T) = \emptyset$ or $\sigma_\mathrm{r}(T) = \{0\}$, $\sigma_\mathrm{c}(T) = \emptyset$).

46.1.3 Symmetric operators in Hilbert spaces

In this section, L denotes a complex Hilbert space. The reader is invited to review §C.3 for basic facts about Hilbert spaces. Let $T \in \mathcal{L}(L)$. The (Hermitian) transpose of T, say $T^\mathsf{H} \in \mathcal{L}(L)$, is defined by setting

$$(T^\mathsf{H}(w), v)_L := (w, T(v))_L, \qquad \forall v, w \in L. \tag{46.7}$$

Let $(J_L^\mathrm{RF})^{-1} : L' \to L$ be the Riesz–Fréchet representation operator (see Theorem C.24), that is, $((J_L^\mathrm{RF})^{-1}(l'), l)_L := \langle l', l \rangle_{L',L}$ for all $l' \in L'$ and $l \in L$. We recall that J_L^RF and $(J_L^\mathrm{RF})^{-1}$ are linear operators because we have chosen dual spaces to be composed of antilinear forms (see Exercise 46.5 and Remark C.26).

Lemma 46.15 (Transpose and adjoint). *Let $T \in \mathcal{L}(L)$. We have $T^\mathsf{H} = (J_L^\mathrm{RF})^{-1} \circ T^* \circ J_L^\mathrm{RF}$, and*

$$\sigma_\mathrm{p}(T^*) = \sigma_\mathrm{p}(T^\mathsf{H}), \quad \sigma_\mathrm{c}(T^*) = \sigma_\mathrm{c}(T^\mathsf{H}), \quad \sigma_\mathrm{r}(T^*) = \sigma_\mathrm{r}(T^\mathsf{H}). \tag{46.8}$$

Finally, if the duality paring is identified with the inner product of L, i.e., if L and L' are identified, we have $T^\mathsf{H} = T^*$.

Proof. The identities $((J_L^\mathrm{RF})^{-1}T^*(l'), l)_L = \langle T^*(l'), l \rangle_{L',L} = \langle l', T(l) \rangle_{L',L} = ((J_L^\mathrm{RF})^{-1}(l'), T(l))_L$ show that $T^\mathsf{H} = (J_L^\mathrm{RF})^{-1} \circ T^* \circ J_L^\mathrm{RF}$. This proves the first assertion. To prove (46.8), we observe that for all $\mu \in \mathbb{C}$, we have $\mu I_{L'} - T^* = \mu I_{L'} - J_L^\mathrm{RF} \circ T^\mathsf{H} \circ (J_L^\mathrm{RF})^{-1} = J_L^\mathrm{RF} \circ (\mu I_L - T^\mathsf{H}) \circ (J_L^\mathrm{RF})^{-1}$. The assertion (46.8) on the spectrum follows readily. Finally, if L and L' are identified, J_L^RF becomes the identity operator so that $T^\mathsf{H} = T^*$. $\qquad\square$

Definition 46.16 (Symmetric operator). *Let $T \in \mathcal{L}(L)$. We say that T is (Hermitian) symmetric if $T = T^\mathsf{H}$.*

Theorem 46.17 (Spectrum, spectral radius, ascent). *Let $T \in \mathcal{L}(L)$ be a symmetric operator. The following holds true:* (i) $\sigma(T) \subset \mathbb{R}$, $\sigma_{\mathrm{r}}(T) = \emptyset$, *and*

$$\{a, b\} \subset \sigma(T) \subset [a, b], \tag{46.9}$$

with $a := \inf_{v \in L, \|v\|_L = 1} (T(v), v)_L$ *and* $b := \sup_{v \in L, \|v\|_L = 1} (T(v), v)_L$. (ii) $\|T\|_{\mathcal{L}(L)} = r(T) = \max(|a|, |b|)$. (iii) *The ascent of any* $\mu \in \sigma_{\mathrm{p}}(T)$ *is equal to 1, i.e., every generalized eigenvector is an eigenvector, and if T is compact, the algebraic multiplicity and the geometric multiplicity of μ are equal.*

Proof. See Lax [278, p. 356], Kreyszig [271, §9.2], and Exercise 46.6 for a proof of (i). See Exercise 46.6(iii) for a proof of (iii). $\qquad\square$

Corollary 46.18 (Characterization of $\sigma(T)$). *Let $T \in \mathcal{L}(L)$ be a symmetric operator. Then $\mu \in \sigma(T)$ iff there is a sequence $(v_n)_{n \in \mathbb{N}}$ in L such that $\|v_n\|_L = 1$ for all $n \in \mathbb{N}$ and $\|T(v_n) - \mu v_n\|_L \to 0$ as $n \to \infty$.*

Proof. Identifying L and L', we apply Corollary C.50 which says that $(\mu I_L - T)$ is not bijective iff there exists a sequence $(v_n)_{n \in \mathbb{N}}$ in L such that $\|v_n\|_L = 1$ and $\|\mu v_n - T(v_n)\|_L \leq \frac{1}{n+1}$. $\qquad\square$

For the reader's convenience, we now recall the notion of Hilbert basis in a separable Hilbert space (separability is defined in Definition C.8).

Definition 46.19 (Hilbert basis). *Let L be a separable Hilbert space. A sequence $(e_n)_{n \in \mathbb{N}}$ in L is said to be a Hilbert basis of L if it satisfies the following two properties:*

(i) $(e_m, e_n)_L = \delta_{mn}$ *for all* $m, n \in \mathbb{N}$.

(ii) *The linear space composed of all the finite linear combinations of the vectors in $(e_n)_{n \in \mathbb{N}}$ is dense in L.*

Not every Hilbert space is separable, but all the Hilbert spaces encountered in this book are separable (or by default are always assumed to be separable).

Lemma 46.20 (Pareseval). *Let L be a separable Hilbert space and let $(e_n)_{n \in \mathbb{N}}$ be a Hilbert basis of L. For all $u \in L$, set $u_n := \sum_{k \in \{0:n\}} (u, e_k)_L e_k$. The following holds true:*

$$\lim_{n \to \infty} \|u - u_n\|_L = 0 \quad and \quad \|u\|_L^2 = \sum_{k \in \mathbb{N}} |(u, e_k)_L|^2. \tag{46.10}$$

Conversely, let $(\alpha_n)_{n \in \mathbb{N}}$ be a sequence in $\ell^2(\mathbb{C})$ and set $u_{\alpha,n} := \sum_{k \in \{0:n\}} \alpha_k e_k$. Then the sequence $(u_{\alpha,n})_{n \in \mathbb{N}}$ converges to some u_α in L such that $(u_\alpha, e_n)_V = \alpha_n$ for all $n \in \mathbb{N}$, and we have $\|u_\alpha\|_L^2 = \lim_{n \to \infty} \sum_{k \in \{0:n\}} \alpha_k^2$.

Proof. See Brezis [89, Cor. 5.10]. $\qquad\square$

Theorem 46.21 (Symmetric compact operator). *Let L be a separable Hilbert space and let $T \in \mathcal{L}(L)$ be a symmetric compact operator. Then there exists a Hilbert basis of L composed of eigenvectors of T.*

Proof. See [89, Thm. 6.11]. $\qquad\square$

The above results mean that the eigenvectors of a symmetric compact operator T form a sequence $(v_n)_{n \in \mathbb{N}}$ s.t. $(v_m, v_n)_L = \delta_{mn}$ for all $m, n \in \mathbb{N}$. Moreover, for all $u \in L$, letting $\alpha_n := (u, v_n)_L$ and $u_n := \sum_{k \in \{0:n\}} \alpha_k v_k$, the sequence $(u_n)_{n \in \mathbb{N}}$ converges to u in L and we have $\|u\|_L^2 = \sum_{k \in \mathbb{N}} |\alpha_k|^2$.

46.2   Introductory examples

We review in this section some typical examples that give rise to an eigenvalue problem, and we illustrate some of the concepts introduced in §46.1.

46.2.1   Example 1: Vibrating string

Consider a vibrating string of linear density ρ, length ℓ, attached at $x = 0$ and $x = \ell$, and maintained under tension with the force τ. Let us set $D := (0, \ell)$, $J := (0, T_{\max})$, $T_{\max} > 0$, and denote by $u : D \times J \to \mathbb{R}$ the displacement of the string in the direction orthogonal to the x-axis. Denoting by $u_0(x)$ and $u_1(x)$ the initial position and the initial velocity (i.e., the time derivative of the displacement), the displacement of the string can be modeled by the linear wave equation

$$\partial_{tt}u(x,t) - c^2 \partial_{xx}u(x,t) = 0 \qquad \text{in } D \times J, \tag{46.11a}$$

$$u(0,t) = 0, \ u(\ell, t) = 0 \qquad \text{in } J, \tag{46.11b}$$

$$u(x,0) = u_0(x), \ \partial_t u(x,0) = u_1(x) \quad \text{in } D, \tag{46.11c}$$

where the wave speed is $c := (\frac{\tau}{\rho})^{\frac{1}{2}}$. The method of the separation of variables gives the following representation of the solution:

$$u(x,t) = \sum_{n \geq 1} \left(\alpha_n \cos(\omega_n t) + \beta_n \sin(\omega_n t) \right) \psi_n(x), \tag{46.12}$$

with $\omega_n := c\lambda_n^{\frac{1}{2}}$, $\lambda_n := \frac{n^2 \pi^2}{\ell^2}$, $\psi_n(x) := \sin(n\pi \frac{x}{\ell})$,

$$\alpha_n := \frac{2}{\ell} \int_0^\ell u_0(x)\psi_n(x)\,\mathrm{d}x,, \qquad \beta_n := \frac{2}{cn\pi} \int_0^\ell u_1(x)\psi_n(x)\,\mathrm{d}x.$$

A remarkable fact is that for all $n \geq 1$, (λ_n, ψ_n) is an eigenpair for the Laplace eigenvalue problem

$$-\partial_{xx}\psi_n(x) = \lambda_n \psi_n(x), \quad \psi_n(0) = 0, \ \psi_n(\ell) = 0. \tag{46.13}$$

The eigenfunctions ψ_n are called normal modes. In musical language, they are called harmonics of the string. Note that $\alpha_n = \int_0^\ell u_0(x)\psi_n(x)\,\mathrm{d}x / \int_0^\ell \psi_n^2(x)\,\mathrm{d}x$, $\omega_n \beta_n = \int_0^\ell u_1(x)\psi_n(x)\,\mathrm{d}x / \int_0^\ell \psi_n^2(x)\,\mathrm{d}x$.

We say that (46.13) is the *spectral problem* associated with the vibrating string. This problem can be reformulated in the following weak form:

$$\begin{cases} \text{Find } \psi \in H_0^1(D)\backslash\{0\} \text{ and } \lambda \in \mathbb{R} \text{ such that} \\ \int_D \partial_x \psi \partial_x w \,\mathrm{d}x = \lambda \int_D \psi w \,\mathrm{d}x, \quad \forall w \in H_0^1(D). \end{cases} \tag{46.14}$$

Let $L := L^2(D)$ and let $T : L \to L$ be defined so that for all $f \in L$, $T(f) \in H_0^1(D)$ solves $\int_D \partial_x(T(f))\partial_x w \,\mathrm{d}x := \int_D fw \,\mathrm{d}x$ for all $w \in H_0^1(D)$. The operator T is compact since the injection $H_0^1(D) \hookrightarrow L^2(D)$ is compact owing to the Rellich–Kondrachov theorem. This compactness property will be important for approximation purposes. Upon observing that $\int_D fT(g) \,\mathrm{d}x = \int_D \partial_x(T(f))\partial_x(T(g)) \,\mathrm{d}x = \int_D T(f)g \,\mathrm{d}x$, we infer that T is symmetric according to Definition 46.16. Owing to Theorem 46.17, all the eigenvalues of T are real and $\sigma_{\mathrm{r}}(T) = 0$. According to Theorem 46.14, the eigenvalues of T are well separated and form a sequence that goes to 0. Note that T is injective, that is, 0 is not an eigenvalue. According to Theorem 46.14, this means that $\sigma_{\mathrm{c}}(T) = \{0\}$. Let (μ, ψ) be an eigenpair of T. Then $\mu \int_D \partial_x \psi \partial_x w \,\mathrm{d}x = \int_D \partial_x(T(\psi))\partial_x w \,\mathrm{d}x = \int_D \psi w \,\mathrm{d}x$. Hence, (μ^{-1}, ψ) solves (46.14). Conversely, one readily sees that if (λ, ψ) solves (46.14), then (λ^{-1}, ψ) is an eigenpair of T. Thus, we have established that (λ, ψ) solves (46.14) iff (λ^{-1}, ψ) is an eigenpair of T. Finally, Theorem 46.21 asserts that there exists a Hilbert basis of L consisting of eigenvectors of T, and the basis in question is $((\frac{2}{\ell})^{\frac{1}{2}}\psi_n)_{n \geq 1}$.

46.2.2   Example 2: Vibrating drum

Consider a two-dimensional elastic homogeneous membrane occupying at rest the domain $D \subset \mathbb{R}^2$ and attached to a rigid frame on ∂D, as shown in Figure 46.1. We assume that D is embedded
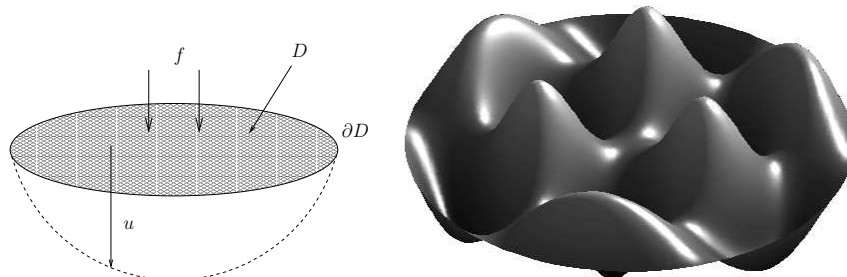

Figure 46.1: Vibrating membrane attached to a rigid frame. Left: reference configuration D, externally applied load f, and equilibrium displacement u. Right: one normal mode.

in $\mathbb{R}^3$ and denote by $\boldsymbol{e}_z$ the third direction. Assume that the membrane is of uniform thickness, has area density ρ, and that the tension tensor in the membrane, $\mathfrak{t}$, is uniform and isotropic, i.e., it is of the form $\mathfrak{t} = \tau \mathbb{I}_2$ for some positive real number τ (force per unit surface). Consider a time-dependent load $f(\boldsymbol{x}, t) := \rho g(\boldsymbol{x}) \cos(\omega t)$ with angular frequency ω for all $(\boldsymbol{x}, t) \in D \times J$ with $J := (0, T_{\max})$, $T_{\max} > 0$. Under the small strain assumption, the time-dependent displacement of the membrane in the e_z direction, $u : D \times J \to \mathbb{R}$, is modeled by the two-dimensional wave equation

$$\partial_{tt} u - c^2 \Delta u = g(\boldsymbol{x}) \cos(\omega t) \qquad \text{in } D \times J, \qquad (46.15\text{a})$$

$$u(\cdot, t)_{|\partial D} = 0 \qquad \text{in } J, \qquad (46.15\text{b})$$

$$u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}), \ \partial_t u(\boldsymbol{x}, 0) = u_1(\boldsymbol{x}) \quad \text{in } D, \qquad (46.15\text{c})$$

where the wave speed is $c := (\frac{\tau}{\rho})^{\frac{1}{2}}$. As in §46.2.1, the solution to this problem can be expressed in terms of the normal modes (eigenmodes) of the membrane, $(\lambda_n, \psi_n)_{n \geq 1}$, which satisfy

$$-\Delta \psi_n = \lambda_n \psi_n \text{ in } D, \quad \psi_{n|\partial D} = 0. \qquad (46.16)$$

Setting $\omega_n := c \lambda_n^{\frac{1}{2}}$, a straightforward calculation shows that if $\omega \notin \{\omega_n\}_{n \geq 1}$,

$$u(\boldsymbol{x}, t) = \sum_{n \geq 1} \left\{ \alpha_n \cos(\omega_n t) + \beta_n \sin(\omega_n t) + \frac{\gamma_n}{2} \frac{\sin\left(\frac{\omega - \omega_n}{2} t\right)}{\frac{\omega - \omega_n}{2}} \frac{\sin\left(\frac{\omega + \omega_n}{2} t\right)}{\frac{\omega + \omega_n}{2}} \right\} \psi_n(\boldsymbol{x}),$$

where

$$\alpha_n := \frac{(u_0, \psi_n)_{L^2(D)}}{\|\psi_n\|_{L^2(D)}^2}, \qquad \omega_n \beta_n := \frac{(u_1, \psi_n)_{L^2(D)}}{\|\psi_n\|_{L^2(D)}^2}, \qquad \gamma_n := \frac{(g, \psi_n)_{L^2(D)}}{\|\psi_n\|_{L^2(D)}^2}.$$

As the forcing angular frequency ω gets close to one of the ω_n's, a resonance phenomenon occurs. When $\omega = \omega_n$, $|u(\boldsymbol{x}, t)|$ grows linearly in time like $t | \sin(\omega_n t)|$.

The *spectral problem* associated with the vibrating drum can be rewritten in weak form as follows:

$$\begin{cases} \text{Find } \psi \in H_0^1(D) \backslash \{0\} \text{ and } \lambda \in \mathbb{R} \text{ such that} \\ \int_D \nabla \psi \cdot \nabla w \, \mathrm{d}x = \lambda \int_D \psi w \, \mathrm{d}x, \quad \forall w \in H_0^1(D). \end{cases} \qquad (46.17)$$

If the tension tensor $\mathfrak{t}$ in the membrane is not uniform and/or not isotropic (think of a membrane made of composite materials), and if the area density ρ is not uniform, the above spectral problem takes the following form:

$$\begin{cases} \text{Find } \psi \in H_0^1(D) \backslash \{0\} \text{ and } \lambda \in \mathbb{R} \text{ such that} \\ \int_D (\mathfrak{t}\nabla\psi)\cdot\nabla w \, \mathrm{d}x = \lambda \int_D \rho\psi w \, \mathrm{d}x, \quad \forall w \in H_0^1(D). \end{cases} \tag{46.18}$$

By proceeding as in §46.2.1 and under reasonable assumptions on $\mathfrak{t}$ and ρ, one can show that the solution operator associated with (46.18) is symmetric and compact from $L^2(D)$ to $L^2(D)$. Hence, the eigenvalues associated with the eigenvalue problem (46.18) are countable, isolated, and grow to infinity.

46.2.3   Example 3: Stability analysis of PDEs

It is common that one has to study the stability of physical systems modeled by PDEs. For instance, the following nonlinear reaction-diffusion equation (sometimes referred to as the Kolmogorov–Petrovsky–Piskounov equation):

$$\partial_t u - \Delta g(u) - f(u) = 0 \quad \text{in } D \times J, \tag{46.19}$$

models the spreading of biological populations when $f(u) := u(1-u)$, the Rayleigh–Benard convection when $f(u) := u(1-u^2)$, and combustion processes when $f(u) := u(1-u)(u-\alpha)$ with $\alpha \in (0,1)$. We assume here that $D := (0,1)^d$, periodic boundary conditions are enforced, f and g are smooth, and g' is bounded from below by some positive constant. Assuming that this problem admits a particular time-independent solution (a standing wave), u_{sw}, the natural question that follows is to determine whether this solution is stable under infinitesimal perturbations. Writing $u(\boldsymbol{x},t) := u_{\mathrm{sw}}(\boldsymbol{x}) + \psi(\boldsymbol{x})e^{-\lambda t}$, $\lambda \in \mathbb{C}$, where ψ is assumed to be small compared to u_{sw}, one obtains the following linearized form of the PDE:

$$-\lambda\psi - \Delta(g'(u_{\mathrm{sw}})\psi) - f'(u_{\mathrm{sw}})\psi = 0 \quad \text{in } D \times J. \tag{46.20}$$

Since $\nabla(g'(u_{\mathrm{sw}})\psi) = g'(u_{\mathrm{sw}})\nabla\psi + \psi g''(u_{\mathrm{sw}})\nabla u_{\mathrm{sw}}$, this problem leads to the following eigenvalue problem:

$$\begin{cases} \text{Find } \psi \in H^1_{\mathrm{per}}(D)\backslash\{0\} \text{ and } \lambda \in \mathbb{C} \text{ such that } \forall w \in H^1_{\mathrm{per}}(D), \\ \int_D \big((g'(u_{\mathrm{sw}})\nabla\psi + \psi\, g''(u_{\mathrm{sw}})\nabla u_{\mathrm{sw}})\cdot\nabla\overline{w} - f'(u_{\mathrm{sw}})\psi\overline{w}\big)\mathrm{d}x = \lambda \int_D \psi\overline{w}\,\mathrm{d}x, \end{cases} \tag{46.21}$$

where $H^1_{\mathrm{per}}(D)$ is composed of the functions in $H^1(D)$ that are periodic over D. The particular solution u_{sw} is said to be linearly stable if all the eigenvalues have a positive real part. Here again, it is the solution operator $T : L^2(D) \to L^2(D)$ that is of interest, where for all $s \in L^2(D)$, $T(s) \in H^1_{\mathrm{per}}(D) \subset L^2(D)$ solves $\int_D \big((g'(u_{\mathrm{sw}})\nabla T(s) + T(s)\, g''(u_{\mathrm{sw}})\nabla u_{\mathrm{sw}})\cdot\nabla\overline{w} - T(s)f'(u_{\mathrm{sw}})\overline{w}\big)\,\mathrm{d}x = \int_D s\overline{w}\,\mathrm{d}x$ for all $w \in H^1_{\mathrm{per}}(D)$. Under reasonable assumptions on f, g, u_{sw}, the operator T can be shown to be compact.

46.2.4   Example 4: Schrödinger equation and hydrogen atom

The vibrating string and the drum are typical examples where compactness directly results from the boundedness of the domain D. We now give an example where compactness results from an additional potential in the PDE.

An important example of eigenvalue problem in physics is the Schrödinger equation. For instance, the normalized Schrödinger equation takes the following form for the one-dimensional quantum harmonic oscillator over $\mathbb{R}$:

$$-\frac{1}{2}\psi'' + \frac{1}{2}x^2\psi = E\psi \quad \text{in } \mathbb{R}. \tag{46.22}$$

The function ψ is the wave function of the oscillator, and the quantity ψ^2 is its probability distribution function. The eigenvalue E is called energy. This problem has a countable (quantified) set of eigenpairs

$$\psi_n(x) := \frac{1}{(2^n n!)^{\frac{1}{2}} \pi^{\frac{1}{4}}} e^{-\frac{x^2}{2}} H_n(x), \qquad E_n := n + \frac{1}{2}, \tag{46.23}$$

where $H_n(x) := (-1)^n e^{x^2} \frac{\mathrm{d}^n}{\mathrm{d}x^n} e^{-x^2}$ is the Hermite polynomial of order n. A natural functional space for this problem is

$$B^1(\mathbb{R}) := \{v \in H^1(\mathbb{R}) \mid xv \in L^2(\mathbb{R})\}. \tag{46.24}$$

In addition to being in $H^1(\mathbb{R})$, functions in $B^1(\mathbb{R})$ satisfy $\int_{\mathbb{R}} x^2 v^2(x)\,\mathrm{d}x < \infty$. It is shown in Exercise 46.8 that the embedding $B^1(\mathbb{R}) \hookrightarrow L^2(\mathbb{R})$ is compact, whereas it is shown in Exercise 46.7 that the embedding $H^1(\mathbb{R}) \hookrightarrow L^2(\mathbb{R})$ is not compact. Hence, the sesquilinear form $a(v,w) = \int_{\mathbb{R}} (v'\overline{w}' + x^2 v\overline{w})\,\mathrm{d}x$ is bounded and coercive on $B^1(\mathbb{R})$, and the operator $T : B^1(\mathbb{R}) \to B^1(\mathbb{R})$ s.t. $a(T(u),w) = \int_{\mathbb{R}} u\overline{w}\,\mathrm{d}x$ for all $w \in B^1(\mathbb{R})$, is symmetric and compact.

The hydrogen atom is a model for which the Schrödinger equation has the following simple form:

$$-\frac{\hbar^2}{2m_{\mathrm{e}}}\Delta\psi - \frac{q^2}{4\pi\epsilon_0 r}\psi = E\psi \quad \text{in } \mathbb{R}^3. \tag{46.25}$$

Here, $\hbar$ is the Planck constant, m_{e} the mass of the electron, ϵ_0 the permittivity of free space, q the electron charge, and $r := \|\boldsymbol{x}\|_{\ell^2}$ the Euclidean distance of the electron to the nucleus. This problem is far more difficult than the one-dimensional quantum harmonic oscillator because the Coulomb potential $-\frac{q^2}{4\pi\epsilon_0 r}$ is negative and vanishes at infinity. The sign problem can be handled as for the Helmholtz problem (see Chapter 35) by invoking Gårding's inequality after making use of Hardy's inequality $|u|^2_{H^1(\mathbb{R}^d)} \geq \frac{(d-2)^2}{4} \int_{\mathbb{R}^d} \frac{u^2}{r^2}\,\mathrm{d}x$ for all $u \in H^1(\mathbb{R}^d)$. The spectrum of the solution operator is composed of the point spectrum and the continuous spectrum. The residual spectrum is empty because the solution operator is symmetric. There is a countable (quantified) set of eigenpairs. Using spherical coordinates, they are given for all $n \geq 1$ by

$$\psi_{n,l,m}(r,\theta,\phi) := C_{n,l} a_0^{-\frac{3}{2}} e^{-\frac{\rho}{2}} \rho^l L_{n-l-1}^{2l+1}(\rho) Y_l^m(\theta,\phi),$$

$$E_n := -\frac{\hbar^2}{2m_{\mathrm{e}}a_0^2} \frac{1}{n^2},$$

where $l \in \{0{:}n{-}1\}$, $m \in \{-l{:}l\}$, $C_{n,l} := \left(\frac{2}{n}\right)^{\frac{3}{2}} \left(\frac{(n-l-1)!}{2n((n+l)!)^3}\right)^{\frac{1}{2}}$, $a_0 := \frac{4\pi\epsilon_0\hbar^2}{m_{\mathrm{e}}q^2}$ is the Bohr radius, $\rho := \frac{2r}{na_0}$, $L_\beta^\gamma(r) := \frac{r^{-\gamma}e^r}{\beta!}\frac{\mathrm{d}^\beta}{\mathrm{d}r^\beta}(e^{-r}r^{\gamma+\beta})$ is the generalized Laguerre polynomial of degree β, and Y_l^m is the spherical harmonic function of degree l and order m.

Exercises

Exercise 46.1 (Spectrum). Let L be a complex Banach space. Let $T \in \mathcal{L}(L)$. (i) Show that $(\lambda T)^* = \overline{\lambda}T^*$ for all $\lambda \in \mathbb{C}$. (ii) Show that $\sigma_{\mathrm{r}}(T) \subset \mathrm{conj}(\sigma_{\mathrm{p}}(T^*)) \subset \sigma_{\mathrm{r}}(T) \cup \sigma_{\mathrm{p}}(T)$. (*Hint*: use

Corollary C.15.) (iii) Show that the spectral radius of T verifies $r(T) \leq \limsup_{n \to \infty} \|T^n\|_{\mathcal{L}(L)}^{\frac{1}{n}}$. (*Hint*: consider $\sum_{n \in \mathbb{N}} (\mu^{-1}T)^n$ and use the root test: the complex-valued series $\sum_{n \in \mathbb{N}} a_n$ converges absolutely if $\limsup_{n \to \infty} |a_n|^{\frac{1}{n}} < 1$.)

Exercise 46.2 (Ascent, algebraic and geometric multiplicities). (i) Let $T \in \mathcal{L}(L)$. Let μ be an eigenvalue of T and let $K_i := \ker(\mu I_L - T)^i$ for all $i \in \mathbb{N} \backslash \{0\}$. Prove that $K_1 \subset K_2 \ldots$, and assuming that there is $j \geq 1$ s.t. $K_j = K_{j+1}$, show that $K_j = K_{j'}$ for all $j' > j$. (ii) Assume that μ has a finite ascent α, and finite algebraic multiplicity m and geometric multiplicity g. Show that $\alpha + g - 1 \leq m \leq \alpha g$. (*Hint*: letting $g_i := \dim(K_i)$ for all $i \in \{1{:}\alpha\}$, prove that $g_1 + i - 1 \leq g_i$ and $g_i \leq g_{i-1} + g_1$.) (iii) Compute the ascent, algebraic multiplicity, and geometric multiplicity of the eigenvalues of following matrices and verify the two inequalities from Step (i):

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Exercise 46.3 (Eigenspaces). The following three questions are independent. (i) Suppose $V = V_1 \oplus V_2$ and consider $T \in \mathcal{L}(V)$ defined by $T(v_1 + v_2) := v_1$ for all $v_1 \in V_1$ and all $v_2 \in V_2$. Find all the eigenvalues and eigenspaces of T. (ii) Let $T \in \mathcal{L}(V)$. Assume that S is invertible. Prove that $S^{-1}TS$ and T have the same eigenvalues. What is the relationship between the eigenvectors of T and those of $S^{-1}TS$? (iii) Let V be a finite-dimensional vector space. Let $\{v_n\}_{n \in \{1{:}m\}} \subset V$, $m \geq 1$. Show that the vectors $\{v_n\}_{n \in \{1{:}m\}}$ are linearly independent iff there exists $T \in \mathcal{L}(V)$ such that $\{v_n\}_{n \in \{1{:}m\}}$ are eigenvectors of T corresponding to distinct eigenvalues.

Exercise 46.4 (Volterra operator). Let $L := L^2((0,1); \mathbb{C})$ and let $T : L \to L$ be s.t. $T(f)(x) := \int_0^x f(t) \, dt$ for a.e. $x \in (0,1)$. Notice that T is a Hilbert–Schmidt operator, but this exercise is meant to be done without using this fact. (i) Show that $T^{\mathsf{H}}(g) = \int_x^1 g(t) \, dt$ for all $g \in L^2((0,1); \mathbb{C})$. (ii) Show that T is injective. (*Hint*: use Theorem 1.32.) (iii) Show that $0 \in \sigma_{\mathrm{c}}(T)$. (iv) Show that $\sigma_{\mathrm{p}}(T) = \emptyset$. (v) Prove that $\mu I_L - T$ is bijective if $\mu \neq 0$. (vi) Determine $\rho(T), \sigma_{\mathrm{p}}(T), \sigma_{\mathrm{c}}(T), \sigma_{\mathrm{r}}(T)$. Do the same for T^{H}.

Exercise 46.5 (Riesz–Fréchet). Let H be a finite-dimensional complex Hilbert space with orthonormal basis $\{e_i\}_{i \in \{1{:}n\}}$ and inner product $(\cdot, \cdot)_H$. (i) Let g be an antilinear form on H, i.e., $g \in H'$. Show that $(J_H^{\mathrm{RF}})^{-1}(g) = \sum_{i \in \{1{:}n\}} g(e_i) e_i$ with $g(e_i) := \langle g, e_i \rangle_{H',H}$, $\forall i \in \{1{:}n\}$. Is $(J_H^{\mathrm{RF}})^{-1} : H' \to H$ linear or antilinear? (ii) Let g be a linear form on H. Show that $x_g := \sum_{i \in \{1{:}n\}} \overline{g(e_i)} e_i$ is s.t. $\langle g, y \rangle_{H',H} = \overline{(x_g, y)_H}$. Is the map $H' \ni g \mapsto x_g \in H$ linear or antilinear?

Exercise 46.6 (Symmetric operator). Let L be a complex Hilbert space and $T \in \mathcal{L}(L)$ be a symmetric operator. (i) Show that $\sigma(T) \subset \mathbb{R}$. (*Hint*: compute $\Im((T(v) - \mu v, v)_L$ and show that $|\Im(\mu)| \|v\|_L^2 \leq |(T(v) - \mu v, v)_L|$ for all $v \in L$.) (ii) Prove that $\sigma_{\mathrm{r}}(T) = \emptyset$. (*Hint*: apply Corollary C.15.) (iii) Show that the ascent of each $\mu \in \sigma_{\mathrm{p}}(T)$ is equal to 1. (*Hint*: compute $\|(\mu I_L - T)(x)\|_L^2$ with $x \in \ker(\mu I_L - T)^2$.)

Exercise 46.7 ($H^1(\mathbb{R}) \hookrightarrow L^2(\mathbb{R})$ is not compact). (i) Let $\chi(x) := 1 + x$ if $-1 \leq x \leq 0$, $\chi(x) := 1 - x$ if $0 \leq x \leq 1$ and $\chi(x) := 0$ if $|x| \geq 1$. Show that $\chi \in H^1(\mathbb{R})$. (ii) Let $v_n(x) := \chi(x - n)$ for all $n \in \mathbb{N}$. Show that $(v_n)_{n \in \mathbb{N}}$ converges weakly to 0 in $L^2(\mathbb{R})$ (see Definition C.28). (iii) Show that the embedding $H^1(\mathbb{R}) \hookrightarrow L^2(\mathbb{R})$ is not compact. (*Hint*: argue by contradiction using Theorem C.23.)

Exercise 46.8 ($B^1(\mathbb{R}) \hookrightarrow L^2(\mathbb{R})$ is compact). (i) Show that the embedding $B^1(\mathbb{R}) \hookrightarrow L^2(\mathbb{R})$ is compact, where $B^1(\mathbb{R}) := \{v \in H^1(\mathbb{R}) \mid xv \in L^2(\mathbb{R})\}$. (*Hint*: let $(u_n)_{n \in \mathbb{N}}$ be a bounded sequence in $B^1(\mathbb{R})$, build nested subsets $J_k \subset \mathbb{N}$, $\forall k \in \mathbb{N}\backslash\{0\}$, s.t. the sequence $(u_{n|(-k,k)})_{n \in J_k}$ converges in $L^2(-k,k)$.) (ii) Give a sufficient condition on $\alpha \in \mathbb{R}$ so that $B^1_\alpha(\mathbb{R}) \hookrightarrow L^2(\mathbb{R})$ is compact, where $B^1_\alpha(\mathbb{R}) := \{v \in H^1(\mathbb{R}) \mid |x|^\alpha v \in L^2(\mathbb{R})\}$.

Exercise 46.9 (Hausdorff–Toeplitz theorem). The goal of this exercise is to prove that the numerical range of a bounded linear operator in a Hilbert space is convex; see also Gustafson [231]. Let L be a complex Hilbert space and let $S_L(1) := \{x \in L \mid \|x\|_L = 1\}$ be the unit sphere in L. Let $T \in \mathcal{L}(L)$ and let $W(T) := \{\alpha \in \mathbb{C} \mid \exists x \in S_L(1),\ \alpha = (T(x), x)_L\}$ be the numerical range of T. Let $\gamma, \mu \in W(T)$, $\gamma \neq \mu$, and $x_1, x_2 \in S_L(1)$ be s.t. $(T(x_1), x_1)_L = \gamma$, $(T(x_2), x_2)_L = \mu$. Let $T' := \frac{1}{\mu - \gamma}(T - \gamma I_L)$. (i) Compute $(T'(x_1), x_1)_L$ and $(T'(x_2), x_2)_L$. (ii) Prove that there exists $\theta \in [0, 2\pi)$ s.t. $\Im(e^{i\theta}(T'(x_1), x_2)_L + e^{-i\theta}(T'(x_2), x_1)_L) = 0$. (iii) Let $x'_1 := e^{i\theta}x_1$. Compute $(T'(x'_1), x'_1)_L$. (iv) Let $\lambda \in [0, 1]$. Show that the following problem has at least one solution: Find $\alpha, \beta \in \mathbb{R}$ s.t. $\|\alpha x'_1 + \beta x_2\|_L = 1$ and $(T'(\alpha x'_1 + \beta x_2), \alpha x'_1 + \beta x_2)_L = \lambda$. (*Hint*: view the two equations as those of an ellipse and an hyperbola, respectively, and determine how these curves cross the axes.) (v) Prove that $W(T)$ is convex. (*Hint*: compute $(T(\alpha x'_1 + \beta x_2), \alpha x'_1 + \beta x_2)_L$.)

Chapter 47

Symmetric operators, conforming approximation

The objective of this chapter is to study the approximation of eigenvalue problems associated with symmetric coercive differential operators using H^1-conforming finite elements. The goal is to derive error estimates on the eigenvalues and the eigenfunctions. The analysis is adapted from Raviart and Thomas [331] and uses relatively simple geometric arguments. The approximation of nonsymmetric eigenvalue problems using nonconforming techniques is studied in Chapter 48 using slightly more involved arguments.

47.1 Symmetric and coercive eigenvalue problems

In this section, we reformulate the eigenvalue problems introduced in §46.2 in a unified setting. This abstract setting will be used in §47.2 to analyze the approximation of these problems using H^1-conforming finite elements. We restrict ourselves to the real-valued setting since we are going to focus on symmetric operators.

47.1.1 Setting

Let D be a Lipschitz domain in $\mathbb{R}^d$. Let $L^2(D)$ be the real Hilbert space equipped with the inner product $(v, w)_{L^2(D)} := \int_D vw \, dx$. Let V be a closed subspace of $H^1(D)$ which, depending on the boundary conditions that are enforced, satisfies $H_0^1(D) \subseteq V \subseteq H^1(D)$. We assume that V is equipped with a norm that is equivalent to that of $H^1(D)$. We also assume that the V-norm is rescaled so that the operator norm of the embedding $V \hookrightarrow L^2(D)$ is at most one, e.g., one could set $\|v\|_V := C_{\mathrm{PS}}^{-1} \ell_D \|\nabla v\|_{\boldsymbol{L}^2(D)}$ if $V := H_0^1(D)$, where C_{PS} is the constant from the Poincaré–Steklov inequality (31.12) in $H_0^1(D)$ and ℓ_D is a characteristic length associated with D, e.g., $\ell_D := \mathrm{diam}(D)$.

Let $a : V \times V \to \mathbb{R}$ be a symmetric bilinear form, i.e., $a(v, w) = a(w, v)$, satisfying the following coercivity and boundedness properties:

$$\alpha \|v\|_V^2 \leq a(v, v), \qquad |a(v, w)| \leq \|a\| \, \|v\|_V \|w\|_V, \tag{47.1}$$

for all $v, w \in V$, with $0 < \alpha \leq \|a\| < \infty$. For instance, we have $a(v, w) := \int_D (\mathbb{t} \nabla v) \cdot \nabla w \, dx$ and

$V := H_0^1(D)$ in (46.18), so that we can take $\alpha := \tau_\flat \ell_D^{-2}$ and $\|a\| := \tau_\sharp \ell_D^{-2}$, where $\tau_\flat$ and $\tau_\sharp$ are the smallest and the largest eigenvalues of $\mathfrak{t}$ in D.

Our goal is to investigate the H^1-conforming approximation of the following eigenvalue problem:

$$\begin{cases} \text{Find } \psi \in V \backslash \{0\} \text{ and } \lambda \in \mathbb{R} \text{ such that} \\ a(\psi, w) = \lambda(\psi, w)_{L^2(D)}, \quad \forall w \in V. \end{cases} \tag{47.2}$$

Let $T : L^2(D) \to L^2(D)$ be the solution operator such that for all $u \in L^2(D)$,

$$a(T(u), w) := (u, w)_{L^2(D)}, \qquad \forall w \in V. \tag{47.3}$$

By proceeding as in §46.2.1, we conclude that T is symmetric and compact. We are then in the setting of Theorem 46.14 and Theorem 46.21.

Theorem 47.1 (Hilbert basis). *Under the above assumptions on the bilinear form a, the following properties hold true:*

(i) *$(\lambda, \psi) \in (0, \infty) \times V$ is an eigenpair for the eigenvalue problem (47.2) iff $(\lambda^{-1}, \psi) \in (0, \infty) \times V$ is an eigenpair for T.*

(ii) *$\sigma_p(T) \subset (0, \frac{1}{\alpha}]$.*

(iii) *The eigenvalue problem (47.2) has a countable sequence of isolated real positive eigenvalues that grows to infinity.*

(iv) *It is possible to construct a Hilbert basis $(\psi_n)_{n \geq 1}$ of $L^2(D)$, where $(\lambda_n, \psi_n)_{n \geq 1}$ are the eigenpairs solving (47.2) (see Definition 46.19). (It is customary to enumerate the eigenpairs starting with $n \geq 1$.)*

(v) *$(\lambda_n^{-\frac{1}{2}} \psi_n)_{n \geq 1}$ is a Hilbert basis of V equipped with the inner product $a(\cdot, \cdot)$.*

Proof. (i) Let (μ, ψ) be an eigenpair of T. Then $\|\psi\|_{L^2(D)}^2 = a(T(\psi), \psi) = \mu a(\psi, \psi)$, which implies that $\mu > 0$. This proves that $\sigma_p(T) = \sigma(T) \backslash \{0\}$ and $\sigma_p(T) \subset (0, \infty)$ (see Theorem 46.14(ii) and recall that $\dim(L^2(D)) = \infty$). Let (μ, ψ) be an eigenpair for T. Then $a(T(\psi), w) = \mu a(\psi, w) = (\psi, w)_{L^2(D)}$ for all $w \in V$. Since $\mu \neq 0$, we conclude that $a(\psi, w) = \mu^{-1}(\psi, w)_{L^2}$ for all $w \in V$, that is, (μ^{-1}, ψ) solves (47.2). The converse is also true: if (λ, ψ) is an eigenpair for (47.2), then the coercivity of a implies that $\lambda \neq 0$, and reasoning as above shows that (λ^{-1}, ψ) is an eigenpair of T.

(ii) Let (μ, ψ) be an eigenpair of T. The coercivity of a implies that $\|\psi\|_{L^2(D)}^2 = a(T(\psi), \psi) = \mu a(\psi, \psi) \geq \mu \alpha \|\psi\|_V^2 \geq \mu \alpha \|\psi\|_{L^2(D)}^2$, where the last bound follows from our assuming that the norm of the embedding $V \hookrightarrow L^2(D)$ is at most one. Hence, $\mu \in (0, \frac{1}{\alpha}]$.

(iii) The number of eigenvalues of T cannot be finite since the eigenspaces are finite-dimensional (see Theorem 46.13(ii)) and there exists a Hilbert basis of $L^2(D)$ composed of eigenvectors of T (see Theorem 46.21). We are then in the third case described in Theorem 46.14(iii): the eigenvalues of T form a (countable) sequence that converges to zero. Hence, the eigenvalues of (47.2) grow to infinity.

(iv) This is a consequence of Theorem 46.21 and Item (iii) proved above.

(v) Let ψ_m, ψ_n be two members of the Hilbert basis $(\psi_k)_{k \geq 1}$ of $L^2(D)$. Recalling that (λ_m, ψ_m) and (λ_n, ψ_n) are eigenpairs of (47.2), we infer that

$$a(\lambda_m^{-\frac{1}{2}} \psi_m, \lambda_n^{-\frac{1}{2}} \psi_n) = \lambda_m^{\frac{1}{2}} \lambda_n^{-\frac{1}{2}} (\psi_m, \psi_n)_{L^2(D)} = \delta_{mn}.$$

Let W be the vector space composed of all the finite linear combinations of vectors in $\{\psi_n\}_{n\geq 1}$. We have to prove that W is dense in V. Let $f \in V'$ and assume that f annihilates W. Denoting by $(J_V^{\mathrm{RF}})^{-1}(f)$ the Riesz–Fréchet representative of f in V equipped with the inner product $a(\cdot, \cdot)$, we have

$$0 = \langle f, \lambda_n^{-\frac{1}{2}}\psi_n \rangle_{V',V} = a((J_V^{\mathrm{RF}})^{-1}(f), \lambda_n^{-\frac{1}{2}}\psi_n) = a(\lambda_n^{-\frac{1}{2}}\psi_n, (J_V^{\mathrm{RF}})^{-1}(f))$$
$$= \lambda_n^{\frac{1}{2}}(\psi_n, (J_V^{\mathrm{RF}})^{-1}(f))_{L^2(D)},$$

for all $n \geq 1$, where we used the symmetry of a. The above identity implies that $(J_V^{\mathrm{RF}})^{-1}(f) = 0$ since W is dense in $L^2(D)$. Hence, $f = 0$. Corollary C.15 then implies that W is dense in V as claimed. □

The eigenvalues are henceforth counted with their multiplicity and ordered as follows: $\lambda_1 \leq \lambda_2 \leq \ldots$. Moreover, the associated eigenfunctions $\psi_1, \psi_2, \ldots$ are chosen and normalized as in Theorem 47.1(iv) so that $\|\psi_n\|_{L^2(D)} = 1$. The coercivity property of a implies that the eigenvalues are all positive and larger than or equal to α. Notice that since T is symmetric, the notions of algebraic and geometric multiplicity coincide, and for every eigenvalue $\lambda^{-1} \in \sigma_{\mathrm{p}}(T)$, the multiplicity of λ is equal to $\dim(\lambda^{-1}I_{L^2(D)} - T)$.

47.1.2 Rayleigh quotient

We introduce in this section the notion of Rayleigh quotient which will be instrumental in the analysis of the H^1-conforming approximation technique presented in §47.2.

Definition 47.2 (Rayleigh quotient). *The Rayleigh quotient of a function $v \in V\backslash\{0\}$, relative to the bilinear form a, is defined as*

$$R(v) := \frac{a(v,v)}{\|v\|_{L^2(D)}^2}. \tag{47.4}$$

In this chapter, all the expressions involving $R(v)$ are understood with $v \neq 0$. For any functional $\mathcal{J} : V \to \mathbb{R}$, we write $\min_{v \in V} \mathcal{J}(v)$ instead of $\inf_{v \in V} \mathcal{J}(v)$ to indicate that the infimum is attained, i.e., if there exists a minimizer $v_* \in V$ such that $\mathcal{J}(v_*) = \inf_{v \in V} \mathcal{J}(v)$.

Proposition 47.3 (First eigenvalue). *Let λ_1 be the smallest eigenvalue of the problem* (47.2) *and let ψ_1 be a corresponding eigenfunction. Then we have*

$$\alpha \leq \lambda_1 = R(\psi_1) = \min_{v \in V} R(v). \tag{47.5}$$

Proof. We have $\lambda_1 = R(\psi_1) \geq \inf_{v \in V} R(v) \geq \alpha$, where the first equality results from $a(\psi_1, \psi_1) = \lambda_1 \|\psi_1\|_{L^2(D)}^2$ and the second from Theorem 47.1(ii). It remains to prove that $\inf_{v \in V} R(v) \geq \lambda_1$ (this also proves that the infimum of R over V is attained at ψ_1 since $\lambda_1 = R(\psi_1)$). Let $v \in V\backslash\{0\}$. Since $(\psi_n)_{n\geq 1}$ is a Hilbert basis of $L^2(D)$ (see Theorem 47.1(iv)), the series $(\sum_{k\in\{1:n\}} \mathsf{W}_k\psi_k)_{n\geq 1}$, with $\mathsf{W}_k := (v, \psi_k)_{L^2(D)}$, converges to v in $L^2(D)$ and we have $\|v\|_{L^2(D)}^2 = \sum_{n\geq 1} \mathsf{W}_n^2$. Furthermore, since $(\lambda_n^{-\frac{1}{2}}\psi_n)_{n\geq 1}$ is a Hilbert basis of V equipped with the inner product $a(\cdot, \cdot)$ (see Theorem 47.1(v)), the series $(\sum_{k\in\{1:n\}} \mathsf{V}_k\lambda_k^{-\frac{1}{2}}\psi_k)_{n\geq 1}$, with $\mathsf{V}_k := a(v, \lambda_k^{-\frac{1}{2}}\psi_k)$, converges to v in V, and we have $a(v,v) = \sum_{n\geq 1} \mathsf{V}_n^2$. But we also have $\mathsf{V}_n = a(v, \lambda_n^{-\frac{1}{2}}\psi_n) = \lambda_n^{\frac{1}{2}}(v, \psi_n)_{L^2(D)} = \lambda_n^{\frac{1}{2}}\mathsf{W}_n$. Since $\lambda_1 \leq \lambda_n$ for all $n \geq 1$, we conclude that

$$R(v) = \frac{\sum_{n\geq 1}\mathsf{V}_n^2}{\sum_{n\geq 1}\mathsf{W}_n^2} = \frac{\sum_{n\geq 1}\lambda_n\mathsf{W}_n^2}{\sum_{n\geq 1}\mathsf{W}_n^2} \geq \lambda_1. \qquad\qquad □$$

Proposition 47.4 (Min-max principle). *Let V_m denote the set of the subspaces of V having dimension m. For all $m \geq 1$, we have*

$$\lambda_m = \min_{E_m \in V_m} \max_{v \in E_m} R(v) = \max_{E_{m-1} \in V_{m-1}} \min_{v \in E_{m-1}^\perp} R(v), \tag{47.6}$$

where for all $m > 1$, $E_{m-1}^\perp$ denotes the orthogonal of E_{m-1} in $L^2(D)$ w.r.t. the L^2-inner product and $E_0 := \{0\}$ by convention.

Proof. Let $W_m := \mathrm{span}\{\psi_1, \ldots, \psi_m\}$. Using the notation $\mathsf{W}_k := (v, \psi_k)_{L^2(D)}$, a direct computation shows that

$$\min_{E_m \in V_m} \max_{v \in E_m} R(v) \leq \max_{v \in W_m} R(v) = \max_{v \in W_m} \frac{\sum_{n \in \{1:m\}} \lambda_n \mathsf{W}_n^2}{\sum_{n \in \{1:m\}} \mathsf{W}_n^2} = \lambda_m.$$

Consider now any $E_m \in V_m$. A dimensional argument shows that there exists $w \neq 0$ in $E_m \cap W_{m-1}^\perp$ (apply the rank nullity theorem to the L^2-orthogonal projection from E_m onto W_{m-1}). Since w can be written in the form $w = \sum_{n \geq m} \mathsf{W}_n \psi_n = \sum_{n \geq m} \lambda_n^{\frac{1}{2}} \mathsf{W}_n \lambda_n^{-\frac{1}{2}} \psi_n$, one shows by proceeding as in the proof of Proposition 47.3 that $R(w) \geq \lambda_m$. As a result, $\max_{v \in E_m} R(v) \geq \lambda_m$. Hence, $\min_{E_m \in V_m} \max_{v \in E_m} R(v) \geq \lambda_m$. This concludes the proof of the first equality in (47.6). See Exercise 47.4 for the proof of the second equality. $\qquad\square$

Remark 47.5 (Poincaré–Steklov constant). The best Poincaré–Steklov constant in $H_0^1(D)$ is $C_{\mathrm{PS}} := \inf_{v \in H_0^1(D) \setminus \{0\}} \frac{\ell_D \|\nabla v\|_{\boldsymbol{L}^2(D)}}{\|v\|_{L^2(D)}}$. Letting λ_1 be the smallest eigenvalue of the Laplacian with Dirichlet boundary conditions, Proposition 47.3 shows that $C_{\mathrm{PS}} = \ell_D \lambda_1^{\frac{1}{2}}$, and the Poincaré–Steklov inequality becomes an equality when applied to the first eigenfunction ψ_1. $\qquad\square$

47.2 H^1-conforming approximation

In this section, we investigate the H^1-conforming finite element approximation of the spectral problem (47.2).

47.2.1   Discrete setting and algebraic viewpoint

We assume that D is a Lipschitz polyhedron in $\mathbb{R}^d$, and we consider a shape-regular sequence $(\mathcal{T}_h)_{h \in \mathcal{H}}$ of affine meshes so that each mesh covers D exactly. Depending on the boundary conditions that are imposed in V, we denote by V_h the H^1-conforming finite element space based on $\mathcal{T}_h$ such that $V_h \subset V$ and $P_{k,0}^{\mathrm{g}}(\mathcal{T}_h) \subseteq V_h \subseteq P_k^{\mathrm{g}}(\mathcal{T}_h)$ with $k \geq 1$ (see §19.2.1 or §19.4). The approximate eigenvalue problem we consider is the following:

$$\begin{cases} \text{Find } \psi_h \in V_h \setminus \{0\} \text{ and } \lambda_h \in \mathbb{R} \text{ such that} \\ a(\psi_h, w_h) = \lambda_h (\psi_h, w_h)_{L^2(D)}, \quad \forall w_h \in V_h. \end{cases} \tag{47.7}$$

Let $I := \dim V_h$, let $\{\varphi_i\}_{i \in \{1:I\}}$ be the global shape functions in V_h, and let $\mathsf{U}_h \in \mathbb{R}^I$ be the coordinate vector of ψ_h relative to this basis. The discrete eigenvalue problem (47.7) can be recast as follows:

$$\begin{cases} \text{Find } \mathsf{U}_h \in \mathbb{R}^I \setminus \{0\} \text{ and } \lambda_h \in \mathbb{R} \text{ such that} \\ \mathcal{A}\mathsf{U}_h = \lambda_h \mathcal{M}\mathsf{U}_h, \end{cases} \tag{47.8}$$

where the *stiffness matrix* $\mathcal{A}$ and the *mass matrix* $\mathcal{M}$ have entries

$$\mathcal{A}_{ij} := a(\varphi_j, \varphi_i) \qquad \text{and} \qquad \mathcal{M}_{ij} := (\varphi_j, \varphi_i)_{L^2(D)}. \tag{47.9}$$

Both matrices are symmetric positive definite since they are Gram matrices (see also §28.1). Because $\mathcal{M}$ is not the identity matrix, the problem (47.8) is called *generalized eigenvalue problem*.

Proposition 47.6 (Spectral problems). (i) (47.7) *and* (47.8) *admit* I *(positive) eigenvalues (counted with their multiplicity)* $\{\lambda_{hi}\}_{i \in \{1:I\}}$. (ii) *The eigenfunctions* $\{\psi_{hi}\}_{i \in \{1:I\}} \subset V_h$ *in* (47.7) *can be chosen so that* $a(\psi_{hj}, \psi_{hi}) = \lambda_{hi}\delta_{ij}$ *and* $(\psi_{hj}, \psi_{hi})_{L^2(D)} = \delta_{ij}$. *Equivalently, the eigenvectors* $\{\mathsf{U}_{hi}\}_{i \in \{1:I\}} \subset \mathbb{R}^I$ *in* (47.8) *can be chosen so that* $\mathsf{U}_{hj}^\mathsf{T}\mathcal{A}\mathsf{U}_{hi} = \lambda_{hi}\delta_{ij}$ *and* $\mathsf{U}_{hj}^\mathsf{T}\mathcal{M}\mathsf{U}_{hi} = \delta_{ij}$.

Proof. (i) Since $\mathcal{A}$ is symmetric and $\mathcal{M}$ is symmetric positive definite, these two matrices can be simultaneously diagonalized. Let us recall the process for completeness. Let $\mathcal{Q}\mathcal{Q}^\mathsf{T}$ be the Cholesky factorization of $\mathcal{M}^{-1}$, i.e., $\mathcal{M} = \mathcal{Q}^{-\mathsf{T}}\mathcal{Q}^{-1}$. Since $\mathcal{Q}^\mathsf{T}\mathcal{A}\mathcal{Q}$ is real and symmetric, there exists an orthogonal matrix $\mathcal{P}$ (with $\mathcal{P}\mathcal{P}^\mathsf{T} = \mathbb{I}_I$), and a diagonal matrix Λ with diagonal entries $(\lambda_{hi})_{i \in \{1:I\}}$, such that $\mathcal{Q}^\mathsf{T}\mathcal{A}\mathcal{Q} = \mathcal{P}\Lambda\mathcal{P}^{-1}$. Then $\mathcal{A}\mathcal{Q}\mathcal{P} = \mathcal{Q}^{-\mathsf{T}}\mathcal{P}\Lambda = \mathcal{M}\mathcal{Q}\mathcal{P}\Lambda$. Let us set $\mathcal{U} := \mathcal{Q}\mathcal{P}$ and let $(\mathsf{U}_{hi})_{i \in \{1:I\}}$ be the columns of the matrix $\mathcal{U}$. The identity $\mathcal{A}\mathcal{U} = \mathcal{M}\mathcal{U}\Lambda$ is equivalent to

$$\mathcal{A}\mathsf{U}_{hi} = \lambda_{hi}\mathcal{M}\mathsf{U}_{hi}, \qquad \forall i \in \{1:I\},$$

showing that the λ_{hi}'s are the eigenvalues of the generalized eigenvalue problem (47.8) and the U_{hi}'s are the corresponding eigenvectors.
(ii) One readily sees that $\mathcal{U}^\mathsf{T}\mathcal{A}\mathcal{U} = \mathcal{P}^\mathsf{T}\mathcal{Q}^\mathsf{T}\mathcal{Q}^{-\mathsf{T}}\mathcal{P}\Lambda = \Lambda$ and $\mathcal{U}^\mathsf{T}\mathcal{M}\mathcal{U} = \mathcal{P}^\mathsf{T}\mathcal{Q}^\mathsf{T}\mathcal{Q}^{-\mathsf{T}}\mathcal{Q}^{-1}\mathcal{Q}\mathcal{P} = \mathbb{I}_I$. This proves the identities on the eigenvectors, and those on the eigenfunctions follow from the definitions of $\mathcal{A}$ and $\mathcal{M}$. $\qquad\square$

It is henceforth assumed that the eigenvalues are enumerated in increasing order $\lambda_{h1} \leq \ldots \leq \lambda_{hI}$, where each eigenvalue appears in this list as many times as its multiplicity. Moreover, the eigenfunctions are chosen and normalized as in Proposition 47.6(ii) so that $\|\psi_{hi}\|_{L^2(D)} = 1$.

47.2.2 Eigenvalue error analysis

Let $m \geq 1$ be a fixed natural number. We assume that h is small enough so that $m \leq I$ (recall that $I := \dim(V_h)$ grows roughly like $(\ell_D/h)^d$ as $h \to 0$). Our objective is to estimate $|\lambda_{hm} - \lambda_m|$. Let us introduce the discrete solution map $G_h : V \to V_h$ defined s.t. $a(G_h(v) - v, v_h) = 0$ for all $v \in V$ and all v_h in V_h (see §26.3.4 and §32.1). Let $W_m := \mathrm{span}\{\psi_i\}_{i \in \{1:m\}}$ and let S_m be the unit sphere of W_m in $L^2(D)$. We define

$$\sigma_{hm} := \min_{v \in W_m \setminus \{0\}} \frac{\|G_h(v)\|_{L^2(D)}}{\|v\|_{L^2(D)}} = \min_{v \in S_m} \|G_h(v)\|_{L^2(D)}. \tag{47.10}$$

(Note that $\|G_h(v)\|_{L^2(D)}$ attains its infimum over S_m since S_m is compact.)

Lemma 47.7 (Comparing λ_m and λ_{hm}). *Let* $m \in \{1:I\}$. *Assume that* $\sigma_{hm} \neq 0$. *The following holds true:*

$$\lambda_m \leq \lambda_{hm} \leq \sigma_{hm}^{-2}\lambda_m. \tag{47.11}$$

Proof. Let $w_h = \sum_{i \in \{1:m\}} \mathsf{W}_i\psi_{hi} \in W_{hm} := \mathrm{span}\{\psi_{hi}\}_{i \in \{1:m\}}$, where the eigenfunctions are chosen and normalized as in Proposition 47.6(ii), so that $\|\psi_{hi}\|_{L^2(D)} = 1$. Then $R(w_h) = \sum_{i \in \{1:m\}} \lambda_{hi}\mathsf{W}_i^2 / \sum_{i \in \{1:m\}} \mathsf{W}_i^2$. We infer that $\lambda_{hm} = \max_{w_h \in W_{hm}} R(w_h)$, and the first inequality in (47.11) is a consequence of Proposition 47.4. Let us now prove the second inequality. We observe

that $\ker(G_h) \cap W_m = \{0\}$ since $\sigma_{hm} \neq 0$ by assumption. Hence, the rank nullity theorem implies that $\dim(G_h(W_m)) = m$. Let $W_{h,m-1} = \operatorname{span}\{\psi_{hi}\}_{i\in\{1:m-1\}}$ and consider the L^2-projection from $G_h(W_m)$ onto $W_{h,m-1}$. The rank nullity theorem implies that there is a nonzero vector $v_h \in G_h(W_m)$ such that v_h is L^2-orthogonal to $W_{h,m-1}$, so that $v_h = \sum_{i\in\{m:I\}} V_i \psi_{hi}$. It follows that $R(v_h) \geq \lambda_{hm}$. As a result, we have

$$\lambda_{hm} \leq R(v_h) \leq \max_{w_h \in G_h(W_m)} \frac{a(w_h, w_h)}{\|w_h\|^2_{L^2(D)}} = \max_{v\in W_m} \frac{a(G_h(v), G_h(v))}{\|G_h(v)\|^2_{L^2(D)}}.$$

Using that $a(G_h(v), G_h(v)) = a(v, G_h(v)) \leq a(v,v)^{\frac{1}{2}} a(G_h(v), G_h(v))^{\frac{1}{2}}$ since a is symmetric and coercive, we infer that $a(G_h(v), G_h(v)) \leq a(v,v)$. Recalling that $\max_{v\in W_m} R(v) = \lambda_m$, we conclude that

$$\lambda_{hm} \leq \max_{v\in W_m} \frac{a(v, v)}{\|G_h(v)\|^2_{L^2(D)}} \leq \max_{v\in W_m} \frac{\|v\|^2_{L^2(D)}}{\|G_h(v)\|^2_{L^2(D)}} \max_{v\in W_m} R(v)$$
$$= \sigma_{hm}^{-2} \max_{v\in W_m} R(v) = \sigma_{hm}^{-2} \lambda_m. \qquad \square$$

Remark 47.8 (Guaranteed upper bound). It is remarkable that independently of the approximation space, but provided conformity holds true, i.e., $V_h \subset V$, each eigenvalue of the discrete problem (47.8) is larger than the corresponding eigenvalue of the exact problem (46.17). In other words, the discrete eigenvalue λ_{hm} is a guaranteed upper bound on the exact eigenvalue λ_m for all $m \in \{1:I\}$. Estimating computable lower bounds on the eigenvalues using conforming elements is more challenging. We refer the reader to Cancès et al. [104] for a literature overview and to Remark 48.13 when the approximation setting is nonconforming. $\qquad \square$

Lemma 47.9 (Lower bound on σ_{hm}). *Let $m \in \{1:I\}$. Recall that S_m is the unit sphere of $W_m := \operatorname{span}\{\psi_i\}_{i\in\{1:m\}}$ in $L^2(D)$ and recall that $G_h : V_h \to V$ is the discrete solution operator. The following holds true:*

$$\sigma_{hm}^2 \geq 1 - 2\sqrt{m} \frac{\|a\|}{\lambda_1} \max_{v\in S_m} \|v - G_h(v)\|^2_V. \tag{47.12}$$

Proof. Let $v \in S_m$. Let $(V_i)_{i\in\{1:m\}}$ be the coordinate vector of v relative to the basis $\{\psi_i\}_{i\in\{1:m\}}$. Since $(\psi_i, \psi_j)_{L^2(D)} = \delta_{ij}$, we have $\sum_{i\in\{1:m\}} V_i^2 = \|v\|^2_{L^2(D)} = 1$. In addition, $\|G_h(v)\|^2_{L^2(D)}$ can be bounded from below as

$$\|G_h(v)\|^2_{L^2(D)} = \|v\|^2_{L^2(D)} - 2(v, v - G_h(v))_{L^2(D)} + \|v - G_h(v)\|^2_{L^2(D)}$$
$$\geq \|v\|^2_{L^2(D)} - 2(v, v - G_h(v))_{L^2(D)}$$
$$= 1 - 2(v, v - G_h(v))_{L^2(D)}. \tag{47.13}$$

Using that (λ_i, ψ_i) is an eigenpair, the symmetry of a, and the Galerkin orthogonality property satisfied by the discrete solution map, we have

$$(v, v - G_h(v))_{L^2(D)} = \sum_{i\in\{1:m\}} V_i (\psi_i, v - G_h(v))_{L^2(D)}$$
$$= \sum_{i\in\{1:m\}} \frac{V_i}{\lambda_i} a(\psi_i, v - G_h(v)) = \sum_{i\in\{1:m\}} \frac{V_i}{\lambda_i} a(\psi_i - G_h(\psi_i), v - G_h(v)).$$

This implies that

$$(v, v - G_h(v))_{L^2(D)} \leq \frac{\|a\|}{\lambda_1} \|v - G_h(v)\|_V \sum_{i \in \{1:m\}} |\mathsf{V}_i| \|\psi_i - G_h(\psi_i)\|_V$$

$$\leq \frac{\|a\|}{\lambda_1} \max_{w \in S_m} \|w - G_h(w)\|_V^2 \sum_{i \in \{1:m\}} |\mathsf{V}_i|$$

$$\leq \sqrt{m} \frac{\|a\|}{\lambda_1} \max_{w \in S_m} \|w - G_h(w)\|_V^2,$$

where we used the boundedness of a and $\lambda_1 \leq \lambda_i$ for all $i \in \{1:m\}$ in the first bound, that $v \cup \{\psi_i\}_{i \in \{1:m\}} \subset S_m$ in the second bound, and the Cauchy–Schwarz inequality and $\sum_{i \in \{1:m\}} \mathsf{V}_i^2 = 1$ in the third bound. The expected estimate is obtained by inserting this bound into (47.13) and taking the infimum over $v \in S_m$ (recall that $\sigma_{hm} := \min_{v \in S_m} \|G_h(v)\|_{L^2(D)}$). $\quad\square$

Theorem 47.10 (Error on eigenvalues). *Let $m \in \mathbb{N}\setminus\{0\}$ and $c_1(m) := 4\sqrt{m}\frac{\|a\|}{\lambda_1}\frac{\|a\|}{\alpha}$. There is $h_0(m) > 0$ s.t. for all $h \in \mathcal{H} \cap (0, h_0(m)]$, we have $\sigma_{hm} \geq \frac{1}{2}$ and*

$$0 \leq \lambda_{hm} - \lambda_m \leq \lambda_m c_1(m) \max_{v \in S_m} \min_{v_h \in V_h} \|v - v_h\|_V^2. \tag{47.14}$$

Proof. (1) Since I grows unboundedly as $h \downarrow 0$, there is $h_0'(m) > 0$ s.t. $m \in \{1:I\}$ for all $h \in \mathcal{H} \cap (0, h_0'(m)]$, i.e., the pair $(\lambda_{hm}, \psi_{hm})$ exists for all $h \in \mathcal{H} \cap (0, h_0'(m)]$. Moreover, since the unit sphere S_m is compact, there is $v_*(m) \in S_m$ such that $\max_{v \in S_m} \|v - G_h(v)\|_V^2 = \|v_*(m) - G_h(v_*(m))\|_V^2$. The approximation property of the sequence $(V_h)_{h \in \mathcal{H}}$ implies that there is $h_0''(m) > 0$ such that $c_0(m)\|v_*(m) - G_h(v_*(m))\|_V^2 \leq \frac{1}{2}$ for all $h \in \mathcal{H} \cap (0, h_0''(m)]$, with $c_0(m) := 2\sqrt{m}\frac{\|a\|}{\lambda_1}$. We now set $h_0(m) := \min(h_0'(m), h_0''(m))$. Observing that $\frac{1}{1-x} \leq 1 + 2x$ for all $x \in [0, \frac{1}{2}]$, and applying this inequality to (47.12) with $x := c_0(m) \max_{v \in S_m} \|v - G_h(v)\|_V^2 \leq \frac{1}{2}$, we infer that $\sigma_{hm}^{-2} \leq 1 + 2c_0(m) \max_{v \in S_m} \|v - G_h(v)\|_V^2$. This implies in particular that $\sigma_{hm} \geq \frac{1}{\sqrt{2}} \geq \frac{1}{2}$ for all $h \in \mathcal{H} \cap (0, h_0(m)]$.
(2) Inserting the above bound into (47.11) yields

$$\lambda_{hm} - \lambda_m \leq (\sigma_{hm}^{-2} - 1)\lambda_m \leq 2\lambda_m c_0(m) \max_{v \in S_m} \|v - G_h(v)\|_V^2.$$

Since a is symmetric and coercive, Céa's lemma (Lemma 26.13) implies that

$$\|v - G_h(v)\|_V \leq \left(\frac{\|a\|}{\alpha}\right)^{\frac{1}{2}} \min_{v_h \in V_h} \|v - v_h\|_V. \tag{47.15}$$

The assertion follows readily. $\quad\square$

Remark 47.11 (Units). One readily sees that $\frac{\|a\|}{\lambda_1}$ scales as $\|\cdot\|_{L^2(D)}^{-2}$, i.e., as ℓ_D^{-2d}. Since $\|\cdot\|_V^2$ also scales like ℓ_D^{2d} owing to our assumption on the boundedness of the embedding $V \hookrightarrow L^2(D)$, we infer that the factor $c_1(m) \max_{v \in S_m} \min_{v_h \in V_h} \|v - v_h\|_V^2$ is nondimensional. $\quad\square$

Remark 47.12 (Double rate). The elliptic regularity theory implies that for all $m \geq 1$, there are $s(m) > 0$ and c_m s.t. $\|\psi_m\|_{H^{1+s(m)}(D)} \leq c_m$. Here, the value of $s(m)$ is not restricted to the interval $(0,1]$ since there is a bootstrapping phenomenon that allows $s(m)$ to be large. To illustrate this property, assume that D is of class $C^{r+1,1}$, $r \in \mathbb{N}$, and the bilinear form a is associated with an operator A satisfying the assumptions of Theorem 31.29. Let $s := r \mod 2 \in$

$\{0,1\}$ and let $l^\sharp \in \mathbb{N}\backslash\{0\}$ be s.t. $2(l^\sharp - 1) + s = r$. Theorem 31.29 implies that there is $c_0(r)$ such that $\|A^{-1}(v)\|_{H^s(D)} \leq c_0(r)\ell_D^2\|v\|_{L^2(D)}$ for all $v \in L^2(D)$, and there are $c_l(r)$, such that $\|A^{-1}(v)\|_{H^{2l+s}(D)} \leq c_l(r)\ell_D^2\|v\|_{L^{2(l-1)+s}(D)}$ for all $v \in H^{2(l-1)+s}(D)$ and all $l \in \{1{:}l^\sharp\}$. Since $A(\psi_m) = \lambda_m\psi_m$, we obtain

$$\|\psi_m\|_{H^{r+2}(D)} = \|\psi_m\|_{H^{2l^\sharp+s}(D)} \leq c_{l^\sharp}(r)\dots c_1(r)c_0(r)(\lambda_m\ell_D^2)^{l^\sharp+1}\|\psi_m\|_{L^2(D)}.$$

Recalling the normalization $\|\psi_m\|_{L^2(D)} = 1$, this argument shows that if D is of class $C^{r+1,1}$, we have $\|\psi_m\|_{H^{1+s(m)}(D)} \leq c_m$ with $s(m) := r+1$ and $c_m := c_{l^\sharp}(r)\dots c_1(r)c_0(r)(\lambda_m\ell_D^2)^{l^\sharp+1}$. Recalling that k is the approximation degree of V_h, let $s_\flat(m) := \min(s(1), \dots, s(m), k)$ for all $m \geq 1$, and $\chi(m) := \max_{v \in S_m} \ell_D^{1+s_\flat(m)}|v|_{H^{1+s_\flat(m)}(D)}$ (recall that S_m is the unit sphere of W_m in $L^2(D)$). The best-approximation estimates established in §22.3 and §22.4 imply that there exists c_{app} such that the following holds true for all $h \in \mathcal{H} \cap (0, h_0(m)]$:

$$\max_{v \in S_m} \min_{v_h \in V_h} \|v - v_h\|_V \leq c_{\text{app}}\chi(m)(h/\ell_D)^{s_\flat(m)}.$$

Owing to Theorem 47.10, this implies that

$$0 \leq \lambda_{hm} - \lambda_m \leq \lambda_m c_1(m)c_{\text{app}}^2\chi(m)^2(h/\ell_D)^{2s_\flat(m)}. \qquad (47.16)$$

In the best-case scenario where $s(n) \geq k$ for all $n \in \{1{:}m\}$, we have $s_\flat(m) = k$ so that the convergence rate for the error on λ_m is $\mathcal{O}(h^{2k})$, i.e., this error converges at a rate that is double that of the best-approximation error on the eigenvectors in the H^1-norm; see Remark 47.16 below. Note that the convergence rate on λ_m in (47.16) depends on the smallest smoothness index of all the eigenfunctions $\{\psi_n\}_{n \in \{1:m\}}$. This shortcoming is circumvented with the more general theory presented in Chapter 48, where the convergence rate on λ_m only depends on the smoothness index of the eigenfunctions associated with λ_m. Note also that since $c_1(m)$ grows unboundedly with m, (47.16) shows that when h is fixed the accuracy of the approximation decreases as m increases. $\qquad\square$

Example 47.13 (1D Laplacian). Let us consider the eigenvalue problem for the one-dimensional Laplacian discretized using $\mathbb{P}_1$ Lagrange elements on a uniform mesh on $D := (0,1)$. It is shown in Exercise 47.5 that $\lambda_m = m^2\pi^2$ and $\lambda_{hm} = \frac{6}{h^2}\frac{1-\cos(m\pi h)}{2+\cos(m\pi h)}$ for all $m \geq 1$. The left panel of Figure 47.1 shows the first 50 exact eigenvalues and the 50 discrete eigenvalues on a mesh having $I := 50$ internal vertices. The exact eigenvalues are approximated from above as predicted in Lemma 47.7. Observe that only the first eigenvalues are approximated accurately. The reason for this is that the eigenfunctions corresponding to large eigenvalues oscillate too much to be represented accurately on the mesh as illustrated in the right panel of Figure 47.1. A rule of thumb is that a meshsize smaller than $\frac{\sqrt{\epsilon}}{m}$ must be used to approximate the m-th eigenvalue with relative accuracy ϵ, i.e., $|\lambda_{hm} - \lambda_m| < \epsilon\lambda_m$. For instance, only the first 10 eigenvalues are approximated within 1% accuracy when $I := 100$. We refer the reader to Exercise 47.5 for further details. $\qquad\square$

47.2.3 Eigenfunction error analysis

The goal of this section is to estimate the approximation error on the eigenfunctions. We first estimate this error in the L^2-norm and then in the H^1-norm. Let $m \geq 1$ be a fixed natural number, and let us assume as in the previous section that the meshsize $h \in \mathcal{H}$ is small enough so that $m \leq I$ and $\sigma_{hm} > 0$ (see Theorem 47.10). For the sake of simplicity, we also assume that the eigenvalue λ_m is simple, and we set $\gamma_m := 2\max_{i \in \mathbb{N}\backslash\{0,m\}} \frac{\lambda_m}{|\lambda_m - \lambda_i|}$. Observe that $\gamma_m = 2\max(\frac{\lambda_m}{\lambda_m - \lambda_{m-1}}, \frac{\lambda_m}{\lambda_{m+1} - \lambda_m})$. Since $\lambda_{hi} \to \lambda_i$ as $h \to 0$ for all $i \in \{1{:}m{+}1\}$ (see Theorem 47.10),
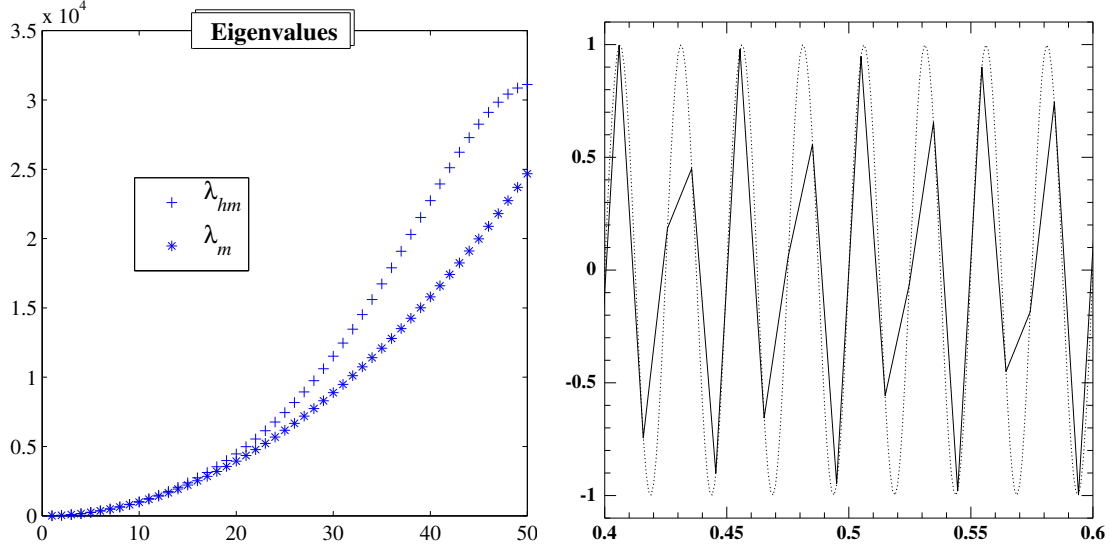
Figure 47.1: $\mathbb{P}_1$ approximation of the eigenvalues of the Laplacian in one dimension. Left: discrete and exact eigenvalues, $I := 50$. Right: Graph of the 80th exact (dashed line) and discrete (solid line) eigenfunctions in the interval $(0.4, 0.6)$, $I := 100$.

there exists $h_0(m) > 0$ so that $\frac{\lambda_m}{|\lambda_m - \lambda_{hi}|} \leq \gamma_m$ for all $i \in \{1{:}m+1\}\backslash\{m\}$ and all $h \in \mathcal{H}\cap(0, h_0(m)]$. Moreover, using that $|\lambda_m - \lambda_{hi}| \leq |\lambda_m - \lambda_{m+1}|$ for all $i \geq m+1$, we infer that the following holds true for all $h \in \mathcal{H}\cap(0, h_0(m)]$:

$$\max_{\substack{i\in\{1:I\}\\ i\neq m}} \frac{\lambda_m}{|\lambda_m - \lambda_{hi}|} \leq \gamma_m. \tag{47.17}$$

Theorem 47.14 (L^2-error on eigenfunctions). *Let $m \in \mathbb{N}\backslash\{0\}$. Assume that λ_m is simple and let $h_0(m) > 0$ be s.t. (47.17) holds true. Let $c_2(m) := 2(1 + \gamma_m)$. There is an eigenfunction ψ_m such that the following holds true for all $h \in \mathcal{H}\cap(0, h_0(m)]$:*

$$\|\psi_m - \psi_{hm}\|_{L^2(D)} \leq c_2(m)\|\psi_m - G_h(\psi_m)\|_{L^2(D)}. \tag{47.18}$$

Proof. Recall that $G_h(\psi_m) = \sum_{i\in\{1:I\}} \mathsf{V}_i\psi_{hi}$ with $\mathsf{V}_i := (G_h(\psi_m), \psi_{hi})_{L^2(D)}$. Let us set $v_{hm} := \mathsf{V}_m\psi_{hm}$ so that $G_h(\psi_m) - v_{hm} = \sum_{i\in\{1:I\}\backslash\{m\}} \mathsf{V}_i\psi_{hi}$. Since the bilinear form a is symmetric and $(\lambda_{hi}, \psi_{hi})$ is a discrete eigenpair, we have

$$\mathsf{V}_i = \frac{1}{\lambda_{hi}}a(\psi_{hi}, G_h(\psi_m)) = \frac{1}{\lambda_{hi}}a(G_h(\psi_m), \psi_{hi})$$
$$= \frac{1}{\lambda_{hi}}a(\psi_m, \psi_{hi}) = \frac{\lambda_m}{\lambda_{hi}}(\psi_m, \psi_{hi})_{L^2(D)},$$

where we used the definition of G_h and that (λ_m, ψ_m) is an eigenpair. This implies that

$$(\lambda_{hi} - \lambda_m)\mathsf{V}_i = \lambda_{hi}\mathsf{V}_i - \lambda_m\mathsf{V}_i = \lambda_m(\psi_m, \psi_{hi})_{L^2(D)} - \lambda_m\mathsf{V}_i$$
$$= \lambda_m(\psi_m, \psi_{hi})_{L^2(D)} - \lambda_m(G_h(\psi_m), \psi_{hi})_{L^2(D)}$$
$$= \lambda_m(\psi_m - G_h(\psi_m), \psi_{hi})_{L^2(D)}.$$

Hence, we have $V_i = \frac{\lambda_m}{\lambda_{hi} - \lambda_m}(\psi_m - G_h(\psi_m), \psi_{hi})_{L^2(D)}$ for all $i \in \{1:I\}\setminus\{m\}$. Since the discrete eigenfunctions $\{\psi_{hi}\}_{i \in \{1:I\}}$ are L^2-orthonormal, we obtain

$$\|G_h(\psi_m) - v_{hm}\|_{L^2(D)}^2 = \sum_{\substack{i \in \{1:I\} \\ i \neq m}} V_i^2 \leq \gamma_m^2 \sum_{\substack{i \in \{1:I\} \\ i \neq m}} (\psi_m - G_h(\psi_m), \psi_{hi})_{L^2(D)}^2$$
$$\leq \gamma_m^2 \|\psi_m - G_h(\psi_m)\|_{L^2(D)}^2, \qquad (47.19)$$

where the first bound follows from (47.17) and the last one from Bessel's inequality $\sum_{i \in \{1:I\}}(\psi_m - G_h(\psi_m), \psi_{hi})_{L^2(D)}^2 \leq \|\psi_m - G_h(\psi_m)\|_{L^2(D)}^2$. Let us now estimate $\|\psi_{hm} - v_{hm}\|_{L^2(D)}$. Since $\|\psi_{hm}\|_{L^2(D)} = 1$, we have

$$\|\psi_{hm} - v_{hm}\|_{L^2(D)} = \|(1 - V_m)\psi_{hm}\|_{L^2(D)} = |V_m - 1|$$
$$= |(G_h(\psi_m), \psi_{hm})_{L^2(D)} - 1|.$$

Assume that ψ_{hm} is chosen so that $V_m = (G_h(\psi_m), \psi_{hm})_{L^2(D)} \geq 0$. Then we have $\|v_{hm}\|_{L^2(D)} = |V_m| = (G_h(\psi_m), \psi_{hm})_{L^2(D)}$, and $\|\psi_{hm} - v_{hm}\|_{L^2(D)} = |\|v_{hm}\|_{L^2(D)} - 1|$. Since the triangle inequality implies that

$$\|\psi_m\|_{L^2(D)} - \|\psi_m - v_{hm}\|_{L^2(D)} \leq \|v_{hm}\|_{L^2(D)} \leq \|\psi_m\|_{L^2(D)} + \|\psi_m - v_{hm}\|_{L^2(D)},$$

and since $\|\psi_m\|_{L^2(D)} = 1$, we infer that $|\|v_{hm}\|_{L^2(D)} - 1| \leq \|\psi_m - v_{hm}\|_{L^2(D)}$. This implies that

$$\|\psi_{hm} - v_{hm}\|_{L^2(D)} = |\|v_{hm}\|_{L^2(D)} - 1| \leq \|\psi_m - v_{hm}\|_{L^2(D)}.$$

Invoking the triangle inequality, the above bound, and the triangle inequality one more time gives

$$\|\psi_m - \psi_{hm}\|_{L^2(D)} \leq \|\psi_m - G_h(\psi_m)\|_{L^2(D)} + \|G_h(\psi_m) - v_{hm}\|_{L^2(D)} + \|\psi_{hm} - v_{hm}\|_{L^2(D)}$$
$$\leq \|\psi_m - G_h(\psi_m)\|_{L^2(D)} + \|G_h(\psi_m) - v_{hm}\|_{L^2(D)} + \|\psi_m - v_{hm}\|_{L^2(D)}$$
$$\leq 2(\|\psi_m - G_h(\psi_m)\|_{L^2(D)} + \|G_h(\psi_m) - v_{hm}\|_{L^2(D)})$$
$$\leq 2(1 + \gamma_m)\|\psi_m - G_h(\psi_m)\|_{L^2(D)},$$

where the last bound follows from (47.19). Using the definition of $c_2(m)$ leads to the expected estimate. $\qquad\square$

Theorem 47.15 (H^1-error on eigenfunctions). *Let $m \in \mathbb{N}\setminus\{0\}$. Assume that λ_m is simple and let $h_0(m) > 0$ be s.t. (47.14) and (47.17) hold for all $h \in \mathcal{H} \cap (0, h_0(m)]$. There is an eigenfunction ψ_m such that the following holds true for all $h \in \mathcal{H} \cap (0, h_0(m)]$:*

$$\|\psi_m - \psi_{hm}\|_V \leq c_3(m) \max_{v \in S_m} \min_{v_h \in V_h} \|v - v_h\|_V, \qquad (47.20)$$

where $c_3(m) := (\frac{\lambda_m}{\alpha})^{\frac{1}{2}}(c_1(m) + c_2(m)^2 \frac{\|a\|}{\alpha})^{\frac{1}{2}}$ is independent of $h \in \mathcal{H}$.

Proof. Owing to the coercivity of a, we infer that

$$\alpha\|\psi_m - \psi_{hm}\|_V^2 \leq a(\psi_m - \psi_{hm}, \psi_m - \psi_{hm})$$
$$= \lambda_{hm} + \lambda_m - 2\lambda_m(\psi_m, \psi_{hm})_{L^2(D)}$$
$$= \lambda_{hm} - \lambda_m + \lambda_m\|\psi_m - \psi_{hm}\|_{L^2(D)}^2,$$

since $\|\psi_m\|_{L^2(D)} = \|\psi_{hm}\|_{L^2(D)} = 1$ implies that $\|\psi_m - \psi_{hm}\|_{L^2(D)}^2 = 2 - 2(\psi_m, \psi_{hm})_{L^2(D)}$. The inequality (47.20) is obtained by estimating $(\lambda_{hm} - \lambda_m)$ and $\|\psi_m - \psi_{hm}\|_{L^2(D)}^2$. The estimate on

$(\lambda_{hm} - \lambda_m)$ is given by (47.14) in Theorem 47.10, and Theorem 47.14 gives $\|\psi_m - \psi_{hm}\|_{L^2(D)} \leq c_2(m)\|\psi_m - G_h(\psi_m)\|_{L^2(D)}$. We observe that

$$\|\psi_m - G_h(\psi_m)\|_{L^2(D)} \leq \|\psi_m - G_h(\psi_m)\|_V \leq \max_{v \in S_m} \|v - G_h(v)\|_v$$

$$\leq \left(\frac{\|a\|}{\alpha}\right)^{\frac{1}{2}} \min_{v_h \in V_h} \|v - v_h\|_V,$$

where the last bound follows from (47.15) (Céa's lemma). Putting everything together leads to the expected estimate. □

Remark 47.16 (Convergence rates). Let us use the notation of Remark 47.12. Assume that the eigenvalue λ_m is simple. We can then invoke the estimates from Theorem 47.14 and Theorem 47.15. The best-approximation estimates in the H^1-norm established in §22.3 and §22.4 and the Aubin–Nitsche lemma (Lemma 32.11) imply that the following holds true for all $h \in \mathcal{H} \cap (0, h_0(m)]$:

$$\|\psi_m - \psi_{hm}\|_{L^2(D)} \leq \check{c}_2(m)\chi(m)(h/\ell_D)^{s_\flat(m)+s}, \tag{47.21a}$$

$$\|\psi_m - \psi_{hm}\|_{H^1(D)} \leq \check{c}_3(m)\chi(m)(h/\ell_D)^{s_\flat(m)}, \tag{47.21b}$$

where the constants $\check{c}_2(m), \check{c}_3(m)$ have the same dependencies w.r.t. m as the constants $c_2(m), c_3(m)$, and $\chi(m)$ is defined in Remark 47.12. The best possible convergence rates are obtained when $s_n(m) \geq k$ for all $n \in \{1:m\}$ so that $s_\flat(m) = k$, yielding the rates $\mathcal{O}(h^{k+1})$ in the L^2-norm and $\mathcal{O}(h^k)$ in the H^1-norm. Moreover, it can be shown that if λ_m has multiplicity p, i.e., $\lambda_m = \lambda_{m+1} = \ldots = \lambda_{m+p-1}$, then there exists an eigenfunction $\psi_m^\dagger \in \mathrm{span}\{\psi_m, \ldots, \psi_{m+p-1}\}$ with $\|\psi_m^\dagger\|_{L^2(D)} = 1$ such that (47.21) holds true with ψ_m replaced by $\psi_m^\dagger$. Note that (47.21) shows that when h is fixed, the accuracy of the approximation decreases as m increases, since $c_2(m), c_3(m)$ grow unboundedly with m. □

Exercises

Exercise 47.1 (Real eigenvalues). Consider the eigenvalue problem: Find $\psi \in H_0^1(D; \mathbb{C})\backslash\{0\}$ and $\lambda \in \mathbb{C}$ s.t. $\int_D (\nabla\psi \cdot \nabla\overline{w} + \psi\overline{w}) \, dx = \lambda \int_D \psi\overline{w} \, dx$ for all $w \in H_0^1(D; \mathbb{C})$. Prove directly that λ is real. (*Hint*: test with $w := \psi$.)

Exercise 47.2 (Smallest eigenvalue). Let $D_1 \subset D_2$ be two Lipschitz domains in $\mathbb{R}^d$. Let $a_i : H_0^1(D_i) \times H_0^1(D_i) \to \mathbb{R}$, $i \in \{1, 2\}$, be two symmetric, coercive, bounded bilinear forms. Assume that $a_1(v, w) = a_2(\widetilde{v}, \widetilde{w})$ for all $v, w \in H_0^1(D_1)$, where $\widetilde{v}, \widetilde{w}$ denote the extension by zero of v, w, respectively. Let $\lambda_1(D_i)$ be the smallest eigenvalue of the eigenvalue problem: Find $\psi \in H_0^1(D_i)\backslash\{0\}$ and $\lambda \in \mathbb{R}$ s.t. $a_i(\psi, w) = \lambda(\psi, w)_{L^2(D_i)}$ for all $w \in H_0^1(D_i)$. Prove that $\lambda_1(D_2) \leq \lambda_1(D_1)$. (*Hint*: use Proposition 47.3.)

Exercise 47.3 (Continuity of eigenvalues). Consider the setting defined in §47.1. Let $a_1, a_2 : V \times V \to \mathbb{R}$ be two symmetric, coercive, bounded bilinear forms. Let $A_1, A_2 : V \to V'$ be the linear operators defined by $\langle A_i(v), w\rangle_{V', V} := a_i(v, w)$, $i \in \{1, 2\}$, for all $v, w \in A$. Let $\lambda_k(a_1)$ and $\lambda_k(a_2)$ be the k-th eigenvalues, respectively. Prove that $|\lambda_k(a_1) - \lambda_k(a_2)| \leq \sup_{v \in S} |\langle(A_1 - A_2)(v), v\rangle_{V', V}|$, where S is the unit sphere in $L^2(D)$. (*Hint*: use the min-max principle.)

Exercise 47.4 (Max-min principle). Prove the second equality in (47.6). (*Hint*: let $E_{m-1} \in V_{m-1}$ and observe that $E_{m-1}^\perp \cap W_m \neq \{0\}$.)

Exercise 47.5 (Laplacian, 1D). Consider the spectral problem for the 1D Laplacian on $D :=$ $(0,1)$. (i) Show that the eigenpairs (λ_m, ψ_m) are $\lambda_m = m^2\pi^2$, $\psi_m(x) = \sin(m\pi x)$, for all $x \in D$ and all $m \geq 1$. (ii) Consider a uniform mesh of D of size $h := \frac{1}{I+1}$ and H^1-conforming $\mathbb{P}_1$ finite elements. Compute the stiffness matrix $\mathcal{A}$ and the mass matrix $\mathcal{M}$. (iii) Show that the eigenvalues of the discrete problem (47.8) are $\lambda_{hm} = \frac{6}{h^2}(\frac{1-\cos(m\pi h)}{2+\cos(m\pi h)})$ for all $m \in \{1{:}I\}$. (*Hint*: consider the vectors $(\sin(\pi hml))_{l\in\{1{:}I\}}$ for all $m \in \{1{:}I\}$.)

Exercise 47.6 (Stiffness matrix). Assume that the mesh sequence $(\mathcal{T}_h)_{h\in\mathcal{H}}$ is quasi-uniform. Estimate from below the smallest eigenvalue of the stiffness matrix $\mathcal{A}$ defined in (47.9) and estimate from above its largest eigenvalue. (*Hint*: see §28.2.3.)

Chapter 48

Nonsymmetric problems

In this chapter, we continue our investigation of the finite element approximation of eigenvalue problems, but this time we do not assume symmetry and we explore techniques that can handle nonconforming approximation settings. The main abstract results used in the present chapter are based on a theory popularized in the landmark review article by Babuška and Osborn [38]. Some results are simplified to avoid invoking spectral projections. Our objective is to show how to apply this abstract theory to the conforming and nonconforming approximation of eigenvalue problems arising from variational formulations.

48.1 Abstract theory

In this section, we present an abstract theory for the approximation of the spectrum of compact operators in complex Banach spaces, and we show how to apply it to spectral problems arising from variational formulations.

48.1.1 Approximation of compact operators

Let L be a complex Banach space and $T \in \mathcal{L}(L)$ be a compact operator. We assume that we have at hand a sequence of compact operators $T_n : L \to L$, $n \in \mathbb{N}$, that converges in norm to T i.e., we assume that

$$\lim_{n \to \infty} \|T - T_n\|_{\mathcal{L}(L)} = 0. \tag{48.1}$$

We want to estimate how the eigenpairs of each member in the sequence $(T_n)_{n \in \mathbb{N}}$ approximate some of the eigenpairs of T.

Recall that $\sigma(T) \backslash \{0\} = \sigma_\mathrm{p}(T) \backslash \{0\}$ and that the nonzero eigenvalues of T are isolated since T is compact; see Items (ii)-(iii) in Theorem 46.14. Let $\mu \in \sigma_\mathrm{p}(T) \backslash \{0\}$ be a nonzero eigenvalue of T. Let α be the ascent of μ. Recall that α is the smallest integer with the property that $\ker(\mu I_L - T)^\alpha = \ker(\mu I_L - T)^{\alpha+1}$. Denoting by $T^* : L' \to L'$ the adjoint of T, we set

$$G_\mu := \ker(\mu I_L - T)^\alpha, \qquad G_\mu^* := \ker(\overline{\mu} I_{L'} - T^*)^\alpha, \tag{48.2a}$$

$$m := \dim(G_\mu) = \dim(G_\mu^*). \tag{48.2b}$$

Members of G_μ and G_μ^* are called *generalized eigenvectors*. The generalized eigenvectors are all eigenvectors only if $\alpha = 1$. Recall that m is the algebraic multiplicity of μ and that $m \geq \alpha$; see

(46.5). Owing to the above assumption on norm convergence, it can be shown that there are m eigenvalues of T_n, say $\{\mu_{n,j}\}_{j\in\{1:m\}}$ (counted with their algebraic multiplicities), that converge to μ as $n \to \infty$. Let $\alpha_{n,j}$ be the ascent of $\mu_{n,j}$ and let us set

$$G_{n,\mu} := \sum_{j\in\{1:m\}} \ker(\mu_{n,j}I_L - T_n)^{\alpha_{n,j}}. \tag{48.3}$$

We want to evaluate how close the subspaces G_μ and $G_{n,\mu}$ are, and for this purpose we define the notion of gap. Given two closed subspaces of L, Y, and Z, we define $\delta(Y, Z) := \sup_{y\in Y; \|y\|_L = 1} \text{dist}(y, Z)$, where $\text{dist}(y, Z) := \inf_{z\in Z} \|y - z\|_L$. The gap between Y and Z is defined by

$$\widehat{\delta}(Y, Z) := \max(\delta(Y, Z), \delta(Z, Y)).$$

Theorem 48.1 (Bound on eigenspace gap). *Assume* (48.1). *Let* $\mu \in \sigma_\text{p}(T)\backslash\{0\}$. *Let* G_μ *be defined in* (48.2a) *and let* $G_{n,\mu}$ *be defined in* (48.3). *There is* c, *depending on* μ, *such that for all* $n \in \mathbb{N}$,

$$\widehat{\delta}(G_\mu, G_{n,\mu}) \leq c\,\|(T - T_n)_{|G_\mu}\|_{\mathcal{L}(G_\mu; L)}. \tag{48.4}$$

Proof. See Osborn [321, Thm. 1] or Babuška and Osborn [38, Thm. 7.1]. □

Let us now examine the convergence of the eigenvalues. When α, the ascent of μ, is larger than one, it is interesting to consider the convergence of the arithmetic mean of the eigenvalues $\mu_{n,j}$. We will see that this quantity converges faster than any of the $\mu_{n,j}$ (for instance, compare (48.5) and (48.6), and see (48.21) in Theorem 48.8).

Theorem 48.2 (Convergence of eigenvalues). *Assume* (48.1). *Let* $\mu \in \sigma_\text{p}(T)\backslash\{0\}$ *with algebraic multiplicity* m. *Let* $\{\mu_{n,j}\}_{j\in\{1:m\}}$ *be the eigenvalues of* T_n *that converge to* μ *and set* $\langle\mu_n\rangle := \frac{1}{m}\sum_{j\in\{1:m\}} \mu_{n,j}$. *There is* c, *depending on* μ, *such that for all* $n \in \mathbb{N}$,

$$\begin{aligned}
|\mu - \langle\mu_n\rangle| \leq &\frac{1}{m} \max_{(v,w)\in G_\mu\times G_\mu^*} \frac{|\langle w, (T - T_n)(v)\rangle_{L',L}|}{\|w\|_{L'}\|v\|_L} \\
&+ c\,\|(T - T_n)_{|G_\mu}\|_{\mathcal{L}(G_\mu; L)}\|(T - T_n)_{|G_\mu^*}^*\|_{\mathcal{L}(G_\mu^*; L')},
\end{aligned} \tag{48.5}$$

and for all $j \in \{1:m\}$,

$$\begin{aligned}
|\mu - \mu_{n,j}| \leq c\Bigg(&\max_{(v,w)\in G_\mu\times G_\mu^*} \frac{|\langle w, (T - T_n)(v)\rangle_{L',L}|}{\|w\|_{L'}\|v\|_L} \\
&+ \|(T - T_n)_{|G_\mu}\|_{\mathcal{L}(G_\mu; L)}\|(T - T_n)_{|G_\mu^*}^*\|_{\mathcal{L}(G_\mu^*; L')}\Bigg)^{\frac{1}{\alpha}}.
\end{aligned} \tag{48.6}$$

Proof. See [321, Thm. 3&4], [38, Thm. 7.2&7.3], and Exercise 48.3. □

Finally, we evaluate how the vectors in $G_{n,\mu}$ approximate those in G_μ.

Theorem 48.3 (Convergence of eigenvectors). *Assume* (48.1). *Let* $\mu \in \sigma_\text{p}(T)\backslash\{0\}$ *with algebraic multiplicity* m. *Let* $\{\mu_{n,j}\}_{j\in\{1:m\}}$ *be the eigenvalues of* T_n *that converge to* μ. *For all integers* $j \in \{1:m\}$ *and* $\ell \in \{1:\alpha\}$, *let* $w_{n,j}$ *be a unit vector in* $\ker(\mu_{n,j}I_L - T_n)^\ell$. *There is* c, *depending on* μ, *such that for every integer* $\ell' \in \{\ell:\alpha\}$, *there is a unit vector* $u_{\ell'} \in \ker(\mu I_L - T)^{\ell'} \subset G_\mu$ *such that for all* $n \in \mathbb{N}$,

$$\|u_{\ell'} - w_{n,j}\|_L \leq c\,\|(T - T_n)_{|G_\mu}\|_{\mathcal{L}(G_\mu; L)}^{\frac{\ell'-\ell+1}{\alpha}}. \tag{48.7}$$

Proof. See [321, Thm. 5] or [38, Thm. 7.4]. □

Remark 48.4 (Literature). The above theory has been developed by Bramble and Osborn [78], Osborn [321], Descloux et al. [161, 162]; see Vainikko [368, 369], Strang and Fix [359] for earlier references. Overviews can also be found in Boffi [62], Chatelin [116, Chap. 6]. □

Remark 48.5 (Sharper bounds). The bounds in Theorem 48.2 are simplified versions of the estimates given in [321, Thm. 3&4]. Therein, instead of $\max_{(v,w)\in G_\mu\times G_\mu^*} \frac{|\langle w,(T-T_n)(v)\rangle_{L',L}|}{\|w\|_{L'}\|v\|_L}$, one has $\sum_{j\in\{1:m\}}|\langle \phi_j^*,(T-T_n)(\phi_j)\rangle_{L',L}|$, where $\{\phi_j\}_{j\in\{1:m\}}$ is a basis of G_μ and $\{\phi_j^*\}_{j\in\{1:m\}}$ is a dual basis of G_μ^*, i.e., $\langle \phi_j^*,\phi_k\rangle_{L',L}=\delta_{jk}$ and the action of the forms ϕ_j^* outside G_μ is defined by selecting an appropriate complement of G_μ. The expressions given in Theorem 48.2 will suffice for our purpose. □

48.1.2 Application to variational formulations

Let $V \hookrightarrow L$ be a complex Banach space with compact embedding and let $a : V\times V \to \mathbb{C}$ be a bounded sesquilinear form. We assume that the sesquilinear form a satisfies the two conditions of the BNB theorem (Theorem 25.9), but we do not assume that a is Hermitian. Let $b : L\times L \to \mathbb{C}$ be another bounded sesquilinear form. We now consider the following eigenvalue problem:

$$\begin{cases} \text{Find } \psi \in V\setminus\{0\} \text{ and } \lambda \in \mathbb{C} \text{ such that} \\ a(\psi,w) = \lambda b(\psi,w), \quad \forall w \in V. \end{cases} \tag{48.8}$$

If (λ,ψ) solves (48.8), we say that (λ,ψ) is an eigenpair of the form a relative to the form b, or simply (λ,ψ) is an eigenpair of (48.8) when the context is unambiguous.

To reformulate (48.8) so as to fit the approximation theory of the spectrum of compact operators from §48.1.1, we define the solution operator $T : L \to V \hookrightarrow L$ such that

$$a(T(v),w) := b(v,w), \qquad \forall v \in L, \ \forall w \in V. \tag{48.9}$$

Note that $T(v)$ is well defined for all $v \in L$ since a satisfies the two BNB conditions. Notice also that $\text{im}(T) \subset V$ and that T is injective.

Proposition 48.6 (Spectrum of T). (i) $0 \notin \sigma_{\mathrm{p}}(T)$. (ii) $(\mu,\psi) \in \mathbb{C}\times V$ is an eigenpair of T iff $(\mu^{-1},\psi) \in \mathbb{C}\times V$ is an eigenpair of (48.8).

Proof. (i) If $(0,\psi)$ is an eigenpair of T (i.e., $\psi \neq 0$), then $a(\psi,v) = 0$ for all $v \in V$, and the inf-sup condition on a implies that $\psi = 0$, which is a contradiction.
(ii) Let (μ,ψ) be an eigenpair of T, i.e., $\mu^{-1}T(\psi) = \psi$ (notice that $\mu \neq 0$ since T is injective). We infer that

$$\mu^{-1}b(\psi,w) = b(\mu^{-1}\psi,w) = a(T(\mu^{-1}\psi),w) = a(\mu^{-1}T(\psi),w) = a(\psi,w),$$

for all $w \in V$. Hence, (μ^{-1},ψ) is an eigenpair of (48.8). The proof of the converse statement is identical. □

We refer the reader to §46.2 for various examples of spectral problems that can be put into the variational form (48.8). For instance, the model problem (46.21) leads to a sesquilinear form a that is not Hermitian since we have $a(v,w) := \int_D (g'(u_{\mathrm{sw}})\nabla v\cdot\nabla\overline{w} + vg''(u_{\mathrm{sw}})\nabla u_{\mathrm{sw}}\cdot\nabla\overline{w} - f'(u_{\mathrm{sw}})v\overline{w})\,\mathrm{d}x$, $V := H^1_{\mathrm{per}}(D)$, and $b(v,w) := \int_D v\overline{w}\,\mathrm{d}x$. An example with a sesquilinear form b that is not the L^2-inner product is obtained from the vibrating string model from §46.2.1 by assuming that the string has a nonuniform bounded linear density ρ. In this case, one recovers the model problem (48.8) with $V := H^1_0(D;\mathbb{R})$, $D := (0,\ell)$, where ℓ is the length of the string, $a(v,w) := \int_D \tau\partial_x v\partial_x w\,\mathrm{d}x$, where $\tau > 0$ is the uniform tension of the string, and $b(v,w) := \int_D \rho vw\,\mathrm{d}x$.

48.2   Conforming approximation

The goal of this section is to illustrate the approximation theory from §48.1 when applied to the conforming approximation of the model problem (48.8). Let V be a closed subspace of $H^1(D)$ which, depending on the boundary conditions that are enforced, satisfies $H_0^1(D) \subseteq V \subseteq H^1(D)$. We assume that V is equipped with a norm that is equivalent to that of $H^1(D)$. We assume also that the V-norm is rescaled so the operator norm of the embedding $V \hookrightarrow L^2(D)$ is at most one, e.g., one could set $\|v\|_V := C_{\mathrm{PS}}^{-1} \ell_D \|\nabla v\|_{L^2(D)}$ if $V := H_0^1(D)$, where C_{PS} is the constant from the Poincaré–Steklov inequality (31.12) in $H_0^1(D)$ and ℓ_D is a characteristic length associated with D, e.g., $\ell_D := \mathrm{diam}(D)$.

Let $T : L^2(D) \to L^2(D)$ be the compact operator defined in (48.9). We identify L and L', so that $T^* = T^{\mathsf{H}}$ (see Lemma 46.15). We want to approximate the spectrum of T assuming that we have at hand an H^1-conforming approximation setting. More precisely, assume that D is a Lipschitz polyhedron and let $(\mathcal{T}_h)_{h \in \mathcal{H}}$ be a shape-regular sequence of affine meshes so that each mesh covers D exactly. Let $k \geq 1$ be the polynomial degree of the approximation. We denote by V_h the H^1-conforming finite element space based on $\mathcal{T}_h$ such that $P_{k,0}^{\mathrm{g}}(\mathcal{T}_h) \subseteq V_h \subseteq P_k^{\mathrm{g}}(\mathcal{T}_h)$ and $V_h \subset V$ (see §19.2.1 or §19.4). To avoid being specific on the type of finite element we use, we assume the following best-approximation result:

$$\min_{v_h \in V_h} \|v - v_h\|_V \leq c\, h^r \ell_D |v|_{H^{1+r}(D)}, \tag{48.10}$$

for all $v \in H^{1+r}(D) \cap V$ and all $r \in [0, k]$. We assume that there is $\alpha_0 > 0$ such that for all $h \in \mathcal{H}$,

$$\inf_{v_h \in V_h} \sup_{w_h \in V_h} \frac{|a(v_h, w_h)|}{\|v_h\|_{H^1(D)} \|w_h\|_{H^1(D)}} \geq \alpha_0. \tag{48.11}$$

Since the sesquilinear form b may differ from the L^2-inner product, we additionally introduce the linear operator $S_* : L^2(D) \to V \hookrightarrow L^2(D)$ s.t.

$$a(v, S_*(w)) = (v, w)_{L^2(D)}, \qquad \forall v \in V,\ \forall w \in L^2(D). \tag{48.12}$$

Notice that we use the L^2-inner product on the right-hand side of (48.12) instead of the sesquilinear form b as we did for the definition of T in (48.9). We also assume that the following elliptic regularity pickup holds true for T and S_* (see §31.4.2): There are real numbers $\tau, \tau^* \in (0, 1]$ such that

$$T \in \mathcal{L}(L^2(D); H^{1+\tau}(D)), \qquad S_* \in \mathcal{L}(L^2(D); H^{1+\tau^*}(D)). \tag{48.13}$$

We have $\tau = \tau^* := 1$ when maximal elliptic regularity occurs.

The discrete counterpart of the eigenvalue problem (48.8) is formulated as follows:

$$\begin{cases} \text{Find } \psi_h \in V_h \backslash \{0\} \text{ and } \lambda_h \in \mathbb{C} \text{ such that} \\ a(\psi_h, w_h) = \lambda_h b(\psi_h, w_h), \quad \forall w_h \in V_h. \end{cases} \tag{48.14}$$

We define the discrete solution operator $T_h : L^2(D) \to V_h \subset L^2(D)$ s.t. for all $v \in L^2(D)$, $T_h(v) \in V_h$ is the unique solution to the following problem:

$$a(T_h(v), w_h) = b(v, w_h), \qquad \forall w_h \in V_h.$$

Notice that 0 cannot be an eigenvalue of (48.14) owing to the inf-sup condition (48.11) satisfied by a on $V_h \times V_h$. Moreover, (λ_h, ψ_h) is an eigenpair of (48.14) iff (λ_h^{-1}, ψ_h) is an eigenpair of T_h.

Lemma 48.7 (Bound on $(T - T_h)$). *There is c such that for all $t, t^* \in [0, k]$, all $v \in L^2(D)$ s.t. $T(v) \in H^{1+t}(D)$, all $w \in L^2(D)$ s.t. $S_*(w) \in H^{1+t^*}(D)$, and all $h \in \mathcal{H}$,*

$$\left|((T - T_h)(v), w)_{L^2(D)}\right| \leq c\, h^{t+t^*} \|a\| \ell_D^2 |T(v)|_{H^{1+t}(D)} |S_*(w)|_{H^{1+t^*}(D)}. \tag{48.15}$$

Proof. Lemma 26.14 and the best-approximation property (48.10) imply that $\|(T - T_h)(v)\|_V \leq ch^t \ell_D |T(v)|_{H^{1+t}(D)}$. Since $(T - T_h)(v) \in V$, the Galerkin orthogonality property and the boundedness of a imply that

$$
\begin{aligned}
\left|((T - T_h)(v), w)_{L^2(D)}\right| &= |a((T - T_h)(v), S_*(w))| \\
&\leq \inf_{w_h \in V_h} |a((T - T_h)(v), S_*(w) - w_h)| \\
&\leq \|a\| \, \|(T - T_h)(v)\|_V \inf_{w_h \in V_h} \|S_*(w) - w_h\|_V.
\end{aligned}
$$

Using the above bound on $(T - T_h)(v)$ and the best-approximation property (48.10) to bound $\|S_*(w) - w_h\|_V$ leads to the expected estimate. $\qquad\square$

The estimate (48.15) with $t := \tau$ and $t^* := \tau^*$ combined with the regularity property (48.13) implies that

$$\|T - T_h\|_{\mathcal{L}(L^2; L^2)} \leq c\, h^{\tau + \tau^*} \big(\|a\| \ell_D^2 \|T\|_{\mathcal{L}(L^2; H^{1+\tau})} \|S_*\|_{\mathcal{L}(L^2; H^{1+\tau^*})}\big). \tag{48.16}$$

Since $\tau + \tau^* > 0$, this means that $T_h \to T$ in operator norm as $h \to 0$, that is, the key assumption (48.1) holds true. It is then legitimate to use the approximation results for compact operators stated in Theorems 48.1 to 48.3.

Let μ be a nonzero eigenvalue of T of ascent α and algebraic multiplicity m, and let

$$G_\mu := \ker(\mu I_{L^2} - T)^\alpha, \qquad G_\mu^* := \ker(\overline{\mu} I_{L^2} - T^{\mathsf{H}})^\alpha, \tag{48.17}$$

so that $m := \dim(G_\mu) = \dim(G_\mu^*)$ (see (48.2)). Recall that Proposition 48.6 implies that $\lambda := \mu^{-1}$ is an eigenvalue for (48.8). Since the smoothness of the generalized eigenvectors may differ from one eigenvalue to the other, we now define τ_μ and τ_μ^* to be the two largest real numbers in $(0, k]$ such that

$$T_{|G_\mu} \in \mathcal{L}(G_\mu; H^{1+\tau_\mu}(D)), \qquad S_{*|G_\mu^*} \in \mathcal{L}(G_\mu^*; H^{1+\tau_\mu^*}(D)), \tag{48.18}$$

where G_μ and G_μ^* are equipped with the L^2-norm. The two real numbers τ_μ and τ_μ^* measure the smoothness of the generalized eigenvectors in G_μ and G_μ^*, respectively. Notice that $\tau_\mu \in [\tau, k]$ and $\tau_\mu^* \in [\tau^*, k]$, where τ and τ^* are defined in (48.13) and are both in $(0, 1]$. We can set $\tau_\mu = \tau_\mu^* := k$ when maximal smoothness is available. It may happen that $\tau_\mu < \tau_\mu^*$ even if a is Hermitian. For instance, this may be the case if $b(v, w) := \int_D \rho v \overline{w} \, dx$, where the function ρ is a bounded discontinuous function.

Owing to the norm convergence of T_h to T as $h \to 0$, there are m eigenvalues of T_h, say $\{\mu_{h,j}\}_{j \in \{1:m\}}$ (counted with their algebraic multiplicities), that converge to μ as $h \to 0$. Let

$$G_{h,\mu} := \sum_{j \in \{1:m\}} \ker(\mu_{h,j} I_{L^2} - T_h)^{\alpha_{h,j}}, \tag{48.19}$$

where $\alpha_{h,j}$ is the ascent of $\mu_{h,j}$. We are now in the position to state the main result of this section.

Theorem 48.8 (Convergence of eigenspace gap, eigenvalues, and eigenvectors). *Let $\mu \in \sigma_{\mathrm{p}}(T) \setminus \{0\}$ with algebraic multiplicity m and let $\{\mu_{h,j}\}_{j \in \{1:m\}}$ be the eigenvalues of T_h that converge to μ. Let G_μ be defined in (48.17) and let $G_{h,\mu}$ be defined in (48.19). There is c, depending on μ, such that for all $h \in \mathcal{H}$,*

$$\widehat{\delta}(G_\mu, G_{h,\mu}) \le c\, h^{\tau_\mu + t^*}, \tag{48.20}$$

and letting $\langle \mu_h \rangle := \frac{1}{m} \sum_{j \in \{1:m\}} \mu_{h,j}$, we have

$$|\mu - \langle \mu_h \rangle| \le c\, h^{\tau_\mu + \tau_\mu^*}, \qquad |\mu - \mu_{h,j}| \le c\, h^{\frac{1}{\alpha}(\tau_\mu + \tau_\mu^*)}, \ \forall j \in \{1:m\}. \tag{48.21}$$

Moreover, for all integers $j \in \{1:m\}$ and $\ell \in \{1:\alpha\}$, let $w_{h,j}$ be a unit vector in $\ker(\mu_{h,j} I_{L^2} - T_h)^\ell$. There is c, depending on μ, such that for every integer $\ell' \in \{\ell:\alpha\}$, there is a unit vector $u_{\ell'} \in \ker(\mu I_{L^2} - T)^{\ell'} \subset G_\mu$ such that for all $h \in \mathcal{H}$,

$$\|u_{\ell'} - w_{h,j}\|_{L^2(D)} \le c\, h^{\frac{\ell' - \ell + 1}{\alpha}(\tau_\mu + \tau^*)}. \tag{48.22}$$

In the above estimates, the constant c depends on $\|a\|_{\ell_D^2}$ and on the operator norms resulting from (48.13) and (48.18).

Proof. Using $t := \tau_\mu$ and $t^* := \tau^*$ in (48.15), we infer that

$$\|(T - T_h)_{|G_\mu}\|_{\mathcal{L}(G_\mu; L^2)} = \sup_{v \in G_\mu} \sup_{w \in L^2} \frac{((T - T_h)(v), w)_{L^2}}{\|v\|_{L^2} \|w\|_{L^2}} \le c\, h^{\tau_\mu + \tau^*}.$$

Similarly, using $t := \tau$ and $t^* := \tau_\mu^*$ in (48.15), and recalling that $T^* = T^{\mathsf{H}}$ in the present case, we infer that

$$\|(T - T_h)^*|_{G_\mu^*}\|_{\mathcal{L}(G_\mu^*; L^2)} = \sup_{v \in L^2} \sup_{w \in G_\mu^*} \frac{(v, (T^{\mathsf{H}} - T_h^{\mathsf{H}})(w))_{L^2}}{\|v\|_{L^2} \|w\|_{L^2}}$$

$$= \sup_{v \in L^2} \sup_{w \in G_\mu^*} \frac{((T - T_h)(v), w)_{L^2}}{\|v\|_{L^2} \|w\|_{L^2}} \le c\, h^{\tau + \tau_\mu^*}.$$

Finally, using $t := \tau_\mu$ and $t^* := \tau_\mu^*$ in (48.15), we infer that

$$\sup_{v \in G_\mu} \sup_{w \in G_\mu^*} \frac{((T - T_h)(v), w)_{L^2}}{\|v\|_{L^2} \|w\|_{L^2}} \le c\, h^{\tau_\mu + \tau_\mu^*}.$$

The conclusion follows by applying Theorems 48.1-48.3. □

Remark 48.9 (Convergence rates). Notice that among the two terms that compose the right-hand side in (48.5), it is the first one that dominates when the meshsize goes to zero. The first term scales like $\mathcal{O}(h^{\tau_\mu + \tau_\mu^*})$, whereas the second one scales like $\mathcal{O}(h^{\tau_\mu + \tau_\mu^* + \tau + \tau^*})$ with $\tau + \tau^* > 0$. The same observation is valid for (48.6). □

Remark 48.10 (Symmetric case). The estimate (48.21) coincides with the estimate (47.16), and the estimate (48.22) (with $\alpha = \ell = \ell' := 1$) coincides with the estimate (47.21) when T is symmetric. Notice though that the estimates from Chapter 47 for the i-th eigenpair depend on the smoothness of all the unit eigenfunctions $\{\psi_n\}_{n \in \{1:i\}}$ (counting the multiplicities), whereas the estimates (48.21)-(48.22) depend only on the smoothness of the unit eigenvectors in G_{μ_i}; see Remark 47.12. □

48.3 Nonconforming approximation

We revisit the theory presented above in a nonconforming context. Typical examples we have in mind are the Crouzeix–Raviart approximation from Chapter 36, Nitsche's boundary penalty technique from Chapter 37, and the discontinuous Galerkin method from Chapter 38. The theory is also applicable to the hybrid high-order method from Chapter 39.

48.3.1 Discrete formulation

We consider again the model problem (48.8) and we want to approximate the spectrum of the operator $T : L^2(D) \to L^2(D)$ defined in (48.9) using an approximation setting that is not conforming in V.

To stay general, we assume that we have at hand a sequence of discrete spaces $(V_h)_{h \in \mathcal{H}}$ with $V_h \not\subset V$. For all $h \in \mathcal{H}$, the sesquilinear form a is approximated by a discrete sesquilinear form $a_h : V_h \times V_h \to \mathbb{C}$, and for simplicity we assume that the sesquilinear form b is meaningful on $V_h \times V_h$, i.e., we assume that $V_h \subset L^2(D)$. The discrete eigenvalue problem is formulated as follows:

$$\begin{cases} \text{Find } \psi_h \in V_h \backslash \{0\} \text{ and } \lambda_h \in \mathbb{C} \text{ such that} \\ a_h(\psi_h, w_h) = \lambda_h b(\psi_h, w_h), \quad \forall w_h \in V_h. \end{cases} \tag{48.23}$$

The discrete solution operator $T_h : L^2(D) \to V_h \subset L^2(D)$ and the adjoint discrete solution operator $S_{*h} : L^2(D) \to V_h \subset L^2(D)$ are defined as follows:

$$a_h(T_h(v), w_h) := b(v, w_h), \qquad\qquad \forall(v, w_h) \in L^2(D) \times V_h, \tag{48.24a}$$

$$a_h(v_h, S_{*h}(w)) := (v_h, w)_{L^2(D)}, \qquad \forall(v_h, w) \in V_h \times L^2(D). \tag{48.24b}$$

We assume that T_h and S_{*h} are both well defined, i.e., we assume that a_h satisfies an inf-sup condition on $V_h \times V_h$ uniformly w.r.t. $h \in \mathcal{H}$. As above, (λ_h, ψ_h) is an eigenpair of (48.23) iff (λ_h^{-1}, ψ_h) is an eigenpair of T_h.

To avoid unnecessary technicalities and to stay general, we make the following assumptions:
(i) There exists a dense subspace $V_{\mathrm{S}} \hookrightarrow V$ such that the solution operators T and S_* satisfy

$$T(v) \in V_{\mathrm{S}}, \qquad S_*(w) \in V_{\mathrm{S}}, \qquad \forall v, w \in L^2(D). \tag{48.25}$$

(ii) There is a sesquilinear form $a_\sharp$ extending a_h to $V_\sharp \times V_\sharp$, with $V_\sharp := V_{\mathrm{S}} + V_h$, i.e., $a_\sharp(v_h, w_h) = a_h(v_h, w_h)$ for all $v_h, w_h \in V_h$. The space $V_\sharp$ is equipped with a norm $\|\cdot\|_{V_\sharp}$ s.t. there is $\|a_\sharp\|$ such that

$$|a_\sharp(v, w)| \le \|a_\sharp\| \, \|v\|_{V_\sharp} \|w\|_{V_\sharp}, \quad \forall v, w \in V_\sharp, \ \forall h \in \mathcal{H}. \tag{48.26}$$

(iii) The sesquilinear forms $a_\sharp$ and a coincide on $V_{\mathrm{S}} \times V_{\mathrm{S}}$ so that

$$a_\sharp(T(v), S_*(w)) = a(T(v), S_*(w)), \qquad \forall v, w \in L^2(D). \tag{48.27}$$

(iv) Restricted Galerkin orthogonality and restricted adjoint Galerkin orthogonality, i.e., we have the following identities:

$$a_\sharp(T(v), w_h) = a_h(T_h(v), w_h), \qquad \forall(v, w_h) \in L^2(D) \times (V_h \cap V), \tag{48.28a}$$

$$a_\sharp(v_h, S_*(w)) = a_h(v_h, S_{*h}(w)), \qquad \forall(v_h, w) \in (V_h \cap V) \times L^2(D). \tag{48.28b}$$

(Notice that discrete test functions are restricted to $V_h \cap V$.)

(v) There is c such that for all $h \in \mathcal{H}$,

$$\|T(v) - T_h(v)\|_{V_\sharp} \leq c \inf_{v_h \in V_h \cap V} \|T(v) - v_h\|_{V_\sharp}, \tag{48.29a}$$

$$\|S_*(w) - S_{*h}(w)\|_{V_\sharp} \leq c \inf_{w_h \in V_h \cap V} \|S_*(w) - w_h\|_{V_\sharp}. \tag{48.29b}$$

Moreover, there is an integer $k \geq 1$, and there is c such that the following best-approximation property holds true for all $t \in [0, k]$, all $v \in H^{1+t}(D) \cap V$, and all $h \in \mathcal{H}$:

$$\inf_{v_h \in V_h \cap V} \|v - v_h\|_{V_\sharp} \leq c \, \ell_D h^t |v|_{H^{1+t}(D)}. \tag{48.30}$$

The reader is invited to verify whether all the above conditions are satisfied, with $V_s := V \cap H^{1+r}(D)$ and $r > \frac{1}{2}$, by the Crouzeix–Raviart approximation from Chapter 36, Nitsche's boundary penalty technique from Chapter 37, and the Discontinuous Galerkin method from Chapter 38.

48.3.2 Error analysis

We are going to use the general approximation results for compact operators stated in Theorems 48.1-48.3. Let $t_0 \geq 0$ be the smallest real number such that $H^{1+t_0}(D) \cap V \subset V_s$. We assume that $t_0 \leq k$, i.e., the interval $[t_0, k]$ is nonempty. In the applications we have in mind, t_0 is a number close to $\frac{1}{2}$ and $k \geq 1$.

Lemma 48.11 (Bound on $(T - T_h)$). *There is c s.t. for all $t, t^* \in [t_0, k]$, all $v \in L^2(D)$ s.t. $T(v) \in H^{1+t}(D)$, all $w \in L^2(D)$ s.t. $S_*(w) \in H^{1+t^*}(D)$, and all $h \in \mathcal{H}$,*

$$\left| ((T - T_h)(v), w)_{L^2(D)} \right| \leq c \, h^{t+t^*} \|a_\sharp\| \ell_D^2 |T(v)|_{H^{1+t}(D)} |S_*(w)|_{H^{1+t^*}(D)}. \tag{48.31}$$

Proof. Let $v \in L^2(D)$ be s.t. $T(v) \in H^{1+t}(D)$, and let $w \in L^2(D)$ be s.t. $S_*(w) \in H^{1+t^*}(D)$. We have $T(v) \in H^{1+t}(D) \cap V \subset V_s$ since $t \geq t_0$, and $S_*(w) \in H^{1+t^*}(D) \cap V \subset V_s$ since $t^* \geq t_0$. Using the definitions of S_* and S_{*h}, the assumption (48.27), i.e., that $a_\sharp$ and a coincide on $V_s \times V_s$ (and that $a_\sharp$ and a_h coincide over $V_h \times V_h$), and elementary manipulations, we infer that

$$\begin{aligned}
((T - T_h)(v), w)_{L^2(D)} &= a(T(v), S_*(w)) - a_h(T_h(v), S_{*h}(w)) \\
&= a_\sharp(T(v), S_*(w)) - a_\sharp(T_h(v), S_{*h}(w)) \\
&= a_\sharp(T(v) - T_h(v), S_*(w)) + a_\sharp(T_h(v), S_*(w) - S_{*h}(w)) \\
&= a_\sharp(T(v) - T_h(v), S_*(w) - S_{*h}(w)) + a_\sharp(T(v) - T_h(v), S_{*h}(w)) \\
&\quad + a_\sharp(T_h(v), S_*(w) - S_{*h}(w)) =: \mathfrak{T}_1 + \mathfrak{T}_2 + \mathfrak{T}_3.
\end{aligned}$$

Owing to the boundedness of $a_\sharp$ on $V_\sharp \times V_\sharp$ and the approximation properties (48.29)-(48.30), we have

$$|\mathfrak{T}_1| \leq c \, h^{t+t^*} \|a_\sharp\| \ell_D^2 |T(v)|_{H^{1+t}(D)} |S_*(w)|_{H^{1+t^*}(D)}.$$

The other two terms have a similar structure that can be dealt with by invoking the restricted Galerkin orthogonality (48.28). For instance, we have

$$\begin{aligned}
|\mathfrak{T}_2| &= \inf_{w_h \in V_h \cap V} |a_\sharp((T - T_h)(v), S_{*h}(w) - w_h)| \\
&\leq \|a_\sharp\| \|(T - T_h)(v)\|_{V_\sharp} \inf_{w_h \in V_h \cap V} \|S_{*h}(w) - w_h\|_{V_\sharp} \\
&\leq c \|a_\sharp\| \|(T - T_h)(v)\|_{V_\sharp} (\|S_{*h}(w) - S_*(w)\|_{V_\sharp} + \inf_{v_h \in V_h \cap V} \|S_*(w) - w_h\|_{V_\sharp}) \\
&\leq c' \, h^{t+t^*} \|a_\sharp\| \ell_D^2 |T(v)|_{H^{1+t}(D)} |S_*(w)|_{H^{1+t^*}(D)}.
\end{aligned}$$

The term $\mathfrak{T}_3$ is estimated similarly. $\qquad\square$

We assume now that the following elliptic regularity pickup holds true for T and S_* (see §31.4.2): There are real numbers $\tau \in (0,1]$ and $\tau^* \in (0,1]$ such that (48.13) holds true. The estimate (48.31) with $t := \tau$ and $t^* := \tau^*$ implies that

$$\|T - T_h\|_{\mathcal{L}(L^2, L^2)} \le c \, \|a_\sharp\| \ell_D^2 \|T\|_{\mathcal{L}(L^2; H^{1+\tau})} \|S_*\|_{\mathcal{L}(L^2; H^{1+\tau^*})} h^{\tau + \tau^*}.$$

Since $\tau + \tau^* > 0$, this means that $T_h \to T$ in operator norm as $h \to 0$, that is, the key assumption (48.1) holds true. It is then legitimate to use the approximation results for compact operators stated in Theorems 48.1-48.3.

Let μ be a nonzero eigenvalue of T of ascent α and algebraic multiplicity m, and let G_μ, G_μ be defined in (48.17). Proposition 48.6 implies that $\lambda := \mu^{-1}$ is an eigenvalue for (48.8). Let τ_μ and τ_μ^* be the two largest real numbers less than or equal to k satisfying (48.18). Recall that $\tau_\mu \in [\tau, k]$ and $\tau_\mu^* \in [\tau^*, k]$. Moreover, we can set $\tau_\mu = \tau_\mu^* := k$ when maximal smoothness is available.

Owing to the norm convergence T_h to T as $h \to 0$, there are m eigenvalues of T_h, say $\{\mu_{h,j}\}_{j \in \{1:m\}}$ (counted with their algebraic multiplicities), that converge to μ as $h \to 0$. Let $G_{h,\mu}$ be defined in (48.19). We are now in the position to state the main result of this section.

Theorem 48.12 (Convergence of eigenspace gap, eigenvalues, and eigenvectors). *Let $\mu \in \sigma_\mathrm{p}(T) \backslash \{0\}$ with algebraic multiplicity m and let $\{\mu_{h,j}\}_{j \in \{1:m\}}$ be the eigenvalues of T_h that converge to μ. There is c, depending on μ, s.t. for all $h \in \mathcal{H}$,*

$$\widehat{\delta}(G_\mu, G_{h,\mu}) \le c \, h^{\tau_\mu + \tau_\mu^*}, \tag{48.32}$$

and letting $\langle \mu_h \rangle := \frac{1}{m} \sum_{j \in \{1:m\}} \mu_{h,j}$, we have

$$|\mu - \langle \mu_h \rangle| \le c \, h^{\tau_\mu + \tau_\mu^*}, \qquad |\mu - \mu_{h,j}| \le c \, h^{\frac{1}{\alpha}(\tau_\mu + \tau_\mu^*)}, \ \forall j \in \{1:m\}. \tag{48.33}$$

Moreover, for all integers $j \in \{1:m\}$ and $\ell \in \{1:\alpha\}$, let $w_{h,j}$ be a unit vector in $\ker(\mu_{h,j} I_{L^2} - T_h)^\ell$. There is c, depending on μ, such that for every integer $\ell' \in \{\ell:\alpha\}$, there is a unit vector $u_{\ell'} \in \ker(\mu I_{L^2} - T)^{\ell'} \subset G_\mu$ such that for all $h \in \mathcal{H}$,

$$\|u_{\ell'} - w_{h,j}\|_{L^2(D)} \le c \, h^{\frac{\ell' - \ell + 1}{\alpha}(\tau_\mu + \tau^*)}. \tag{48.34}$$

In the above estimates, the constant c depends on $\|a_\sharp\| \ell_D^2$ and on the operator norms defined in (48.13) and (48.18).

Proof. See Exercise 48.4. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

Remark 48.13 (Literature). The nonconforming approximation of the elliptic eigenvalue problem has been studied in Antonietti et al. [12] for discontinuous Galerkin (dG) methods, Gopalakrishnan et al. [220] for hybridizable discontinuous Galerkin (HDG) methods, and Calo et al. [103] for hybrid high-order (HHO) methods. We refer the reader to Canuto [105], Mercier et al. [301], Durán et al. [182], Boffi et al. [63] for mixed and hybrid mixed methods and to Carstensen and Gedicke [109], Liu [287] for guaranteed eigenvalue lower bounds using Crouzeix–Raviart elements. $\qquad$ $\square$

Exercises

Exercise 48.1 (Linearity). Consider the setting of §48.1.2. Let $V \hookrightarrow L$ be two complex Banach spaces and $a : V \times V \to \mathbb{C}$ be a bounded sesquilinear form satisfying the two conditions of the

BNB theorem. Let $b : L \times L \to \mathbb{C}$ be bounded sesquilinear form. (i) Let $T : L \to L$ be such that $a(T(v), w) := b(v, w)$ for all $v \in L$ and all $w \in V$. Show that T is well defined and linear. (ii) Let $T_* : L \to L$ be such that $a(v, T_*(w)) := b(v, w)$ for all $v \in V$ and all $w \in L$. Show that T_* is well defined and linear.

Exercise 48.2 (Invariant sets). (i) Let $S, T \in \mathcal{L}(V)$ be such that $ST = TS$. Prove that $\ker(S)$ and $\mathrm{im}(S)$ are invariant under T. (ii) Let $T \in \mathcal{L}(V)$ and let $W_1, \dots, W_m$ be subspaces of V that are invariant under T. Prove that $W_1 + \dots + W_m$ and $\bigcap_{i \in \{1:m\}} W_i$ are invariant under T. (iii) Let $T \in \mathcal{L}(V)$ and let $\{v_1, \dots, v_n\}$ be a basis of V. Show that the following statements are equivalent: (a) The matrix of T with respect to $\{v_1, \dots, v_n\}$ is upper triangular; (b) $T(v_j) \in \mathrm{span}\{v_1, \dots, v_j\}$ for all $j \in \{1:n\}$; (c) $\mathrm{span}\{v_1, \dots, v_j\}$ is invariant under T for all $j \in \{1:n\}$. (iv) Let $T \in \mathcal{L}(V)$. Let μ be an eigenvalue of T. Prove that $\mathrm{im}(\mu I_V - T)$ is invariant under T. Prove that $\ker(\mu I_V - T)^\alpha$ is invariant under T for every integer $\alpha \geq 1$.

Exercise 48.3 (Trace). (i) Let V be a complex Banach space. Let $G \subset V$ be a subspace of V of dimension m. Let $\{\phi_j\}_{j \in \{1:m\}}$ and $\{\psi_j\}_{j \in \{1:m\}}$ be two bases of G, and let $\{\phi'_j\}_{j \in \{1:m\}}$ and $\{\psi'_j\}_{j \in \{1:m\}}$ be corresponding dual bases, i.e., $\langle \phi'_i, \phi_j \rangle_{V',V} = \delta_{ij}$, etc. (the way the antilinear forms $\{\phi'_j\}_{j \in \{1:m\}}$ and $\{\psi'_j\}_{j \in \{1:m\}}$ are extended to V does not matter). Let $T \in \mathcal{L}(V)$ and assume that G is invariant under T. Show that $\sum_{j \in \{1:m\}} \langle \psi'_j, T(\psi_j) \rangle_{V',V} = \sum_{j \in \{1:m\}} \langle \phi'_j, T(\phi_j) \rangle_{V',V}$. (ii) Let $B \in \mathbb{C}^{m \times m}$ be s.t. $T(\phi_i) =: \sum_{j \in \{1:m\}} B_{ji} \phi_j$ (recall that G is invariant under T). Let $\mathsf{V} := (\langle \phi'_j, v \rangle_{V',V})_{j \in \{1:m\}}^{\mathsf{T}}$ for all $v \in G$. Prove that $T^\alpha(v) = \sum_{j \in \{1:m\}} (B^\alpha \mathsf{V})_j \phi_j$ for all $\alpha \in \mathbb{N}$. (*Hint*: use an induction argument.) (iii) Let $\mu \in \mathbb{C}$, $\alpha \geq 1$, and $S \in \mathcal{L}(V)$. Assume that $G := \ker(\mu I_V - S)^\alpha$ is finite-dimensional and nontrivial (i.e., $\dim(G) := m \geq 1$). Prove that $\sum_{j \in \{1:m\}} \langle \phi'_j, S(\phi_j) \rangle_{V,V'} = m\mu$. (*Hint*: consider the $m \times m$ matrix A with entries $\langle \phi'_i, (\mu I_V - S)(\phi_j) \rangle_{V',V}$ and show that $A^\alpha = 0$.)

Exercise 48.4 (Theorem 48.12). Prove the estimates in Theorem 48.12. (*Hint*: see the proof of Theorem 48.8.)

Exercise 48.5 (Nonconforming approximation). Consider the Laplace operator with homogeneous Dirichlet boundary conditions in a Lipschitz polyhedron D with $b(v, w) := \int_D \rho v w \, \mathrm{d}x$, where $\rho \in C^\infty(D; \mathbb{R})$. Verify that the assumptions (48.25) to (48.30) hold true for the Crouzeix–Raviart approximation.

Chapter 49

Well-posedness for PDEs in mixed form

In Part XI, composed of Chapters 49 to 55, we study the well-posedness and the finite element approximation of PDEs formulated in mixed form. Mixed formulations are often obtained from elliptic PDEs by introducing one or more auxiliary variables. One reason for introducing these variables can be that they have some physical relevance. For instance, one can think of the flux in Darcy's equations (see Chapter 51). Another reason can be to relax a constraint imposed on a variational formulation. This is the case for the Stokes equations where the pressure results from the incompressibility constraint enforced on the velocity field (see Chapter 53). The PDEs considered in this part enjoy a coercivity property on the primal variable, but not on the auxiliary variable, so that the analysis relies on inf-sup conditions. The goal of the present chapter is to identify necessary and sufficient conditions for the well-posedness of a model problem that serves as a prototype for PDEs in mixed form. The finite element approximation of this model problem is investigated in Chapter 50.

49.1 Model problems

We introduce in this section some model problems illustrating the concept of PDEs in mixed form. Let D be a domain in $\mathbb{R}^d$, i.e., D is a nonempty, open, bounded, connected subset of $\mathbb{R}^d$ (see Definition 3.1).

49.1.1 Darcy

Consider the elliptic PDE $-\nabla\cdot(\mathbb{d}\nabla p) = f$ in D; see §31.1. Introducing the flux (or filtration velocity) $\boldsymbol{\sigma} := -\mathbb{d}\nabla p$ leads to the following mixed formulation known as Darcy's equations (see §24.1.2):

$$\mathbb{d}^{-1}\boldsymbol{\sigma} + \nabla p = \mathbf{0} \qquad \text{in } D, \tag{49.1a}$$

$$\nabla\cdot\boldsymbol{\sigma} = f \qquad \text{in } D. \tag{49.1b}$$

Here, (49.1a) is a phenomenological law relating the flux to the gradient of the primal unknown p (a nonzero right-hand side can be considered as well). The equation (49.1b) expresses mass

conservation. For simplicity, we assume that (49.1) is equipped with the boundary condition $p_{|\partial D} = 0$.

Let us give a more abstract form to the above problem by setting

$$\boldsymbol{V} := \boldsymbol{L}^2(D), \qquad Q := H_0^1(D). \tag{49.2}$$

Consider the linear operators $A : \boldsymbol{V} \to \boldsymbol{V}' = \boldsymbol{L}^2(D)$ (owing to the Riesz–Fréchet representation theorem) and $B : \boldsymbol{V} \to Q' = H^{-1}(D)$ defined by setting $A(\boldsymbol{\tau}) := \mathbb{d}^{-1}\boldsymbol{\tau}$ and $B(\boldsymbol{\tau}) := -\nabla{\cdot}\boldsymbol{\tau}$. Under appropriate boundedness assumptions on $\mathbb{d}^{-1}$, the linear operators A and B are bounded. Using the identification $(H_0^1(D))'' = H_0^1(D)$, we have $B^* : Q \to \boldsymbol{V}'$ and $\langle B^*(q), \boldsymbol{\tau} \rangle_{\boldsymbol{V}',\boldsymbol{V}} = \langle q, B(\boldsymbol{\tau}) \rangle_{H_0^1(D),H^{-1}(D)} = (\nabla q, \boldsymbol{\tau})_{\boldsymbol{L}^2(D)}$ for all $q \in H_0^1(D)$ and all $\boldsymbol{\tau} \in \boldsymbol{L}^2(D)$. Hence, $B^*(q) = \nabla q$ for all $q \in H_0^1(D)$.

An alternative point of view consists of setting

$$\boldsymbol{V} := \boldsymbol{H}(\mathrm{div}; D), \qquad Q := L^2(D). \tag{49.3}$$

Then $A : \boldsymbol{V} \to \boldsymbol{V}'$ is defined by setting $A(\boldsymbol{\tau}) := \mathbb{d}^{-1}\boldsymbol{\tau}$ (where we use that $\boldsymbol{V} \hookrightarrow \boldsymbol{L}^2(D) \equiv \boldsymbol{L}^2(D)' \hookrightarrow \boldsymbol{V}'$) and $B : \boldsymbol{V} \to Q' = L^2(D)$ (owing to the Riesz–Fréchet representation theorem) is defined by setting $B(\boldsymbol{\tau}) := -\nabla{\cdot}\boldsymbol{\tau}$. The adjoint operator $B^* : Q \to \boldsymbol{V}'$ is s.t. $\langle B^*(q), \boldsymbol{\tau} \rangle_{\boldsymbol{V}',\boldsymbol{V}} = (q, B(\boldsymbol{\tau}))_{L^2(D)} = (q, -\nabla{\cdot}\boldsymbol{\tau})_{L^2(D)}$ for all $q \in L^2(D)$ and all $\boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div}; D)$. Let us have a closer look at B^*. Let $q \in L^2(D)$ and assume that there exists $\boldsymbol{g} \in \boldsymbol{L}^2(D)$ so that $\langle B^*(q), \boldsymbol{\tau} \rangle_{\boldsymbol{V}',\boldsymbol{V}} = (\boldsymbol{g}, \boldsymbol{\tau})_{\boldsymbol{L}^2(D)}$. This implies that $(q, -\nabla{\cdot}\boldsymbol{\tau})_{L^2(D)} = (\boldsymbol{g}, \boldsymbol{\tau})_{\boldsymbol{L}^2(D)}$ for all $\boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div}; D)$. Taking $\boldsymbol{\tau} \in \boldsymbol{C}_0^\infty(D)$ shows that q has a weak derivative and $\nabla q = \boldsymbol{g}$. Hence, $q \in H^1(D)$ and $(q, \nabla{\cdot}\boldsymbol{\tau})_{L^2(D)} + (\nabla q, \boldsymbol{\tau})_{\boldsymbol{L}^2(D)} = 0$ for all $\boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div}; D)$. Moreover, considering the trace of q on ∂D, $\gamma^{\mathrm{g}}(q) \in H^{\frac{1}{2}}(\partial D)$, and using the surjectivity of the normal trace operator $\gamma^{\mathrm{d}} : \boldsymbol{H}(\mathrm{div}; D) \to H^{-\frac{1}{2}}(\partial D)$ (see Theorem 4.15), we infer that for all $\phi \in H^{-\frac{1}{2}}(\partial D)$, there is $\boldsymbol{\tau}_\phi \in \boldsymbol{H}(\mathrm{div}; D)$ s.t. $\gamma^{\mathrm{d}}(\boldsymbol{\tau}_\phi) = \phi$. Then $\langle \phi, \gamma^{\mathrm{g}}(q) \rangle_{H^{-\frac{1}{2}}, H^{\frac{1}{2}}} = \langle \gamma^{\mathrm{d}}(\boldsymbol{\tau}_\phi), \gamma^{\mathrm{g}}(q) \rangle_{H^{-\frac{1}{2}}, H^{\frac{1}{2}}} = (q, \nabla{\cdot}\boldsymbol{\tau}_\phi)_{L^2(D)} + (\nabla q, \boldsymbol{\tau}_\phi)_{\boldsymbol{L}^2(D)} = 0$ for all $\phi \in H^{-\frac{1}{2}}(\partial D)$. Hence, $\gamma^{\mathrm{g}}(q) = 0$, i.e., $q_{|\partial D} = 0$. This shows that $B^*(q) \in \boldsymbol{L}^2(D)$ encodes the boundary condition $q_{|\partial D} = 0$ in a weak sense.

In conclusion, regardless of the chosen setting, the Darcy problem (49.1) can be reformulated as follows:

$$\begin{cases} \text{Find } \boldsymbol{\sigma} \in \boldsymbol{V} \text{ and } p \in Q \text{ such that} \\ A(\boldsymbol{\sigma}) + B^*(p) = \boldsymbol{0}, \\ B(\boldsymbol{\sigma}) = -f. \end{cases} \tag{49.4}$$

The mixed finite element approximation of (49.4) is studied in Chapter 51.

49.1.2 Stokes

The Stokes equations model steady incompressible flows in which inertia forces are negligible. The problem is written in the following mixed form:

$$\nabla{\cdot}(-\mu\mathbb{e}(\boldsymbol{u})) + \nabla p = \boldsymbol{f} \qquad \text{in } D, \tag{49.5a}$$

$$\nabla{\cdot}\boldsymbol{u} = 0 \qquad \text{in } D, \tag{49.5b}$$

where $\mu > 0$ is the viscosity, $\boldsymbol{u} : D \to \mathbb{R}^d$ the velocity field with the (linearized) strain rate tensor $\mathbb{e}(\boldsymbol{u}) := \frac{1}{2}(\nabla\boldsymbol{u} + (\nabla\boldsymbol{u})^\mathsf{T}) : D \to \mathbb{R}^{d\times d}$, $p : D \to \mathbb{R}$ the pressure, and $\boldsymbol{f} : D \to \mathbb{R}^d$ the body force. The equation (49.5a) expresses the momentum balance in the flow, and (49.5b) the mass balance. For simplicity, we assume that (49.5) is equipped with the boundary condition $\boldsymbol{u}_{|\partial D} = \boldsymbol{0}$.

Let us set

$$\boldsymbol{V} := \boldsymbol{H}_0^1(D), \qquad Q := L^2(D), \tag{49.6}$$

and let us define $A : \boldsymbol{V} \to \boldsymbol{V}' = \boldsymbol{H}^{-1}(D)$, $B : \boldsymbol{V} \to Q' = L^2(D)$ (owing to the Riesz–Fréchet representation theorem) by setting $A(\boldsymbol{v}) := -\nabla{\cdot}(\mu\mathbb{e}(\boldsymbol{v}))$ and $B(\boldsymbol{v}) := -\nabla{\cdot}\boldsymbol{v}$. The adjoint operator $B^* : Q \to \boldsymbol{V}'$ is s.t. $\langle B^*(q), \boldsymbol{v}\rangle_{\boldsymbol{V}',\boldsymbol{V}} = (q, -\nabla{\cdot}\boldsymbol{v})_{L^2(D)}$ for all $\boldsymbol{v} \in \boldsymbol{V}$ and all $q \in Q$. This means that $B^*(q) = \nabla q$ for all $q \in L^2(D)$. In conclusion, the Stokes problem (49.5) can be reformulated as follows:

$$\begin{cases} \text{Find } \boldsymbol{v} \in V \text{ and } p \in Q \text{ such that} \\ A(\boldsymbol{v}) + B^*(p) = \boldsymbol{f}, \\ B(\boldsymbol{v}) = 0. \end{cases} \tag{49.7}$$

The finite element approximation of (49.7) is studied in Chapters 53 to 55.

49.1.3 Maxwell

Consider the model problem (43.4) for Maxwell's equations in the time-harmonic regime (see §43.1.1) in the limit $\omega \to 0$ and with the boundary condition $\boldsymbol{H}_{|\partial D}{\times}\boldsymbol{n} = \boldsymbol{0}$. Ampère's equation (43.4a) gives $-\boldsymbol{E} + \frac{1}{\sigma}\nabla{\times}\boldsymbol{H} = \frac{1}{\sigma}\boldsymbol{j}_{\mathrm{s}}$, and Faraday's equation (43.4b) gives $\nabla{\cdot}(\mu\boldsymbol{H}) = 0$ and $\nabla{\times}\boldsymbol{E} = \boldsymbol{0}$ (since $\omega \to 0$). Setting $\kappa := \sigma^{-1}$ the strong form of this problem consists of looking for a field $\boldsymbol{H} : D \to \mathbb{R}^3$ such that $\nabla{\times}(\kappa\nabla{\times}\boldsymbol{H}) = \boldsymbol{f}$, with $\boldsymbol{f} := \nabla{\times}(\kappa\boldsymbol{j}_{\mathrm{s}}) : D \to \mathbb{R}^3$, $\boldsymbol{H}_{|\partial D}{\times}\boldsymbol{n} = \boldsymbol{0}$, and under the constraint $\nabla{\cdot}(\mu\boldsymbol{H}) = \boldsymbol{0}$. The dual variable of this constraint is a scalar-valued function $\phi : D \to \mathbb{R}$ with the boundary condition $\phi_{|\partial D} = 0$, leading to the following mixed formulation (see, e.g., Kanayama et al. [263]):

$$\nabla{\times}(\kappa\nabla{\times}\boldsymbol{H}) + \nu\nabla\phi = \boldsymbol{f} \qquad \text{in } D, \tag{49.8a}$$
$$\nabla{\cdot}(\mu\boldsymbol{H}) = 0 \qquad \text{in } D. \tag{49.8b}$$

Let us set

$$\boldsymbol{V} := \boldsymbol{H}_0(\mathrm{curl}; D), \qquad Q := H_0^1(D), \tag{49.9}$$

and let us define $A : \boldsymbol{V} \to \boldsymbol{V}'$, $B : \boldsymbol{V} \to Q' = H^{-1}(D)$ by setting $A(\boldsymbol{v}) := \nabla{\times}(\kappa\nabla{\times}\boldsymbol{v})$ and $B(\boldsymbol{v}) := -\nabla{\cdot}(\nu\boldsymbol{v})$. Using the identification $(H_0^1(D))'' = H_0^1(D)$, the adjoint operator $B^* : Q \to \boldsymbol{V}'$ is s.t. $\langle B^*(\psi), \boldsymbol{v}\rangle_{\boldsymbol{V}',\boldsymbol{V}} = \langle \psi, -\nabla{\cdot}(\nu\boldsymbol{v})\rangle_{H_0^1(D),H^{-1}(D)} = (\nu\nabla\psi, \boldsymbol{v})_{\boldsymbol{L}^2(D)}$ for all $\boldsymbol{v} \in \boldsymbol{V}$ and all $\psi \in Q$. This means that $B^*(\psi) = \nu\nabla\psi$ for all $\psi \in H_0^1(D)$. In conclusion, the Maxwell problem (49.8) can be reformulated as follows:

$$\begin{cases} \text{Find } \boldsymbol{H} \in \boldsymbol{V} \text{ and } \phi \in Q \text{ such that} \\ A(\boldsymbol{H}) + B^*(\phi) = \boldsymbol{f}, \\ B(\boldsymbol{H}) = 0. \end{cases} \tag{49.10}$$

Some further aspects of this problem are considered in Exercise 49.6.

49.2 Well-posedness in Hilbert spaces

Consider two real Hilbert spaces V and Q, and two operators $A \in \mathcal{L}(V; V')$, $B \in \mathcal{L}(V; Q')$. We identify $Q'' = Q$. Our goal in this section is to investigate the well-posedness of the following

mixed model problem:

$$
\begin{cases}
\text{Find } u \in V \text{ and } p \in Q \text{ such that} \\
A(u) + B^*(p) = f, \\
B(u) = g,
\end{cases}
\tag{49.11}
$$

for all $(f, g) \in V' \times Q'$. We assume in the entire section that A is self-adjoint and coercive. This assumption simplifies many arguments. In particular, we establish the well-posedness of (49.11) by means of a coercivity argument on the Schur complement. The complete theory for well-posedness in Banach spaces is done in §49.4. Let α be the coercivity constant of A,

$$
\inf_{v \in V} \frac{\langle A(v), v \rangle_{V', V}}{\|v\|_V^2} =: \alpha > 0.
\tag{49.12}
$$

We also assume that B is surjective, i.e., recalling Lemma C.40,

$$
\inf_{q \in Q} \frac{\|B^*(q)\|_{V'}}{\|q\|_Q} =: \beta > 0.
\tag{49.13}
$$

We denote $\|a\| := \|A\|_{\mathcal{L}(V; V')}$ and $\|b\| := \|B\|_{\mathcal{L}(V'; Q)}$.

49.2.1 Schur complement

Let $J_Q : Q \to Q'$ be the Riesz–Fréchet isometric isomorphism (see Theorem C.24), i.e.,

$$
\langle J_Q(q), r \rangle_{Q', Q} := (q, r)_Q, \qquad \forall q, r \in Q.
$$

We call Schur complement of A on Q the linear operator $S : Q \to Q$ defined by

$$
S := J_Q^{-1} B A^{-1} B^*.
\tag{49.14}
$$

(S is sometimes defined with the opposite sign in the literature.)

Lemma 49.1 (Coercivity and boundedness of S). *Let S be defined in* (49.14). *Then S is symmetric and bijective with*

$$
\frac{\beta^2}{\|a\|} \|q\|_Q^2 \leq (S(q), q)_Q \leq \frac{\|b\|^2}{\alpha} \|q\|_Q^2, \qquad \forall q \in Q.
\tag{49.15}
$$

Proof. (1) Symmetry. Since A^{-1} is self-adjoint, we infer that for all $q, r \in Q$,

$$
(S(q), r)_Q = \langle B A^{-1} B^*(q), r \rangle_{Q', Q} = \langle A^{-1} B^*(q), B^*(r) \rangle_{V, V'}
$$
$$
= \langle A^{-1} B^*(r), B^*(q) \rangle_{V, V'} = (S(r), q)_Q.
$$

(2) Bounds (49.15). The self-adjointness and coercivity of A imply that $\|A^{-1}\|_{\mathcal{L}(V'; V)} = \alpha^{-1}$ (see Lemma C.51) and $\langle A^{-1}(\phi), \phi \rangle_{V, V'} \geq \frac{1}{\|a\|} \|\phi\|_{V'}^2$ for all $\phi \in V'$ (see Lemma C.63). Moreover, the definitions of $\|b\|$ and β mean that $\|B^*\|_{\mathcal{L}(Q; V')} = \|b\|$ and $\|B^*(q)\|_{V'} \geq \beta \|q\|_Q$ for all $q \in Q$. Since $(S(q), q)_Q = \langle A^{-1}(B^*(q)), B^*(q) \rangle_{V, V'}$ for all $q \in Q$, we conclude that (49.15) holds true. Finally, S is bijective since S is Q-coercive and bounded. $\qquad\square$

Lemma 49.2 (Equivalence with (49.11)**).** *Let $(u, p) \in V \times Q$. Then the pair (u, p) solves* (49.11) *iff (u, p) solves*

$$
S(p) = J_Q^{-1}(B A^{-1}(f) - g), \qquad A(u) = f - B^*(p).
\tag{49.16}
$$

Proof. Let $(u,p) \in V \times Q$ solve (49.11). Since A is bijective, we have $u = A^{-1}(f - B^*(p))$, so that $B(u) = BA^{-1}(f - B^*(p)) = g$. This in turn implies that $J_Q^{-1}BA^{-1}(f - B^*(p)) = J_Q^{-1}(g)$, finally giving $S(p) = J_Q^{-1}(BA^{-1}(f) - g)$ and $A(u) = f - B^*(p)$. This means that (u,p) solves (49.16). Conversely, assume that $(u,p) \in V \times Q$ solves (49.16). Then $BA^{-1}B^*(p) = BA^{-1}(f) - g$, that is, $BA^{-1}(f - B^*(p)) = g$. But $A^{-1}(f - B^*(p)) = u$. Hence, $B(u) = g$ and $A(u) = f - B^*(p)$, which means that (u,p) solves (49.11). □

Theorem 49.3 (Well-posedness). *The problem* (49.11) *is well-posed.*

Proof. Owing to Lemma 49.2, it suffices to show that (49.16) is well-posed, but this is a consequence of Lemma 49.1 and the Lax–Milgram lemma. □

49.2.2 Formulation with bilinear forms

We now reformulate the mixed problem (49.11) using bilinear forms. This formalism will be used in Chapters 50 to 55 where we consider various Galerkin approximations to (49.11). Let us set

$$a(v,w) := \langle A(v), w \rangle_{V',V}, \qquad b(w,q) := \langle q, B(w) \rangle_{Q,Q'}, \tag{49.17}$$

for all $v, w \in V$ and all $q \in Q$ (recall that we have identified Q'' and Q). The assumed boundedness of A and B implies that a and b are bounded on $V \times V$ and on $V \times Q$, respectively. The abstract problem (49.11) can then be reformulated in the following equivalent form:

$$\begin{cases} \text{Find } u \in V \text{ and } p \in Q \text{ such that} \\ a(u,w) + b(w,p) = f(w), & \forall w \in V, \\ b(u,q) = g(q), & \forall q \in Q, \end{cases} \tag{49.18}$$

with the shorthand notation $f(w) := \langle f, w \rangle_{V',V}$ and $g(q) := \langle g, q \rangle_{Q',Q}$. The definitions (49.12) and (49.13) of α and β are then equivalent to

$$\inf_{v \in V} \frac{|a(v,v)|}{\|v\|_V^2} =: \alpha > 0, \qquad \inf_{q \in Q} \sup_{v \in V} \frac{|b(v,q)|}{\|v\|_V \|q\|_Q} =: \beta > 0. \tag{49.19}$$

Let $X := V \times Q$ and consider the bilinear form $t : X \times X \to \mathbb{R}$ defined by

$$t((v,q),(w,r)) := a(v,w) + b(w,q) + b(v,r), \tag{49.20}$$

for all $(v,q),(w,r) \in X$. Then $(u,p) \in X$ solves (49.11) iff

$$t((u,p),(w,r)) = f(w) + g(r), \qquad \forall (w,r) \in X. \tag{49.21}$$

49.2.3 Sharper a priori estimates

We collect in this section some additional results regarding the operator S, and we give a priori estimates on the solution to the mixed problem (49.11). Recall from Definition 46.1 the notions of spectrum and eigenvalues.

Corollary 49.4 (Spectrum of S). *The spectrum of S is such that* $\sigma(S) \subset \left[\frac{\beta^2}{\|a\|}, \frac{\|b\|^2}{\alpha} \right]$, *and* $\|S\|_{\mathcal{L}(Q;Q)} \le \frac{\|b\|^2}{\alpha}$, $\|S^{-1}\|_{\mathcal{L}(Q;Q)} \le \frac{\|a\|}{\beta^2}$.

Proof. These statements are consequences of (49.15) and Theorem 46.17. Recall that Theorem 46.17 asserts in particular that $\sigma(S) \subset \mathbb{R}$, $\|S\|_{\mathcal{L}(Q;Q)} = \sup_{\lambda \in \sigma(S)} |\lambda|$ and $\|S^{-1}\|_{\mathcal{L}(Q;Q)} = \sup_{\lambda \in \sigma(S)} |\lambda|^{-1}$. See also Exercise 49.3. □

Remark 49.5 (Spectrum of S). Corollary 49.4 can be refined by equipping V with the energy norm $\|\cdot\|_a^2 := a(\cdot,\cdot)^{\frac{1}{2}}$ (recall that a is symmetric and coercive) which is equivalent to $\|\cdot\|_V$. Setting $\beta_a := \inf_{q\in Q} \sup_{v\in V} \frac{|b(v,q)|}{\|v\|_a\|q\|_Q}$ and $\|b\|_a := \sup_{q\in Q} \sup_{v\in V} \frac{|b(v,q)|}{\|v\|_a\|q\|_Q}$, we have $\|b\|_a^2 = \|S\|_{\mathcal{L}(Q;Q)}$, $\beta_a^{-2} = \|S^{-1}\|_{\mathcal{L}(Q;Q)}$, and $\{\beta_a^2, \|b\|_a^2\} \subset \sigma(S) \subset [\beta_a^2, \|b\|_a^2]$. $\square$

We define the linear operator $T : X \to X$ such that

$$T(v,q) := (v + A^{-1}B^*(q), S^{-1}J_Q^{-1}B(v)), \tag{49.22}$$

for all $(v,q) \in V \times Q$. With a slight abuse of notation regarding the column vector convention, we can also write $T := \begin{pmatrix} I_V & A^{-1}B^* \\ S^{-1}J_Q^{-1}B & 0 \end{pmatrix}$, where I_V is the identity in V. We have $T \in \mathcal{L}(X;X)$, and upon introducing the weighted inner product $(x,y)_{\widetilde{X}} := a(v,w) + (S(q),r)_Q$ for all $x := (v,q)$ and $y := (w,r)$ in X, we also have $(T(x),y)_{\widetilde{X}} = a(v,w) + b(v,r) + b(w,q)$, that is, $(T(x),y)_{\widetilde{X}} = t(x,y)$. This identity implies that T is symmetric with respect to the weighted inner product $(x,y)_{\widetilde{X}}$. The following result, due to Bacuta [42], provides a complete characterization of the spectrum of T. We refer the reader to §50.3.2 for the algebraic counterpart of this result.

Theorem 49.6 (Spectrum of T). Let $\varrho := \frac{1+\sqrt{5}}{2}$ be the golden ratio. Assume that $\ker(B)$ is nontrivial. Then

$$\sigma(T) = \sigma_{\mathrm{p}}(T) = \{-\varrho^{-1}, 1, \varrho\}. \tag{49.23}$$

Proof. Let $\lambda \in \sigma(T)$. Owing to Corollary 46.18, there is a sequence $(x_n)_{n\in\mathbb{N}}$ in X such that $\|x_n\|_X = 1$ for all $n \in \mathbb{N}$ and $T(x_n) - \lambda x_n \to 0$ in X as $n \to \infty$. Writing $x_n := (v_n, q_n)$, we infer that $(1-\lambda)v_n + A^{-1}B^*(q_n) \to 0$ in V and $S^{-1}J_Q^{-1}B(v_n) - \lambda q_n \to 0$ in Q as $n \to \infty$. Applying the bounded operator $J_Q^{-1}B$ to the first limit and the bounded operator S to the second one, we infer that

$$\begin{pmatrix} 1-\lambda & 1 \\ 1 & -\lambda \end{pmatrix} \begin{pmatrix} J_Q^{-1}B(v_n) \\ S(q_n) \end{pmatrix} \to 0.$$

Assume that $\lambda \notin \{-\varrho^{-1}, 1, \varrho\}$. The matrix on the left-hand side is invertible since $\lambda \notin \{-\varrho^{-1}, \varrho\}$. This implies that $S(q_n) \to 0$ in Q, so that $\|q_n\|_Q \to 0$ since S is a bounded bijective operator. Since $\lambda \neq 1$ and recalling that $(1-\lambda)v_n + A^{-1}B^*(q_n) \to 0$ in V, we conclude that $\|v_n\|_V \to 0$, providing the expected contradiction with $\|x_n\|_X = 1$. Hence, $\sigma(T) \subset \{-\varrho^{-1}, 1, \varrho\}$. Finally, we observe that $\lambda = 1$ is an eigenvalue associated with the eigenvectors $(v,0)^{\mathsf{T}}$ for all $v \in \ker(B) \setminus \{0\}$, and $\pm\varrho^{\pm 1}$ is an eigenvalue associated with the eigenvectors $(\pm\varrho^{\pm 1}A^{-1}B^*(q), q)^{\mathsf{T}}$ for all $q \in Q \setminus \{0\}$. This proves (49.23). $\square$

Theorem 49.6 allows us to derive sharp stability estimates for the solution of (49.18) in the weighted norm $\|(v,q)\|_{\widetilde{X}} := (a(v,v) + (S(q),q)_Q)^{\frac{1}{2}}$ induced by the weighted inner product for which T is symmetric. Equipping X with this weighted norm and since $\|T\|_{\mathcal{L}(X;X)} = \sup_{\lambda\in\sigma(T)} |\lambda|$ and $\|T^{-1}\|_{\mathcal{L}(X;X)} = \sup_{\lambda\in\sigma(T)} |\lambda|^{-1}$ (owing to Theorem 46.17), we infer from Theorem 49.6 that

$$\|T\|_{\mathcal{L}(X;X)} = \|T^{-1}\|_{\mathcal{L}(X;X)} = \varrho. \tag{49.24}$$

Recalling that $t((v,q),(w,r)) = (T(v,q),(w,r))_{\widetilde{X}}$, we infer that

$$\inf_{x\in X} \sup_{y\in Y} \frac{|t(x,y)|}{\|x\|_{\widetilde{X}}\|y\|_{\widetilde{X}}} = \|T^{-1}\|_{\mathcal{L}(X;X)}^{-1} = \varrho^{-1}, \tag{49.25a}$$

$$\sup_{x\in X} \sup_{y\in Y} \frac{|t(x,y)|}{\|x\|_{\widetilde{X}}\|y\|_{\widetilde{X}}} = \|T\|_{\mathcal{L}(X;X)} = \varrho. \tag{49.25b}$$

Corollary 49.7 (Stability). *Let $(u, p) \in X$ solve* (49.18). *The following holds true:*

$$\varrho^{-1} \left(\frac{1}{\|a\|} \|f\|_{V'}^2 + \frac{\alpha}{\|b\|^2} \|g\|_{Q'}^2 \right)^{\frac{1}{2}} \leq \|(u, p)\|_{\widetilde{X}} \leq \varrho \left(\frac{1}{\alpha} \|f\|_{V'}^2 + \frac{\|a\|}{\beta^2} \|g\|_{Q'}^2 \right)^{\frac{1}{2}}.$$

Proof. Let us set $x := (u, p)$. Then (49.18) amounts to $T(x) = y$ with $y := (A^{-1}(f), S^{-1} J_Q^{-1}(g))$. Hence, we have

$$\|T\|_{\mathcal{L}(X;X)}^{-1} \|y\|_{\widetilde{X}} \leq \|x\|_{\widetilde{X}} \leq \|T^{-1}\|_{\mathcal{L}(X;X)} \|y\|_{\widetilde{X}},$$

and we use (49.24) to infer that $\varrho^{-1} \|y\|_{\widetilde{X}} \leq \|x\|_{\widetilde{X}} \leq \varrho \|y\|_{\widetilde{X}}$. Finally, the bounds in the proof of Lemma 49.1 imply that $\|y\|_{\widetilde{X}}^2 \geq \frac{1}{\|a\|} \|f\|_{V'}^2 + \frac{\alpha}{\|b\|^2} \|g\|_{Q'}^2$ and $\|y\|_{\widetilde{X}}^2 \leq \frac{1}{\alpha} \|f\|_{V'}^2 + \frac{\|a\|}{\beta^2} \|g\|_{Q'}^2$. $\qquad\square$

Proposition 49.8 (Stability). *Let $(u, p) \in X$ solve* (49.18). *The following holds true:*

$$\frac{4}{((4\frac{\|b\|^2}{\alpha} + 1)^{\frac{1}{2}} + 1)^2} \left(\frac{1}{\|a\|} \|f\|_{V'}^2 + \|g\|_{Q'}^2 \right) \leq a(u, u) + \|p\|_Q^2$$

$$\leq \frac{4}{((4\frac{\beta^2}{\|a\|} + 1)^{\frac{1}{2}} - 1)^2} \left(\frac{1}{\alpha} \|f\|_{V'}^2 + \|g\|_{Q'}^2 \right). \quad (49.26)$$

Proof. The proof is similar to that of Corollary 49.7 but uses the operator $\widetilde{T} \in \mathcal{L}(X; X)$ s.t. $\widetilde{T}(v, q) := (v + A^{-1}B^*(q), J_Q^{-1}B(v))$ for all $(v, q) \in V \times Q$; see [42] and Exercise 49.4. $\qquad\square$

49.3 Saddle point problems in Hilbert spaces

In this section, we assume again that V and Q are real Hilbert spaces and the bilinear form a is symmetric and coercive, i.e., A is self-adjoint and coercive. We show that the mixed problem (49.18) has a saddle point structure.

49.3.1 Finite-dimensional constrained minimization

We start by introducing some simple ideas in the finite-dimensional setting of linear algebra. Let N, M be two positive integers, let $\mathcal{A}$ be a symmetric positive definite matrix in $\mathbb{R}^{N \times N}$ and let $\mathsf{F} \in \mathbb{R}^N$. Consider the functional $\mathfrak{E} : \mathbb{R}^N \to \mathbb{R}$ such that $\mathfrak{E}(\mathsf{V}) := \frac{1}{2}(\mathcal{A}\mathsf{V}, \mathsf{V})_{\ell^2(\mathbb{R}^N)} - (\mathsf{F}, \mathsf{V})_{\ell^2(\mathbb{R}^N)}$. Then $\mathfrak{E}$ admits a unique global minimizer over $\mathbb{R}^N$, say U, which is characterized by the Euler condition $D\mathfrak{E}(\mathsf{U})(\mathsf{W}) = (\mathcal{A}\mathsf{U} - \mathsf{F}, \mathsf{W})_{\ell^2(\mathbb{R}^N)} = 0$ for all $\mathsf{W} \in \mathbb{R}^N$, i.e., $\mathcal{A}\mathsf{U} = \mathsf{F}$ (see Proposition 25.8 and Remark 26.5 for similar results expressed in terms of bilinear forms).

Let now $\mathcal{B} \in \mathbb{R}^{M \times N}$, let $\mathsf{G} \in \mathrm{im}(\mathcal{B}) \subset \mathbb{R}^M$, and consider the affine subspace $K := \{\mathsf{V} \in \mathbb{R}^N \mid \mathcal{B}\mathsf{V} = \mathsf{G}\}$. Then $\mathfrak{E}$ admits a unique global minimizer over K, say U, which is characterized by the Euler condition $D\mathfrak{E}(\mathsf{U})(\mathsf{W}) = (\mathcal{A}\mathsf{U} - \mathsf{F}, \mathsf{W})_{\ell^2(\mathbb{R}^N)} = 0$ for all $\mathsf{W} \in \ker(\mathcal{B})$, i.e., $\mathcal{A}\mathsf{U} - \mathsf{F} \in \ker(\mathcal{B})^\perp$. Since $\ker(\mathcal{B})^\perp = \mathrm{im}(\mathcal{B}^\mathsf{T})$, we infer that there is $\mathsf{P} \in \mathbb{R}^M$ such that $\mathcal{A}\mathsf{U} + \mathcal{B}^\mathsf{T}\mathsf{P} = \mathsf{F}$. Recalling that $\mathcal{B}\mathsf{U} = \mathsf{G}$, the optimality condition is equivalent to solving the system

$$\begin{pmatrix} \mathcal{A} & \mathcal{B}^\mathsf{T} \\ \mathcal{B} & \mathbb{O} \end{pmatrix} \begin{pmatrix} \mathsf{U} \\ \mathsf{P} \end{pmatrix} = \begin{pmatrix} \mathsf{F} \\ \mathsf{G} \end{pmatrix}, \quad (49.27)$$

where $\mathbb{O}$ is the zero matrix in $\mathbb{R}^{M \times M}$. Moreover, P is unique if $\ker(\mathcal{B}^\mathsf{T}) = \{0\}$, i.e., if $\mathcal{B}$ has full row rank (note that this implies that $N \geq M$). This argument shows that the problem (49.27)

(which is similar to (49.11)) is actually the optimality condition characterizing the minimizer of the functional $\mathsf{V} \mapsto \mathfrak{E}(\mathsf{V})$ under the constraint $\mathsf{V} \in K$.

Another way to look at the above problem consists of introducing the Lagrange multiplier associated with the constraint $\mathsf{V} \in K$, say Q, and considering the *Lagrangian* functional

$$\mathfrak{L}(\mathsf{V}, \mathsf{Q}) := \mathfrak{E}(\mathsf{V}) + (\mathsf{Q}, \mathcal{B}\mathsf{V} - \mathsf{G})_{\ell^2(\mathbb{R}^M)}. \tag{49.28}$$

Then the optimality conditions for a saddle point of $\mathfrak{L}$, say (U, P), are $D_\mathsf{V}\mathfrak{L}(\mathsf{U}, \mathsf{P})(\mathsf{W}) = (\mathcal{A}\mathsf{U} - \mathsf{F}, \mathsf{W})_{\ell^2(\mathbb{R}^N)} + (\mathcal{B}^\mathsf{T}\mathsf{P}, \mathsf{W})_{\ell^2(\mathbb{R}^M)} = 0$ for all $\mathsf{W} \in \mathbb{R}^N$, and $D_\mathsf{Q}\mathfrak{L}(\mathsf{U}, \mathsf{P})(\mathsf{R}) = (\mathsf{R}, \mathcal{B}\mathsf{U} - \mathsf{G})_{\ell^2(\mathbb{R}^M)} = 0$ for all $\mathsf{R} \in \mathbb{R}^M$, which again gives (49.27).

49.3.2 Lagrangian

Let us now reformulate in a more general framework the computations we have done in the previous section in finite dimension.

Definition 49.9 (Saddle point). *Let V and Q be two sets and consider a map $\mathcal{F} : V \times Q \to \mathbb{R}$. A pair $(u, p) \in V \times Q$ is said to be a* saddle point *of $\mathcal{F}$ if*

$$\forall q \in Q, \quad \mathcal{F}(u, q) \leq \mathcal{F}(u, p) \leq \mathcal{F}(v, p), \quad \forall v \in V. \tag{49.29}$$

Equivalently, we have $\sup_{q \in Q} \mathcal{F}(u, q) = \mathcal{F}(u, p) = \inf_{v \in V} \mathcal{F}(v, p)$.

Lemma 49.10 (Inf-sup). *The pair $(u, p) \in V \times Q$ is a saddle point of $\mathcal{F}$ iff*

$$\inf_{v \in V} \sup_{q \in Q} \mathcal{F}(v, q) = \mathcal{F}(u, p) = \sup_{q \in Q} \inf_{v \in V} \mathcal{F}(v, q). \tag{49.30}$$

Proof. According to Definition 49.9, $(u, p) \in V \times Q$ is a saddle point of $\mathcal{F}$ iff

$$\inf_{v \in V} \sup_{q \in Q} \mathcal{F}(v, q) \leq \sup_{q \in Q} \mathcal{F}(u, q) = \mathcal{F}(u, p) = \inf_{v \in V} \mathcal{F}(v, p) \leq \sup_{q \in Q} \inf_{v \in V} \mathcal{F}(v, q).$$

But independently of the existence of a saddle point, one can prove that

$$\sup_{q \in Q} \inf_{v \in V} \mathcal{F}(v, q) \leq \inf_{v \in V} \sup_{q \in Q} \mathcal{F}(v, q). \tag{49.31}$$

Indeed, $\inf_{v \in V} \mathcal{F}(v, r) \leq \mathcal{F}(w, r) \leq \sup_{q \in Q} \mathcal{F}(w, q)$ for all $(w, r) \in V \times Q$. The assertion (49.31) follows by taking the supremum over $r \in Q$ on the left and the infimum over $w \in V$ on the right. Thus, the existence of a saddle point is equivalent to $\sup_{q \in Q} \inf_{v \in V} \mathcal{F}(v, q) = \inf_{v \in V} \sup_{q \in Q} \mathcal{F}(v, q)$. □

Proposition 49.11 (Lagrangian). *Let V and Q be two real Hilbert spaces. Let a be a bounded, symmetric, and coercive bilinear form on $V \times V$. Let b be a bounded bilinear form on $V \times Q$ satisfying (49.19). Let $f \in V'$ and $g \in Q'$. The following three statements are equivalent: (i) u minimizes the quadratic functional $\mathfrak{E}(v) := \frac{1}{2}a(v, v) - f(v)$ on the affine subspace $V_g := \{v \in V \mid b(v, q) = g(q), \forall q \in Q\}$. (ii) There is (a unique) $p \in Q$ such that the pair $(u, p) \in V \times Q$ is a saddle point of the Lagrangian $\mathcal{L}$ s.t.*

$$\mathcal{L}(v, q) := \frac{1}{2}a(v, v) + b(v, q) - f(v) - g(q). \tag{49.32}$$

(iii) The pair (u, p) is the unique solution of (49.18).

Proof. See Exercise 49.2.   □

49.4   Babuška–Brezzi theorem

Let V and M be two real Banach spaces. Consider two bounded linear operators $A : V \to V'$ and $B : V \to M$, and the model problem

$$\begin{cases} \text{Find } u \in V \text{ and } p \in M' \text{ such that} \\ A(u) + B^*(p) = f, \\ B(u) = g, \end{cases} \tag{49.33}$$

where $B^* : M' \to V'$ is the adjoint of B, $f \in V'$, and $g \in M$. The goal of this section is to characterize the well-posedness of (49.33), reformulate it in terms of inf-sup conditions and bilinear forms associated with the operators A and B, and relate this well-posedness result to the BNB theorem (Theorem 25.9). The theory exposed here is due to Babuška and Brezzi [34, 90]. (In the Hilbert setting considered in §49.2-§49.3, the spaces M and Q are related by $Q = M'$.)

49.4.1   Setting with Banach operators

Let $\ker(B)$ be the null space of B and let J_B be the canonical injection from $\ker(B)$ into V and $J_B^* : V' \to \ker(B)'$ be its adjoint. Let $A_\pi : \ker(B) \to \ker(B)'$ be such that $\langle A_\pi(v), w \rangle_{V',V} := \langle A(v), w \rangle_{V',V}$ for all $v, w \in \ker(B)$, i.e., $A_\pi := J_B^* A J_B$.

Theorem 49.12 (Well-posedness). *Problem* (49.33) *is well-posed if and only if A_π is an isomorphism and B is surjective.*

Proof. (1) Assume first that (49.33) is well-posed.
(1.a) Let $g \in M$ and let us denote by (u, p) the solution to (49.33) with data $(0, g)$. Since u satisfies $B(u) = g$, we infer that B is surjective.
(1.b) Let us show that A_π is surjective. Let $h \in \ker(B)'$. Owing to the Hahn–Banach theorem, there is an extension $\widetilde{h} \in V'$ s.t. $\langle \widetilde{h}, v \rangle_{V',V} = \langle h, v \rangle_{V',V}$ for all v in $\ker(B)$ and $\|\widetilde{h}\|_{V'} = \|h\|_{\ker(B)'}$. Let (u, p) be the solution to (49.33) with $f := \widetilde{h}$ and $g := 0$. Then $u \in \ker(B)$. Since $\langle B^*(p), v \rangle_{V',V} = \langle p, B(v) \rangle_{M',M} = 0$ for all $v \in \ker(B)$, we infer that $\langle A_\pi(u), v \rangle_{V',V} = \langle A(u), v \rangle_{V',V} = \langle \widetilde{h}, v \rangle_{V',V} = \langle h, v \rangle_{V',V}$ for all $v \in \ker(B)$. As a result, $A_\pi(u) = h$.
(1.c) Let us show that A_π is injective. Let $u \in \ker(B)$ be s.t. $A_\pi(u) = 0$. Then $\langle A(u), v \rangle_{V',V} = 0$ for all $v \in \ker(B)$, so that $A(u) \in \ker(B)^\perp$. B being surjective, $\mathrm{im}(B)$ is closed and owing to Banach's theorem (Theorem C.35), $\mathrm{im}(B^*) = \ker(B)^\perp$. As a result, $A(u) \in \mathrm{im}(B^*)$, i.e., there is $p \in M'$ such that $A(u) = -B^*(p)$. Hence, $A(u) + B^*(p) = 0$ and $B(u) = 0$, which shows that (u, p) solves (49.33) with $f := 0$ and $g := 0$. Uniqueness of the solution to (49.33) implies that $u = 0$.
(2) Conversely, assume that A_π is an isomorphism and B is surjective.
(2.a) For all $f \in V'$ and all $g \in M$, let us show that there is at least one solution to (49.33). The operator B being surjective, there is $u_g \in V$ s.t. $B(u_g) = g$. Denote by $h_{f,g}$ the bounded linear form on $\ker(B)$ s.t. $\langle h_{f,g}, v \rangle_{V',V} = \langle f, v \rangle_{V',V} - \langle A(u_g), v \rangle_{V',V}$ for all $v \in \ker(B)$. Since $A_\pi : \ker(B) \to \ker(B)'$ is an isomorphism, A_π is surjective, so that there is $\phi \in \ker(B)$ s.t. $A_\pi(\phi) = h_{f,g}$. Set $u := \phi + u_g$. The linear form $f - A(u)$ is in $\ker(B)^\perp$. Since B is surjective, $\ker(B)^\perp = \mathrm{im}(B^*)$, i.e., there is $p \in M'$ such that $B^*(p) = f - A(u)$. Moreover, $B(u) = B(\phi + u_g) = B(u_g) = g$. Hence, we have constructed a solution to (49.33).
(2.b) Let us show that the solution is unique. Let (u, p) be such that $B(u) = 0$ and $A(u) + B^*(p) = 0$, so that $u \in \ker(B)$ and $A_\pi(u) = 0$. Since πA is injective, $u = 0$. As a result, $B^*(p) = 0$. Since B is surjective, B^* is injective, which implies $p = 0$. $\square$

49.4.2   Setting with bilinear forms and reflexive spaces

Let us now assume that V and M are reflexive Banach spaces and let us set $Q := M'$. Notice that this implies that $Q' = M'' = M$. Thus, we have $B \in \mathcal{L}(V; Q')$ and $B^* \in \mathcal{L}(Q; V')$. Consider the two bounded bilinear forms a and b defined, respectively, on $V \times V$ and on $V \times Q$ s.t. $a(v, w) := \langle A(v), w \rangle_{V',V}$ and $b(v, q) := \langle B(v), q \rangle_{Q',Q}$. Let us set

$$\|a\| := \sup_{v \in V} \sup_{w \in W} \frac{|a(v, w)|}{\|v\|_V \|w\|_V}, \qquad \|b\| := \sup_{v \in V} \sup_{q \in Q} \frac{|b(v, q)|}{\|v\|_V \|q\|_Q}. \tag{49.34}$$

Let $f \in V'$ and $g \in Q'$. With the shorthand notation $f(v) := \langle f, v \rangle_{V',V}$ and $g(q) := \langle g, q \rangle_{Q',Q}$, the abstract problem (49.33) is reformulated as follows:

$$\begin{cases} \text{Find } u \in V \text{ and } p \in Q \text{ such that} \\ a(u, w) + b(w, p) = f(w), & \forall w \in V, \\ b(u, q) = g(q), & \forall q \in Q. \end{cases} \tag{49.35}$$

Theorem 49.13 (Babuška–Brezzi). (49.35) *is well-posed if and only if*

$$\begin{cases} \inf_{v \in \ker(B)} \sup_{w \in \ker(B)} \frac{|a(v, w)|}{\|v\|_V \|w\|_V} =: \alpha > 0, \\ \forall w \in \ker(B), \quad [\forall v \in \ker(B), \, a(v, w) = 0] \implies [w = 0], \end{cases} \tag{49.36}$$

and the following inequality, usually called Babuška–Brezzi condition, *holds:*

$$\inf_{q \in Q} \sup_{v \in V} \frac{|b(v, q)|}{\|v\|_V \|q\|_Q} =: \beta > 0. \tag{49.37}$$

Furthermore, we have the following a priori estimates:

$$\|u\|_V \le c_1 \|f\|_{V'} + c_2 \|g\|_{Q'}, \tag{49.38a}$$
$$\|p\|_Q \le c_3 \|f\|_{V'} + c_4 \|g\|_{Q'}, \tag{49.38b}$$

with $c_1 := \frac{1}{\alpha}$, $c_2 := \frac{1}{\beta}(1 + \frac{\|a\|}{\alpha})$, $c_3 := \frac{1}{\beta}(1 + \frac{\|a\|}{\alpha})$, *and* $c_4 := \frac{\|a\|}{\beta^2}(1 + \frac{\|a\|}{\alpha})$.

Proof. (1) Since $\ker(B) \subset V$ is reflexive, we infer form Theorem C.49 that (49.36) is equivalent to A_π being an isomorphism. Furthermore, (49.37) is equivalent to B being surjective owing to (C.17) in Lemma C.40 since Q is reflexive. We invoke Theorem 49.12 to conclude that (49.35) is well-posed iff (49.36)-(49.37) hold true.
(2) Let us now prove the a priori estimates (49.38). From the condition (49.37) and Lemma C.42 (since Q is reflexive), we deduce that there exists $u_g \in V$ such that $B(u_g) = g$ and $\beta \|u_g\|_V \le \|g\|_{Q'}$. Setting $\phi := u - u_g \in \ker(B)$ yields $a(\phi, v) = f(v) - a(u_g, v)$ for all $v \in \ker(B)$. Since

$$|a(\phi, v)| \le (\|f\|_{V'} + \|a\| \, \|u_g\|_V) \|v\|_V \le \left(\|f\|_{V'} + \frac{\|a\|}{\beta} \|g\|_{Q'} \right) \|v\|_V,$$

taking the supremum over v in $\ker(B)$ yields $\alpha \|\phi\|_V \le \|f\|_{V'} + \frac{\|a\|}{\beta} \|g\|_{Q'}$ owing to (49.36). The estimate on u results from this inequality and the triangle inequality. To prove the estimate on p, we deduce from (49.37) and Lemma C.40 that $\beta \|p\|_Q \le \|B^*(p)\|_{V'}$, yielding $\beta \|p\|_Q \le \|a\| \|u\|_V + \|f\|_{V'}$. The estimate on $\|p\|_Q$ then results from that on $\|u\|_V$. $\qquad \square$

Remark 49.14 (Coercivity). The conditions in (49.36) are automatically fulfilled if the bilinear form a is coercive on $\ker(B)$ or coercive on X. □

Let us recall that we have adopted the convention that suprema and infima in expressions like (49.34)-(49.36)-(49.37) are taken over nonzero arguments. To relate the conditions (49.36) and (49.37) with the conditions (BNB1) and (BNB2) from the BNB theorem (Theorem 25.9), let us introduce the space $X := V \times Q$ equipped with the norm $\|(v,q)\|_X := \|v\|_V + \|q\|_Q$ and let us recall from (49.20) the bounded bilinear form t on $X \times X$ defined by

$$t((v,q),(w,r)) := a(v,w) + b(w,q) + b(v,r). \tag{49.39}$$

Theorem 49.15 (Link with BNB). *The bilinear form t satisfies the conditions* (BNB1) *and* (BNB2) *if and only if* (49.36) *and* (49.37) *are satisfied.*

Proof. (1) Let us prove that (49.36) and (49.37) imply (BNB1) and (BNB2). (1a) Proof of (BNB1). Let $(v,q) \in X$ and set $\mathbb{S} := \sup_{(w,r) \in X} \frac{|t((v,q),(w,r))|}{\|(w,r)\|_X}$. Lemma C.42 implies that there exists $\widehat{v} \in V$ such that $B(\widehat{v}) = B(v)$ and $\beta \|\widehat{v}\|_V \leq \|B(v)\|_{Q'}$. We infer that

$$\beta \|\widehat{v}\|_V \leq \|B(v)\|_{Q'} = \sup_{r \in Q} \frac{|b(v,r)|}{\|r\|_Q} = \sup_{(0,r) \in X} \frac{|t((v,q),(0,r))|}{\|(0,r)\|_X} \leq \mathbb{S}.$$

Observing that $v - \widehat{v} \in \ker(B)$ we also infer that

$$
\begin{aligned}
\alpha \|v - \widehat{v}\|_V &\leq \sup_{w \in \ker(B)} \frac{|a(v - \widehat{v}, w)|}{\|w\|_V} \\
&= \sup_{w \in \ker(B)} \frac{|a(v - \widehat{v}, w) + b(w,q) + b(v,0)|}{\|w\|_V} \\
&\leq \sup_{(w,0) \in X} \frac{|t((v,q),(w,0))|}{\|(w,0)\|_X} + \sup_{w \in \ker(B)} \frac{|a(\widehat{v}, w)|}{\|w\|_V} \\
&\leq \mathbb{S} + \|a\| \, \|\widehat{v}\|_V \leq \left(1 + \frac{\|a\|}{\beta}\right) \mathbb{S}.
\end{aligned}
$$

Using the triangle inequality yields $\|v\|_V \leq \left(\frac{1}{\beta} + \frac{1}{\alpha}\left(1 + \frac{\|a\|}{\beta}\right)\right) \mathbb{S}$. Then we proceed as follows to bound $\|q\|_Q$:

$$\beta \|q\|_Q \leq \sup_{w \in V} \frac{|b(w,q)|}{\|w\|_V} \leq \sup_{w \in V} \frac{|a(v,w) + b(w,q) + b(v,0)|}{\|(w,0)\|_X} + \sup_{w \in V} \frac{|a(v,w)|}{\|w\|_V}.$$

This estimate implies that $\beta \|q\|_Q \leq \mathbb{S} + \|a\| \|v\|_V$, and we conclude that

$$\|q\|_Q \leq \frac{1}{\beta}\left(1 + \|a\|\left(\frac{1}{\beta} + \frac{1}{\alpha}\left(1 + \frac{\|a\|}{\beta}\right)\right)\right) \mathbb{S}.$$

This proves (BNB1).
(1b) Let $(w,r) \in X$ be s.t. $t((v,q),(w,r)) = 0$ for all $(v,q) \in X$, i.e.,

$$a(v,w) + b(v,r) = 0, \quad \forall v \in V, \tag{49.40a}$$
$$b(w,r) = 0, \qquad \quad \forall r \in Q. \tag{49.40b}$$

Then (49.40b) implies that $w \in \ker(B)$, and taking $v \in \ker(B)$ in (49.40a), we infer that $a(v, w) = 0$, for all $v \in \ker(B)$. The second statement in (49.36) implies that $w = 0$. Finally, (49.40a) yields $b(v, r) = 0$ for all $v \in V$, and (49.37) implies that $r = 0$. This proves (BNB2).

(2) Let us now prove that the conditions (BNB1) and (BNB2) on the bilinear form t imply the conditions (49.36) and (49.37) on the bilinear forms a and b. Let γ denote the inf-sup constant of the bilinear form t on $X \times X$.

(2a) Let us start with (49.37). For all $q \in Q$, we have

$$\gamma \|q\|_Q = \gamma \|(0, q)\|_X \leq \sup_{(w, r) \in X} \frac{|t((0, q), (w, r))|}{\|(w, r)\|_X} = \sup_{(w, r) \in X} \frac{|b(w, q)|}{\|(w, r)\|_X}$$

$$= \sup_{w \in V} \sup_{r \in Q} \frac{|b(w, q)|}{\|(w, r)\|_X} = \sup_{w \in V} \frac{|b(w, q)|}{\|w\|_V},$$

since the supremum over $r \in Q$ is reached for $r = 0$. This proves (49.37) with $\beta \geq \gamma > 0$.

(2b) Let us prove the first statement in (49.36). For all $w \in V$, we define $(w'_w, r'_w) \in X$ to be the solution to the adjoint problem $t((v, q), (w'_w, r'_w)) = a(v, w)$ for all $(v, q) \in X$. Owing to Lemma C.53, this problem is well-posed. Moreover, we have $w'_w \in \ker(B)$ and $\gamma \|w'_w\|_V \leq \|a\| \|w\|_V$. Let $v \in \ker(B)$. We have $a(v, w) = t((v, q), (w'_w, r'_w)) = a(v, w'_w)$ for all $q \in Q$. We infer that

$$\gamma \|v\|_V = \gamma \|(v, 0)\|_X \leq \sup_{(w, r) \in X} \frac{|t((v, 0), (w, r))|}{\|(w, r)\|_X} = \sup_{(w, r) \in X} \frac{|a(v, w)|}{\|(w, r)\|_X}$$

$$= \sup_{w \in V} \frac{|a(v, w)|}{\|w\|_V} = \sup_{w \in V} \frac{|a(v, w'_w)|}{\|w\|_V} \leq \frac{\|a\|}{\gamma} \sup_{w \in V} \frac{|a(v, w'_w)|}{\|w'_w\|_V}.$$

Since $w'_w \in \ker(B)$, this finally gives $\frac{\gamma^2}{\|a\|} \|v\|_V \leq \sup_{w \in \ker(B)} \frac{|a(v, w)|}{\|w\|_V}$, which is the first statement in (49.36) with $\alpha \geq \frac{\gamma^2}{\|a\|} > 0$.

(2c) Let us now prove the second statement in (49.36). We first recall that we have already established that (49.37) holds true. From Lemma C.40, we then infer that $\mathrm{im}(B^*)$ is closed in V'. Let $w \in \ker(B)$ be s.t. $a(v, w) = 0$ for all $v \in \ker(B)$. This implies that $A^*(w) \in \ker(B)^\perp = \overline{\mathrm{im}(B^*)} = \mathrm{im}(B^*)$. Then there is $r_w \in Q$ s.t. $B^*(r_w) = A^*(w)$. For all $(v, q) \in X$, we then have

$$t((v, q), (w, -r_w)) = a(v, w) + b(v, q) - b(v, r_w) = a(v, w) - b(v, r_w)$$
$$= \langle A^*(w) - B^*(r_w), v \rangle_{V', V} = 0.$$

The condition (BNB2) on t implies that $(w, -r_w) = 0$, so that $w = 0$. □

Remark 49.16 (Sharper estimate). Sharper estimates on the inf-sup stability constant of t have been derived in Corollary 49.7 and Proposition 49.8 under the assumption that the bilinear form a is symmetric and coercive. □

Remark 49.17 (Direct sums). Notice that the map $V \ni w \mapsto w'_w \in \ker(B)$ introduced in step (2b) of the proof of Theorem 49.15 implies that any $w \in V$ can be uniquely decomposed into $w = w'_w + (A^*)^{-1} B^* r'_w$. This means that we have the direct decomposition $V = \ker(B) \oplus \mathrm{im}((A^*)^{-1} B^*)$. Note also that the same argument implies that $V = \ker(B) \oplus \mathrm{im}(A^{-1} B^*)$. □

Exercises

Exercise 49.1 (Algebraic setting). (i) Derive the counterpart of Theorem 49.12 in the setting of §49.3.1. (*Hint*: assume that the matrix $\mathcal{B}$ has full row rank and consider a basis of $\ker(\mathcal{B})$.) (ii)

What happens if the matrix $\mathcal{A}$ is symmetric positive definite?

Exercise 49.2 (Constrained minimization). The goal is to prove Proposition 49.11. (i) Prove that if u minimizes $\mathfrak{E}$ over V_g, there is (a unique) $p \in Q$ such that (u,p) solves (49.35). (*Hint*: proceed as in §49.3.1.) (ii) Prove that (u,p) solves (49.35) if and only if (u,p) is a saddle point of $\mathcal{L}$. (*Hint*: consider $\mathfrak{E}_p : V \to \mathbb{R}$ s.t. $\mathfrak{E}_p(v) := \mathcal{L}(v,p)$ with fixed $p \in Q$.) (iii) Prove that if (u,p) is a saddle point of $\mathcal{L}$, then u minimizes $\mathfrak{E}$ over V_g. (iv) Application: minimize $\mathfrak{E}(v) := 2v_1^2 + 2v_2^2 - 6v_1 + v_2$ over $\mathbb{R}^2$ under the constraint $2v_1 + 3v_2 = -1$.

Exercise 49.3 (Symmetric operator). Let X be a Hilbert space and let $T \in \mathcal{L}(X;X)$ be a bijective symmetric operator. (i) Prove that T^{-1} is symmetric. (ii) Prove that $[\lambda \in \sigma(T)] \iff [\lambda^{-1} \in \sigma(T^{-1})]$. (*Hint*: use Corollary 46.18.) (iii) Prove that $\sigma(T) \subset \mathbb{R}$. (*Hint*: consider the sesquilinear form $t_\lambda(x,y) := ((T - \lambda I_X)(x), y)_X$ and use the Lax–Milgram lemma.)

Exercise 49.4 (Sharp stability). The goal is to prove Proposition 49.8. (i) Assume that $\ker(B)$ is nontrivial. Verify that $1 \in \sigma_{\mathrm{p}}(\widetilde{T})$. (ii) Let $\lambda \neq 1$ be in $\sigma(\widetilde{T})$. Prove that $\lambda(\lambda - 1) \in \sigma(S)$. (*Hint*: consider the sequence $x_n := (v_n, q_n)$ in X from Corollary 46.18, then observe that $(S(q_n), q_n)_Q = (1 - \lambda)^2 \langle A(v_n), v_n \rangle_{V',V} + \delta_n$, with $\delta_n := \langle B^*(q_n) + (1 - \lambda)A(v_n), A^{-1}B^*(q_n) - (1 - \lambda)v_n \rangle_{V',V}$, and prove that $S(q_n) - \lambda(\lambda - 1)q_n \to 0$ and $\liminf_{n\to\infty} \|q_n\|_Q > 0$.) (iii) Prove that $\sigma(\widetilde{T}) \subset [\lambda_\sharp^-, \lambda_\flat^-] \cup \{1\} \cup [\lambda_\flat^+, \lambda_\sharp^+]$ with $\lambda_\flat^\pm = \frac{1}{2}(1 \pm (4\frac{\beta^2}{\|a\|} + 1)^{\frac{1}{2}})$, and $\lambda_\sharp^\pm = \frac{1}{2}(1 \pm (4\frac{\|b\|^2}{\alpha} + 1)^{\frac{1}{2}})$. (*Hint*: use Lemma 49.1.) (iv) Conclude. (*Hint*: $\widetilde{T}$ is symmetric with respect to the weighted inner product $(x,y)_{\widetilde{X}} := a(v,w) + (q,r)_Q$.)

Exercise 49.5 (Abstract Helmholtz decomposition). Consider the setting of §49.2 and equip V with the bilinear form a as inner product. (i) Prove that $\mathrm{im}(A^{-1}B^*)$ is closed and that $V = \ker(B) \oplus \mathrm{im}(A^{-1}B^*)$, the sum being a-orthogonal. (*Hint*: use Lemma C.39.) (ii) Let $f \in \ker(B)^\perp$. Prove that solving $b(v,p) = f(v)$ for all $v \in V$ is equivalent to solving $(S(p), q)_Q = (J_Q^{-1}BA^{-1}(f), q)_Q$ for all $q \in Q$.

Exercise 49.6 (Maxwell's equations). Consider the following problem: For $\boldsymbol{f} \in \boldsymbol{L}^2(D)$, find $\boldsymbol{A}$ and ϕ such that

$$\begin{cases} \nabla\times(\kappa\nabla\times\boldsymbol{A}) + \nu\nabla\phi = \boldsymbol{f}, \\ \nabla\cdot(\nu\boldsymbol{A}) = 0, \\ \boldsymbol{A}_{|\partial D_\mathrm{d}}\times\boldsymbol{n} = \boldsymbol{0}, \ \phi_{|\partial D_\mathrm{d}} = 0, \ (\kappa\nabla\times\boldsymbol{A})_{|\partial D_\mathrm{n}}\times\boldsymbol{n} = \boldsymbol{0}, \ \boldsymbol{A}_{|\partial D_\mathrm{n}}\cdot\boldsymbol{n} = 0, \end{cases}$$

where κ, ν are real and positive constants (for simplicity), and $|\partial D_\mathrm{d}| > 0$ (see §49.1.3; here we write $\boldsymbol{A}$ in lieu of $\boldsymbol{H}$ and we consider mixed Dirichlet–Neumann conditions). (i) Give a mixed weak formulation of this problem. (*Hint*: use the spaces $\boldsymbol{V}_\mathrm{d} := \{\boldsymbol{v} \in \boldsymbol{H}(\mathrm{curl}; D) \mid \gamma^\mathrm{c}(\boldsymbol{v})_{|\partial D_\mathrm{d}} = \boldsymbol{0}\}$, where the meaning of the boundary condition is specified in §43.2.1, and $Q_\mathrm{d} := \{q \in H^1(D) \mid \gamma^\mathrm{g}(q)_{|\partial D_\mathrm{d}} = 0\}$.) (ii) Let $B : \boldsymbol{V}_\mathrm{d} \to Q_\mathrm{d}'$ be s.t. $\langle B(\boldsymbol{v}), q \rangle_{Q_\mathrm{d}', Q_\mathrm{d}} := (\nu\boldsymbol{v}, \nabla q)_{\boldsymbol{L}^2(D)}$. Let $\boldsymbol{v} \in \ker(B)$. Show that $\nabla\cdot\boldsymbol{v} = 0$ and, if $\boldsymbol{v} \in \boldsymbol{H}^1(D)$, $\gamma^\mathrm{g}(\boldsymbol{v})_{|\partial D_\mathrm{n}}\cdot\boldsymbol{n} = 0$. (*Hint*: recall that ν is constant.) (iii) Accept as a fact that D, ∂D_d, ∂D_n have topological and smoothness properties such that there exists $c > 0$ s.t. $\ell_D\|\nabla\times\boldsymbol{v}\|_{\boldsymbol{L}^2(D)} \geq c\|\boldsymbol{v}\|_{\boldsymbol{L}^2(D)}$, for all $\boldsymbol{v} \in \ker(B)$, with $\ell_D := \mathrm{diam}(D)$. Show that the above weak problem is well-posed. (*Hint*: use Theorem 49.13.) (iv) Let $(\mathcal{T}_h)_{h\in\mathcal{H}}$ be a shape-regular sequence of affine meshes. Let $k \geq 0$, let $\boldsymbol{V}_h := \boldsymbol{P}_k^\mathrm{c}(\mathcal{T}_h) \cap \boldsymbol{V}_\mathrm{d}$, and let $Q_h := P_{k+1}^\mathrm{g}(\mathcal{T}_h) \cap Q_\mathrm{d}$. Show that $\nabla Q_h \subset \boldsymbol{V}_h$. (v) Show that the discrete mixed problem is well-posed in $\boldsymbol{V}_h\times Q_h$ assuming that $\partial D_\mathrm{d} = \partial D$. (*Hint*: invoke Theorem 44.6.)

Chapter 50

Mixed finite element approximation

This chapter is concerned with the approximation of the model problem analyzed in Chapter 49. We focus on the Galerkin approximation in the conforming setting. We establish necessary and sufficient conditions for well-posedness, and we derive error bounds in terms of the best-approximation error. Then we consider the algebraic viewpoint, and we discuss augmented Lagrangian methods in the context of saddle point problems. Finally, we examine iterative solvers, including Uzawa iterations and Krylov subspace methods.

50.1 Conforming Galerkin approximation

Let V and Q be two reflexive (real) Banach spaces. Let a and b be two bounded bilinear forms on $V \times V$ and on $V \times Q$ respectively. Let $f \in V'$ and let $g \in Q'$. We consider the following model problem:

$$\begin{cases} \text{Find } u \in V \text{ and } p \in Q \text{ such that} \\ a(u,w) + b(w,p) = f(w), & \forall w \in V, \\ b(u,q) = g(q), & \forall q \in Q. \end{cases} \tag{50.1}$$

We introduce the associated operators $A \in \mathcal{L}(V;V')$ and $B \in \mathcal{L}(V;Q')$ such that $a(v,w) := \langle A(v), w \rangle_{V',V}$ and $b(v,q) := \langle B(v), q \rangle_{Q',Q}$. We assume that (50.1) is well-posed. Owing to Theorem 49.13, this means that the bilinear form a satisfies the conditions (49.36) (implying the inf-sup condition $\inf_{v \in \ker(B)} \sup_{w \in \ker(B)} \frac{|a(v,w)|}{\|v\|_V \|w\|_V} =: \alpha > 0$) and that the bilinear form b satisfies the inf-sup condition (49.37), i.e., $\inf_{q \in Q} \sup_{v \in V} \frac{|b(v,q)|}{\|v\|_V \|q\|_Q} =: \beta > 0$.

A conforming Galerkin approximation of (50.1) is obtained by considering two finite-dimensional subspaces $V_h \subset V$, $Q_h \subset Q$. The discrete problem is

$$\begin{cases} \text{Find } u_h \in V_h \text{ and } p_h \in Q_h \text{ such that} \\ a(u_h, w_h) + b(w_h, p_h) = f(w_h), & \forall w_h \in V_h, \\ b(u_h, q_h) = g(q_h), & \forall q_h \in Q_h. \end{cases} \tag{50.2}$$

50.1.1   Well-posedness

Let $B_h : V_h \to Q'_h$ be the discrete counterpart of the operator $B : V \to Q'$, that is,

$$\langle B_h(v_h), q_h \rangle_{Q'_h, Q_h} := \langle B(v_h), q_h \rangle_{Q', Q} = b(v_h, q_h), \qquad \forall (v_h, q_h) \in V_h \times Q_h.$$

The null space of B_h is such that

$$\ker(B_h) = \{v_h \in V_h \mid \forall q_h \in Q_h, \, b(v_h, q_h) = 0\}. \tag{50.3}$$

One important aspect of the discretization is that the surjectivity of B does not imply that of B_h. One rare occasion where this is nevertheless the case is when $B^*(Q_h) \subset V_h$, i.e., $B_h^* = B_{|Q_h}^*$. This exceptional situation is illustrated in Exercise 49.6. Note also that in general, $\ker(B_h)$ is not necessarily a subspace of $\ker(B)$.

Proposition 50.1 (Well-posedness). (50.2) *is well-posed if and only if*

$$\inf_{v_h \in \ker(B_h)} \sup_{w_h \in \ker(B_h)} \frac{|a(v_h, w_h)|}{\|v_h\|_V \|w_h\|_V} := \alpha_h > 0, \tag{50.4a}$$

$$\inf_{q_h \in Q_h} \sup_{v_h \in V_h} \frac{|b(v_h, q_h)|}{\|v_h\|_V \|q_h\|_Q} := \beta_h > 0. \tag{50.4b}$$

Proof. Apply Theorem 49.13 and use the fact that (50.4a) implies both conditions in (49.36) since V_h is finite-dimensional; see Remark 26.7. □

The condition (50.4a) holds true for all conforming subspaces V_h and Q_h if a is V-coercive (the coercivity of a on $\ker(B)$ may not be sufficient since it may happen that $\ker(B_h) \not\subset \ker(B)$). Note that verifying the inf-sup condition for a on $V_h \times V_h$ is not sufficient to prove (50.4a) (think of an invertible matrix having a square diagonal sub-block that is not invertible; see Exercise 50.1). Furthermore, the condition (50.4b) is equivalent to B_h being surjective, which is also equivalent to B_h^* being injective since the setting is finite-dimensional. In practice, it is important that both (50.4a) and (50.4b) hold true uniformly w.r.t. $h \in \mathcal{H}$, i.e., $\inf_{h \in \mathcal{H}} \alpha_h =: \alpha_0 > 0$ and $\inf_{h \in \mathcal{H}} \beta_h =: \beta_0 > 0$.

50.1.2   Error analysis

Our goal is to estimate the approximation errors $(u - u_h)$ and $(p - p_h)$ in terms of the best-approximation error on u by a discrete field in V_h and the best-approximation error on p by a discrete function in Q_h. Céa's lemma (Lemma 26.13) could be applied to the bilinear form $t((v, q), (w, r)) := a(v, w) + b(w, q) + b(v, r)$ introduced in §49.4.2 (see Exercise 50.2). But here, we present a more specific analysis distinguishing the errors on u and on p. We say that $\Pi_h \in \mathcal{L}(V; V_h)$ is a Fortin operator for the bilinear form b if $b(\Pi_h(v) - v, q_h) = 0$ for all $q_h \in Q_h$ and all $v \in V$. (We do not assume V_h to be pointwise invariant under Π_h.) This class of operators is investigated in §26.2.3. In particular, Lemma 26.9 shows that the inf-sup condition (50.4b) implies the existence of a Fortin operator with $\|\Pi_h\|_{\mathcal{L}(V; V_h)} \leq \frac{\|b\|}{\beta_h}$.

Lemma 50.2 (Error estimate). *Let (u, p) solve (50.1). Assume (50.4) and let (u_h, p_h) solve the discrete problem (50.2). Let $\Pi_h \in \mathcal{L}(V; V_h)$ be any Fortin operator. The following error estimates hold true:*

$$\|u - u_h\|_V \leq c_{1h} \inf_{v_h \in \Pi_h(u) + \ker(B_h)} \|u - v_h\|_V + c_{2h} \inf_{q_h \in Q_h} \|p - q_h\|_Q, \tag{50.5a}$$

$$\|p - p_h\|_Q \leq c_{3h} \inf_{v_h \in \Pi_h(u) + \ker(B_h)} \|u - v_h\|_V + c_{4h} \inf_{q_h \in Q_h} \|p - q_h\|_Q, \tag{50.5b}$$

with $c_{1h} := (1 + \frac{\|a\|}{\alpha_h})$, $c_{2h} := \frac{\|b\|}{\alpha_h}$ *if* $\ker(B_h) \not\subset \ker(B)$ *and* $c_{2h} := 0$ *otherwise*, $c_{3h} := c_{1h}\frac{\|a\|}{\beta_h}$, *and* $c_{4h} := 1 + \frac{\|b\|}{\beta_h} + c_{2h}\frac{\|a\|}{\beta_h}$.

Proof. (1) Estimate on $(u - u_h)$. Let $v_h \in \Pi_h(u) + \ker(B_h)$, i.e., $v_h := \Pi_h(u) + \gamma_h$ with $\gamma_h \in \ker(B_h)$. Then $u_h - v_h \in \ker(B_h)$ since we have

$$b(u_h - v_h, q_h) = b(u_h - \Pi_h(u), q_h) + b(\gamma_h, q_h) = b(u_h - u, q_h) = 0,$$

for all $q_h \in Q_h$, where we used the Galerkin orthogonality property for the second equation in (50.2). Owing the inf-sup condition (50.4a), we infer that

$$\alpha_h \|u_h - v_h\|_V \leq \sup_{y_h \in \ker(B_h)} \frac{|a(u_h - v_h, y_h)|}{\|y_h\|_V}$$

$$= \sup_{y_h \in \ker(B_h)} \frac{|b(y_h, p - p_h) + a(u - v_h, y_h)|}{\|y_h\|_V},$$

where the equality follows from the Galerkin orthogonality property for the first equation in (50.2), i.e., we have $a(u - u_h, y_h) + b(y_h, p - p_h) = 0$ for all $y_h \in V_h$. If $\ker(B_h) \subset \ker(B)$, then $b(y_h, p - p_h) = 0$ for all $y_h \in \ker(B_h)$, yielding

$$\alpha_h \|u_h - v_h\|_V \leq \|a\| \, \|u - v_h\|_V.$$

In the general case, we have $b(y_h, p_h) = 0 = b(y_h, q_h)$ for all $q_h \in Q_h$, since y_h is in $\ker(B_h)$. This implies that

$$\alpha_h \|u_h - v_h\|_V \leq \|a\| \, \|u - v_h\|_V + \|b\| \, \|p - q_h\|_Q.$$

Hence, both cases are summarized by the following estimate:

$$\|u_h - v_h\|_V \leq \frac{\|a\|}{\alpha_h} \|u - v_h\|_V + c_{2h}\|p - q_h\|_Q,$$

with c_{2h} as in the assertion. Using the triangle inequality and taking the infimum over $v_h \in \Pi_h(u) + \ker(B_h)$ and over $q_h \in Q_h$ leads to (50.5a).
(2) Estimate on $(p - p_h)$. Using again the Galerkin orthogonality property for the first equation in (50.2), we have

$$b(v_h, q_h - p_h) = a(u_h - u, v_h) + b(v_h, q_h - p), \qquad \forall(v_h, q_h) \in V_h \times Q_h.$$

Combined with the inf-sup condition (50.4b), this implies that

$$\beta_h \|q_h - p_h\|_Q \leq \sup_{v_h \in V_h} \frac{|b(v_h, q_h - p_h)|}{\|v_h\|_V}$$

$$\leq \|a\| \, \|u - u_h\|_V + \|b\| \, \|p - q_h\|_Q.$$

The bound (50.5b) follows from the triangle inequality, the bound on $(u - u_h)$, and by taking the infimum over $q_h \in Q_h$. $\qquad\square$

The estimate on $(u - u_h)$ involves the best-approximation error on u by a member of the affine subspace $\Pi_h(u) + \ker(B_h)$. This error may not be easy to estimate in practice, and it is sometimes preferable to bound it by the best-approximation error on u by a member of V_h since $\Pi_h(u) + \ker(B_h) \subset V_h$. Of course, the best-approximation error in $\Pi_h(u) + \ker(B_h)$ is larger than the best-approximation error in V_h. The following lemma quantifies the discrepancy. (Recall that (50.4b) is equivalent to the existence of Fortin operators.)

Lemma 50.3 (Best-approximation in V_h). *Assume* (50.4b). *The following holds true for all* $u \in V$ *and any Fortin operator* $\Pi_h \in \mathcal{L}(V; V_h)$:

$$\inf_{v_h \in \Pi_h(u) + \ker(B_h)} \|u - v_h\|_V \leq (1 + \|\Pi_h\|_{\mathcal{L}(V;V_h)}) \inf_{y_h \in V_h} \|u - y_h\|_V. \tag{50.6}$$

Proof. Let $u \in V$. Let $y_h \in V_h$ and set $z_h := \Pi_h(u - y_h)$. Then $y_h + z_h = \Pi_h(u) + y_h - \Pi_h(y_h) \in \Pi_h(u) + \ker(B_h)$ since $b(y_h - \Pi_h(y_h), q_h) = 0$ for all $q_h \in Q_h$. This implies that

$$\inf_{v_h \in \Pi_h(u) + \ker(B_h)} \|u - v_h\|_V \leq \|u - (y_h + z_h)\|_V \leq \|u - y_h\|_V + \|z_h\|_V$$

$$\leq (1 + \|\Pi_h\|_{\mathcal{L}(V;V_h)}) \|u - y_h\|_V,$$

and we conclude by taking the infimum over $y_h \in V_h$. □

Remark 50.4 ($g = 0$). If $g = 0$, then $\Pi_h(u) \in \ker(B_h)$, and the infimum in (50.5) and (50.6) reduces to $v_h \in \ker(B_h)$. □

Corollary 50.5 (Error estimate). *Let* (u, p) *solve* (50.1). *Assume* (50.4) *and let* (u_h, p_h) *solve the discrete problem* (50.2). *The following error estimates hold true:*

$$\|u - u_h\|_V \leq c'_{1h} \inf_{v_h \in V_h} \|u - v_h\|_V + c_{2h} \inf_{q_h \in Q_h} \|p - q_h\|_Q, \tag{50.7a}$$

$$\|p - p_h\|_Q \leq c'_{3h} \inf_{v_h \in V_h} \|u - v_h\|_V + c_{4h} \inf_{q_h \in Q_h} \|p - q_h\|_Q, \tag{50.7b}$$

with $c'_{1h} := (1 + \frac{\|a\|}{\alpha_h})(1 + \|\Pi_h\|_{\mathcal{L}(V;V_h)})$ *for every Fortin operator* $\Pi_h \in \mathcal{L}(V; V_h)$, $c'_{3h} := c'_{1h} \frac{\|a\|}{\beta_h}$, *and* c_{2h}, c_{4h} *are as in Lemma* 50.2.

Proof. Combine Lemma 50.2 with Lemma 50.3. □

Remark 50.6 (c'_{1h}). Lemma 26.9 asserts the existence of a Fortin operator with $\|\Pi_h\|_{\mathcal{L}(V;V_h)} \leq \frac{\|b\|}{\beta_h}$. Hence, the upper bound $c'_{1h} \leq (1 + \frac{\|a\|}{\alpha_h})(1 + \frac{\|b\|}{\beta_h})$ always holds true. However, the estimate $\|\Pi_h\|_{\mathcal{L}(V;V_h)} \leq \frac{\|b\|}{\beta_h}$ can be pessimistic. For instance, for the Stokes equations in elongated domains, the boundedness constant of the bilinear form $b(\boldsymbol{v}, p) = (\nabla \cdot \boldsymbol{v}, p)_{L^2(D)}$ on $\boldsymbol{H}_0^1(D) \times L^2(D)$ is $\|b\| = 1$, and the inf-sup constant β_h can be shown to be very small (see Chizhonkov and Olshanskii [119], Dobrowolski [170]), whereas for some of these domains it is possible to construct a Fortin operator with norm of order unity (see Mardal et al. [294], Linke et al. [284]). □

Remark 50.7 ($\ker(B_h)$). We refer the reader to Theorem 51.16 for an example of error estimate exploiting the approximation properties in $\ker(B_h)$ in the context of Darcy's equations. □

Remark 50.8 (c_{2h}). The constant c_{2h} vanishes whenever $\ker(B_h) \subset \ker(B)$. Using a discrete pair (V_h, Q_h) that guarantees that $\ker(B_h) \subset \ker(B)$ may be interesting when the best approximation error on p is (much) larger than that on u. A simple example where this occurs is when $f = B^*(p)$ for some $p \in Q$ and $g = 0$, so that the solution to (50.1) is $(0, p)$. If $\ker(B_h) \subset \ker(B)$, the estimate (50.7a) implies that $u_h = u = 0$. But if $\ker(B_h) \not\subset \ker(B)$, then u_h is generally nonzero and grows linearly with the size of p, which is not a desirable property. More generally, the well-posedness of (50.1) with $g := 0$ implies the abstract *Helmholtz decomposition* $V' = Y_0 \oplus Y_1$ with $Y_0 := A(\ker(B))$ and $Y_1 = \text{im}(B^*)$. Whenever the component of f in Y_1 is much larger than that in Y_0, the best-approximation error on p dominates the approximation error on u unless the discretization satisfies $\ker(B_h) \subset \ker(B)$. See also Remark 53.22 for further insight in the context of the Stokes equations. □

Remark 50.9 (Stabilization). It is possible to approximate (50.1) using discrete spaces V_h and Q_h that violate the inf-sup condition (50.4b) by replacing the bilinear forms a and b by some stabilized versions a_h and b_h; see Chapters 62 and 63. $\qquad\square$

We now establish an error estimate on u in a norm that is weaker than that in V. We do so by using a duality argument in the spirit of the Aubin–Nitsche lemma (Lemma 32.11).

Definition 50.10 (Smoothing property). *The problem* (50.1) *is said to have a* smoothing property *if there is a Hilbert space $H \hookrightarrow V$ with inner product $(\cdot, \cdot)_H$, two Banach spaces $Y \hookrightarrow V$ and $N \hookrightarrow Q$, and a constant c_{smo} such that the following adjoint problem:*

$$
\begin{cases}
\textit{Find } \varphi(g) \in V \textit{ and } \vartheta(g) \in Q \textit{ such that} \\
a(v, \varphi(g)) + b(v, \vartheta(g)) = (g, v)_H, & \forall v \in V, \\
b(\varphi(g), q) = 0, & \forall q \in Q,
\end{cases}
$$

has a unique solution for all $g \in H$ and satisfies the a priori estimate $\|\varphi(g)\|_Y + \|\vartheta(g)\|_N \leq c_{\mathrm{smo}}\|g\|_H$.

In addition to the smoothing property, we assume that the spaces H, Y, and N satisfy an additional approximation property, i.e., there are $s > 0$ and c such that the following holds true for all $(v, q) \in Y \times N$ and all $h \in \mathcal{H}$:

$$
\inf_{(v_h, q_h) \in V_h \times Q_h} (\|v - v_h\|_V + \|q - q_h\|_Q) \leq c\, h^s (\|v\|_Y + \|q\|_N). \tag{50.8}
$$

Lemma 50.11 (Improved error estimate in weaker norm). *Let (u, p) solve* (50.1)*. Assume* (50.4) *and let (u_h, p_h) solve the discrete problem* (50.2)*. Assume that* (50.1) *has a smoothing property and that* (50.8) *holds true. Then we have*

$$
\|u - u_h\|_H \leq c\, h^s (\|u - u_h\|_V + \|p - p_h\|_Q),
$$

where c is independent of (u, p), (u_h, p_h) and $h \in \mathcal{H}$.

Proof. Set $\mathcal{V} := V \times Q$, $\mathcal{Z} := Y \times N$, and $\mathcal{L} := H \times Q$, each equipped with the product norm. Define the symmetric positive bilinear form $l((v, q), (w, r)) := (v, w)_H$ and the seminorm $|(v, q)|_L := \|v\|_H$. Apply Lemma 32.11 in the conforming setting with the bilinear form $t((u, p), (v, q)) := a(u, v) + b(v, p) + b(u, q)$ to conclude. $\qquad\square$

50.2 Algebraic viewpoint

In this section, we study the linear system associated with the discrete problem (50.2) assuming that the well-posedness conditions (50.4a)-(50.4b) are satisfied. We also assume that the bilinear form a satisfies an inf-sup condition on $V_h \times V_h$. For simplicity, we consider real vector spaces.

50.2.1 The coupled linear system

Let $N := \dim(V_h)$ and $M := \dim(Q_h)$. Let $\{\varphi_i\}_{i \in \{1:N\}}$ be a basis for V_h and let $\{\psi_k\}_{k \in \{1:M\}}$ be a basis for Q_h. Recall that these bases consist of global shape functions when V_h and Q_h are finite element spaces. Proceeding as in §28.1.1, for every column vectors $\mathsf{U} = (\mathsf{U}_1, \ldots, \mathsf{U}_N)^\mathsf{T}$ in $\mathbb{R}^N$ and $\mathsf{P} = (\mathsf{P}_1, \ldots, \mathsf{P}_M)^\mathsf{T}$ in $\mathbb{R}^M$, we define the functions $\mathsf{R}_\varphi(\mathsf{U}) \in V_h$ and $\mathsf{R}_\psi(\mathsf{P}) \in Q_h$ by

$\mathsf{R}_\varphi(\mathsf{U}) := \sum_{i\in\{1:N\}} \mathsf{U}_i\varphi_i$ and $\mathsf{R}_\psi(\mathsf{P}) := \sum_{k\in\{1:M\}} \mathsf{P}_k\psi_k$. The correspondences between $\mathsf{R}_\varphi(\mathsf{U})$ and U and between $\mathsf{R}_\psi(\mathsf{P})$ and P are one-to-one since $\{\varphi_i\}_{i\in\{1:N\}}$ and $\{\psi_k\}_{k\in\{1:M\}}$ are bases.

Inserting the expansions of $\mathsf{R}_\varphi(\mathsf{U})$ and $\mathsf{R}_\psi(\mathsf{P})$ into (50.2) and choosing the basis functions of V_h and Q_h to test (50.2), we obtain the linear system

$$\mathcal{C}\begin{pmatrix}\mathsf{U}\\\mathsf{P}\end{pmatrix} = \begin{pmatrix}\mathsf{F}\\\mathsf{G}\end{pmatrix}, \qquad \mathcal{C} := \begin{pmatrix}\mathcal{A} & \mathcal{B}^\mathsf{T}\\\mathcal{B} & \mathbb{O}\end{pmatrix}, \tag{50.9}$$

where the matrices $\mathcal{A} \in \mathbb{R}^{N\times N}$ and $\mathcal{B} \in \mathbb{R}^{M\times N}$ are such that $\mathcal{A}_{ij} := a(\varphi_j,\varphi_i)$ and $\mathcal{B}_{ki} := b(\varphi_i,\psi_k)$ for all $i \in \{1:N\}$ and all $k \in \{1:M\}$, $\mathbb{O}$ is the zero matrix in $\mathbb{R}^{M\times M}$, and the vectors $\mathsf{F} \in \mathbb{R}^N$ and $\mathsf{G} \in \mathbb{R}^M$ are such that $\mathsf{F}_i = f(\varphi_i)$ and $\mathsf{G}_k = g(\psi_k)$ for all $i,j \in \{1:N\}$ and all $k \in \{1:M\}$.

The matrix $\mathcal{C}$ is invertible since (50.2) is well-posed owing to (50.4a)-(50.4b). Notice also that (50.4b) implies that $\mathcal{B}^\mathsf{T}$ has full column rank and $\mathcal{B}$ has full row rank (these ranks are equal to M). Moreover, $\mathcal{A}$ is invertible since we additionally assumed that a satisfies an inf-sup condition on $V_h\times V_h$. Algebraic counterparts of the boundedness and inf-sup conditions on the bilinear forms a and b can be established by using the dual norm

$$\|\mathsf{U}\|_{\ell^2_\varphi} := \sup_{\mathsf{Y}\in\mathbb{R}^N} \frac{\mathsf{U}^\mathsf{T}\mathsf{Y}}{\|\mathsf{R}_\varphi(\mathsf{Y})\|_V}, \qquad \forall\mathsf{U} \in \mathbb{R}^N. \tag{50.10}$$

Proposition 50.12 (Norm equivalence). *The following holds true:*

$$\alpha_h \|\mathsf{R}_\varphi(\mathsf{U})\|_V \le \|\mathcal{A}\mathsf{U}\|_{\ell^2_\varphi} \le \|a\|\,\|\mathsf{R}_\varphi(\mathsf{U})\|_V, \qquad \forall\mathsf{U} \in \mathbb{R}^N, \tag{50.11a}$$

$$\beta_h \|\mathsf{R}_\psi(\mathsf{P})\|_Q \le \|\mathcal{B}^\mathsf{T}\mathsf{P}\|_{\ell^2_\varphi} \le \|b\|\,\|\mathsf{R}_\psi(\mathsf{P})\|_Q, \qquad \forall\mathsf{P} \in \mathbb{R}^M. \tag{50.11b}$$

Proof. See Exercise 50.4. $\square$

50.2.2   Schur complement

Since the matrix $\mathcal{A}$ is invertible, the vector U can be eliminated from the linear system (50.9) yielding

$$\mathcal{S}\mathsf{P} = \mathcal{B}\mathcal{A}^{-1}\mathsf{F} - \mathsf{G}, \qquad \mathcal{S} := \mathcal{B}\mathcal{A}^{-1}\mathcal{B}^\mathsf{T}. \tag{50.12}$$

Once P is known, U is obtained by solving $\mathcal{A}\mathsf{U} = F - \mathcal{B}^\mathsf{T}\mathsf{P}$. The matrix $\mathcal{S} \in \mathbb{R}^{M\times M}$ (up to a sign convention) is called *Schur complement* of $\mathcal{A}$; see §49.2.1 for the infinite-dimensional counterpart. Notice that the matrix $\mathcal{S}$ is invertible (if $\mathcal{S}\mathsf{P} = 0$, setting $\mathsf{U} := -\mathcal{A}^{-1}\mathcal{B}^\mathsf{T}\mathsf{P}$, we infer that $\mathcal{C}(\mathsf{U},\mathsf{P})^\mathsf{T} = (0,0)^\mathsf{T}$, and $\mathcal{C}$ being invertible, this implies that $\mathsf{U} = 0$ and $\mathsf{P} = 0$).

Additional properties of the Schur complement matrix $\mathcal{S}$ are available when the bilinear form a is symmetric and coercive, since in this case the matrix $\mathcal{A}$ is symmetric positive definite.

Proposition 50.13 (Symmetry and positivity of $\mathcal{S}$). *If $\mathcal{A}$ is symmetric positive definite, so is $\mathcal{S}$.*

Proof. The definition of $\mathcal{S}$ implies that $\mathcal{S}^\mathsf{T} = \mathcal{B}(\mathcal{A}^{-1})^\mathsf{T}\mathcal{B}^\mathsf{T}$, but $(\mathcal{A}^{-1})^\mathsf{T} = (\mathcal{A}^\mathsf{T})^{-1}$. Hence, $\mathcal{S}$ is symmetric if $\mathcal{A}$ is symmetric. Let now $\mathsf{P} \in \mathbb{R}^M$. Then $\mathsf{P}^\mathsf{T}\mathcal{S}\mathsf{P} = \mathsf{P}^\mathsf{T}\mathcal{B}\mathcal{A}^{-1}\mathcal{B}^\mathsf{T}\mathsf{P} = (\mathcal{B}^\mathsf{T}\mathsf{P})^\mathsf{T}\mathcal{A}^{-1}\mathcal{B}^\mathsf{T}\mathsf{P} \ge 0$. This proves that $\mathcal{S}$ is positive semidefinite. Moreover, $\mathcal{S}\mathsf{P} = 0$ implies that $\mathcal{B}^\mathsf{T}\mathsf{P} = 0$, so that $\mathsf{P} = 0$ since $\mathcal{B}^\mathsf{T}$ has full column rank. Hence, $\mathcal{S}$ is positive definite. $\square$

Note that even if $\mathcal{A}$ is symmetric positive definite, the matrix $\mathcal{C}$ is symmetric but indefinite. Observing that

$$\mathcal{C} = \begin{pmatrix}\mathcal{I}_N & \mathbb{O}_{N,M}\\\mathcal{B}\mathcal{A}^{-1} & \mathcal{I}_M\end{pmatrix}\begin{pmatrix}\mathcal{A} & \mathbb{O}_{N,M}\\\mathbb{O}_{M,N} & -\mathcal{S}\end{pmatrix}\begin{pmatrix}\mathcal{I}_N & \mathcal{A}^{-1}\mathcal{B}^\mathsf{T}\\\mathbb{O}_{N,M} & \mathcal{I}_M\end{pmatrix},$$

we infer from the Sylvester Law of Inertia (stating that two symmetric matrices $\mathcal{C}$ and $\mathcal{C}'$ satisfying $\mathcal{C} = \mathcal{P}\mathcal{C}'\mathcal{P}^\mathsf{T}$ with $\mathcal{P}$ invertible have the same number of positive, zero, and negative eigenvalues; see Golub and van Loan [218, p. 403]) that $\mathcal{C}$ has N positive eigenvalues and M negative ones. Upper and lower bounds on the clusters of positive and negative eigenvalues of $\mathcal{C}$ are derived in Rusten and Winther [338], Wathen and Silvester [390]. In practice, the matrix $\mathcal{C}$ can be very poorly conditioned. We return to this issue in §50.3.2. Note that changing the lower-left block of $\mathcal{C}$ into $-\mathcal{B}$ produces a positive semidefinite, but nonsymmetric, matrix.

Let us now examine more closely the eigenvalues of $\mathcal{S}$ (see Verfürth [375]). To this purpose, let $\mathcal{M}_Q \in \mathbb{R}^{M \times M}$ be the matrix with entries $\mathcal{M}_{Q,kl} := (\psi_k, \psi_l)_Q$ for all $k, l \in \{1{:}M\}$. The matrix $\mathcal{M}_Q$ is symmetric by construction, and the identity $\mathsf{P}^\mathsf{T}\mathcal{M}_Q\mathsf{P} = (\mathsf{R}_\psi(\mathsf{P}), \mathsf{R}_\psi(\mathsf{P}))_Q = \|\mathsf{R}_\psi(\mathsf{P})\|_Q^2$ for all $\mathsf{P} \in \mathbb{R}^M$ shows that $\mathcal{M}_Q$ is positive definite. Since Q is the L^2-space in many applications, the matrix $\mathcal{M}_Q$ is called mass matrix (see §28.1.1). Let $\mu_{\min}$ and $\mu_{\max}$ be the lowest and largest eigenvalues of $\mathcal{M}_Q$. Recall from §28.2.1 that the (Euclidean) condition number $\kappa(\mathcal{Z})$ of a symmetric invertible matrix $\mathcal{Z}$ is the ratio of the largest to the smallest eigenvalues of $\mathcal{Z}$ in absolute value.

Proposition 50.14 (Spectrum of $\mathcal{S}$). *Assume that the bilinear form a is symmetric and coercive on V_h with constant α_h and that the inf-sup condition (50.4b) for b is satisfied with constant β_h. Then the matrices $\mathcal{S}$ and $\mathcal{M}_Q$ are spectrally equivalent, i.e., the following holds true for all $\mathsf{P} \in \mathbb{R}^M$:*

$$\frac{\beta_h^2}{\|a\|} \le \frac{\mathsf{P}^\mathsf{T}\mathcal{S}\mathsf{P}}{\mathsf{P}^\mathsf{T}\mathcal{M}_Q\mathsf{P}} \le \frac{\|b\|^2}{\alpha_h}. \tag{50.13}$$

Moreover, $\sigma(\mathcal{M}_Q^{-1}\mathcal{S}) \subset \left[\frac{\beta_h^2}{\|a\|}, \frac{\|b\|^2}{\alpha_h}\right]$, and $\sigma(\mathcal{S}) \subset \left[\mu_{\min}\frac{\beta_h^2}{\|a\|}, \mu_{\max}\frac{\|b\|^2}{\alpha_h}\right]$, which implies that $\kappa(\mathcal{M}_Q^{-1}\mathcal{S}) \le \frac{\|a\|}{\alpha_h}\left(\frac{\|b\|}{\beta_h}\right)^2$ and $\kappa(\mathcal{S}) \le \frac{\|a\|}{\alpha_h}\left(\frac{\|b\|}{\beta_h}\right)^2 \kappa(\mathcal{M}_Q)$.

Proof. (1) Proof of (50.13). For all $\mathsf{P} \in \mathbb{R}^M$, we observe that

$$\sup_{\mathsf{Y}\in\mathbb{R}^N} \frac{(\mathcal{B}^\mathsf{T}\mathsf{P})^\mathsf{T}\mathsf{Y}}{(\mathsf{Y}^\mathsf{T}\mathcal{A}\mathsf{Y})^{\frac{1}{2}}} = \sup_{\mathsf{Y}\in\mathbb{R}^N} \frac{(\mathcal{B}^\mathsf{T}\mathsf{P})^\mathsf{T}\mathcal{A}^{-\frac{1}{2}}\mathsf{Y}}{\|\mathsf{Y}\|_{\ell^2(\mathbb{R}^N)}} = \sup_{\mathsf{Y}\in\mathbb{R}^N} \frac{(\mathcal{A}^{-\frac{1}{2}}\mathcal{B}^\mathsf{T}\mathsf{P})^\mathsf{T}\mathsf{Y}}{\|\mathsf{Y}\|_{\ell^2(\mathbb{R}^N)}}$$
$$= \|\mathcal{A}^{-\frac{1}{2}}\mathcal{B}^\mathsf{T}\mathsf{P}\|_{\ell^2(\mathbb{R}^N)} = (\mathsf{P}^\mathsf{T}\mathcal{S}\mathsf{P})^{\frac{1}{2}},$$

since $\mathcal{A}$ is symmetric positive definite. Observing that $\frac{1}{\|a\|} \le \frac{\|\mathsf{R}_\varphi(\mathsf{Y})\|_V^2}{\mathsf{Y}^\mathsf{T}\mathcal{A}\mathsf{Y}} \le \frac{1}{\alpha_h}$ for all $\mathsf{Y} \in \mathbb{R}^N$, we infer that

$$\frac{1}{\|a\|}\|\mathcal{B}^\mathsf{T}\mathsf{P}\|_{\ell_\varphi^2}^2 \le \mathsf{P}^\mathsf{T}\mathcal{S}\mathsf{P} \le \frac{1}{\alpha_h}\|\mathcal{B}^\mathsf{T}\mathsf{P}\|_{\ell_\varphi^2}^2.$$

Finally, (50.13) follows from (50.11b) using $\mathsf{P}^\mathsf{T}\mathcal{M}_Q\mathsf{P} = \|\mathsf{R}_\psi(\mathsf{P})\|_Q^2$.
(2) The spectrum and condition number for $\mathcal{M}_Q^{-1}\mathcal{S}$ readily follow from (50.13), and the results for $\mathcal{S}$ follow from the fact that $\mu_{\min}\|\mathsf{P}\|_{\ell^2(\mathbb{R}^M)}^2 \le \mathsf{P}^\mathsf{T}\mathcal{M}_Q\mathsf{P} \le \mu_{\max}\|\mathsf{P}\|_{\ell^2(\mathbb{R}^M)}^2$ for all $\mathsf{P} \in \mathbb{R}^M$. $\qquad\square$

50.2.3 Augmented Lagrangian for saddle point problems

Assume that the matrix $\mathcal{A}$ is symmetric positive definite and that $\mathcal{B}^\mathsf{T}$ has full column rank. Referring to §49.3.2 for the infinite-dimensional setting we infer that the pair (U, P) solves the linear system (50.9) iff it is a saddle point of the Lagrangian

$$\mathcal{L}(\mathsf{Y}, \mathsf{R}) := \frac{1}{2}\mathsf{Y}^\mathsf{T}\mathcal{A}\mathsf{Y} - \mathsf{F}^\mathsf{T}\mathsf{Y} + \mathsf{R}^\mathsf{T}(\mathcal{B}\mathsf{Y} - \mathsf{G}). \tag{50.14}$$

Recall that

$$\inf_{\mathsf{Y}\in\mathbb{R}^N} \sup_{\mathsf{R}\in\mathbb{R}^M} \mathcal{L}(\mathsf{Y},\mathsf{R}) = \mathcal{L}(\mathsf{U},\mathsf{P}) = \sup_{\mathsf{R}\in\mathbb{R}^M} \inf_{\mathsf{Y}\in\mathbb{R}^N} \mathcal{L}(\mathsf{Y},\mathsf{R}). \tag{50.15}$$

The optimization problem on the left-hand side of (50.15) amounts to minimizing the convex energy functional $\mathfrak{E}(\mathsf{Y}) := \frac{1}{2}\mathsf{Y}^\mathsf{T}\mathcal{A}\mathsf{Y} - \mathsf{F}^\mathsf{T}\mathsf{Y}$ over the affine subspace $\{\mathsf{Y} \in \mathbb{R}^N \mid B\mathsf{Y} = \mathsf{G}\}$ since $\sup_{\mathsf{R}\in\mathbb{R}^M} \mathcal{L}(\mathsf{Y},\mathsf{R}) = \infty$ if $B\mathsf{Y} \neq \mathsf{G}$. Consider now the optimization problem on the right-hand side of (50.15). The minimization of $\mathcal{L}(\mathsf{Y},\mathsf{R})$ over $\mathsf{Y} \in \mathbb{R}^N$ leads to the optimal solution $\mathsf{Y}_* := \mathcal{A}^{-1}(\mathsf{F} - \mathcal{B}^\mathsf{T}\mathsf{R})$, and we are left with maximizing the following concave functional over $\mathbb{R}^M$:

$$\mathsf{R} \mapsto \mathcal{L}(\mathsf{Y}_*,\mathsf{R}) = -\frac{1}{2}\mathsf{R}^\mathsf{T}\mathcal{S}\mathsf{R} + (\mathcal{B}\mathcal{A}^{-1}\mathsf{F} - \mathsf{G})^\mathsf{T}\mathsf{R} - \frac{1}{2}\mathsf{F}^\mathsf{T}\mathcal{A}^{-1}\mathsf{F},$$

where $\mathcal{S} := \mathcal{B}\mathcal{A}^{-1}\mathcal{B}^\mathsf{T}$. The optimal solution to this maximization problem solves $\mathcal{S}\mathsf{R} = \mathcal{B}\mathcal{A}^{-1}\mathsf{F} - \mathsf{G}$, i.e., we recover the Schur complement system (50.12).

The main idea of augmented Lagrangian methods (see Fortin and Glowinski [203]) is to add to the Lagrangian a least-squares penalty on the constraint. Specifically, letting $\rho > 0$ be a real parameter and recalling the mass matrix $\mathcal{M}_Q \in \mathbb{R}^{M \times M}$, the *augmented Lagrangian* is defined as

$$\mathcal{L}_\rho(\mathsf{Y},\mathsf{R}) := \mathcal{L}(\mathsf{Y},\mathsf{R}) + \frac{\rho}{2}(\mathcal{B}\mathsf{Y} - \mathsf{G})^\mathsf{T}\mathcal{M}_Q^{-1}(\mathcal{B}\mathsf{Y} - \mathsf{G}).$$

Since we also have $B\mathsf{U} = \mathsf{G}$, the solution to (50.9) is also the unique saddle point of the augmented Lagrangian $\mathcal{L}_\rho$, i.e., (U,P) can be found by solving the following linear system:

$$\begin{pmatrix} \mathcal{A}_\rho & \mathcal{B}^\mathsf{T} \\ \mathcal{B} & \mathbb{0} \end{pmatrix} \begin{pmatrix} \mathsf{U} \\ \mathsf{P} \end{pmatrix} = \begin{pmatrix} \mathsf{F}_\rho \\ \mathsf{G} \end{pmatrix}, \qquad \begin{aligned} \mathcal{A}_\rho &:= \mathcal{A} + \rho\mathcal{B}^\mathsf{T}\mathcal{M}_Q^{-1}\mathcal{B}, \\ \mathsf{F}_\rho &:= \mathsf{F} + \rho\mathcal{B}^\mathsf{T}\mathcal{M}_Q^{-1}\mathsf{G}. \end{aligned} \tag{50.16}$$

The augmented Schur complement is defined as $\mathcal{S}_\rho := \mathcal{B}\mathcal{A}_\rho^{-1}\mathcal{B}^\mathsf{T}$. Recall that $\sigma(\mathcal{M}_Q^{-1}\mathcal{S}) \subset [s_\flat, s_\sharp]$, with $s_\flat := \frac{\beta_h^2}{\|a\|}$ and $s_\sharp := \frac{\|b\|^2}{\alpha_h}$.

Proposition 50.15 (Spectrum of $\mathcal{S}_\rho$). *The following holds true:*

$$\mathcal{S}_\rho^{-1} = \rho\mathcal{M}_Q^{-1} + \mathcal{S}^{-1}, \tag{50.17}$$

and $\sigma(\mathcal{M}_Q^{-1}\mathcal{S}_\rho) \subset [(\rho + s_\flat^{-1})^{-1}, (\rho + s_\sharp^{-1})^{-1}]$ *and* $\kappa(\mathcal{M}_Q^{-1}\mathcal{S}_\rho) \leq \frac{\rho + s_\flat^{-1}}{\rho + s_\sharp^{-1}}$.

Proof. See Exercise 50.5 for the proof of (50.17). The properties on the spectrum and the condition number of $\mathcal{S}_\rho$ follow readily.   □

Remark 50.16 (Value of ρ). Proposition 50.15 shows that taking $\rho \gg 1$ improves the condition number of the Schur complement $\mathcal{S}_\rho$. A large value of ρ however deteriorates the conditioning of the matrix $\mathcal{A}_\rho$ which makes it more difficult to invert iteratively. In practice, it is necessary to strike a balance between these two criteria.   □

Remark 50.17 (Hilbert setting). The notion of augmented Lagrangian can be extended to the infinite-dimensional setting. The mass matrix $\mathcal{M}_Q$ is then replaced by the Riesz–Fréchet isomorphism $J_Q : Q \to Q'$.   □

The augmented Lagrangian technique is in general preferable to the following unconstrained penalty method:

$$\begin{pmatrix} \mathcal{A} & \mathcal{B}^\mathsf{T} \\ \mathcal{B} & -\epsilon\mathcal{M}_Q \end{pmatrix} \begin{pmatrix} \mathsf{U}_\epsilon \\ \mathsf{P}_\epsilon \end{pmatrix} = \begin{pmatrix} \mathsf{F} \\ \mathsf{G} \end{pmatrix}, \tag{50.18}$$

where $\epsilon > 0$ is a small parameter. This technique is often referred to as artificial compressibility in the fluid mechanics literature. Eliminating P_ϵ from the first equation yields

$$\mathcal{A}_{\frac{1}{\epsilon}} U_\epsilon = F + \epsilon^{-1} \mathcal{B}^\mathsf{T} \mathcal{M}_Q^{-1} G. \tag{50.19}$$

The advantage of (50.19) with respect to (50.9) (or to (50.16)) is that the system matrix (i.e., $\mathcal{A}_{\frac{1}{\epsilon}}$) is symmetric positive definite. The solution (U_ϵ, P_ϵ) however differs from (U, P). In particular, U fails to satisfy the constraint $\mathcal{B}U = G$, although the difference $(U - U_\epsilon, P - P_\epsilon)$ tends to zero as $\epsilon \to 0$. Unfortunately, taking $\epsilon \ll 1$ makes the linear system (50.19) ill-conditioned.

Proposition 50.18 (Penalty). *Let $\epsilon > 0$. Let (U, P) solve (50.9) and (U_ϵ, P_ϵ) solve (50.18). The following holds true:*

$$\tfrac{\alpha_h \beta_h}{\|a\|} \|R_\varphi(U - U_\epsilon)\|_V + \tfrac{\alpha_h \beta_h^2}{\|a\|^2} \|R_\psi(P - P_\epsilon)\|_Q \leq \epsilon \|R_\psi(P)\|_Q. \tag{50.20}$$

Proof. See Exercise 50.6.   □

50.3   Iterative solvers

In this section, we discuss iterative solvers for the linear system (50.9) (or its augmented Lagrangian version (50.16)). First, we discuss the Uzawa algorithm as an example of a technique based on stationary iterations. Then, we present more efficient techniques based on preconditioned Krylov subspaces. We assume that the matrix $\mathcal{A}$ is invertible and that the matrix $\mathcal{B}^\mathsf{T}$ has full column rank.

50.3.1   Uzawa algorithm

The Uzawa algorithm is an iterative method where U and P are updated one after the other. Given $P_0 \in \mathbb{R}^M$ and a parameter $\eta > 0$, the algorithm consists of constructing the iterates (U_m, P_m) for $m = 1, 2, \dots$ as follows:

$$\mathcal{A} U_m = F - \mathcal{B}^\mathsf{T} P_{m-1}, \tag{50.21a}$$
$$\mathcal{M}_Q P_m = \mathcal{M}_Q P_{m-1} + \eta(\mathcal{B} U_m - G). \tag{50.21b}$$

This makes sense since $\mathcal{A}$ and $\mathcal{M}_Q$ are invertible. Eliminating U_m gives

$$\mathcal{M}_Q P_m = \mathcal{M}_Q P_{m-1} - \eta(\mathcal{S} P_{m-1} - \mathcal{B} \mathcal{A}^{-1} F + G). \tag{50.22}$$

In other words, the Uzawa algorithm is equivalent to the Richardson iteration applied to the linear system (50.12) left-preconditioned by the mass matrix $\mathcal{M}_Q$. (Recall that for a generic linear system $\mathcal{Z}X = Y$, the Richardson iteration reads $X_m = X_{m-1} + \eta(\mathcal{Z}X_{m-1} - Y)$.) If $\mathcal{A}$ is symmetric positive definite, we can use the bounds on the spectrum of $\mathcal{M}_Q^{-1}\mathcal{S}$ from Proposition 50.14, that is, $s_\flat := \frac{\beta_h^2}{\|a\|} \leq \mathcal{M}_Q^{-1}\mathcal{S} \leq \frac{\|b\|^2}{\alpha_h} := s_\sharp$ in the sense of quadratic forms. We then infer that the Richardson iteration (50.22) converges geometrically provided we take $0 < \eta < \frac{2}{s_\sharp}$, and the error reduction factor is maximized by taking the optimal value $\eta_{\text{opt}} := \frac{2}{s_\flat + s_\sharp}$; see Saad [339, p. 106]. It is often easier to estimate $s_\sharp$ than $s_\flat$ since β_h is more difficult to estimate than α_h.

Remark 50.19 (Implementation). The matrices $\mathcal{A}$ and $\mathcal{B}$ are sparse (see §29.1), but $\mathcal{S}$ is a dense matrix owing to the presence of $\mathcal{A}^{-1}$ in the definition of $\mathcal{S}$. Since precomputing $\mathcal{A}^{-1}$ is generally too expensive, an inner iteration has to be employed to compute the action of $\mathcal{A}^{-1}$ on vectors in $\mathbb{R}^N$. The matrix $\mathcal{B}$ can be assembled and stored once and for all, or its action on a given vector in $\mathbb{R}^M$ can be computed on the fly whenever needed. In practice, one must often find a compromise between many (often conflicting) criteria: the memory space available; the number of times the linear system has to be solved; the ratio between the speed to access memory and the speed to perform floating point operations; parallelization; etc. □

Remark 50.20 (Variants). The mass matrix $\mathcal{M}_Q$ can be replaced by the identity matrix $\mathcal{I}_M$ in (50.21b). The advantage is that this avoids computing the inverse of the mass matrix (although this matrix is generally easy to invert since it is well-conditioned). The drawback is that the choice of the relaxation parameter η now depends on the spectrum of the unpreconditioned Schur complement matrix, which requires some information on the (mesh-dependent) spectrum of $\mathcal{M}_Q$. Another variant is to consider an approximate inverse of $\mathcal{A}$ that is easy to compute, say $\mathcal{H}$, and to replace (50.21a) by $\mathsf{U}_m = \mathsf{U}_{m-1} + \mathcal{H}(\mathsf{F} - \mathcal{A}\mathsf{U}_{m-1} - \mathcal{B}^\mathsf{T}\mathsf{P}_{m-1})$ leading to an inexact Uzawa algorithm; see Bacuta [41] for a convergence analysis. □

50.3.2 Krylov subspace methods

Krylov subspace methods for solving (preconditioned) linear systems of the form (50.9) or variations thereof constitute an active area of research. In this section, we sketch a few important ideas and refer to Benzi et al. [52, §9] for a broader treatment and to Elman et al. [185, Chap. 6&8], Turek [366] for applications to fluid mechanics.

In the context of saddle point problems, the matrix $\mathcal{C}$ in (50.9) is symmetric, but indefinite (recall that the matrix $\mathcal{A}$ is symmetric positive definite by assumption). In this case, MINRES is a method of choice to solve (50.9); see [185, p. 289]. The attractive feature of MINRES is that it achieves an optimality property on the residual while employing only short-term recurrences. Specifically, at step $m \geq 1$, the iterate $\mathsf{X}_m \in \mathbb{R}^{N+M}$ with residual $\mathsf{R}_m := (\mathsf{F}, \mathsf{G})^\mathsf{T} - \mathcal{C}\mathsf{X}_m$ satisfies the following optimality property (compare with Proposition 28.20 for the conjugate gradient method applied to symmetric positive definite linear systems):

$$\|\mathsf{R}_m\|_{\ell^2(\mathbb{R}^{N+M})} = \min_{\mathsf{Y} \in \mathsf{U}_0 + K_m} \|(\mathsf{F}, \mathsf{G})^\mathsf{T} - \mathcal{C}\mathsf{Y}\|_{\ell^2(\mathbb{R}^{N+M})}, \qquad (50.23)$$

with the Krylov subspace $K_m := \mathrm{span}\{\mathsf{R}_0, \mathcal{C}\mathsf{R}_0, \ldots, \mathcal{C}^{m-1}\mathsf{R}_0\}$. The convergence rate of MINRES depends on the spectrum of $\mathcal{C}$. More precisely, defining $\tilde{c}_m := \min_{p \in \mathbb{P}_m, p(0)=1} \max_{\lambda \in \sigma(\mathcal{C})} |p(\lambda)|$, one can prove that $\|\mathsf{R}_m\|_{\ell^2(\mathbb{R}^{N+M})} \leq \tilde{c}_m \|\mathsf{R}_0\|_{\ell^2(\mathbb{R}^{N+M})}$ (this bound is sharp). The constant $\tilde{c}_m$ can be estimated under the assumption that $\sigma(\mathcal{C}) \subset [-a, -b] \cup [c, d]$ with positive real numbers a, b, c, d such that the two intervals have the same length (i.e., $d - c = a - b$). One can show that (see Greenbaum [221, Chap. 3])

$$\|\mathsf{R}_{2m}\|_{\ell^2(\mathbb{R}^{N+M})} \leq 2 \left(\frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}} \right)^m \|\mathsf{R}_0\|_{\ell^2(\mathbb{R}^{N+M})}. \qquad (50.24)$$

The minimization property of MINRES implies that $\|\mathsf{R}_{2m+1}\|_{\ell^2(\mathbb{R}^{N+M})} \leq \|\mathsf{R}_{2m}\|_{\ell^2(\mathbb{R}^{N+M})}$, but it is possible that no reduction of the norm of the residual occurs in every other step, leading to a staircasing behavior of the iterates. A comparison of (28.23) with (50.24) shows that MINRES requires twice as many iterations as the Conjugate Gradient to reach a given threshold for a symmetric positive definite matrix with condition number κ^2. Hence, solving linear systems like (50.9) is a

significant computational challenge, and preconditioning is essential. Before addressing this question let us observe that MINRES is bound to fail if $\mathcal{A}$ is not symmetric, since symmetry is essential for MINRES to work properly. This happens, for instance, in fluid mechanics when solving the Oseen or (linearized) Navier–Stokes equations. One alternative is to use the GMRES method which retains an optimality property over the Krylov subspace at the price of storing a complete basis thereof; see Saad [339, §6.5] for a thorough description.

   Preconditioning is a very important ingredient of Krylov subspace methods, especially for linear systems of the form (50.9). Here, we only discuss block preconditioners and refer the reader to [52, §10] and references therein for further insight into preconditioned Krylov methods. Block diagonal and triangular preconditioners are, respectively, of the form

$$\mathcal{P}_{\mathrm{d}} := \begin{pmatrix} \hat{\mathcal{A}} & \mathbb{O} \\ \mathbb{O} & \hat{\mathcal{S}} \end{pmatrix}, \qquad \mathcal{P}_{\mathrm{t}} := \begin{pmatrix} \hat{\mathcal{A}} & \mathcal{B}^{\mathsf{T}} \\ \mathbb{O} & \hat{\mathcal{S}} \end{pmatrix}, \tag{50.25}$$

where $\hat{\mathcal{A}}$ and $\hat{\mathcal{S}}$ are easy-to-invert approximations of $\mathcal{A}$ and $\mathcal{S}$. In the ideal case where $\hat{\mathcal{A}} := \mathcal{A}$ and $\hat{\mathcal{S}} := \mathcal{S}$, a direct calculation shows that the left-preconditioned matrices $\mathcal{P}_{\mathrm{d}}^{-1}\mathcal{C}$ and $\mathcal{P}_{\mathrm{t}}^{-1}\mathcal{C}$ are zeroes of the polynomials $p_{\mathrm{d}}(\lambda) := (\lambda - 1)(\lambda - \frac{1 \pm \sqrt{5}}{2})$ and $p_{\mathrm{t}}(\lambda) := \lambda^2 - 1$, respectively (see Kuznetsov [272], Murphy et al. [309]; see also (49.22) and Theorem 49.6), implying convergence in at most three (resp., two) steps for every preconditioned Krylov subspace method. The block triangular preconditioner $\mathcal{P}_{\mathrm{t}}$ breaks the symmetry of the system even if $\mathcal{A}$ is symmetric, but this preconditioner is still quite effective in many cases, particularly for Oseen and Navier–Stokes flows (where the convective term breaks the symmetry of $\mathcal{A}$ anyway). Note also that the costs of the two preconditioners in (50.25) are essentially identical since the cost of the additional multiplication by $\mathcal{B}^{\mathsf{T}}$ is often marginal.

   Effective choices for $\hat{\mathcal{A}}$ and $\hat{\mathcal{S}}$ are often driven by the application at hand. For Darcy's and Maxwell's equations (see Examples in §49.1.1 and §49.1.3), $\mathcal{A}$ represents a zeroth-order differential operator (multiplication by a material property), and choosing a diagonal lumping for $\hat{\mathcal{A}}$ together with some multilevel technique for $\hat{\mathcal{S}}$ often works well if the material coefficients are smooth (see, e.g., Perugia and Simoncini [324] for magnetostatics problems). For the Stokes equations (see §49.1.2), $\mathcal{A}$ represents a second-order differential operator, and the preconditioner $\hat{\mathcal{A}}$ is typically based on some multilevel technique. The mass matrix associated with p can be used for $\hat{\mathcal{S}}$ and a detailed eigenvalue analysis of the resulting block-diagonal preconditioned system can be found in Silvester and Wathen [347]. The approximation of the Schur complement becomes more delicate in the unsteady case and in the presence of convection. Preconditioners devised from the structure of the steady Navier–Stokes equations can be found in Elman et al. [185, Chap. 8] and the references therein. Furthermore, an attractive idea for transient and high-Reynolds number flows is to consider a block triangular preconditioner based on the augmented Lagrangian formulation (50.16) for the $(1,1)$-block, together with the (scaled) mass matrix for the $(2,2)$-block (thereby avoiding to consider the Schur complement); see Benzi and Olshanskii [51], Benzi et al. [53].

Exercises

Exercise 50.1 (Algebraic setting). Let $\mathcal{A} := \begin{pmatrix} 1 & \sqrt{2} \\ \sqrt{2} & 0 \end{pmatrix}$ and $\mathcal{B} := (1, 0)^{\mathsf{T}}$. Show that

$$\inf_{\mathsf{V} \in \ker(\mathcal{B})} \sup_{\mathsf{W} \in \ker(\mathcal{B})} \frac{\mathsf{W}^{\mathsf{T}} \mathcal{A} \mathsf{V}}{\|\mathsf{W}\|_{\ell^2(\mathbb{R}^2)} \|\mathsf{V}\|_{\ell^2(\mathbb{R}^2)}} < \inf_{\mathsf{V} \in \mathbb{R}^2} \sup_{\mathsf{W} \in \mathbb{R}^2} \frac{\mathsf{W}^{\mathsf{T}} \mathcal{A} \mathsf{V}}{\|\mathsf{W}\|_{\ell^2(\mathbb{R}^2)} \|\mathsf{V}\|_{\ell^2(\mathbb{R}^2)}}.$$

(*Hint*: one number is equal to 0 and the other is equal to 1.)

Exercise 50.2 (Saddle point problem). Let V, Q be Hilbert spaces and let a be a symmetric, coercive, bilinear form. Consider the discrete problem (50.2) and the bilinear form $t(y, z) := a(v, w) + b(w, q) + b(v, r)$ for all $y := (v, q), z := (w, r) \in X := V \times Q$. Let $X_h := V_h \times Q_h$ and consider the linear map $P_h \in \mathcal{L}(X; X_h)$ such that for all $x \in X$, $P_h(x) \in X_h$ is the unique solution of $t(P_h(x), y_h) = t(x, y_h)$ for all $y_h \in X_h$. Equip X and X_h with the norm $\|(v, q)\|_{\widetilde{X}} := (\|v\|_a^2 + \|q\|_Q^2)^{\frac{1}{2}}$ with $\|v\|_a^2 := a(v, v)$. (i) Prove that $\|P_h\|_{\mathcal{L}(X;X)} \leq \tilde{c}_h := \frac{(4\frac{\|b\|^2}{\alpha} + 1)^{\frac{1}{2}} + 1}{(4\frac{\beta_h^2}{\|a\|} + 1)^{\frac{1}{2}} - 1}$. (*Hint*: use Proposition 49.8.) (ii) Prove that $\|u - u_h\|_a^2 + \|p - p_h\|_Q^2 \leq \tilde{c}_h^2 (\inf_{v_h \in V_h} \|u - u_h\|_a^2 + \inf_{q_h \in Q_h} \|p - q_h\|_Q^2)$. (*Hint*: see the proof of Theorem 5.14.)

Exercise 50.3 (Error estimate). (i) Prove directly the estimate (50.7a) with c'_{1h} replaced by $c''_{1h} := (1 + \frac{\|a\|}{\alpha_h})(1 + \frac{\|b\|}{\beta_h})$. (*Hint*: consider $z_h \in V_h$ s.t. $B_h(z_h) := B_h(u_h - v_h)$ with $v_h \in V_h$ arbitrary.) (ii) Assume that V is a Hilbert space, $\ker(B_h) \subset \ker(B)$, and $g := 0$. Prove that $\|u - u_h\|_V \leq \frac{\|a\|}{\alpha_h} \inf_{v_h \in \ker(B_h)} \|u - v_h\|_V$.

Exercise 50.4 (Bound on $\mathcal{A}$ and $\mathcal{B}$). (i) Prove Proposition 50.12. (*Hint*: observe that $(\mathcal{A}\mathsf{U})^\mathsf{T}\mathsf{Y} = a(\mathsf{R}_\varphi(\mathsf{U}), \mathsf{R}_\varphi(\mathsf{Y}))$.) (ii) Let $\mathcal{J}_V \in \mathbb{R}^{N \times N}$ be the symmetric positive definite matrix with entries $\mathcal{J}_{V,ij} := (\varphi_i, \varphi_j)_X$ for all $i, j \in \{1:N\}$. Let $\|\cdot\|_{\ell^2(\mathbb{R}^N)}$ denote the Euclidean norm in $\mathbb{R}^N$. Verify that $\|\mathsf{R}_\varphi(\mathsf{U})\|_V = \|\mathcal{J}_V^{\frac{1}{2}}\mathsf{U}\|_{\ell^2(\mathbb{R}^N)}$ and $\|\mathsf{U}\|_{\ell^2(\mathbb{R}^N)} = \|\mathcal{J}_V^{-\frac{1}{2}}\mathsf{U}\|_{\ell^2(\mathbb{R}^N)}$ for all $\mathsf{U} \in \mathbb{R}^N$.

Exercise 50.5 ($\mathcal{S}_\rho$). The goal is to prove the identity (50.17). (i) Verify that $\mathcal{A}_\rho^{-1} = \mathcal{A}^{-1} - \rho\mathcal{A}^{-1}\mathcal{B}^\mathsf{T}(\mathcal{M}_Q + \rho\mathcal{S})^{-1}\mathcal{B}\mathcal{A}^{-1}$. (*Hint*: multiply the right-hand side by $\mathcal{A}_\rho$ and develop the product.) (ii) Infer that $\mathcal{S}_\rho = \mathcal{S} - \rho\mathcal{S}(\mathcal{M}_Q + \rho\mathcal{S})^{-1}\mathcal{S}$. (iii) Conclude. (*Hint*: multiply the right-hand side by $\rho\mathcal{M}_Q^{-1} + \mathcal{S}^{-1}$.)

Exercise 50.6 (Penalty). (i) Prove Proposition 50.18. (*Hint*: verify that $\mathcal{C}(\mathsf{U} - \mathsf{U}_\epsilon, \mathsf{P} - \mathsf{P}_\epsilon)^\mathsf{T} = (0, -\epsilon\mathcal{M}_Q\mathsf{P}_\epsilon)^\mathsf{T}$ and use Proposition 50.12.) (ii) Replace the mass matrix $\mathcal{M}_Q$ by the identity matrix $\mathcal{I}_M$ times a positive coefficient λ in (50.18). Does the method still converge? Is there any interest of doing so? Can you think of another choice?

Exercise 50.7 (Inexact Minres and DPG). Let V, Y be Hilbert spaces and $B \in \mathcal{L}(V; Y')$ be s.t. $\beta\|v\|_V \leq \|B(v)\|_{Y'} \leq \|b\|\|v\|_V$ for all $v \in V$ with $0 < \beta \leq \|b\| < \infty$. Set $b(v, y) := \langle B(v), y \rangle_{Y',Y}$. Let $f \in Y'$. Let $J_Y : Y \to Y'$ denote the isometric Riesz–Fréchet isomorphism. (i) Show that the Minres problem $\min_{v \in V} \|f - B(v)\|_{Y'}$ has a unique solution $u \in V$. (*Hint*: introduce the sesquilinear form $a(v, w) := \langle B(v), J_Y^{-1}(B(w)) \rangle_{Y',Y}$ and invoke the Lax–Milgram Lemma.) (ii) Let $\{V_h \subset V\}_{h \in \mathcal{H}}$ and $\{Y_h \subset Y\}_{h \in \mathcal{H}}$ be sequences of subspaces approximating V and Y, respectively. Assume that there is $\beta_0 > 0$ s.t. for all $h \in \mathcal{H}$,

$$\inf_{v_h \in V_h} \sup_{y_h \in Y_h} \frac{|b(v_h, y_h)|}{\|v_h\|_V \|y_h\|_Y} \geq \beta_0. \tag{50.26}$$

Let $I_h : Y_h \to Y$ be the canonical injection and $I_h^* : Y' \to Y_h'$. Show that the inexact Minres problem $\min_{v_h \in V_h} \|I_h^*(f - B(v_h))\|_{Y_h'}$ has a unique solution $u_h \in V_h$. (*Hint*: introduce the residual representative $r_h := J_{Y_h}^{-1}I_h^*(f - B(u_h)) \in V_h$ and show that the pair $(u_h, r_h) \in V_h \times Y_h$ solves a saddle point problem.) (iii) Show that the residual representative $r_h \in Y_h$ is the unique solution of the following constrained minimization problem: $\min_{z_h \in Y_h \cap (I_h^*(B(V_h)))^\perp} \frac{1}{2}\|z_h\|_Y^2 - \langle I_h^*(f), z_h \rangle_{Y_h',Y_h}$. (*Hint*: see Proposition 49.11.) (iv) Assume now that $f \in \operatorname{im}(B)$ so that $B(u) = f$. Prove that there is c s.t. $\|u - u_h\|_V \leq c \inf_{w_h \in V_h} \|u - w_h\|_V$ for all $h \in \mathcal{H}$. (*Hint*: use a Fortin operator.) *Note*: since $\beta\|v_h\|_V \leq \|B(v_h)\|_{Y'}$ for all $v_h \in V_h$, it is natural to expect that the inf-sup condition (50.26) is satisfied if the subspace $Y_h \subset Y$ is chosen rich enough. The inexact residual minimization in

a discrete dual norm is at the heart of the discontinuous Petrov–Galerkin (dPG) method; see Demkowicz and Gopalakrishnan [158], Gopalakrishnan and Qiu [219], Carstensen et al. [111]. The extension to reflexive Banach spaces is studied in Muga and van der Zee [308].

Chapter 51

Darcy's equations

Darcy's equations consist of the following PDEs in the domain $D \subset \mathbb{R}^d$:

$$\mathbb{d}^{-1}\boldsymbol{\sigma} + \nabla p = \boldsymbol{f} \quad \text{in } D, \tag{51.1a}$$

$$\nabla{\cdot}\boldsymbol{\sigma} = g \qquad \text{in } D. \tag{51.1b}$$

The unknowns are the *primal* variable p and the *dual* variable $\boldsymbol{\sigma}$. In the literature, p is also called *potential* and $\boldsymbol{\sigma}$ *flux*. The PDEs (51.1) are used to model porous media flows, e.g., fluid flows in aquifers and petroleum reservoirs. In this context, $\boldsymbol{\sigma}$ is the seepage velocity, p the pressure, and $\mathbb{d}$ the material permeability, the equation (51.1a) is called *Darcy's law*, and (51.1b) expresses mass conservation. Eliminating the dual variable $\boldsymbol{\sigma}$ leads to $-\nabla{\cdot}(\mathbb{d}\nabla p) = g - \nabla{\cdot}(\mathbb{d}\boldsymbol{f})$ in D, which is a PDE where the only unknown is the primal variable p. This PDE can be approximated using, e.g., H^1-conforming finite elements as in Chapter 32. The approach we follow here is conceptually different since our aim is to approximate simultaneously the primal and the dual variables. In this chapter, we derive well-posed weak formulations for (51.1) with various boundary conditions. Then we study mixed finite element approximations using $\boldsymbol{H}(\text{div})$-conforming spaces for the dual variable.

51.1   Weak mixed formulation

The data in (51.1) are $\mathbb{d}$, $\boldsymbol{f}$, and g, where $\mathbb{d}$ is a second-order tensor if the material is anisotropic and it may depend on $\boldsymbol{x}$ if the material is non-homogeneous. We assume that $\boldsymbol{f} \in \boldsymbol{L}^2(D)$, $g \in L^2(D)$ and that $\mathbb{d}$ is symmetric and the eigenvalues of $\mathbb{d}$ are bounded from below and from above, respectively, by $\lambda_\flat$ and $\lambda_\sharp$ uniformly in D. We assume that $\lambda_\flat > 0$. We will consider Dirichlet, Neumann, and mixed Dirichlet–Neumann conditions for (51.1), and we will see that contrary to the primal formulation studied in Chapter 31, Dirichlet conditions on p are enforced weakly, whereas Neumann conditions on $\boldsymbol{\sigma}$ are enforced strongly.

51.1.1   Dirichlet boundary condition

In this section, we consider the Dirichlet condition $\gamma^{\text{g}}(p) = a_{\text{d}}$ on ∂D. Let us first proceed informally by assuming that all the functions are smooth enough. Multiplying (51.1a) by a smooth vector-valued test function $\boldsymbol{\tau}$, integrating over D, integrating by parts the term with ∇p, and using

the Dirichlet boundary condition, we infer that

$$\int_D ((\mathrm{d}^{-1}\boldsymbol{\sigma})\cdot\boldsymbol{\tau} - p\nabla\cdot\boldsymbol{\tau})\,\mathrm{d}x = \int_D \boldsymbol{f}\cdot\boldsymbol{\tau}\,\mathrm{d}x - \int_{\partial D} a_{\mathrm{d}}\boldsymbol{\tau}\cdot\boldsymbol{n}\,\mathrm{d}s. \tag{51.2}$$

Furthermore, multiplying (51.1b) by a smooth scalar-valued test function q and integrating over D gives

$$\int_D (\nabla\cdot\boldsymbol{\sigma})q\,\mathrm{d}x = \int_D gq\,\mathrm{d}x. \tag{51.3}$$

Since $\boldsymbol{f} \in \boldsymbol{L}^2(D)$ and $g \in L^2(D)$, the volume integrals make sense if we assume that $\boldsymbol{\sigma},\boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div}; D) := \{\boldsymbol{\varsigma} \in \boldsymbol{L}^2(D) \mid \nabla\cdot\boldsymbol{\varsigma} \in L^2(D)\}$ and we assume that $p, q \in L^2(D)$. To be dimensionally coherent, we equip $\boldsymbol{H}(\mathrm{div}; D)$ with the norm $\|\boldsymbol{\varsigma}\|_{\boldsymbol{H}(\mathrm{div};D)} := (\|\boldsymbol{\varsigma}\|_{\boldsymbol{L}^2(D)}^2 + \ell_D^2\|\nabla\cdot\boldsymbol{\varsigma}\|_{L^2(D)}^2)^{\frac{1}{2}}$, where ℓ_D is a length scale associated with D, e.g., $\ell_D := \mathrm{diam}(D)$.

A rigorous meaning can be given to the boundary integral in (51.2) once we recall that any vector field in $\boldsymbol{H}(\mathrm{div}; D)$ has a normal component over ∂D that can be defined by the normal trace map $\gamma^{\mathrm{d}} : \boldsymbol{H}(\mathrm{div}; D) \to H^{-\frac{1}{2}}(\partial D)$ such that the following integration by parts formula holds true (see (4.12)):

$$\langle\gamma^{\mathrm{d}}(\boldsymbol{\tau}), \gamma^{\mathrm{g}}(q)\rangle_{\partial D} := \int_D (q\nabla\cdot\boldsymbol{\tau} + \boldsymbol{\tau}\cdot\nabla q)\,\mathrm{d}x, \tag{51.4}$$

for all $q \in H^1(D)$ and all $\boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div}; D)$, where $\langle\cdot,\cdot\rangle_{\partial D}$ denotes the duality pairing between $H^{-\frac{1}{2}}(\partial D)$ and $H^{\frac{1}{2}}(\partial D)$. More precisely, assuming that $a_{\mathrm{d}} \in H^{\frac{1}{2}}(\partial D)$, the boundary integral in (51.2) is understood as $\langle\gamma^{\mathrm{d}}(\boldsymbol{\tau}), a_{\mathrm{d}}\rangle_{\partial D}$. We can now define the linear forms $F_{\mathrm{d}}(\boldsymbol{\tau}) := \int_D \boldsymbol{f}\cdot\boldsymbol{\tau}\,\mathrm{d}x - \langle\gamma^{\mathrm{d}}(\boldsymbol{\tau}), a_{\mathrm{d}}\rangle_{\partial D}$ and $G_{\mathrm{d}}(q) := -\int_D gq\,\mathrm{d}x$ and the bilinear forms

$$a(\boldsymbol{\varsigma},\boldsymbol{\tau}) := \int_D (\mathrm{d}^{-1}\boldsymbol{\varsigma})\cdot\boldsymbol{\tau}\,\mathrm{d}x, \qquad b(\boldsymbol{\varsigma},q) := -\int_D (\nabla\cdot\boldsymbol{\varsigma})q\,\mathrm{d}x, \tag{51.5}$$

and we consider the following problem:

$$\begin{cases} \text{Find } \boldsymbol{\sigma} \in \boldsymbol{V} := \boldsymbol{H}(\mathrm{div}; D) \text{ and } p \in Q := L^2(D) \text{ such that} \\ a(\boldsymbol{\sigma},\boldsymbol{\tau}) + b(\boldsymbol{\tau}, p) = F_{\mathrm{d}}(\boldsymbol{\tau}), \quad \forall\boldsymbol{\tau} \in \boldsymbol{V}, \\ b(\boldsymbol{\sigma}, q) = G_{\mathrm{d}}(q), \qquad\qquad \forall q \in Q. \end{cases} \tag{51.6}$$

The above assumptions imply that $F_{\mathrm{d}} \in \boldsymbol{V}'$, $G_{\mathrm{d}} \in Q' \equiv Q$, a is bounded on $\boldsymbol{V}\times\boldsymbol{V}$, and b is bounded on $\boldsymbol{V}\times Q$. The negative sign in the definition of b and G_{d} is not essential. This choice leads to a symmetric weak problem.

Proposition 51.1 (Well-posedness). *Assume $\boldsymbol{f} \in \boldsymbol{L}^2(D)$, $g \in L^2(D)$, and $a_{\mathrm{d}} \in H^{\frac{1}{2}}(\partial D)$.* (i) *The problem* (51.6) *is well-posed.* (ii) *The pair $(\boldsymbol{\sigma}, p)$ satisfies the PDEs (51.1) a.e. in D, p is in $H^1(D)$, and p satisfies the boundary condition $\gamma^{\mathrm{g}}(p) = a_{\mathrm{d}}$ a.e. on ∂D, where $\gamma^{\mathrm{g}} : H^1(D) \to H^{\frac{1}{2}}(\partial D)$ is the trace map defined in Theorem 3.10.*

Proof. (i) We apply Theorem 49.13. Set $B := \nabla\cdot : \boldsymbol{V} \to Q$ so that $\ker(B) = \{\boldsymbol{v} \in \boldsymbol{V} \mid \nabla\cdot\boldsymbol{v} = 0\}$. We have already seen that the bilinear forms a and b are continuous. Moreover, the inequality $a(\boldsymbol{\varsigma},\boldsymbol{\varsigma}) \geq \lambda_\sharp^{-1}\|\boldsymbol{\varsigma}\|_{\boldsymbol{L}^2(D)}^2$ implies that the bilinear form a is coercive on $\ker(B)$ since $\|\boldsymbol{\varsigma}\|_{\boldsymbol{L}^2(D)} = \|\boldsymbol{\varsigma}\|_{\boldsymbol{H}(\mathrm{div};D)}$ if $\boldsymbol{\varsigma} \in \ker(B)$. Hence, the two conditions (49.36) on a hold true. Moreover, the inf-sup condition (49.37) on b follows from Lemma 51.2 below. Hence, all the required assumptions for well-posedness are met.
(ii) Testing (51.6) with an arbitrary function $\boldsymbol{\tau} \in \boldsymbol{C}_0^\infty(D) := C_0^\infty(D; \mathbb{R}^d)$ and with $q := 0$, we

infer that $\int_D p\nabla\cdot\boldsymbol{\tau}\,\mathrm{d}x = \int_D(\mathrm{d}^{-1}\boldsymbol{\sigma} - \boldsymbol{f})\cdot\boldsymbol{\tau}\,\mathrm{d}x$. This proves that p has a weak derivative in $L^2(D)$ and $\nabla p = \boldsymbol{f} - \mathrm{d}^{-1}\boldsymbol{\sigma}$. Hence, $p \in H^1(D)$. Since $p \in H^1(D)$, we invoke the integration by parts formula (51.4) and infer that

$$\langle\gamma^{\mathrm{d}}(\boldsymbol{\tau}), \gamma^{\mathrm{g}}(p)\rangle_{\partial D} = \int_D (p\nabla\cdot\boldsymbol{\tau} + \boldsymbol{\tau}\cdot\nabla p)\,\mathrm{d}x = \int_D \left(p\nabla\cdot\boldsymbol{\tau} - \boldsymbol{\tau}\cdot(\mathrm{d}^{-1}\boldsymbol{\sigma} - \boldsymbol{f})\right)\,\mathrm{d}x$$
$$= -a(\boldsymbol{\sigma}, \boldsymbol{\tau}) - b(p, \boldsymbol{\tau}) + F_{\mathrm{d}}(\boldsymbol{\tau}) + \langle\gamma^{\mathrm{d}}(\boldsymbol{\tau}), a_{\mathrm{d}}\rangle_{\partial D} = \langle\gamma^{\mathrm{d}}(\boldsymbol{\tau}), a_{\mathrm{d}}\rangle_{\partial D},$$

for all $\boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div}; D)$. The surjectivity of γ^{d} from Theorem 4.15 implies that $\langle\phi, \gamma^{\mathrm{g}}(p) - a_{\mathrm{d}}\rangle_{\partial D} = 0$ for all $\phi \in H^{-\frac{1}{2}}(\partial D)$. Hence, $\gamma^{\mathrm{g}}(p) = a_{\mathrm{d}}$ in $H^{\frac{1}{2}}(\partial D)$. Testing (51.6) with $\boldsymbol{\tau} := \boldsymbol{0}$ and an arbitrary function $q \in C_0^\infty(D)$ finally yields $\int_D(\nabla\cdot\boldsymbol{\sigma} - g)q\,\mathrm{d}x = 0$. Invoking Theorem 1.32 proves $\nabla\cdot\boldsymbol{\sigma} = g$. $\qquad\square$

Lemma 51.2 (Surjectivity of divergence). *Let D be a Lipschitz domain in $\mathbb{R}^d$. The operator $\nabla\cdot : \boldsymbol{H}(\mathrm{div}; D) \to L^2(D)$ is surjective, and we have*

$$\inf_{q\in L^2(D)} \sup_{\boldsymbol{\tau}\in\boldsymbol{H}(\mathrm{div};D)} \frac{|\int_D q\nabla\cdot\boldsymbol{\tau}\,\mathrm{d}x|}{\|q\|_{L^2(D)}\|\boldsymbol{\tau}\|_{\boldsymbol{H}(\mathrm{div};D)}} \geq \ell_D^{-1}\beta_D, \tag{51.7}$$

with $\beta_D := (C_{\mathrm{PS}}^{-2} + 1)^{-\frac{1}{2}}$ where C_{PS} is the constant from the Poincaré–Steklov inequality (3.11) with $p := 2$, i.e., $C_{\mathrm{PS}}\|v\|_{L^2(D)} \leq \ell_D\|\nabla v\|_{\boldsymbol{L}^2(D)}$ for all $v \in H_0^1(D)$.

Proof. Let $q \in L^2(D)$. Let $\phi \in H_0^1(D)$ be such that $(\nabla\phi, \nabla\psi)_{\boldsymbol{L}^2(D)} = (q, \psi)_{L^2(D)}$ for all $\psi \in H_0^1(D)$, so that $\|\nabla\phi\|_{\boldsymbol{L}^2(D)} \leq C_{\mathrm{PS}}^{-1}\ell_D\|q\|_{L^2(D)}$. Setting $\boldsymbol{\varsigma}_q := -\nabla\phi$, we have $\boldsymbol{\varsigma}_q \in \boldsymbol{H}(\mathrm{div}; D)$, $\nabla\cdot\boldsymbol{\varsigma}_q = q$, and

$$\|\boldsymbol{\varsigma}_q\|_{\boldsymbol{H}(\mathrm{div};D)}^2 = \|\nabla\phi\|_{\boldsymbol{L}^2(D)}^2 + \ell_D^2\|q\|_{L^2(D)}^2 \leq (C_{\mathrm{PS}}^{-2} + 1)\ell_D^2\|q\|_{L^2(D)}^2,$$

so that $\|\boldsymbol{\varsigma}_q\|_{\boldsymbol{H}(\mathrm{div};D)} \leq \ell_D\beta_D^{-1}\|q\|_{L^2(D)}$. As a result, we have

$$\sup_{\boldsymbol{\tau}\in\boldsymbol{H}(\mathrm{div};D)} \frac{\int_D q\nabla\cdot\boldsymbol{\tau}\,\mathrm{d}x}{\|\boldsymbol{\tau}\|_{\boldsymbol{H}(\mathrm{div};D)}} \geq \frac{\int_D q\nabla\cdot\boldsymbol{\varsigma}_q\,\mathrm{d}x}{\|\boldsymbol{\varsigma}_q\|_{\boldsymbol{H}(\mathrm{div};D)}} = \frac{\|q\|_{L^2(D)}^2}{\|\boldsymbol{\varsigma}_q\|_{\boldsymbol{H}(\mathrm{div};D)}} \geq \ell_D^{-1}\beta_D\|q\|_{L^2(D)},$$

and this proves (51.7). $\qquad\square$

51.1.2 Neumann boundary condition

We now consider the Neumann boundary condition $\boldsymbol{\sigma}\cdot\boldsymbol{n} = a_{\mathrm{n}}$ on ∂D, and we assume that $a_{\mathrm{n}} \in H^{-\frac{1}{2}}(\partial D)$. We still look for $\boldsymbol{\sigma} \in \boldsymbol{H}(\mathrm{div}; D)$, and we interpret the boundary condition as $\gamma^{\mathrm{d}}(\boldsymbol{\sigma}) = a_{\mathrm{n}}$. Since $\langle\gamma^{\mathrm{d}}(\boldsymbol{\sigma}), 1\rangle_{\partial D} = \int_D \nabla\cdot\boldsymbol{\sigma}\,\mathrm{d}x$, the data a_{n} and g must satisfy the compatibility condition

$$\langle a_{\mathrm{n}}, 1\rangle_{\partial D} = \int_D g\,\mathrm{d}x. \tag{51.8}$$

Since only the gradient of the primal variable p now appears in the problem, we additionally require that $p \in L_*^2(D) := \{q \in L^2(D) \mid \int_D q\,\mathrm{d}x = 0\}$. Since $\gamma^{\mathrm{d}} : \boldsymbol{H}(\mathrm{div}; D) \to H^{-\frac{1}{2}}(\partial D)$ is surjective (see Theorem 4.15(ii) or Corollary 31.20), there exists a field $\boldsymbol{\sigma}_{\mathrm{n}} \in \boldsymbol{H}(\mathrm{div}; D)$ s.t. $\gamma^{\mathrm{d}}(\boldsymbol{\sigma}_{\mathrm{n}}) = a_{\mathrm{n}}$. We now make the change of variable $\boldsymbol{\sigma}_0 := \boldsymbol{\sigma} - \boldsymbol{\sigma}_{\mathrm{n}}$. Note that $\boldsymbol{\sigma}_0$ satisfies the homogeneous Neumann boundary condition $\gamma^{\mathrm{d}}(\boldsymbol{\sigma}_0) = 0$, i.e.,

$$\boldsymbol{\sigma}_0 \in \boldsymbol{H}_0(\mathrm{div}; D) := \{\boldsymbol{\varsigma} \in \boldsymbol{H}(\mathrm{div}; D) \mid \gamma^{\mathrm{d}}(\boldsymbol{\varsigma}) = 0\} = \ker(\gamma^{\mathrm{d}}). \tag{51.9}$$

$\boldsymbol{H}_0(\text{div}; D)$ is a Hilbert space when equipped with the norm $\|\cdot\|_{\boldsymbol{H}(\text{div};D)}$. The weak formulation is as follows:

$$\begin{cases} \text{Find } \boldsymbol{\sigma}_0 \in \boldsymbol{V} := \boldsymbol{H}_0(\text{div}; D) \text{ and } p \in Q := L^2_*(D) \text{ such that} \\ a(\boldsymbol{\sigma}_0, \boldsymbol{\tau}) + b(\boldsymbol{\tau}, p) = F_{\text{n}}(\boldsymbol{\tau}), \qquad \forall \boldsymbol{\tau} \in \boldsymbol{V}, \\ b(\boldsymbol{\sigma}_0, q) = G_{\text{n}}(q), \qquad\qquad \forall q \in Q, \end{cases} \qquad (51.10)$$

with the linear forms $F_{\text{n}}(\boldsymbol{\tau}) := \int_D (\boldsymbol{f} - \mathbb{d}^{-1}\boldsymbol{\sigma}_{\text{n}}) \cdot \boldsymbol{\tau} \, \mathrm{d}x$ and $G_{\text{n}}(q) := -\int_D (g - \nabla \cdot \boldsymbol{\sigma}_{\text{n}}) q \, \mathrm{d}x$. The bilinear forms a and b have been defined in (51.5).

Proposition 51.3 (Well-posedness). *Assume that $\boldsymbol{f} \in \boldsymbol{L}^2(D)$, $g \in L^2(D)$, $a_{\text{n}} \in H^{-\frac{1}{2}}(\partial D)$ and that the compatibility condition (51.8) holds true. (i) The problem (51.10) is well-posed. (ii) The pair $(\boldsymbol{\sigma} := \boldsymbol{\sigma}_0 + \boldsymbol{\sigma}_{\text{n}}, p)$ satisfies the PDEs (51.1) a.e. in D, and the boundary condition $\gamma^{\text{d}}(\boldsymbol{\sigma}) = a_{\text{n}}$ is satisfied in $H^{-\frac{1}{2}}(\partial D)$.*

Proof. See Exercise 51.2. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\Box$

Remark 51.4 (Choice of $\boldsymbol{\sigma}_{\text{n}}$). One possibility to define $\boldsymbol{\sigma}_{\text{n}} \in \boldsymbol{H}(\text{div}; D)$ is to set $\boldsymbol{\sigma}_{\text{n}} := \nabla\phi$, where $\phi \in H^1_*(D) := \{q \in H^1(D) \mid \int_D q \, \mathrm{d}x = 0\}$ solves the pure Neumann problem $\int_D \nabla\phi \cdot \nabla r \, \mathrm{d}x := -\int_D gr \, \mathrm{d}x + \langle a_{\text{n}}, \gamma^{\text{g}}(r)\rangle_{\partial D}$ for all $r \in H^1_*(D)$. The compatibility condition (51.8) implies that it is legitimate to take any test function r in $H^1(D)$ in the above equation. Taking first $r \in C_0^\infty(D)$ yields $\nabla \cdot \boldsymbol{\sigma}_{\text{n}} = \Delta\phi = g$, whence $\langle \gamma^{\text{d}}(\boldsymbol{\sigma}_{\text{n}}) - a_{\text{n}}, \gamma^{\text{g}}(r)\rangle_{\partial D} = 0$ for all $r \in H^1(D)$. The trace map $\gamma^{\text{g}} : H^1(D) \to H^{\frac{1}{2}}(\partial D)$ being surjective, we conclude that $\gamma^{\text{d}}(\boldsymbol{\sigma}_{\text{n}}) = a_{\text{n}}$. This construction of $\boldsymbol{\sigma}_{\text{n}}$ gives $G_{\text{n}} = 0$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\Box$

51.1.3   Mixed Dirichlet–Neumann boundary conditions

Let $\partial D_{\text{d}} \cup \partial D_{\text{n}}$ be a partition of the boundary ∂D with $|\partial D_{\text{d}}| \neq 0$ and $|\partial D_{\text{n}}| \neq 0$. We want to enforce the mixed Dirichlet–Neumann conditions $p = a_{\text{d}}$ on ∂D_{d} and $\boldsymbol{\sigma} \cdot \boldsymbol{n} = a_{\text{n}}$ on ∂D_{n}. A rigorous mathematical setting for these conditions entails some subtleties.

Concerning the Dirichlet condition, we assume that there exists a bounded extension operator $H^{\frac{1}{2}}(\partial D_{\text{d}}) \to H^{\frac{1}{2}}(\partial D)$ (see §31.3.3). We assume that $a_{\text{d}} \in H^{\frac{1}{2}}(\partial D_{\text{d}})$, and we denote by $\breve{a}_{\text{d}} \in H^{\frac{1}{2}}(\partial D)$ the extension of a_{d}. Concerning the Neumann condition, we have seen in §51.1.2 that Neumann conditions on ∂D are enforced using the normal trace operator $\gamma^{\text{d}} : \boldsymbol{H}(\text{div}; D) \to H^{-\frac{1}{2}}(\partial D)$. When the Neumann condition is enforced only on ∂D_{n}, we need to consider the restriction to ∂D_{n} of linear forms in $H^{-\frac{1}{2}}(\partial D)$. Let $\widetilde{H}^{\frac{1}{2}}(\partial D_{\text{n}})$ be composed of the functions θ defined on ∂D_{n} whose extension by zero to ∂D, say $\tilde{\theta}$, is in $H^{\frac{1}{2}}(\partial D)$. Let us denote by $\langle \cdot, \cdot \rangle_{\partial D_{\text{n}}}$ the duality pairing between $\widetilde{H}^{\frac{1}{2}}(\partial D_{\text{n}})'$ and $\widetilde{H}^{\frac{1}{2}}(\partial D_{\text{n}})$ i.e., the action of $a \in \widetilde{H}^{\frac{1}{2}}(\partial D_{\text{n}})'$ on $r \in \widetilde{H}^{\frac{1}{2}}(\partial D_{\text{n}})$ is denoted by $\langle a, r\rangle_{\partial D_{\text{n}}}$. Then for all $\boldsymbol{\varsigma} \in \boldsymbol{H}(\text{div}; D)$, the restriction $\gamma^{\text{d}}(\boldsymbol{\varsigma})_{|\partial D_{\text{n}}}$ is defined in $\widetilde{H}^{\frac{1}{2}}(\partial D_{\text{n}})'$ by setting

$$\langle \gamma^{\text{d}}(\boldsymbol{\varsigma})_{|\partial D_{\text{n}}}, \theta \rangle_{\partial D_{\text{n}}} := \langle \gamma^{\text{d}}(\boldsymbol{\varsigma}), \tilde{\theta} \rangle_{\partial D}. \qquad (51.11)$$

Lemma 51.5 (Surjectivity of restricted normal trace). *The restricted normal trace operator $\gamma^{\text{d}}(\cdot)_{|\partial D_{\text{n}}} : \boldsymbol{H}(\text{div}; D) \to \widetilde{H}^{\frac{1}{2}}(\partial D_{\text{n}})'$ is surjective.*

Proof. We proceed as in Remark 51.4. Let $a_{\text{n}} \in \widetilde{H}^{\frac{1}{2}}(\partial D_{\text{n}})'$ and let us set $H^1_{\text{d}}(D) := \{r \in H^1(D) \mid \gamma^{\text{g}}(r)_{|\partial D_{\text{d}}} = 0\}$. Notice that for all $r \in H^1_{\text{d}}(D)$, the zero extension of $\gamma^{\text{g}}(r)_{|\partial D_{\text{n}}}$ to ∂D coincides with $\gamma^{\text{g}}(r)$, which is in $H^{\frac{1}{2}}(\partial D)$. Hence, $\gamma^{\text{g}}(r)_{|\partial D_{\text{n}}} \in \widetilde{H}^{\frac{1}{2}}(\partial D_{\text{n}})$. Let $\phi \in H^1_{\text{d}}(D)$ solve the mixed Dirichlet–Neumann problem $\int_D \nabla\phi \cdot \nabla r \, \mathrm{d}x = \langle a_{\text{n}}, \gamma^{\text{g}}(r)_{|\partial D_{\text{n}}}\rangle_{\partial D_{\text{n}}}$ for all $r \in H^1_{\text{d}}(D)$. We

now set $\boldsymbol{\sigma}_\mathrm{n} := \nabla\phi$ and observe that $\nabla{\cdot}\boldsymbol{\sigma}_\mathrm{n} = 0$. Since for all $\theta \in \widetilde{H}^{\frac{1}{2}}(\partial D_\mathrm{n})$ there is $r_\theta \in H_\mathrm{d}^1(D)$ such that $\gamma^\mathrm{g}(r_\theta) = \tilde{\theta}$ (the zero-extension of θ to ∂D), we infer that

$$\langle \gamma^\mathrm{d}(\boldsymbol{\sigma}_\mathrm{n})_{|\partial D_\mathrm{n}} - a_\mathrm{n}, \theta \rangle_{\partial D_\mathrm{n}} = \langle \gamma^\mathrm{d}(\boldsymbol{\sigma}_\mathrm{n}), \tilde{\theta} \rangle_{\partial D} - \int_D \nabla\phi{\cdot}\nabla r_\theta = \int_D r_\theta \nabla{\cdot}\boldsymbol{\sigma}_\mathrm{n}\,\mathrm{d}x = 0.$$

Since this identity holds true for all $\theta \in \widetilde{H}^{\frac{1}{2}}(\partial D_\mathrm{n})$, we conclude that $\gamma^\mathrm{d}(\boldsymbol{\sigma}_\mathrm{n})_{|\partial D_\mathrm{n}} = a_\mathrm{n}$, i.e., $\gamma^\mathrm{d}(\cdot)_{|\partial D_\mathrm{n}}$ is surjective. $\qquad\square$

Owing to Lemma 51.5, it is natural to assume that $a_\mathrm{n} \in \widetilde{H}^{\frac{1}{2}}(\partial D_\mathrm{n})'$. Referring again to this lemma and its proof, we then infer the existence of $\boldsymbol{\sigma}_\mathrm{n} \in \boldsymbol{H}(\mathrm{div};D)$ with $\nabla{\cdot}\boldsymbol{\sigma}_\mathrm{n} = 0$ such that $\gamma^\mathrm{d}(\boldsymbol{\sigma}_\mathrm{n})_{|\partial D_\mathrm{n}} = a_\mathrm{n}$. Making the change of variable $\boldsymbol{\sigma}_0 := \boldsymbol{\sigma} - \boldsymbol{\sigma}_\mathrm{n}$ gives

$$\boldsymbol{\sigma}_0 \in \boldsymbol{H}_\mathrm{n}(\mathrm{div};D) := \{\boldsymbol{\varsigma} \in \boldsymbol{H}(\mathrm{div};D) \mid \gamma^\mathrm{d}(\boldsymbol{\varsigma})_{|\partial D_\mathrm{n}} = 0\}. \tag{51.12}$$

Notice that $\boldsymbol{H}_\mathrm{n}(\mathrm{div};D)$ is a Hilbert space when equipped with the natural norm. The weak formulation we now consider is as follows:

$$\begin{cases} \text{Find } \boldsymbol{\sigma}_0 \in \boldsymbol{V} := \boldsymbol{H}_\mathrm{n}(\mathrm{div};D) \text{ and } p \in Q := L^2(D) \text{ such that} \\ a(\boldsymbol{\sigma}_0, \boldsymbol{\tau}) + b(\boldsymbol{\tau}, p) = F_\mathrm{dn}(\boldsymbol{\tau}), \qquad \forall \boldsymbol{\tau} \in \boldsymbol{V}, \\ b(\boldsymbol{\sigma}_0, q) = G_\mathrm{n}(q), \qquad\qquad \forall q \in Q, \end{cases} \tag{51.13}$$

with the linear forms $F_\mathrm{dn}(\boldsymbol{\tau}) := \int_D (\boldsymbol{f} - \mathrm{d}^{-1}\boldsymbol{\sigma}_\mathrm{n}){\cdot}\boldsymbol{\tau}\,\mathrm{d}x - \langle \gamma^\mathrm{d}(\boldsymbol{\tau}), \breve{a}_\mathrm{d} \rangle_{\partial D}$ and $G_\mathrm{n}(q) := -\int_D (g - \nabla{\cdot}\boldsymbol{\sigma}_\mathrm{n})q\,\mathrm{d}x$. The bilinear forms a, b are defined in (51.5). Recall that $\breve{a}_\mathrm{d} \in H^{\frac{1}{2}}(\partial D)$ is an extension of a_d over ∂D, and notice that $\langle \gamma^\mathrm{d}(\boldsymbol{\tau}), \breve{a}_\mathrm{d} \rangle_{\partial D}$ is independent on the way $\breve{a}_\mathrm{d}$ is extended to $H^{\frac{1}{2}}(\partial D)$. Indeed, considering two extensions $\breve{a}_\mathrm{d}$ and $\hat{a}_\mathrm{d}$, we have $\breve{a}_\mathrm{d} - \hat{a}_\mathrm{d} \in \widetilde{H}^{\frac{1}{2}}(\partial D_\mathrm{n})$, so that $\langle \gamma^\mathrm{d}(\boldsymbol{\tau}), \breve{a}_\mathrm{d} - \hat{a}_\mathrm{d} \rangle_{\partial D} = 0$ for all $\boldsymbol{\tau} \in \boldsymbol{V}$.

Proposition 51.6 (Well-posedness). *Assume $\boldsymbol{f} \in \boldsymbol{L}^2(D)$, $g \in L^2(D)$, $a_\mathrm{d} \in H^{\frac{1}{2}}(\partial D_\mathrm{d})$, and $a_\mathrm{n} \in \widetilde{H}^{\frac{1}{2}}(\partial D_\mathrm{n})'$. (i) The problem (51.13) is well-posed. (ii) The pair $(\boldsymbol{\sigma} := \boldsymbol{\sigma}_0 + \boldsymbol{\sigma}_\mathrm{n}, p)$ satisfies the PDEs (51.1) a.e. in D, the Dirichlet condition $\gamma^\mathrm{g}(p)_{|\partial D_\mathrm{d}} = a_\mathrm{d}$ is satisfied a.e. on ∂D_d, and the Neumann condition $\gamma^\mathrm{d}(\boldsymbol{\sigma})_{|\partial D_\mathrm{n}} = a_\mathrm{n}$ is satisfied in $(\widetilde{H}^{\frac{1}{2}}(\partial D_\mathrm{n}))'$.*

Proof. We just sketch the differences with the proof of Proposition 51.1. The surjectivity of $\nabla{\cdot} : \boldsymbol{H}_\mathrm{n}(\mathrm{div};D) \to L^2(D)$ follows by defining $\phi_r \in H^1(D)$ such that $\Delta\phi_r = r$, $\phi_{r|\partial D_\mathrm{d}} = 0$, $\partial_n\phi_{r|\partial D_\mathrm{n}} = 0$ for all $r \in L^2(D)$ and observing that $\nabla{\cdot}(\nabla\phi_r) = r$ and $\nabla\phi_r \in \boldsymbol{H}_\mathrm{n}(\mathrm{div};D)$. To recover the Dirichlet boundary condition, we observe, as in the proof of Proposition 51.1, that $\nabla p = \boldsymbol{f} - \mathrm{d}^{-1}(\boldsymbol{\sigma}_\mathrm{n} - \boldsymbol{\sigma}_0)$, which in turn implies that $p \in H^1(D)$ and

$$\langle \gamma^\mathrm{d}(\boldsymbol{\tau}), \gamma^\mathrm{g}(p) \rangle_{\partial D} = \langle \gamma^\mathrm{d}(\boldsymbol{\tau}), \breve{a}_\mathrm{d} \rangle_{\partial D}, \qquad \forall \boldsymbol{\tau} \in \boldsymbol{H}_\mathrm{n}(\mathrm{div};D). \tag{51.14}$$

Let $\psi \in C_0^\infty(\partial D_\mathrm{d})$ and let $\tilde{\psi}$ be the extension by zero of ψ to ∂D. Recalling that $\gamma^\mathrm{d} : \boldsymbol{H}(\mathrm{div};D) \to H^{-\frac{1}{2}}(\partial D)$ is surjective (see Theorem 4.15(ii) or Corollary 31.20), there is $\boldsymbol{\tau}_\psi \in \boldsymbol{H}(\mathrm{div};D)$ s.t. $\gamma^\mathrm{d}(\boldsymbol{\tau}_\psi) = \tilde{\psi}$. Notice that $\gamma^\mathrm{d}(\boldsymbol{\tau}_\psi)_{|\partial D_\mathrm{n}} = \tilde{\psi}_{|\partial D_\mathrm{n}} = 0$, i.e., $\boldsymbol{\tau}_\psi \in \boldsymbol{H}_\mathrm{n}(\mathrm{div};D)$. Using $\boldsymbol{\tau}_\psi$ in (51.14) shows that $\int_{\partial D_\mathrm{d}} (\gamma^\mathrm{g}(p) - a_\mathrm{d})\psi\,\mathrm{d}s = 0$, which in turn gives $\gamma^\mathrm{g}(p)_{|\partial D_\mathrm{d}} = a_\mathrm{d}$ since ψ is arbitrary in $C_0^\infty(\partial D_\mathrm{d})$. $\qquad\square$

51.2 Primal, dual, and dual mixed formulations

In this section, we consider alternative formulations where either the primal variable p or the dual variable $\boldsymbol{\sigma}$ is eliminated. We focus on homogeneous Dirichlet boundary conditions for simplicity. The material readily extends to other types of (non-homogeneous) boundary conditions.

The primal formulation of the PDEs (51.1) with the boundary condition $p = 0$ on ∂D is obtained by eliminating $\boldsymbol{\sigma}$ using the identity $\boldsymbol{\sigma} = \mathrm{d}(\boldsymbol{f} - \nabla p)$. This leads to the following formulation:

$$\begin{cases} \text{Find } p \in H_0^1(D) \text{ such that} \\ a_\sharp(p, r) := \int_D \mathrm{d}\nabla p{\cdot}\nabla r \, \mathrm{d}x = F_\sharp(r), \quad \forall r \in H_0^1(D), \end{cases} \tag{51.15}$$

with the linear form $F_\sharp(r) := \int_D (\mathrm{d}\boldsymbol{f}{\cdot}\nabla r + gr) \, \mathrm{d}x$. This problem has been analyzed in Chapter 31 (see Remark 31.7 since $F_\sharp \in H^{-1}(D)$). In particular, p is the unique minimizer of the energy functional

$$\mathfrak{E}_\sharp(q) := \frac{1}{2} a_\sharp(q, q) - F_\sharp(q) \tag{51.16}$$

over $H_0^1(D)$ (see Remark 31.10).

The dual formulation is obtained by eliminating p using divergence-free test functions in Darcy's law (observe that $\int_D \nabla p{\cdot}\boldsymbol{\tau} \, \mathrm{d}x = 0$ if $\boldsymbol{\tau}$ is divergence-free since $p_{|\partial D} = 0$), and enforcing the mass conservation equation explicitly. This leads to the following formulation:

$$\begin{cases} \text{Find } \boldsymbol{\sigma} \in \boldsymbol{H}(\mathrm{div}; D) \text{ with } \nabla{\cdot}\boldsymbol{\sigma} = g \text{ such that} \\ a_\flat(\boldsymbol{\sigma}, \boldsymbol{\tau}) := \int_D (\mathrm{d}^{-1}\boldsymbol{\sigma}){\cdot}\boldsymbol{\tau} \, \mathrm{d}x = F_\flat(\boldsymbol{\tau}), \quad \forall \boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div} = 0; D), \end{cases} \tag{51.17}$$

with the space $\boldsymbol{H}(\mathrm{div} = 0; D) := \{\boldsymbol{\varsigma} \in \boldsymbol{H}(\mathrm{div}; D) \mid \nabla{\cdot}\boldsymbol{\varsigma} = 0\}$ and the linear form $F_\flat(\boldsymbol{\tau}) := \int_D \boldsymbol{f}{\cdot}\boldsymbol{\tau} \, \mathrm{d}x$. The well-posedness of (51.17) can be established by lifting the divergence constraint using $\boldsymbol{\sigma}_g \in \boldsymbol{H}(\mathrm{div}; D)$ such that $\nabla{\cdot}\boldsymbol{\sigma}_g = g$, making the change of variable $\boldsymbol{\sigma}_0 := \boldsymbol{\sigma} - \boldsymbol{\sigma}_g \in \boldsymbol{H}(\mathrm{div} = 0; D)$, and observing that the bilinear form $a_\flat$ is coercive on $\boldsymbol{H}(\mathrm{div} = 0; D)$ equipped with the natural norm. Moreover, defining the complementary energy functional

$$\mathfrak{E}_\flat(\boldsymbol{\varsigma}) := -\frac{1}{2} \int_D (\boldsymbol{\varsigma} - \mathrm{d}\boldsymbol{f}){\cdot}\mathrm{d}^{-1}(\boldsymbol{\varsigma} - \mathrm{d}\boldsymbol{f}) \, \mathrm{d}x, \tag{51.18}$$

and since (51.17) amounts to $D\mathfrak{E}_\flat(\boldsymbol{\sigma})(\boldsymbol{\tau}) = 0$ for all $\boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div} = 0; D)$, we infer that the field $\boldsymbol{\sigma}$ solving (51.17) is the unique maximizer of $\mathfrak{E}_\flat$ over the affine subspace of $\boldsymbol{H}(\mathrm{div}; D)$ with divergence equal to g. The dual formulation is seldom used for approximation purposes since it requires to manipulate divergence-free vector fields. An interesting application related to flux recovery for H^1-conforming finite elements is presented in §52.2.

We can now relate the primal formulation (51.15) and the dual formulation (51.17) to the mixed formulation (51.6), which in the present context is called *dual mixed formulation*.

Proposition 51.7 (Equivalence, energy identity). *The primal formulation* (51.15), *the dual formulation* (51.17), *and the dual mixed formulation* (51.6) *are equivalent in the sense that the solutions p from* (51.15) *and* (51.6) *coincide, the solutions $\boldsymbol{\sigma}$ from* (51.17) *and* (51.6) *coincide, and we have $\mathrm{d}^{-1}\boldsymbol{\sigma} + \nabla p = \boldsymbol{f}$. Moreover, the following energy identity holds true:*

$$\min_{q \in H_0^1(D)} \mathfrak{E}_\sharp(q) = \mathfrak{E}_\sharp(p) = \mathfrak{E}_\flat(\boldsymbol{\sigma}) = \max_{\boldsymbol{\varsigma} \in \boldsymbol{H}(\mathrm{div}; D), \nabla{\cdot}\boldsymbol{\varsigma} = g} \mathfrak{E}_\flat(\boldsymbol{\varsigma}). \tag{51.19}$$

Proof. See Exercise 51.4. $\square$

Remark 51.8 (Lagrangian). Proposition 49.11 implies that the pair $(\boldsymbol{\sigma}, p)$ solving the dual mixed formulation (51.6) is the unique saddle point of the Lagrangian

$$\mathcal{L}(\boldsymbol{\varsigma}, q) := \int_D (\frac{1}{2}\boldsymbol{\varsigma}{\cdot}\mathrm{d}^{-1}\boldsymbol{\varsigma} - q\nabla{\cdot}\boldsymbol{\varsigma}) \, \mathrm{d}x - F_\mathrm{d}(\boldsymbol{\varsigma}) - G_\mathrm{d}(q).$$

Since $\mathcal{L}(\boldsymbol{\varsigma}, q) = -\mathfrak{E}_\flat(\boldsymbol{\varsigma}) - \frac{1}{2}\int_D \boldsymbol{f}{\cdot}\mathrm{d}\boldsymbol{f} \, \mathrm{d}x + \int_D q(g - \nabla{\cdot}\boldsymbol{\varsigma}) \, \mathrm{d}x$, we infer that $\mathfrak{E}_\sharp(p) = \mathfrak{E}_\flat(\boldsymbol{\sigma}) = -\mathcal{L}(\boldsymbol{\sigma}, p) - \frac{1}{2}\int_D \boldsymbol{f}{\cdot}\mathrm{d}\boldsymbol{f} \, \mathrm{d}x$. $\square$

Remark 51.9 (Linear elasticity). The above formalism can be applied to the linear elasticity equations studied in Chapter 42. Denoting by $\boldsymbol{u}$ the displacement and $\mathbb{s}$ the stress tensor, Darcy's law is replaced by the constitutive equation $\mathbb{C}{:}\mathbb{s}(\boldsymbol{u}) = \mathbb{e}(\boldsymbol{u}) = \frac{1}{2}(\nabla\boldsymbol{u} + (\nabla\boldsymbol{u})^{\mathsf{T}})$, where $\mathbb{C}$ is the (fourth-order) compliance tensor (see Exercise 42.1), and the mass conservation equation is replaced by the equilibrium equation $\nabla{\cdot}\mathbb{s}(\boldsymbol{u}) = \boldsymbol{g}$. We refer the reader to, e.g., Gatica [212, §2.4.3] for some weak mixed formulations and to §42.4.2 for a brief literature review of their mixed finite element approximation. □

51.3   Approximation of the mixed formulation

In this section, we analyze an $\boldsymbol{H}(\mathrm{div})$-conforming approximation of the weak formulation (51.6) focusing on Dirichlet boundary conditions for simplicity.

51.3.1   Discrete problem and well-posedness

Let $(\mathcal{T}_h)_{h\in\mathcal{H}}$ be a shape-regular sequence of affine simplicial meshes so that each mesh covers D exactly, let $\boldsymbol{P}_k^{\mathrm{d}}(\mathcal{T}_h)$ be the $\boldsymbol{H}(\mathrm{div})$-conforming Raviart–Thomas finite element space of order $k \geq 0$ from §19.2.3, and let $P_k^{\mathrm{b}}(\mathcal{T}_h)$ be the broken finite element space built using piecewise polynomials in $\mathbb{P}_{k,d}$. We recall that

$$\boldsymbol{P}_k^{\mathrm{d}}(\mathcal{T}_h) := \{\boldsymbol{\varsigma}_h \in \boldsymbol{H}(\mathrm{div};D) \mid \boldsymbol{\psi}_K^{\mathrm{d}}(\boldsymbol{\varsigma}_{h|K}) \in \mathbb{R}\mathbb{T}_{k,d}, \ \forall K \in \mathcal{T}_h\}, \tag{51.20a}$$

$$P_k^{\mathrm{b}}(\mathcal{T}_h) := \{q_h \in L^2(D) \mid \psi_K^{\mathrm{g}}(q_{h|K}) \in \mathbb{P}_{k,d}, \ \forall K \in \mathcal{T}_h\}, \tag{51.20b}$$

where $\boldsymbol{\psi}_K^{\mathrm{d}}$ is the contravariant Piola transformation and ψ_K^{g} is the pullback by the geometric mapping $\boldsymbol{T}_K$ (see Definition 9.8). Notice that $\boldsymbol{\varsigma}_{h|K} \in \mathbb{R}\mathbb{T}_{k,d}$ and $q_{h|K} \in \mathbb{P}_{k,d}$ since $\mathcal{T}_h$ is affine. The discrete counterpart of (51.6) is

$$\begin{cases} \text{Find } \boldsymbol{\sigma}_h \in \boldsymbol{V}_h := \boldsymbol{P}_k^{\mathrm{d}}(\mathcal{T}_h) \text{ and } p_h \in Q_h := P_k^{\mathrm{b}}(\mathcal{T}_h) \text{ such that} \\ a(\boldsymbol{\sigma}_h, \boldsymbol{\tau}_h) + b(\boldsymbol{\tau}_h, p_h) = F_{\mathrm{d}}(\boldsymbol{\tau}_h), \qquad \forall \boldsymbol{\tau}_h \in \boldsymbol{V}_h, \\ b(\boldsymbol{\sigma}_h, q_h) = G_{\mathrm{d}}(q_h), \qquad\qquad\quad \forall q_h \in Q_h. \end{cases} \tag{51.21}$$

Let $\ker(B_h) = \{\boldsymbol{\varsigma}_h \in \boldsymbol{V}_h \mid b(\boldsymbol{\varsigma}_h, q_h) = 0, \ \forall q_h \in Q_h\}$ and recall that $\ker(B) = \{\boldsymbol{\varsigma} \in \boldsymbol{H}(\mathrm{div};D) \mid \nabla{\cdot}\boldsymbol{\varsigma} = 0\}$.

Lemma 51.10 (Discrete inf-sup). *We have*

$$\ker(B_h) \subset \ker(B), \tag{51.22}$$

and the following holds true:

$$\inf_{q_h \in Q_h} \sup_{\boldsymbol{\varsigma}_h \in \boldsymbol{V}_h} \frac{|\int_D q_h \nabla{\cdot}\boldsymbol{\varsigma}_h \, \mathrm{d}x|}{\|q_h\|_{L^2(D)}\|\boldsymbol{\varsigma}_h\|_{\boldsymbol{H}(\mathrm{div};D)}} \geq \ell_D^{-1}\beta_D^{\flat}, \tag{51.23}$$

with $\beta_D^{\flat} := (C_{\mathrm{PS}}^{-2}\|\mathcal{J}_h^{\mathrm{d}}\|_{\mathcal{L}(\boldsymbol{L}^2;\boldsymbol{L}^2)}^2 + 1)^{-\frac{1}{2}} > 0$, where $\mathcal{J}_h^{\mathrm{d}}$ is the $\boldsymbol{L}^2$-stable commuting projection from §23.3.3.

Proof. Since $\nabla{\cdot} : \boldsymbol{V}_h := \boldsymbol{P}_k^{\mathrm{d}}(\mathcal{T}_h) \to P_k^{\mathrm{b}}(\mathcal{T}_h) =: Q_h$, we have $\ker(B_h) \subset \ker(B)$. Let us prove (51.23). Let $q_h \in Q_h$. Using Lemma 51.2, we know that there is $\boldsymbol{\varsigma}_{q_h} \in \boldsymbol{H}(\mathrm{div};D)$ such that $\nabla{\cdot}\boldsymbol{\varsigma}_{q_h} =$

q_h and $\|\boldsymbol{\varsigma}_{q_h}\|_{\boldsymbol{L}^2(D)} \leq C_{\mathrm{PS}}^{-1}\ell_D\|q_h\|_{L^2(D)}$. Let us set $\boldsymbol{\varsigma}_h^* := \mathcal{J}_h^{\mathrm{d}}(\boldsymbol{\varsigma}_{q_h})$. The commuting property in Theorem 23.12 implies that $\nabla{\cdot}\boldsymbol{\varsigma}_h^* = \mathcal{J}_h^{\mathrm{b}}(\nabla{\cdot}\boldsymbol{\varsigma}_{q_h}) = \mathcal{J}_h^{\mathrm{b}}(q_h) = q_h$ (since $P_k^{\mathrm{b}}(\mathcal{T}_h)$ is pointwise invariant under $\mathcal{J}_h^{\mathrm{b}}$). Since $\|\boldsymbol{\varsigma}_h^*\|_{\boldsymbol{L}^2(D)} \leq \|\mathcal{J}_h^{\mathrm{d}}\|_{\mathcal{L}(\boldsymbol{L}^2;\boldsymbol{L}^2)}\|\boldsymbol{\varsigma}_{q_h}\|_{\boldsymbol{L}^2(D)}$, we infer that $\|\boldsymbol{\varsigma}_h^*\|_{\boldsymbol{H}(\mathrm{div};D)} \leq \ell_D(\beta_D^\flat)^{-1}\|q_h\|_{L^2(D)}$. As a result, we have

$$\sup_{\boldsymbol{\varsigma}_h \in \boldsymbol{V}_h} \frac{\int_D q_h \nabla{\cdot}\boldsymbol{\varsigma}_h \,\mathrm{d}x}{\|\boldsymbol{\varsigma}_h\|_{\boldsymbol{H}(\mathrm{div};D)}} \geq \frac{\int_D q_h \nabla{\cdot}\boldsymbol{\varsigma}_h^* \,\mathrm{d}x}{\|\boldsymbol{\varsigma}_h^*\|_{\boldsymbol{H}(\mathrm{div};D)}} = \frac{\|q_h\|_{L^2(D)}^2}{\|\boldsymbol{\varsigma}_h^*\|_{\boldsymbol{H}(\mathrm{div};D)}} \geq \ell_D^{-1}\beta_D^\flat\|q_h\|_{L^2(D)},$$

and this proves (51.23). $\qquad\square$

Corollary 51.11 (Well-posedness). (51.21) *is well-posed.*

Proof. We apply Proposition 50.1. The condition (50.4a) on the bilinear form a follows from the coercivity of a on $\ker(B)$ and $\ker(B_h) \subset \ker(B)$, whereas the condition (50.4b) on the bilinear form b is just (51.23). $\qquad\square$

Remark 51.12 (Discrete inf-sup). Using the $\boldsymbol{L}^2$-norm of $\boldsymbol{\varsigma}_h$ instead of the $\boldsymbol{H}(\mathrm{div})$-norm in the proof of Lemma 51.10, one can show that

$$\inf_{q_h \in Q_h} \sup_{\boldsymbol{\varsigma}_h \in \boldsymbol{V}_h} \frac{|\int_D q_h \nabla{\cdot}\boldsymbol{\varsigma}_h \,\mathrm{d}x|}{\|q_h\|_{L^2(D)}\|\boldsymbol{\varsigma}_h\|_{\boldsymbol{L}^2(D)}} \geq \ell_D^{-1}\beta_D^\sharp,$$

where $\beta_D^\sharp := C_{\mathrm{PS}}\|\mathcal{J}_h^{\mathrm{d}}\|_{\mathcal{L}(\boldsymbol{L}^2;\boldsymbol{L}^2)}^{-1} > \beta_D^\flat$. We will use this somewhat sharper bound in the proof of the error estimate in Theorem 51.16. $\qquad\square$

Remark 51.13 ($\mathcal{I}_h^{\mathrm{d}}$ vs. $\mathcal{J}_h^{\mathrm{d}}$). One can use the canonical interpolation operator $\mathcal{I}_h^{\mathrm{d}}$ instead of $\mathcal{J}_h^{\mathrm{d}}$ to prove (51.23). Owing to the theory of elliptic regularity in Lipschitz domains, the function $\boldsymbol{\varsigma}_{q_h}$ constructed in Lemma 51.2 is indeed in $\boldsymbol{L}^p(D)$ for some $p > 2$. Proposition 17.3 then implies that $\boldsymbol{\varsigma}_{q_h}$ is in the domain of $\mathcal{I}_h^{\mathrm{d}}$, and the commuting property results from Lemma 19.6. $\qquad\square$

Remark 51.14 (Fortin operator). Since $\nabla{\cdot} : \boldsymbol{H}(\mathrm{div};D) \to L^2(D)$ is surjective, $\nabla{\cdot}$ admits a bounded right inverse which we denote by $(\nabla{\cdot})^\dagger$. Then $\Pi_h = \mathcal{J}_h^{\mathrm{d}} \circ (\nabla{\cdot})^\dagger \circ \mathcal{I}_h^{\mathrm{b}} \circ (\nabla{\cdot})$ is a Fortin operator, where $\mathcal{I}_h^{\mathrm{b}}$ is the L^2-orthogonal projection onto the broken polynomial space $P_k^{\mathrm{b}}(\mathcal{T}_h)$; see Exercise 51.6. $\qquad\square$

Remark 51.15 (Variants). Other boundary conditions can be treated. For mixed Dirichlet–Neumann conditions, for instance, one assumes that the boundary faces are located on either ∂D_{d} or ∂D_{n}. Then an $\boldsymbol{H}_{\mathrm{n}}(\mathrm{div};D)$-conforming subspace is built by taking fields in $\boldsymbol{P}_k^{\mathrm{d}}(\mathcal{T}_h)$ with zero normal component on ∂D_{n}. It is also possible to work with rectangular meshes using Cartesian Raviart–Thomas elements of degree $k \geq 0$ for the dual variable and $\mathbb{Q}_{k,d}$ polynomials for the primal variable. $\qquad\square$

51.3.2 Error analysis

The error analysis presented in this section follows the general ideas of §50.1.2. We also exploit the particular structure of Darcy's equations to derive more specific estimates that give bounds on the dual variable in $\boldsymbol{H}(\mathrm{div})$ that are independent of the discrete inf-sup constant, which is not the case of the estimate derived in §50.1.2.

Theorem 51.16 (Error estimate). *Let* $(\boldsymbol{\sigma}, p)$ *and* $(\boldsymbol{\sigma}_h, p_h)$ *solve* (51.6) *and* (51.21), *respectively. Let* $\boldsymbol{V}_{h,g} := \{\varsigma_h \in \boldsymbol{V}_h \mid B_h(\varsigma_h) = \mathcal{I}_h^{\mathrm{b}}(g)\}$. (i) *We have*

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)} \leq c_1 \inf_{\varsigma_h \in \boldsymbol{V}_{h,g}} \|\boldsymbol{\sigma} - \varsigma_h\|_{\boldsymbol{L}^2(D)},$$

$$\|\nabla\cdot(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_{L^2(D)} = \inf_{\phi_h \in P_k^{\mathrm{b}}(\mathcal{T}_h)} \|\nabla\cdot\boldsymbol{\sigma} - \phi_h\|_{L^2(D)},$$

$$\|p - p_h\|_{L^2(D)} \leq c_3 \inf_{\varsigma_h \in \boldsymbol{V}_{h,g}} \ell_D \|\boldsymbol{\sigma} - \varsigma_h\|_{\boldsymbol{L}^2(D)} + 2 \inf_{q_h \in Q_h} \|p - q_h\|_{L^2(D)},$$

with $c_1 := \frac{\lambda_\sharp}{\lambda_\flat}$ *and* $c_3 := \frac{c_1}{\lambda_\flat}(\beta_D^\sharp)^{-1}$ *with* $\beta_D^\sharp$ *from Remark 51.12.* (ii) *If* $\boldsymbol{\sigma} \in \boldsymbol{H}^r(D)$, $\nabla\cdot\boldsymbol{\sigma} \in H^r(D)$, *and* $p \in H^r(D)$ *with* $r \in (0, k+1]$, *then*

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)} \leq c \left(\sum_{K \in \mathcal{T}_h} h_K^{2r} |\boldsymbol{\sigma}|_{\boldsymbol{H}^r(D_K)}^2 \right)^{\frac{1}{2}},$$

$$\|\nabla\cdot(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_{L^2(D)} \leq c \left(\sum_{K \in \mathcal{T}_h} h_K^{2r} |\nabla\cdot\boldsymbol{\sigma}|_{\boldsymbol{H}^r(K)}^2 \right)^{\frac{1}{2}},$$

$$\|p - p_h\|_{L^2(D)} \leq c \left(\sum_{K \in \mathcal{T}_h} h_K^{2r} \left(\ell_D^2 |\boldsymbol{\sigma}|_{\boldsymbol{H}^r(D_K)}^2 + |p|_{H^r(K)}^2 \right) \right)^{\frac{1}{2}},$$

where D_K *is the set of the points composing the mesh cells that share at least one face with* $K \in \mathcal{T}_h$ *(one can replace* D_K *by* K *if* $r > \frac{1}{2}$*). In particular, we have* $\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)} \leq ch^r |\boldsymbol{\sigma}|_{\boldsymbol{H}^r(D)}$, $\|\nabla\cdot(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_{L^2(D)} \leq ch^r |\nabla\cdot\boldsymbol{\sigma}|_{H^r(D)}$, *and* $\|p - p_h\|_{L^2(D)} \leq ch^r(\ell_D |\boldsymbol{\sigma}|_{\boldsymbol{H}^r(D)} + |p|_{H^r(D)})$.

Proof. (1) We first observe that we have the following Galerkin orthogonality:

$$a(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \boldsymbol{\tau}_h) + b(\boldsymbol{\tau}_h, p - p_h) = 0, \qquad \forall \boldsymbol{\tau}_h \in \boldsymbol{V}_h, \tag{51.24a}$$

$$b(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, q_h) = 0, \qquad \forall q_h \in Q_h. \tag{51.24b}$$

Since $(B_h(\boldsymbol{\sigma}_h) - \mathcal{I}_h^{\mathrm{b}}(g), q_h)_{L^2(D)} = b(\boldsymbol{\sigma}_h - \boldsymbol{\sigma}, q_h) = 0$ for all $q_h \in Q_h$, owing to (51.24b), and $B_h(\boldsymbol{\sigma}_h) - \mathcal{I}_h^{\mathrm{b}}(g) \in Q_h$, we infer that $\boldsymbol{\sigma}_h \in \boldsymbol{V}_{h,g}$.
(2) Let $\varsigma_h \in \boldsymbol{V}_{h,g}$ so that $\boldsymbol{\sigma}_h - \varsigma_h \in \ker(B_h) \subset \ker(B)$. Since (51.24a) implies that $a(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \boldsymbol{\tau}_h) = 0$ for all $\boldsymbol{\tau}_h \in \ker(B_h)$, we infer that

$$\lambda_\sharp^{-1} \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)}^2 \leq a(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \boldsymbol{\sigma} - \boldsymbol{\sigma}_h)$$
$$= a(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \boldsymbol{\sigma} - \varsigma_h) + a(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \varsigma_h - \boldsymbol{\sigma}_h)$$
$$= a(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \boldsymbol{\sigma} - \varsigma_h) \leq \lambda_\flat^{-1} \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)} \|\boldsymbol{\sigma} - \varsigma_h\|_{\boldsymbol{L}^2(D)}.$$

Thus, $\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)} \leq c_1 \|\boldsymbol{\sigma} - \varsigma_h\|_{\boldsymbol{L}^2(D)}$, and the expected bound on $\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)}$ follows by taking the infimum over $\varsigma_h \in \boldsymbol{V}_{h,g}$.
(3) Since $\nabla\cdot\boldsymbol{\sigma}_h = B_h(\boldsymbol{\sigma}_h) = \mathcal{I}_h^{\mathrm{b}}(g)$ and $\mathcal{I}_h^{\mathrm{b}}$ is the L^2-orthogonal projection onto $P_k^{\mathrm{b}}(\mathcal{T}_h)$, we have $\nabla\cdot(\boldsymbol{\sigma}_h - \boldsymbol{\sigma}) = \mathcal{I}_h^{\mathrm{b}}(g) - g$. The optimal bound on $\|\nabla\cdot(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_{L^2(D)}$ follows readily.
(4) Let $q_h \in Q_h$. Owing to Remark 51.12, we infer that there is $\boldsymbol{\tau}_h \in \boldsymbol{V}_h$ such that $\nabla\cdot\boldsymbol{\tau}_h = -(p_h - q_h)$ and $\ell_D^{-1} \beta_D^\sharp \|\boldsymbol{\tau}_h\|_{\boldsymbol{L}^2(D)} \leq \|p_h - q_h\|_{L^2(D)}$. We infer that

$$\|p_h - q_h\|_{L^2(D)}^2 = b(\boldsymbol{\tau}_h, p_h - q_h) = b(\boldsymbol{\tau}_h, p_h - p) + b(\boldsymbol{\tau}_h, p - q_h)$$
$$= a(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \boldsymbol{\tau}_h) + b(\boldsymbol{\tau}_h, p - q_h)$$
$$\leq \lambda_\flat^{-1} \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)} \|\boldsymbol{\tau}_h\|_{\boldsymbol{L}^2(D)} + \|\nabla\cdot\boldsymbol{\tau}_h\|_{L^2(D)} \|p - q_h\|_{L^2(D)},$$

owing to (51.24a). The above properties of $\boldsymbol{\tau}_h$ combined with the above bound on $\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)}$ and the triangle inequality lead to the expected bound on $\|p - p_h\|_{L^2(D)}$.

(5) The convergence rate on $\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)}$ is obtained by taking $\boldsymbol{\varsigma}_h := \mathcal{J}_h^{\mathrm{d}}(\boldsymbol{\sigma})$ and by using the approximation properties of $\mathcal{J}_h^{\mathrm{d}}$, which result from Item (iii) in Theorem 23.12 and from Theorem 22.6. Note that $\boldsymbol{\varsigma}_h \in \boldsymbol{V}_{h,g}$ since $\nabla{\cdot}\mathcal{J}_h^{\mathrm{d}}(\boldsymbol{\sigma}) = \mathcal{I}_h^{\mathrm{b}}(\nabla{\cdot}\boldsymbol{\sigma}) = \mathcal{I}_h^{\mathrm{b}}(g)$. The other two estimates follow from the estimate on $\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)}$ and the approximation properties of $\mathcal{I}_h^{\mathrm{b}}$. $\square$

We now use duality techniques to derive an estimate on the primal variable with a better rate of convergence. One difference with the primal formulation analyzed in §32.2 is that we now bound the (discrete) error $\mathcal{I}_h^{\mathrm{b}}(p) - p_h$, instead of the full error $p - p_h$. As in §32.2, we assume that the following *smoothing property* is satisfied: There are real numbers $s \in (0,1]$ and c_{smo} such that the (adjoint) solution $z_\phi \in H_0^1(D)$ to the PDE $-\nabla{\cdot}(\mathbb{d}\nabla z_\phi) = \phi$ for all $\phi \in L^2(D)$, satisfies the a priori bound $\|z_\phi\|_{H^{1+s}(D)} \le c_{\mathrm{smo}}\ell_D^{-2}\|\phi\|_{L^2(D)}$. Sufficient conditions for this smoothing property are given by the elliptic regularity theory; see §31.4. We also assume that $\mathbb{d}$ is s.t. the map $\boldsymbol{H}^s(D) \ni \boldsymbol{\xi} \mapsto \mathbb{d}{\cdot}\boldsymbol{\xi} \in \boldsymbol{H}^s(D)$ is bounded (see (31.34)).

Theorem 51.17 (Potential supercloseness). *Under the above smoothing property and multiplier assumption, the following holds true:*

$$\|\mathcal{I}_h^{\mathrm{b}}(p) - p_h\|_{L^2(D)} \le c\,h^s\ell_D^{1-s}(\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)} + h\|\nabla{\cdot}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_{L^2(D)}). \tag{51.25}$$

Proof. Let $z_\phi \in H^{1+s}(D) \cap H_0^1(D)$ be the adjoint solution with data $\phi := \mathcal{I}_h^{\mathrm{b}}(p) - p_h$, i.e., $-\nabla{\cdot}(\mathbb{d}\nabla z_\phi) = \mathcal{I}_h^{\mathrm{b}}(p) - p_h$, and let us set $\boldsymbol{\xi} := -\mathbb{d}\nabla z_\phi$ so that $\nabla{\cdot}\boldsymbol{\xi} = \mathcal{I}_h^{\mathrm{b}}(p) - p_h$. We observe that

$$\begin{aligned}
\|\mathcal{I}_h^{\mathrm{b}}(p) - p_h\|_{L^2(D)}^2 &= (p - p_h, \mathcal{I}_h^{\mathrm{b}}(p) - p_h)_{L^2(D)} = (p - p_h, \nabla{\cdot}\boldsymbol{\xi})_{L^2(D)} \\
&= (p - p_h, \nabla{\cdot}\mathcal{J}_h^{\mathrm{d}}(\boldsymbol{\xi}))_{L^2(D)} = (\mathbb{d}^{-1}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), \mathcal{J}_h^{\mathrm{d}}(\boldsymbol{\xi}))_{\boldsymbol{L}^2(D)} \\
&= (\mathbb{d}^{-1}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), \mathcal{J}_h^{\mathrm{d}}(\boldsymbol{\xi}) - \boldsymbol{\xi})_{\boldsymbol{L}^2(D)} + (\mathbb{d}^{-1}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), \boldsymbol{\xi})_{\boldsymbol{L}^2(D)},
\end{aligned}$$

where in the first line we used that $(p - \mathcal{I}_h^{\mathrm{b}}(p), q_h)_{L^2(D)} = 0$ for all $q_h \in P_k^{\mathrm{b}}(\mathcal{T}_h)$ and $\nabla{\cdot}\boldsymbol{\xi} = \mathcal{I}_h^{\mathrm{b}}(p) - p_h$, and in the second line we used that $\nabla{\cdot}\mathcal{J}_h^{\mathrm{d}}(\boldsymbol{\xi}) = \mathcal{J}_h^{\mathrm{b}}(\nabla{\cdot}\boldsymbol{\xi}) = \nabla{\cdot}\boldsymbol{\xi}$, since $\nabla{\cdot}\boldsymbol{\xi} \in P_k^{\mathrm{b}}(\mathcal{T}_h)$, and the identity (51.24a) with $\boldsymbol{\tau}_h := \mathcal{J}_h^{\mathrm{d}}(\boldsymbol{\xi})$. The first term on the right-hand side, say $\mathfrak{T}_1$, is bounded as follows:

$$|\mathfrak{T}_1| \le \lambda_\flat^{-1}\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)}\|\mathcal{J}_h^{\mathrm{d}}(\boldsymbol{\xi}) - \boldsymbol{\xi}\|_{\boldsymbol{L}^2(D)} \le c\,\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)}h^s|\boldsymbol{\xi}|_{\boldsymbol{H}^s(D)},$$

and $|\boldsymbol{\xi}|_{\boldsymbol{H}^s(D)} \le c|z_\phi|_{H^{1+s}(D)}$ owing to the multiplier assumption. For the second term, say $\mathfrak{T}_2$, (51.24b) implies that

$$\begin{aligned}
(\boldsymbol{\sigma}_h - \boldsymbol{\sigma}, \nabla z_\phi)_{\boldsymbol{L}^2(D)} &= (\nabla{\cdot}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), z_\phi)_{L^2(D)} \\
&= (\nabla{\cdot}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), z_\phi - \mathcal{I}_h^{\mathrm{b}}(z_\phi))_{L^2(D)}.
\end{aligned}$$

Hence, $|\mathfrak{T}_2| \le c\|\nabla{\cdot}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_{L^2(D)}h^{1+s}|z_\phi|_{H^{1+s}(D)}$. Finally, we have

$$|z_\phi|_{H^{1+s}(D)} \le \ell_D^{-1-s}\|z_\phi\|_{H^{1+s}(D)} \le c_{\mathrm{smo}}\,\ell_D^{1-s}\|\mathcal{I}_h^{\mathrm{b}}(p) - p_h\|_{L^2(D)},$$

owing to the smoothing property. $\square$

Remark 51.18 (L^∞-bounds). L^∞-bounds on the dual and primal errors can be derived as in, e.g., Gastaldi and Nochetto [211]. $\square$

Exercises

Exercise 51.1 (Compactness). Let $D := (0,1)^3$ be the unit cube in $\mathbb{R}^3$. Show that the embedding $\boldsymbol{H}_0(\text{div}; D) \hookrightarrow \boldsymbol{L}^2(D)$ is not compact. (*Hint*: let

$$\boldsymbol{\phi}_{1,n}(x_1, x_2, x_3) := \frac{1}{n\pi} \sin(n\pi x_2) \sin(n\pi x_3),$$

$$\boldsymbol{\phi}_{2,n}(x_1, x_2, x_3) := \frac{1}{n\pi} \sin(n\pi x_3) \sin(n\pi x_1),$$

$$\boldsymbol{\phi}_{3,n}(x_1, x_2, x_3) := \frac{1}{n\pi} \sin(n\pi x_1) \sin(n\pi x_2),$$

for all $n \geq 1$, set $\boldsymbol{v}_n := \nabla \times \boldsymbol{\phi}_n$, and prove first that $(\boldsymbol{v}_n)_{n \geq 1}$ weakly converges to zero in $\boldsymbol{L}^2(D)$ (see Definition C.28), then compute $\|\boldsymbol{v}_n\|_{\boldsymbol{L}^2(D)}$ and argue by contradiction.)

Exercise 51.2 (Neumann condition). Prove Proposition 51.3. (*Hint*: for the surjectivity of the divergence, solve a pure Neumann problem.)

Exercise 51.3 (Integration by parts). Let $H^1_\text{d}(D)$ and $\boldsymbol{H}_\text{n}(\text{div}; D)$ be defined in §51.1.3. Prove that $\int_D (\nabla q \cdot \boldsymbol{\varsigma} + q\nabla\cdot\boldsymbol{\varsigma}) \, dx = 0$ for all $q \in H^1_\text{d}(D)$ and all $\boldsymbol{\varsigma} \in \boldsymbol{H}_\text{n}(\text{div}; D)$. (*Hint*: observe that $\gamma^\text{g}(q)_{|\partial D_\text{n}} \in \widetilde{H}^{\frac{1}{2}}(\partial D_\text{n})$.)

Exercise 51.4 (Primal, dual formulations). Prove Proposition 51.7.

Exercise 51.5 (Primal mixed formulation). Consider the problem: Find $p \in H^1(D)$ such that $-\Delta p = f$ and $\gamma^\text{g}(p) = g$ with $f \in L^2(D)$ and $g \in H^{\frac{1}{2}}(\partial D)$. Derive a mixed formulation of this problem with unknowns $(p, \lambda) \in H^1(D) \times H^{-\frac{1}{2}}(\partial D)$ and show that it is well-posed. (*Hint*: set $b(v, \mu) := \langle \mu, \gamma^\text{g}(v) \rangle_{\partial D}$ and observe that $B = \gamma^\text{g} : H^1(D) \to H^{\frac{1}{2}}(\partial D)$.) Recover the PDE and the boundary condition. *Note*: this method is introduced in Babuška [34].

Exercise 51.6 (Fortin operator). Justify Remark 51.14. (*Hint*: use arguments similar to those of the proof of Lemma 51.10.)

Exercise 51.7 (Inf-sup condition). The goal is to prove the inf-sup condition (51.23) using the canonical Raviart–Thomas interpolation operator. (i) Do this by using elliptic regularity. (*Hint*: solve a Dirichlet problem.) (ii) Do this again by using the surjectivity of $\nabla\cdot : \boldsymbol{H}^1(D) \to L^2(D)$.

Exercise 51.8 (Error estimate). (i) Prove that

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\boldsymbol{H}(\text{div};D)} \leq c'_1 \inf_{\boldsymbol{\varsigma}_h \in \boldsymbol{V}_h} \|\boldsymbol{\sigma} - \boldsymbol{\varsigma}_h\|_{\boldsymbol{H}(\text{div};D)},$$

$$\|p - p_h\|_{L^2(D)} \leq c'_3 \inf_{\boldsymbol{\varsigma}_h \in \boldsymbol{V}_h} \|\boldsymbol{\sigma} - \boldsymbol{\varsigma}_h\|_{\boldsymbol{H}(\text{div};D)} + 2 \inf_{q_h \in Q_h} \|p - q_h\|_{L^2(D)},$$

with $c'_1 := (1 + \frac{\lambda_\sharp}{\lambda_\flat})(1 + \frac{1}{\beta})$ and $c'_3 := \frac{c'_1}{\lambda_\flat \beta'_{L^2}}$. (ii) Assuming that $\boldsymbol{\sigma} \in \boldsymbol{H}^r(D)$, $\nabla\cdot\boldsymbol{\sigma} \in H^r(D)$, and $p \in H^r(D)$ with $r \in (0, k+1]$, prove that

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\boldsymbol{H}(\text{div};D)} \leq c\,h^r(|\boldsymbol{\sigma}|_{\boldsymbol{H}^r(D)} + |\nabla\cdot\boldsymbol{\sigma}|_{H^r(D)}),$$

$$\|p - p_h\|_{L^2(D)} \leq c\,h^r(|\boldsymbol{\sigma}|_{\boldsymbol{H}^r(D)} + |\nabla\cdot\boldsymbol{\sigma}|_{H^r(D)} + |p|_{H^r(D)}).$$

(*Hint*: use the commuting projection $\mathcal{J}^\text{d}_h$.)

Exercise 51.9 (Box scheme). Let $\mathbb{d} := \lambda_0 \mathbb{I}_d$, $\lambda_0 > 0$, and enforce the boundary condition $\gamma^{\mathrm{g}}(p) = 0$. Let $V_h := \boldsymbol{P}_0^{\mathrm{d}}(\mathcal{T}_h) \times P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$, where $P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$ is the Crouzeix–Raviart space defined in (36.8). Let $W_h := \boldsymbol{P}_0^{\mathrm{b}}(\mathcal{T}_h) \times P_0^{\mathrm{b}}(\mathcal{T}_h)$. Consider the bilinear form $a_h : V_h \times W_h \to \mathbb{R}$ defined by $a_h(v_h, w_h) := \lambda_0^{-1}(\boldsymbol{\sigma}_h, \boldsymbol{\tau}_h)_{\boldsymbol{L}^2(D)} + (\nabla \cdot \boldsymbol{\sigma}_h, q_h)_{L^2(D)} + (\nabla_h p_h, \boldsymbol{\tau}_h)_{\boldsymbol{L}^2(D)}$ with $v_h := (\boldsymbol{\sigma}_h, p_h)$ and $w_h := (\boldsymbol{\tau}_h, q_h)$ (see Definition 36.3 for the broken gradient ∇_h). (i) Prove that $\dim(V_h) = \dim(W_h)$ and that there is $\alpha > 0$ s.t. for all $v_h \in V_h$ and all $h \in \mathcal{H}$, $\alpha \|v_h\|_{V_h} \leq \sup_{w_h \in W_h} \frac{|a_h(v_h, w_h)|}{\|w_h\|_{W_h}}$ with $\|v_h\|_{V_h}^2 := \lambda_0^{-1} \|\boldsymbol{\sigma}_h\|_{\boldsymbol{H}(\mathrm{div};D)}^2 + \lambda_0 \|\nabla_h p_h\|_{\boldsymbol{L}^2(D)}^2$ and $\|w_h\|_{W_h}^2 := \lambda_0^{-1} \|\boldsymbol{\tau}_h\|_{\boldsymbol{L}^2(D)}^2 + \lambda_0 \ell_D^{-2} \|q_h\|_{L^2(D)}^2$. (*Hint*: test with $(\underline{\boldsymbol{\sigma}}_h + \lambda_0 \nabla_h p_h, 2\underline{p}_h + \ell_D^2 \lambda_0^{-1} \nabla \cdot \boldsymbol{\sigma}_h)$, where $(\underline{\boldsymbol{\sigma}}_h, \underline{p}_h)$ is the L^2-orthogonal projection of $(\boldsymbol{\sigma}_h, p_h)$ onto W_h.) (ii) Consider the discrete problem: Find $u_h \in V_h$ such that $a_h(u_h, w_h) = (\boldsymbol{f}, \boldsymbol{\tau}_h)_{\boldsymbol{L}^2(D)} + (g, q_h)_{L^2(D)}$ for all $w_h \in W_h$. Show that this problem is well-posed, prove a quasi-optimal error estimate, and show that the error converges to zero with rate h if the exact solution is smooth enough. (*Hint*: use Lemma 27.5.) *Note*: the scheme has been introduced in Croisille [149] to approximate (51.1). It is a Petrov–Galerkin scheme with only local test functions.

Chapter 52

Potential and flux recovery

This chapter addresses topics related to the approximation of Darcy's equations using either mixed or H^1-conforming finite elements. Mixed finite elements approximate the flux (i.e., the dual variable $\boldsymbol{\sigma}$) in $\boldsymbol{H}(\mathrm{div}; D)$, but the connection to the gradient of the potential (i.e., the primal variable p) sitting in $H^1_0(D)$ is enforced weakly. We show here how this connection can be made explicit using hybridization techniques. Alternatively, H^1-conforming finite elements approximate the primal variable in $H^1_0(D)$, but the connection to the dual variable $\boldsymbol{\sigma}$ sitting in $\boldsymbol{H}(\mathrm{div}; D)$ is enforced weakly. We show here how this connection can be made explicit by using a local post-processing technique. In the whole chapter, we consider homogeneous Dirichlet boundary conditions on the potential for simplicity, and we assume that $(\mathcal{T}_h)_{h\in\mathcal{H}}$ is a shape-regular sequence of affine simplicial meshes so that each mesh covers the domain $D \subset \mathbb{R}^d$ exactly.

52.1 Hybridization of mixed finite elements

Hybridization was introduced by Fraejis de Veubeke in 1965 (see [207] for a reprint) as a computationally effective technique to transform the symmetric indefinite linear system on the potential and the flux into a symmetric positive definite system on an auxiliary variable playing the role of a Lagrange multiplier associated with the continuity constraint on the normal component of the flux. As shown in the seminal work of Arnold and Brezzi [16], viewing the auxiliary variable as a potential trace on the mesh faces allows one to devise a post-processed potential with better approximation properties.

52.1.1 From hybridization to static condensation

Let us focus on the same discrete setting as in §51.3, where we employed simplicial $\mathbb{RT}_{k,d}$ Raviart–Thomas elements for the flux and broken $\mathbb{P}_{k,d}$ finite elements for the potential with some polynomial degree $k \geq 0$, i.e.,

$$\boldsymbol{V}_h := \boldsymbol{P}^{\mathrm{d}}_k(\mathcal{T}_h) := \{\boldsymbol{\varsigma}_h \in \boldsymbol{H}(\mathrm{div}; D) \mid \boldsymbol{\psi}^{\mathrm{d}}_K(\boldsymbol{\varsigma}_{h|K}) \in \mathbb{RT}_{k,d}, \ \forall K \in \mathcal{T}_h\}, \tag{52.1a}$$

$$Q_h := P^{\mathrm{b}}_k(\mathcal{T}_h) := \{q_h \in L^2(D) \mid \psi^{\mathrm{g}}_K(q_{h|K}) \in \mathbb{P}_{k,d}, \ \forall K \in \mathcal{T}_h\}, \tag{52.1b}$$

where $\boldsymbol{\psi}^{\mathrm{d}}_K$ is the contravariant Piola transformation and ψ^{g}_K is the pullback by the geometric mapping $\boldsymbol{T}_K$ (see Definition 9.8). Recall that the bilinear forms are $a(\boldsymbol{\varsigma}_h, \boldsymbol{\tau}_h) := \int_D (\mathbb{d}^{-1}\boldsymbol{\varsigma}_h)\cdot\boldsymbol{\tau}_h\,\mathrm{d}x$

and $b(\boldsymbol{\varsigma}_h, q_h) := -\int_D (\nabla \cdot \boldsymbol{\varsigma}_h) q_h \, \mathrm{d}x$, and that the linear forms are $F(\boldsymbol{\tau}_h) := \int_D \boldsymbol{\tau}_h \cdot \boldsymbol{f} \, \mathrm{d}x$ and $G(q_h) := -\int_D q_h g \, \mathrm{d}x$. Our starting point is the discrete problem (51.21) which consists of seeking the pair $(\boldsymbol{\sigma}_h, p_h) \in \boldsymbol{V}_h \times Q_h$ such that

$$a(\boldsymbol{\sigma}_h, \boldsymbol{\tau}_h) + b(\boldsymbol{\tau}_h, p_h) = F(\boldsymbol{\tau}_h), \quad \forall \boldsymbol{\tau}_h \in \boldsymbol{V}_h, \tag{52.2a}$$

$$b(\boldsymbol{\sigma}_h, q_h) \quad\quad\quad = G(q_h), \quad \forall q_h \in Q_h. \tag{52.2b}$$

Recall that this problem is well-posed (see Corollary 51.11) and gives optimal error estimates (see Theorem 51.16).

Let Λ_h be the space composed of the functions that are piecewise polynomials of degree at most k on the mesh interfaces and are extended by zero on all the boundary faces, i.e.,

$$\Lambda_h := \{\lambda_h \in L^2(\mathcal{F}_h) \mid \lambda_h \circ \boldsymbol{T}_F \in \mathbb{P}_{k,d-1}, \, \forall F \in \mathcal{F}_h^\circ, \, \lambda_{h|\partial D} = 0\}, \tag{52.3}$$

where $\boldsymbol{T}_F$ is an affine bijective mapping from the unit simplex of $\mathbb{R}^{d-1}$ to F. Let $\boldsymbol{V}_h^{\mathrm{hy}} := \boldsymbol{P}_k^{\mathrm{d,b}}(\mathcal{T}_h) := \{\boldsymbol{\varsigma}_h \in \boldsymbol{L}^2(D) \mid \boldsymbol{\psi}_K^{\mathrm{d}}(\boldsymbol{\varsigma}_{h|K}) \in \mathbb{RT}_{k,d}, \, \forall K \in \mathcal{T}_h\}$ be the broken Raviart–Thomas space (this space is composed of piecewise $\mathbb{RT}_{k,d}$ polynomials in each mesh cell since the mesh is affine). Recall that $\boldsymbol{V}_h = \boldsymbol{H}(\mathrm{div}; D) \cap \boldsymbol{V}_h^{\mathrm{hy}}$ (see §18.2.3). We define the bilinear form $b_h(\boldsymbol{\varsigma}_h, q_h) := -\sum_{K \in \mathcal{T}_h} \int_K (\nabla \cdot \boldsymbol{\varsigma}_{h|K}) q_h \, \mathrm{d}x$ on $\boldsymbol{V}_h^{\mathrm{hy}} \times Q_h$ and observe that $b_{h|\boldsymbol{V}_h \times Q_h} = b$. The hybridized version of the discrete problem (51.21) consists of seeking the triple $(\boldsymbol{\sigma}_h', p_h', \lambda_h) \in \boldsymbol{V}_h^{\mathrm{hy}} \times Q_h \times \Lambda_h$ such that

$$a(\boldsymbol{\sigma}_h', \boldsymbol{\tau}_h) + b_h(\boldsymbol{\tau}_h, p_h') + c_h(\boldsymbol{\tau}_h, \lambda_h) = F(\boldsymbol{\tau}_h), \quad \forall \boldsymbol{\tau}_h \in \boldsymbol{V}_h^{\mathrm{hy}}, \tag{52.4a}$$

$$b_h(\boldsymbol{\sigma}_h', q_h) \quad\quad\quad\quad\quad = G(q_h), \quad \forall q_h \in Q_h, \tag{52.4b}$$

$$c_h(\boldsymbol{\sigma}_h', \mu_h) \quad\quad\quad\quad\quad = 0, \quad\quad \forall \mu_h \in \Lambda_h, \tag{52.4c}$$

with the bilinear form

$$c_h(\boldsymbol{\tau}_h, \mu_h) := \sum_{F \in \mathcal{F}_h^\circ} \int_F [\![\boldsymbol{\tau}_h]\!] \cdot \boldsymbol{n}_F \mu_h \, \mathrm{d}s. \tag{52.5}$$

Proposition 52.1 (Equivalence). *The discrete problem* (52.4) *admits a unique solution*

$$(\boldsymbol{\sigma}_h', p_h', \lambda_h) \in \boldsymbol{V}_h^{\mathrm{hy}} \times Q_h \times \Lambda_h.$$

Moreover, $\sigma_h' \in \boldsymbol{V}_h$, *and the pair* $(\boldsymbol{\sigma}_h', p_h')$ *is the unique solution to* (52.2).

Proof. The well-posedness of (52.4) is treated in Exercise 52.1. Assume that $(\boldsymbol{\sigma}_h', p_h', \lambda_h) \in \boldsymbol{V}_h^{\mathrm{hy}} \times Q_h \times \Lambda_h$ is the unique solution to (52.4). Equation (52.4c) implies that the normal component of $\boldsymbol{\sigma}_h'$ is continuous, since Lemma 14.7 shows that $(\boldsymbol{\sigma}_h' \cdot \boldsymbol{n}_F) \circ \boldsymbol{T}_F$ is in $\mathbb{P}_{k,d-1}$ for all $F \in \mathcal{F}_h$. Owing to Theorem 18.10, we infer that $\boldsymbol{\sigma}_h' \in \boldsymbol{H}(\mathrm{div}; D)$, i.e., $\boldsymbol{\sigma}_h' \in \boldsymbol{V}_h$. Restricting the test functions in (52.4a) to be in $\boldsymbol{V}_h$ so that the bilinear form b_h can be replaced by b, and using (52.4b), shows that the pair $(\boldsymbol{\sigma}_h', p_h')$ solves (52.2). Uniqueness of the solution to (52.2) shows that $(\boldsymbol{\sigma}_h', p_h') = (\boldsymbol{\sigma}_h, p_h)$. $\square$

Owing to the equivalence result stated in Proposition 52.1, we drop the primes from now on in the discrete problem (52.4). The advantage of (52.4) over (52.2) is that the pair $(\boldsymbol{\sigma}_h, p_h)$ can be eliminated locally. This operation is called *static condensation* (see §28.1.2 and §39.2.2). Let $\mathbb{V}_K^k := (\boldsymbol{\psi}_K^{\mathrm{d}})^{-1}(\mathbb{RT}_{k,d}) \times (\psi_K^{\mathrm{g}})^{-1}(\mathbb{P}_{k,d})$ (since the mesh is affine, we have $\mathbb{V}_K^k := \mathbb{RT}_{k,d} \times \mathbb{P}_{k,d}$). We define the following local bilinear form on $\mathbb{V}_K^k \times \mathbb{V}_K^k$:

$$\hat{a}_K((\boldsymbol{\sigma}, p), (\boldsymbol{\tau}, q)) := (\mathrm{d}^{-1}\boldsymbol{\sigma}, \boldsymbol{\tau})_{\boldsymbol{L}^2(K)} - (\nabla \cdot \boldsymbol{\tau}, p)_{L^2(K)} - (\nabla \cdot \boldsymbol{\sigma}, q)_{L^2(K)}.$$

For all $\mu \in L^2(\partial K)$, we define the polynomial pair $(\boldsymbol{S}_\mu, P_\mu) \in \mathbb{V}_K^k$ by solving the following local problem: For all $(\boldsymbol{\tau}, q) \in \mathbb{V}_K^k$,

$$\hat{a}_K((\boldsymbol{S}_\mu, P_\mu), (\boldsymbol{\tau}, q)) = -(\mu, \boldsymbol{\tau} \cdot \boldsymbol{n}_K)_{L^2(\partial K)}. \tag{52.6}$$

Notice that this definition implies, in particular, that $\nabla \cdot \boldsymbol{S}_\mu = 0$. For all $(\boldsymbol{\phi}, \psi) \in \boldsymbol{L}^2(K) \times L^2(K)$, we define the polynomial pair $(\boldsymbol{S}_{\boldsymbol{\phi}, \psi}, P_{\boldsymbol{\phi}, \psi}) \in \mathbb{V}_K^k$ by solving the following local problem: For all $(\boldsymbol{\tau}, q) \in \mathbb{V}_K^k$,

$$\hat{a}_K((\boldsymbol{S}_{\boldsymbol{\phi}, \psi}, P_{\boldsymbol{\phi}, \psi}), (\boldsymbol{\tau}, q)) = (\boldsymbol{\phi}, \boldsymbol{\tau})_{\boldsymbol{L}^2(K)} - (\psi, q)_{L^2(K)}. \tag{52.7}$$

Notice that both local problems are well-posed since $\hat{a}_K$ satisfies an inf-sup condition on $\mathbb{V}_K^k \times \mathbb{V}_K^k$. For all $\mu_h \in \Lambda_h$ and all $K \in \mathcal{T}_h$, we denote by $\mu_{\partial K} := (\mu_{h|F})_{F \in \mathcal{F}_K}$ the restriction of μ_h to the mesh faces in ∂K.

Proposition 52.2 (Static condensation). (i) $(\boldsymbol{\sigma}_h, p_h, \lambda_h)$ solves (52.4) if and only if $(\boldsymbol{\sigma}_h, p_h)_{|K} = (\boldsymbol{S}_{\lambda_{\partial K}}, P_{\lambda_{\partial K}}) + (\boldsymbol{S}_{\boldsymbol{f}_{|K}, g_{|K}}, P_{\boldsymbol{f}_{|K}, g_{|K}})$ for all $K \in \mathcal{T}_h$, where $\lambda_h \in \Lambda_h$ is the unique solution of

$$\sum_{K \in \mathcal{T}_h} (\mathrm{d}^{-1} \boldsymbol{S}_{\lambda_{\partial K}}, \boldsymbol{S}_{\mu_{\partial K}})_{\boldsymbol{L}^2(K)} = \ell(\mu_h), \qquad \forall \mu_h \in \Lambda_h, \tag{52.8}$$

with $\ell(\mu_h) := \sum_{K \in \mathcal{T}_h} (g, P_{\mu_{\partial K}})_{L^2(K)} - (\boldsymbol{f}, \boldsymbol{S}_{\mu_{\partial K}})_{\boldsymbol{L}^2(K)}$. (ii) The algebraic realization of (52.8) leads to a symmetric positive definite matrix.

Proof. (i) Assume that $(\boldsymbol{\sigma}_h, p_h, \lambda_h)$ solves (52.4) and let us show that $(\boldsymbol{\sigma}_h, p_h)_{|K} = (\boldsymbol{S}_{\lambda_{\partial K}}, P_{\lambda_{\partial K}}) + (\boldsymbol{S}_{\boldsymbol{f}_{|K}, g_{|K}}, P_{\boldsymbol{f}_{|K}, g_{|K}})$. Let $K \in \mathcal{T}_h$ and $(\boldsymbol{\tau}, q) \in \mathbb{V}_K^k$. Extending these functions by zero to D leads to a pair $(\tilde{\boldsymbol{\tau}}_h, \tilde{q}_h) \in \boldsymbol{V}_h^{\mathrm{hy}} \times Q_h$ that we can use as a test function in (52.4a)-(52.4b). This leads to

$$\begin{aligned}
\hat{a}_K((\boldsymbol{\sigma}_{h|K}, p_{h|K}), (\boldsymbol{\tau}, q)) &= a(\boldsymbol{\sigma}_h, \tilde{\boldsymbol{\tau}}_h) + b(\tilde{\boldsymbol{\tau}}_h, p_h) + b(\boldsymbol{\sigma}_h, \tilde{q}_h) \\
&= F(\tilde{\boldsymbol{\tau}}_h) + G(\tilde{q}_h) - c_h(\tilde{\boldsymbol{\tau}}_h, \lambda_h) \\
&= (\boldsymbol{f}, \boldsymbol{\tau})_{\boldsymbol{L}^2(K)} - (g, q)_{L^2(K)} - (\lambda_h, \boldsymbol{\tau} \cdot \boldsymbol{n}_K)_{L^2(\partial K)} \\
&= \hat{a}_K((\boldsymbol{S}_{\boldsymbol{f}_{|K}, g_{|K}}, P_{\boldsymbol{f}_{|K}, g_{|K}}), (\boldsymbol{\tau}, q)) + \hat{a}_K((\boldsymbol{S}_{\lambda_{\partial K}}, P_{\lambda_{\partial K}}), (\boldsymbol{\tau}, q)).
\end{aligned}$$

Since $(\boldsymbol{\tau}, q) \in \mathbb{V}_K^k$ is arbitrary, this proves that $(\boldsymbol{\sigma}_h, p_h)_{|K} = (\boldsymbol{S}_{\lambda_{\partial K}}, P_{\lambda_{\partial K}}) + (\boldsymbol{S}_{\boldsymbol{f}_{|K}, g_{|K}}, P_{\boldsymbol{f}_{|K}, g_{|K}})$. Let us now establish (52.8). We first observe that (52.4c) together with the definition (52.5) of the bilinear form c_h implies that the following identity holds true for all $\mu_h \in \Lambda_h$:

$$\begin{aligned}
0 = c_h(\boldsymbol{\sigma}_h, \mu_h) &= \sum_{K \in \mathcal{T}_h} (\boldsymbol{\sigma}_{h|K} \cdot \boldsymbol{n}_K, \mu_{\partial K})_{L^2(\partial K)} \tag{52.9} \\
&= \sum_{K \in \mathcal{T}_h} (\boldsymbol{S}_{\lambda_{\partial K}} \cdot \boldsymbol{n}_K, \mu_{\partial K})_{L^2(\partial K)} + (\boldsymbol{S}_{\boldsymbol{f}_{|K}, g_{|K}} \cdot \boldsymbol{n}_K, \mu_{\partial K})_{L^2(\partial K)}.
\end{aligned}$$

Using the definition of $(\boldsymbol{S}_{\mu_{\partial K}}, P_{\mu_{\partial K}})$, observing that $\nabla \cdot \boldsymbol{S}_{\lambda_{\partial K}} = \nabla \cdot \boldsymbol{S}_{\mu_{\partial K}} = 0$, and since d is symmetric, we also infer that

$$\begin{aligned}
(\boldsymbol{S}_{\lambda_{\partial K}} \cdot \boldsymbol{n}_K, \mu_{\partial K})_{L^2(\partial K)} &= -\hat{a}_K((\boldsymbol{S}_{\mu_{\partial K}}, P_{\mu_{\partial K}}), (\boldsymbol{S}_{\lambda_{\partial K}}, P_{\lambda_{\partial K}})) \\
&= -(\mathrm{d}^{-1} \boldsymbol{S}_{\lambda_{\partial K}}, \boldsymbol{S}_{\mu_{\partial K}})_{\boldsymbol{L}^2(K)}. \tag{52.10}
\end{aligned}$$

Using the definition of $(\boldsymbol{S}_{\mu_{\partial K}}, P_{\mu_{\partial K}})$, the symmetry of $\hat{a}_K$, and the definition of $(\boldsymbol{S}_{\boldsymbol{f}_{|K}, g_{|K}}, P_{\boldsymbol{f}_{|K}, g_{|K}})$, we finally infer that

$$\begin{aligned}
(\boldsymbol{S}_{\boldsymbol{f}_{|K}, g_{|K}} \cdot \boldsymbol{n}_K, \mu_{\partial K})_{L^2(\partial K)} &= -\hat{a}_K((\boldsymbol{S}_{\mu_{\partial K}}, P_{\mu_{\partial K}}), (\boldsymbol{S}_{\boldsymbol{f}_{|K}, g_{|K}}, P_{\boldsymbol{f}_{|K}, g_{|K}})) \\
&= -\hat{a}_K((\boldsymbol{S}_{\boldsymbol{f}_{|K}, g_{|K}}, P_{\boldsymbol{f}_{|K}, g_{|K}}), (\boldsymbol{S}_{\mu_{\partial K}}, P_{\mu_{\partial K}})) \\
&= -(\boldsymbol{f}, \boldsymbol{S}_{\mu_{\partial K}})_{\boldsymbol{L}^2(K)} + (g, P_{\mu_{\partial K}})_{L^2(K)}. \tag{52.11}
\end{aligned}$$

Adding the identities (52.10)-(52.11), summing the result over $K \in \mathcal{T}_h$, and adding the new identity thus obtained to (52.9) leads to (52.8). The proof of the converse statement employs similar arguments and is left to the reader.

(ii) The matrix associated with (52.8) is symmetric positive semidefinite by construction. Let us show the definiteness. Assume that $\lambda_h \in \Lambda_h$ satisfies $\sum_{K \in \mathcal{T}_h} (\mathrm{d}^{-1} \boldsymbol{S}_{\lambda_{\partial K}}, \boldsymbol{S}_{\lambda_{\partial K}})_{\boldsymbol{L}^2(K)} = 0$. Then $\boldsymbol{S}_{\lambda_{\partial K}} = \boldsymbol{0}$ for all $K \in \mathcal{T}_h$. Owing to (52.6) and the definition of the spaces $\boldsymbol{V}_h$ and $\mathbb{V}_K^k$, we infer that $\int_K P_{\lambda_{\partial K}} \nabla \cdot \boldsymbol{\tau}_{|K} \, \mathrm{d}x = \int_{\partial K} \lambda_{\partial K} \boldsymbol{\tau}_{|K} \cdot \boldsymbol{n}_K \, \mathrm{d}s$ for all $\boldsymbol{\tau} \in \boldsymbol{V}_h$. Summing over the mesh cells and since $\boldsymbol{\tau} \in \boldsymbol{V}_h$, we infer that $\sum_{K \in \mathcal{T}_h} \int_K P_{\lambda_{\partial K}} \nabla \cdot \boldsymbol{\tau} \, \mathrm{d}x = 0$. Since $\nabla \cdot \boldsymbol{V}_h = Q_h$, choosing $\boldsymbol{\tau}$ so that $\nabla \cdot \boldsymbol{\tau}_{|K} = P_{\lambda_{\partial K}}$ for all $K \in \mathcal{T}_h$ implies that $P_{\lambda_{\partial K}} = 0$. This in turn implies that $\int_{\partial K} \lambda_{\partial K} \boldsymbol{\tau} \cdot \boldsymbol{n}_K \, \mathrm{d}s = 0$ for all $\boldsymbol{\tau} \in (\boldsymbol{\psi}_K^{\mathrm{d}})^{-1}(\mathbb{RT}_{k,d})$. Since $\lambda_{\partial K} \in \gamma_{\partial K}^{\mathrm{d}}((\boldsymbol{\psi}_K^{\mathrm{d}})^{-1}(\mathbb{RT}_{k,d}))$, this argument shows that $\lambda_{\partial K} = 0$. $\qquad\square$

Remark 52.3 (Literature). The above proof is inspired from Cockburn [130], Boffi et al. [65, §7.2-7.3], and Cockburn and Gopalakrishnan [131]. $\qquad\square$

Remark 52.4 (Lowest-order $(k = 0)$). There is a close link between the lowest-order Raviart–Thomas elements and the Crouzeix–Raviart elements from Chapter 36 when d and g are piecewise constant and $\boldsymbol{f} := \boldsymbol{0}$ in (52.2); see Marini [295] and Exercise 52.2. The implementation of the lowest-order Raviart–Thomas method with one unknown per cell and connections to finite volume and mimetic finite difference methods are discussed in Younes et al. [399], Vohralík and Wohlmuth [385]. $\qquad\square$

52.1.2 From hybridization to post-processing

Let λ_h be the solution to the global "skeleton" problem (52.8). Let $(\boldsymbol{\sigma}_h, p_h) := (\boldsymbol{S}_{\lambda_h} + \boldsymbol{S}_{\boldsymbol{f}_{|K}, g_{|K}}, P_{\lambda_h} + P_{\boldsymbol{f}_{|K}, g_{|K}})$. Recall that we have shown that $(\boldsymbol{\sigma}_h, p_h)$ solves (52.2) and $(\boldsymbol{\sigma}_h, p_h, \lambda_h)$ solves (52.4). We are now going to post-process p_h and λ_h to construct a potential m_h^{nc} in a piecewise polynomial space of higher order. The superscript refers to the fact that m_h^{nc} is a nonconforming function, i.e., $m_h^{\mathrm{nc}} \notin H_0^1(D)$. However, we will see below that the jumps of m_h^{nc} across the mesh interfaces and its trace on the boundary faces have vanishing moments against polynomials of degree at most k.

Let Q_K and $\Lambda_{\partial K}$ be composed of the restriction to K and ∂K of the functions in Q_h and Λ_h, respectively. For Raviart–Thomas elements, $Q_K \circ \boldsymbol{T}_K^{-1} := \mathbb{P}_{k,d}$ and $\Lambda_{\partial K}$ is composed of piecewise polynomials of degree k on the faces of K, i.e., $\lambda_{\partial K | F} \circ \boldsymbol{T}_F \in \mathbb{P}_{k,d-1}$ for all $F \in \mathcal{F}_K$ (recall that the mesh is affine by assumption). Let Π_{Q_K} and $\Pi_{\Lambda_{\partial K}}$ denote the corresponding L^2-orthogonal projections. The post-processed potential m_h^{nc} is built locally in each mesh cell. One first picks a polynomial space $\mathbb{M}_{k,k'}$ satisfying $\mathbb{P}_{k,d} \subset \mathbb{M}_{k,k'} \subset \mathbb{P}_{k',d}$ and such that the problem of seeking $m_K \circ \boldsymbol{T}_K^{-1} \in \mathbb{M}_{k,k'}$ s.t. $\Pi_{Q_K}(m_K) := q_K$ and $\Pi_{\Lambda_{\partial K}}(m_K) := \lambda_{\partial K}$ is solvable for all $q_K \in Q_K$ and all $\lambda_{\partial K} \in \Lambda_{\partial K}$. Then m_h^{nc} is defined such that for all $K \in \mathcal{T}_h$,

$$ m_{h|K}^{\mathrm{nc}} \in \mathbb{M}_{k,k'}, \quad \Pi_{Q_K}(m_{h|K}^{\mathrm{nc}}) := p_{h|K}, \quad \Pi_{\Lambda_{\partial K}}(m_{h|K}^{\mathrm{nc}}) := \lambda_{h|\partial K}. \tag{52.12} $$

There are in general various admissible choices for the polynomial space $\mathbb{M}_{k,k'}$. For instance, if one works with simplicial Raviart–Thomas elements of degree k, one can set $\mathbb{M}_{k,k'} := \mathbb{P}_{k',d}$ with $k' := k + 2$ since we have $\dim(\mathbb{P}_{k',d}) = \binom{k'+d}{d} > \binom{k+d}{d} + (d+1)\binom{k+d-1}{d-1} = \dim(\mathbb{P}_{k,d}) + (d+1)\dim(\mathbb{P}_{k,d-1}) = \dim(Q_K) + \dim(\Lambda_{\partial K})$. Alternatively, one can take a smaller space $\mathbb{M}_{k,k'}$ so that $\dim(\mathbb{M}_{k,k'}) = \dim(Q_K) + \dim(\Lambda_{\partial K})$. For the lowest-order Raviart–Thomas elements in $\mathbb{R}^2$, one can set $\mathbb{M}_{0,2} := \mathrm{span}\{1, \widehat{\lambda}_0 \widehat{\lambda}_1, \widehat{\lambda}_1 \widehat{\lambda}_2, \widehat{\lambda}_2 \widehat{\lambda}_0\}$, where $\{\widehat{\lambda}_0, \widehat{\lambda}_1, \widehat{\lambda}_2\}$ are the barycentric coordinates on the reference element, i.e., $k = 0$ and $k' = 2$. A similar choice can be made in dimension 3 with $k' = 3$. Further examples can be found in Arnold and Brezzi [16], Vohralík [382] for $k := 0$ and more generally in Arbogast and Chen [13] for all $k \geq 0$.

Proposition 52.5 (Post-processed potential). *Let m_h^{nc} satisfy (52.12). The following holds true for all $\boldsymbol{\tau} \in \mathbb{RT}_{k,d}$ and all $K \in \mathcal{T}_h$:*

$$(\mathbb{d}^{-1}\boldsymbol{\sigma}_h - \boldsymbol{f} + \nabla m_h^{\mathrm{nc}}, \boldsymbol{\tau})_{\boldsymbol{L}^2(K)} = 0. \tag{52.13}$$

Moreover, recalling the convention that $[\![v]\!]_F = v_{|F}$ if $F \in \mathcal{F}_h^{\partial}$, the following holds true for all $\zeta \in \mathbb{P}_{k,d-1}$ and all $F \in \mathcal{F}_h$:

$$\int_F [\![m_h^{\mathrm{nc}}]\!]_F (\zeta \circ \boldsymbol{T}_F^{-1}) \, \mathrm{d}s = 0. \tag{52.14}$$

Proof. Let us take a test function $\boldsymbol{\tau}_K$ supported in a single mesh cell $K \in \mathcal{T}_h$ in (52.4a). Integrating by parts in K and using (52.12) together with $\nabla{\cdot}\boldsymbol{\tau}_K \in Q_K$ and $\boldsymbol{\tau}_{K|\partial K}{\cdot}\boldsymbol{n}_K \in \Lambda_{\partial K}$, we infer that

$$\begin{aligned}
&(\mathbb{d}^{-1}\boldsymbol{\sigma}_h - \boldsymbol{f} + \nabla m_h^{\mathrm{nc}}, \boldsymbol{\tau}_K)_{\boldsymbol{L}^2(K)} \\
&= (\mathbb{d}^{-1}\boldsymbol{\sigma}_h - \boldsymbol{f}, \boldsymbol{\tau}_K)_{\boldsymbol{L}^2(K)} - (m_h^{\mathrm{nc}}, \nabla{\cdot}\boldsymbol{\tau}_K)_{L^2(K)} + (m_h^{\mathrm{nc}}, \boldsymbol{\tau}_K{\cdot}\boldsymbol{n}_K)_{L^2(\partial K)} \\
&= (\mathbb{d}^{-1}\boldsymbol{\sigma}_h - \boldsymbol{f}, \boldsymbol{\tau}_K)_{\boldsymbol{L}^2(K)} - (p_h, \nabla{\cdot}\boldsymbol{\tau}_K)_{L^2(K)} + (\lambda_h, \boldsymbol{\tau}_K{\cdot}\boldsymbol{n}_K)_{L^2(\partial K)} = 0.
\end{aligned}$$

This proves (52.13). Observing that λ_h is single-valued on the interfaces and vanishes on the boundary faces, the rightmost equation in (52.12) implies that $0 = \int_F [\![m_h^{\mathrm{nc}} - \lambda_{h|F}]\!]_F (\zeta \circ \boldsymbol{T}_F^{-1}) \, \mathrm{d}s = \int_F [\![m_h^{\mathrm{nc}}]\!]_F (\zeta \circ \boldsymbol{T}_F^{-1}) \, \mathrm{d}s$. This proves (52.14). $\qquad\square$

The post-processed potential m_h^{nc} can be used in the a priori and a posteriori analysis of mixed finite element methods; see Vohralík [383].

52.2 Flux recovery for H^1-conforming elements

In this section, we return to the primal formulation of the model elliptic problem considered in Chapter 32, i.e., $-\nabla{\cdot}(\mathbb{d}\nabla p) = g$ in D with $p_{|\partial D} = 0$ (we write p instead of u and g instead of f for consistency with the present notation). As before, we assume that the eigenvalues of $\mathbb{d}$ are in the interval $[\lambda_\flat, \lambda_\sharp]$ a.e. in D with $\lambda_\flat > 0$. The exact flux is $\boldsymbol{\sigma} := -\mathbb{d}\nabla p \in \boldsymbol{L}^2(D)$ (this corresponds to setting $\boldsymbol{f} := \boldsymbol{0}$ in Darcy's law). A crucial observation is that

$$\boldsymbol{\sigma} \in \boldsymbol{H}(\mathrm{div}; D), \qquad \nabla{\cdot}\boldsymbol{\sigma} = g. \tag{52.15}$$

Let $p_h \in P_{k,0}^{\mathrm{g}}(\mathcal{T}_h)$ be the discrete solution obtained from the $H_0^1(D)$-conforming finite element approximation of order $k \geq 1$. Recall that p_h satisfies $(\mathbb{d}\nabla p_h, \nabla w_h)_{\boldsymbol{L}^2(D)} = (g, w_h)_{L^2(D)}$ for all $w_h \in P_{k,0}^{\mathrm{g}}(\mathcal{T}_h)$; see (32.5). The approximate flux $\boldsymbol{\sigma}_h := -\mathbb{d}\nabla p_h \in \boldsymbol{L}^2(D)$ delivers an accurate approximation of the exact flux $\boldsymbol{\sigma}$. We indeed have $\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)} \leq \lambda_\sharp |p - p_h|_{H^1(D)}$, and we have seen in §32.2 that the error $|p - p_h|_{H^1(D)}$ converges to zero as $h \to 0$ with the rate $\mathcal{O}(h^r)$ provided $p \in H^{1+r}(D)$ and $r \in (0, k]$. But for this approximation we do not have $\boldsymbol{\sigma}_h \in \boldsymbol{H}(\mathrm{div}; D)$. Since it is desirable for some applications to have a discrete flux in $\boldsymbol{H}(\mathrm{div}; D)$, we now present a post-processing technique to build a post-processed flux $\boldsymbol{\sigma}_h^* \in \boldsymbol{H}(\mathrm{div}; D)$ s.t.

$$(\nabla{\cdot}\boldsymbol{\sigma}_h^*, q)_{L^2(K)} = (g, q)_{L^2(K)}, \qquad \forall q \in \mathbb{P}_{l,d}, \tag{52.16}$$

for all $K \in \mathcal{T}_h$, $l \in \{k-1, k\}$, and such that $\|\boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^*\|_{\boldsymbol{L}^2(K)}$ is bounded by the local H^1-seminorm of the error $p - p_h$ (up to some data oscillation). Notice that the post-processing is local and does not require any additional global solve. The flux $\boldsymbol{\sigma}_h^*$ can be used as the transport velocity field in underground flow applications (see Bastian and Rivière [47]). The post-processed flux can also be used to evaluate sharp a posteriori error estimates (see §52.2.3).

52.2.1 Local flux equilibration

We locally construct a flux σ_h^* that satisfies (52.16) in this section, and we show that $\|\sigma_h - \sigma_h^*\|_{L^2(K)}$ behaves as expected in §52.2.2. Let $z \in \mathcal{V}_h$ be a mesh vertex, let ψ_z be the corresponding $\mathbb{P}_1$ Lagrange basis function (also called hat or Courant basis function). The support of ψ_z is denoted by D_z and consists of all the mesh cells in the set $\mathcal{T}_z$ having z as vertex. This set is often called *finite element star*.

   Let $l \geq 0$ and $\boldsymbol{P}_{l,*}^{\mathrm{d}}(\mathcal{T}_z)$ be the (local) Raviart–Thomas finite element space of order l in the star D_z with the additional requirement that every function $\boldsymbol{\tau}_z \in \boldsymbol{P}_{l,*}^{\mathrm{d}}(\mathcal{T}_z)$ is such that $(\boldsymbol{\tau}_z \cdot \boldsymbol{n}_{D_z})_{|\partial D_z} = 0$ if $z \in \mathcal{V}_h^\circ$ or $(\boldsymbol{\tau}_z \cdot \boldsymbol{n}_{D_z})_{|\partial D_z \setminus \partial D} = 0$ if $z \in \mathcal{V}_h^\partial$, where $\boldsymbol{n}_{D_z}$ is the outward unit normal to D_z. Let $P_{l,*}^{\mathrm{b}}(\mathcal{T}_z)$ be the (local) broken space of scalar-valued finite elements of order l in the star D_z with the constraint that every function $q_z \in P_{l,*}^{\mathrm{b}}(\mathcal{T}_z)$ satisfies $(q_z, 1)_{L^2(D_z)} = 0$ if $z \in \mathcal{V}_h^\circ$. Let $\mathcal{I}_l^{\mathrm{d,b}}$ be the interpolation operator in the broken Raviart–Thomas space $\boldsymbol{P}_{l,*}^{\mathrm{d,b}}(\mathcal{T}_z)$ (without boundary conditions) and let $\mathcal{I}_l^{\mathrm{b}}$ be the L^2-orthogonal projection onto the broken space $P_l^{\mathrm{b}}(\mathcal{T}_z)$. Let us set $\boldsymbol{f}_z := -\psi_z \mathrm{d}\nabla p_h$, $g_z := \psi_z g - (\mathrm{d}\nabla p_h) \cdot \nabla \psi_z$, and consider the constrained minimization problem

$$\sigma_z^* := \underset{\boldsymbol{\tau}_z \in \boldsymbol{P}_{l,*}^{\mathrm{d}}(\mathcal{T}_z), \nabla \cdot \boldsymbol{\tau}_z = \mathcal{I}_l^{\mathrm{b}}(g_z)}{\arg \min} \|\boldsymbol{\tau}_z - \mathcal{I}_l^{\mathrm{d,b}}(\boldsymbol{f}_z)\|_{L^2(D_z)}. \tag{52.17}$$

Following the discussion in §51.2, the problem (52.17) can be efficiently solved by considering the following dual mixed formulation:

$$\begin{cases} \text{Find } \sigma_z^* \in \boldsymbol{P}_{l,*}^{\mathrm{d}}(\mathcal{T}_z) \text{ and } r_z^* \in P_{l,*}^{\mathrm{b}}(\mathcal{T}_z) \text{ such that} \\ (\sigma_z^*, \boldsymbol{\tau}_z)_{L^2(D_z)} - (\nabla \cdot \boldsymbol{\tau}_z, r_z^*)_{L^2(D_z)} = (\mathcal{I}_l^{\mathrm{d,b}}(\boldsymbol{f}_z), \boldsymbol{\tau}_z)_{L^2(D_z)}, & \forall \boldsymbol{\tau}_z \in \boldsymbol{P}_{l,*}^{\mathrm{d}}(\mathcal{T}_z), \\ (\nabla \cdot \sigma_z^*, q_z)_{L^2(D_z)} = (g_z, q_z)_{L^2(D_z)}, & \forall q_z \in P_{l,*}^{\mathrm{b}}(\mathcal{T}_z). \end{cases}$$

We obtain a pure Neumann problem if $z \in \mathcal{V}_h^\circ$ and a mixed Dirichlet–Neumann problem if $z \in \mathcal{V}_h^\partial$. The pure Neumann problem is well-posed owing to the compatibility condition

$$(\mathcal{I}_l^{\mathrm{b}}(g_z), 1)_{L^2(D_z)} = (g_z, 1)_{L^2(D_z)} = (g, \psi_z)_{L^2(D_z)} - (\mathrm{d}\nabla p_h, \nabla \psi_z)_{\boldsymbol{L}^2(D_z)} = 0,$$

which is the Galerkin orthogonality property on the hat basis functions.

Theorem 52.6 (Equilibrated flux). *Let $l \geq 0$. Let σ_z^* be defined by (52.17) for all $z \in \mathcal{V}_h$, and let $\widetilde{\sigma}_z^*$ be the zero extension of σ_z^* outside D_z. Set $\sigma_h^* := \sum_{z \in \mathcal{V}_h} \widetilde{\sigma}_z^*$. Then $\sigma_h^* \in \boldsymbol{H}(\mathrm{div}; D)$, and the divergence of σ_h^* satisfies (52.16).*

Proof. For every $z \in \mathcal{V}_h$, the normal component of σ_z^* is continuous across all the interfaces in the mesh $\mathcal{T}_z$ since $\sigma_z^* \in \boldsymbol{P}_{l,*}^{\mathrm{d}}(\mathcal{T}_z)$. Recall also that by definition the normal component of σ_z^* is zero on all the boundary faces F of $\mathcal{T}_z$ that are not in $\mathcal{F}_h^\partial$ (i.e., those in $\mathcal{F}_h^\circ$). Invoking Theorem 18.10 shows that $\widetilde{\sigma}_z^* \in \boldsymbol{H}(\mathrm{div}; D)$. This argument implies that $\sigma_h^* \in \boldsymbol{H}(\mathrm{div}; D)$. Furthermore, after observing that $\sigma_{h|K}^* = \sum_{z \in \mathcal{V}_K} \sigma_{z|K}^*$, where $\mathcal{V}_K$ is the collection of all the vertices of K, we have

$$(\nabla \cdot \sigma_h^*, q)_{L^2(K)} = \sum_{z \in \mathcal{V}_K} (\nabla \cdot \sigma_z^*, q)_{L^2(K)} = \sum_{z \in \mathcal{V}_K} (g_z, q)_{L^2(K)} = (g, q)_{L^2(K)},$$

for all $q \in \mathbb{P}_{l,d}$, since the local partition of unity $\sum_{z \in \mathcal{V}_K} \psi_{z|K} = 1$ implies that $\sum_{z \in \mathcal{V}_K} g_z = g$. $\qquad \square$

Remark 52.7 (Local vs. global). The post-processed flux σ_h^* is a member of the Raviart–Thomas finite element space $\boldsymbol{P}_l^{\mathrm{d}}(\mathcal{T}_h)$ of order l (see (51.20a)). Yet, the construction of σ_h^* is local, so the technique discussed above is an inexpensive alternative to the global equilibration procedure defined by $\sigma_h^{*\mathrm{glob}} := \arg \min_{\boldsymbol{\tau}_h \in \boldsymbol{P}_l^{\mathrm{d}}(\mathcal{T}_h), \nabla \cdot \boldsymbol{\tau}_h = \mathcal{I}_h^{\mathrm{b}}(g)} \|\boldsymbol{\tau}_h - \sigma_h\|_{\boldsymbol{L}^2(D)}$ which requires to solve a global Darcy problem using mixed elements. $\qquad \square$

Remark 52.8 (Literature). That H^1-conforming finite elements can be post-processed to give quantities that enjoy local conservation properties has been highlighted in Hughes et al. [250]. The present construction of $\boldsymbol{\sigma}_h^*$ is inspired by the seminal ideas in Braess and Schöberl [76], Braess et al. [77]; see also Ern and Vohralík [190, 191, 192] for further results. A somewhat simpler construction in the lowest-order case can be found in Destuynder and Métivet [163]; see also Larson and Niklasson [275]. An alternative approach consists of performing the flux equilibration on a dual (barycentric) mesh; see Luce and Wohlmuth [290], Ern and Vohralík [189], Vohralík [384] for the lowest-order case and see Hannukainen et al. [238] for a higher-order extension. □

52.2.2 L^2-norm estimate

In this section, we assume for simplicity that $\mathbb{d}$ is piecewise constant. We also assume that there exists a polynomial space $\mathbb{M}_{k,k'}$ as considered in §52.1.2.

Lemma 52.9 (L^2-estimate). *Let $\boldsymbol{\sigma}_h^*$ be defined as in Theorem 52.6 with $l \geq k-1$. Recall that $\boldsymbol{\sigma}_h := -\mathbb{d}\nabla p_h$. There is a constant c such that for all $K \in \mathcal{T}_h$ and all $h \in \mathcal{H}$,*

$$\|\boldsymbol{\sigma}_h^* - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(K)} \leq c \Bigg(\sum_{K' \in \check{\mathcal{T}}_K} h_{K'} \|g + \nabla{\cdot}(\mathbb{d}\nabla p_h)\|_{L^2(K')}$$

$$+ \sum_{F' \in \check{\mathcal{F}}_K^\circ} h_{F'}^{\frac{1}{2}} \|[\![\mathbb{d}\nabla p_h]\!]{\cdot}\boldsymbol{n}_F\|_{L^2(F')} \Bigg), \tag{52.18}$$

where $\check{\mathcal{T}}_K$ and $\check{\mathcal{F}}_K^\circ$ are the collections of those cells and interfaces that share at least one vertex with K, respectively. Moreover, defining the oscillation term $\omega_{K'}^{\mathrm{v}} := h_{K'}\|g - \mathcal{I}_l^{\mathrm{b}}(g)\|_{L^2(K')}$, we have

$$\|\boldsymbol{\sigma}_h^* - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(K)} \leq c \sum_{K' \in \check{\mathcal{T}}_K} (|p - p_h|_{H^1(K')} + \omega_{K'}^{\mathrm{v}}). \tag{52.19}$$

Proof. Let $K \in \mathcal{T}_h$ and let us estimate $\|\boldsymbol{\sigma}_h^* + \mathbb{d}\nabla p_h\|_{\boldsymbol{L}^2(K)}$. Since $\mathbb{d}$ is piecewise constant and $l \geq k-1$, we infer that $(\mathbb{d}\nabla p_h)_{|K}$ is in $\mathbb{RT}_{l,d}$. Hence, $\mathcal{I}_l^{\mathrm{d,b}}(\mathbb{d}\nabla p_h) = \mathbb{d}\nabla p_h$. Recalling that $\boldsymbol{f}_{\boldsymbol{z}} := -\psi_{\boldsymbol{z}}\mathbb{d}\nabla p_h$, using the local partition of unity and the linearity of $\mathcal{I}_l^{\mathrm{d,b}}$, and since $\boldsymbol{\sigma}_{h|K}^* = \sum_{\boldsymbol{z} \in \mathcal{V}_K} \boldsymbol{\sigma}_{\boldsymbol{z}}^*$, we infer that

$$(\boldsymbol{\sigma}_h^* + \mathbb{d}\nabla p_h)_{|K} = \sum_{\boldsymbol{z} \in \mathcal{V}_K} \boldsymbol{\sigma}_{\boldsymbol{z}}^* + \mathcal{I}_l^{\mathrm{d,b}}\Bigg(\sum_{\boldsymbol{z} \in \mathcal{V}_K} \psi_{\boldsymbol{z}}\mathbb{d}\nabla p_h \Bigg) = \sum_{\boldsymbol{z} \in \mathcal{V}_K} (\boldsymbol{\sigma}_{\boldsymbol{z}}^* - \mathcal{I}_l^{\mathrm{d,b}}(\boldsymbol{f}_{\boldsymbol{z}})).$$

Invoking the triangle inequality leads to

$$\|\boldsymbol{\sigma}_h^* + \mathbb{d}\nabla p_h\|_{\boldsymbol{L}^2(K)} \leq \sum_{\boldsymbol{z} \in \mathcal{V}_K} \|\boldsymbol{\sigma}_{\boldsymbol{z}}^* - \mathcal{I}_l^{\mathrm{d,b}}(\boldsymbol{f}_{\boldsymbol{z}})\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})},$$

which shows that we are left with estimating $\|\boldsymbol{\sigma}_{\boldsymbol{z}}^* - \mathcal{I}_l^{\mathrm{d,b}}(\boldsymbol{f}_{\boldsymbol{z}})\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})}$ for all $\boldsymbol{z} \in \mathcal{V}_K$. Owing to Exercise 52.5, we infer that there is $c > 0$ such that for all $\boldsymbol{z} \in \mathcal{V}_h$ and all $h \in \mathcal{H}$,

$$c\|\boldsymbol{\sigma}_{\boldsymbol{z}}^* - \mathcal{I}_l^{\mathrm{d,b}}(\boldsymbol{f}_{\boldsymbol{z}})\|_{\boldsymbol{L}^2(D_{\boldsymbol{z}})} \leq \sum_{K' \in \mathcal{T}_{\boldsymbol{z}}} h_{K'}\|\delta_{\boldsymbol{z}}^{\mathrm{v}}\|_{L^2(K')} + \sum_{F' \in \mathcal{F}_{\boldsymbol{z}}^\circ} h_{F'}^{\frac{1}{2}}\|\delta_{\boldsymbol{z}}^{\mathrm{s}}\|_{L^2(F')}, \tag{52.20}$$

with $\delta_{\boldsymbol{z}}^{\mathrm{v}} := \nabla{\cdot}\boldsymbol{f}_{\boldsymbol{z}} - g_{\boldsymbol{z}}$ and $\delta_{\boldsymbol{z}}^{\mathrm{s}} := [\![\boldsymbol{f}_{\boldsymbol{z}}]\!]{\cdot}\boldsymbol{n}_{F'}$. This leads to the bound (52.18) since $\delta_{\boldsymbol{z}}^{\mathrm{v}} = -\psi_{\boldsymbol{z}}(g + \nabla{\cdot}(\mathbb{d}\nabla p_h))$, $\delta_{\boldsymbol{z}}^{\mathrm{s}} = \psi_{\boldsymbol{z}}[\![\mathbb{d}\nabla p_h]\!]{\cdot}\boldsymbol{n}_F$, and $\|\psi_{\boldsymbol{z}}\|_{L^\infty(D_{\boldsymbol{z}})} = 1$. We refer the reader to Exercise 52.4 for the proof of (52.19). □

Remark 52.10 (Choice of l). Lemma 52.9 shows that $\boldsymbol{\sigma}_h^* \in \boldsymbol{P}_l^{\mathrm{d}}(\mathcal{T}_h)$ approximates the exact flux in $\boldsymbol{L}^2$ with the rate $\mathcal{O}(h^k)$. Hence, the choice $l := k - 1$ is optimal from this viewpoint. Choosing $l := k$ leads to a (slightly) more precise recovered flux since (52.16) is satisfied up to $l = k$. Moreover, if g is smooth, i.e., $g \in H^l(\mathcal{T}_h)$, the data oscillation term in (52.19) converges like $\mathcal{O}(h^{l+2})$, which for $l = k - 1$ and for $l = k$, respectively, is one order or two orders faster than the approximation error $p - p_h$. $\square$

Remark 52.11 (Literature). The proof of (52.18) essentially follows Ern and Vohralík [190] (up to minor variations). A different proof of (52.19) is devised in Braess et al. [77] (see also Ern and Vohralík [191]) in dimension two with uniform diffusion, allowing one to prove that the constant c is independent of the polynomial degree k. The proof in dimension three can be found in Ern and Vohralík [192]. $\square$

52.2.3 Application to a posteriori error analysis

An important application of local flux recovery is the a posteriori error analysis of H^1-conforming finite elements. Recall from Chapter 34 that a posteriori error estimates provide two-sided, fully computable bounds on the approximation error $(p - p_h)$. We continue to denote the primal variable by p and the source term by g instead of using u and f as in Chapter 34.

Lemma 52.12 (Two-sided bound). *Let $p \in H_0^1(D)$ solve $-\nabla\cdot(\mathrm{d}\nabla p) = g$, and let $p_h \in H_0^1(D)$ be its H^1-conforming approximation. Set $\boldsymbol{\sigma}_h := -\mathrm{d}\nabla p_h$. We have the following upper and lower bounds on $\|\nabla(p - p_h)\|_{\boldsymbol{L}^2(D)}$:*

$$\lambda_\flat \|\nabla(p - p_h)\|_{\boldsymbol{L}^2(D)} \leq \inf_{\boldsymbol{\sigma}^* \in \boldsymbol{H}(\mathrm{div};D),\, \nabla\cdot\boldsymbol{\sigma}^* = g} \|\boldsymbol{\sigma}^* - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)}$$

$$\leq \lambda_\sharp \|\nabla(p - p_h)\|_{\boldsymbol{L}^2(D)}. \tag{52.21}$$

Proof. Recall that Lemma 34.3 gives $\lambda_\flat \|\nabla(p - p_h)\|_{\boldsymbol{L}^2(D)} \leq \|\rho(p_h)\|_{H^{-1}(D)}$ with the residual defined s.t. $\langle \rho(p_h), \varphi \rangle := (g, \varphi)_{L^2(D)} - (\mathrm{d}\nabla p_h, \nabla\varphi)_{\boldsymbol{L}^2(D)}$ and $\|\varphi\|_{H_0^1(D)} := |\varphi|_{H^1(D)}$ for all $\varphi \in H_0^1(D)$. For all $\boldsymbol{\sigma}^* \in \boldsymbol{H}(\mathrm{div};D)$ such that $\nabla\cdot\boldsymbol{\sigma}^* = g$, we then have

$$\langle \rho(p_h), \varphi \rangle = (g, \varphi)_{L^2(D)} + (\boldsymbol{\sigma}_h, \nabla\varphi)_{\boldsymbol{L}^2(D)}$$
$$= (\nabla\cdot\boldsymbol{\sigma}^*, \varphi)_{L^2(D)} + (\boldsymbol{\sigma}_h, \nabla\varphi)_{\boldsymbol{L}^2(D)} = (\boldsymbol{\sigma}_h - \boldsymbol{\sigma}^*, \nabla\varphi)_{\boldsymbol{L}^2(D)}.$$

Hence, $\|\rho(p_h)\|_{H^{-1}(D)} \leq \|\boldsymbol{\sigma}^* - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(D)}$, and the first bound in (52.21) follows since $\boldsymbol{\sigma}^*$ is arbitrary. For the second bound, it suffices to pick $\boldsymbol{\sigma}^* := -\mathrm{d}\nabla p$, which is in $\boldsymbol{H}(\mathrm{div};D)$ and its weak divergence is equal to g. $\square$

Solving the infinite-dimensional constrained minimization problem in (52.21) is unfeasible. Let us consider instead the milder constraint

$$(\nabla\cdot\boldsymbol{\sigma}^*, 1)_{L^2(K)} = (g, 1)_{L^2(K)}, \qquad \forall K \in \mathcal{T}_h. \tag{52.22}$$

Lemma 52.13 (Flux-equilibrated upper bound). *The following holds true for all $\boldsymbol{\sigma}^* \in \boldsymbol{H}(\mathrm{div};D)$ satisfying (52.22):*

$$\lambda_\flat \|\nabla(p - p_h)\|_{\boldsymbol{L}^2(D)} \leq \left(\sum_{K \in \mathcal{T}_h} \left(\|\boldsymbol{\sigma}^* - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(K)} + \eta_{\mathrm{osc},K} \right)^2 \right)^{\frac{1}{2}}, \tag{52.23}$$

with the data oscillation term $\eta_{\mathrm{osc},K} := \frac{1}{\pi} h_K \|g - \nabla\cdot\boldsymbol{\sigma}^*\|_{L^2(K)}$.

Proof. Let $\varphi \in H_0^1(D)$. Owing to (52.22), we infer that $(g - \nabla \cdot \boldsymbol{\sigma}^*, \varphi)_{L^2(D)} = \sum_{K \in \mathcal{T}_h} (g - \nabla \cdot \boldsymbol{\sigma}^*, \varphi - c_K)_{L^2(D)}$, and we choose the constant c_K equal to the mean value of φ in K. Recalling that the mesh cells are convex sets, the Poincaré–Steklov inequality (12.13) gives $\|\varphi - c_K\|_{L^2(K)} \le \frac{1}{\pi} h_K \|\nabla \varphi\|_{\boldsymbol{L}^2(K)}$. We now adapt the proof of Lemma 52.12 to infer that

$$\langle \rho(p_h), \varphi \rangle \le \sum_{K \in \mathcal{T}_h} (\|\boldsymbol{\sigma}^* - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(K)} + \eta_{\mathrm{osc}, K}) \|\nabla \varphi\|_{\boldsymbol{L}^2(K)}.$$

We conclude by using the Cauchy–Schwarz inequality. $\square$

An ideal candidate for $\boldsymbol{\sigma}^*$ is the locally post-processed flux $\boldsymbol{\sigma}_h^*$ introduced in Theorem 52.6 and satisfying (52.16) for some $l \ge 0$, which is a (possibly) higher-order version of (52.22). Moreover, Lemma 52.9 shows that $\|\boldsymbol{\sigma}_h^* - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(K)}$ is also a local lower bound on the error, up to the oscillation term $\eta_{\mathrm{osc}, K}$. Lemma 52.13 is actually valid for all $p_h \in H_0^1(D)$. That p_h solves a discrete problem is exploited in the actual construction of $\boldsymbol{\sigma}_h^*$ by means of the Galerkin orthogonality property on the hat basis functions.

Remark 52.14 (Comparison). The constants in Lemma 52.13 are simpler to estimate than those in Corollary 34.14. Indeed, the leading term in (52.23) has constant 1, and the data oscillation term only depends on the constant from the Poincaré–Steklov inequality in mesh cells (as opposed to the vertex-based stars which have a more complex geometry). However, equilibrated-flux a posteriori error estimates depend on data oscillations both for the lower and the upper bounds, i.e., not just for the lower bound as in §34.3. $\square$

Remark 52.15 (Literature). Equilibrated-flux a posteriori error estimation for H^1-conforming finite elements has a long history. Invoking $\boldsymbol{H}(\mathrm{div})$-fluxes leads to guaranteed upper bounds on the error, as shown by Prager and Synge [327], Hlaváček et al. [245] (see also Exercise 52.6). Building the flux by means of a local equilibration procedure on finite element stars leads in turn to local efficiency, i.e., to local lower bounds on the error; see Ladevèze and Leguillon [273], Ainsworth and Oden [7], Parés et al. [322]. Unfortunately, the estimators proposed in these references are not computable since they require solving an infinite-dimensional problem locally. Inexpensive local flux equilibration where local finite-dimensional problems are solved on finite element stars are devised in Destuynder and Métivet [163], Braess and Schöberl [76]. The idea of working on stars for a posteriori error analysis can be traced back to Babuška and Miller [36], where infinite-dimensional Dirichlet problems are posed on stars. Finite-dimensional Dirichlet problems inspired by Carstensen and Funken [107] are considered in Morin et al. [306]. $\square$

Exercises

Exercise 52.1 (Hybridization). Consider the discrete problem (52.4). (i) Let $\widetilde{Q}_h := Q_h \times \Lambda_h$ and $\widetilde{B}_h : \boldsymbol{V}_h^{\mathrm{hy}} \to \widetilde{Q}_h'$ s.t. $\langle \widetilde{B}_h(\boldsymbol{\tau}_h), (q_h, \mu_h) \rangle_{\widetilde{Q}_h', \widetilde{Q}_h} := b_h(\boldsymbol{\tau}_h, q_h) + c_h(\boldsymbol{\tau}_h, \mu_h)$ for all $\boldsymbol{\tau}_h \in \boldsymbol{V}_h^{\mathrm{hy}}$ and $(q_h, \mu_h) \in \widetilde{Q}_h$. Prove that $\widetilde{B}_h^*$ is injective. (*Hint*: integrate by parts and use the degrees of freedom of the $\mathbb{RT}_{k,d}$ element.) (ii) Prove that (52.4) admits a unique solution.

Exercise 52.2 (Crouzeix–Raviart). Assume that $\mathrm{d}_{|K}$ and $g_{|K}$ are constant over each mesh cell $K \in \mathcal{T}_h$. Let ∇_h denote the broken gradient (see Definition 36.3). Let $P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$ be the nonconforming Crouzeix–Raviart finite element space with homogeneous Dirichlet conditions (see

(36.8)) and let $p_h^{\mathrm{CR}} \in P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$ solve $\int_D (\mathbb{d}\nabla_h p_h^{\mathrm{CR}}) \cdot \nabla_h q_h^{\mathrm{CR}} \, \mathrm{d}x = \int_D g q_h^{\mathrm{CR}} \, \mathrm{d}x$ for all $q_h^{\mathrm{CR}} \in P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$. Let $\boldsymbol{x}_K$ be the barycenter of K for all $K \in \mathcal{T}_h$. Define

$$\boldsymbol{\sigma}_{h|K} := -(\mathbb{d}\nabla p_h^{\mathrm{CR}})_{|K} + d^{-1} g_{|K}(\boldsymbol{x} - \boldsymbol{x}_K)_{|K},$$

$$p_{h|K} := p_h^{\mathrm{CR}}(\boldsymbol{x}_K) + d^{-2}|K|^{-1} g_{|K}(\mathbb{d}^{-1}(\boldsymbol{x} - \boldsymbol{x}_K), \boldsymbol{x} - \boldsymbol{x}_K)_{L^2(K)}.$$

(i) Prove that $\boldsymbol{\sigma}_h \in \boldsymbol{P}_0^{\mathrm{d}}(\mathcal{T}_h)$. (*Hint*: compute $\int_F [\![\boldsymbol{\sigma}_h]\!] \cdot \boldsymbol{n}_F \varphi_F^{\mathrm{CR}} \, \mathrm{d}s$ with φ_F^{CR} the Crouzeix–Raviart basis function attached to F.) (ii) Prove that $\int_D (q_h^{\mathrm{CR}} \nabla \cdot \boldsymbol{\tau}_h + \nabla_h q_h^{\mathrm{CR}} \cdot \boldsymbol{\tau}_h) \, \mathrm{d}x = 0$ for all $q_h^{\mathrm{CR}} \in P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$ and all $\boldsymbol{\tau}_h \in \boldsymbol{P}_0^{\mathrm{d}}(\mathcal{T}_h)$. (iii) Prove that the pair $(\boldsymbol{\sigma}_h, p_h)$ solves (51.21) for $k := 0$ and $\boldsymbol{f} := \boldsymbol{0}$. (*Hint*: any function $\boldsymbol{\tau}_h \in \boldsymbol{P}_0^{\mathrm{d}}(\mathcal{T}_h)$ is such that $\boldsymbol{\tau}_{h|K} = \boldsymbol{\tau}_K + d^{-1}(\nabla \cdot \boldsymbol{\tau}_h)_{|K}(\boldsymbol{x} - \boldsymbol{x}_K)$, where $\boldsymbol{\tau}_K$ is the mean value of $\boldsymbol{\tau}_h$ on K.)

Exercise 52.3 (Post-processed potential). Let $k \geq 0$. Consider the simplicial Raviart–Thomas element $\mathbb{RT}_{k,d}$. Assume that it is possible to find a polynomial space $\mathbb{M}_{k,k'}$ so that for all $m \in \mathbb{M}_{k,k'}$, $\Pi_{Q_K}(m) = \Pi_{\Lambda_{\partial K}}(m) = 0$ implies that $m = 0$ for all $K \in \mathcal{T}_h$. Prove that $(\nabla m, \boldsymbol{\tau})_{\boldsymbol{L}^2(K)} = 0$ for all $\boldsymbol{\tau} \in \mathbb{RT}_{k,d}$ implies that $m = 0$. (*Hint*: integrate by parts and use the degrees of freedom in $\mathbb{RT}_{k,d}$.) Let now m_h^{nc} be the post-processed potential from the dual mixed formulation (52.2). Show that $\|\nabla m_h^{\mathrm{nc}}\|_{\boldsymbol{L}^2(K)} \leq c \|\mathbb{d}^{-1}\boldsymbol{\sigma}_h - \boldsymbol{f}\|_{\boldsymbol{L}^2(K)}$ for all $K \in \mathcal{T}_h$. (*Hint*: use norm equivalence on the reference element, then (52.13); see also Vohralík [383, Lem. 5.4].)

Exercise 52.4 (Bound (52.19)). Prove (52.19). (*Hint*: use Theorem 34.19.)

Exercise 52.5 (Inverse inequality). Prove (52.20). (*Hint*: consider the dual mixed formulation of (52.17) and introduce the post-processed variable $m_{\boldsymbol{z}}^{\mathrm{nc}}$, use (52.13), accept as a fact that $\|m_{\boldsymbol{z}}^{\mathrm{nc}}\|_{L^2(D_{\boldsymbol{z}})} \leq c h_{D_{\boldsymbol{z}}} \|\nabla_h m_{\boldsymbol{z}}^{\mathrm{nc}}\|_{L^2(D_{\boldsymbol{z}})}$, and bound traces of $m_{\boldsymbol{z}}^{\mathrm{nc}}$ using Lemma 12.15.)

Exercise 52.6 (Prager–Synge equality). Let $u \in H_0^1(D)$ be such that $-\Delta u = f$ in $L^2(D)$. Let $u_h \in H_0^1(D)$, and let $\boldsymbol{\sigma}^* \in \boldsymbol{H}(\mathrm{div}; D)$ be such that $\nabla \cdot \boldsymbol{\sigma}^* = f$. Prove that $\|\nabla(u - u_h)\|_{\boldsymbol{L}^2(D)}^2 + \|\nabla u + \boldsymbol{\sigma}^*\|_{\boldsymbol{L}^2(D)}^2 = \|\nabla u_h + \boldsymbol{\sigma}^*\|_{\boldsymbol{L}^2(D)}^2$. (*Hint*: compute $(\nabla(u - u_h), \nabla u + \boldsymbol{\sigma}^*)_{\boldsymbol{L}^2(D)}$.)

Chapter 53

Stokes equations: Basic ideas

The Stokes equations constitute the basic linear model for incompressible fluid mechanics. We first derive a weak formulation of the Stokes equations and establish its well-posedness. The approximation is then realized by means of mixed finite elements, that is, we consider a pair of finite elements, where the first component of the pair is used to approximate the velocity and the second component is used to approximate the pressure. Following the ideas of Chapter 50, the finite element pair is said to be stable whenever the discrete velocity and the discrete pressure spaces satisfy an inf-sup condition. In this chapter, we list some classical unstable pairs. Examples of stable pairs are reviewed in the following two chapters.

53.1 Incompressible fluid mechanics

Let D be a Lipschitz domain in $\mathbb{R}^d$. We are interested in modeling the behavior of incompressible fluid flows in D in the time-independent Stokes regime, i.e., the inertial forces are assumed to be negligible. Given a vector-valued field $\boldsymbol{f} : D \to \mathbb{R}^d$ (the body force acting on the fluid) and a scalar-valued field $g : D \to \mathbb{R}$ (the mass production rate), the Stokes problem consists of seeking the velocity field $\boldsymbol{u} : D \to \mathbb{R}^d$ and the pressure field $p : D \to \mathbb{R}$ such that the following balance equations hold true:

$$-\nabla\!\cdot\!\mathbb{s}(\boldsymbol{u}) + \nabla p = \boldsymbol{f} \qquad \text{in } D, \tag{53.1a}$$

$$\nabla\!\cdot\!\boldsymbol{u} = g \qquad \text{in } D, \tag{53.1b}$$

$$\boldsymbol{u}_{|\partial D_{\mathrm{d}}} = \boldsymbol{a}_{\mathrm{d}}, \quad \mathbb{s}(\boldsymbol{u})_{|\partial D_{\mathrm{n}}}\boldsymbol{n} - p_{|\partial D_{\mathrm{n}}}\boldsymbol{n} = \boldsymbol{a}_{\mathrm{n}} \quad \text{on } \partial D. \tag{53.1c}$$

The equations (53.1a)-(53.1b) express, respectively, the balance of momentum and mass. The second-order tensor $\mathbb{s}(\boldsymbol{u})$ in (53.1a) is the *viscous stress tensor*. Notice that we abuse the notation in (53.1a) since we should write $\nabla\!\cdot\!(\mathbb{s}(\boldsymbol{u}))$ instead of $\nabla\!\cdot\!\mathbb{s}(\boldsymbol{u})$. As for linear elasticity (see §42.1), the principle of conservation of angular momentum implies that $\mathbb{s}(\boldsymbol{u})$ is symmetric and, assuming the fluid to be Newtonian, Galilean invariance implies that

$$\mathbb{s}(\boldsymbol{u}) = 2\mu\mathbb{e}(\boldsymbol{u}) + \lambda(\nabla\!\cdot\!\boldsymbol{u})\mathbb{I}, \qquad \mathbb{e}(\boldsymbol{u}) := \frac{1}{2}(\nabla\boldsymbol{u} + (\nabla\boldsymbol{u})^{\mathsf{T}}), \tag{53.2}$$

where $\mathbb{I}$ is the $d\times d$ identity tensor. The quantity $\mathbb{e}(\boldsymbol{u})$ is called (linearized) *strain rate tensor*, and the constants $\mu > 0$, $\lambda \geq 0$ are the dynamic and bulk viscosities, respectively. In (53.1c),

the subsets $\partial D_{\mathrm{d}}, \partial D_{\mathrm{n}}$ form a partition of the boundary ∂D, and we assume for simplicity that $|\partial D_{\mathrm{d}}| > 0$. The boundary data are the prescribed velocity $\boldsymbol{a}_{\mathrm{d}}$ on ∂D_{d} (Dirichlet condition) and the prescribed normal force $\boldsymbol{a}_{\mathrm{n}}$ on ∂D_{n} (Neumann condition).

Remark 53.1 (Total stress tensor). After introducing the total stress tensor $\mathbb{r}(\boldsymbol{u}, p) := \mathbb{s}(\boldsymbol{u}) - p\mathbb{I}$, one can rewrite the momentum balance equation (53.1a) in the form $-\nabla{\cdot}\mathbb{r}(\boldsymbol{u}, p) = \boldsymbol{f}$, and the Neumann condition on ∂D_{n} as $\mathbb{r}(\boldsymbol{u}, p)_{|\partial D_{\mathrm{n}}}\boldsymbol{n} = \boldsymbol{a}_{\mathrm{n}}$. $\square$

Remark 53.2 (Incompressibility). The field $\boldsymbol{u}$ is said to be incompressible, or divergence-free, if $\nabla{\cdot}\boldsymbol{u} = g = 0$. In the incompressible regime, (53.2) simplifies to $\mathbb{s}(\boldsymbol{u}) = 2\mu\mathbb{e}(\boldsymbol{u})$. $\square$

Remark 53.3 (Laplacian/Cauchy–Navier form). When $g = 0$ and the dynamic viscosity is constant, the momentum equation can be simplified by observing that $\nabla{\cdot}((\nabla\boldsymbol{u})^{\mathsf{T}}) = \nabla(\nabla{\cdot}\boldsymbol{u}) = \boldsymbol{0}$. The momentum equation can then be rewritten in the Laplacian (or *Cauchy–Navier*) form $-\mu\Delta\boldsymbol{u} + \nabla p = \boldsymbol{f}$, and the Neumann boundary condition becomes $\mu\partial_n\boldsymbol{u}_{|\partial D_{\mathrm{n}}} - p_{|\partial D_{\mathrm{n}}}\boldsymbol{n} = \boldsymbol{a}_{\mathrm{n}}$. $\square$

Remark 53.4 (Pressure constant). When $\partial D = \partial D_{\mathrm{d}}$, the data fields g and $\boldsymbol{a}_{\mathrm{d}}$ must satisfy the compatibility condition $\int_D g\,\mathrm{d}x = \int_{\partial D} \boldsymbol{a}_{\mathrm{d}}{\cdot}\boldsymbol{n}\,\mathrm{d}s$, and the pressure is determined up to an additive constant. This indetermination is usually removed by assuming that $\int_D p\,\mathrm{d}x = 0$. $\square$

Remark 53.5 ($\lambda = 0$). Since $\nabla{\cdot}(\lambda(\nabla{\cdot}\boldsymbol{u})\mathbb{I}) = \nabla(\lambda\nabla{\cdot}\boldsymbol{u})$, we can redefine the pressure and the viscous stress tensor by setting $p' := p - \lambda\nabla{\cdot}\boldsymbol{u}$ and $\mathbb{s}'(\boldsymbol{u}) := 2\mu\mathbb{e}(\boldsymbol{u})$. Then the momentum balance equation (53.1a) becomes $-\nabla{\cdot}\mathbb{s}'(\boldsymbol{u}) + \nabla p' = \boldsymbol{f}$. We adopt this change of variable in what follows, i.e., we assume that $\mathbb{s}(\boldsymbol{u}) := 2\mu\mathbb{e}(\boldsymbol{u})$ from now on. $\square$

Remark 53.6 (Homogeneous Dirichlet condition). Let us assume that there is a function $\boldsymbol{u}_{\mathrm{d}}$ (smooth enough) s.t. $(\boldsymbol{u}_{\mathrm{d}})_{|\partial D_{\mathrm{d}}} = \boldsymbol{a}_{\mathrm{d}}$. Then we can make the change of variable $\boldsymbol{u}' := \boldsymbol{u} - \boldsymbol{u}_{\mathrm{d}}$ so that $\boldsymbol{u}'$ satisfies the homogeneous boundary condition $\boldsymbol{u}'_{|\partial D_{\mathrm{d}}} = \boldsymbol{0}$. Upon denoting $\boldsymbol{f}' := \boldsymbol{f} + \nabla{\cdot}(\mathbb{s}(\boldsymbol{u}_{\mathrm{d}}))$, $g' := g - \nabla{\cdot}\boldsymbol{u}_{\mathrm{d}}$, and inserting the definition $\boldsymbol{u} = \boldsymbol{u}' + \boldsymbol{u}_{\mathrm{d}}$ into (53.1), one observes that the pair $(\boldsymbol{u}', p)$ solves a Stokes problem with homogeneous Dirichlet data and with source terms $\boldsymbol{f}'$ and g'. From now on, we abuse the notation and use the symbols $\boldsymbol{u}$, $\boldsymbol{f}$, g instead of $\boldsymbol{u}'$, $\boldsymbol{f}'$, g'. This is equivalent to assuming that $\boldsymbol{a}_{\mathrm{d}} = \boldsymbol{0}$. $\square$

53.2   Weak formulation and well-posedness

In this section, we present a weak formulation of the Stokes equations and we establish its well-posedness using the Babuška–Brezzi theorem (Theorem 49.13).

53.2.1   Weak formulation

Let $\boldsymbol{w}$ be a sufficiently smooth $\mathbb{R}^d$-valued test function. Since the velocity $\boldsymbol{u}$ vanishes on ∂D_{d}, we only consider test functions $\boldsymbol{w}$ that vanish on ∂D_{d}. Multiplying (53.1a) by $\boldsymbol{w}$ and integrating over D gives

$$-\int_D (\nabla{\cdot}\mathbb{s}(\boldsymbol{u})){\cdot}\boldsymbol{w}\,\mathrm{d}x + \int_D \nabla p{\cdot}\boldsymbol{w}\,\mathrm{d}x = \int_D \boldsymbol{f}{\cdot}\boldsymbol{w}\,\mathrm{d}x.$$

Integrating by parts the term involving the viscous stress tensor, we obtain

$$-\int_D (\nabla{\cdot}\mathbb{s}(\boldsymbol{u})){\cdot}\boldsymbol{w}\,\mathrm{d}x = \int_D \mathbb{s}(\boldsymbol{u}){:}\nabla\boldsymbol{w}\,\mathrm{d}x - \int_{\partial D_{\mathrm{n}}} (\mathbb{s}(\boldsymbol{u})\boldsymbol{n}){\cdot}\boldsymbol{w}\,\mathrm{d}s,$$

where $\boldsymbol{n} := (n_1, \ldots, n_d)^\mathsf{T}$ is the outward unit normal to D. The boundary integral over ∂D_d is zero since $\boldsymbol{w}$ vanishes on ∂D_d. The symmetry of $\mathbb{s}(\boldsymbol{u})$ implies that $\mathbb{s}(\boldsymbol{u}){:}\nabla\boldsymbol{w} = \mathbb{s}(\boldsymbol{u}){:}\mathbb{e}(\boldsymbol{w})$. Similarly, the term $\int_D \nabla p{\cdot}\boldsymbol{w}\,\mathrm{d}x$ is equal to $-\int_D p\nabla{\cdot}\boldsymbol{w}\,\mathrm{d}x + \int_{\partial D_\mathrm{n}} p\boldsymbol{n}{\cdot}\boldsymbol{w}\,\mathrm{d}s$. Combining the above equations and using the Neumann boundary condition $\mathbb{s}(\boldsymbol{u})_{|\partial D_\mathrm{n}}\boldsymbol{n} - p_{|\partial D_\mathrm{n}}\boldsymbol{n} = \boldsymbol{a}_\mathrm{n}$, the weak form of the momentum equation is

$$\int_D (\mathbb{s}(\boldsymbol{u}){:}\mathbb{e}(\boldsymbol{w}) - p\nabla{\cdot}\boldsymbol{w})\,\mathrm{d}x = \int_D \boldsymbol{f}{\cdot}\boldsymbol{w}\,\mathrm{d}x + \int_{\partial D_\mathrm{n}} \boldsymbol{a}_\mathrm{n}{\cdot}\boldsymbol{w}\,\mathrm{d}s.$$

The three integrals are well defined if $p \in L^2(D)$, $\boldsymbol{f} \in \boldsymbol{L}^2(D)$, $\boldsymbol{a}_\mathrm{n} \in L^2(\partial D_\mathrm{n})$, and if $\boldsymbol{u}$, $\boldsymbol{w}$ are in the space

$$\boldsymbol{V}_\mathrm{d}(D) := \{\boldsymbol{v} \in \boldsymbol{H}^1(D) \mid \gamma^\mathrm{g}(\boldsymbol{v})_{|\partial D_\mathrm{d}} = \boldsymbol{0}\}, \tag{53.3}$$

with the $\mathbb{R}^d$-valued trace operator $\gamma^\mathrm{g} : \boldsymbol{H}^1(D) \to \boldsymbol{H}^{\frac{1}{2}}(\partial D)$ acting componentwise as the scalar-valued trace operator $\gamma^\mathrm{g} : H^1(D) \to H^{\frac{1}{2}}(\partial D)$. We equip the space $\boldsymbol{V}_\mathrm{d}$ with the norm $\|\boldsymbol{v}\|_{\boldsymbol{V}_\mathrm{d}} := |\boldsymbol{v}|_{\boldsymbol{H}^1(D)} = \|\nabla\boldsymbol{v}\|_{\mathbb{L}^2(D)}$. Since $|\partial D_\mathrm{d}| > 0$, we infer from the Poincaré–Steklov inequality (42.9) that there is a constant $\tilde{C}_\mathrm{PS} > 0$ s.t. $\tilde{C}_\mathrm{PS}\|\boldsymbol{v}\|_{\boldsymbol{L}^2(D)} \le \ell_D\|\nabla\boldsymbol{v}\|_{\mathbb{L}^2(D)}$ for all $\boldsymbol{v} \in \boldsymbol{V}_\mathrm{d}$ (recall that ℓ_D is a length scale associated with D, e.g., $\ell_D := \mathrm{diam}(D)$). This argument shows that $\|{\cdot}\|_{\boldsymbol{V}_\mathrm{d}}$ is a norm on $\boldsymbol{V}_\mathrm{d}$, equivalent to the $\|{\cdot}\|_{\boldsymbol{H}^1(D)}$-norm.

A weak formulation of the mass conservation (53.1b) is obtained as above by testing the equation against a sufficiently smooth scalar-valued function q. No integration by parts needs to be performed, and we simply write

$$\int_D q\nabla{\cdot}\boldsymbol{u}\,\mathrm{d}x = \int_D gq\,\mathrm{d}x.$$

The left-hand side is well defined provided $q \in L^2(D)$ and $\boldsymbol{u}$ is in $\boldsymbol{V}_\mathrm{d}$. Note that if $\partial D = \partial D_\mathrm{d}$, the compatibility condition $\int_D g\,\mathrm{d}x = 0$ implies the equality $\int_D \nabla{\cdot}\boldsymbol{u}\,\mathrm{d}x = \int_D g\,\mathrm{d}x$, meaning that the mass conservation equation need not be tested against constant functions. In this particular case, the test functions q must be restricted to be of zero mean over D. This motivates the following definition:

$$Q := \begin{cases} L^2(D) & \text{if } \partial D \ne \partial D_\mathrm{d}, \\ L^2_*(D) := \{q \in L^2(D) \mid \int_D q\,\mathrm{d}x = 0\} & \text{if } \partial D = \partial D_\mathrm{d}. \end{cases} \tag{53.4}$$

We equip the space Q with the L^2-norm. Let us define the bilinear forms

$$a(\boldsymbol{v}, \boldsymbol{w}) := \int_D \mathbb{s}(\boldsymbol{v}){:}\mathbb{e}(\boldsymbol{w})\,\mathrm{d}x, \qquad b(\boldsymbol{w}, q) := -\int_D q\nabla{\cdot}\boldsymbol{w}\,\mathrm{d}x, \tag{53.5}$$

on $\boldsymbol{V}_\mathrm{d}{\times}\boldsymbol{V}_\mathrm{d}$ and $\boldsymbol{V}_\mathrm{d}{\times}Q$, respectively. We also define the linear forms $F(\boldsymbol{w}) := \int_D \boldsymbol{f}{\cdot}\boldsymbol{w}\,\mathrm{d}x + \int_{\partial D_\mathrm{n}} \boldsymbol{a}_\mathrm{n}{\cdot}\boldsymbol{w}\,\mathrm{d}s$, $G(q) := -\int_D gq\,\mathrm{d}x$ on $\boldsymbol{V}_\mathrm{d}$ and Q, respectively. Assuming enough smoothness on $\boldsymbol{f}$, $\boldsymbol{a}_\mathrm{n}$, and g, it is reasonable to expect that $F \in \mathcal{L}(\boldsymbol{V}_\mathrm{d}; \mathbb{R})$ and $G \in \mathcal{L}(Q; \mathbb{R})$. We obtain the following weak formulation:

$$\begin{cases} \text{Find } \boldsymbol{u} \in \boldsymbol{V}_\mathrm{d} \text{ and } p \in Q \text{ such that} \\ a(\boldsymbol{u}, \boldsymbol{w}) + b(\boldsymbol{w}, p) = F(\boldsymbol{w}), & \forall \boldsymbol{w} \in \boldsymbol{V}_\mathrm{d}, \\ b(\boldsymbol{u}, q) = G(q), & \forall q \in Q. \end{cases} \tag{53.6}$$

Proposition 53.7 (Weak solution). *Assume $\boldsymbol{f} \in \boldsymbol{L}^2(D)$, $g \in Q$, and $\boldsymbol{a}_\mathrm{n} \in \boldsymbol{L}^2(\partial D_\mathrm{n})$. Any weak solution $(\boldsymbol{u}, p)$ to (53.6) satisfies (53.1a)-(53.1b) a.e. in D and satisfies the boundary condition (53.1c) a.e. on ∂D.*

Proof. Let us set $\mathbb{r}(\boldsymbol{u}, p) := \mathbb{s}(\boldsymbol{u}) - p\mathbb{I} \in \mathbb{L}^2(D)$. Testing the momentum equation in (53.6) against an arbitrary function $\boldsymbol{w} \in \boldsymbol{C}_0^\infty(D)$, we infer that $\mathbb{r}(\boldsymbol{u}, p)$ has a weak divergence in $\boldsymbol{L}^2(D)$ equal to $-\boldsymbol{f}$. Since $\nabla \cdot \mathbb{r}(\boldsymbol{u}, p) = \nabla \cdot \mathbb{s}(\boldsymbol{u}) - \nabla p$, we infer that (53.1a) is satisfied a.e. in D. Testing the mass equation in (53.6) against an arbitrary function $q \in C_0^\infty(D)$, we infer that (53.1b) is satisfied a.e. in D (if $\partial D = \partial D_\mathrm{d}$, the compatibility condition $\int_D g \, \mathrm{d}x = 0$ implies that $b(\boldsymbol{u}, q) = G(q)$ for all $q \in L^2(D)$). The Dirichlet boundary condition $\boldsymbol{u}_{|\partial D_\mathrm{d}} = \boldsymbol{0}$ is a natural consequence of the trace theorem (Theorem 3.10) and $\boldsymbol{u}$ being in $\boldsymbol{V}_\mathrm{d}$. To derive the Neumann condition, we proceed as in §31.3.3. Since $\nabla \cdot \mathbb{r}(\boldsymbol{u}, p) = -\boldsymbol{f} \in \boldsymbol{L}^2(D)$, we have $\mathbb{r}(\boldsymbol{u}, p) \in \mathbb{H}(\mathrm{div}; D)$ (i.e., each row of $\mathbb{r}(\boldsymbol{u}, p)$ is in $\boldsymbol{H}(\mathrm{div}; D)$). Owing to Theorem 4.15, we infer that $\mathbb{r}(\boldsymbol{u}, p)\boldsymbol{n} \in \boldsymbol{H}^{-\frac{1}{2}}(\partial D)$. As a result, we have

$$\langle \mathbb{r}(\boldsymbol{u}, p)\boldsymbol{n}, \gamma^\mathrm{g}(\boldsymbol{w}) \rangle_{\partial D} = \int_D (\mathbb{r}(\boldsymbol{u}, p){:}\nabla\boldsymbol{w} + (\nabla \cdot \mathbb{r}(\boldsymbol{u}, p)) \cdot \boldsymbol{w}) \, \mathrm{d}x$$

$$= \int_D (\mathbb{s}(\boldsymbol{u}){:}\mathbb{e}(\boldsymbol{w}) - p\nabla \cdot \boldsymbol{w} - \boldsymbol{f} \cdot \boldsymbol{w}) \, \mathrm{d}x = \int_{\partial D_\mathrm{n}} \boldsymbol{a}_\mathrm{n} \cdot \gamma^\mathrm{g}(\boldsymbol{w}) \, \mathrm{d}s, \qquad \forall \boldsymbol{w} \in \boldsymbol{V}_\mathrm{d},$$

which implies that the Neumann condition $\mathbb{r}(\boldsymbol{u}, p)\boldsymbol{n} = \boldsymbol{a}_\mathrm{n}$ is satisfied in $\widetilde{\boldsymbol{H}}^{\frac{1}{2}}(\partial D_\mathrm{n})'$, with $\widetilde{\boldsymbol{H}}^{\frac{1}{2}}(\partial D_\mathrm{n}) := \{\boldsymbol{v} \in \boldsymbol{H}^{\frac{1}{2}}(\partial D_\mathrm{n}) \mid \widetilde{\boldsymbol{v}} \in \boldsymbol{H}^{\frac{1}{2}}(\partial D)\}$ (recall that $\widetilde{\boldsymbol{v}}$ is the zero extension of $\boldsymbol{v}$ to ∂D). Actually, the Neumann condition is satisfied a.e. on ∂D_n since we assumed $\boldsymbol{a}_\mathrm{n} \in \boldsymbol{L}^2(\partial D_\mathrm{n})$. $\qquad\square$

Remark 53.8 (Neumann data). The above proof shows that it is possible to take more generally $\boldsymbol{a}_\mathrm{n} \in \widetilde{\boldsymbol{H}}^{\frac{1}{2}}(\partial D_\mathrm{n})'$. $\qquad\square$

53.2.2 Well-posedness

One readily sees that the bilinear form $a(\boldsymbol{v}, \boldsymbol{w}) := (\mathbb{s}(\boldsymbol{v}), \mathbb{e}(\boldsymbol{w}))_{\mathbb{L}^2(D)}$ defined in (53.5) is coercive and bounded on $\boldsymbol{V}_\mathrm{d} \times \boldsymbol{V}_\mathrm{d}$. The coercivity of a has been established in Theorem 42.11 as a consequence of Korn's inequalities. In particular, there is $C_\mathrm{K} > 0$ s.t. $\|\mathbb{e}(\boldsymbol{v})\|_{\mathbb{L}^2(D)} \geq C_\mathrm{K}|\boldsymbol{v}|_{\boldsymbol{H}^1(D)}$ for all $\boldsymbol{v} \in \boldsymbol{V}_\mathrm{d}$, and this implies that (see (42.15) with $\rho_{\min} := 2\mu$ in the present setting)

$$a(\boldsymbol{v}, \boldsymbol{v}) \geq 2\mu C_\mathrm{K}^2 |\boldsymbol{v}|_{\boldsymbol{H}^1(D)}^2, \qquad \forall \boldsymbol{v} \in \boldsymbol{V}_\mathrm{d}. \tag{53.7}$$

Moreover, the Cauchy–Schwarz inequality and the bound $\|\mathbb{e}(\boldsymbol{v})\|_{\mathbb{L}^2(D)} \leq |\boldsymbol{v}|_{\boldsymbol{H}^1(D)}$ show that the boundedness constant of the bilinear form a satisfies $\|a\| \leq 2\mu$. Hence, the key argument for the well-posedness of the Stokes problem is the surjectivity of the divergence operator $\nabla \cdot : \boldsymbol{V}_\mathrm{d} \to Q$. This result is a bit more subtle than Lemma 51.2 since $\boldsymbol{V}_\mathrm{d}$ is a smaller space than $\boldsymbol{H}(\mathrm{div}; D)$.

Lemma 53.9 ($\nabla\cdot$ is surjective). *Let D be a Lipschitz domain in $\mathbb{R}^d$. (i) Case $\partial D = \partial D_\mathrm{d}$. $\nabla \cdot : \boldsymbol{H}_0^1(D) \to L_*^2(D)$ is surjective. (ii) Case $\partial D \neq \partial D_\mathrm{d}$. Consider the partition $\partial D = \partial D_\mathrm{d} \cup \partial D_\mathrm{n}$ with $|\partial D_\mathrm{d}| > 0$. Assume that $|\partial D_\mathrm{n}| > 0$ and that there exists a subset $\mathcal{O}$ of ∂D_n with $|\mathcal{O}| > 0$ and $\boldsymbol{n}_{|\mathcal{O}} \in \boldsymbol{H}^{\frac{1}{2}}(\mathcal{O})$. Then the operator $\nabla \cdot : \boldsymbol{X} := \{\boldsymbol{v} \in \boldsymbol{V}_\mathrm{d} \mid \gamma^\mathrm{g}(\boldsymbol{v})_{|\partial D_\mathrm{n}} \times \boldsymbol{n} = \boldsymbol{0}\} \to L^2(D)$ is surjective. (iii) In all the cases, identifying Q' with Q we have*

$$\inf_{q \in Q} \sup_{\boldsymbol{v} \in \boldsymbol{V}_\mathrm{d}} \frac{|\int_D q\nabla \cdot \boldsymbol{v} \, \mathrm{d}x|}{\|q\|_{L^2(D)}|\boldsymbol{v}|_{\boldsymbol{H}^1(D)}} := \beta_D > 0. \tag{53.8}$$

Proof. (i) We refer the reader to Girault and Raviart [217, pp. 18-26] for a proof of the surjectivity of $\nabla \cdot : \boldsymbol{H}_0^1(D) \to L_*^2(D)$, (see also Exercise 53.1 if D is a smooth domain). (ii) Let us now consider the second case. Let q be in $L^2(D)$. Let ρ be a smooth nonnegative function compactly supported in $\mathcal{O}$ such that $\int_\mathcal{O} \rho \, \mathrm{d}s > 0$ (this is possible since $|\mathcal{O}| > 0$). Let $\boldsymbol{g} := c\rho\boldsymbol{n}$ be a vector field in $\mathcal{O}$, where the constant c is chosen s.t. $\int_\mathcal{O} \boldsymbol{g} \cdot \boldsymbol{n} \, \mathrm{d}s = \int_D q \, \mathrm{d}x$. Let $\widetilde{\boldsymbol{g}}$ be the zero extension of $\boldsymbol{g}$ to ∂D.

Since $\boldsymbol{n}_{|\mathcal{O}} \in \boldsymbol{H}^{\frac{1}{2}}(\mathcal{O})$, we have $\rho\boldsymbol{n} \in \widetilde{\boldsymbol{H}}^{\frac{1}{2}}(\mathcal{O})$. Hence, $\widetilde{\boldsymbol{g}}$ is in $\boldsymbol{H}^{\frac{1}{2}}(\partial D)$ so that it is possible to find a function $\boldsymbol{w}$ in $\boldsymbol{H}^1(D)$ s.t. $\gamma^{\mathrm{g}}(\boldsymbol{w}) = \widetilde{\boldsymbol{g}}$ on ∂D. We have $\gamma^{\mathrm{g}}(\boldsymbol{w})_{|\partial D_{\mathrm{d}}} = \boldsymbol{0}$ and $\gamma^{\mathrm{g}}(\boldsymbol{w})_{|\partial D_{\mathrm{n}}} \times \boldsymbol{n} = \boldsymbol{0}$, i.e., $\boldsymbol{w} \in \boldsymbol{X}$. Now let $q_0 := \nabla{\cdot}\boldsymbol{w} - q$. The above definitions and the divergence formula imply that $q_0 \in L^2(D)$ and $\int_D q_0 \, \mathrm{d}x = 0$. Hence, q_0 is in $L_*^2(D)$. Since $\nabla{\cdot} : \boldsymbol{H}_0^1(D) \to L_*^2(D)$ is surjective, there is $\boldsymbol{w}_0 \in \boldsymbol{H}_0^1(D)$ such that $\nabla{\cdot}\boldsymbol{w}_0 = -q_0$. Thus, for all q in $L^2(D)$ the function $\boldsymbol{w} + \boldsymbol{w}_0$ is in $\boldsymbol{X}$ with $\nabla{\cdot}(\boldsymbol{w} + \boldsymbol{w}_0) = q$, that is, $\nabla{\cdot}\boldsymbol{X} \to L^2(D)$ is surjective. This also implies that $\nabla{\cdot} : \boldsymbol{V}_{\mathrm{d}} \to Q$ is surjective. (iii) The inf-sup condition (53.8) follows from the surjectivity of $\nabla{\cdot} : \boldsymbol{V}_{\mathrm{d}} \to Q$ and Lemma C.40. □

Remark 53.10 (Inf-sup condition in $W^{1,p}$-$L^{p'}$). Let $p \in (1,\infty)$ and let $p' \in (1,\infty)$ be s.t. $\frac{1}{p} + \frac{1}{p'} = 1$. Then the operator $\nabla{\cdot} : \boldsymbol{W}_0^{1,p}(D) \to L_*^p(D) := \{q \in L^p(D) \mid \int_D q \, \mathrm{d}x = 0\}$ is surjective (see Auscher et al. [30, Lem. 10]), that is, identifying $(L_*^p(D))'$ with $L_*^{p'}(D)$, we have

$$\inf_{q \in L_*^{p'}(D)} \sup_{\boldsymbol{v} \in \boldsymbol{W}_0^{1,p}(D)} \frac{|\int_D q \nabla{\cdot}\boldsymbol{v} \, \mathrm{d}x|}{\|q\|_{L^{p'}(D)} |\boldsymbol{v}|_{\boldsymbol{W}^{1,p}(D)}} := \beta_{D,p} > 0. \tag{53.9}$$

The assumption that D is Lipschitz can be weakened. For instance, the inf-sup condition (53.9) holds true also if D is a bounded open set in $\mathbb{R}^d$ and if D is star-shaped with respect to an open ball $B \subset D$, i.e., for all $\boldsymbol{x} \in D$ and $\boldsymbol{z} \in B$, the segment joining $\boldsymbol{x}$ and $\boldsymbol{z}$ is contained in D; see Bogovskiĭ [66], Galdi [210, Lem. 3.1, Chap. III], Durán and Muschietti [181], Durán et al. [180], Solonnikov [349, Prop. 2.1], Costabel and McIntosh [146]. □

Let $B : \boldsymbol{V}_{\mathrm{d}} \to Q'$ be s.t. $\langle B(\boldsymbol{v}), q \rangle_{Q',Q} := b(\boldsymbol{v}, q) := -\int_D q(\nabla{\cdot}\boldsymbol{v}) \, \mathrm{d}x$. Identifying Q and Q', we have $B(\boldsymbol{v}) = -\nabla{\cdot}\boldsymbol{v}$, and $\ker(B) := \{\boldsymbol{v} \in \boldsymbol{V}_{\mathrm{d}} \mid \nabla{\cdot}\boldsymbol{v} = 0\}$.

Theorem 53.11 (Well-posedness). (i) *The weak formulation* (53.6) *of the Stokes problem is well-posed.* (ii) T*here is c such that for all $\boldsymbol{f} \in \boldsymbol{L}^2(D)$, all $g \in Q$, and all $\boldsymbol{a}_{\mathrm{n}} \in \boldsymbol{L}^2(\partial D_{\mathrm{n}})$,*

$$2\mu|\boldsymbol{u}|_{\boldsymbol{H}^1(D)} + \|p\|_{L^2(D)} \le c \left(\ell_D \|\boldsymbol{f}\|_{\boldsymbol{L}^2(D)} + \mu\|g\|_{L^2(D)} + \ell_D^{\frac{1}{2}} \|\boldsymbol{a}_{\mathrm{n}}\|_{\boldsymbol{L}^2(\partial D_{\mathrm{n}})} \right).$$

Proof. We apply the Babuška–Brezzi theorem (Theorem 49.13). The inf-sup condition (49.37) on the bilinear form b follows from Lemma 53.9. The two conditions in (49.36) are satisfied owing to the coercivity of the bilinear form a on $\boldsymbol{V}_{\mathrm{d}}$ (see (53.7)). Finally, the stability estimate follows from (49.38). □

One can formulate a more precise stability result on the product space $Y := \boldsymbol{V}_{\mathrm{d}} \times Q$ equipped with the norm $\|(\boldsymbol{v}, q)\|_Y^2 := \mu|\boldsymbol{v}|_{\boldsymbol{H}^1(D)}^2 + \mu^{-1}\|p\|_{L^2(D)}^2$, and the bilinear form $t((\boldsymbol{v}, q), (\boldsymbol{w}, r)) := a(\boldsymbol{v}, \boldsymbol{w}) + b(\boldsymbol{w}, q) - b(\boldsymbol{v}, r)$ on $Y \times Y$.

Lemma 53.12 (Inf-sup condition). *The following holds true:*

$$\inf_{(\boldsymbol{v},q) \in Y} \sup_{(\boldsymbol{w},r) \in Y} \frac{|t((\boldsymbol{v}, q), (\boldsymbol{w}, r))|}{\|(\boldsymbol{v}, q)\|_Y \|(\boldsymbol{w}, r)\|_Y} =: \gamma > 0, \tag{53.10}$$

where γ is uniform w.r.t. $\mu > 0$.

Proof. Let $(\boldsymbol{v}, q) \in Y$ and let us set $\mathbb{S} := \sup_{(\boldsymbol{w},r) \in Y} \frac{|t((\boldsymbol{v},q),(\boldsymbol{w},r))|}{\|(\boldsymbol{w},r)\|_Y}$. Owing to (53.7), we have

$$2\mu C_{\mathrm{K}}^2 |\boldsymbol{v}|_{\boldsymbol{H}^1(D)}^2 \le a(\boldsymbol{v}, \boldsymbol{v}) = t((\boldsymbol{v}, q), (\boldsymbol{v}, q)) \le \mathbb{S}\|(\boldsymbol{v}, q)\|_Y. \tag{53.11}$$

Moreover, owing to Lemma 53.9, there is $\boldsymbol{w}_q \in \boldsymbol{V}_{\mathrm{d}}$ s.t.

$$\nabla{\cdot}\boldsymbol{w}_q = -\mu^{-1}q, \qquad |\boldsymbol{w}_q|_{\boldsymbol{H}^1(D)} \le (\beta_D\mu)^{-1}\|q\|_{L^2(D)}.$$

We obtain

$$\mu^{-1}\|q\|_{L^2(D)}^2 = -(q, \nabla\cdot\boldsymbol{w}_q) = -t((\boldsymbol{v}, q), (\boldsymbol{w}_q, 0)) + a(\boldsymbol{v}, \boldsymbol{w}_q)$$

$$\leq \mathbb{S}\mu^{\frac{1}{2}}|\boldsymbol{w}_q|_{\boldsymbol{H}^1(D)} + 2\mu^{\frac{1}{2}}|\boldsymbol{v}|_{\boldsymbol{H}^1(D)}\mu^{\frac{1}{2}}|\boldsymbol{w}_q|_{\boldsymbol{H}^1(D)}$$

$$\leq c'\,(\mathbb{S} + \mathbb{S}^{\frac{1}{2}}\|(\boldsymbol{v}, q)\|_Y^{\frac{1}{2}})\mu^{\frac{1}{2}}|\boldsymbol{w}_q|_{\boldsymbol{H}^1(D)},$$

where we used that $|a(\boldsymbol{v}, \boldsymbol{w})| \leq 2\mu|\boldsymbol{v}|_{\boldsymbol{H}^1(D)}|\boldsymbol{w}|_{\boldsymbol{H}^1(D)}$ and then (53.11). Using the bound on $|\boldsymbol{w}_q|_{\boldsymbol{H}^1(D)}$ and Young's inequality leads to

$$\mu^{-1}\|q\|_{L^2(D)}^2 \leq c\,(\mathbb{S}^2 + \mathbb{S}\|(\boldsymbol{v}, q)\|_Y).$$

We can now combine this bound with (53.11) to infer that

$$\|(\boldsymbol{v}, q)\|_Y^2 \leq c\,(\mathbb{S}^2 + \mathbb{S}\|(\boldsymbol{v}, q)\|_Y).$$

Applying one more time Young's inequality yields $\|(\boldsymbol{v}, q)\|_Y \leq c\mathbb{S}$. $\qquad\square$

Remark 53.13 (Helmholtz decomposition). Letting $H_*^1(D) := H^1(D) \cap L_*^2(D)$ and $\boldsymbol{\mathcal{H}} := \{\boldsymbol{v} \in \boldsymbol{L}^2(D) \mid \nabla\cdot\boldsymbol{v} = 0,\ \boldsymbol{v}_{|\partial D}\cdot\boldsymbol{n} = 0\}$, the following $\boldsymbol{L}^2$-orthogonal Helmholtz decomposition holds true: $\boldsymbol{L}^2(D) = \boldsymbol{\mathcal{H}} \oplus \nabla(H_*^1(D))$ (see Lemma 74.1). The $\boldsymbol{L}^2$-orthogonal projection $\boldsymbol{P}_{\boldsymbol{\mathcal{H}}} : \boldsymbol{L}^2(D) \to \boldsymbol{\mathcal{H}}$ resulting from this decomposition is often called *Leray projection*. Let $(\boldsymbol{u}, p)$ solve (53.6). Assume for simplicity that the homogeneous Dirichlet condition $\boldsymbol{u}_{|\partial D} = \boldsymbol{0}$ is enforced over the whole boundary and assume that $g = 0$. Since $\boldsymbol{u}$ is divergence-free and vanishes at the boundary, we have $(\boldsymbol{f}, \boldsymbol{u})_{\boldsymbol{L}^2(D)} = (\boldsymbol{P}_{\boldsymbol{\mathcal{H}}}(\boldsymbol{f}), \boldsymbol{u})_{\boldsymbol{L}^2(D)}$. Then taking $\boldsymbol{w} := \boldsymbol{u}$ in (53.6) and invoking the coercivity of a shows that $2\mu C_{\mathrm{K}}^2|\boldsymbol{u}|_{\boldsymbol{H}^1(D)}^2 \leq a(\boldsymbol{u}, \boldsymbol{u}) = (\boldsymbol{P}_{\boldsymbol{\mathcal{H}}}(\boldsymbol{f}), \boldsymbol{u})_{\boldsymbol{L}^2(D)}$. Owing to the Cauchy–Schwarz inequality and the Poincaré–Steklov inequality, we get

$$2\mu|\boldsymbol{u}|_{\boldsymbol{H}^1(D)} \leq C_{\mathrm{PS}}^{-1}C_{\mathrm{K}}^{-2}\ell_D\|\boldsymbol{P}_{\boldsymbol{\mathcal{H}}}(\boldsymbol{f})\|_{\boldsymbol{L}^2(D)}.$$

This a priori estimate on the velocity is sharper than the one from Theorem 53.11 since $\|\boldsymbol{P}_{\boldsymbol{\mathcal{H}}}(\boldsymbol{f})\|_{\boldsymbol{L}^2(D)}$ appears on the right-hand side instead of $\|\boldsymbol{f}\|_{\boldsymbol{L}^2(D)}$. One should bear in mind that, even if $p \in H_*^1(D)$, the fields $-\nabla\cdot\mathbb{s}(\boldsymbol{u})$ and $\boldsymbol{P}_{\boldsymbol{\mathcal{H}}}(\boldsymbol{f})$ are generally different since the normal component of $\nabla\cdot\mathbb{s}(\boldsymbol{u})$ at ∂D is generally nonzero. $\qquad\square$

53.2.3 Regularity pickup

Regularity properties for the Stokes problems can be established when μ and λ are both constant (or smooth) and $|\partial D_{\mathrm{n}}| = 0$. For instance, if ∂D is of class C^∞, for all $s > 0$ there is c, depending on D and s, such that

$$\mu\ell_D^{-1}\|\boldsymbol{u}\|_{\boldsymbol{H}^{1+s}(D)} + \|p\|_{H^s(D)} \leq c\,(\ell_D\|\boldsymbol{f}\|_{\boldsymbol{H}^{s-1}(D)} + \mu\|g\|_{H^s(D)}). \qquad (53.12)$$

There is an upper limit on s when D is not smooth. For instance, let D be a two-dimensional convex polygon. Let $\rho : D \to \mathbb{R}$ be the distance to the closest vertex of D. It is shown in Kellogg and Osborn [266, Thm. 2] that there is a constant c that depends only on D such that

$$\mu|\boldsymbol{u}|_{\boldsymbol{H}^2(D)} + |p|_{H^1(D)} \leq c\,\big(\|\boldsymbol{f}\|_{\boldsymbol{L}^2(D)} + \mu\ell_D^{-1}(|g|_{H^1(D)} + \|\rho^{-1}g\|_{L^2(D)})\big). \qquad (53.13)$$

The situation is a bit more complicated in dimension three. We refer to Dauge [153] for an overview of the problem. Assuming that $g = 0$, it is shown in [153, p. 75] that (53.12) holds true in the following situations: (i) For all $s \leq 1$ if D is a convex polyhedron; (ii) For all $s < \frac{3}{2}$ if D is any convex domain with wedge angles $\leq \frac{2}{3}\pi$; (iii) For all $s < \frac{1}{2}$ if D has a piecewise smooth boundary, and its faces meet two by two or three by three with independent normal vectors at the meeting points.

53.3 Conforming approximation

In the rest of this chapter, we assume that D is a polyhedron in $\mathbb{R}^d$ and $(\mathcal{T}_h)_{h \in \mathcal{H}}$ is a shape-regular sequence of matching meshes so that each mesh covers D exactly. We also assume that ∂D_{d} is a union of mesh faces. Let $(\boldsymbol{V}_{h\mathrm{d}} \subset \boldsymbol{V}_{\mathrm{d}})_{h \in \mathcal{H}}$ and $(Q_h \subset Q)_{h \in \mathcal{H}}$ be sequences of finite-dimensional spaces built using $(\mathcal{T}_h)_{h \in \mathcal{H}}$. Notice that the inclusion $\boldsymbol{V}_{h\mathrm{d}} \subset \boldsymbol{V}_{\mathrm{d}}$ means that the homogeneous Dirichlet condition on the velocity is strongly enforced on ∂D_{d}. The discrete counterpart of the problem (53.6) is as follows:

$$\begin{cases} \text{Find } \boldsymbol{u}_h \in \boldsymbol{V}_{h\mathrm{d}} \text{ and } p_h \in Q_h \text{ such that} \\ a(\boldsymbol{u}_h, \boldsymbol{v}_h) + b(\boldsymbol{v}_h, p_h) = F(\boldsymbol{v}_h), & \forall \boldsymbol{v}_h \in \boldsymbol{V}_{h\mathrm{d}}, \\ b(\boldsymbol{u}_h, q_h) = G(q_h), & \forall q_h \in Q_h. \end{cases} \tag{53.14}$$

Since $\boldsymbol{V}_{h\mathrm{d}}$ is $\boldsymbol{V}_{\mathrm{d}}$-conforming, the discrete formulation inherits the coercivity of a. Unfortunately, there is no reason a priori for the discrete formulation to inherit the surjectivity of the divergence operator established in Lemma 53.9. Verifying this condition is the crucial step in devising stable mixed finite elements for the Stokes problem.

Proposition 53.14 (Well-posedness). *The discrete problem (53.14) is well-posed if and only if the following inf-sup condition holds true:*

$$\inf_{q_h \in Q_h} \sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h\mathrm{d}}} \frac{|\int_D q_h \nabla \cdot \boldsymbol{v}_h \, \mathrm{d}x|}{\|q_h\|_{L^2(D)} |\boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}} =: \beta_h > 0. \tag{53.15}$$

Proof. Apply Proposition 50.1. □

We henceforth say that the inf-sup condition (53.15) holds uniformly w.r.t. $h \in \mathcal{H}$ if $\inf_{h \in \mathcal{H}} \beta_h =: \beta_0 > 0$.

Definition 53.15 (Stable/unstable pair). *We say that a pair of finite elements used to approximate the velocity and the pressure is* stable *if the inf-sup condition (53.15) holds true uniformly w.r.t. $h \in \mathcal{H}$, and we say that it is* unstable *otherwise.*

Remark 53.16 (Inf-sup condition in $\boldsymbol{W}^{1,p}$-$L^{p'}$). Let $p \in (1, \infty)$ and let $p' \in (1, \infty)$ be s.t. $\frac{1}{p} + \frac{1}{p'} = 1$. As in Remark 53.10, a more general variant of the inf-sup condition (53.15) is

$$\inf_{q_h \in Q_h} \sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h\mathrm{d}}} \frac{|\int_D q_h \nabla \cdot \boldsymbol{v}_h \, \mathrm{d}x|}{\|q_h\|_{L^{p'}(D)} |\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(D)}} =: \beta_h > 0. \tag{53.16}$$

We will see in the next chapters that many stable finite element pairs for the Stokes equations satisfy this more general inf-sup condition. □

Let us define the discrete operator $B_h : \boldsymbol{V}_{h\mathrm{d}} \to Q_h'$ s.t. $\langle B_h(\boldsymbol{v}_h), q_h \rangle_{Q_h', Q_h} := b(\boldsymbol{v}_h, q_h) = -\int_D q_h \nabla \cdot \boldsymbol{v}_h \, \mathrm{d}x$ for all $(\boldsymbol{v}_h, q_h) \in \boldsymbol{V}_{h\mathrm{d}} \times Q_h$. We have

$$(53.15) \iff B_h \text{ is surjective}, \tag{53.17a}$$
$$\ker(B_h) = \{ \boldsymbol{v}_h \in \boldsymbol{V}_{h\mathrm{d}} \mid (q_h, \nabla \cdot \boldsymbol{v}_h)_{L^2(D)} = 0, \ \forall q_h \in Q_h \}. \tag{53.17b}$$

The operator $B_h : \boldsymbol{V}_{h\mathrm{d}} \to Q_h'$ is the discrete counterpart of the divergence operator $B : \boldsymbol{V}_{\mathrm{d}} \to Q'$ introduced just above Theorem 53.11. We observe that the inf-sup condition (53.15) is equivalent to asserting the surjectivity of B_h. Moreover, assuming for simplicity that $g = 0$ in the mass conservation equation, the discrete Stokes problem (53.14) produces a velocity field $\boldsymbol{u}_h \in \ker(B_h)$.

One then says that the discrete velocity field is weakly divergence-free. However, $\ker(B_h)$ may not be a subspace of $\ker(B)$, i.e., the discrete velocity field $\boldsymbol{u}_h$ is not necessarily strongly (or pointwise) divergence-free.

Several techniques are available to prove the inf-sup condition (53.15), and we refer the reader to the next two chapters for various examples. Recall in particular that (53.15) is equivalent to the existence of a Fortin operator $\boldsymbol{\Pi}_h \in \mathcal{L}(\boldsymbol{V}_{\mathrm{d}}; \boldsymbol{V}_{h\mathrm{d}})$ s.t. $b(\boldsymbol{\Pi}_h(\boldsymbol{v}) - \boldsymbol{v}, q_h) = 0$ for all $q_h \in Q_h$ (see Lemma 26.9).

Theorem 53.17 (Error estimate). *Let $(\boldsymbol{u}, p)$ solve (53.6). Assume (53.15) and let $(\boldsymbol{u}_h, p_h)$ solve (53.14). Then we have*

$$|\boldsymbol{u} - \boldsymbol{u}_h|_{\boldsymbol{H}^1(D)} \leq c_{1h} \inf_{\boldsymbol{v}_h \in \boldsymbol{V}_{h\mathrm{d}}} |\boldsymbol{u} - \boldsymbol{v}_h|_{\boldsymbol{H}^1(D)} + c_{2h} \inf_{q_h \in Q_h} \|p - q_h\|_{L^2(D)},$$

$$\|p - p_h\|_{L^2(D)} \leq c_{3h} \inf_{\boldsymbol{v}_h \in \boldsymbol{V}_{h\mathrm{d}}} |\boldsymbol{u} - \boldsymbol{v}_h|_{\boldsymbol{H}^1(D)} + c_{4h} \inf_{q_h \in Q_h} \|p - q_h\|_{L^2(D)},$$

where $c_{1h} := (1 + \frac{\|a\|}{\alpha})(1 + \|\boldsymbol{\Pi}_h\|_{\mathcal{L}(\boldsymbol{V}_{\mathrm{d}}; \boldsymbol{V}_{h\mathrm{d}})})$ for any Fortin operator $\boldsymbol{\Pi}_h \in \mathcal{L}(\boldsymbol{V}_{\mathrm{d}}; \boldsymbol{V}_{h\mathrm{d}})$, $c_{2h} := 0$ if $\ker(B_h) \subset \ker(B)$ and $c_{2h} := \frac{\|b\|}{\alpha}$ otherwise, $c_{3h} := c_{1h} \frac{\|a\|}{\beta_h}$, and $c_{4h} := 1 + \frac{\|b\|}{\beta_h} + c_{2h} \frac{\|a\|}{\beta_h}$. Here, $\alpha \geq 2\mu C_{\mathrm{K}}^2$ is the coercivity constant of the bilinear form a on $\boldsymbol{V}_{\mathrm{d}} \times \boldsymbol{V}_{\mathrm{d}}$, $\|a\| \leq 2\mu$ its norm, and $\|b\| \leq 1$ the norm of the bilinear form b on $\boldsymbol{V}_{\mathrm{d}} \times Q$.

Proof. This is a direct application of Corollary 50.5. $\square$

Remark 53.18 (β_h vs. β_0). The estimates from Theorem 53.17 show that it is important that the inf-sup condition (53.15) be satisfied uniformly w.r.t. $h \in \mathcal{H}$. Indeed, the factor $\frac{1}{\beta_h}$ appears in the coefficients c_{3h} and c_{4h} in the pressure error bound, and a factor $\frac{1}{\beta_h}$ may appear in the constant c_{1h} affecting both error bounds if $\|\boldsymbol{\Pi}_h\|_{\mathcal{L}(\boldsymbol{V}_{\mathrm{d}}; \boldsymbol{V}_{h\mathrm{d}})} \sim \frac{\|b\|}{\beta_h}$ for every Fortin operator. $\square$

We say that the pair $(\boldsymbol{\xi}(\boldsymbol{r}), \phi(\boldsymbol{r})) \in \boldsymbol{V}_{\mathrm{d}} \times Q$ is the solution to the adjoint problem of (53.6) with source term $\boldsymbol{r} \in \boldsymbol{L}^2(D)$ if $a(\boldsymbol{v}, \boldsymbol{\xi}(\boldsymbol{r})) + b(\boldsymbol{v}, \phi(\boldsymbol{r})) = \int_D \boldsymbol{r} \cdot \boldsymbol{v} \, \mathrm{d}x$ for all $\boldsymbol{v} \in \boldsymbol{V}_{\mathrm{d}}$ and $b(\boldsymbol{\xi}(\boldsymbol{r}), q) = 0$ for all $q \in Q$.

Theorem 53.19 ($\boldsymbol{L}^2$-velocity error estimate). *Let $(\boldsymbol{u}, p)$ solve (53.6). Assume (53.15) and let $(\boldsymbol{u}_h, p_h)$ solve (53.14). Assume that there exist real numbers c_{smo} and $s \in (0, 1]$ s.t.*

$$\mu \ell_D^{-1} \|\boldsymbol{\xi}(\boldsymbol{r})\|_{\boldsymbol{H}^{1+s}(D)} + \|\phi(\boldsymbol{r})\|_{H^s(D)} \leq c_{\mathrm{smo}} \ell_D \|\boldsymbol{r}\|_{\boldsymbol{L}^2(D)}, \qquad \forall \boldsymbol{r} \in \boldsymbol{L}^2(D),$$

and that there is c such that for all $h \in \mathcal{H}$, $\inf_{\boldsymbol{v}_h \in \boldsymbol{V}_{h\mathrm{d}}} |\boldsymbol{v} - \boldsymbol{v}_h|_{\boldsymbol{H}^1(D)} \leq ch^s |\boldsymbol{v}|_{\boldsymbol{H}^{1+s}(D)}$ for all $\boldsymbol{v} \in \boldsymbol{V}_{\mathrm{d}} \cap \boldsymbol{H}^{1+s}(D)$ and $\inf_{q_h \in Q_h} \|q - q_h\|_{L^2(D)} \leq ch^s |q|_{H^s(D)}$ for all $q \in Q \cap H^s(D)$. Then there is c s.t. for all $h \in \mathcal{H}$,

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{L}^2(D)} \leq c\, h^s \ell_D^{1-s} \left(\inf_{\boldsymbol{v}_h \in \boldsymbol{V}_{h\mathrm{d}}} |\boldsymbol{u} - \boldsymbol{v}_h|_{\boldsymbol{H}^1(D)} + \frac{\|b\|}{\|a\|} \inf_{q_h \in Q_h} \|p - q_h\|_{L^2(D)} \right).$$

Proof. Apply Lemma 50.11 or see Exercise 53.3. $\square$

Let us give some further insight into the velocity error estimate from Theorem 53.17. For simplicity, we assume that $g = 0$. Let us define the projection operator $\boldsymbol{P}_h^{\mathrm{s}} : \boldsymbol{V}_{\mathrm{d}} \to \ker(B_h)$ such that

$$a(\boldsymbol{P}_h^{\mathrm{s}}(\boldsymbol{v}), \boldsymbol{w}_h) = a(\boldsymbol{v}, \boldsymbol{w}_h), \qquad \forall (\boldsymbol{v}, \boldsymbol{w}_h) \in \boldsymbol{V}_{\mathrm{d}} \times \ker(B_h). \tag{53.18}$$

Lemma 53.20 (Quasi-optimality). *Assume* (53.15). *The following holds true for all* $\boldsymbol{v} \in \boldsymbol{V}_{\mathrm{d}}$ *and any Fortin operator* $\boldsymbol{\Pi}_h \in \mathcal{L}(\boldsymbol{V}_{\mathrm{d}}; \boldsymbol{V}_{h\mathrm{d}})$:

$$|\boldsymbol{v} - \boldsymbol{P}_h^{\mathrm{S}}(\boldsymbol{v})|_{\boldsymbol{H}^1(D)} \leq \tilde{c}_{1h} \inf_{\boldsymbol{v}_h \in \boldsymbol{V}_{h\mathrm{d}}} |\boldsymbol{v} - \boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}, \qquad (53.19)$$

with $\tilde{c}_{1h} := \frac{\|a\|}{\alpha}(1 + \|\boldsymbol{\Pi}_h\|_{\mathcal{L}(\boldsymbol{V}_{\mathrm{d}}; \boldsymbol{V}_{h\mathrm{d}})})$.

Proof. Since the bilinear form a is bounded and coercive, we have

$$|\boldsymbol{v} - \boldsymbol{P}_h^{\mathrm{S}}(\boldsymbol{v})|_{\boldsymbol{H}^1(D)} \leq \frac{\|a\|}{\alpha} \inf_{\boldsymbol{v}_h \in \ker(B_h)} |\boldsymbol{v} - \boldsymbol{v}_h|_{\boldsymbol{H}^1(D)}.$$

The assertion then follows from Lemma 50.3 (notice that $\boldsymbol{\Pi}_h(\boldsymbol{u}) \in \ker(B_h)$ since $\nabla \cdot \boldsymbol{u} = g = 0$ by assumption, see Remark 50.4). $\qquad\square$

Lemma 53.21 (Discrete velocity estimate). *Let* $(\boldsymbol{u}, p)$ *solve* (53.6). *Assume* (53.15) *and let* $(\boldsymbol{u}_h, p_h)$ *solve* (53.14). *As in Theorem 53.17, set* $c_{2h} := 0$ *if* $\ker(B_h) \subset \ker(B)$ *and* $c_{2h} := \frac{\|b\|}{\alpha}$ *otherwise. The following holds true:*

$$|\boldsymbol{u}_h - \boldsymbol{P}_h^{\mathrm{S}}(\boldsymbol{u})|_{\boldsymbol{H}^1(D)} \leq c_{2h} \inf_{q_h \in Q_h} \|p - q_h\|_{L^2(D)}. \qquad (53.20)$$

Proof. The proof follows a similar, yet simpler, path to that of Lemma 50.2. Since $a(\boldsymbol{u}_h, \boldsymbol{w}_h) + b(\boldsymbol{w}_h, p_h) = F(\boldsymbol{w}_h) = a(\boldsymbol{u}, \boldsymbol{w}_h) + b(\boldsymbol{w}_h, p) = a(\boldsymbol{P}_h^{\mathrm{S}}(\boldsymbol{u}), \boldsymbol{w}_h) + b(\boldsymbol{w}_h, p)$ for all $\boldsymbol{w}_h \in \ker(B_h) \subset \boldsymbol{V}_{h\mathrm{d}} \subset \boldsymbol{V}_{\mathrm{d}}$, setting $\boldsymbol{e}_h := \boldsymbol{u}_h - \boldsymbol{P}_h^{\mathrm{S}}(\boldsymbol{u}) \in \ker(B_h)$, we infer that $a(\boldsymbol{e}_h, \boldsymbol{w}_h) = b(\boldsymbol{w}_h, p - p_h)$ for all $\boldsymbol{w}_h \in \ker(B_h)$. Since $\boldsymbol{e}_h \in \ker(B_h)$, invoking the coercivity of a then yields

$$\alpha \, |\boldsymbol{e}_h|_{\boldsymbol{H}^1(D)}^2 \leq b(\boldsymbol{e}_h, p - p_h).$$

If $\ker(B_h) \subset \ker(B)$, then $|\boldsymbol{e}_h|_{\boldsymbol{H}^1(D)} = 0$ which proves (53.20). Otherwise, we use that $\boldsymbol{e}_h \in \ker(B_h)$ to write $\alpha|\boldsymbol{e}_h|_{\boldsymbol{H}^1(D)}^2 \leq b(\boldsymbol{e}_h, p - q_h)$ for all $q_h \in Q_h$, and invoke the boundedness of b to prove (53.20). $\qquad\square$

The bound (53.20) implies that $\boldsymbol{u}_h = \boldsymbol{P}_h^{\mathrm{S}}(\boldsymbol{u})$ whenever $\ker(B_h) \subset \ker(B)$. Moreover, in the general case, combining (53.20) with (53.19) and using the triangle inequality we obtain again the velocity error estimate from Theorem 53.17 with the slightly sharper constant $\tilde{c}_{1h}$ instead of c_{1h}.

Remark 53.22 (Well-balanced scheme). In the particular case where $\boldsymbol{f} = \nabla \phi$ for some $\phi \in H^1(D) \cap L_*^2(D)$, the solution to the Stokes problem (53.6) is $(\boldsymbol{u}, p) = (\boldsymbol{0}, \phi)$. This situation is encountered with hydrostatic (or curl-free) body forces. One says that the discrete problem (53.14) is *well-balanced* w.r.t. hydrostatic body forces if $\boldsymbol{u}_h = \boldsymbol{0}$ as well. (One also sometimes says that the discretization is *pressure robust*.) A well-balanced discretization of the Stokes equations can be desirable even if $\boldsymbol{f}$ is not curl-free, but has a relatively large curl-free component. In this case, a discretization that is not well-balanced can lead to a rather poor velocity approximation, even on meshes that seem rather fine. Lemma 53.21 shows that (53.14) is well-balanced whenever $\ker(B_h) \subset \ker(B)$. The scheme can be made well-balanced when $\ker(B_h) \not\subset \ker(B)$ by slightly modifying the discrete momentum equation. Considering Dirichlet conditions over the whole boundary for simplicity, one introduces a lifting operator $L : \boldsymbol{V}_{h\mathrm{d}} \to \boldsymbol{V}_{\mathrm{d}}$ such that $L(\ker(B_h)) \subset \ker(B)$ and then replaces the first equation in (53.14) by $a(\boldsymbol{u}_h, \boldsymbol{w}_h) + b(\boldsymbol{w}_h, p_h) = (\boldsymbol{f}, L(\boldsymbol{w}_h))_{\boldsymbol{L}^2(D)}$ for all $\boldsymbol{w}_h \in \boldsymbol{V}_{h\mathrm{d}}$. The lifting operator L must satisfy some consistency conditions to preserve the optimal decay rates of the error estimate. This idea has been introduced by Linke [283] and explored more thoroughly by Lederer et al. [279] in the context of mixed finite elements with continuous pressures; see also

John et al. [261] for an overview. Examples of curl-free body forces in fluid mechanics are the Coriolis force if $d = 2$, the gravity, and the centrifugal force. Obviously, if $\boldsymbol{f} \approx \nabla\phi$ and ϕ is explicitly known, one can always make the change of variable $p \to p - \phi$ to alleviate the above difficulty if the scheme is not well-balanced. $\square$

53.4   Classical examples of unstable pairs

We study in this section three pairs of finite elements that look appealing at first sight, but that unfortunately do not satisfy the inf-sup condition (53.15). For simplicity, we consider a homogeneous Dirichlet condition on the velocity over the whole boundary, so that $\boldsymbol{V}_{\mathrm{d}} := \boldsymbol{H}_0^1(D)$ and we write $\boldsymbol{V}_{h0}$ instead of $\boldsymbol{V}_{h\mathrm{d}}$ for the discrete velocity space. Since the approximation setting is conforming, we have $\boldsymbol{V}_{h0} \subset \boldsymbol{H}_0^1(D)$ in all cases.

Recall that the inf-sup condition (53.15) is not satisfied if and only if $B_h^* : Q_h \to \boldsymbol{V}_{h0}'$ is not injective (or, once global shape functions have been chosen, the associated matrix does not have full column rank). In this case, a nonzero pressure field in $\ker(B_h^*)$ is called *spurious pressure mode*. Equivalently, the inf-sup condition is not satisfied if and only if $B_h : \boldsymbol{V}_{h0} \to Q_h'$ is not surjective.

53.4.1   The $(\mathbb{Q}_1, \mathbb{P}_0)$ pair: Checkerboard instability

A well-known pair of incompatible finite elements is the $(\mathbb{Q}_1, \mathbb{P}_0)$ pair obtained when approximating the velocity with continuous piecewise bilinear polynomials and the pressure with piecewise constants. This pair produces an instability often called *checkerboard instability*.

Let us restrict ourselves to the two-dimensional setting and assume that $D := (0, 1)^2$. We define a uniform Cartesian mesh on D as follows: Let N be an integer larger than 2. Set $h := \frac{1}{I}$, and for all $i, j \in \{0 : I-1\}$, denote by $\boldsymbol{a}_{ij}$ the point with Cartesian coordinates (ih, jh). Let K_{ij} be the square cell whose bottom left node is $\boldsymbol{a}_{ij}$; see Figure 53.1. The resulting mesh is denoted by $\mathcal{T}_h := \bigcup_{i,j} K_{ij}$. Consider the following finite element spaces:

$$\boldsymbol{V}_{h0} := \{\boldsymbol{v}_h \in \boldsymbol{C}^0(\overline{D}) \mid \forall K_{ij} \in \mathcal{T}_h,\, \boldsymbol{v}_h \circ \boldsymbol{T}_{K_{ij}} \in \mathbb{Q}_{1,d},\, \boldsymbol{v}_{h|\partial D} = \boldsymbol{0}\}, \tag{53.21a}$$

$$Q_h := \{q_h \in L_*^2(D) \mid \forall K_{ij} \in \mathcal{T}_h,\, q_h \circ \boldsymbol{T}_{K_{ij}} \in \mathbb{P}_{0,d}\}. \tag{53.21b}$$

Recall that for all $K \in \mathcal{T}_h$, $\boldsymbol{T}_K : \widehat{K} \to K$ denotes the geometric mapping; see §8.1. For all $p_h \in Q_h$, set $p_{i+\frac{1}{2}, j+\frac{1}{2}} := p_{h|K_{ij}}$, and for all $\boldsymbol{v}_h \in \boldsymbol{V}_{h0}$, denote by (u_{ij}, v_{ij}) the values of the two Cartesian components of $\boldsymbol{v}_h$ at the node $\boldsymbol{a}_{ij}$.

To prove that the inf-sup constant is zero, it is sufficient to prove the existence of a nonzero pressure field $p_h \in \ker(B_h^*)$, i.e., $\int_D p_h \nabla{\cdot}\boldsymbol{v}_h \,\mathrm{d}x = 0$ for all $\boldsymbol{v}_h \in \boldsymbol{V}_{h0}$. Since p_h is constant on each cell, we have

$$\int_{K_{ij}} p_h \nabla{\cdot}\boldsymbol{v}_h \,\mathrm{d}x = p_{i+\frac{1}{2}, j+\frac{1}{2}} \int_{\partial K_{ij}} \boldsymbol{v}_h{\cdot}\boldsymbol{n} \,\mathrm{d}s$$

$$= \tfrac{1}{2}h p_{i+\frac{1}{2}, j+\frac{1}{2}} \left(u_{i+1,j} + u_{i+1,j+1} + v_{i+1,j+1} + v_{i,j+1}\right.$$

$$\left. -u_{i,j} - u_{i,j+1} - v_{i,j} - v_{i+1,j}\right).$$

Summing over all the cells and rearranging the sum yields

$$\int_D p_h \nabla{\cdot}\boldsymbol{v}_h \,\mathrm{d}x = -h^2 \sum_{i,j\in\{0\,:\,N-1\}} (u_{i,j} G_{1,ij}(p_h) + v_{i,j} G_{2,ij}(p_h)),$$
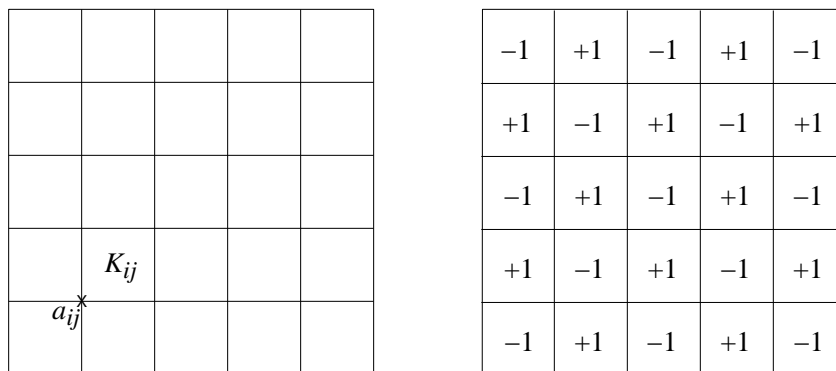
Figure 53.1: $(\mathbb{Q}_1, \mathbb{P}_0)$ pair: mesh (left) and spurious pressure mode (right).

where

$$G_{1,ij}(p_h) := \tfrac{1}{2h}(p_{i+\frac{1}{2},j+\frac{1}{2}} + p_{i+\frac{1}{2},j-\frac{1}{2}} - p_{i-\frac{1}{2},j+\frac{1}{2}} - p_{i-\frac{1}{2},j-\frac{1}{2}}),$$
$$G_{2,ij}(p_h) := \tfrac{1}{2h}(p_{i+\frac{1}{2},j+\frac{1}{2}} + p_{i-\frac{1}{2},j+\frac{1}{2}} - p_{i+\frac{1}{2},j-\frac{1}{2}} - p_{i-\frac{1}{2},j-\frac{1}{2}}).$$

We infer that $\int_D p_h \nabla \cdot \boldsymbol{v}_h \, \mathrm{d}x = 0$ for all $\boldsymbol{v}_h \in \boldsymbol{V}_{h0}$ if and only if

$$p_{i+\frac{1}{2},j+\frac{1}{2}} = p_{i-\frac{1}{2},j-\frac{1}{2}} \qquad \text{and} \qquad p_{i-\frac{1}{2},j+\frac{1}{2}} = p_{i+\frac{1}{2},j-\frac{1}{2}}.$$

The solution set of this linear system is a two-dimensional vector space. One dimension is spanned by the constant field $p_h = 1$, but span$\{1\}$ must be excluded from the solution set since the elements in Q_h must have a zero mean. The other dimension is spanned by the field whose value is alternatively $+1$ and -1 on adjacent cells in a checkerboard pattern, as shown on the right panel of Figure 53.1. This is a spurious pressure mode, and if N is even, this spurious mode is in Q_h (i.e., it satisfies the zero-mean condition). In this case, the inf-sup condition is not satisfied, i.e., the $(\mathbb{Q}_1, \mathbb{P}_0)$ pair is incompatible for the Stokes problem.

Remark 53.23 (Filtering). Since the $(\mathbb{Q}_1, \mathbb{P}_0)$ pair is very simple to program, one may be tempted to cure its deficiencies by restricting the size of Q_h. For instance, one could enforce the pressure to be orthogonal (in the L^2-sense) to the space spanned by the spurious pressure mode. Unfortunately, this remedy is not strong enough to produce a healthy finite element pair, since it can be shown that in this case there are positive constants c, c' s.t. $ch \leq \beta_h \leq c'h$ uniformly w.r.t. $h \in \mathcal{H}$; see Boland and Nicolaides [68] or Girault and Raviart [217, p. 164]. This shows that the method may not converge since the factor $\frac{1}{\beta_h}$ appears in the error bound on the velocity and the factor $\frac{1}{\beta_h^2}$ appears in the error bound on the pressure (see Theorem 53.17). $\qquad \square$

53.4.2 The $(\mathbb{P}_1, \mathbb{P}_1)$ pair: Checkerboard-like instability

Because it is very simple to program, the continuous $\mathbb{P}_1$ finite element for both the velocity and the pressure is a natural choice for approximating the Stokes problem. Unfortunately, the $(\mathbb{P}_1, \mathbb{P}_1)$ pair does not satisfy the inf-sup condition (53.15). To understand the origin of the problem, let us construct a two-dimensional counterexample in $D := (0,1)^2$. Consider a uniform Cartesian mesh composed of squares of side h and split each square along one diagonal as shown in the left panel of Figure 53.2. Let $\mathcal{T}_h$ be the resulting triangulation and let the velocity and the pressure finite

element spaces be

$$\boldsymbol{V}_{h0} := \boldsymbol{P}_{1,0}^{\mathrm{g}}(\mathcal{T}_h), \qquad Q_h := P_1^{\mathrm{g}}(\mathcal{T}_h) \cap L_*^2(D), \tag{53.22}$$

where $\boldsymbol{P}_{1,0}^{\mathrm{g}}(\mathcal{T}_h)$ is the vector-valued version of the space $P_{1,0}^{\mathrm{g}}(\mathcal{T}_h)$ defined in §19.4. Let $\{\boldsymbol{z}_{n,K}\}_{n \in \{0:2\}}$

Figure 53.2: $(\mathbb{P}_1, \mathbb{P}_1)$ pair: the mesh (left) and one spurious pressure mode (right).

be the three vertices of the mesh cell $K \in \mathcal{T}_h$. Now consider a pressure field p_h such that $\sum_{n \in \{0:2\}} p_h(\boldsymbol{z}_{n,K})$ is zero on each triangle K. An example of such a *spurious pressure mode* is shown in the right panel of Figure 53.2. Then we have for all $\boldsymbol{v}_h \in \boldsymbol{V}_{h0}$,

$$\int_D p_h \nabla \cdot \boldsymbol{v}_h \, \mathrm{d}x = \sum_{K \in \mathcal{T}_h} (\nabla \cdot \boldsymbol{v}_h)_{|K} \int_K p_h \, \mathrm{d}x,$$

$$= \sum_{K \in \mathcal{T}_h} (\nabla \cdot \boldsymbol{v}_h)_{|K} \frac{|K|}{3} \sum_{n \in \{0:2\}} p_h(\boldsymbol{z}_{n,K}) = 0.$$

Hence, p_h satisfies $\int_D p_h \nabla \cdot \boldsymbol{v}_h \, \mathrm{d}x = 0$ for all $\boldsymbol{v}_h \in \boldsymbol{V}_{h0}$. In other words, the field p_h is a spurious pressure mode, and the inf-sup constant is zero.

53.4.3 The $(\mathbb{P}_1, \mathbb{P}_0)$ pair: Locking effect

A simple alternative to the $(\mathbb{Q}_1, \mathbb{P}_0)$ pair consists of using the $(\mathbb{P}_1, \mathbb{P}_0)$ pair, i.e., assuming that $\mathcal{T}_h$ is composed of simplices, the velocity is approximated with continuous, piecewise linear polynomials and the pressure with piecewise constants. We observe that $\ker(B_h) \subset \ker(B)$ in this case since the divergence of the velocity is piecewise constant. Unfortunately, the $(\mathbb{P}_1, \mathbb{P}_0)$ pair does not satisfy the inf-sup condition (53.15). Let us produce a two-dimensional counterexample. Assume that D is a simply connected polygon. Let N_{c}, $N_{\mathrm{v}}^{\mathrm{i}}$, and $N_{\mathrm{e}}^{\partial}$ denote the number of elements, internal vertices, and boundary edges in $\mathcal{T}_h$, respectively. The Euler relations give $N_{\mathrm{c}} = 2N_{\mathrm{v}}^{\mathrm{i}} + N_{\mathrm{e}}^{\partial} - 2$ (see Remark 8.13 and Exercise 8.2). Since $\dim(Q_h) = N_{\mathrm{c}} - 1$ and $\dim(\boldsymbol{V}_{h0}) = 2N_{\mathrm{v}}^{\mathrm{i}}$, the rank nullity theorem implies that

$$\dim(\ker(B_h^*)) = \dim(Q_h) - \dim(\mathrm{im}(B_h^*)) \geq \dim(Q_h) - \dim(\boldsymbol{V}_{h0})$$

$$= N_{\mathrm{c}} - 1 - 2N_{\mathrm{v}}^{\mathrm{i}} = N_{\mathrm{e}}^{\partial} - 3.$$

Hence, there are at least $N_{\mathrm{e}}^{\partial} - 3$ spurious pressure modes. This means that the space Q_h is far too rich for B_h to be surjective. Actually, in some cases, it can be shown that B_h is injective, i.e., the only member of $\ker(B_h)$ is zero. This situation is referred to as *locking* in the literature.

Remark 53.24 (Comparison with $(\mathbb{P}_1, \mathbb{P}_1)$). Note that the dimension of the pressure finite element space is smaller for the $(\mathbb{P}_1, \mathbb{P}_1)$ pair (where $\dim(Q_h) = N_v - 1$) than for the $(\mathbb{P}_1, \mathbb{P}_0)$ pair (where $\dim(Q_h) = N_c - 1$). Indeed, we have $N_c \sim 2N_v$ on fine meshes (see Exercise 8.2). $\qquad\square$

Exercises

Exercise 53.1 ($\nabla\cdot$ is surjective). Let $D \subset \mathbb{R}^2$ be a domain of class C^2. Prove that $\nabla\cdot : \boldsymbol{H}_0^1(D) \to L_*^2(D)$ is continuous and surjective. (*Hint*: construct $\boldsymbol{v} \in \boldsymbol{H}_0^1(D)$ such that $\boldsymbol{v} = \nabla q + \nabla\times\psi$, where q solves a Poisson problem, ψ solves a biharmonic problem, and $\nabla\times\psi := (\partial_2\psi, -\partial_1\psi)^\mathsf{T}$.)

Exercise 53.2 (de Rham). Let D be a bounded open set in $\mathbb{R}^d$ and assume that D is star-shaped with respect to an open ball $B \subset D$. Prove that the continuous linear forms on $\boldsymbol{W}_0^{1,p}(D)$ that are zero on $\ker(\nabla\cdot)$ are gradients of functions in $L_*^{p'}(D)$. (*Hint*: use Remark 53.10 and the closed range theorem.)

Exercise 53.3 (L^2-estimate). Prove Theorem 53.19 directly, i.e., without invoking Lemma 50.11.

Exercise 53.4 (Projection). Let $(\boldsymbol{V}_{h0}, Q_h)_{h\in\mathcal{H}}$ be a sequence of pairs of finite element spaces. Let $p \in [1, \infty]$ and let $p' \in [1, \infty]$ be s.t. $\frac{1}{p} + \frac{1}{p'} = 1$. Let $\Pi_h^Z : Q_h \to Z_h$ be an operator, where Z_h is a finite-dimensional subspace of $L^p(D)$. Assume that there are $\beta_1, \beta_2 > 0$ such that for all $h \in \mathcal{H}$, $\sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}} \frac{|\int_D q_h \nabla\cdot\boldsymbol{v}_h \, \mathrm{d}x|}{|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(D)}} \geq \beta_1 \|q_h - \Pi_h^Z(q_h)\|_{L^{p'}(D)}$ for all $q_h \in Q_h$ and $\sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}} \frac{|\int_D q_h \nabla\cdot\boldsymbol{v}_h \, \mathrm{d}x|}{|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(D)}} \geq \beta_2 \|q_h\|_{L^{p'}(D)}$ for all $q_h \in Z_h$. (i) Show that Π_h^Z is bounded uniformly w.r.t. $h \in \mathcal{H}$. (ii) Show that the $(\boldsymbol{V}_{h0}, Q_h)$ pair satisfies an inf-sup condition uniformly w.r.t. $h \in \mathcal{H}$.

Exercise 53.5 (Spurious mode for the $(\mathbb{Q}_1, \mathbb{Q}_1)$ pair). (i) Let $\widehat{K} := [0,1]^2$ be the unit square. Let $\widehat{\boldsymbol{a}}_{ij} := (\frac{i}{2}, \frac{j}{2})$, for all $i, j \in \{0{:}2\}$. Show that the quadrature $\int_{\widehat{K}} f(\widehat{\boldsymbol{x}}) \, \mathrm{d}\widehat{x} \approx \sum_{i,j} w_{ij} f(\widehat{\boldsymbol{a}}_{ij})$, where $w_{ij} := \frac{1}{36}(3i(2-i)+1)(3j(j-2)+1)$ ($w_{ij} := \frac{1}{36}$ for the four vertices of $\widehat{K}$, $w_{ij} := \frac{1}{9}$ for the four edge midpoints, and $w_{ij} := \frac{4}{9}$ at the barycenter of $\widehat{K}$) is exact for all $f \in \mathbb{Q}_2$. (*Hint*: write the $\mathbb{Q}_2$ Lagrange shape functions in tensor-product form and use Simpson's rule in each direction.) (ii) Consider $D := (0,1)^2$ and a mesh composed of $I{\times}I$ squares, $I \geq 2$. Consider the points $\boldsymbol{a}_{lm} := (\frac{l}{2I}, \frac{m}{2I})$ for all $l, m \in \{0{:}2I\}$. Let p_h be the continuous, piecewise bilinear function such that $p_h(\boldsymbol{a}_{2k,2n}) := (-1)^{k+n}$ for all $k, n \in \{0{:}I\}$. Show that p_h is a spurious pressure mode for the $(\mathbb{Q}_1, \mathbb{Q}_1)$ pair (continuous velocity and pressure).

Chapter 54

Stokes equations: Stable pairs (I)

This chapter reviews various stable finite element pairs that are suitable to approximate the Stokes equations, i.e., the discrete velocity space and the discrete pressure space satisfy the inf-sup condition (53.15) (or its $\boldsymbol{W}^{1,p}$-$L^{p'}$ version (53.16)) uniformly with respect to $h \in \mathcal{H}$. We first review two standard techniques to prove the inf-sup condition, one based on the Fortin operator and one hinging on a weak control of the pressure gradient. Then we show how these techniques can be applied to finite element pairs where the discrete pressure space is H^1-conforming. The two main examples are the mini element based on the $(\mathbb{P}_1\text{-bubble}, \mathbb{P}_1)$ pair and the Taylor–Hood element based on the $(\mathbb{P}_2, \mathbb{P}_1)$ pair. In the next chapter, we introduce another technique based on macroelements to prove the inf-sup condition and we review stable finite element pairs where the discrete pressures are discontinuous. We assume in the entire chapter that Dirichlet conditions are enforced on the velocity over the whole boundary, that D is a polyhedron in $\mathbb{R}^d$, and that $(\mathcal{T}_h)_{h \in \mathcal{H}}$ is a shape-regular sequence of affine meshes so that each mesh covers D exactly.

54.1 Proving the inf-sup condition

We briefly review two standard techniques to prove the inf-sup condition (53.15): one uses a Fortin operator and the other uses a weak control on the pressure gradient. Since this section is only meant to be a short introduction, the reader is referred to Boffi et al. [65, Chap. 8], Girault and Raviart [217, §II.1.4] for thorough reviews of the topic.

54.1.1 Fortin operator

One way to prove the inf-sup condition (53.15) consists of using the notion of Fortin operator. The theory behind the Fortin operator theory is investigated in detail §26.2.3. We now briefly summarize the main features of this theory and adapt the notation to the setting of the Stokes equations.

   Let $\boldsymbol{V}, Q$ be two complex Banach spaces and let b be a bounded sesquilinear form on $\boldsymbol{V} \times Q$. Let β and $\|b\|$ be the inf-sup and the boundedness constants of b. Let $\boldsymbol{V}_{h0} \subset \boldsymbol{V}$ and let $Q_h \subset Q$ be finite-dimensional subspaces equipped, respectively, with the norms of $\boldsymbol{V}$ and Q. A map $\boldsymbol{\Pi}_h : \boldsymbol{V} \to \boldsymbol{V}_{h0}$, is called a *Fortin operator* if $b(\boldsymbol{\Pi}_h(\boldsymbol{v}) - \boldsymbol{v}, q_h) = 0$ for all $(\boldsymbol{v}, q_h) \in \boldsymbol{V} \times Q_h$, and there is real number $\gamma_h > 0$ such that $\gamma_h \|\boldsymbol{\Pi}_h(\boldsymbol{v})\|_{\boldsymbol{V}} \leq \|\boldsymbol{v}\|_{\boldsymbol{V}}$ for all $\boldsymbol{v} \in \boldsymbol{V}$. The key result we are going to

use is the following statement (see Lemma 26.9, Boffi et al. [65, Prop. 8.4.1], and the work by the authors [187, Thm. 1]).

Lemma 54.1 (Fortin operator). *If there exists a Fortin operator, then the inf-sup condition*

$$\inf_{q_h \in Q_h} \sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}} \frac{|b(\boldsymbol{v}_h, q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}} \|q_h\|_Q} =: \beta_h > 0, \tag{54.1}$$

holds true with $\beta_h \geq \gamma_h \beta$. *Conversely, if the inf-sup condition* (54.1) *holds true, then there exists a Fortin operator with* $\gamma_h \geq \frac{\beta_h}{\|b\|}$.

Hence, proving the inf-sup condition (54.1) can be done by constructing a Fortin operator. A practical way to do this is given by the following result.

Lemma 54.2 (Decomposition). *Let* $\boldsymbol{\Pi}_{1h}, \boldsymbol{\Pi}_{2h} : \boldsymbol{V} \to \boldsymbol{V}_{h0}$ *be two operators. Assume the following:* (i) $\boldsymbol{\Pi}_{2h}$ *is linear;* (ii) $b(\boldsymbol{v} - \boldsymbol{\Pi}_{2h}(\boldsymbol{v}), q_h) = 0$ *for all* $(\boldsymbol{v}, q_h) \in \boldsymbol{V} \times Q_h$; (iii) *The real numbers*

$$c_{1h} := \sup_{\boldsymbol{v} \in \boldsymbol{V}} \frac{\|\boldsymbol{\Pi}_{1h}(\boldsymbol{v})\|_{\boldsymbol{V}}}{\|\boldsymbol{v}\|_{\boldsymbol{V}}} \quad \text{and} \quad c_{2h} := \sup_{\boldsymbol{v} \in \boldsymbol{V}} \frac{\|\boldsymbol{\Pi}_{2h}(\boldsymbol{v} - \boldsymbol{\Pi}_{1h}(\boldsymbol{v}))\|_{\boldsymbol{V}}}{\|\boldsymbol{v}\|_{\boldsymbol{V}}} \tag{54.2}$$

are finite. Then (recalling that $I_{\boldsymbol{V}} : \boldsymbol{V} \to \boldsymbol{V}$ *is the identity)*

$$\boldsymbol{\Pi}_h := \boldsymbol{\Pi}_{1h} + \boldsymbol{\Pi}_{2h}(I_{\boldsymbol{V}} - \boldsymbol{\Pi}_{1h}) \tag{54.3}$$

is a Fortin operator with $\gamma_h \geq (c_{1h} + c_{2h})^{-1}$.

Proof. Since the operator $\boldsymbol{\Pi}_{2h}$ is linear owing to the assumption (i), we have

$$b(\boldsymbol{v} - \boldsymbol{\Pi}_h(\boldsymbol{v}), q_h) = b(\boldsymbol{v} - \boldsymbol{\Pi}_{2h}(\boldsymbol{v}), q_h) - b(\boldsymbol{\Pi}_{1h}(\boldsymbol{v}) - \boldsymbol{\Pi}_{2h}(\boldsymbol{\Pi}_{1h}(\boldsymbol{v})), q_h),$$

for all $(\boldsymbol{v}, q_h) \in \boldsymbol{V} \times Q_h$, and both terms on the right-hand side are zero owing to the assumption (ii). Furthermore, we have $\sup_{\boldsymbol{v} \in \boldsymbol{V}} \frac{\|\boldsymbol{\Pi}_h(\boldsymbol{v})\|_{\boldsymbol{V}}}{\|\boldsymbol{v}\|_{\boldsymbol{V}}} \leq c_{1h} + c_{2h}$, i.e., $\gamma_h \|\boldsymbol{\Pi}_h(\boldsymbol{v})\|_{\boldsymbol{V}} \leq \|\boldsymbol{v}\|_{\boldsymbol{V}}$ for all $\boldsymbol{v} \in \boldsymbol{V}$ with $\gamma_h \geq (c_{1h} + c_{2h})^{-1} > 0$ owing to the assumption (iii). □

54.1.2 Weak control on the pressure gradient

A second possibility to prove the inf-sup condition (54.1) consists of establishing a weak control on the gradient of the pressure. This technique can be used when the discrete pressure space is H^1-conforming. Let us focus more specifically on the bilinear form $b(\boldsymbol{v}, q) := -(\nabla \cdot \boldsymbol{v}, q)_{L^2(D)}$. Let $p \in (1, \infty)$, $\boldsymbol{V} := \boldsymbol{W}_0^{1,p}(D)$ equipped with the norm $\|\boldsymbol{v}\|_{\boldsymbol{V}} := |\boldsymbol{v}|_{\boldsymbol{W}^{1,p}(D)}$, and $Q := L_*^{p'}(D) := \{q \in L^{p'}(D) \mid \int_D q \, dx = 0\}$ equipped with the norm $\|q\|_Q := \|q\|_{L^{p'}(D)}$ with $p' \in (1, \infty)$ s.t. $\frac{1}{p} + \frac{1}{p'} = 1$. The discrete velocity space is $\boldsymbol{V}_{h0} \subset \boldsymbol{V}$, and the discrete pressure space is $Q_h \subset Q$.

Lemma 54.3 (Pressure gradient control). *Assume that the discrete pressure space* Q_h *is* H^1-*conforming, and that there is* c *such that the following holds true for all* $p \in (1, \infty)$ *and all* $h \in \mathcal{H}$:

$$\sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}} \frac{|b(\boldsymbol{v}_h, q_h)|}{|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(D)}} \geq c \left(\sum_{K \in \mathcal{T}_h} h_K^{p'} \|\nabla q_h\|_{\boldsymbol{L}^{p'}(K)}^{p'} \right)^{\frac{1}{p'}}. \tag{54.4}$$

Then the inf-sup condition (54.1) *holds true uniformly w.r.t.* $h \in \mathcal{H}$.

Proof. Let $q_h \in Q_h$. Since $Q_h \subset Q$, the continuous inf-sup condition (53.9) implies that

$$\beta_D \|q_h\|_Q \leq \sup_{\boldsymbol{v} \in \boldsymbol{V}} \frac{|b(\boldsymbol{v}, q_h)|}{|\boldsymbol{v}|_{\boldsymbol{W}^{1,p}(D)}} \leq \sup_{\boldsymbol{v} \in \boldsymbol{V}} \frac{|b(\boldsymbol{\mathcal{I}}_h^{\mathrm{av}}(\boldsymbol{v}), q_h)|}{|\boldsymbol{v}|_{\boldsymbol{W}^{1,p}(D)}} + \sup_{\boldsymbol{v} \in \boldsymbol{V}} \frac{|b(\boldsymbol{v} - \boldsymbol{\mathcal{I}}_h^{\mathrm{av}}(\boldsymbol{v}), q_h)|}{|\boldsymbol{v}|_{\boldsymbol{W}^{1,p}(D)}},$$

where $\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}$ is the $\mathbb{R}^d$-valued version of the $W_0^{1,p}$-conforming quasi-interpolation operator introduced in §22.4.2. This means that $\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}(\boldsymbol{v}) := \sum_{i \in \{1:d\}} \mathcal{I}_{h0}^{\mathrm{av}}(v_i) \boldsymbol{e}_i$, where $\boldsymbol{v} := \sum_{i \in \{1:d\}} v_i \boldsymbol{e}_i$ and $\{\boldsymbol{e}_i\}_{i \in \{1:d\}}$ is the canonical Cartesian basis of $\mathbb{R}^d$. Let $\mathfrak{T}_1, \mathfrak{T}_2$ denote the two terms on the right-hand side. Owing to the $\boldsymbol{W}_0^{1,p}$-stability of $\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}$, we have $|\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}(\boldsymbol{v})|_{\boldsymbol{W}^{1,p}(D)} \leq c_{\mathcal{I}} |\boldsymbol{v}|_{\boldsymbol{W}^{1,p}(D)}$. Since $\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}(\boldsymbol{v}) \in \boldsymbol{V}_{h0}$, we infer that

$$|\mathfrak{T}_1| \leq c_{\mathcal{I}} \sup_{\boldsymbol{v} \in \boldsymbol{V}} \frac{|b(\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}(\boldsymbol{v}), q_h)|}{\|\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}(\boldsymbol{v})\|_{\boldsymbol{W}^{1,p}(D)}} \leq c_{\mathcal{I}} \sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}} \frac{|b(\boldsymbol{v}_h, q_h)|}{|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(D)}}.$$

Moreover, using that Q_h is H^1-conforming to integrate by parts, and then invoking Hölder's inequality and the approximation properties of $\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}$, we infer that

$$|b(\boldsymbol{v} - \boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}(\boldsymbol{v}), q_h)| = \left|(\nabla q_h, \boldsymbol{v} - \boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}(\boldsymbol{v}))_{\boldsymbol{L}^2(D)}\right|$$
$$\leq c \sum_{K \in \mathcal{T}_h} \|\nabla q_h\|_{\boldsymbol{L}^{p'}(K)} h_K \|\nabla \boldsymbol{v}\|_{\mathbb{L}^p(D_K)},$$

where D_K is the set of the points composing the mesh cells touching K. Since $\sum_{K \in \mathcal{T}_h} \|\nabla \boldsymbol{v}\|_{\mathbb{L}^p(D_K)}^p \leq c\|\nabla \boldsymbol{v}\|_{\mathbb{L}^p(D)}^p = c|\boldsymbol{v}|_{\boldsymbol{W}^{1,p}(D)}^p$ owing to the regularity of the mesh sequence, Hölder's inequality combined with the assumption (54.4) implies that

$$|\mathfrak{T}_2| \leq c' \left(\sum_{K \in \mathcal{T}_h} h_K^{p'} \|\nabla q_h\|_{\boldsymbol{L}^{p'}(K)}^{p'}\right)^{\frac{1}{p'}} \leq c'' \sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}} \frac{|b(\boldsymbol{v}_h, q_h)|}{|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(D)}}.$$

This completes the proof of the inf-sup condition. $\qquad\square$

Remark 54.4 (Literature). The technique presented above is based on Bercovier and Pironneau [54, Prop. 1], Verfürth [376] (for $p := 2$). $\qquad\square$

54.2 Mini element: the $(\mathbb{P}_1\text{-bubble}, \mathbb{P}_1)$ pair

Let $(\mathcal{T}_h)_{h \in \mathcal{H}}$ be a shape-regular sequence of affine simplicial meshes. Recall from §53.4.2 that the reason for which the $(\mathbb{P}_1, \mathbb{P}_1)$ pair does not satisfy the inf-sup condition (53.15) is that the velocity space is not rich enough (or equivalently the pressure space is too rich). To circumvent this difficulty, we are going to enlarge the velocity space by adding one more degree of freedom per simplex for each Cartesian component of the velocity.

Let $\widehat{K}$ be the reference simplex and $\widehat{\boldsymbol{x}}_{\widehat{K}}$ be its barycenter, and let $\widehat{b}$ be a function such that

$$\widehat{b} \in W_0^{1,\infty}(\widehat{K}), \qquad 0 \leq \widehat{b} \leq 1, \qquad \widehat{b}(\widehat{\boldsymbol{x}}_{\widehat{K}}) = 1. \tag{54.5}$$

One can use $\widehat{b}(\widehat{\boldsymbol{x}}) := (d+1)^{d+1} \prod_{i \in \{0:d\}} \widehat{\lambda}_i(\widehat{\boldsymbol{x}})$, where $\{\widehat{\lambda}_i\}_{i \in \{0:d\}}$ are the barycentric coordinates on $\widehat{K}$. This function is usually called *bubble function* in reference to the shape of its graph as shown in Figure 54.1. Another possibility consists of dividing the simplex $\widehat{K}$ into $(d+1)$ subsimplices
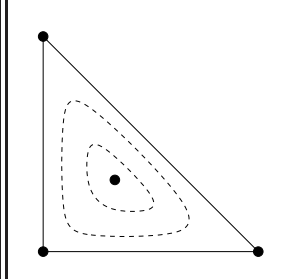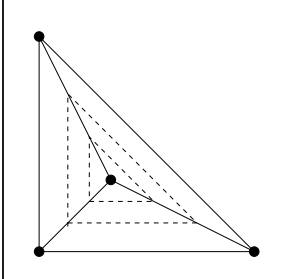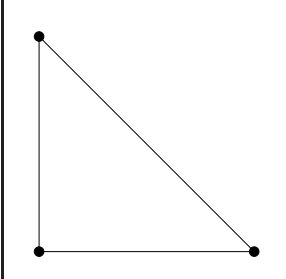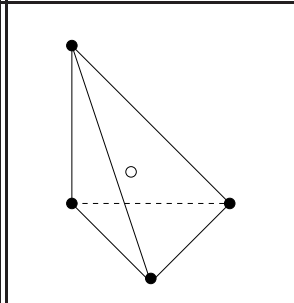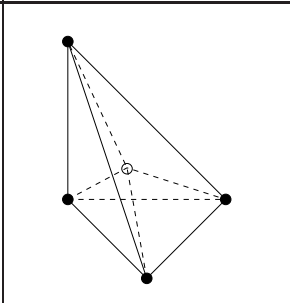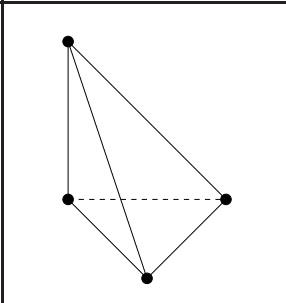
Velocity		Pressure
$\mathbb{P}_1$-bubble	$3\mathbb{P}_1$ or $4\mathbb{P}_1$	$\mathbb{P}_1$

Figure 54.1: Conventional representation of the $(\mathbb{P}_1\text{-bubble}, \mathbb{P}_1)$ pair in dimensions two (top) and three (bottom). The degrees of freedom for the velocity are shown in the first column ($\mathbb{P}_1$-bubble) and in the second column ($3\mathbb{P}_1$ in dimension two and $4\mathbb{P}_1$ in dimension three). Some isolines of the two-dimensional bubble function are drawn. The pressure degrees of freedom are shown in the third column.

by connecting the $(d+1)$ vertices of $\widehat{K}$ to $\widehat{\boldsymbol{x}}_{\widehat{K}}$. Then $\widehat{b}$ is defined to be the continuous piecewise affine function on $\widehat{K}$ that is equal to one at $\widehat{\boldsymbol{x}}_{\widehat{K}}$ and zero at the vertices of $\widehat{K}$. We introduce the finite-dimensional space $\widehat{\boldsymbol{P}} := \mathbb{P}_{1,d} \oplus (\mathrm{span}\{\widehat{b}\})^d$ and define $\widehat{\Sigma}$ to be the Lagrange degrees of freedom associated with the vertices of $\widehat{K}$ plus $\widehat{\boldsymbol{x}}_{\widehat{K}}$ for each Cartesian component of the velocity.

Let $(\mathcal{T}_h)_{h \in \mathcal{H}}$ be a shape-regular sequence of affine simplicial meshes so that each mesh covers D exactly. Recalling that we are enforcing homogeneous Dirichlet conditions on the velocity, the approximation spaces are defined by

$$\boldsymbol{V}_{h0} := \mathbb{P}_{1,0}^{\mathrm{g}}(\mathcal{T}_h) \oplus \boldsymbol{B}_h, \qquad Q_h := P_1^{\mathrm{g}}(\mathcal{T}_h) \cap L_*^2(D), \tag{54.6}$$

where $\boldsymbol{B}_h := \bigoplus_{K \in \mathcal{T}_h} (\mathrm{span}\{b_K\})^d$ and $b_K := \widehat{b} \circ \boldsymbol{T}_K$ being the bubble function associated with the mesh cell $K \in \mathcal{T}_h$. Notice that

$$\boldsymbol{V}_{h0} = \{\boldsymbol{v}_h \in \boldsymbol{C}^0(\overline{D}) \mid \forall K \in \mathcal{T}_h, \; \boldsymbol{v}_h \circ \boldsymbol{T}_K \in \widehat{\boldsymbol{P}}, \; \boldsymbol{v}_{h|\partial D} = \boldsymbol{0}\}, \tag{54.7}$$

and that $\boldsymbol{V}_{h0} \subset \boldsymbol{W}_0^{1,p}(D)$ for all $p \in (1, \infty)$. We now show that the $(\mathbb{P}_1\text{-bubble}, \mathbb{P}_1)$ pair is stable. We do so by constructing a Fortin operator as in Lemma 54.2.

Lemma 54.5 (Stability). *Let $p \in (1, \infty)$ and let $p' \in (1, \infty)$ be s.t. $\frac{1}{p} + \frac{1}{p'} = 1$. Let $\boldsymbol{V}_{h0}$ and Q_h*

be defined in (54.6). *There is* β_0 *such that for all* $h \in \mathcal{H}$,

$$\inf_{q_h \in Q_h} \sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}} \frac{|\int_D q_h \nabla \cdot \boldsymbol{v}_h \, \mathrm{d}x|}{\|q_h\|_{L^{p'}(D)} |\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(D)}} \geq \beta_0 > 0. \tag{54.8}$$

Proof. Let us build a Fortin operator by means of the construction devised in Lemma 54.2 with $\boldsymbol{V} := \boldsymbol{W}_0^{1,p}(D)$ equipped with the norm $\|\boldsymbol{v}\|_{\boldsymbol{V}} := |\boldsymbol{v}|_{\boldsymbol{W}^{1,p}(D)}$. We define the operator $\boldsymbol{\Pi}_{2h} : \boldsymbol{V} \to \boldsymbol{V}_{h0}$ by setting

$$\boldsymbol{\Pi}_{2h}(\boldsymbol{v}) := \sum_{K \in \mathcal{T}_h} \frac{\int_K \boldsymbol{v} \, \mathrm{d}x}{\int_K b_K \, \mathrm{d}x} b_K \in \boldsymbol{B}_h \subset \boldsymbol{V}_{h0}.$$

This operator is linear in agreement with the assumption (i) of Lemma 54.2. Moreover, the definition of $\boldsymbol{\Pi}_{2h}$ implies that $\int_K \boldsymbol{\Pi}_{2h}(\boldsymbol{v}) \, \mathrm{d}x = \int_K \boldsymbol{v} \, \mathrm{d}x$ for all $\boldsymbol{v} \in \boldsymbol{V}$. Then for all $(\boldsymbol{v}, q_h) \in \boldsymbol{V} \times Q_h$ we infer that

$$b(\boldsymbol{v}, q_h) = -\int_D q_h \nabla \cdot \boldsymbol{v} \, \mathrm{d}x = \int_D \boldsymbol{v} \cdot \nabla q_h \, \mathrm{d}x = \sum_{K \in \mathcal{T}_h} \nabla q_{h|K} \cdot \int_K \boldsymbol{v} \, \mathrm{d}x$$

$$= \sum_{K \in \mathcal{T}_h} \nabla q_{h|K} \cdot \int_K \boldsymbol{\Pi}_{2h}(\boldsymbol{v}) \, \mathrm{d}x = \int_D \boldsymbol{\Pi}_{2h}(\boldsymbol{v}) \cdot \nabla q_h \, \mathrm{d}x = b(\boldsymbol{\Pi}_{2h}(\boldsymbol{v}), q_h),$$

which proves the assumption (ii) of Lemma 54.2. We now set $\boldsymbol{\Pi}_{1h} := \boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}$, where $\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}} : \boldsymbol{V} \to \boldsymbol{V}_{h0}$ is the $\mathbb{R}^d$-valued version of the $\boldsymbol{W}_0^{1,p}$-conforming quasi-interpolation operator introduced in §22.4.2. We observe that the real number $c_{1h} := \sup_{\boldsymbol{v} \in \boldsymbol{V}} \frac{|\boldsymbol{\Pi}_{1h}(\boldsymbol{v})|_{\boldsymbol{W}^{1,p}(D)}}{|\boldsymbol{v}|_{\boldsymbol{W}^{1,p}(D)}}$ is uniformly bounded w.r.t. $h \in \mathcal{H}$. Moreover, the regularity of the mesh sequence and Lemma 11.7 imply that for all $K \in \mathcal{T}_h$,

$$|b_K|_{W^{1,p}(K)} \leq c \, \|\mathbb{J}_K^{-1}\|_{\ell^2} |\det(\mathbb{J}_K)|^{\frac{1}{p}} |\widehat{b}|_{W^{1,p}(\widehat{K})} \leq c' \, h_K^{-1} |K|^{\frac{1}{p}}.$$

Similar arguments show that $\int_K b_K \, \mathrm{d}x \geq c|K|$ and Hölder's inequality implies that $|\int_K \boldsymbol{v} \, \mathrm{d}x| \leq |K|^{\frac{1}{p'}} \|\boldsymbol{v}\|_{\boldsymbol{L}^p(K)}$. Putting these estimates together shows that

$$|\boldsymbol{\Pi}_{2h}(\boldsymbol{v})|_{\boldsymbol{W}^{1,p}(K)} \leq c \, h_K^{-1} \|\boldsymbol{v}\|_{\boldsymbol{L}^p(K)}.$$

Then the approximation properties of $\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}$ (see Theorem 22.14) yield

$$|\boldsymbol{\Pi}_{2h}(\boldsymbol{v} - \boldsymbol{\Pi}_{1h}(\boldsymbol{v}))|_{\boldsymbol{W}^{1,p}(K)} = |\boldsymbol{\Pi}_{2h}(\boldsymbol{v} - \boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}(\boldsymbol{v}))|_{\boldsymbol{W}^{1,p}(K)}$$
$$\leq c \, h_K^{-1} \|\boldsymbol{v} - \boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}(\boldsymbol{v})\|_{\boldsymbol{L}^p(K)} \leq c' \, |\boldsymbol{v}|_{\boldsymbol{W}^{1,p}(D_K)},$$

where D_K is the set of the points composing the mesh cells touching K. Summing the above bound over $K \in \mathcal{T}_h$ and using the regularity of the mesh sequence, we infer that

$$|\boldsymbol{\Pi}_{2h}(\boldsymbol{v} - \boldsymbol{\Pi}_{1h}(\boldsymbol{v}))|_{\boldsymbol{W}^{1,p}(D)} \leq c \, |\boldsymbol{v}|_{\boldsymbol{W}^{1,p}(D)}.$$

This shows that the real number $c_{2h} := \sup_{\boldsymbol{v} \in \boldsymbol{V}} \frac{|\boldsymbol{\Pi}_{2h}(\boldsymbol{v} - \boldsymbol{\Pi}_{1h}(\boldsymbol{v}))|_{\boldsymbol{W}^{1,p}(D)}}{|\boldsymbol{v}|_{\boldsymbol{W}^{1,p}(D)}}$ is uniformly bounded w.r.t. $h \in \mathcal{H}$. In conclusion, all the assumptions of Lemma 54.2 are met, showing that $\boldsymbol{\Pi}_h := \boldsymbol{\Pi}_{1h} + \boldsymbol{\Pi}_{2h}(I_{\boldsymbol{V}} - \boldsymbol{\Pi}_{1h})$ is a Fortin operator with $\gamma_h \geq (c_{1h} + c_{2h})^{-1}$. Notice that γ_h is bounded from below away from zero uniformly w.r.t. $h \in \mathcal{H}$. Invoking Lemma 54.1, we conclude that the inf-sup condition (54.8) holds true uniformly w.r.t. $h \in \mathcal{H}$. $\qquad\square$

Remark 54.6 (Convergence rate). Assume that the solution to (53.6) is such that $\boldsymbol{u} \in \boldsymbol{H}^2(D) \cap \boldsymbol{H}_0^1(D)$ and $p \in H^1(D) \cap L^2_*(D)$. Owing to Theorem 53.17, the discrete solution $(\boldsymbol{u}_h, p_h)$ to (53.14) with $(\boldsymbol{V}_{h0}, Q_h)$ defined in (54.6) satisfies $\mu|\boldsymbol{u} - \boldsymbol{u}_h|_{\boldsymbol{H}^1(D)} + \|p - p_h\|_{L^2(D)} \le ch(\mu|\boldsymbol{u}|_{\boldsymbol{H}^2(D)} + |p|_{H^1(D)})$. If the assumptions of Theorem 53.19 additionally hold true with $s := 1$, then $\mu\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{L}^2(D)} \le ch^2(\mu|\boldsymbol{u}|_{\boldsymbol{H}^2(D)} + |p|_{H^1(D)})$. Notice that the convergence rate of the error on the velocity is that associated with the finite element space $\mathbb{P}^{\mathrm{g}}_{1,0}(\mathcal{T}_h)$, i.e., the bubble functions introduced to approximate the velocity do not contribute to the approximation error, they contribute only to the stability of the discretization (see also Exercise 54.2). □

Remark 54.7 (Literature). The idea of using bubble functions has been introduced by Crouzeix and Raviart [151]. The analysis of the mini element is due to Arnold et al. [20]. □

54.3 Taylor–Hood element: the $(\mathbb{P}_2, \mathbb{P}_1)$ pair

This section is dedicated to the analysis of the Taylor–Hood element based on the $(\mathbb{P}_2, \mathbb{P}_1)$ pair. Compared to the mini element which is based on the $(\mathbb{P}_1\text{-bubble}, \mathbb{P}_1)$ pair, the idea is to further enrich the discrete velocity space so as to improve by one order the convergence rate of the error. Let $(\mathcal{T}_h)_{h \in \mathcal{H}}$ be a shape-regular family of affine simplicial meshes. Recalling that we are enforcing homogeneous Dirichlet conditions on the velocity, the approximation spaces are defined by

$$\boldsymbol{V}_{h0} := \boldsymbol{P}^{\mathrm{g}}_{2,0}(\mathcal{T}_h), \qquad Q_h := P^{\mathrm{g}}_1(\mathcal{T}_h) \cap L^2_*(D), \tag{54.9}$$

i.e., the velocity is approximated using continuous $\mathbb{P}_2$ elements and the pressure is approximated using continuous $\mathbb{P}_1$ elements. The conventional representation of this element is shown in Figure 54.2. We are going to prove the inf-sup condition (54.1) by using the technique described in §54.1.2, i.e., we first establish a weak control on the pressure gradient, then we invoke Lemma 54.3. As above, we set $\boldsymbol{V} := \boldsymbol{W}^{1,p}_0(D)$ and $Q := L^{p'}_*(D)$ with $p, p' \in (1, \infty)$ and $\frac{1}{p} + \frac{1}{p'} = 1$. Notice that $\boldsymbol{V}_{h0} \subset \boldsymbol{V}$ and $Q_h \subset Q$.

Lemma 54.8 (Bound on pressure gradient). *Let $\boldsymbol{V}_{h0}$, Q_h be defined in (54.9). Assume that $d \in \{2, 3\}$ and that every mesh cell has at least d internal edges (i.e., at most one face in ∂D). There is c such that the following holds true for all $p \in (1, \infty)$ and all $h \in \mathcal{H}$:*

$$\sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}} \frac{|\int_D q_h \nabla \cdot \boldsymbol{v}_h \, \mathrm{d}x|}{|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(D)}} \ge c \left(\sum_{K \in \mathcal{T}_h} h_K^{p'} \|\nabla q_h\|^{p'}_{\boldsymbol{L}^{p'}(K)} \right)^{\frac{1}{p'}}. \tag{54.10}$$

Proof. We only give the proof for $d = 3$ since the proof for $d = 2$ is similar. Let us number all the internal mesh edges from 1 to $N^{\mathrm{i}}_{\mathrm{e}}$. Consider an oriented edge E_i with $i \in \{1 : N^{\mathrm{i}}_{\mathrm{e}}\}$, and denote its two endpoints by $\boldsymbol{z}_i^\pm$ and its midpoint by $\boldsymbol{m}_i$. Set $l_i := \|\boldsymbol{z}_i^+ - \boldsymbol{z}_i^-\|_{\ell^2}$ and $\boldsymbol{\tau}_i := l_i^{-1}(\boldsymbol{z}_i^+ - \boldsymbol{z}_i^-)$, so that l_i is the length of E_i and $\boldsymbol{\tau}_i$ is the unit tangent vector orienting E_i. Let q_h be a function in Q_h and let sgn be the sign function. Let $\boldsymbol{v}_h \in \boldsymbol{V}_{h0}$ be (uniquely) defined by prescribing its global degrees of freedom in $\boldsymbol{V}_{h0}$ as follows:

$$\begin{cases} \boldsymbol{v}_h(\boldsymbol{a}_j) := \boldsymbol{0} & \text{if } \boldsymbol{a}_j \text{ is a mesh vertex,} \\ \boldsymbol{v}_h(\boldsymbol{m}_i) := -l_i^{p'} \mathrm{sgn}(\partial_{\boldsymbol{\tau}_i} q_h)|\partial_{\boldsymbol{\tau}_i} q_h|^{p'-1} \boldsymbol{\tau}_i & \text{if } E_i \not\subset \partial D, \\ \boldsymbol{v}_h(\boldsymbol{m}_i) := \boldsymbol{0} & \text{if } E_i \subset \partial D, \end{cases}$$

Velocity	Pressure	Velocity	Pressure
$\mathbb{P}_2$	$\mathbb{P}_1$	$\mathbf{Q}_2$	$\mathbb{Q}_1$
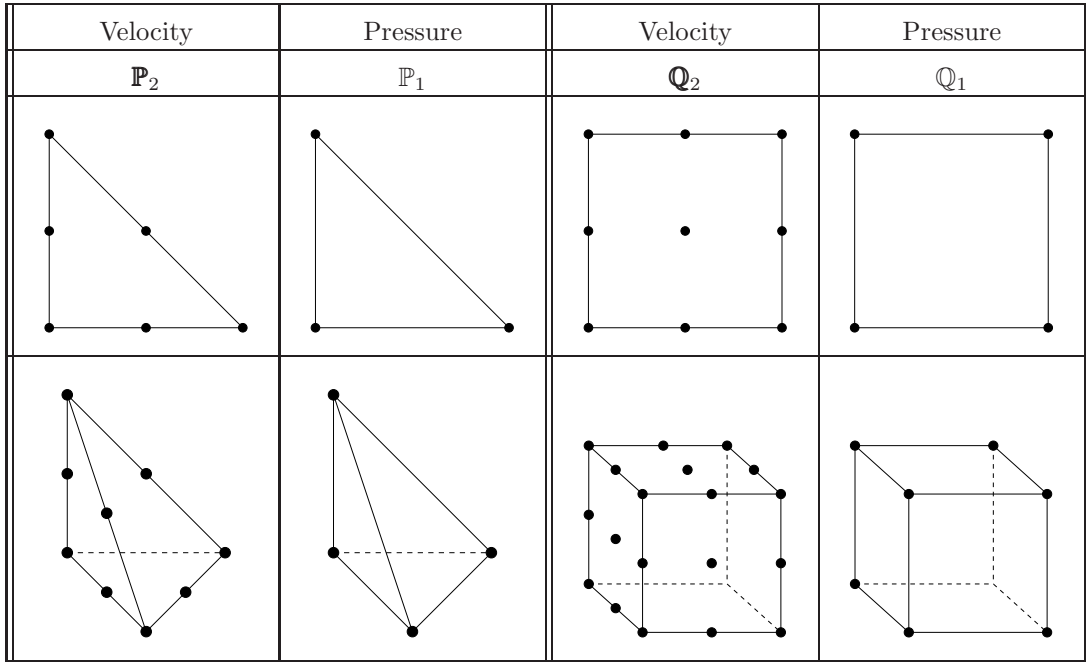
Figure 54.2: Conventional representation of the $(\mathbb{P}_2, \mathbb{P}_1)$ pair (left) and of the $(\mathbf{Q}_2, \mathbb{Q}_1)$ pair (right) in dimensions two (top) and three (bottom, only visible degrees of freedom are shown).

where $\partial_{\tau_i} q_h := \boldsymbol{\tau}_i \cdot \nabla q_h$ denotes the tangential derivative of q_h along the oriented edge E_i. Note that $\boldsymbol{v}_h(\boldsymbol{m}_i)$ depends only on the values of q_h on E_i. Let $K \in \mathcal{T}_h$. Using the quadrature formula

$$\int_K \phi \, dx = |K| \left(\sum_{\boldsymbol{m} \in \mathcal{M}_K} \frac{\phi(\boldsymbol{m})}{5} - \sum_{\boldsymbol{a} \in \mathcal{V}_K} \frac{\phi(\boldsymbol{a})}{20} \right), \qquad \forall \phi \in \mathbb{P}_2,$$

where $\mathcal{M}_K$ is the set of the midpoints of the edges of K and $\mathcal{V}_K$ is the set of the vertices of K and since Q_h is H^1-conforming, we infer that

$$\int_D q_h \nabla \cdot \boldsymbol{v}_h \, dx = - \int_D \boldsymbol{v}_h \cdot \nabla q_h \, dx = - \sum_{K \in \mathcal{T}_h} \int_K \boldsymbol{v}_h \cdot \nabla q_h \, dx$$

$$= - \sum_{K \in \mathcal{T}_h} |K| \sum_{\boldsymbol{m}_i \in K} \frac{1}{5} \boldsymbol{v}_h(\boldsymbol{m}_i) \cdot \nabla q_h(\boldsymbol{m}_i)$$

$$= \sum_{K \in \mathcal{T}_h} |K| \sum_{\boldsymbol{m}_i \in K} \frac{1}{5} |\partial_{\tau_i} q_h(\boldsymbol{m}_i)|^{p'} l_i^{p'} \geq c \sum_{K \in \mathcal{T}_h} h_K^{p'} \|\nabla q_h\|_{\boldsymbol{L}^{p'}(K)}^{p'}.$$

The last inequality results from the fact that $l_i \geq ch_K$ owing to the regularity of the mesh sequence, and that every tetrahedron $K \in \mathcal{T}_h$ has at least three edges in D, i.e., the quantities $|\partial_{\tau_i} q_h(\boldsymbol{m}_i)|$, where $\boldsymbol{m}_i$ spans the midpoints of the edges of K that are not in ∂D, control $\|\nabla q_h\|_{\ell^2}$. Finally, the inverse inequality from Lemma 12.1 (with $r := p$, $l := 1$, $m := 0$) together with Proposition 12.5 implies that for all $K \in \mathcal{T}_h$,

$$|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(K)}^p \leq c \, h_K^{-p} |K| \sum_{\boldsymbol{m} \in \mathcal{M}_K} \|\boldsymbol{v}_h(\boldsymbol{m})\|_{\ell^2}^p,$$

and since $l_i \leq ch_K$, we have $\|\boldsymbol{v}_h(\boldsymbol{m})\|_{\ell^2} \leq ch_K^{p'}\|\nabla q_h\|_{\ell^2}^{p'-1}$. Since $p(p'-1) = p'$, combining these bounds shows that $|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(K)}^p \leq ch_K^{p'}\|\nabla q_h\|_{\boldsymbol{L}^{p'}(K)}^{p'}$ for all $K \in \mathcal{T}_h$. This proves (54.10). $\qquad\square$

Lemma 54.9 (Stability). *For all $p \in (1,\infty)$ and under the hypotheses of Lemma 54.8, the $(\mathbb{P}_2, \mathbb{P}_1)$ pair satisfies the inf-sup condition (54.8) uniformly w.r.t. $h \in \mathcal{H}$.*

Proof. Apply Lemma 54.3. $\qquad\square$

Remark 54.10 (Convergence rate). Owing to Theorem 53.17 and assuming that the solution to (53.6) is smooth enough, the solution to (53.14) with $(\boldsymbol{V}_{h0}, Q_h)$ defined in (54.9) satisfies $\mu|\boldsymbol{u}-\boldsymbol{u}_h|_{\boldsymbol{H}^1(D)}+\|p-p_h\|_{L^2(D)} \leq ch^2(\mu|\boldsymbol{u}|_{\boldsymbol{H}^3(D)}+|p|_{H^2(D)})$. Moreover, if the assumptions of Theorem 53.19 are met for some $s \in (0,1]$, then $\mu\|\boldsymbol{u}-\boldsymbol{u}_h\|_{\boldsymbol{L}^2(D)} \leq ch^{2+s}\ell_D^{1-s}(\mu|\boldsymbol{u}|_{\boldsymbol{H}^3(D)}+|p|_{H^2(D)})$. $\qquad\square$

Remark 54.11 (Literature). Further insight and alternative proofs can be found in Bercovier and Pironneau [54, Prop. 1], Girault and Raviart [217, p. 176], Stenberg [354]. We refer the reader to Mardal et al. [294] for the construction of a Fortin operator associated with the Taylor–Hood element in dimension two. Well-balanced schemes (see Remark 53.22) using Taylor–Hood mixed finite elements are analyzed in Lederer et al. [279]. $\qquad\square$

54.4   Generalizations of the Taylor–Hood element

In this section, we briefly review some generalizations of the Taylor–Hood element: extension to quadrangles, higher-order extensions, and the use of a submesh to build the discrete velocity space.

54.4.1   The $(\mathbb{P}_k, \mathbb{P}_{k-1})$ and $(\mathbb{Q}_k, \mathbb{Q}_{k-1})$ pairs

It is possible to generalize the Taylor–Hood element to quadrangles and hexahedra. For instance, the $(\mathbb{Q}_2, \mathbb{Q}_1)$ pair has the same properties as the Taylor–Hood element; see Figure 54.2.

It is also possible to use higher-degree polynomials. For $k \geq 2$, the $(\mathbb{P}_k, \mathbb{P}_{k-1})$ pair and the $(\mathbb{Q}_k, \mathbb{Q}_{k-1})$ pair are stable in dimensions two and three. Provided the solution to (53.6) is smooth enough, these elements yield the error estimates $\mu|\boldsymbol{u} - \boldsymbol{u}_h|_{\boldsymbol{H}^1(D)} + \|p - p_h\|_{L^2(D)} \leq ch^k(\mu|\boldsymbol{u}|_{\boldsymbol{H}^{k+1}(D)} + |p|_{H^k(D)})$ and $\mu\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{L}^2(D)} \leq ch^{k+s}\ell_D^{1-s}(\mu|\boldsymbol{u}|_{\boldsymbol{H}^{k+1}(D)} + |p|_{H^k(D)})$ if the assumptions of Theorem 53.19 are met for some $s \in (0,1]$. Proofs and further insight can be found in Stenberg [352, p. 18], Brezzi and Falk [92], Boffi et al. [65, p. 494], Boffi [60].

54.4.2   The $(\mathbb{P}_1\text{-iso-}\mathbb{P}_2, \mathbb{P}_1)$ and $(\mathbb{Q}_1\text{-iso-}\mathbb{Q}_2, \mathbb{Q}_1)$ pairs

An alternative to the Taylor–Hood element consists of replacing the $\mathbb{P}_2$ approximation of the velocity by a $\mathbb{P}_1$ approximation on a finer simplicial mesh. This finer mesh, say $\mathcal{T}_{\frac{h}{2}}$, is constructed as follows. In two dimensions, each triangle in $\mathcal{T}_h$ is divided into four new triangles by connecting the midpoints of the three edges. In three dimensions, each tetrahedron in $\mathcal{T}_h$ is divided into eight new tetrahedra (all having the same volume) by dividing each face into four new triangles and by connecting the midpoints of one pair of nonintersecting edges (there are three pairs of nonintersecting edges). This construction is illustrated in the top and bottom left panels of Figure 54.3. The discrete spaces are

$$\boldsymbol{V}_{h0} := \boldsymbol{P}_{1,0}^{\mathrm{g}}(\mathcal{T}_{\frac{h}{2}}), \qquad Q_h := P_1^{\mathrm{g}}(\mathcal{T}_h) \cap L_*^2(D). \tag{54.11}$$

Velocity	Pressure	Velocity	Pressure
$\mathbb{P}_1$-iso-$\mathbb{P}_2$	$\mathbb{P}_1$	$\mathbb{Q}_1$-iso-$\mathbb{Q}_2$	$\mathbb{Q}_1$
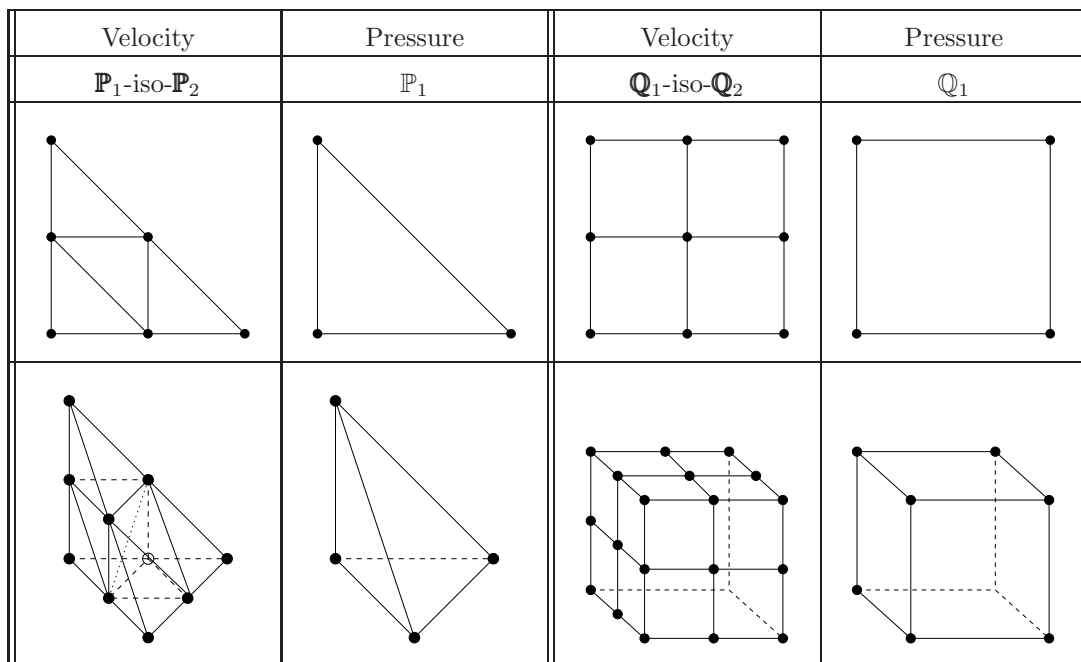
Figure 54.3: $(\mathbb{P}_1\text{-iso-}\mathbb{P}_2, \mathbb{P}_1)$ (left) and $(\mathbb{Q}_1\text{-iso-}\mathbb{Q}_2, \mathbb{Q}_1)$ (right) pairs in dimensions two (top) and three (bottom, only visible degrees of freedom are shown for the $(\mathbb{Q}_1\text{-iso-}\mathbb{Q}_2, \mathbb{Q}_1)$ pair).

These finite element pairs are often called $(\mathbb{P}_1\text{-iso-}\mathbb{P}_2, \mathbb{P}_1)$, or $(4\,\mathbb{P}_1, \mathbb{P}_1)$ in two dimensions and $(8\,\mathbb{P}_1, \mathbb{P}_1)$ in three dimensions.

The $(\mathbb{P}_1\text{-iso-}\mathbb{P}_2, \mathbb{P}_1)$ pair can be generalized to quadrangles in two dimensions and hexahedra in three dimensions. Assume that $(\mathcal{T}_h)_{h \in \mathcal{H}}$ is a shape-regular sequence of meshes composed of quadrangles or hexahedra. A new mesh $\mathcal{T}_{\frac{h}{2}}$ is defined in two dimensions by dividing each quadrangle in $\mathcal{T}_h$ into four new quadrangles and by connecting the midpoints of all the pairs of nonintersecting edges. In three dimensions, we divide each hexahedron in $\mathcal{T}_h$ into eight new hexahedra by dividing each face into four quadrangles and by connecting the barycenters of all the pairs of nonintersecting faces. This construction is illustrated in the top and bottom right panels of Figure 54.3. The discrete spaces are

$$\boldsymbol{V}_{h0} := \{\boldsymbol{v}_h \in \boldsymbol{C}^0(\overline{D}) \mid \forall K \in \mathcal{T}_{\frac{h}{2}},\ \boldsymbol{v}_h \circ \boldsymbol{T}_K \in \mathbb{Q}_1,\ \boldsymbol{v}_{h|\partial D} = \boldsymbol{0}\}, \tag{54.12a}$$

$$Q_h := \{q_h \in C^0(\overline{D}) \cap L^2_*(D) \mid \forall K \in \mathcal{T}_h,\ q_h \circ \boldsymbol{T}_K \in \mathbb{Q}_1\}. \tag{54.12b}$$

These finite elements are often called $(\mathbb{Q}_1\text{-iso-}\mathbb{Q}_2, \mathbb{Q}_1)$, or $(4\,\mathbb{Q}_1, \mathbb{Q}_1)$ in dimension two and $(8\,\mathbb{Q}_1, \mathbb{Q}_1)$ in dimension three.

Lemma 54.12 (Stability). *For all $p \in (1, \infty)$, and under the hypotheses of Lemma 54.8 if $\mathcal{T}_h$ is composed of simplices, the $(\mathbb{P}_1\text{-iso-}\mathbb{P}_2, \mathbb{P}_1)$ and $(\mathbb{Q}_1\text{-iso-}\mathbb{Q}_2, \mathbb{Q}_1)$ pairs satisfy the inf-sup condition (54.8) uniformly w.r.t. $h \in \mathcal{H}$.*

Proof. Adapt the proof of Lemma 54.8; see Bercovier and Pironneau [54] (for $d = 2$ and $p = 2$) and Exercise 54.4. $\qquad\square$

Remark 54.13 (Convergence rate). Owing to Theorem 53.17 and assuming that the solution to (53.6) is smooth enough, the discrete solution to (53.14) with $(\boldsymbol{V}_{h0}, Q_h)$ defined in either (54.11) or (54.12) satisfies $\mu|\boldsymbol{u}-\boldsymbol{u}_h|_{\boldsymbol{H}^1(D)}+\|p-p_h\|_{L^2(D)} \leq ch(\mu|\boldsymbol{u}|_{\boldsymbol{H}^2(D)}+|p|_{H^1(D)})$, and if the assumptions of Theorem 53.19 are met for some $s \in (0,1]$, we have $\mu\|\boldsymbol{u}-\boldsymbol{u}_h\|_{\boldsymbol{L}^2(D)} \leq ch^{1+s}\ell_D^{1-s}(\mu|\boldsymbol{u}|_{\boldsymbol{H}^2(D)}+|p|_{H^1(D)})$. $\qquad\square$

Exercises

Exercise 54.1 (Mini element). Show that the Fortin operator $\boldsymbol{\Pi}_h$ constructed in the proof of Lemma 54.5 is of the form $\boldsymbol{\Pi}_h(\boldsymbol{v}) := \boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}(\boldsymbol{v}) + \sum_{K\in\mathcal{T}_h}\sum_{i\in\{1:d\}}\gamma_K^i(\boldsymbol{v})b_K\boldsymbol{e}_i$, for some coefficients $\gamma_K^i(\boldsymbol{v})$ to be determined. Here, $\{\boldsymbol{e}_i\}_{i\in\{1:d\}}$ is the canonical Cartesian basis of $\mathbb{R}^d$.

Exercise 54.2 (Bubble⇔Stabilization). Consider the mini element defined in §54.2 and assume that the viscosity μ is constant over D. Recall that $\boldsymbol{V}_{h0} := \boldsymbol{V}_{h0}^1 \oplus \boldsymbol{B}_h$ and $Q_h := P_1^{\mathrm{g}}(\mathcal{T}_h) \cap L_*^2(D)$ with $\boldsymbol{V}_{h0}^1 := \mathbb{P}_{1,0}^{\mathrm{g}}(\mathcal{T}_h)$. Let $(\boldsymbol{u}_h, p_h)$ be the solution to the discrete Stokes problem (53.14). (i) Show that $a(\boldsymbol{v}_h, \boldsymbol{b}_h) = 0$ for all $\boldsymbol{v}_h \in \boldsymbol{V}_{h0}^1$ and all $\boldsymbol{b}_h \in \boldsymbol{B}_h$. (ii) Set $\boldsymbol{u}_h := \boldsymbol{u}_h^1 + \boldsymbol{u}_h^b \in \boldsymbol{V}_{h0}$. Show that

$$a(\boldsymbol{u}_h^1, \boldsymbol{v}_h) + b(\boldsymbol{v}_h, p_h) = F(\boldsymbol{v}_h), \qquad \forall \boldsymbol{v}_h \in \boldsymbol{V}_{h0}^1. \tag{54.13}$$

(iii) Let $b_K := \widehat{b}\circ\boldsymbol{T}_K$ be the bubble function on $K \in \mathcal{T}_h$. Let $\{\boldsymbol{e}_i\}_{i\in\{1:d\}}$ be the canonical Cartesian basis of $\mathbb{R}^d$. Let $\mathcal{S}^K \in \mathbb{R}^{d\times d}$ be defined by $\mathcal{S}_{ij}^K := \frac{1}{\int_K b_K\,\mathrm{d}x}a(b_K\boldsymbol{e}_j, b_K\boldsymbol{e}_i)$ for all $i,j \in \{1:d\}$. Let $\boldsymbol{u}_{h|K}^b := \sum_{i\in\{1:d\}}c_K^i\boldsymbol{e}_ib_K$. Show that $\boldsymbol{c}_K = (\mathcal{S}^K)^{-1}(\boldsymbol{F}_K - \nabla p_{h|K})$, where $F_K^i := \frac{1}{\int_K b_K\,\mathrm{d}x}F(b_K\boldsymbol{e}_i)$, for all $i \in \{1:d\}$. (iv) Set $c_h(p_h, q_h) := \sum_{K\in\mathcal{T}_h}\nabla q_{h|K}(\mathcal{S}^K)^{-1}\nabla p_{h|K}\int_K b_K\,\mathrm{d}x$ and $R_h(q_h) := \sum_{K\in\mathcal{T}_h}\nabla q_{h|K}(\mathcal{S}^K)^{-1}\boldsymbol{F}_K\int_K b_K\,\mathrm{d}x$. Show that the mass conservation equation becomes

$$b(\boldsymbol{u}_h^1, q_h) - c_h(p_h, q_h) = G(q_h) - R_h(q_h), \qquad \forall q_h \in Q_h. \tag{54.14}$$

Note: since $(\mathcal{S}_K)^{-1}$ scales like $\mu^{-1}h_K^2$, $c_h(p_h, q_h)$ behaves like $\sum_{K\in\mathcal{T}_h}\frac{h_K^2}{\mu}\int_K\nabla q_h\cdot\nabla p_h\,\mathrm{d}x$, and $R_h(q_h)$ scales like $\sum_{K\in\mathcal{T}_h}\frac{h_K^2}{\mu}\int_K\nabla q_{h|K}\cdot\boldsymbol{F}_K\,\mathrm{d}x$. This shows that, once the bubbles are eliminated, the system (54.13)-(54.14) is equivalent to a stabilized form of the Stokes system for the $(\mathbb{P}_1,\mathbb{P}_1)$ pair; see Chapters 62 and 63.

Exercise 54.3 (Singular vertex). Let $K \subset \mathbb{R}^2$ be a quadrangle and let $\boldsymbol{z}$ be the intersection of the two diagonals of K. Let $K_1,\ldots,K_4$ be the four triangles formed by dividing K along its two diagonals (assume that $K_1 \cap K_3 = \{\boldsymbol{z}\}$ and $K_2 \cap K_4 = \{\boldsymbol{z}\}$). (i) Let ϕ be a scalar field continuous over K and of class C^1 over the triangles $K_1,\ldots,K_4$. Prove that $\sum_{i\in\{1:4\}}(-1)^i\boldsymbol{n}\cdot\nabla\phi_{|K_i}(\boldsymbol{z}) = 0$ for every unit vector $\boldsymbol{n}$. (ii) Let $\boldsymbol{v}$ be a vector field continuous over K and of class C^1 over the triangles $K_1,\ldots,K_4$. Prove that $\sum_{i\in\{1:4\}}(-1)^i\nabla\cdot\boldsymbol{v}_{|K_i}(\boldsymbol{z}) = 0$. (iii) Assume that $\boldsymbol{v}$ is linear over each triangle. Show that the four equations $\int_{K_i}\nabla\cdot\boldsymbol{v}\,\mathrm{d}x = 0$ for all $i \in \{1:4\}$ are linearly dependent.

Exercise 54.4 ($\mathbb{P}_1$-iso-$\mathbb{P}_2$, $\mathbb{P}_1$). Consider the setting of Lemma 54.12 with the $(\mathbb{P}_1$-iso-$\mathbb{P}_2, \mathbb{P}_1)$ pair in dimension three. (i) Let $K \in \mathcal{T}_h$. Let $\mathcal{V}_K$ be the set of the vertices of K. Let $\mathcal{M}_K$ be the midpoints of the six edges of K. Let $\mathcal{M}_K^1$ be the set of the two midpoints that are connected to create the 8 new tetrahedra. Let $\mathcal{M}_K^2$ be the set of the remaining midpoints. Let $\boldsymbol{V}_{h0}$ be the $\mathbb{P}_1$ velocity space based of $\mathcal{T}_{h/2}$. Find the coefficients α, β, γ so that the following quadrature is exact for all $\boldsymbol{w}_h \in V_{h0}$: $\int_K\boldsymbol{w}_h\,\mathrm{d}x = |K|(\alpha\sum_{\boldsymbol{z}\in\mathcal{V}_K}\boldsymbol{w}_h(\boldsymbol{z}) + \beta\sum_{\boldsymbol{m}\in\mathcal{M}_K^1}\boldsymbol{w}_h(\boldsymbol{m}) + \gamma\sum_{\boldsymbol{m}\in\mathcal{M}_K^2}\boldsymbol{w}_h(\boldsymbol{m}))$. (*Hint*:

on a tetrahedron K' with vertices $\{\boldsymbol{z}'\}_{\boldsymbol{z}' \in \mathcal{V}_{K'}}$, the quadrature $\int_{K'} \boldsymbol{w}_h \, \mathrm{d}x = |K'| \sum_{\boldsymbol{z}' \in \mathcal{V}_{K'}} \frac{1}{4} \boldsymbol{w}_h(\boldsymbol{z}')$ is exact on $\mathbb{P}_1$.) (ii) Prove Lemma 54.12 for the $(\mathbb{P}_1\text{-iso-}\mathbb{P}_2, \mathbb{P}_1)$ pair in dimension three for all $p \in (1, \infty)$. (*Hint*: adapt the proof of Lemma 54.8.)

Chapter 55

Stokes equations: Stable pairs (II)

In this chapter, we continue the study of stable finite element pairs that are suitable to approximate the Stokes equations. In doing so, we introduce another technique to prove the inf-sup condition that is based on a notion of macroelement. Recall that we assume that Dirichlet conditions are enforced on the velocity over the whole boundary, that D is a polyhedron in $\mathbb{R}^d$, and that $(\mathcal{T}_h)_{h \in \mathcal{H}}$ is a shape-regular sequence of affine meshes so that each mesh covers D exactly. In this chapter, we focus more specifically on the case where the discrete pressure space is a broken finite element space.

55.1 Macroelement techniques

In addition to the Fortin operator technique described in Lemma 54.1 and the method consisting of weakly controlling the pressure gradient described in Lemma 54.3, we now present a third method to establish the inf-sup condition between the discrete velocity space and the discrete pressure space. This method is based on a notion of macroelement.

We return to the abstract setting and consider two complex Banach spaces $\boldsymbol{V}$ and Q and a bounded sesquilinear form b on $\boldsymbol{V} \times Q$. Let $\boldsymbol{V}_{h0} \subset \boldsymbol{V}$ and $Q_h \subset Q$. Recall that $\|b\|$ denotes the boundedness constant of b on $\boldsymbol{V} \times Q$ and that the inf-sup condition (54.1) takes the form

$$\inf_{q_h \in Q_h} \sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}} \frac{|b(\boldsymbol{v}_h, q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}} \|q_h\|_Q} =: \beta_h > 0. \tag{55.1}$$

Lemma 55.1 (Partition lemma). *Let $\boldsymbol{V}_{h0}^1, \boldsymbol{V}_{h0}^2$ be two subspaces of $\boldsymbol{V}_{h0}$ and Q_h^1, Q_h^2 be two subspaces of Q such that $Q_h = Q_h^1 + Q_h^2$. Let*

$$\beta_1 := \inf_{q_h \in Q_h^1} \sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}^1} \frac{|b(\boldsymbol{v}_h, q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}} \|q_h\|_Q}, \qquad \beta_2 := \inf_{q_h \in Q_h^2} \sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}^2} \frac{|b(\boldsymbol{v}_h, q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}} \|q_h\|_Q},$$

$$b_{12} := \sup_{q_h \in Q_h^1} \sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}^2} \frac{|b(\boldsymbol{v}_h, q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}} \|q_h\|_Q}, \qquad b_{21} := \sup_{q_h \in Q_h^2} \sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}^1} \frac{|b(\boldsymbol{v}_h, q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}} \|q_h\|_Q}.$$

Assume that $0 < \beta_1 \beta_2$ and $\lambda_1 \lambda_2 < 1$ with $\lambda_1 := \frac{b_{12}}{\beta_2}$, $\lambda_2 := \frac{b_{21}}{\beta_1}$. Then the inf-sup condition (55.1) holds true with $\beta_h \geq \frac{1}{4} \min(\beta_1, \beta_2)$ if $\lambda_1 + \lambda_2 \leq 1$ and with $\beta_h \geq \frac{1}{64}(1 - \lambda_1 \lambda_2)\|b\|^{-2} \min(\beta_1, \beta_2)^3$ otherwise.

Proof. Let $q_h := q_h^1 + q_h^2 \in Q_h \backslash \{0\}$. The definition of β_1, β_2 together with the assumption $0 < \beta_1\beta_2$ implies that there exists $\boldsymbol{v}_h^l \in \boldsymbol{V}_h^l$ so that $b(\boldsymbol{v}_h^l, q_h^l) = \|q_h^l\|_Q^2$ and $\beta_l \|\boldsymbol{v}_h^l\|_{\boldsymbol{V}} \le \|q_h^l\|_Q$ for all $l \in \{1, 2\}$. We now investigate two cases: either $\lambda_1 + \lambda_2 \le 1$ or $\lambda_1 + \lambda_2 > 1$.
(1) Let us assume that $\lambda_1 + \lambda_2 \le 1$. Then, setting $\boldsymbol{v}_h := \boldsymbol{v}_h^1 + \boldsymbol{v}_h^2$ we have

$$
\begin{aligned}
b(\boldsymbol{v}_h, q_h) &= b(\boldsymbol{v}_h^1, q_h^1) + b(\boldsymbol{v}_h^2, q_h^1) + b(\boldsymbol{v}_h^1, q_h^2) + b(\boldsymbol{v}_h^2, q_h^2) \\
&\ge \|q_h^1\|_Q^2 + \|q_h^2\|_Q^2 - b_{12} \|\boldsymbol{v}_h^2\|_{\boldsymbol{V}} \|q_h^1\|_Q - b_{21} \|\boldsymbol{v}_h^1\|_{\boldsymbol{V}} \|q_h^2\|_Q \\
&\ge \|q_h^1\|_Q^2 + \|q_h^2\|_Q^2 - (\beta_2^{-1} b_{12} + \beta_1^{-1} b_{21}) \|q_h^1\|_Q \|q_h^2\|_Q.
\end{aligned}
$$

Using that $\beta_2^{-1} b_{12} + \beta_1^{-1} b_{21} = \lambda_1 + \lambda_2 \le 1$, we infer that

$$
\begin{aligned}
b(\boldsymbol{v}_h, q_h) &\ge \frac{1}{2} \|q_h^1\|_Q^2 + \frac{1}{2} \|q_h^2\|_Q^2 \ge \frac{1}{4} \left(\|q_h^1\|_Q + \|q_h^2\|_Q \right)^2 \\
&\ge \frac{1}{4} \|q_h\|_Q \left(\beta_1 \|\boldsymbol{v}_h^1\|_{\boldsymbol{V}} + \beta_2 \|\boldsymbol{v}_h^2\|_{\boldsymbol{V}} \right) \ge \frac{1}{4} \min(\beta_1, \beta_2) \|q_h\|_Q \|\boldsymbol{v}_h\|_{\boldsymbol{V}},
\end{aligned}
$$

where we used the triangle inequality and the above bounds on $\|\boldsymbol{v}_h^l\|_{\boldsymbol{V}}$ for all $l \in \{1, 2\}$. The assertion then follows with $\beta_h \ge \frac{1}{4} \min(\beta_1, \beta_2)$.
(2) Let us now assume that $\lambda_1 + \lambda_2 > 1$. Without loss of generality, we assume that $\lambda_2 \ge \lambda_1$. Let $\sigma \in \mathbb{R}$, let $\boldsymbol{v}_h := \boldsymbol{v}_h^1 + \sigma \boldsymbol{v}_h^2$, let $\epsilon > 0$, and let us minorize $b(\boldsymbol{v}_h, q_h)$ as follows:

$$
\begin{aligned}
b(\boldsymbol{v}_h, q_h) &= b(\boldsymbol{v}_h^1, q_h^1) + \sigma b(\boldsymbol{v}_h^2, q_h^1) + b(\boldsymbol{v}_h^1, q_h^2) + \sigma b(\boldsymbol{v}_h^2, q_h^2) \\
&\ge \|q_h^1\|_Q^2 + \sigma \|q_h^2\|_Q^2 - b_{12} \|\boldsymbol{v}_h^2\|_{\boldsymbol{V}} \|q_h^1\|_Q - b_{21} \sigma \|\boldsymbol{v}_h^1\|_{\boldsymbol{V}} \|q_h^2\|_Q \\
&\ge \|q_h^1\|_Q^2 + \sigma \|q_h^2\|_Q^2 - (\beta_2^{-1} b_{12} + \sigma \beta_1^{-1} b_{21}) \|q_h^1\|_Q \|q_h^2\|_Q \\
&\ge \left(1 - \frac{\epsilon}{2}(\lambda_1 + \sigma \lambda_2) \right) \|q_h^1\|_Q^2 + \left(\sigma - \frac{1}{2\epsilon}(\lambda_1 + \sigma \lambda_2) \right) \|q_h^2\|_Q^2.
\end{aligned}
$$

Let us show that we can choose σ and ϵ so that $\frac{\epsilon}{2}(\lambda_1 + \sigma \lambda_2) < 1$ and $\frac{1}{2\epsilon}(\lambda_1 + \sigma \lambda_2) < \sigma$. We consider the quadratic equation $\Psi(t) := (\lambda_1 + t\lambda_2)^2 - 4t = 0$. Since the discriminant, $16(1 - \lambda_1\lambda_2)$, is positive and $\lambda_2 \ne 0$, $\Psi(t)$ has two distinct roots, t_-, t_+, and Ψ is minimal at $\frac{1}{2}(t_- + t_+) = \frac{2 - \lambda_1\lambda_2}{\lambda_2^2}$. Therefore, if we choose $\sigma := \frac{2 - \lambda_1\lambda_2}{\lambda_2^2}$, we have $\Psi(\sigma) < 0$, i.e., $\frac{1}{2}(\lambda_1 + \sigma \lambda_2) < \frac{2\sigma}{\lambda_1 + \sigma \lambda_2}$. We then define ϵ by setting $\epsilon\sigma := \frac{1}{2}(\frac{1}{2}(\lambda_1 + \sigma \lambda_2) + \frac{2\sigma}{\lambda_1 + \sigma \lambda_2})$. This choice in turn implies that $\epsilon\sigma < \frac{2\sigma}{\lambda_1 + \sigma \lambda_2}$, i.e., $\frac{\epsilon}{2}(\lambda_1 + \sigma \lambda_2) < 1$ and that $\frac{1}{2}(\lambda_1 + \sigma \lambda_2) < \epsilon\sigma$, i.e., $\frac{1}{2\epsilon}(\lambda_1 + \sigma \lambda_2) < \sigma$. We have thus proved that $c_1 := 1 - \frac{\epsilon}{2}(\lambda_1 + \sigma \lambda_2) > 0$ and $c_2 := \sigma - \frac{1}{2\epsilon}(\lambda_1 + \sigma \lambda_2) > 0$. Then we conclude as above

$$
\begin{aligned}
b(\boldsymbol{v}_h, q_h) &\ge \frac{1}{2} \min(c_1, c_2) \|q_h\|_Q (\beta_1 \|\boldsymbol{v}_h^1\|_{\boldsymbol{V}} + \beta_2 \|\boldsymbol{v}_h^2\|_{\boldsymbol{V}}) \\
&\ge \frac{1}{2} \min(c_1, c_2) \min(\beta_1, \sigma^{-1}\beta_2) \|q_h\|_Q \|\boldsymbol{v}_h\|_{\boldsymbol{V}},
\end{aligned}
$$

and the assertion follows with $\beta_h \ge \frac{1}{2} \min(c_1, c_2) \min(\beta_1, \sigma^{-1}\beta_2)$. Notice that $\lambda_2 \in [\frac{1}{2}, \frac{\|b\|}{\beta_1}]$ because $2\lambda_2 \ge \lambda_1 + \lambda_2 \ge 1$ and $b_{21} \le \|b\|$. Moreover, since $\sigma = \frac{2 - \lambda_1\lambda_2}{\lambda_2^2}$ and $\epsilon = \frac{\lambda_2(3 - \lambda_1\lambda_2)}{2(2 - \lambda_1\lambda_2)}$, we obtain

$$
c_1 = \frac{1 - \lambda_1\lambda_2}{2(2 - \lambda_1\lambda_2)}, \qquad c_2 = \frac{(1 - \lambda_1\lambda_2)(2 - \lambda_1\lambda_2)}{\lambda_2^2(3 - \lambda_1\lambda_2)},
$$

so that $c_1 \ge \frac{1}{4}(1 - \lambda_1\lambda_2)$, $c_2 \ge \frac{\beta_1^2}{2\|b\|^2}(1 - \lambda_1\lambda_2)$, $\sigma^{-1} \ge \frac{1}{8}$. Hence, we have $\beta_h \ge \frac{1}{32} \min(\frac{1}{2}, \frac{\beta_1^2}{\|b\|^2})(1 - \lambda_1\lambda_2) \min(\beta_1, \beta_2) \ge \frac{1}{64}(1 - \lambda_1\lambda_2) \frac{\min(\beta_1, \beta_2)^3}{\|b\|^2}$. $\qquad \square$

Remark 55.2 (Inequality $\lambda_1\lambda_2 < 1$). This inequality, which amounts to $b_{12}b_{21} < \beta_1\beta_2$, is trivially satisfied if $b_{12}b_{21} = 0$, which is the case in many applications; see, e.g., Corollary 55.3 below. □

Let us illustrate the above result with the Stokes problem. We set $\boldsymbol{V} := \boldsymbol{H}_0^1(D)$, $Q := L_*^2(D)$, $\|\boldsymbol{v}\|_{\boldsymbol{V}} := |\boldsymbol{v}|_{\boldsymbol{H}^1(D)}$, $\|q\|_Q := \|q\|_{L^2(D)}$, and $b(\boldsymbol{v}, q) := -(\nabla\cdot\boldsymbol{v}, q)_{L^2(D)}$. Let $\mathcal{T}_h$ be a mesh in the sequence $(\mathcal{T}_h)_{h\in\mathcal{H}}$. Let $\mathcal{U}_h$ be a partition of the set $\mathcal{T}_h$. We call $\mathcal{U}_h$ *macroelement partition* and the members of $\mathcal{U}_h$ *macroelements*. For every macroelement $U \in \mathcal{U}_h$, we abuse the notation by writing U also for the set of the points composing the cells in the macroelement U. For all $U \in \mathcal{U}_h$, we define the following spaces:

$$\boldsymbol{V}_{h0}(U) := \{\boldsymbol{v}_h \in \boldsymbol{V}_{h0} \mid \boldsymbol{v}_{h|U} \in \boldsymbol{H}_0^1(U),\ \boldsymbol{v}_{h|D\setminus U} = 0\} \subset \boldsymbol{V}_{h0}, \tag{55.2a}$$

$$Q_h(U) := \{\mathbb{1}_U q_h \mid q_h \in Q_h\}, \tag{55.2b}$$

$$\overline{Q}_h(U) := \mathrm{span}(\mathbb{1}_U), \qquad \widetilde{Q}_h(U) := \{q_h \in Q_h(U) \mid \textstyle\int_U q_h \,\mathrm{d}x = 0\}, \tag{55.2c}$$

where $\mathbb{1}_U$ is the indicator function of U. We additionally define

$$\widetilde{Q}_h := \sum_{U\in\mathcal{U}_h} \widetilde{Q}_h(U), \qquad\qquad \overline{Q}_h := \sum_{U\in\mathcal{U}_h} \overline{Q}_h(U). \tag{55.3}$$

Corollary 55.3 (Macroelement partition). *Assume that for all $h \in \mathcal{H}$, there exists a partition of $\mathcal{T}_h$, say $\mathcal{U}_h$, such that*

$$\forall U \in \mathcal{U}_h, \quad \inf_{q_h\in\widetilde{Q}_h(U)} \sup_{\boldsymbol{v}_h\in\boldsymbol{V}_{h0}(U)} \frac{|b(\boldsymbol{v}_h, q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}}\|q_h\|_Q} =: \beta_{1h}(U) > 0, \tag{55.4a}$$

$$\inf_{q_h\in\overline{Q}_h} \sup_{\boldsymbol{v}_h\in\boldsymbol{V}_{h0}} \frac{|b(\boldsymbol{v}_h, q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}}\|q_h\|_Q} =: \beta_{2h} > 0. \tag{55.4b}$$

(i) *The inf-sup condition* (55.1) *is satisfied.* (ii) *If* $\inf_{h\in\mathcal{H}} \beta_{2h} > 0$ *and* $\inf_{h\in\mathcal{H}} \min_{U\in\mathcal{U}_h} \beta_{1h}(U) > 0$, *the inf-sup condition* (55.1) *holds uniformly w.r.t.* $h \in \mathcal{H}$.

Proof. The idea is to show that the assumptions of Lemma 55.1 are met.
(1) For all $q_h \in Q_h$ and all $U \in \mathcal{U}_h$, let us denote $\overline{q}_{hU} := \frac{1}{|U|}\int_U q_h \,\mathrm{d}x$. The identities $q_h = \sum_{U\in\mathcal{U}_h} \mathbb{1}_U q_h$ and $\mathbb{1}_U q_h = \mathbb{1}_U(q_h - \overline{q}_{hU}) + \overline{q}_{hU}\mathbb{1}_U$ show that $Q_h = Q_h^1 + Q_h^2$, with $Q_h^1 := \widetilde{Q}_h$ and $Q_h^2 := \overline{Q}_h$. Notice that this decomposition holds true whether Q_h is composed of discontinuous functions or not.
(2) Let us prove the first inf-sup condition from Lemma 55.1. Let $q_h \in Q_h^1 = \widetilde{Q}_h$. Then (55.4a) implies that for all $U \in \mathcal{U}_h$ there is $\boldsymbol{v}_h(U) \in \boldsymbol{V}_{h0}(U)$ s.t. $\nabla\cdot(\boldsymbol{v}_h(U)) = \mathbb{1}_U q_h$ and $\beta_{1h}(U)\|\boldsymbol{v}_h(U)\|_{\boldsymbol{V}} \le \|\mathbb{1}_U q_h\|_Q = \|q_h\|_{L^2(U)}$. Set $\boldsymbol{v}_h := \sum_{U\in\mathcal{U}_h} \boldsymbol{v}_h(U) \in \boldsymbol{V}_{h0}^1 := \sum_{U\in\mathcal{U}_h} \boldsymbol{V}_{h0}(U)$. Notice that $\boldsymbol{V}_{h0}^1 \subset \boldsymbol{V}_{h0}$ by construction. Using that $\left(\sum_{U\in\mathcal{U}_h} \|\boldsymbol{v}_h(U)\|_{\boldsymbol{V}}^2\right)^{\frac{1}{2}} = \|\boldsymbol{v}_h\|_{\boldsymbol{V}}$, we infer that

$$\int_D q_h \nabla\cdot\boldsymbol{v}_h \,\mathrm{d}x = \sum_{U\in\mathcal{U}_h} \int_U q_h \nabla\cdot\boldsymbol{v}_h(U) \,\mathrm{d}x = \sum_{U\in\mathcal{U}_h} \|q_h\|_{L^2(U)}^2$$

$$= \|q_h\|_{L^2(D)} \left(\sum_{U\in\mathcal{U}_h} \|q_h\|_{L^2(U)}^2\right)^{\frac{1}{2}} \ge \|q_h\|_Q \left(\sum_{U\in\mathcal{U}_h} (\beta_{1h}(U))^2 \|\boldsymbol{v}_h(U)\|_{\boldsymbol{V}}^2\right)^{\frac{1}{2}}$$

$$\ge \beta_{1h}\|q_h\|_Q \left(\sum_{U\in\mathcal{U}_h} \|\boldsymbol{v}_h(U)\|_{\boldsymbol{V}}^2\right)^{\frac{1}{2}} = \beta_{1h}\|q_h\|_Q\|\boldsymbol{v}_h\|_{\boldsymbol{V}},$$

$\beta_{1h} := \min_{U \in \mathcal{U}_h} \beta_{1h}(U) > 0$. Hence, $\inf_{q_h \in Q_h^1} \sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}^1} \frac{|b(\boldsymbol{v}_h, q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}} \|q_h\|_Q} \geq \beta_{1h}$.

(3) The second inf-sup condition from Lemma 55.1 holds by assumption with $\boldsymbol{V}_{h0}^2 := \boldsymbol{V}_{h0}$, $Q_h^2 := \overline{Q}_h$, and the constant $\beta_{2h} > 0$.

(4) Finally, let us verify the last assumption by showing that $\lambda_1 \lambda_2 := \frac{b_{12} b_{21}}{\beta_{1h} \beta_{2h}} = 0 < 1$. Let $\boldsymbol{v}_h := \sum_{U \in \mathcal{U}_h} \boldsymbol{v}_h(U) \in \boldsymbol{V}_{h0}^1$ and $q_h := \sum_{U \in \mathcal{U}_h} q_U \mathbb{1}_U \in Q_h^2$. We obtain

$$b(\boldsymbol{v}_h, q_h) = \sum_{U \in \mathcal{U}_h} q_U \int_U \nabla \cdot \boldsymbol{v}_h(U) \, \mathrm{d}x = 0,$$

since $\boldsymbol{v}_h(U) \in \boldsymbol{H}_0^1(U)$ implies that $\int_U \nabla \cdot \boldsymbol{v}_h(U) \, \mathrm{d}x = 0$ for all $U \in \mathcal{U}_h$. Hence, $b_{21} = 0$. This completes the proof. $\square$

Remark 55.4 (Assumption (55.4a)**).** For all $q_h \in Q_h$, let $\overline{q}_h \in \overline{Q}_h$ be defined s.t. $\overline{q}_{h|U} := \overline{q}_{hU} := \frac{1}{|U|} \int_U q_h \, \mathrm{d}x$ for all $U \in \mathcal{U}_h$. Since $\int_U q_h \nabla \cdot \boldsymbol{v}_h \, \mathrm{d}x = \int_U (q_h - \overline{q}_{hU}) \nabla \cdot \boldsymbol{v}_h \, \mathrm{d}x$ for all $\boldsymbol{v}_h \in \boldsymbol{V}_{h0}(U)$ and all $U \in \mathcal{U}_h$, the assumption (55.4a) means that for all $q_h \in Q_h$, we have $\sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}(U)} \frac{|b(\boldsymbol{v}_h, q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}}} \geq \beta_{1h}(U) \|q_{h|U} - \overline{q}_{hU}\|_Q$. Then the argument in Step (2) of the proof of Corollary 55.3 shows that $\sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}^1} \frac{|b(\boldsymbol{v}_h, q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}}} \geq \beta_{1h} \|q_h - \overline{q}_h\|_Q$ for all $q_h \in Q_h$, where we have set $\beta_{1h} := \min_{U \in \mathcal{U}_h} \beta_{1h}(U)$. $\square$

Notice that $\widetilde{Q}_h \subset Q_h$ and $\overline{Q}_h \subset Q_h$ when Q_h is composed of discontinuous functions, but the above theory does not require that Q_h be composed of discontinuous finite elements. It turns out that the assumption (55.4b) can be relaxed if Q_h is H^1-conforming.

Proposition 55.5 (Macroelement, continuous pressures). *Let* $(\mathcal{T}_h)_{h \in \mathcal{H}}$ *be a shape-regular mesh sequence. Assume that there exists a macroelement partition* $\mathcal{U}_h$ *for every mesh* $\mathcal{T}_h$. *Assume that every* $U \in \mathcal{U}_h$ *can be mapped by an affine mapping to a reference set* $\widehat{U}$ *and that the sequence* $\{\mathcal{U}_h\}_{h \in \mathcal{H}}$ *is shape-regular. Assume that* $\inf_{h \in \mathcal{H}} \max_{U \in \mathcal{U}_h} \operatorname{card}\{K \subset U\} < \infty$. *Assume that* $Q_h \subset H^1(D) \cap L_*^2(D)$ *and that the following holds true that for all* $h \in \mathcal{H}$:

$$\forall U \in \mathcal{U}_h, \quad \inf_{q_h \in \widetilde{Q}_h(U)} \sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}(U)} \frac{|b(\boldsymbol{v}_h, q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}} \|q_h\|_Q} =: \beta_{1h}(U) > 0. \tag{55.5}$$

(i) *The inf-sup condition* (55.1) *is satisfied.* (ii) *If* $\inf_{h \in \mathcal{H}} \min_{U \in \mathcal{U}_h} \beta_{1h}(U) > 0$, *the inf-sup condition* (55.1) *holds uniformly w.r.t.* $h \in \mathcal{H}$.

Proof. See Brezzi and Bathe [91, Prop. 4.1] and Exercise 55.7. $\square$

Remark 55.6 (Literature). Macroelement techniques have been introduced in a series of works by Boland and Nicolaides [67], Girault and Raviart [217, §II.1.4], Stenberg [352, 354, 353]. This theory is further refined in Qin [328, Chap. 3]. In particular, Lemma 55.1 is established in [328, Thm. 3.4.1]. It is possible to generalize the macroelement technique to situations where the macroelements are not disjoint provided one assumes that each cell K belongs to a finite set of macroelements with cardinality bounded from above uniformly w.r.t. $h \in \mathcal{H}$. This type of technique can be used in particular to prove the stability of the generalized Taylor–Hood elements $(\mathbb{P}_k, \mathbb{P}_{k-1})$, $(\mathbb{Q}_k, \mathbb{Q}_{k-1})$, $k \geq 2$. We refer the reader to Boffi et al. [65, §8.8] for a thorough discussion on this topic. $\square$

55.2 Discontinuous pressures and bubbles

We investigate in this section finite element pairs based on simplicial meshes. The pressure approximation is discontinuous and stability is achieved by enriching the velocity space.

55.2.1 Discontinuous pressures

Since the functional space for the pressure is $Q := L^2_*(D)$, the approximation setting remains conforming for the pressure. The discrete pressure space is typically the broken polynomial space (see §18.1.2)

$$P^{\mathrm{b}}_{l,*}(\mathcal{T}_h) := \{q_h \in L^2_*(D) \mid \forall K \in \mathcal{T}_h,\, q_h \circ \boldsymbol{T}_K \in \mathbb{P}_{l,d}\}, \tag{55.6}$$

for some $l \in \mathbb{N}$ and where $\boldsymbol{T}_K : \widehat{K} \to K$ is the geometric mapping. The $(\mathbb{P}_k, \mathbb{P}^{\mathrm{b}}_l)$ pair refers to the choice of finite element space $\boldsymbol{V}_{h0} := \boldsymbol{P}^{\mathrm{g}}_{k,0}(\mathcal{T}_h)$ for the velocity and $Q_h := P^{\mathrm{b}}_{l,*}(\mathcal{T}_h)$ for the pressure. The stable finite element pairs investigated herein are the $(\mathbb{P}_2, \mathbb{P}^{\mathrm{b}}_0)$ and the $(\mathbb{P}_2\text{-bubble}, \mathbb{P}^{\mathrm{b}}_1)$ pairs.

Remark 55.7 (Local mass balance). Working with discontinuous pressures is interesting since it becomes possible to test the discrete mass conservation equation against a function supported in a single mesh cell $K \in \mathcal{T}_h$. This leads to the local mass balance $\int_K (\psi^{\mathrm{g}}_K)^{-1}(q)\nabla{\cdot}\boldsymbol{u}_h \,\mathrm{d}x = \int_K (\psi^{\mathrm{g}}_K)^{-1}(q)g \,\mathrm{d}x$ for all $q \in \mathbb{P}_{k,d}$ with $\psi^{\mathrm{g}}_K(q) := q \circ \boldsymbol{T}_K$, see Exercise 55.1. $\qquad\square$

55.2.2 The $(\mathbb{P}_2, \mathbb{P}^{\mathrm{b}}_0)$ pair

Let $(\mathcal{T}_h)_{h\in\mathcal{H}}$ be a shape-regular family of affine simplicial meshes. Recalling that we are enforcing homogeneous Dirichlet conditions on the velocity, the $(\mathbb{P}_2, \mathbb{P}^{\mathrm{b}}_0)$ pair gives to the following approximation spaces:

$$\boldsymbol{V}_{h0} := \boldsymbol{P}^{\mathrm{g}}_{2,0}(\mathcal{T}_h), \qquad Q_h := P^{\mathrm{b}}_{0,*}(\mathcal{T}_h). \tag{55.7}$$

This simple finite element pair satisfies the inf-sup condition (55.1) uniformly w.r.t. $h \in \mathcal{H}$ in dimension two, but it has little practical interest since it is does not provide optimal convergence results. Nevertheless it is an important building block for other more useful finite element pairs. Let $\boldsymbol{V} := \boldsymbol{W}^{1,p}_0(D)$ be equipped with the norm $\|\boldsymbol{v}\|_{\boldsymbol{V}} := |\boldsymbol{v}|_{\boldsymbol{W}^{1,p}(D)}$ and let $Q := L^{p'}_*(D)$ be equipped with the norm $\|q\|_Q := \|q\|_{L^{p'}(D)}$, where $p, p' \in (1, \infty)$ are s.t. $\frac{1}{p} + \frac{1}{p'} = 1$.

Lemma 55.8 (Stability). *Assume that $d = 2$. The $(\mathbb{P}_2, \mathbb{P}^{\mathrm{b}}_0)$ pair satisfies the inf-sup condition (55.1) uniformly w.r.t. $h \in \mathcal{H}$.*

Proof. We construct a Fortin operator by using the decomposition defined in Lemma 54.2 and by invoking Lemma 54.1 to conclude. The operator $\boldsymbol{\Pi}_{2h} : \boldsymbol{V} \to \boldsymbol{V}_{h0}$ is defined as follows. Let $\boldsymbol{v} \in \boldsymbol{V}_{h0}$. We set $\boldsymbol{\Pi}_{2h}(\boldsymbol{v})(\boldsymbol{z}) := \boldsymbol{0}$ for all $\boldsymbol{z} \in \mathcal{V}^{\circ}_h$ ($\mathcal{V}^{\circ}_h$ is the collection of the internal vertices of the mesh), and $\boldsymbol{\Pi}_{2h}(\boldsymbol{v})(\boldsymbol{m}_F) := \frac{3}{2|F|}\int_F \boldsymbol{v} \,\mathrm{d}s$ for all $F \in \mathcal{F}^{\circ}_h$ ($\mathcal{F}^{\circ}_h$ is the collection of the mesh interfaces), where $\boldsymbol{m}_F$ is the barycenter of F. This entirely defines $\boldsymbol{\Pi}_{2h}(\boldsymbol{v})$ in $\boldsymbol{V}_{h0}$ since $d = 2$. Notice that $\boldsymbol{v}_{|F} \in \boldsymbol{L}^1(F)$ for all $\boldsymbol{v} \in \boldsymbol{W}^{1,p}_0(D)$ and all $F \in \mathcal{F}^{\circ}_h$ so that the above construction is meaningful. Then we set $\boldsymbol{\Pi}_{1h} := \boldsymbol{\mathcal{I}}^{\mathrm{av}}_{h0}$, where $\boldsymbol{\mathcal{I}}^{\mathrm{av}}_{h0}$ is the $\mathbb{R}^d$-valued version of the $W^{1,p}_0$-conforming quasi-interpolation operator introduced in §22.4.2. This means that $\boldsymbol{\mathcal{I}}^{\mathrm{av}}_{h0}(\boldsymbol{v}) := \sum_{i\in\{1:d\}} \mathcal{I}^{\mathrm{av}}_{h0}(v_i)\boldsymbol{e}_i$, where $\boldsymbol{v} := \sum_{i\in\{1:d\}} v_i\boldsymbol{e}_i$ and $\{\boldsymbol{e}_i\}_{i\in\{1:d\}}$ is the canonical Cartesian basis of $\mathbb{R}^d$. The rest of the proof consists of verifying that the assumptions (i)–(iii) from Lemma 54.2 are met; see Exercise 55.2. $\qquad\square$

Remark 55.9 (Literature). The reader is referred to Boffi et al. [65, §8.4.3] for other details on the $(\mathbb{P}_2, \mathbb{P}^{\mathrm{b}}_0)$ pair. In general, this pair is not stable in dimension 3, but it is shown in Zhang and Zhang [404] that one can construct special families of tetrahedral meshes for which stability holds. $\qquad\square$

55.2.3   The $(\mathbb{P}_2\text{-bubble}, \mathbb{P}_1^{\mathrm{b}})$ pair

Let $\widehat{b}$ be the bubble function defined in (54.5) and $\widehat{\boldsymbol{P}} := \mathbb{P}_{2,d} \oplus (\mathrm{span}\{\widehat{b}\})^d$. Let $(\mathcal{T}_h)_{h\in\mathcal{H}}$ be a shape-regular family of affine simplicial meshes. Recalling that we are enforcing homogeneous Dirichlet conditions on the velocity, the $(\mathbb{P}_2\text{-bubble}, \mathbb{P}_1^{\mathrm{b}})$ pair gives the following approximation spaces:

$$\boldsymbol{V}_{h0} := \boldsymbol{P}_{2,0}^{\mathrm{g}}(\mathcal{T}_h) \oplus \boldsymbol{B}_h, \qquad Q_h := P_{1,*}^{\mathrm{b}}(\mathcal{T}_h), \tag{55.8}$$

with $\boldsymbol{B}_h := \bigoplus_{K\in\mathcal{T}_h} (\mathrm{span}\{b_K\})^d$ and $b_K := \widehat{b} \circ \boldsymbol{T}_K$ is the bubble function associated with the mesh cell $K \in \mathcal{T}_h$. Notice that

$$\boldsymbol{V}_{h0} := \{\boldsymbol{v}_h \in \boldsymbol{C}^0(\overline{D}) \mid \forall K \in \mathcal{T}_h,\ \boldsymbol{v}_h \circ \boldsymbol{T}_K \in \widehat{\boldsymbol{P}},\ \boldsymbol{v}_{h|\partial D} = \boldsymbol{0}\}. \tag{55.9}$$

Since the pressure is locally $\mathbb{P}_1$ on each simplex and globally discontinuous, its local degrees of freedom can be taken to be its mean value and its gradient in each mesh cell. A conventional representation is shown in Figure 55.1. We have the following result (see Boffi et al. [65, p. 488]).

Proposition 55.10 ($\mathbb{P}_2\text{-bubble}, \mathbb{P}_1^{\mathrm{b}}$). *The $(\mathbb{P}_2\text{-bubble}, \mathbb{P}_1^{\mathrm{b}})$ pair satisfies the inf-sup condition (55.1) uniformly w.r.t. $h \in \mathcal{H}$. Moreover, this pair leads to the same error estimates as the Taylor–Hood element, that is, $\mu|\boldsymbol{u} - \boldsymbol{u}_h|_{\boldsymbol{H}^1(D)} + \|p - p_h\|_{L^2(D)} \leq ch^2(\mu|\boldsymbol{u}|_{\boldsymbol{H}^3(D)} + |p|_{H^2(D)})$, and if the assumptions of Theorem 53.19 are met for some $s \in (0,1]$, then $\mu\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{L}^2(D)} \leq ch^{2+s}\ell_D^{1-s}(\mu|\boldsymbol{u}|_{\boldsymbol{H}^3(D)} + |p|_{H^2(D)})$.*
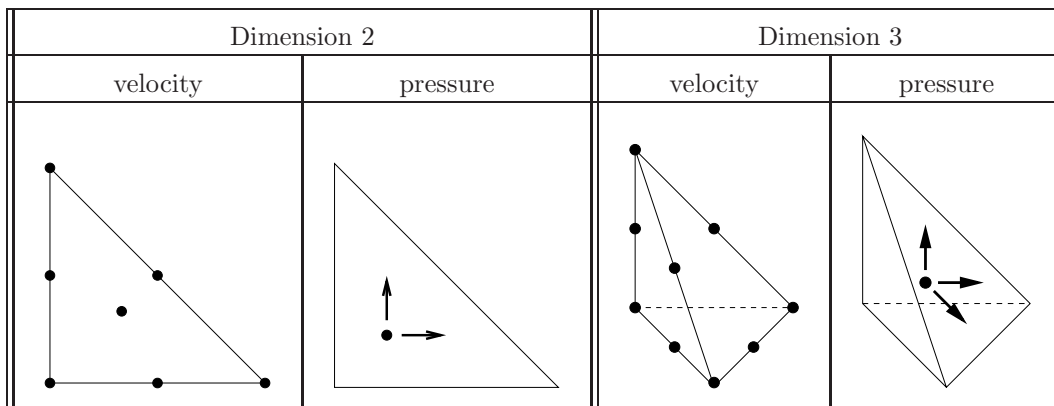
Figure 55.1: Conventional representation of the $(\mathbb{P}_2\text{-bubble}, \mathbb{P}_1^{\mathrm{b}})$ pair in dimensions two (left) and three (right, only visible degrees of freedom of the velocity are shown). Among various possibilities, the degrees of freedom for the pressure here are the mean value (indicated by a dot) and the d components of the gradient (indicated by arrows).

Remark 55.11 (Literature). The $(\mathbb{P}_2\text{-bubble}, \mathbb{P}_1^{\mathrm{b}})$ pair is also called conforming *Crouzeix–Raviart mixed finite element* [151]. $\square$

55.3   Scott–Vogelius elements and generalizations

Let $k \geq 1$. The $(\mathbb{P}_k, \mathbb{P}_{k-1}^{\mathrm{b}})$ pair is interesting since $\nabla\cdot\boldsymbol{P}_{k,0}^{\mathrm{g}}(\mathcal{T}_h) \subset P_{k-1,*}^{\mathrm{b}}(\mathcal{T}_h)$, which implies that any vector field in $\boldsymbol{P}_{k,0}^{\mathrm{g}}(\mathcal{T}_h)$ whose divergence is L^2-orthogonal to $P_{k-1,*}^{\mathrm{b}}(\mathcal{T}_h)$ is exactly divergence-free.

55.3.1 Special meshes

In general, the $(\mathbb{P}_k, \mathbb{P}_{k-1}^{\mathrm{b}})$ pair does not satisfy the inf-sup condition (55.1) (e.g., we have seen in §53.4.3 that for $k = 1$, this pair suffers from locking). However, it is possible to construct special meshes so that this element satisfies the inf-sup condition (55.1) uniformly w.r.t. $h \in \mathcal{H}$ for some k. Let us now introduce some special meshes to substantiate this claim. Various two-dimensional examples of such meshes are shown in Figure 55.2.

Irregular crisscross: A two-dimensional triangulation $\mathcal{T}_h$ is said to be an *irregular crisscross* mesh if it is obtained from a matching mesh of $D \subset \mathbb{R}^2$ composed of quadrangles, where each quadrilateral cell is divided along its two diagonals; see the leftmost panel in Figure 55.2.

Simplicial barycentric $(d+1)$-sected: We say that $\mathcal{T}_h$ is a *simplicial barycentric $(d+1)$-sected* mesh in $\mathbb{R}^d$ if $\mathcal{T}_h$ is obtained after refinement of a simplicial matching mesh by subdividing each initial simplex into $(d+1)$ sub-simplices by connecting the barycenter with the $(d+1)$ vertices. Simplicial barycentric $(d+1)$-sected meshes are also called *Hsieh–Clough–Tocher (HCT)* meshes in the literature; see the second panel from the left in Figure 55.2.

Twice quadrisected crisscrossed: We say that a two-dimensional triangulation $\mathcal{T}_h$ is *twice quadrisected crisscrossed* if it is formed as follows. First, the polygon D is partitioned into a matching mesh of quadrangles, say $\mathcal{Q}_{4h}$. Then, each quadrangle in $\mathcal{Q}_{4h}$ is divided into four new quadrangles by connecting the point at the intersection of its two diagonals with the midpoint on each of its edges. The mesh Q_{2h} thus formed is subdivided once more by repeating this process. Finally, $\mathcal{T}_h$ is obtained by dividing each quadrangle in $\mathcal{Q}_h$ along its two diagonals, thereby giving 4 triangles per quadrangular cell in $\mathcal{Q}_h$, or 64 triangles for each quadrangle in Q_{4h}; see the third panel from the left in Figure 55.2.

Powell–Sabin: A simplicial mesh of a polygon or polyhedron D is said to be a *Powell–Sabin* mesh if it is constructed as follows. For instance, assuming that the space dimension is two, let $\mathcal{T}_h$ be an affine simplicial matching mesh of D. For each triangle $K \in \mathcal{T}_h$, let $\boldsymbol{c}_K$ be the center of the inscribed circle of K and assume that $\boldsymbol{c}_K \in K$ for all $K \in \mathcal{T}_h$. We then divide K into three triangles by connecting $\boldsymbol{c}_K$ to the three vertices of K (this is similar to an HCT triangulation). Each of the newly created triangles is divided again by connecting $\boldsymbol{c}_K$ to $\boldsymbol{c}_{K_1}$, $\boldsymbol{c}_{K_2}$, and $\boldsymbol{c}_{K_3}$, where K_1, K_2, and K_3 are the three neighbors of K (or $\boldsymbol{c}_K$ is connected to the midpoint of the edge if the corresponding neighbor does not exist). The same construction can be done in $\mathbb{R}^d$ as shown in Zhang [403, Fig. 1]. This construction is illustrated in the rightmost panel in Figure 55.2.
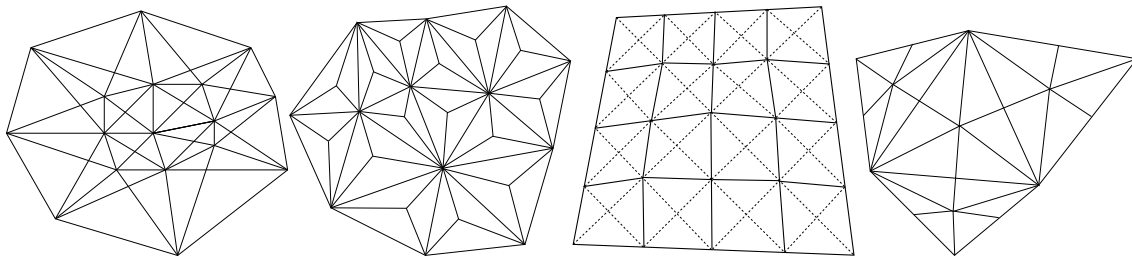
Figure 55.2: Irregular crisscross mesh (left). Simplicial barycentric trisected mesh also called Hsieh–Clough–Tocher (HCT) mesh (center left). One quadrangular cell that is twice quadrisected and crisscrossed (center right). Powell–Sabin mesh (right).

55.3.2   Stable $(\mathbb{P}_k, \mathbb{P}^{\mathrm{b}}_{k-1})$ pairs on special meshes

The stability of the $(\mathbb{P}_k, \mathbb{P}^{\mathrm{b}}_{k-1})$ pair has been thoroughly investigated in dimension two by Scott and Vogelius [345].

Lemma 55.12 ($(\mathbb{P}_k, \mathbb{P}^{\mathrm{b}}_{k-1})$, $k \geq 4$, $d = 2$). *Let $d = 2$ and $k \geq 4$. Assume that the mesh sequence $(\mathcal{T}_h)_{h \in \mathcal{H}}$ is quasi-uniform. Assume also that any pair of edges meeting at an internal vertex does not form a straight line. (An internal vertex violating this property is called singular vertex; see Exercise 54.3.) The $(\mathbb{P}_k, \mathbb{P}^{\mathrm{b}}_{k-1})$ pair satisfies the inf-sup condition (55.1) uniformly w.r.t. $h \in \mathcal{H}$.*

Proof. See [345, Thm. 5.1].   □

There are extensions of the above result to the $(\mathbb{P}_3, \mathbb{P}^{\mathrm{b}}_2)$ pair, the $(\mathbb{P}_2, \mathbb{P}^{\mathrm{b}}_1)$ pair, and the $(\mathbb{P}_1, \mathbb{P}^{\mathrm{b}}_0)$ pair in dimension two on some of the special meshes described above; see Qin [328].

Lemma 55.13 (Crisscross meshes, $k \in \{2, 3\}$, $d = 2$). *Let $(\mathcal{T}_h)_{h \in \mathcal{H}}$ be a shape-regular sequence of irregular crisscross meshes. Then the $(\mathbb{P}_2, \mathbb{P}^{\mathrm{b}}_1)$ pair and the $(\mathbb{P}_3, \mathbb{P}^{\mathrm{b}}_2)$ pair have as many spurious pressure modes as singular vertices, but the velocity approximation is optimal, and the pressure approximation in the L^2-orthogonal complement to the spurious modes is optimal.*

Proof. See [328, Thm. 4.3.1 & 6.2.1].   □

Lemma 55.14 (HCT meshes, $k \in \{2, 3\}$, $d = 2$). *Let $(\mathcal{T}_h)_{h \in \mathcal{H}}$ be a shape-regular sequence of barycentric trisected triangulations. Then the $(\mathbb{P}_2, \mathbb{P}^{\mathrm{b}}_1)$ pair and the $(\mathbb{P}_3, \mathbb{P}^{\mathrm{b}}_2)$ pair satisfy the inf-sup condition (55.1) uniformly w.r.t. $h \in \mathcal{H}$, and therefore lead to optimal error estimates.*

Proof. These statements are proved in Qin [328, Thm. 4.6.1 & 6.4.1]. We detail the proof for the $(\mathbb{P}_2, \mathbb{P}^{\mathrm{b}}_1)$ pair since it illustrates the use of the macroelement technique from Corollary 55.3. Here, $\boldsymbol{V}_{h0} := \mathbb{P}^{\mathrm{g}}_{2,0}(\mathcal{T}_h)$ and $Q_h := \mathbb{P}^{\mathrm{b}}_{1,*}(\mathcal{T}_h)$.
(1) Let $(\mathcal{U}_h)_{h \in \mathcal{H}}$ be the sequence of triangulations that is used to create $(\mathcal{T}_h)_{h \in \mathcal{H}}$ by barycentric trisection. For every triangle $U \in \mathcal{U}_h$, we consider the spaces $\boldsymbol{V}_{h0}(U)$, $Q_h(U)$, $\overline{Q}_h(U)$, and $\widetilde{Q}_h(U)$ defined in (55.2). We also consider the spaces $\widetilde{Q}_h$, $\overline{Q}_h$ defined in (55.3). We are going to prove the inf-sup conditions (55.4a) and (55.4b) in Corollary 55.3.
(2) Proof of (55.4b). We have $\overline{Q}_h := \sum_{U \in \mathcal{U}_h} \overline{Q}_h(U) = P^{\mathrm{b}}_{0,*}(\mathcal{U}_h)$. Since, as established in Lemma 55.8, the $(\boldsymbol{P}^{\mathrm{g}}_{2,0}(\mathcal{U}_h), P^{\mathrm{b}}_{0,*}(\mathcal{U}_h))$ pair satisfies an inf-sup condition uniformly w.r.t. $h \in \mathcal{H}$, and $\boldsymbol{P}^{\mathrm{g}}_{2,0}(\mathcal{U}_h) \subset \boldsymbol{P}^{\mathrm{g}}_{2,0}(\mathcal{T}_h) =: \boldsymbol{V}_{h0}$, we infer that the inf-sup condition (55.4b) is satisfied uniformly w.r.t. $h \in \mathcal{H}$.
(3) Proof of (55.4a). Let $\widehat{U}$ be the reference simplex in $\mathbb{R}^2$. For every $U \in \mathcal{U}_h$, let $\boldsymbol{T}_U : \widehat{U} \to U$ be the corresponding affine geometric mapping. Let us set

$$
\begin{aligned}
\boldsymbol{V}(\widehat{U}) &:= \{\boldsymbol{\psi}^{\mathrm{d}}_U(\boldsymbol{v}_h) \mid \boldsymbol{v}_h \in \boldsymbol{V}_{h0}(U)\}, \\
Q(\widehat{U}) &:= \{\psi^{\mathrm{g}}_U(q_h) \mid q_h \in Q_h(U)\}, \qquad \widetilde{Q}(\widehat{U}) := \{\psi^{\mathrm{g}}_U(q_h) \mid q_h \in \widetilde{Q}_h(U)\},
\end{aligned}
$$

where ψ^{g}_U is the pullback by $\boldsymbol{T}_U$ and $\boldsymbol{\psi}^{\mathrm{d}}_U$ is the contravariant Piola transformation, i.e., $\psi^{\mathrm{g}}_U(q) := q \circ \boldsymbol{T}_U$ and $\boldsymbol{\psi}^{\mathrm{d}}_U(\boldsymbol{v}) := \det(\mathbb{J}_U)\mathbb{J}^{-1}_U(\boldsymbol{v} \circ \boldsymbol{T}_U)$ (see Definition 9.8). One can verify that both spaces $\boldsymbol{V}(\widehat{U})$ and $\widetilde{Q}(\widehat{U})$ are 8-dimensional, whereas the space $Q(\widehat{U})$ is 9-dimensional. Let $\widehat{B} : \boldsymbol{V}(\widehat{U}) \to Q(\widehat{U})$ be defined by $(\widehat{B}(\widehat{\boldsymbol{v}}), \widehat{q})_{L^2(\widehat{U})} = \int_{\widehat{U}} \widehat{q}(\widehat{\boldsymbol{x}}) \nabla \cdot \widehat{\boldsymbol{v}}(\widehat{\boldsymbol{x}}) \, \mathrm{d}\widehat{x}$ for all $(\widehat{\boldsymbol{v}}, \widehat{q}) \in \boldsymbol{V}(\widehat{U}) \times Q(\widehat{U})$. A lengthy but straightforward computation (see Exercise 55.5) shows that $\mathrm{im}(\widehat{B})^\perp = \mathrm{span}(\mathbb{1}_{\widehat{U}})$, where $\perp$ means the L^2-orthogonal complement in $Q(\widehat{U})$. Since $\widetilde{Q}(\widehat{U}) = (\mathrm{span}(\mathbb{1}_{\widehat{U}}))^\perp$, this result implies that

$\widehat{B} : \boldsymbol{V}(\widehat{U}) \to \widetilde{Q}(\widehat{U})$ is surjective. (Actually, $\widehat{B}$ is bijective since $\dim(\boldsymbol{V}(\widehat{U})) = \dim(\widetilde{Q}(\widehat{U}))$.) Hence, we have

$$\inf_{\widehat{q} \in \widetilde{Q}(\widehat{U})} \sup_{\widehat{\boldsymbol{v}} \in \boldsymbol{V}(\widehat{U})} \frac{|\int_{\widehat{U}} \widehat{q}(\widehat{\boldsymbol{x}}) \nabla \cdot \widehat{\boldsymbol{v}}(\widehat{\boldsymbol{x}}) \, \mathrm{d}\widehat{x}|}{\|\widehat{q}\|_{Q(\widehat{U})} \|\widehat{\boldsymbol{v}}\|_{\boldsymbol{V}(\widehat{U})}} =: \widehat{\beta}_1 > 0,$$

with $\|\boldsymbol{v}\|_{\boldsymbol{V}(\widehat{U})} := |\widehat{\boldsymbol{v}}|_{\boldsymbol{H}^1(\widehat{U})}$ and $\|\widehat{q}\|_{Q(\widehat{U})} := \|\widehat{q}\|_{L^2(\widehat{U})}$. Using the scaling inequality (11.7b) and the regularity of the mesh sequence $(\mathcal{U}_h)_{h \in \mathcal{H}}$, we infer that there is $c_1 > 0$ s.t. $c_1 \|\boldsymbol{v}\|_{\boldsymbol{V}} \|q\|_Q \leq \|\widehat{\boldsymbol{v}}\|_{\boldsymbol{V}(\widehat{U})} \|\widehat{q}\|_{Q(\widehat{U})}$ for all $\boldsymbol{v} \in \boldsymbol{V}_{h0}(U)$, all $q \in Q_h(U)$, all $U \in \mathcal{U}_h$, and all $h \in \mathcal{H}$. Observing that $\int_{\widehat{U}} \widehat{q}(\widehat{\boldsymbol{x}}) \nabla \cdot \widehat{\boldsymbol{v}}(\widehat{\boldsymbol{x}}) \, \mathrm{d}\widehat{x} = \int_U q(\boldsymbol{x}) \nabla \cdot \boldsymbol{v}(\boldsymbol{x}) \, \mathrm{d}x$ (see Exercise 14.3(i)), we infer that

$$\inf_{q \in \widetilde{Q}(U)} \sup_{\boldsymbol{v} \in \boldsymbol{V}_{h0}(U)} \frac{|\int_U q(\boldsymbol{x}) \nabla \cdot \boldsymbol{v}(\boldsymbol{x}) \, \mathrm{d}x|}{\|q\|_{Q_h} \|\boldsymbol{v}\|_{\boldsymbol{V}}} =: \beta_1 \geq c_1 \widehat{\beta}_2 > 0, \tag{55.10}$$

i.e., the inf-sup condition (55.4a) is satisfied uniformly w.r.t. $h \in \mathcal{H}$. $\qquad \square$

The analysis of the $(\mathbb{P}_1, \mathbb{P}_0^{\mathrm{b}})$ pair is a little bit more subtle since filtering the spurious pressure modes is not enough to approximate the velocity and the pressure properly on general meshes, but filtering is sufficient on twice quadrisected crisscrossed meshes or Powell–Sabin meshes.

Lemma 55.15 ($(\mathbb{P}_1, \mathbb{P}_0^{\mathrm{b}})$). *Let $(\mathcal{T}_h)_{h \in \mathcal{H}}$ be a shape-regular mesh sequence of either twice quadrisected crisscrossed meshes or Powell–Sabin meshes. Then the $(\mathbb{P}_1, \mathbb{P}_0^{\mathrm{b}})$ pair optimally approximates the velocity of the Stokes problem (i.e., first-order in the $\boldsymbol{H}^1$-seminorm) and the approximation of the pressure is optimal as well after post-processing the spurious pressure modes.*

Proof. See Qin [328, Thm. 7.4.2], Zhang [402]. $\qquad \square$

Three-dimensional extensions of the above results are available.

Lemma 55.16 ($(\mathbb{P}_k, \mathbb{P}_{k-1}^{\mathrm{b}})$, $d = 3$). *Let $(\mathcal{T}_h)_{h \in \mathcal{H}}$ be a shape-regular sequence of simplicial barycentric quadrisected meshes in $\mathbb{R}^3$. The $(\mathbb{P}_k, \mathbb{P}_{k-1}^{\mathrm{b}})$ pair is uniformly stable for all $k \geq 3$.*

Proof. See Zhang [401, Thm. 5]. $\qquad \square$

Lemma 55.17 ($(\mathbb{P}_2, \mathbb{P}_1^{\mathrm{b}})$, $d = 3$). *Let $(\mathcal{T}_h)_{h \in \mathcal{H}}$ be a shape-regular sequence of Powell–Sabin simplicial meshes in $\mathbb{R}^3$. The $(\mathbb{P}_2, \mathbb{P}_1^{\mathrm{b}})$ pair optimally approximates the velocity and after post-processing the spurious modes, the approximation of the pressure is optimal as well.*

Proof. See Zhang [403, Thm. 4.1]. $\qquad \square$

55.4 Nonconforming and hybrid methods

In this section, we review some nonconforming and some hybrid discretization methods. Let us start with a nonconforming approximation technique based on the Crouzeix–Raviart finite element studied in Chapter 36. Let $(\mathcal{T}_h)_{h \in \mathcal{H}}$ be a shape-regular sequence of affine simplicial meshes. Let $P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$ be the Crouzeix–Raviart finite element space with homogeneous Dirichlet conditions (see (36.8)). Recall that $P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$ is composed of piecewise affine functions with continuous mean value across the mesh interfaces and zero mean value at the boundary faces. The $(\mathbb{P}_1^{\mathrm{CR}}, \mathbb{P}_0^{\mathrm{b}})$ pair gives the following approximation spaces:

$$\boldsymbol{V}_{h0} := \boldsymbol{P}_{1,0}^{\mathrm{CR}}(\mathcal{T}_h), \qquad Q_h := P_{0,*}^{\mathrm{b}}(\mathcal{T}_h), \tag{55.11}$$

where $\boldsymbol{P}_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$ is composed of vector-valued functions with each Cartesian component in $P_{1,0}^{\mathrm{CR}}(\mathcal{T}_h)$. Observe that $\boldsymbol{V}_{h0}$ is nonconforming in $\boldsymbol{W}_0^{1,p}(D)$. The conventional representation of the $(\mathbb{P}_1^{\mathrm{CR}}, \mathbb{P}_0^{\mathrm{b}})$ pair is shown in Figure 55.3.

Dimension 2		Dimension 3	
velocity	pressure	velocity	pressure
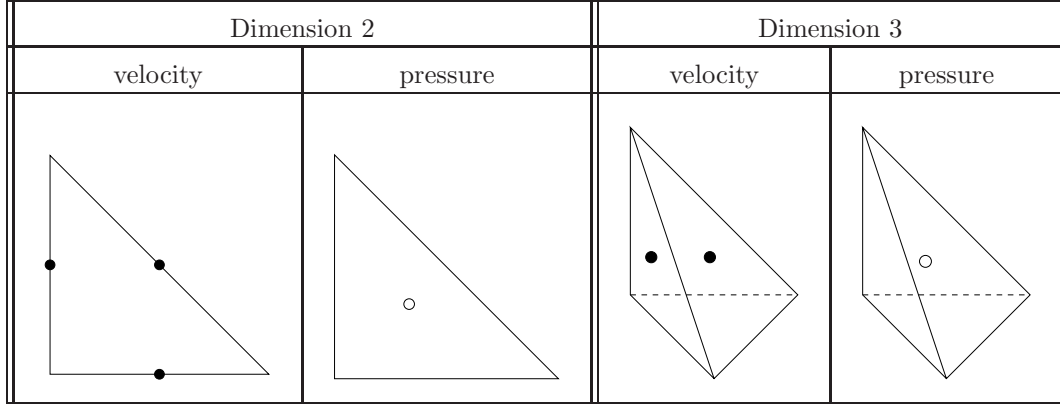
Figure 55.3: Conventional representation of the $(\mathbb{P}_1^{\mathrm{CR}}, \mathbb{P}_0^{\mathrm{b}})$ pair in dimensions two (left) and three (right, only visible velocity degrees of freedom are shown). The pressure degree of freedom is the average over each mesh cell.

To avoid technicalities related to the discrete version of Korn's inequality in $\boldsymbol{V}_{h0}$ (see §42.4.1), we assume in this section that the momentum equation in the Stokes equations is written in the Laplacian (or Cauchy–Navier) form (see Remark 53.3), i.e., we replace the bilinear form a defined in (53.5) by $a(\boldsymbol{v}, \boldsymbol{w}) := \int_D \mu \nabla \boldsymbol{v} : \nabla \boldsymbol{w} \, \mathrm{d}x$. Since $\boldsymbol{V}_{h0}$ is nonconforming, we define the following discrete counterparts of the bilinear forms a and b:

$$a_h(\boldsymbol{v}_h, \boldsymbol{w}_h) := \sum_{K \in \mathcal{T}_h} \int_K \mu \nabla \boldsymbol{v}_h : \nabla \boldsymbol{w}_h \, \mathrm{d}x, \quad b_h(\boldsymbol{v}_h, q_h) := - \sum_{K \in \mathcal{T}_h} \int_K q_h \nabla \cdot \boldsymbol{v}_h \, \mathrm{d}x,$$

and consider the following discrete problem:

$$\begin{cases} \text{Find } \boldsymbol{u}_h \in \boldsymbol{V}_{h0} \text{ and } p_h \in Q_h \text{ such that} \\ a_h(\boldsymbol{u}_h, \boldsymbol{v}_h) + b_h(\boldsymbol{v}_h, p_h) = F(\boldsymbol{v}_h), & \forall \boldsymbol{v}_h \in \boldsymbol{V}_{h0}, \\ b_h(\boldsymbol{u}_h, q_h) = G(q_h), & \forall q_h \in Q_h, \end{cases} \quad (55.12)$$

where the linear forms on the right-hand side are defined as before as $F(\boldsymbol{v}_h) := \int_D \boldsymbol{f} \cdot \boldsymbol{v}_h \, \mathrm{d}x$ and $G(q_h) := - \int_D g q_h \, \mathrm{d}x$. Let $p \in (1, \infty)$ and let us equip $\boldsymbol{V}_{h0}$ with the mesh-dependent norm $|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(\mathcal{T}_h)}^p := \sum_{K \in \mathcal{T}_h} |\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(K)}^p$ (the same reasoning as in the proof of Lemma 36.4 shows that $\boldsymbol{v}_h \mapsto |\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(\mathcal{T}_h)}$ is indeed a norm on $\boldsymbol{V}_{h0}$).

Lemma 55.18 (Stability). *Let $p, p' \in (1, \infty)$ be s.t. $\frac{1}{p} + \frac{1}{p'} = 1$. There is β_0 such that for all $h \in \mathcal{H}$,*

$$\inf_{q_h \in Q_h} \sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}} \frac{|b_h(\boldsymbol{v}_h, q_h)|}{|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(\mathcal{T}_h)} \|q_h\|_{L^{p'}(D)}} \geq \beta_0 > 0. \quad (55.13)$$

Proof. For all $r \in L_*^p(D)$, there is $\boldsymbol{v}_r \in \boldsymbol{W}_0^{1,p}(D)$ s.t. $\nabla \cdot \boldsymbol{v}_r = r$ and $|\boldsymbol{v}_r|_{\boldsymbol{W}^{1,p}(D)} \leq c \|r\|_{L^p(D)}$ (see Remark 53.10). Let $\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{CR}} : \boldsymbol{W}_0^{1,p}(D) \to \boldsymbol{V}_{h0}$ be the vector-valued Crouzeix–Raviart interpolation operator. Owing to the local commuting property established in Exercise 36.1, we have

$b_h(\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{CR}}(\boldsymbol{v}_r) - \boldsymbol{v}_r, q_h) = 0$ for all $q_h \in Q_h$. Since

$$\int_D q_h r \, \mathrm{d}x = b(\boldsymbol{v}_r, q_h) = b_h(\boldsymbol{v}_r, q_h) = b_h(\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{CR}}(\boldsymbol{v}_r), q_h),$$

we infer that

$$\|q_h\|_{L^{p'}(D)} \leq \sup_{r \in L_*^p(D)} \frac{|\int_D q_h r \, \mathrm{d}x|}{\|r\|_{L^p(D)}} = \sup_{r \in L_*^p(D)} \frac{|b_h(\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{CR}}(\boldsymbol{v}_r), q_h)|}{\|r\|_{L^p(D)}}$$

$$\leq \sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}} \frac{|b_h(\boldsymbol{v}_h, q_h)|}{|\boldsymbol{v}_h|_{\boldsymbol{W}^{1,p}(\mathcal{T}_h)}} \times \sup_{r \in L_*^p(D)} \frac{|\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{CR}}(\boldsymbol{v}_r)|_{\boldsymbol{W}^{1,p}(\mathcal{T}_h)}}{\|r\|_{L^p(D)}}.$$

Using the $\boldsymbol{W}_0^{1,p}$-stability of $\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{CR}}$ (see Lemma 36.1 with $r := 0$) together with the above bound on $\boldsymbol{v}_r$, we conclude that $\sup_{r \in L_*^p(D)} \frac{|\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{CR}}(\boldsymbol{v}_r)|_{\boldsymbol{W}^{1,p}(\mathcal{T}_h)}}{\|r\|_{L^p(D)}}$ is uniformly bounded w.r.t. $h \in \mathcal{H}$. This proves the expected inf-sup condition. $\qquad\square$

Remark 55.19 (Convergence rate). The $(\mathbb{P}_1^{\mathrm{CR}}, \mathbb{P}_0^{\mathrm{b}})$ pair is first-order accurate. More precisely, let $(\boldsymbol{u}, p)$ solve (53.6) and assume that $\boldsymbol{u} \in \boldsymbol{H}^2(D) \cap \boldsymbol{H}_0^1(D)$, $p \in H^1(D) \cap L_*^2(D)$. Then the solution to (53.14) with $(\boldsymbol{V}_{h0}, Q_h)$ defined in (55.11) satisfies $\mu\|\nabla_h(\boldsymbol{u} - \boldsymbol{u}_h)\|_{\mathbb{L}^2(D)} + \|p - p_h\|_{L^2(D)} \leq ch(\mu|\boldsymbol{u}|_{\boldsymbol{H}^2(D)} + |p|_{H^1(D)})$. Moreover, if the assumptions of Theorem 53.19 are met for some $s \in (0, 1]$, we have $\mu\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\boldsymbol{L}^2(D)} \leq ch^{1+s}\ell_D^{1-s}(\mu|\boldsymbol{u}|_{\boldsymbol{H}^2(D)} + |p|_{H^1(D)})$; see Exercise 55.4. $\qquad\square$

Remark 55.20 (Literature). The $(\mathbb{P}_1^{\mathrm{CR}}, \mathbb{P}_0^{\mathrm{b}})$ pair has been introduced by Crouzeix and Raviart [151]. A quadrilateral nonconforming mixed finite element has been introduced by Rannacher and Turek [330, 366]. $\qquad\square$

Remark 55.21 (Fortin operator). The proof of Lemma 55.18 shows that the Crouzeix–Raviart interpolation operator acts as a nonconforming Fortin operator. Indeed, we have $\nabla{\cdot}(\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{CR}}(\boldsymbol{v})) = \Pi_K^0(\nabla{\cdot}\boldsymbol{v})$ for all $\boldsymbol{v} \in \boldsymbol{W}_0^{1,p}(D)$ and all $K \in \mathcal{T}_h$ (see Exercise 36.1), and since any $q_h \in Q_h$ is piecewise constant, this implies that $b_h(\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{CR}}(\boldsymbol{v}) - \boldsymbol{v}, q_h) = 0$. Moreover, there is $\gamma_0 > 0$ s.t. $\gamma_0|\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{CR}}(\boldsymbol{v})|_{\boldsymbol{W}^{1,p}(\mathcal{T}_h)} \leq |\boldsymbol{v}|_{\boldsymbol{W}^{1,p}(D)}$ for all $\boldsymbol{v} \in \boldsymbol{W}^{1,p}(D)$ and all $h \in \mathcal{H}$. $\qquad\square$

An arbitrary-order discretization of the Stokes equations can be done by using the hybrid high-order (HHO) method introduced in §39.1. The method uses face-based and cell-based velocities together with discontinuous cell-based pressures. Let $k \in \mathbb{N}$ denote the degree of the velocity and pressure unknowns. As in Di Pietro et al. [169], one can take any $k \geq 0$ if one uses the Cauchy–Navier form of the momentum equation (see Remark 53.3). If one uses instead the formulation based on the linearized strain tensor (i.e., (53.1a) with (53.2)), then one can adapt the HHO method for the linear elasticity equations from Di Pietro and Ern [166] (see §42.4.3). In this case, one takes $k \geq 1$ since the analysis invokes a Korn inequality in each mesh cell. In practice, the size of the linear system can be significantly reduced since one can eliminate locally all the cell-based velocities and all the (cell-based) pressures up to a constant in each cell. The size of the linear system is thus reduced to $\dim(\mathbb{P}_{k,d-1}){\times}d{\times}N_{\mathrm{f}} + N_{\mathrm{c}}$, where N_{f} and N_{c} are the number of mesh faces and cells, respectively. Other methods using similar discrete unknowns are the hybridizable discontinuous Galerkin (HDG) methods developed by Egger and Waluga [184], Cockburn and Shi [132], and the related weak Galerkin methods from Wang and Ye [387]. See also Lehrenfeld and Schöberl [281] for HDG methods with $\boldsymbol{H}(\mathrm{div})$-velocities and Jeon et al. [254] for hybridized finite elements.

Remark 55.22 (Well-balanced scheme). For the $(\mathbb{P}_1^{\mathrm{CR}}, \mathbb{P}_0^{\mathrm{b}})$ pair, the discrete velocity fields are divergence-free locally in each mesh cell, but since $\boldsymbol{V}_{h0}$ is nonconforming in $\boldsymbol{H}(\mathrm{div}; D)$ (the

normal component of fields in $\boldsymbol{V}_{h0}$ can jump across the mesh interfaces), these fields are generally not divergence-free in D. Recalling Remark 53.22, this means that the discretization is not well-balanced, and this can lead to a poor velocity approximation in problems with large curl-free body forces. This issue has been addressed in Linke [283], where a well-balanced scheme is designed by using a lifting operator mapping the velocity test functions to the lowest-order Raviart–Thomas space in order to test the body forces in the discrete momentum balance equation. A similar modification is possible for the HHO discretization by using a lifting operator mapping the velocity test functions to the Raviart–Thomas space of the same degree as the face-based velocities; see [169]. □

55.5   Stable pairs with $\mathbb{Q}_k$-based velocities

It is possible to used mixed finite elements based on quadrangular and hexahedral meshes. Since the literature on the topic is vast and this chapter is just meant to be a brief overview of the field, we only mention a few results. We assume in the entire section that $(\mathcal{T}_h)_{h\in\mathcal{H}}$ is a shape-regular sequence of affine meshes composed of cuboids. We start with a negative result.

Lemma 55.23 ($(\mathbb{Q}_k, \mathbb{Q}_{k-1}^b)$). *The $(\mathbb{Q}_k, \mathbb{Q}_{k-1}^b)$ pair composed of continuous $\mathbb{Q}_k$ elements for the velocity and discontinuous $\mathbb{Q}_{k-1}$ elements for the pressure does not satisfy the inf-sup condition for all $k \geq 1$.*

Proof. This result is established in Brezzi and Falk [92, Thm. 3.2]. A proof is proposed in Exercise 55.3. □

It is possible to save the situation by removing some degrees of freedom in the pressure space. This can be done by considering the polynomial space $\mathbb{P}_l^b$ instead of $\mathbb{Q}_l^b$ with $l \in \{0, 1\}$.

Lemma 55.24 ($(\mathbb{Q}_2, \mathbb{P}_0^b)$). *The $(\mathbb{Q}_2, \mathbb{P}_0^b)$ pair satisfies the inf-sup condition (53.15) uniformly w.r.t. $h \in \mathcal{H}$ in $\mathbb{R}^2$.*

Proof. The proof is the same as that for the $(\mathbb{P}_2, \mathbb{P}_0^b)$ pair. For every face/edge $F \in \mathcal{F}_h$ and every $\boldsymbol{v}_h \in \mathbb{Q}_{2,0}^g(\mathcal{T}_h)$, $\boldsymbol{v}_{|F} \cdot \boldsymbol{n}_F$ is quadratic and one can use Simpson's quadrature rule to compute $\int_F \boldsymbol{v}_h \cdot \boldsymbol{n}_F \, \mathrm{d}s$; see Exercise 55.2. □

Lemma 55.25 ($(\mathbb{Q}_2, \mathbb{P}_1^b)$). *The $(\mathbb{Q}_2, \mathbb{P}_1^b)$ pair satisfies the inf-sup condition (53.15) uniformly w.r.t. $h \in \mathcal{H}$ in $\mathbb{R}^2$ and yields the same error estimates as the Taylor–Hood mixed finite element.*

Proof. The proof is similar to that of the $(\mathbb{P}_2, \mathbb{P}_1^b)$ pair. The reader is referred to Boffi et al. [65, §8.6.3.1] for other details and a literature review. □

Remark 55.26 ($\mathbb{Q}_1$ geometric transformation). Let us assume that for all $K \in \mathcal{T}_h$, the geometric finite element that is used to construct the cells in $\mathcal{T}_h$ is the Lagrange $\mathbb{Q}_1$ element; see §8.1. Then the $(\mathbb{Q}_2, \mathbb{P}_1^b)$ pair satisfies the inf-sup condition (53.15) uniformly w.r.t. $h \in \mathcal{H}$ in $\mathbb{R}^2$ (the proof is the same as that of Lemma 55.25), but, as shown in Arnold et al. [22], the approximation properties are suboptimal since in this case the polynomial space $\mathbb{P}_1$ is not rich enough to ensure optimal approximability of the pressure. □

Exercises

Exercise 55.1 (Local mass balance). Let $\boldsymbol{u}_h \in \boldsymbol{V}_{h0}$ and $g \in L_*^2(D)$ satisfy $\int_D q_h \nabla\cdot\boldsymbol{u}_h \, \mathrm{d}x = \int_D q_h g \, \mathrm{d}x$ for all $q_h \in P_{k,*}^{\mathrm{b}}(\mathcal{T}_h)$. Show that $\int_K (\psi_K^{\mathrm{g}})^{-1}(q)\nabla\cdot\boldsymbol{u}_h \, \mathrm{d}x = \int_K (\psi_K^{\mathrm{g}})^{-1}(q)g \, \mathrm{d}x$ for all $q \in \mathbb{P}_{k,d}$ and all $K \in \mathcal{T}_h$ with $\psi_K^{\mathrm{g}}(q) := q \circ \boldsymbol{T}_K$. (*Hint*: use that $\int_D \nabla\cdot\boldsymbol{u}_h \, \mathrm{d}x = \int_D g \, \mathrm{d}x = 0$.)

Exercise 55.2 ($(\mathbb{P}_2, \mathbb{P}_0^{\mathrm{b}})$). Complete the proof of Lemma 55.8. (*Hint*: to show that the assumption (ii) from Lemma 54.2 is met, prove that $\int_F (\boldsymbol{v} - \boldsymbol{\Pi}_{2h}(\boldsymbol{v})) \, \mathrm{d}s = \boldsymbol{0}$ for all $F \in \mathcal{F}_h^{\circ}$ using Simpson's quadrature rule; to show that the assumption (iii) is met, show first that $|\boldsymbol{\Pi}_{2h}(\boldsymbol{v})|_{\boldsymbol{W}^{1,p}(K)} \leq ch_K^{\frac{1}{p}-1}\sum_{F\in\mathcal{F}_K^{\circ}} \|\boldsymbol{v}\|_{\boldsymbol{L}^p(F)}$ and then invoke the multiplicative trace inequality (12.16).)

Exercise 55.3 ($(\mathbb{Q}_k, \mathbb{Q}_{k-1}^{\mathrm{b}})$). (i) Justify Lemma 55.23 for $k := 2$ by constructing a counterexample. (*Hint*: given an interior vertex of a uniform Cartesian mesh, consider the patch composed of the four square cells sharing this vertex, and find an oscillating pressure field using (ii) from Exercise 54.3.) (ii) Generalize the argument for all $k \geq 2$.

Exercise 55.4 ($(\mathbb{P}_1^{\mathrm{CR}}, \mathbb{P}_0^{\mathrm{b}})$). Justify the claim in Remark 55.19. (*Hint*: see the proof of Theorem 36.11.)

Exercise 55.5 ($(\mathbb{P}_2, \mathbb{P}_1^{\mathrm{b}})$, HCT mesh). Using the notation from the proof of Lemma 55.14, the goal is to prove that $\mathrm{im}(\widehat{B})^{\perp} = \mathrm{span}(\mathbb{1}_{\widehat{U}})$. Let $\widehat{\boldsymbol{z}}_1 := (0,0)$, $\widehat{\boldsymbol{z}}_2 := (1,0)$, $\widehat{\boldsymbol{z}}_3 := (0,1)$, $\widehat{\boldsymbol{z}}_4 := (\frac{1}{3}, \frac{1}{3})$. Consider the triangles $\widehat{K}_1 := \mathrm{conv}(\widehat{\boldsymbol{z}}_1, \widehat{\boldsymbol{z}}_2, \widehat{\boldsymbol{z}}_4)$, $\widehat{K}_2 := \mathrm{conv}(\widehat{\boldsymbol{z}}_2, \widehat{\boldsymbol{z}}_3, \widehat{\boldsymbol{z}}_4)$, and $\widehat{K}_3 := \mathrm{conv}(\widehat{\boldsymbol{z}}_3, \widehat{\boldsymbol{z}}_1, \widehat{\boldsymbol{z}}_4)$. Let $p \in P_1^{\mathrm{b}}(\widehat{U})$ with the reference macroelement $\widehat{U} := \{\widehat{K}_1, \widehat{K}_2, \widehat{K}_3\}$, and set

$$p_1 := p_{|\widehat{K}_1}(\widehat{\boldsymbol{z}}_1), \ p_2 := p_{|\widehat{K}_1}(\widehat{\boldsymbol{z}}_2), \ p_3 := p_{|\widehat{K}_1}(\widehat{\boldsymbol{z}}_4),$$
$$q_1 := p_{|\widehat{K}_2}(\widehat{\boldsymbol{z}}_2), \ q_2 := p_{|\widehat{K}_2}(\widehat{\boldsymbol{z}}_3), \ q_3 := p_{|\widehat{K}_2}(\widehat{\boldsymbol{z}}_4),$$
$$s_1 := p_{|\widehat{K}_3}(\widehat{\boldsymbol{z}}_3), \ s_2 := p_{|\widehat{K}_3}(\widehat{\boldsymbol{z}}_1), \ s_3 := p_{|\widehat{K}_3}(\widehat{\boldsymbol{z}}_4).$$

Let $\widehat{\boldsymbol{m}}_{14} := \frac{1}{2}(\widehat{\boldsymbol{z}}_1 + \widehat{\boldsymbol{z}}_4)$, $\widehat{\boldsymbol{m}}_{24} := \frac{1}{2}(\widehat{\boldsymbol{z}}_2 + \widehat{\boldsymbol{z}}_4)$, $\widehat{\boldsymbol{m}}_{34} := \frac{1}{2}(\widehat{\boldsymbol{z}}_3 + \widehat{\boldsymbol{z}}_4)$. Let $\boldsymbol{u} \in \boldsymbol{P}_{2,0}^{\mathrm{g}}(\widehat{U})$ and set $(u_7, v_7)^{\mathsf{T}} := \boldsymbol{u}(\widehat{\boldsymbol{m}}_{14})$, $(u_8, v_8)^{\mathsf{T}} := \boldsymbol{u}(\widehat{\boldsymbol{m}}_{24})$, $(u_9, v_9)^{\mathsf{T}} := \boldsymbol{u}(\widehat{\boldsymbol{m}}_{34})$, $(u_{10}, v_{10})^{\mathsf{T}} := \boldsymbol{u}(\widehat{\boldsymbol{z}}_4)$. (i) Show (or accept as a fact) that

$$\int_{\widehat{K}_1} p\nabla\cdot\boldsymbol{u} \, \mathrm{d}\widehat{x} = (-u_7 + u_8 + 4v_7 + 2v_8)p_1$$
$$+ (-u_7 + u_8 + v_7 + 5v_8)p_2 + (-2u_7 + 2u_8 - v_7 + v_8 + 3v_{10})p_3.$$

(*Hint*: compute the $\mathbb{P}_2$ shape functions on $\widehat{K}_1$ associated with the nodes $\widehat{\boldsymbol{m}}_{14}$, $\widehat{\boldsymbol{m}}_{24}$, and $\widehat{\boldsymbol{z}}_4$.) (ii) Let $\boldsymbol{T}_{\widehat{K}_2} : \widehat{K}_1 \to \widehat{K}_2$, $\boldsymbol{T}_{\widehat{K}_3} : \widehat{K}_1 \to \widehat{K}_3$ be the geometric mappings s.t.

$$\boldsymbol{T}_{\widehat{K}_2}(\widehat{\boldsymbol{x}}) := \widehat{\boldsymbol{z}}_2 + \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix} (\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{z}}_1), \quad \boldsymbol{T}_{\widehat{K}_3}(\widehat{\boldsymbol{x}}) := \widehat{\boldsymbol{z}}_3 + \begin{pmatrix} 0 & 1 \\ -1 & -1 \end{pmatrix} (\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{z}}_1).$$

Verify that $\boldsymbol{T}_{\widehat{K}_i}$ maps the vertices of $\widehat{K}_1$ to the vertices of $\widehat{K}_i$ for $i \in \{2,3\}$. (iii) Compute the contravariant Piola tranformations $\psi_{\widehat{K}_2}^{\mathrm{d}}(\boldsymbol{v})$ and $\psi_{\widehat{K}_3}^{\mathrm{d}}(\boldsymbol{v})$. (iv) Compute $\int_{\widehat{K}_i} p\nabla\cdot\boldsymbol{u} \, \mathrm{d}\widehat{x}$ for $i \in \{2,3\}$. (*Hint*: use Steps (i) and (iii), and $\int_{\widehat{K}_i} q\nabla\cdot\boldsymbol{v} \, \mathrm{d}\widehat{x} = \int_{\widehat{K}_1} \psi_{\widehat{K}_i}^{\mathrm{g}}(q)\nabla\cdot(\psi_{\widehat{K}_i}^{\mathrm{d}}(\boldsymbol{v})) \, \mathrm{d}\widehat{x}$ (see Exercise 14.3(i)).) (v) Write the linear system corresponding to the statement $(\widehat{B}(\boldsymbol{u}), p)_{L^2(\widehat{U})} := \int_{\widehat{U}} p\nabla\cdot\boldsymbol{u} \, \mathrm{d}\widehat{x} = 0$ for all $\boldsymbol{u} \in \boldsymbol{P}_{2,0}^{\mathrm{g}}(\widehat{U})$, and compute $\mathrm{im}(\widehat{B})^{\perp}$.

Exercise 55.6 (Macroelement partition). Reprove Corollary 55.3 without invoking the partition lemma (Lemma 55.1). (*Hint*: see Brezzi and Bathe [91, Prop.4.2].)

Exercise 55.7 (Macroelement, continuous pressure). Let the assumptions of Proposition 55.5 hold true. (i) Show that there are $c_1, c_2 > 0$ s.t.

$$\sup_{\boldsymbol{v}_h \in \boldsymbol{V}_{h0}} \frac{|b(\boldsymbol{v}_h, q_h)|}{\|\boldsymbol{v}_h\|_{\boldsymbol{V}}} \geq c_1 \beta_D \|q_h\|_Q - c_2 \Big(\sum_{U \in \mathcal{U}_h} h_U^2 |q_h|_{H^1(U)}^2 \Big)^{\frac{1}{2}},$$

for all $q_h \in Q_h$ and all $h \in \mathcal{H}$. (*Hint:* use the quasi-interpolation operator $\boldsymbol{\mathcal{I}}_{h0}^{\mathrm{av}}$ and proceed as in the proof of Lemma 54.3.) (ii) Setting $\overline{q}_{hU} := \frac{1}{|U|} \int_U q_h \, dx$, show that there is c s.t. $|q_{h|U}|_{H^1(U)} \leq c \|q_h - \overline{q}_{hU}\|_{L^2(\widehat{U})}$ for all $U \in \mathcal{U}_h$ and all $h \in \mathcal{H}$. (*Hint*: use Lemma 11.7 and the affine geometric mapping $\boldsymbol{T}_U : \widehat{U} \to U$.) (iii) Prove Corollary 55.5. (*Hint*: use Remark 55.4. See also Brezzi and Bathe [91, Prop 4.1].)

Appendix C

Bijective operators in Banach spaces

The goal of this appendix is to recall fundamental results on linear operators (that is, bounded linear maps) in Banach and Hilbert spaces, and in particular to state conditions allowing us to assert the bijectivity of these operators. The results collected herein provide a theoretical framework for the mathematical analysis of the finite element method. We refer the reader to Aubin [29], Brezis [89], Lax [278], Rudin [337], Yosida [398], Zeidler [400] for further reading.

C.1 Injection, surjection, bijection

Since we are interested in asserting the bijectivity of bounded linear maps in Banach and Hilbert spaces, let us first recall some basic notions concerning injectivity, surjectivity, and bijectivity, as well as left and right inverses.

Definition C.1 (Injection, surjection, bijection). *Let E and G be two nonempty sets. A function (or map) $f : E \to G$ is said to be* injective *if every element of the codomain (i.e., G) is mapped to by at most one element of the domain (i.e., E). The function is said to be* surjective *if every element of the codomain is mapped to by at least one element of the domain. Finally, f is said to be* bijective *if every element of the codomain is mapped to by exactly one element of the domain (i.e., f is both injective and surjective).*

Definition C.2 (Left and right inverse). *Let E and G be two nonempty sets and let $f : E \to G$ be a function. We say that $f^{\ddagger} : G \to E$ is a* left inverse *of f if $(f^{\ddagger} \circ f)(e) = e$ for all $e \in E$, and that $f^{\dagger} : G \to E$ is a* right inverse *of f if $(f \circ f^{\dagger})(g) = g$ for all $g \in G$.*

A map with a left inverse is necessarily injective. Conversely, if the map $f : E \to G$ is injective, the following holds true: (i) The map $\tilde{f} : E \to f(E)$ such that $\tilde{f}(e) = f(e)$ for all $e \in E$ has a unique left inverse; (ii) One can construct a left inverse $f^{\ddagger} : G \to E$ of f by setting $f^{\ddagger}(g) := e$ (with $e \in E$ arbitrary) if $g \notin f(E)$ and $f^{\ddagger}(g) := (\tilde{f})^{\ddagger}(g)$ otherwise; (iii) If E, G are vector spaces and the map f is linear, the left inverse of $\tilde{f}$ is also linear. A map with a right inverse is necessarily surjective. Conversely, one can construct right inverse maps for every surjective map by invoking the axiom of choice.

C.2   Banach spaces

Basic properties of Banach and Hilbert spaces are collected in Appendix A. In this section, we recall these properties and give more details. To stay general, we consider complex vector spaces, i.e., vector spaces over the field $\mathbb{C}$ of complex numbers. The case of real vector spaces is recovered by replacing the field $\mathbb{C}$ by $\mathbb{R}$, by removing the real part symbol $\Re(\cdot)$ and the complex conjugate symbol $\bar{\cdot}$, and by interpreting $|\cdot|$ as the absolute value instead of the complex modulus. Recall that a complex vector space V equipped with a norm $\|\cdot\|_V$ is said to be a Banach space if every Cauchy sequence in V has a limit in V.

Let V, W be complex vector spaces. The complex vector space composed of the bounded linear maps from V to W is denoted by $\mathcal{L}(V; W)$. Members of $\mathcal{L}(V; W)$ are often called *operators*. This space is equipped with the norm

$$\|A\|_{\mathcal{L}(V;W)} := \sup_{v \in V} \frac{\|A(v)\|_W}{\|v\|_V} < \infty, \qquad \forall A \in \mathcal{L}(V; W). \tag{C.1}$$

In this book, we systematically abuse the notation by implicitly assuming that the argument in this type of supremum or infimum is nonzero. If W is a Banach space, then $\mathcal{L}(V; W)$ equipped with the above norm is also a Banach space (see Rudin [337, p. 87], Yosida [398, p. 111]).

Theorem C.3 (Banach–Steinhaus). *Let V, W be Banach spaces and let $\{A_i\}_{i \in \mathcal{I}}$ be a collection of operators in $\mathcal{L}(V; W)$ (the set $\mathcal{I}$ is not necessarily countable). Assume that $\sup_{i \in \mathcal{I}} \|A_i(v)\|_W$ is a finite number for all $v \in V$. Then there is a real number C such that*

$$\sup_{i \in \mathcal{I}} \|A_i(v)\|_W \le C\|v\|_V, \qquad \forall v \in V. \tag{C.2}$$

Proof. See Brezis [89, p. 32], Lax [278, Chap. 10]. $\qquad\square$

Corollary C.4 (Pointwise convergence). *Let V, W be Banach spaces. Let $(A_n)_{n \in \mathbb{N}}$ be a sequence in $\mathcal{L}(V; W)$ such that for all $v \in V$, the sequence $(A_n(v))_{n \in \mathbb{N}}$ converges as $n \to \infty$ to a limit in W denoted by $A(v)$ (one says that the sequence $(A_n)_{n \in \mathbb{N}}$ converges pointwise to A). The following holds true:*

(i) $\sup_{n \in \mathbb{N}} \|A_n\|_{\mathcal{L}(V;W)} < \infty$.

(ii) $A \in \mathcal{L}(V; W)$.

(iii) $\|A\|_{\mathcal{L}(V;W)} \le \liminf_{n \to \infty} \|A_n\|_{\mathcal{L}(V;W)}$.

Proof. The statement (i) follows from the Banach–Steinhaus theorem. Owing to (C.2), we infer that $\|A_n(v)\|_W \le C\|v\|_V$ for all $v \in V$ and all $n \in \mathbb{N}$. Letting $n \to \infty$ yields $\|A(v)\|_W \le C\|v\|_V$, and since A is obviously linear, we infer that the statement (ii) holds true. The statement (iii) results from the bound $\|A_n(v)\|_W \le \|A_n\|_{\mathcal{L}(V;W)}\|v\|_V$ for all $v \in V$ and all $n \in \mathbb{N}$. $\qquad\square$

Remark C.5 (Uniform convergence on compact sets). Corollary C.4 does not claim that $(A_n)_{n \in \mathbb{N}}$ converges to A in $\mathcal{L}(V; W)$, i.e., uniformly on bounded sets. A standard argument shows however that $(A_n)_{n \in \mathbb{N}}$ converges uniformly to A on compact sets. Let indeed $K \subset V$ be a compact set. Let $\epsilon > 0$. Set $C := \sup_{n \in \mathbb{N}} \|A_n\|_{\mathcal{L}(V;W)}$. The real number C is finite owing to Corollary C.4(i). The set K being compact, there is a finite set of points $\{x_i\}_{i \in \mathcal{I}}$ in K such that for all $v \in K$, there is $i \in \mathcal{I}$ such that $\|v - x_i\|_V \le (3C)^{-1}\epsilon$. Owing to the pointwise convergence of $(A_n)_{n \in \mathbb{N}}$ to A, there is N_i such that $\|A_n(x_i) - A(x_i)\|_W \le \frac{1}{3}\epsilon$ for all $n \ge N_i$. Using the triangle inequality and the statement (iii) above, we infer that

$$\|A_n(v) - A(v)\|_W \le \|A_n(v - x_i)\|_W + \|A_n(x_i) - A(x_i)\|_W + \|A(v - x_i)\|_W \le \epsilon,$$

for all $v \in K$ and all $n \ge \max_{i \in \mathcal{I}} N_i$. $\qquad\square$

C.3   Hilbert spaces

Let V be a complex vector space equipped with an inner product $(\cdot, \cdot)_V : V \times V \to \mathbb{C}$. Recall that the inner product is linear w.r.t. its first argument and antilinear w.r.t. its second argument, i.e., $(\lambda v, w)_V = \lambda(v, w)_V$ and $(v, \lambda w)_V = \overline{\lambda}(v, w)_V$ for all $\lambda \in \mathbb{C}$ and all $v, w \in V$, and that Hermitian symmetry means that $(v, w)_V = \overline{(w, v)_V}$. The space V is said to be a Hilbert space if it is a Banach space when equipped with the induced norm $\|v\|_V := (v, v)_V^{\frac{1}{2}}$ for all $v \in V$. Recall the Cauchy–Schwarz inequality

$$|(v, w)_V| \leq \|v\|_V \|w\|_V, \qquad \forall v, w \in V. \tag{C.3}$$

Notice that we obtain an equality in (C.3) iff v and w are collinear. This follows from $\|v\|_V \|w\|_V - \Re(\xi(v, w)_V) = \frac{\|v\|_V \|w\|_V}{2} \left\| \frac{v}{\|v\|_V} - \overline{\xi} \frac{w}{\|w\|_V} \right\|_V^2$ for all nonzero $v, w \in V$ and all $\xi \in \mathbb{C}$ with $|\xi| = 1$.

Remark C.6 (Arithmetic-geometric and Young's inequalities). Let $x_1, \ldots, x_n$ be non-negative real numbers. Using the convexity of the function $x \mapsto e^x$, one can show the following arithmetic-geometric inequality:

$$(x_1 x_2 \ldots x_n)^{\frac{1}{n}} \leq \frac{1}{n}(x_1 + \ldots + x_n). \tag{C.4}$$

Moreover, Young's inequality states that for every positive real number $\gamma > 0$,

$$|(v, w)_V| \leq \frac{\gamma}{2} \|v\|_V^2 + \frac{1}{2\gamma} \|w\|_V^2, \qquad \forall v, w \in V. \tag{C.5}$$

This follows from the Cauchy–Schwarz inequality and (C.4) with $n := 2$, $x_1 := \gamma \|v\|_V^2$, and $x_2 := \gamma^{-1} \|w\|_V^2$. $\qquad \square$

Definition C.7 (Hilbert basis). *A sequence $(e_n)_{n \in \mathbb{N}}$ in V is said to be a Hilbert basis of V if it satisfies the following two properties:*

(i) *$(e_m, e_n)_V = \delta_{mn}$ for all $m, n \in \mathbb{N}$.*

(ii) *The linear space composed of all the finite linear combinations of the vectors in $(e_n)_{n \in \mathbb{N}}$ is dense in V.*

The existence of Hilbert bases is not a natural consequence of the Hilbert space structure, but the question of the existence of Hilbert bases can be given a positive answer by introducing the notion of separability.

Definition C.8 (Separability). *A Hilbert space V is said to be separable if it admits a countable dense subset $(v_n)_{n \in \mathbb{N}}$.*

Not every Hilbert space is separable, but all the Hilbert spaces encountered in this book are separable (or by default are always assumed to be separable). The main motivation for the notion of separability is the following result.

Theorem C.9 (Separability and Hilbert basis). *Every separable Hilbert space has a Hilbert basis.*

Proof. See [89, Thm. 5.11]. $\qquad \square$

Lemma C.10 (Pareseval). *Let $(e_n)_{n\in\mathbb{N}}$ be a Hilbert basis of V. For all $u \in V$, set $u_n := \sum_{k\in\{0:n\}}(u, e_k)_V e_k$. The following holds true:*

$$\lim_{n\to\infty} \|u - u_n\|_V = 0 \quad and \quad \|u\|_V^2 = \sum_{k\in\mathbb{N}} |(u, e_k)_V|^2. \tag{C.6}$$

Conversely, let $(\alpha_n)_{n\in\mathbb{N}}$ be a sequence in $\ell^2(\mathbb{C})$ and set $u_{\alpha,n} := \sum_{k\in\{0:n\}} \alpha_k e_k$ for all $n \in \mathbb{N}$. Then the sequence $(u_{\alpha,n})_{n\in\mathbb{N}}$ converges to some u_α in V such that $(u_\alpha, e_n)_V = \alpha_n$ for all $n \in \mathbb{N}$, and we have $\|u_\alpha\|_V^2 = \lim_{n\to\infty} \sum_{k\in\{0:n\}} \alpha_k^2$.

Proof. See Brezis [89, Thm. 5.9]. $\square$

A striking consequence of Lemma C.10 is that all separable Hilbert spaces are isomorphic and isometric with $\ell^2(\mathbb{C})$.

Remark C.11 (Space $V_\mathbb{R}$). Let V be a complex vector space. By restricting the scaling operation $(\lambda, v) \mapsto \lambda v$ to $(\lambda, v) \in \mathbb{R}\times V$, V can also be equipped with a vector space structure over $\mathbb{R}$, which we denote by $V_\mathbb{R}$ (V and $V_\mathbb{R}$ are the same sets, but they are equipped with different vector space structures). For instance, if $V = \mathbb{C}^m$, then $\dim(V) = m$ but $\dim(V_\mathbb{R}) = 2m$. Moreover, the canonical set $\{e_k\}_{k\in\{1:m\}}$, where the Cartesian components of e_k in $\mathbb{C}^m$ are $e_{k,l} = \delta_{kl}$ (the Kronecker symbol) for all $l \in \{1:m\}$, is a basis of V, whereas the set $\{e_k, \mathsf{i}e_k\}_{k\in\{1:m\}}$ with $\mathsf{i}^2 = -1$ is a basis of $V_\mathbb{R}$. Finally, if V is a complex Hilbert space with inner product $(\cdot, \cdot)_V$, then $V_\mathbb{R}$ is a real Hilbert space with inner product $\Re(\cdot, \cdot)_V$. $\square$

C.4 Duality, reflexivity, and adjoint operators

Let V be a complex Banach space. Its dual space V' is composed of all the antilinear forms $A : V \to \mathbb{C}$ that are bounded. The reason we consider antilinear forms is that we employ the complex conjugate of test functions in the weak formulation of complex-valued PDEs. The action of $A \in V'$ on $v \in V$ is denoted by $\langle A, v\rangle_{V',V} \in \mathbb{C}$ (and sometimes also $A(v)$). Equipped with the norm

$$\|A\|_{V'} := \sup_{v\in V} \frac{|\langle A, v\rangle_{V',V}|}{\|v\|_V}, \qquad \forall A \in V', \tag{C.7}$$

V' is a Banach space. In the real case, the absolute value can be omitted at the numerator since $\pm v$ can be considered in the supremizing set. In the complex case, the modulus can be replaced by the real part since v can be multiplied by any $\xi \in \mathbb{C}$ with $|\xi| = 1$.

Remark C.12 (Linear vs. antilinear form). If $A : V \to \mathbb{C}$ is an antilinear form, then $\overline{A}$ (defined by $\overline{A}(v) := \overline{A(v)} \in \mathbb{C}$ for all $v \in V$) is a linear form. $\square$

C.4.1 Fundamental results in Banach spaces

Theorem C.13 (Hahn–Banach). *Let V be a normed vector space over $\mathbb{C}$ and let W be a subspace of V. Let $B \in W'$. There exists $A \in V'$ that extends B, i.e., $A(w) = B(w)$ for all $w \in W$, and such that $\|A\|_{V'} = \|B\|_{W'}$.*

Proof. For the real case, see Brezis [89, p. 3], Lax [278, Chap. 3], Rudin [337, p. 56], Yosida [398, p. 102]. The above statement is a simplified version of the actual Hahn–Banach theorem. For the complex case, see Lax [278, p. 27], Brezis [89, Prop. 11.23]. $\square$

Corollary C.14 (Norm by duality). *The following holds true:*

$$\|v\|_V = \sup_{A \in V'} \frac{|A(v)|}{\|A\|_{V'}} = \sup_{A \in V'} \frac{|\langle A, v \rangle_{V'V}|}{\|A\|_{V'}}, \tag{C.8}$$

for all $v \in V$, and the supremum is attained.

Proof. Assume $v \neq 0$ (the assertion is obvious for $v = 0$). We first observe that $\sup_{A \in V'} \frac{|A(v)|}{\|A\|_{V'}} \leq \|v\|_V$ since $|A(v)| \leq \|A\|_{V'}\|v\|_V$. Let $W := \text{span}\{v\}$ and let $B \in W'$ be defined as $B(\lambda v) := \bar{\lambda}\|v\|_V$ for all $\lambda \in \mathbb{C}$. By construction, $B \in W'$ and $\|B\|_{W'} = 1$. Owing to the Hahn–Banach theorem, there exists $A \in V'$ such that $\|A\|_{V'} = 1$ and $A(v) = B(v) = \|v\|_V$. □

Corollary C.15 (Characterization of density). *Let V be a normed vector space over $\mathbb{C}$ and W be a subspace of V. Then $\overline{W} \neq V$ (i.e., W is not dense in V) if and only if there exists $f \in V'\backslash\{0\}$ such that $f(w) = 0$ for all $w \in W$.*

Proof. See Brezis [89, p. 8], Rudin [337, Thm. 5.19]. □

Definition C.16 (Double dual). *The* double dual *of a Banach space V is denoted by V'' and is defined to be the dual space of its dual space V'.*

Proposition C.17 (Isometry into double dual). *The bounded linear map $J_V : V \to V''$ such that*

$$\langle J_V(v), \phi' \rangle_{V'',V'} = \overline{\langle \phi', v \rangle_{V',V}}, \qquad \forall (v, \phi') \in V \times V', \tag{C.9}$$

is an isometry.

Proof. The claim follows from Corollary C.14 since

$$\|J_V(v)\|_{V''} = \sup_{\phi' \in V'} \frac{|\langle J_V(v), \phi' \rangle_{V'',V'}|}{\|\phi'\|_{V'}} = \sup_{\phi' \in V'} \frac{|\langle \phi', v \rangle_{V',V}|}{\|\phi'\|_{V'}} = \|v\|_V. \qquad □$$

Definition C.18 (Reflexivity). *A Banach space V is said to be* reflexive *if J_V is an isomorphism.*

Remark C.19 (Map J_V). Since J_V is an isometry, it is injective. Thus, V can be identified with the subspace $J_V(V) \subset V''$. It may happen that the map J_V is not surjective. In this case, V is a proper subspace of V''. □

Example C.20 (Lebesgue spaces). One important consequence of Theorem 1.41 is that the Lebesgue space $L^p(D)$ is reflexive for all $p \in (1, \infty)$. However, $L^1(D)$ and $L^\infty(D)$ are not reflexive. Indeed, $L^\infty(D) = L^1(D)'$, but $L^1(D) \subsetneq L^\infty(D)'$ with strict inclusion; see §1.4 and Brezis [89, p. 102]. □

Remark C.21 (Space $V_\mathbb{R}$). Let V be a complex vector space and let $V_\mathbb{R}$ be defined in Remark C.11. Let $V'_\mathbb{R}$ be the dual space of $V_\mathbb{R}$, i.e., the normed real vector space composed of the bounded $\mathbb{R}$-linear maps from V to $\mathbb{R}$. Then the map $I : V' \to V'_\mathbb{R}$ s.t. for all $\ell \in V'$, $I(\ell)(v) := \Re(\ell(v))$, for all $v \in V$, is a bijective isometry; see [89, Prop. 11.22]. □

Definition C.22 (Weak convergence). *Let V be a Banach space. The sequence $(v_n)_{n \in \mathbb{N}}$ in V is said to* converge weakly *to $v \in V$ if*

$$\langle A, v_n \rangle_{V',V} \to \langle A, v \rangle_{V',V}, \qquad \forall A \in V'. \tag{C.10}$$

It is shown in Brezis [89, Prop. 3.5] that if the sequence $(v_n)_{n \in \mathbb{N}}$ converges strongly to v (that is, in the norm topology, i.e., $\|v_n - v\|_V \to 0$ as $n \to \infty$), then it also converges weakly to v. The converse is true if V is finite-dimensional (see [89, Prop. 3.6]). Furthermore, if the sequence $(v_n)_{n \in \mathbb{N}}$ converges weakly to v, then it is bounded and $\|v\|_V \le \liminf_{n \to \infty} \|v_n\|_V$. One important result on weak convergence is the following (see [89, Thm. 3.18]).

Theorem C.23 (Reflexivity and weak compactness). *Let V be a reflexive Banach space. Then from every bounded sequence $(v_n)_{n \in \mathbb{N}}$ of V, there exists a subsequence $(v_{n_k})_{k \in \mathbb{N}}$ that is weakly convergent.*

C.4.2 Further results in Hilbert spaces

Theorem C.24 (Riesz–Fréchet). *The operator $J_V^{\mathrm{RF}} : V \to V'$ such that*

$$\langle J_V^{\mathrm{RF}}(v), w \rangle_{V',V} := (v, w)_V, \qquad \forall v, w \in V, \tag{C.11}$$

is a linear isometric isomorphism.

Proof. See Brezis [89, Thm. 5.5], Lax [278, p. 56], Yosida [398, p. 90], or Exercise 25.1. $\square$

Remark C.25 (Riesz–Fréchet representation). Theorem C.24 is often called *Riesz–Fréchet representation* theorem. It states that for every antilinear form $v' \in V'$, there exists a unique vector $v \in V$ such that $v' = J_V^{\mathrm{RF}}(v)$. The vector $(J_V^{\mathrm{RF}})^{-1}(v') \in V$ is called *Riesz–Fréchet representative* of the antilinear form $v' \in V'$. The action of v' on V is represented by $(J_V^{\mathrm{RF}})^{-1}(v')$ with the identity $\langle v', w \rangle_{V',V} = \left((J_V^{\mathrm{RF}})^{-1}(v'), w \right)_V$ for all $w \in V$. $\square$

Remark C.26 (Linear vs. antilinear). Notice that J_V^{RF} is a linear operator. If we had adopted the convention that dual spaces were composed of linear forms, we would have had to define J_V^{RF} by setting $\langle J_V^{\mathrm{RF}}(v), w \rangle_{V',V} := \overline{(v, w)_V}$ for all $v, w \in V$, or, equivalently, $\langle v', w \rangle_{V',V} := \overline{((J_V^{\mathrm{RF}})^{-1}(v'), w)_V}$ for all $w \in V$ and $v' \in V'$. In this case, J_V^{RF} would have been antilinear. $\square$

Corollary C.27 (Reflexivity). *Hilbert spaces are reflexive.*

Owing to the Riesz–Fréchet theorem, the notion of weak convergence (see Definition C.22) can be reformulated as follows in Hilbert spaces.

Definition C.28 (Weak convergence). *Let V be a Hilbert space. The sequence $(v_n)_{n \in \mathbb{N}}$ in V is said to* converge weakly *to $v \in V$ if $(w, v_n)_V \to (w, v)_V$ as $n \to \infty$, for all $w \in V$.*

A useful connection between weak and strong convergence in Hilbert spaces is that if the sequence $(v_n)_{n \in \mathbb{N}}$ converges weakly to $v \in V$ and if additionally, $\|v_n\|_V \to \|v\|_V$ as $n \to \infty$, then the sequence $(v_n)_{n \in \mathbb{N}}$ converges strongly to v, i.e., $\|v_n - v\|_V \to 0$ as $n \to \infty$ (see, e.g., Brezis [89, Prop. 3.32]).

C.4.3 Adjoint

Definition C.29 (Adjoint operator). *Let V, W be complex Banach spaces. Let $A \in \mathcal{L}(V; W)$. The* adjoint operator *of A is the bounded linear operator $A^* \in \mathcal{L}(W'; V')$ such that*

$$\langle A^*(w'), v \rangle_{V',V} := \langle w', A(v) \rangle_{W',W}, \qquad \forall (v, w') \in V \times W'. \tag{C.12}$$

Note that $(\lambda A)^* = \overline{\lambda} A^*$ for all $\lambda \in \mathbb{C}$.

Lemma C.30 (Norm of adjoint). *Let $A \in \mathcal{L}(V;W)$ and let $A^* \in \mathcal{L}(W';V')$ be its adjoint. Then $\|A^*\|_{\mathcal{L}(W';V')} = \|A\|_{\mathcal{L}(V;W)}$.*

Proof. We have

$$\|A^*\|_{\mathcal{L}(W';V')} = \sup_{w' \in W'} \frac{\|A^*(w')\|_{V'}}{\|w'\|_{W'}} = \sup_{w' \in W'} \sup_{v \in V} \frac{|\langle A^*(w'), v \rangle_{V',V}|}{\|v\|_V \|w'\|_{W'}}$$

$$= \sup_{v \in V} \sup_{w' \in W'} \frac{|\langle w', A(v) \rangle_{W',W}|}{\|v\|_V \|w'\|_{W'}} = \sup_{v \in V} \frac{\|A(v)\|_W}{\|v\|_V} = \|A\|_{\mathcal{L}(V;W)},$$

where we used that $\sup_{w' \in W'} \sup_{v \in V} = \sup_{v \in V} \sup_{w' \in W'}$, the definition of A^*, and Corollary C.14. $\square$

Definition C.31 (Self-adjoint operator). *Let V be a reflexive Banach space. Let $A \in \mathcal{L}(V;V')$, so that $A^* \in \mathcal{L}(V'';V')$. The operator A is said to be* self-adjoint *if $A = A^* \circ J_V$, i.e., if the following holds true:*

$$\langle A(v), w \rangle_{V',V} = \overline{\langle A(w), v \rangle_{V',V}}, \qquad \forall v, w \in V. \tag{C.13}$$

In particular, $\langle A(v), v \rangle_{V',V}$ takes real values if A is self-adjoint. (Notice that if A is self-adjoint, λA is not self-adjoint if the imaginary part of $\lambda \in \mathbb{C}$ is nonzero.) If the spaces V and V'' are actually identified, we write $A^* \in \mathcal{L}(V;V')$ and say that A is self-adjoint if $A = A^*$.

Remark C.32 (Hermitian transpose). If V and W are finite-dimensional and after choosing one basis for V and one for W, A can be represented by a matrix with complex-valued entries. Then A^* is represented in the same bases by the Hermitian transpose of this matrix. Self-adjoint operators are represented by Hermitian matrices. $\square$

C.5 Open mapping and closed range theorems

Let V, W be complex Banach spaces. For $A \in \mathcal{L}(V;W)$, we denote by $\ker(A)$ its kernel and by $\mathrm{im}(A)$ its range. The operator A being bounded, $\ker(A)$ is closed in V. Hence, the quotient of V by $\ker(A)$, $V/\ker(A)$, can be defined. This space is composed of equivalence classes $\check{v}$ such that v and w are in the same class $\check{v}$ if and only if $v - w \in \ker(A)$, i.e., $A(v) = A(w)$.

Theorem C.33 (Quotient space). *The space $V/\ker(A)$ is a Banach space when equipped with the norm $\|\check{v}\| := \inf_{v \in \check{v}} \|v\|_V$. Moreover, the operator $\check{A} : V/\ker(A) \to \mathrm{im}(A)$ s.t. $\check{A}(\check{v}) := A(v)$ for all v in $\check{v}$, is an isomorphism.*

Proof. See Brezis [89, §11.2], Yosida [398, p. 60]. $\square$

For subspaces $M \subset V$ and $N \subset V'$, we define the *annihilators* of M and N as follows:

$$M^\perp := \{v' \in V' \mid \forall m \in M, \ \langle v', m \rangle_{V',V} = 0\}, \tag{C.14a}$$

$$N^\perp := \{v \in V \mid \forall n' \in N, \ \langle n', v \rangle_{V',V} = 0\}. \tag{C.14b}$$

Let $\overline{M}$ denote the closure of the subspace M in V. A characterization of $\ker(A)$ and $\mathrm{im}(A)$ is given by the following result.

Lemma C.34 (Kernel and range). *Let $A \in \mathcal{L}(V;W)$. The following holds true:*

(i) $\ker(A) = (\mathrm{im}(A^*))^\perp$.

(ii) $\ker(A^*) = (\operatorname{im}(A))^{\perp}$.

(iii) $\overline{\operatorname{im}(A)} = (\ker(A^*))^{\perp}$.

(iv) $\overline{\operatorname{im}(A^*)} \subset (\ker(A))^{\perp}$.

Proof. See Brezis [89, Cor. 2.18], Yosida [398, pp. 202-209]. □

Showing that the range of an operator is closed is a crucial step towards proving that this operator is surjective. This is the purpose of the following fundamental theorem.

Theorem C.35 (Banach or closed range). *Let $A \in \mathcal{L}(V;W)$. The following statements are equivalent:*

(i) $\operatorname{im}(A)$ *is closed.*

(ii) $\operatorname{im}(A^*)$ *is closed.*

(iii) $\operatorname{im}(A) = (\ker(A^*))^{\perp}$.

(iv) $\operatorname{im}(A^*) = (\ker(A))^{\perp}$.

Proof. See Brezis [89, Thm. 2.19], Yosida [398, p. 205]. □

We now put in place the second keystone of the edifice.

Theorem C.36 (Open mapping). *If $A \in \mathcal{L}(V;W)$ is surjective and U is an open set in V, then $A(U)$ is an open set in W.*

Proof. See Brezis [89, Thm. 2.6], Lax [278, p. 168], Rudin [337, p. 47], Yosida [398, p. 75]. □

Theorem C.36, also due to Banach, has far-reaching consequences. In particular, it leads to the following characterization of the closedness of $\operatorname{im}(A)$.

Lemma C.37 (Characterization of closed range). *Let $A \in \mathcal{L}(V;W)$. The following statements are equivalent:*

(i) $\operatorname{im}(A)$ *is closed in W.*

(ii) *A has a bounded right inverse map $A^{\dagger} : \operatorname{im}(A) \to V$, i.e., $(A \circ A^{\dagger})(w) = w$ for all $w \in \operatorname{im}(A)$, and there exists $\alpha > 0$ such that $\alpha \|A^{\dagger}(w)\|_V \leq \|w\|_W$ for all $w \in \operatorname{im}(A)$ ($A^{\dagger}$ is not necessarily linear).*

Proof. (i) $\Rightarrow$ (ii). Since $\operatorname{im}(A)$ is closed in W, $\operatorname{im}(A)$ is a Banach space. Applying the open mapping theorem to $A : V \to \operatorname{im}(A)$ and $U := B_V(0,1)$ (the open unit ball in V) proves that $A(B_V(0,1))$ is open in $\operatorname{im}(A)$. Since $0 \in A(B_V(0,1))$, there is $\gamma > 0$ s.t. $B_W(0,\gamma) \subset A(B_V(0,1))$. Let $w \in \operatorname{im}(A)$. Since $\frac{\gamma}{2} \frac{w}{\|w\|_W} \in B_W(0,\gamma)$, there is $z \in B_V(0,1)$ s.t. $A(z) = \frac{\gamma}{2} \frac{w}{\|w\|_W}$. Setting $A^{\dagger}(w) := \frac{2\|w\|_W}{\gamma} z$ leads to $A(A^{\dagger}(w)) = w$ and $\frac{\gamma}{2} \|A^{\dagger}(w)\|_V \leq \|w\|_W$.
(ii) $\Rightarrow$ (i). Let $(w_n)_{n \in \mathbb{N}}$ be a sequence in $\operatorname{im}(A)$ that converges to some $w \in W$. The sequence $(v_n := A^{\dagger}(w_n))_{n \in \mathbb{N}}$ in V is such that $A(v_n) = w_n$ and $\alpha \|v_n\|_V \leq \|w_n\|_W$. Thus, $(v_n)_{n \in \mathbb{N}}$ is a Cauchy sequence in V. Since V is a Banach space, $(v_n)_{n \in \mathbb{N}}$ converges to a certain $v \in V$. Owing to the boundedness of A, $(A(v_n))_{n \in \mathbb{N}}$ converges to $A(v)$. Hence, $w = A(v) \in \operatorname{im}(A)$. □

Corollary C.38 (Bounded inverse). *If $A \in \mathcal{L}(V;W)$ is bijective, then $A^{-1} \in \mathcal{L}(W;V)$.*

Proof. Since A is bijective, $\operatorname{im}(A) = W$ is closed. Moreover, the right inverse $A^{\dagger}$ is necessarily equal to A^{-1} (apply A^{-1} to $A \circ A^{\dagger} = I_W$). Lemma C.37(ii) shows that $A^{-1} \in \mathcal{L}(W;V)$ with $\|A^{-1}\|_{\mathcal{L}(W;V)} \leq \alpha^{-1}$. □

C.6   Characterization of surjectivity

As a consequence of the closed range theorem and of the open mapping theorem, we deduce two characterizations of surjective operators.

Lemma C.39 (Surjectivity of A^*). *Let $A \in \mathcal{L}(V;W)$. The following statements are equivalent:*

(i) $A^* : W' \to V'$ *is surjective.*

(ii) $A : V \to W$ *is injective and* $\mathrm{im}(A)$ *is closed in W.*

(iii) *There exists $\alpha > 0$ such that*

$$\|A(v)\|_W \geq \alpha\|v\|_V, \qquad \forall v \in V. \tag{C.15}$$

   Equivalently, there exists $\alpha > 0$ such that

$$\inf_{v \in V} \sup_{w' \in W'} \frac{|\langle w', A(v) \rangle_{W',W}|}{\|w'\|_{W'}\|v\|_V} \geq \alpha. \tag{C.16}$$

Proof. (i) $\Rightarrow$ (iii). Since the map A^* is surjective, Lemma C.37 implies that A^* has a bounded right inverse map $A^{*\dagger} : V' \to W'$. In particular, $A^*(A^{*\dagger}(v')) = v'$ for all $v' \in V'$, and there is $\alpha > 0$ such that $\alpha\|A^{*\dagger}(v')\|_{W'} \leq \|v'\|_{V'}$. Let now $v \in V$. We infer that

$$\frac{|\langle v', v \rangle_{V',V}|}{\|v'\|_{V'}} = \frac{|\langle A^*(A^{*\dagger}(v')), v \rangle_{V',V}|}{\|v'\|_{V'}} \leq \alpha^{-1} \frac{|\langle A^{*\dagger}(v'), A(v) \rangle_{W',W}|}{\|A^{*\dagger}(v')\|_{W'}}$$

$$\leq \alpha^{-1} \sup_{w' \in W'} \frac{|\langle w', A(v) \rangle_{W',W}|}{\|w'\|_{W'}}.$$

Since $\|v\|_V = \sup_{v' \in V'} \frac{|\langle v', v \rangle_{V',V}|}{\|v'\|_{V'}}$, taking the supremum w.r.t. $v' \in V'$ followed by the infimum w.r.t. $v \in V$ proves (C.16). Moreover, (C.15) and (C.16) are equivalent owing to Corollary C.14. (iii) $\Rightarrow$ (ii). The bound (C.15) implies that A is injective. Consider a sequence $(v_n)_{n \in \mathbb{N}}$ such that $(A(v_n))_{n \in \mathbb{N}}$ is a Cauchy sequence in W. Then (C.15) implies that $(v_n)_{n \in \mathbb{N}}$ is a Cauchy sequence in V. Let v be its limit. A being bounded implies that $A(v_n) \to A(v)$. Hence, $\mathrm{im}(A)$ is closed. (ii) $\Rightarrow$ (i). Since $\mathrm{im}(A)$ is closed, we use Theorem C.35(iv) together with the injectivity of A to infer that $\mathrm{im}(A^*) = (\ker(A))^\perp = \{0\}^\perp = V'$. This shows that A^* is surjective. $\square$

Lemma C.40 (Surjectivity of A). *Let $A \in \mathcal{L}(V;W)$. The following statements are equivalent:*

(i) $A : V \to W$ *is surjective.*

(ii) $A^* : W' \to V'$ *is injective and* $\mathrm{im}(A^*)$ *is closed in V'.*

(iii) *There exists $\alpha > 0$ such that*

$$\|A^*(w')\|_{V'} \geq \alpha\|w'\|_{W'}, \qquad \forall w' \in W'. \tag{C.17}$$

   Equivalently, there exists $\alpha > 0$ such that

$$\inf_{w' \in W'} \sup_{v \in V} \frac{|\langle A^*(w'), v \rangle_{V',V}|}{\|w'\|_{W'}\|v\|_V} \geq \alpha. \tag{C.18}$$

Proof. We only prove the implication (i) $\Rightarrow$ (iii) since the rest of the proof proceeds as above (the equivalence between (C.17) and (C.18) now follows from the definition of $\|\cdot\|_{V'}$). Since the map A is surjective, Lemma C.37 implies that A has a bounded right inverse map $A^\dagger : W \to V$. In particular, $A(A^\dagger(w)) = w$ for all $w \in W$, and there is $\alpha > 0$ such that $\alpha \|A^\dagger(w)\|_V \le \|w\|_W$. Then for all $w' \in W'$, we have

$$
\begin{aligned}
\|A^*(w')\|_{V'} &= \sup_{v \in V} \frac{|\langle A^*(w'), v \rangle_{V',V}|}{\|v\|_V} \ge \sup_{w \in W} \frac{|\langle A^*(w'), A^\dagger(w) \rangle_{V',V}|}{\|A^\dagger(w)\|_V} \\
&= \sup_{w \in W} \frac{|\langle w', w \rangle_{W',W}|}{\|A^\dagger(w)\|_V} \ge \alpha \sup_{w \in W} \frac{|\langle w', w \rangle_{W',W}|}{\|w\|_W} = \alpha \|w'\|_{W'}. \qquad \square
\end{aligned}
$$

Remark C.41 (Lions' theorem). The assertion (i) $\Leftrightarrow$ (iii) in Lemma C.40 is sometimes called *Lions' theorem.* It means that establishing the a priori estimate (C.17) is a necessary and sufficient condition to prove that the problem $A(u) = f$ has at least one solution $u \in V$ for all $f \in W$. $\square$

Lemma C.42 (Right inverse). *Let V, W be Banach spaces and let $A \in \mathcal{L}(V; W)$ be a surjective operator. Assume that V is reflexive. Then A has a bounded right inverse $A^\dagger : W \to V$ satisfying $\alpha \|A^\dagger(w)\|_V \le \|w\|_W$, where α is the same constant as in the equivalent statements (C.17) and (C.18).*

Proof. The proof is inspired from ideas by P. Azerad (private communication). Let $A \in \mathcal{L}(V; W)$ be a surjective operator. Lemma C.34(ii) shows that the adjoint operator $A^* : W' \to V'$ is injective. Let us equip the subspace $R := \mathrm{im}(A^*) \subset V'$ with the norm $\|\cdot\|_{V'}$. The injectivity of A^* implies the existence of a linear left inverse $A^{*\ddagger} : R \to W'$. Consider its adjoint $A^{*\ddagger *} : W'' \to R'$. Let $E_{R'V''}^{\mathrm{HB}}$ be one Hahn–Banach extension operator from R' to V'' (see Theorem C.13). Using the reflexivity of V to invoke the inverse of the canonical isometry $J_V : V \to V''$, we set

$$
A^\dagger := J_V^{-1} \circ E_{R'V''}^{\mathrm{HB}} \circ A^{*\ddagger *} \circ J_W : W \to V.
$$

Let us verify that $A^\dagger$ satisfies the expected properties. We have for all $(w', w) \in W' \times W$,

$$
\begin{aligned}
\langle w', A(A^\dagger(w)) \rangle_{W',W} &= \langle A^*(w'), A^\dagger(w) \rangle_{V',V} \\
&= \overline{\langle E_{R'V''}^{\mathrm{HB}}(A^{*\ddagger *}(J_W(w))), A^*(w') \rangle_{V'',V'}} \\
&= \overline{\langle A^{*\ddagger *}(J_W(w)), A^*(w') \rangle_{R',R}} = \overline{\langle J_W(w), A^{*\ddagger}(A^*(w')) \rangle_{W'',W'}} \\
&= \overline{\langle J_W(w), w' \rangle_{W'',W'}} = \langle w', w \rangle_{W',W},
\end{aligned}
$$

where to pass from the second to the third line we used that $A^*(w') \in R$. Since w' is arbitrary in W', this proves that $A \circ A^\dagger = I_W$. Moreover, since $R := \mathrm{im}(A^*)$, we infer that for all $w \in W$,

$$
\begin{aligned}
\|A^\dagger(w)\|_V &= \|A^{*\ddagger *}(J_W(w))\|_{R'} = \sup_{w' \in W'} \frac{|\langle A^{*\ddagger *}(J_W(w)), A^*(w') \rangle_{R',R}|}{\|A^*(w')\|_{V'}} \\
&= \sup_{w' \in W'} \frac{|\langle J_W(w), w' \rangle_{W'',W'}|}{\|A^*(w')\|_{V'}} \le \sup_{w' \in W'} \frac{\|w'\|_{W'}}{\|A^*(w')\|_{V'}} \|w\|_W.
\end{aligned}
$$

Since $A \in \mathcal{L}(V; W)$ is surjective, we have $\sup_{w' \in W'} \frac{\|w'\|_{W'}}{\|A^*(w')\|_{V'}} \le \alpha^{-1}$ owing to (C.17), and this shows that $\|A^\dagger(w)\|_V \le \alpha^{-1} \|w\|_W$. $\square$

Remark C.43 (Counterexample). The assumption that V be reflexive in Lemma C.42 cannot be removed if one insists on having the bound $\alpha \sup_{\|w\|_W=1} \|A^\dagger(w)\|_V \leq 1$. Let us consider the real sequence spaces ℓ^p, $p \in [1, \infty]$. Since $V := \ell^1$ is not reflexive, there exists a linear form $A : \ell^1 \to W := \mathbb{R}$ that does not attain its norm on the unit ball of V (this is James's theorem [253, Thm. 1]). Notice that $A \neq 0$, hence A is necessarily surjective. Using ℓ^2 as pivot space, it is well known that ℓ^∞ can be identified with the dual of V (see e.g., Brezis [89, Thm. 4.14]). Let t be the nonzero sequence in ℓ^∞ such that $A(v) = (v, t)_{\ell^2} := \sum_{i \in \mathbb{N}} v_i t_i$ for all $v \in V$. A simple computation shows that the adjoint $A^* : \mathbb{R} \to V' \equiv \ell^\infty$ is such that $A^*(s) = st$ for all $s \in \mathbb{R}$. Let us define $\alpha := \inf_{w' \in \mathbb{R}} \sup_{v \in \ell^1} \frac{|(A^*(w'), v)_{\ell^2}|}{|w'| \|v\|_{\ell^1}}$. We have $\alpha = \sup_{v \in \ell^1} \frac{|(t, v)_{\ell^2}|}{\|v\|_{\ell^1}} = \|A\|_{V'}$. Let $A^\dagger$ be a right inverse of A. Then for all $s \in \mathbb{R}$, we have $s = (A \circ A^\dagger)(s) = (t, A^\dagger(s))_{\ell^2}$. For all $s \in \mathbb{R} \backslash \{0\}$, $\frac{A^\dagger(s)}{\|A^\dagger(s)\|_V}$ is in the unit ball of V. Since A does not attain its norm on this ball by assumption, we infer that $|A(\frac{A^\dagger(s)}{\|A^\dagger(s)\|_V})| < \alpha$. Since $A(\frac{A^\dagger(s)}{\|A^\dagger(s)\|_V}) = \frac{1}{\|A^\dagger(s)\|_V} s$, we can rewrite the above bound as $\frac{1}{\alpha}|s| < \|A^\dagger(s)\|_V$ for all $s \in \mathbb{R} \backslash \{0\}$, that is, $1 < \alpha \sup_{\|w\|_W=1} \|A^\dagger(w)\|_V$. $\square$

We observe that nothing is said in Lemma C.42 on the linearity of the right inverse $A^\dagger$. A slightly different construction of $A^\dagger$ that guarantees linearity is possible in the Hilbertian setting.

Lemma C.44 (Right inverse in Hilbert spaces). *Let Y, Z be two nontrivial Hilbert spaces. Let $B : Y \to Z'$ be a bounded linear operator such that there exists $\beta > 0$ s.t.*

$$\|B(y)\|_{Z'} \geq \beta \|y\|_Y, \qquad \forall y \in Y. \tag{C.19}$$

Then $B^* : Z \to Y'$ has a linear right inverse $B^{*\dagger} : Y' \to Z$ such that $\|B^{*\dagger}\|_{\mathcal{L}(Y';Z)} \leq \beta^{-1}$.

Proof. Owing to Lemma C.39, the assumption (C.19) is equivalent to $B^* : Z \to Y'$ being surjective. Let us set $M := \ker(B^*)^\perp \subset Z$, where the orthogonality is defined using the inner product of Z (note that $M \neq \{0\}$ since otherwise $\ker(B^*) = Z$, i.e., $B^* = B = 0$ implying by (C.19) that $Y = \{0\}$ would be trivial). Let $J : M \to Z$ be the canonical injection, and note that $J^* : Z' \to M'$ is s.t. for all $z' \in Z'$ and all $m \in M$,

$$\langle J^*(z'), m \rangle_{M',M} = \langle z', J(m) \rangle_{Z',Z} = \langle z', m \rangle_{Z',Z}.$$

Let us set $S := J^* \circ B : Y \to M'$. Let $y' \in Y'$. The surjectivity of B^* together with the definition of M implies that there is $z := m + m^\perp \in M \oplus M^\perp = Z$ s.t. $y' = B^*(z) = B^*(m) = B^*(J(m)) = S^*(m)$, which proves that S^* is surjective. Let $m \in M$ and assume that $0 = S^*(m) = B^*(J(m)) = B^*(m)$. Then $m \in \ker(B^*) \cap \ker(B^*)^\perp$, i.e., $m = 0$, which proves that $S^* : M \to Y'$ is injective. Hence, S^* and S are isomorphisms. Moreover, since $\|(S^*)^{-1}\|_{\mathcal{L}(Y';M)} = \sup_{y' \in Y'} \frac{\|(S^*)^{-1}(y')\|_Z}{\|y'\|_{Y'}} = \sup_{m \in M} \frac{\|m\|_Z}{\|S^*(m)\|_{Y'}}$, we have

$$\|(S^*)^{-1}\|_{\mathcal{L}(Y';M)}^{-1} = \inf_{m \in M} \frac{\|S^*(m)\|_{Y'}}{\|m\|_Z} = \inf_{m \in M} \sup_{y \in Y} \frac{|\langle S^*(m), y \rangle_{Y',Y}|}{\|m\|_Z \|y\|_Y}$$

$$= \inf_{y \in Y} \sup_{m \in M} \frac{|\langle S(y), m \rangle_{M',M}|}{\|y\|_Y \|m\|_Z} = \inf_{y \in Y} \sup_{m \in M} \frac{|\langle B(y), m \rangle_{Z',Z}|}{\|y\|_Y \|m\|_Z},$$

where the first equality on the second line follows from (C.25) below and the bijectivity of S. Using that $Z = M \oplus M^\perp$ and $M^\perp = \ker(B^*)$, we obtain

$$\|(S^*)^{-1}\|_{\mathcal{L}(Y';M)}^{-1} = \inf_{y \in Y} \sup_{m \in M} \frac{|\langle B(y), m \rangle_{Z',Z}|}{\|y\|_Y \|m\|_Z}$$

$$\geq \inf_{y \in Y} \sup_{m + m^\perp \in M \oplus M^\perp} \frac{|\langle B(y), m + m^\perp \rangle_{Z',Z}|}{\|y\|_Y (\|m\|_Z^2 + \|m^\perp\|_Z^2)^{1/2}} = \beta,$$

which proves that $\|(S^*)^{-1}\|_{\mathcal{L}(Y';M)} \leq \beta^{-1}$. (Note that we actually have $\|(S^*)^{-1}\|_{\mathcal{L}(Y';M)} = \beta^{-1}$ since $\sup_{m \in M} \frac{|\langle B(y),m \rangle_{Z',Z}|}{\|m\|_Z} \leq \sup_{z \in Z} \frac{|\langle B(y),z \rangle_{Z',Z}|}{\|z\|_Z}$.) Let us now set

$$B^{*\dagger} := J \circ (B^* \circ J)^{-1} = J \circ (S^*)^{-1} : Y' \to Z.$$

Then $B^* \circ B^{*\dagger} = B^* \circ J \circ (S^*)^{-1} = S^* \circ (S^*)^{-1} = I_{Y'}$, which proves that $B^{*\dagger}$ is indeed a right inverse of B^*. Moreover, $\|B^{*\dagger}\|_{\mathcal{L}(Y',Z)} = \|J \circ (S^*)^{-1}\|_{\mathcal{L}(Y',Z)} \leq \|(S^*)^{-1}\|_{\mathcal{L}(Y',Z)} = \beta^{-1}$. $\square$

Remark C.45 (Lemma C.44 vs. Lemma C.42). Without the statement on the linearity of $B^{*\dagger}$, Lemma C.44 would be a direct consequence of Lemma C.42 applied with $A := B^*$, $V := Z$, and $W := Y'$. Indeed, the condition (C.19) implies that A is a surjective operator satisfying the inf-sup condition (C.18) with constant β. $\square$

Remark C.46 (Left inverse). The operator $B^{*\ddagger} := (J^* \circ B)^{-1} \circ J^* = S^{-1} \circ J^* : Z' \to Y$ is a left inverse of B s.t. $\|B^{*\ddagger}\|_{\mathcal{L}(Z';Y)} \leq \beta^{-1}$. $\square$

Finally, let us recall two important results on compactness.

Lemma C.47 (Peetre–Tartar). *Let X, Y, Z be Banach spaces. Let $A \in \mathcal{L}(X;Y)$ be injective and let $T \in \mathcal{L}(X;Z)$ be compact. Assume that there is $c > 0$ such that $c\|x\|_X \leq \|A(x)\|_Y + \|T(x)\|_Z$ for all $x \in X$. Then $\mathrm{im}(A)$ is closed. Equivalently, there is $\alpha > 0$ such that*

$$\alpha\|x\|_X \leq \|A(x)\|_Y, \qquad \forall x \in X. \tag{C.20}$$

Proof. Owing to Lemma C.39 and since A is injective, $\mathrm{im}(A)$ is closed iff (C.20) holds true. This inequality has already been proved in Lemma A.20 (see (A.6)). $\square$

Theorem C.48 (Schauder). *A bounded linear operator between Banach spaces is compact if and only if its adjoint is compact.*

Proof. See Brezis [89, Thm. 6.4]. $\square$

C.7 Characterization of bijectivity

The following theorem provides the theoretical foundation of the BNB theorem stated in §25.3 and which is often invoked in this book.

Theorem C.49 (Bijectivity of A). *Let $A \in \mathcal{L}(V;W)$. The following statements are equivalent:*

(i) $A : V \to W$ *is bijective.*

(ii) A *is injective, $\mathrm{im}(A)$ is closed, and $A^* : W' \to V'$ is injective.*

(iii) A^* *is injective and there exists $\alpha > 0$ such that*

$$\|A(v)\|_W \geq \alpha\|v\|_V, \qquad \forall v \in V. \tag{C.21}$$

Equivalently, A^* is injective and

$$\inf_{v \in V} \sup_{w' \in W'} \frac{|\langle w', A(v) \rangle_{W',W}|}{\|w'\|_{W'}\|v\|_V} =: \alpha > 0. \tag{C.22}$$

Proof. (1) The statements (ii) and (iii) are equivalent since (C.21) is equivalent to A injective and $\mathrm{im}(A)$ closed owing to Lemma C.39.

(2) Let us first prove that (i) implies (ii). Since A is surjective, $\ker(A^*) = \mathrm{im}(A)^\perp = \{0\}$, i.e., A^* is injective. Since $\mathrm{im}(A) = W$ is closed and A is injective, this yields (ii). Finally, to prove that (ii) implies (i), we only need to prove that (ii) implies the surjectivity of A. The injectivity of A^* implies that $\overline{\mathrm{im}(A)} = (\ker(A^*))^\perp = W$. Since $\mathrm{im}(A)$ is closed, $\mathrm{im}(A) = W$, i.e., A is surjective. $\square$

Corollary C.50 (Self-adjoint bijective operator). *Assume that V is reflexive. Let $A \in \mathcal{L}(V; V')$ be a self-adjoint operator. Then A is bijective iff there is a real number $\alpha > 0$ such that*

$$\|A(v)\|_{V'} \geq \alpha \|v\|_V, \qquad \forall v \in V. \tag{C.23}$$

Proof. Owing to Theorem C.49, the bijectivity of A implies that A satisfies the inequality (C.23). Conversely, (C.23) means that A is injective. It follows that A^* is injective since $A^* = A \circ J_V^{-1}$ owing to the reflexivity hypothesis. The bijectivity of A then follows from Theorem C.49(iii). $\square$

Let $A \in \mathcal{L}(V; W)$ be a bijective operator. We have seen in Corollary C.38 that $A^{-1} \in \mathcal{L}(W; V)$. We can now characterize more precisely the constants associated with the boundedness of A^{-1} and the closedness of its range.

Lemma C.51 (Bounds on A^{-1}). *Let $A \in \mathcal{L}(V; W)$ be a bijective operator. Then $\|A^{-1}\|_{\mathcal{L}(W;V)} = \alpha^{-1}$ with α defined in (C.22), and*

$$\inf_{w \in W} \frac{\|A^{-1}(w)\|_V}{\|w\|_W} = \inf_{w \in W} \sup_{v' \in V'} \frac{|\langle v', A^{-1}(w)\rangle_{V',V}|}{\|v'\|_{V'} \|w\|_W} = \|A\|_{\mathcal{L}(V;W)}^{-1}. \tag{C.24}$$

Proof. (1) Using the bijectivity of A, we have

$$\left(\sup_{w \in W} \frac{\|A^{-1}(w)\|_V}{\|w\|_W} \right)^{-1} = \left(\sup_{v \in V} \frac{\|v\|_V}{\|A(v)\|_W} \right)^{-1}$$

$$= \inf_{v \in V} \frac{\|A(v)\|_W}{\|v\|_V} = \inf_{v \in V} \sup_{w' \in W'} \frac{|\langle w', A(v)\rangle_{W',W}|}{\|w'\|_{W'} \|v\|_V},$$

which shows using (C.22) that $\|A^{-1}\|_{\mathcal{L}(W;V)} = \alpha^{-1}$.

(2) Similarly, we have

$$\left(\inf_{w \in W} \frac{\|A^{-1}(w)\|_V}{\|w\|_W} \right)^{-1} = \sup_{v \in V} \frac{\|A(v)\|_W}{\|v\|_V} = \|A\|_{\mathcal{L}(V;W)}.$$

Since $\|A^{-1}(w)\|_V = \sup_{v' \in V'} \frac{|\langle v', A^{-1}(w)\rangle_{V',V}|}{\|v'\|_{V'}}$ owing to Corollary C.14, this proves the inf-sup condition in (C.24). $\square$

Let us finish this section with some useful results concerning the bijectivity of the adjoint operator and some bounds on its inverse.

Corollary C.52 (Bijectivity of A^*). *Let $A \in \mathcal{L}(V; W)$ and consider its adjoint $A^* \in \mathcal{L}(W'; V')$. Then A is bijective if and only if A^* is bijective.*

Proof. Assume first that A is bijective. Since A is injective and $\mathrm{im}(A) = W$, the equivalence of Items (i) and (ii) in Lemma C.39 implies that A^* is surjective. Since A is surjective, the equivalence of Items (i) and (ii) in Lemma C.40 implies that A^* is injective. Hence, A^* is bijective. The converse statement is proved by invoking the same arguments. $\square$

Lemma C.53 (Inf-sup condition). *Let $A \in \mathcal{L}(V; W)$ be a bijective operator. Assume that V is reflexive. The following holds true:*

$$\inf_{v \in V} \sup_{w' \in W'} \frac{|\langle w', A(v) \rangle_{W',W}|}{\|w'\|_{W'} \|v\|_V} = \inf_{w' \in W'} \sup_{v \in V} \frac{|\langle w', A(v) \rangle_{W',W}|}{\|w'\|_{W'} \|v\|_V}. \tag{C.25}$$

In other words, the inf-sup constant of $A \in \mathcal{L}(V; W)$ on $V \times W'$ is equal to the inf-sup constant of $A^* \in \mathcal{L}(W'; V')$ on $W' \times V$.

Proof. The left-hand side, l, and the right-hand side, r, in (C.25) are two positive finite numbers since A is a bijective bounded operator. The left-hand side being equal to l means that l is the largest number such that $\|A(v)\|_W \geq l \|v\|_V$ for all v in V. Let $w' \in W'$ and $w \in W$. Since A is surjective, we can consider its right inverse $A^\dagger$, and the previous statement regarding l implies that $l \|A^\dagger(w)\|_V \leq \|w\|_W$. Since $A(A^\dagger(w)) = w$, this implies that

$$\|w'\|_{W'} = \sup_{w \in W} \frac{|\langle w', w \rangle_{W',W}|}{\|w\|_W} = \sup_{w \in W} \frac{|\langle A^*(w'), A^\dagger(w) \rangle_{V',V}|}{\|w\|_W}$$

$$\leq \|A^*(w')\|_{V'} \sup_{w \in W} \frac{\|A^\dagger(w)\|_V}{\|w\|_W} \leq \frac{1}{l} \|A^*(w')\|_{V'} = \frac{1}{l} \sup_{v \in V} \frac{|\langle w', A(v) \rangle_{W',W}|}{\|v\|_V}.$$

Taking the infimum w.r.t. $w' \in W'$ proves that $l \leq r$. The converse inequality $r \leq l$ is proved similarly by working with W' in lieu of V, V' in lieu of W and A^* in lieu of A (notice that A^* is bijective owing to Corollary C.52), leading to

$$\inf_{w' \in W'} \sup_{v'' \in V''} \frac{|\langle v'', A^*(w') \rangle_{V'',V'}|}{\|v''\|_{V''} \|w'\|_{W'}} \leq \inf_{v'' \in V''} \sup_{w' \in W'} \frac{|\langle v'', A^*(w') \rangle_{V'',V'}|}{\|v''\|_{V''} \|w'\|_{W'}}.$$

Owing to the reflexivity of V, this inequality becomes $r \leq l$. $\square$

Remark C.54 (Counterexample). The identity (C.25) can fail if $A \neq 0$ is not bijective. For instance, if $A : (x_0, x_1, x_2 \ldots) \mapsto (0, x_0, x_1, x_2, \ldots)$ is the right shift operator in ℓ^2, then $A^* : (x_0, x_1, x_2 \ldots) \mapsto (x_1, x_2, x_3, \ldots)$ is the left shift operator. It can be verified that A is injective but not surjective, whereas A^* is surjective but not injective. Using the notation of the proof of Lemma C.53, it can also be shown that $l = 1$ and $r = 0$. $\square$

Lemma C.55 (Bounds on A^{-*}). *Let $A \in \mathcal{L}(V; W)$ be a bijective operator. Assume that V is reflexive. Let $A^* \in \mathcal{L}(W'; V')$ be the adjoint of A and let $A^{-*} \in \mathcal{L}(V'; W')$ denote its inverse. Then $\|A^{-*}\|_{\mathcal{L}(V';W')} = \alpha^{-1}$ with α defined in (C.22), and*

$$\inf_{v' \in V'} \frac{\|A^{-*}(v')\|_{W'}}{\|v'\|_{V'}} = \inf_{v' \in V'} \sup_{w \in W} \frac{|\langle A^{-*}(v'), w \rangle_{W',W}|}{\|v'\|_{V'} \|w\|_W} = \frac{1}{\|A\|_{\mathcal{L}(V;W)}}. \tag{C.26}$$

Proof. Notice that the notation A^{-*} is meant to reflect that $(A^{-1})^* = (A^*)^{-1}$. Combining the results from Lemma C.30 and Lemma C.51, we infer that $\|A^{-*}\|_{\mathcal{L}(V';W')} = \|A^{-1}\|_{\mathcal{L}(W;V)} = \alpha^{-1}$. Moreover, the first equality in (C.26) follows from the definition of $\|\cdot\|_{W'}$ and the second one from $\langle A^{-*}(v'), w \rangle_{W',W} = \langle v', A^{-1}(w) \rangle_{V',V}$, the identity (C.25) (since A^{-1} is bijective), and the identity (C.24). $\square$

C.8 Coercive operators

We now focus on the more specific class of coercive operators. The notion of coercivity plays a central role in the analysis of PDEs involving the Laplace operator, and more generally elliptic operators (see Chapter 31).

Definition C.56 (Coercive operator). *Let V be a complex Banach space. The operator $A \in \mathcal{L}(V; V')$ is said to be a* coercive *if there exist a real number $\alpha > 0$ and a complex number $\xi \in \mathbb{C}$ with $|\xi| = 1$ such that*

$$\Re\big(\xi \langle A(v), v\rangle_{V',V}\big) \geq \alpha \|v\|_V^2, \qquad \forall v \in V. \tag{C.27}$$

In the real case, we have either $\xi = 1$ or $\xi = -1$.

Remark C.57 (Self-adjoint case). Let A be a coercive self-adjoint operator (see Definition C.31). Since $\langle A(v), v\rangle_{V',V}$ is real for all $v \in V$, coercivity means that $\Re(\xi)\langle A(v), v\rangle_{V',V} \geq \alpha\|v\|_V^2$. Thus, up to rescaling α, one can always take either $\xi = 1$ or $\xi = -1$ when A is self-adjoint. $\square$

The coercivity condition is sometimes defined as follows: There exists a real number $\alpha > 0$ such that $|\langle A(v), v\rangle_{V',V}| \geq \alpha\|v\|_V^2$ for all $v \in V$. Although this variant looks slightly more general since $\Re\big(\xi\langle A(v), v\rangle_{V',V}\big) \leq |\langle A(v), v\rangle_{V',V}|$, it is equivalent to (C.27). More precisely, we have the following result.

Lemma C.58 (Real part vs. module). *Let $\alpha > 0$ and let V be a Hilbert space. The following two statements are equivalent:* (i) *$|\langle A(v), v\rangle_{V',V}| \geq \alpha\|v\|_V^2$ for all $v \in V$.* (ii) *There is $\xi \in \mathbb{C}$ with $|\xi| = 1$ s.t. (C.27) holds true.*

Proof. Let us prove the claim in the real case. It suffices to show that the statement (i) implies that $\langle A(v), v\rangle_{V',V}$ has always the same sign for all nonzero $v \in V$. Reasoning by contradiction, if there are nonzero $v, w \in V$ such that $\langle A(v), v\rangle_{V',V} < 0$ and $\langle A(w), w\rangle_{V',V} > 0$, then the second-order polynomial $\mathbb{R} \ni \lambda \mapsto \langle A(v + \lambda w), v + \lambda w\rangle_{V',V} \in \mathbb{R}$ has at least one root $\lambda_* \in \mathbb{R}$. The statement (i) yields $v + \lambda_* w = 0$, so that $\langle A(v), v\rangle_{V',V} = \lambda_*^2 \langle A(w), w\rangle_{V',V} > 0$, which contradicts $\langle A(v), v\rangle_{V',V} < 0$. We refer the reader to Brezis [89, p. 366] for the proof in the complex case (see also Exercise 46.9 for a proof of the Hausdorff–Toeplitz theorem). $\square$

It turns out that the notion of coercivity is relevant *only in Hilbert spaces*.

Proposition C.59 (Hilbert structure). *Let V be a Banach space. V can be equipped with a Hilbert structure with the same topology if and only if there is a coercive operator in $\mathcal{L}(V; V')$.*

Proof. Setting $((v, w))_V := \frac{1}{2}(\xi\langle A(v), w\rangle_{V',V} + \overline{\xi\langle A(w), v\rangle_{V',V}})$, we define a sesquilinear form on $V \times V$ that is Hermitian. The coercivity and boundedness of A imply that

$$\alpha \|v\|_V^2 \leq ((v, v))_V \leq \|A\|_{\mathcal{L}(V;V')} \|v\|_V^2,$$

for all $v \in V$. This shows positive definiteness (so that $((\cdot, \cdot))_V$ is an inner product in V) and that the induced norm is equivalent to $\|\cdot\|_V$. $\square$

Corollary C.60 (Coercivity as a sufficient condition). *If the operator $A \in \mathcal{L}(V; V')$ is coercive, then it is bijective.*

Proof. This is the Lax–Milgram lemma which is proved in §25.2. $\square$

Definition C.61 (Monotone operator). *The operator $A \in \mathcal{L}(V; V')$ is said to be* monotone *if*

$$\Re\big(\langle A(v), v\rangle_{V',V}\big) \geq 0, \qquad \forall v \in V. \tag{C.28}$$

Corollary C.62 (Coercivity as a necessary and sufficient condition). *Assume that V is reflexive. Let $A \in \mathcal{L}(V; V')$ be a monotone self-adjoint operator. Then A is bijective iff it is coercive (with $\xi := 1$).*

Proof. See Exercise 25.7. $\square$

From now on, we assume that V is a Hilbert space. If the operator $A \in \mathcal{L}(V; V')$ is coercive (and therefore bijective), its inverse $A^{-1} \in \mathcal{L}(V'; V)$ turns out to be coercive as well. Indeed, using the coercivity of A and the lower bound on A^{-1} resulting from (C.24), we infer that for all $\phi \in V'$,

$$
\begin{aligned}
\Re\big(\xi\langle \phi, A^{-1}(\phi)\rangle_{V',V}\big) &= \Re\big(\xi\langle A(A^{-1}(\phi)), A^{-1}(\phi)\rangle_{V',V}\big) \\
&\geq \alpha \|A^{-1}(\phi)\|_V^2 \geq \frac{\alpha}{\|A\|^2}\|\phi\|_{V'}^2,
\end{aligned}
\tag{C.29}
$$

with the shorthand notation $\|A\| := \|A\|_{\mathcal{L}(V;V')}$. The following results provide more precise characterizations of the coercivity constant of A^{-1}.

Lemma C.63 (Coercivity of A^{-1}, self-adjoint case). *Let $A \in \mathcal{L}(V; V')$ be a self-adjoint coercive operator (i.e., (C.27) holds true with either $\xi = 1$ or $\xi = -1$ according to Remark C.57). Then A^{-1} is coercive with coercivity constant $\|A\|^{-1}$, and we have more precisely*

$$
\inf_{\phi \in V'} \frac{\xi\langle \phi, A^{-1}(\phi)\rangle_{V',V}}{\|\phi\|_{V'}^2} = \frac{1}{\|A\|}.
\tag{C.30}
$$

Proof. Assume that $\xi = 1$ (the case $\xi = -1$ is identical). The coercivity of A together with $A = A^*$ implies that $((v,w))_A := \langle A(v), w\rangle_{V',V}$ is an inner product on V. Let $v \in V$ and $\phi \in V'$. Since $\langle \phi, v\rangle_{V',V} = ((A^{-1}(\phi), v))_A$, the Cauchy–Schwarz inequality implies that

$$
\begin{aligned}
\Re\big(\langle \phi, v\rangle_{V',V}\big) &\leq ((v,v))_A^{\frac{1}{2}} ((A^{-1}(\phi), A^{-1}(\phi)))_A^{\frac{1}{2}} \\
&= \langle A(v), v\rangle_{V',V}^{\frac{1}{2}} \langle \phi, A^{-1}(\phi)\rangle_{V',V}^{\frac{1}{2}} \leq \|A\|^{\frac{1}{2}} \|v\|_V \langle \phi, A^{-1}(\phi)\rangle_{V',V}^{\frac{1}{2}},
\end{aligned}
$$

where we used the boundedness of A. This implies that

$$
\|\phi\|_{V'} = \sup_{v \in V} \frac{|\langle \phi, v\rangle_{V',V}|}{\|v\|_V} \leq \|A\|^{\frac{1}{2}} \langle \phi, A^{-1}(\phi)\rangle_{V',V}^{\frac{1}{2}}.
$$

Taking the infimum over $\phi \in V'$, we infer that

$$
\frac{1}{\|A\|} \leq \inf_{\phi \in V'} \frac{\langle \phi, A^{-1}(\phi)\rangle_{V',V}}{\|\phi\|_{V'}^2} \leq \inf_{\phi \in V'} \sup_{\psi \in V'} \frac{|\langle \psi, A^{-1}(\phi)\rangle_{V',V}|}{\|\psi\|_{V'}\|\phi\|_{V'}} = \frac{1}{\|A\|},
$$

where the last equality follows from (C.24). Thus, all the terms are equal, and this concludes the proof. $\qquad\square$

Let us now consider the case where the operator $A \in \mathcal{L}(V; V')$ is not necessarily self-adjoint. Since V is Hilbert space, V is reflexive. Hence, the adjoint of A is $A^* \in \mathcal{L}(V; V')$, and we have $\langle A^*(v), w\rangle_{V',V} = \overline{\langle A(w), v\rangle_{V',V}}$.

Lemma C.64 (Coercivity of A^{-1}, general case). *Let $A \in \mathcal{L}(V; V')$ be a coercive operator with parameters $\alpha > 0$ and $\xi \in \mathbb{C}$ with $|\xi| = 1$. Let the self-adjoint part of ξA be defined as $(\xi A)_s := \frac{1}{2}(\xi A + (\xi A)^*) = \frac{1}{2}(\xi A + \overline{\xi}A^*)$. The following holds true:*

$$
\frac{\alpha}{\|A\|^2} \leq \inf_{\phi \in V'} \frac{\Re\big(\xi\langle \phi, A^{-1}(\phi)\rangle_{V',V}\big)}{\|\phi\|_{V'}^2} \leq \frac{1}{\|(\xi A)_s\|}.
\tag{C.31}
$$

Proof. The lower bound in (C.31) is a restatement of (C.29). To establish the upper bound, let us set $B := \xi A$. Then B are B_s are coercive (and therefore invertible) operators since

$$
\langle B_s(v), v\rangle_{V',V} = \Re\big(\langle B(v), v\rangle_{V',V}\big) = \Re\big(\xi\langle A(v), v\rangle_{V',V}\big) \geq \alpha \|v\|_V^2,
$$

for all $v \in V$. A direct calculation shows that

$$
\begin{aligned}
B^{-1}(B - B_\mathrm{s})B_\mathrm{s}^{-1}(B^* - B_\mathrm{s})B^{-*} &= (B_\mathrm{s}^{-1} - B^{-1})(I - B_\mathrm{s}B^{-*}) \\
&= B_\mathrm{s}^{-1} - B^{-1} - B^{-*} + B^{-1}B_\mathrm{s}B^{-*} \\
&= B_\mathrm{s}^{-1} - B^{-1} - B^{-*} + \frac{1}{2}B^{-1}(B + B^*)B^{-*} \\
&= B_\mathrm{s}^{-1} - \frac{1}{2}(B^{-1} + B^{-*}).
\end{aligned}
$$

This implies that for all $\phi \in V'$,

$$
\begin{aligned}
\langle \phi, B_\mathrm{s}^{-1}(\phi) \rangle_{V',V} &= \frac{1}{2}\langle \phi, (B^{-1} + B^{-*})(\phi) \rangle_{V',V} + \langle \phi, B^{-1}(B - B_\mathrm{s})B_\mathrm{s}^{-1}(B^* - B_\mathrm{s})B^{-*}(\phi) \rangle_{V',V} \\
&= \Re\big(\langle \phi, B^{-1}(\phi) \rangle_{V',V}\big) + \langle \psi, B_\mathrm{s}^{-1}(\psi) \rangle_{V',V} \geq \Re\big(\langle \phi, B^{-1}(\phi) \rangle_{V',V}\big),
\end{aligned}
$$

with $\psi := (B^* - B_\mathrm{s})B^{-*}(\phi)$ and where we used that $\langle \psi, B_\mathrm{s}^{-1}(\psi) \rangle_{V',V} \geq 0$. Applying Lemma C.63 to the operator B_s, which is coercive and self-adjoint, we conclude that

$$
\frac{1}{\|B_\mathrm{s}\|} = \inf_{\phi \in V'} \frac{\langle \phi, B_\mathrm{s}^{-1}(\phi) \rangle_{V',V}}{\|\phi\|_{V'}^2} \geq \inf_{\phi \in V'} \frac{\Re\big(\langle \phi, B^{-1}(\phi) \rangle_{V',V}\big)}{\|\phi\|_{V'}^2}.
$$

Since $\langle \phi, B^{-1}(\phi) \rangle_{V',V} = (\overline{\xi})^{-1}\langle \phi, A^{-1}(\phi) \rangle_{V',V}$ and $(\overline{\xi})^{-1} = \xi$, this proves the upper bound in (C.31). $\qquad\square$

Bibliography

[1] M. Abbas, A. Ern, and N. Pignet. Hybrid high-order methods for finite deformations of hyperelastic materials. *Comput. Mech.*, 62(4):909–928, 2018. pages 221

[2] M. Abbas, A. Ern, and N. Pignet. A hybrid high-order method for incremental associative plasticity with small deformations. *Comput. Methods Appl. Mech. Engrg.*, 346:891–912, 2019. pages 221

[3] M. Abramowitz and I. Stegun. *Handbook of mathematical functions.* Dover Publications Inc., New York, NY, 9th edition, 1972. pages 70

[4] Y. Achdou, C. Bernardi, and F. Coquel. A priori and a posteriori analysis of finite volume discretizations of Darcy's equations. *Numer. Math.*, 96(1):17–42, 2003. pages 151

[5] L. M. Adams and H. F. Jordan. Is SOR color-blind? *SIAM J. Sci. Statist. Comput.*, 7(2):490–506, 1986. pages 66

[6] S. Agmon, A. Douglis, and L. Nirenberg. Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions. I. *Comm. Pure Appl. Math.*, 12:623–727, 1959. pages 89

[7] M. Ainsworth and J. T. Oden. *A posteriori error estimation in finite element analysis.* Wiley, New York, NY, 2000. pages 335

[8] M. Amara and J. M. Thomas. Equilibrium finite elements for the linear elastic problem. *Numer. Math.*, 33(4):367–383, 1979. pages 219

[9] C. Amrouche and R. Ratsimahalo. Conditions "inf sup" dans les espaces de Banach non réflexifs. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(12):1069–1072, 2000. pages 16

[10] C. Amrouche, C. Bernardi, M. Dauge, and V. Girault. Vector potentials in three-dimensional non-smooth domains. *Math. Methods Appl. Sci.*, 21(9):823–864, 1998. pages 229, 237

[11] R. Andreev. Stability of sparse space-time finite element discretizations of linear parabolic evolution equations. *IMA J. Numer. Anal.*, 33(1):242–260, 2013. pages 31

[12] P. F. Antonietti, A. Buffa, and I. Perugia. Discontinuous Galerkin approximation of the Laplace eigenproblem. *Comput. Methods Appl. Mech. Engrg.*, 195(25):3483–3503, 2006. pages 285

[13] T. Arbogast and Z. Chen. On the implementation of mixed methods as nonconforming methods for second-order elliptic problems. *Math. Comp.*, 64(211):943–972, 1995. pages 330

[14] J. H. Argyris and S. Kelsey. *Energy theorems and structural analysis*. Buttelworths, London, UK, third edition, 1967. Originally published in Aircraft Engrg., 26, pp. 347–356, 383–387, 394, 410–422, (1954) and 27, pp. 42–58, 80–94, 125–134, 145–158 (1955). pages 212

[15] D. N. Arnold. An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, 19:742–760, 1982. pages 169, 171

[16] D. N. Arnold and F. Brezzi. Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates. *RAIRO Modél. Math. Anal. Numér.*, 19(1):7–32, 1985. pages 327, 330

[17] D. N. Arnold and R. Winther. Mixed finite elements for elasticity. *Numer. Math.*, 92(3): 401–419, 2002. pages 219

[18] D. N. Arnold, I. Babuška, and J. Osborn. Finite element methods: principles for their selection. *Comput. Methods Appl. Mech. Engrg.*, 45(1-3):57–96, 1984. pages 30

[19] D. N. Arnold, F. Brezzi, and J. Douglas, Jr. PEERS: a new mixed finite element for plane elasticity. *Japan J. Appl. Math.*, 1(2):347–367, 1984. pages 219

[20] D. N. Arnold, F. Brezzi, and M. Fortin. A stable finite element for the Stokes equations. *Calcolo*, 21:337–344, 1984. pages 356

[21] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.*, 39(5):1749–1779, 2001/02. pages 177

[22] D. N. Arnold, D. Boffi, and R. S. Falk. Approximation by quadrilateral finite elements. *Math. Comp.*, 71(239):909–922, 2002. pages 374

[23] D. N. Arnold, R. S. Falk, and R. Winther. Finite element exterior calculus, homological techniques, and applications. *Acta Numer.*, 15:1–155, 2006. pages 237, 241

[24] D. N. Arnold, R. S. Falk, and R. Winther. Mixed finite element methods for linear elasticity with weakly imposed symmetry. *Math. Comp.*, 76(260):1699–1723, 2007. pages 219

[25] D. N. Arnold, G. Awanou, and R. Winther. Finite elements for symmetric tensors in three dimensions. *Math. Comp.*, 77(263):1229–1251, 2008. pages 219

[26] D. N. Arnold, R. S. Falk, and R. Winther. Finite element exterior calculus: from Hodge theory to numerical stability. *Bull. Amer. Math. Soc. (N.S.)*, 47(2):281–354, 2010. pages 237

[27] F. Assous, P. Ciarlet, Jr., and S. Labrunie. *Mathematical foundations of computational electromagnetism*, volume 198 of *Applied Mathematical Sciences*. Springer, Cham, Switzerland, 2018. pages 225

[28] J.-P. Aubin. Behavior of the error of the approximate solutions of boundary value problems for linear elliptic operators by Galerkin's and finite difference methods. *Ann. Scuola Norm. Sup. Pisa (3)*, 21:599–637, 1967. pages 98

[29] J.-P. Aubin. *Applied functional analysis*. Pure and Applied Mathematics. Wiley-Interscience, New York, NY, second edition, 2000. With exercises by B. Cornet and J.-M. Lasry, Translated from the French by C. Labrousse. pages 98, 377

[30] P. Auscher, E. Russ, and P. Tchamitchian. Hardy Sobolev spaces on strongly Lipschitz domains of $\mathbb{R}^n$. *J. Funct. Anal.*, 218(1):54–109, 2005. pages 341

[31] O. Axelsson and A. Kucherov. Real valued iterative methods for solving complex symmetric linear systems. *Numer. Linear Algebra Appl.*, 7(4):197–218, 2000. pages 55

[32] B. Ayuso de Dios, K. Lipnikov, and G. Manzini. The nonconforming virtual element method. *ESAIM Math. Model. Numer. Anal.*, 50(3):879–904, 2016. pages 183

[33] I. Babuška. Error-bounds for finite element method. *Numer. Math.*, 16:322–333, 1970/1971. pages 14, 27

[34] I. Babuška. The finite element method with Lagrangian multipliers. *Numer. Math.*, 20: 179–192, 1973. pages 295, 325

[35] I. Babuška. The finite element method with penalty. *Math. Comp.*, 27:221–228, 1973. pages 107

[36] I. Babuška and A. Miller. A feedback finite element method with a posteriori error estimation. I. The finite element method and some basic properties of the a posteriori error estimator. *Comput. Methods Appl. Mech. Engrg.*, 61(1):1–40, 1987. pages 120, 335

[37] I. Babuška and J. Osborn. Analysis of finite element methods for second order boundary value problems using mesh dependent norms. *Numer. Math.*, 34(1):41–62, 1980. pages 99

[38] I. Babuška and J. Osborn. Eigenvalue problems. In *Handbook of Numerical Analysis, Vol. II*, pages 641–787. North-Holland, Amsterdam, The Netherlands, 1991. pages 277, 278, 279

[39] I. Babuška and W. C. Rheinbolt. Error estimates for adaptive finite element method computations. *SIAM J. Numer. Anal.*, 15:736–754, 1978. pages 121

[40] I. Babuška and M. Suri. Locking effects in the finite element approximation of elasticity problems. *Numer. Math.*, 62(4):439–463, 1992. pages 217

[41] C. Bacuta. A unified approach for Uzawa algorithms. *SIAM J. Numer. Anal.*, 44(6):2633–2649, 2006. pages 310

[42] C. Bacuta. Sharp stability and approximation estimates for symmetric saddle point systems. *Appl. Anal.*, 95(1):226–237, 2016. pages 292, 293

[43] S. Badia and R. Codina. A nodal-based finite element approximation of the Maxwell problem suitable for singular solutions. *SIAM J. Numer. Anal.*, 50(2):398–417, 2012. pages 252

[44] G. A. Baker. Finite element methods for elliptic equations using nonconforming elements. *Math. Comp.*, 31(137):45–59, 1977. pages 169

[45] G. R. Barrenechea, E. Burman, and F. Karakatsani. Edge-based nonlinear diffusion for finite element approximations of convection-diffusion equations and its relation to algebraic flux-correction schemes. *Numer. Math.*, 135(2):521–545, 2017. pages 111

[46] F. Bassi and S. Rebay. A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations. *J. Comput. Phys.*, 131(2): 267–279, 1997. pages 174

[47] P. Bastian and B. Rivière. Superconvergence and $H(\mathrm{div})$ projection for discontinuous Galerkin methods. *Internat. J. Numer. Methods Fluids*, 42(10):1043–1057, 2003. pages 331

[48] R. Becker and R. Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numerica*, 10:1–102, 2001. pages 123

[49] H. Beirão da Veiga. On a stationary transport equation. *Ann. Univ. Ferrara Sez. VII*, 32: 79–91, 1986. pages 92

[50] L. Beirão da Veiga, F. Brezzi, A. Cangiani, G. Manzini, L. D. Marini, and A. Russo. Basic principles of virtual element methods. *M3AS Math. Models Methods Appl. Sci.*, 199(23): 199–214, 2013. pages 185

[51] M. Benzi and M. A. Olshanskii. An augmented Lagrangian-based approach to the Oseen problem. *SIAM J. Sci. Comput.*, 28(6):2095–2113, 2006. pages 311

[52] M. Benzi, G. H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numer.*, 14:1–137, 2005. pages 310, 311

[53] M. Benzi, M. A. Olshanskii, and Z. Wang. Modified augmented Lagrangian preconditioners for the incompressible Navier-Stokes equations. *Internat. J. Numer. Methods Fluids*, 66(4): 486–508, 2011. pages 311

[54] M. Bercovier and O. Pironneau. Error estimates for finite element solution of the Stokes problem in the primitive variables. *Numer. Math.*, 33:211–224, 1979. pages 353, 358, 359

[55] C. Bernardi and R. Verfürth. Adaptive finite element methods for elliptic equations with non-smooth coefficients. *Numer. Math.*, 85(4):579–608, 2000. pages 91

[56] P. Binev, W. Dahmen, and R. DeVore. Adaptive finite element methods with convergence rates. *Numer. Math.*, 97(2):219–268, 2004. pages 126, 127, 128

[57] M. S. Birman and M. Z. Solomyak. L_2-theory of the Maxwell operator in arbitrary domains. *Russian Mathematical Surveys*, 42(6):75, 1987. pages 229

[58] Å. Björck. *Numerical methods in matrix computations*, volume 59 of *Texts in Applied Mathematics*. Springer, Cham, Switzerland, 2015. pages 53

[59] J. Blechta, J. Málek, and M. Vohralík. Localization of the $W^{-1,q}$ norm for local a posteriori efficiency. *IMA J. Numer. Anal.*, 40(2):914–950, 2020. pages 120

[60] D. Boffi. Three-dimensional finite element methods for the Stokes problem. *SIAM J. Numer. Anal.*, 34(2):664–670, 1997. pages 358

[61] D. Boffi. A note on the de Rham complex and a discrete compactness property. *Appl. Math. Lett.*, 14(1):33–38, 2001. pages 237

[62] D. Boffi. Finite element approximation of eigenvalue problems. *Acta Numer.*, 19:1–120, 2010. pages 279

[63] D. Boffi, F. Brezzi, and L. Gastaldi. On the problem of spurious eigenvalues in the approximation of linear elliptic problems in mixed form. *Math. Comp.*, 69(229):121–140, 2000. pages 285

[64] D. Boffi, F. Brezzi, and M. Fortin. Reduced symmetry elements in linear elasticity. *Commun. Pure Appl. Anal.*, 8(1):95–121, 2009. pages 219

[65] D. Boffi, F. Brezzi, and M. Fortin. *Mixed finite element methods and applications*, volume 44 of *Springer Series in Computational Mathematics*. Springer, Heidelberg, Germany, 2013. pages 330, 351, 352, 358, 366, 367, 368, 374

[66] M. E. Bogovskiĭ. Solutions of some problems of vector analysis, associated with the operators div and grad. In *Theory of cubature formulas and the application of functional analysis to problems of mathematical physics*, volume 1980 of *Trudy Sem. S. L. Soboleva, No. 1*, pages 5–40, 149. Akad. Nauk SSSR Sibirsk. Otdel. Inst. Mat., Novosibirsk, Russia, 1980. pages 341

[67] J. M. Boland and R. A. Nicolaides. Stability of finite elements under divergence constraints. *SIAM J. Numer. Anal.*, 20(4):722–731, 1983. pages 366

[68] J. M. Boland and R. A. Nicolaides. Stability and semistable low order finite elements for viscous flows. *SIAM J. Numer. Anal.*, 22:474–492, 1985. pages 347

[69] A. Bonito and J.-L. Guermond. Approximation of the eigenvalue problem for the time harmonic Maxwell system by continuous Lagrange finite elements. *Math. Comp.*, 80(276): 1887–1910, 2011. pages 249, 252

[70] A. Bonito, J.-L. Guermond, and F. Luddens. Regularity of the Maxwell equations in heterogeneous media and Lipschitz domains. *J. Math. Anal. Appl.*, 408:498–512, 2013. pages 91, 235

[71] A. Bonito, J.-L. Guermond, and F. Luddens. An interior penalty method with C^0 finite elements for the approximation of the Maxwell equations in heterogeneous media: convergence analysis with minimal regularity. *ESAIM Math. Model. Numer. Anal.*, 50(5):1457–1489, 2016. pages 249, 252

[72] A.-S. Bonnet-Ben Dhia, P. Ciarlet, Jr., and C. M. Zwölf. Time harmonic wave diffraction problems in materials with sign-shifting coefficients. *J. Comput. Appl. Math.*, 234(6):1912–1919, 2010. pages 15

[73] A.-S. Bonnet-Ben Dhia, L. Chesnel, and P. Ciarlet, Jr. T-coercivity for scalar interface problems between dielectrics and metamaterials. *ESAIM Math. Model. Numer. Anal.*, 46(6): 1363–1387, 2012. pages 15

[74] A. Bossavit. *Computational electromagnetism, variational formulations, complementary, edge elements*, volume 2 of *Electromagnetism*. Academic Press, New York, NY, 1998. pages 225, 226

[75] M. Botti, D. A. Di Pietro, and P. Sochala. A hybrid high-order method for nonlinear elasticity. *SIAM J. Numer. Anal.*, 55(6):2687–2717, 2017. pages 221

[76] D. Braess and J. Schöberl. Equilibrated residual error estimator for edge elements. *Math. Comp.*, 77(262):651–672, 2008. pages 333, 335

[77] D. Braess, V. Pillwein, and J. Schöberl. Equilibrated residual error estimates are p-robust. *Comput. Methods Appl. Mech. Engrg.*, 198(13-14):1189–1197, 2009. pages 333, 334

[78] J. H. Bramble and J. E. Osborn. Rate of convergence estimates for nonselfadjoint eigenvalue approximations. *Math. Comp.*, 27:525–549, 1973. pages 279

[79] J. H. Bramble and J. E. Pasciak. A new approximation technique for div-curl systems. *Math. Comp.*, 73(248):1739–1762, 2004. pages 252

[80] J. H. Bramble, J. E. Pasciak, and J. Xu. Parallel multilevel preconditioners. *Math. Comp.*, 55(191):1–22, 1990. pages 54

[81] J. H. Bramble, J. E. Pasciak, and P. S. Vassilevski. Computational scales of Sobolev norms with application to preconditioning. *Math. Comp.*, 69(230):463–480, 2000. pages 51, 54

[82] J. H. Bramble, T. V. Kolev, and J. E. Pasciak. The approximation of the Maxwell eigenvalue problem using a least-squares method. *Math. Comp.*, 74(252):1575–1598, 2005. pages 252

[83] J. Brandts, S. Korotov, and M. Křížek. The discrete maximum principle for linear simplicial finite element approximations of a reaction-diffusion problem. *Linear Algebra Appl.*, 429(10): 2344–2357, 2008. pages 111

[84] J. Brandts, S. Korotov, M. Křížek, and J. Šolc. On nonobtuse simplicial partitions. *SIAM Rev.*, 51(2):317–335, 2009. pages 110

[85] H. Brass and K. Petras. *Quadrature theory*, volume 178 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2011. pages 70

[86] S. C. Brenner. Poincaré-Friedrichs inequalities for piecewise H^1 functions. *SIAM J. Numer. Anal.*, 41(1):306–324, 2003. pages 171

[87] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, NY, third edition, 2008. pages 27

[88] S. C. Brenner and L.-Y. Sung. Linear finite element methods for planar linear elasticity. *Math. Comp.*, 59(200):321–338, 1992. pages 218

[89] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, NY, 2011. pages 108, 253, 256, 257, 258, 377, 378, 379, 380, 381, 382, 383, 384, 387, 388, 391

[90] F. Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from Lagrange multipliers. *RAIRO, Anal. Num.*, pages 129–151, 1974. pages 295

[91] F. Brezzi and K.-J. Bathe. A discourse on the stability conditions for mixed finite element formulations. *Comput. Methods Appl. Mech. Engrg.*, 82(1-3):27–57, 1990. pages 366, 376

[92] F. Brezzi and R. S. Falk. Stability of a higher-order Hood–Taylor method. *SIAM J. Numer. Anal.*, 28:581–590, 1991. pages 358, 374

[93] F. Brezzi, G. Manzini, L. D. Marini, P. Pietra, and A. Russo. Discontinuous Galerkin approximations for elliptic problems. *Numer. Methods Partial Differential Equations*, 16(4): 365–378, 2000. pages 174

[94] W. L. Briggs. *A multigrid tutorial*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1987. pages 54

[95] A. Buffa and C. Ortner. Compact embeddings of broken Sobolev spaces and applications. *IMA J. Numer. Anal.*, 4(29):827–855, 2009. pages 171, 176

[96] A. Buffa, P. Ciarlet, Jr., and E. Jamelot. Solving electromagnetic eigenvalue problems in polyhedral domains with nodal finite elements. *Numer. Math.*, 113:497–518, 2009. pages 251

[97] E. Burman. Robust error estimates in weak norms for advection dominated transport problems with rough data. *Math. Models Methods Appl. Sci.*, 24(13):2663–2684, 2014. pages 92

[98] E. Burman and A. Ern. Discrete maximum principle for Galerkin approximations of the Laplace operator on arbitrary meshes. *C. R. Math. Acad. Sci. Paris*, 338(8):641–646, 2004. pages 111

[99] E. Burman and A. Ern. Stabilized Galerkin approximation of convection-diffusion-reaction equations: discrete maximum principle and convergence. *Math. Comp.*, 74(252):1637–1652, 2005. pages 111

[100] E. Burman and A. Ern. Discontinuous Galerkin approximation with discrete variational principle for the nonlinear Laplacian. *C. R. Math. Acad. Sci. Paris*, 346(17–18):1013–1016, 2008. pages 176

[101] E. Burman and P. Zunino. A domain decomposition method for partial differential equations with non-negative form based on interior penalties. *SIAM J. Numer. Anal.*, 44:1612–1638, 2006. pages 200

[102] E. Burman, H. Wu, and L. Zhu. Linear continuous interior penalty finite element method for Helmholtz equation with high wave number: one-dimensional analysis. *Numer. Methods Partial Differential Equations*, 32(5):1378–1410, 2016. pages 142

[103] V. Calo, M. Cicuttin, Q. Deng, and A. Ern. Spectral approximation of elliptic operators by the hybrid high-order method. *Math. Comp.*, 88(318):1559–1586, 2019. pages 285

[104] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík. Guaranteed and robust a posteriori bounds for Laplace eigenvalues and eigenvectors: conforming approximations. *SIAM J. Numer. Anal.*, 55(5):2228–2254, 2017. pages 270

[105] C. Canuto. Eigenvalue approximations by mixed methods. *RAIRO Anal. Numér.*, 12(1):27–50, 1978. pages 285

[106] S. Caorsi, P. Fernandes, and M. Raffetto. On the convergence of Galerkin finite element approximations of electromagnetic eigenproblems. *SIAM J. Numer. Anal.*, 38(2):580–607, 2000. pages 237

[107] C. Carstensen and S. A. Funken. Fully reliable localized error control in the FEM. *SIAM J. Sci. Comput.*, 21(4):1465–1484, 1999/00. pages 335

[108] C. Carstensen and S. A. Funken. Constants in Clément-interpolation error and residual based a posteriori error estimates in finite element methods. *East-West J. Numer. Math.*, 8 (3):153–175, 2000. pages 120

[109] C. Carstensen and J. Gedicke. Guaranteed lower bounds for eigenvalues. *Math. Comp.*, 83 (290):2605–2629, 2014. pages 285

[110] C. Carstensen and F. Hellwig. Low-order discontinuous Petrov-Galerkin finite element methods for linear elasticity. *SIAM J. Numer. Anal.*, 54(6):3388–3410, 2016. pages 218

[111] C. Carstensen, L. Demkowicz, and J. Gopalakrishnan. A posteriori error control for DPG methods. *SIAM J. Numer. Anal.*, 52(3):1335–1353, 2014. pages 25, 313

[112] C. Carstensen, M. Feischl, M. Page, and D. Praetorius. Axioms of adaptivity. *Comput. Math. Appl.*, 67(6):1195–1253, 2014. pages 126, 128

[113] J. M. Cascón, C. Kreuzer, R. H. Nochetto, and K. G. Siebert. Quasi-optimal convergence rate for an adaptive finite element method. *SIAM J. Numer. Anal.*, 46(5):2524–2550, 2008. pages 125, 126, 127, 128

[114] J. Céa. Approximation variationnelle des problèmes aux limites. *Ann. Inst. Fourier (Grenoble)*, 14(2):345–444, 1964. pages 27

[115] S. N. Chandler-Wilde, D. P. Hewett, and A. Moiola. Interpolation of Hilbert and Sobolev spaces: quantitative estimates and counterexamples. *Mathematika*, 61(2):414–443, 2015. pages 88

[116] F. Chatelin. *Spectral approximation of linear operators*. Computer Science and Applied Mathematics. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, NY, 1983. With a foreword by P. Henrici, with solutions to exercises by M. Ahués. pages 253, 279

[117] Z. Chen and H. Chen. Pointwise error estimates of discontinuous Galerkin methods with penalty for second-order elliptic problems. *SIAM J. Numer. Anal.*, 42(3):1146–1166, 2004. pages 174

[118] L. Chesnel and P. Ciarlet, Jr. *T*-coercivity and continuous Galerkin methods: application to transmission problems with sign changing coefficients. *Numer. Math.*, 124(1):1–29, 2013. pages 19

[119] E. V. Chizhonkov and M. A. Olshanskii. On the domain geometry dependence of the LBB condition. *M2AN Math. Model. Numer. Anal.*, 34(5):935–951, 2000. pages 304

[120] P. Ciarlet, Jr. *T*-coercivity: application to the discretization of Helmholtz-like problems. *Comput. Math. Appl.*, 64(1):22–34, 2012. pages 139

[121] P. Ciarlet, Jr. On the approximation of electromagnetic fields by edge finite elements. Part 1: Sharp interpolation results for low-regularity fields. *Comput. Math. Appl.*, 71(1):85–104, 2016. pages 231, 235

[122] P. Ciarlet, Jr. and M. Vohralík. Localization of global norms and robust a posteriori error control for transmission problems with sign-changing coefficients. *ESAIM Math. Model. Numer. Anal.*, 52(5):2037–2064, 2018. pages 118, 120

[123] P. G. Ciarlet. *Mathematical elasticity II: Theory of plates*, volume 27 of *Studies in Mathematics and its Applications*. Elsevier, Amsterdam, 1997. pages 215

[124] P. G. Ciarlet. *The finite element method for elliptic problems*, volume 40 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. Reprint of the 1978 original [North-Holland, Amsterdam, The Netherlands]. pages 27, 114

[125] P. G. Ciarlet and P. Ciarlet, Jr. Another approach to linearized elasticity and a new proof of Korn's inequality. *Math. Models Methods Appl. Sci.*, 15(2):259–271, 2005. pages 216

[126] P. G. Ciarlet and P.-A. Raviart. The combined effect of curved boundaries and numerical integration in isoparametric finite element methods. In *The mathematical foundations of the finite element method with applications to partial differential equations (Proc. Sympos., Univ. Maryland, Baltimore, MD, 1972)*, pages 409–474. Academic Press, New York, NY, 1972. pages 114

[127] P. G. Ciarlet and P.-A. Raviart. Maximum principle and uniform convergence for the finite element method. *Comput. Methods Appl. Mech. Engrg.*, 2:17–31, 1973. pages 109

[128] M. Cicuttin, D. A. Di Pietro, and A. Ern. Implementation of discontinuous skeletal methods on arbitrary-dimensional, polytopal meshes using generic programming. *J. Comput. Appl. Math.*, 344:852–874, 2018. pages 183

[129] R. W. Clough. The finite element method in plane stress analysis. In *Proc. 2nd ASCE Conference on Electronic Computation*, Pittsburgh, PA, 1960. pages 212

[130] B. Cockburn. Static condensation, hybridization, and the devising of the HDG methods. In G. R. Barrenechea, F. Brezzi, A. Cangiani, and E. H. Georgoulis, editors, *Building bridges: Connections and challenges in modern approaches to numerical partial differential equations*, volume 114 of *Lecture Notes in Computational Science and Engineering*, pages 129–178, Cham, Switzerland, 2016. Springer. pages 186, 188, 330

[131] B. Cockburn and J. Gopalakrishnan. A characterization of hybridized mixed methods for second order elliptic problems. *SIAM J. Numer. Anal.*, 42(1):283–301, 2004. pages 330

[132] B. Cockburn and K. Shi. Devising HDG methods for Stokes flow: An overview. *Comput. & Fluids*, 98:221–229, 2014. pages 373

[133] B. Cockburn, D. Schötzau, and J. Wang. Discontinuous Galerkin methods for incompressible elastic materials. *Comput. Methods Appl. Mech. Engrg.*, 195(25-28):3184–3204, 2006. pages 218

[134] B. Cockburn, J. Gopalakrishnan, and R. Lazarov. Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems. *SIAM J. Numer. Anal.*, 47(2):1319–1365, 2009. pages 188

[135] B. Cockburn, J. Gopalakrishnan, and F.-J. Sayas. A projection-based error analysis of HDG methods. *Math. Comp.*, 79(271):1351–1367, 2010. pages 188

[136] B. Cockburn, W. Qiu, and K. Shi. Conditions for superconvergence of HDG methods for second-order elliptic problems. *Math. Comp.*, 81(279):1327–1353, 2012. pages 188

[137] B. Cockburn, D. A. Di Pietro, and A. Ern. Bridging the hybrid high-order and hybridizable discontinuous Galerkin methods. *ESAIM Math. Model Numer. Anal.*, 50(3):635–650, 2016. pages 183, 188

[138] A. Cohen, R. DeVore, and R. H. Nochetto. Convergence rates of AFEM with H^{-1} data. *Found. Comput. Math.*, 12(5):671–718, 2012. pages 118, 120

[139] R. Cools. Monomial cubature rules since "Stroud": a compilation. II. *J. Comput. Appl. Math.*, 112(1-2):21–27, 1999. pages 70

[140] R. Cools and P. Rabinowitz. Monomial cubature rules since "Stroud": a compilation. *J. Comput. Appl. Math.*, 48(3):309–326, 1993. pages 70

[141] H. O. Cordes. Vereinfachter Beweis der Existenz einer Apriori-Hölderkonstanten. *Math. Ann.*, 138:155–178, 1959. pages 80

[142] M. Costabel. A remark on the regularity of solutions of Maxwell's equations on Lipschitz domains. *Math. Methods Appl. Sci.*, 12(4):365–368, 1990. pages 229

[143] M. Costabel. A coercive bilinear form for Maxwell's equations. *J. Math. Anal. Appl.*, 157 (2):527–541, 1991. pages 251

[144] M. Costabel and M. Dauge. Maxwell and Lamé eigenvalues on polyhedra. *Math. Methods Appl. Sci.*, 22(3):243–258, 1999. pages 251

[145] M. Costabel and M. Dauge. Weighted regularization of Maxwell equations in polyhedral domains. A rehabilitation of nodal finite elements. *Numer. Math.*, 93(2):239–277, 2002. pages 251

[146] M. Costabel and A. McIntosh. On Bogovskiĭ and regularized Poincaré integral operators for de Rham complexes on Lipschitz domains. *Math. Z.*, 265(2):297–320, 2010. pages 341

[147] M. Costabel, M. Dauge, and S. Nicaise. Analytic regularity for linear elliptic systems in polygons and polyhedra. *Math. Models Methods Appl. Sci.*, 22(8):1250015, 63 pp., 2012. pages 90

[148] M. Costabel, M. Dauge, and S. Nicaise. Weighted analytic regularity in polyhedra. *Comput. Math. Appl.*, 67(4):807–817, 2014. pages 90

[149] J.-P. Croisille. Finite volume box schemes and mixed methods. *M2AN Math. Model. Numer. Anal.*, 34(2):1087–1106, 2000. pages 326

[150] J.-P. Croisille and I. Greff. Some nonconforming mixed box schemes for elliptic problems. *Numer. Methods Partial Differential Equations*, 18(3):355–373, 2002. pages 150

[151] M. Crouzeix and P.-A. Raviart. Conforming and nonconforming finite element methods for solving the stationary Stokes equations. I. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge*, 7(R-3):33–75, 1973. pages 146, 356, 368, 373

[152] M. Dauge. *Elliptic boundary value problems on corner domains*, volume 1341 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1988. pages 89, 90, 91

[153] M. Dauge. Stationary Stokes and Navier-Stokes systems on two- or three-dimensional domains with corners. I. Linearized equations. *SIAM J. Math. Anal.*, 20(1):74–97, 1989. pages 342

[154] R. Dautray and J.-L. Lions. *Mathematical analysis and numerical methods for science and technology. Vol. 4. Integral equations and numerical methods*. Springer-Verlag, Berlin, Germany, 1990. pages 114

[155] P. J. Davis and P. Rabinowitz. *Methods of numerical integration*. Computer Science and Applied Mathematics. Academic Press, New York, NY, 1975. pages 70

[156] T. A. Davis. *Direct methods for sparse linear systems*, volume 2 of *Fundamentals of Algorithms*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006. pages 53, 61

[157] C. Dawson, S. Sun, and M. F. Wheeler. Compatible algorithms for coupled flow and transport. *Comp. Meth. Appl. Mech. Eng.*, 193:2565–2580, 2004. pages 174

[158] L. F. Demkowicz and J. Gopalakrishnan. An overview of the discontinuous Petrov Galerkin method. In *Recent developments in discontinuous Galerkin finite element methods for partial differential equations*, volume 157 of *The IMA Volumes in Mathematics and its Applications*, pages 149–180. Springer, Cham, Switzerland, 2014. pages 313

[159] A. Demlow, D. Leykekhman, A. H. Schatz, and L. B. Wahlbin. Best approximation property in the W_∞^1 norm for finite element methods on graded meshes. *Math. Comp.*, 81(278): 743–764, 2012. pages 95

[160] J. W. Demmel, S. C. Eisenstat, J. R. Gilbert, X. S. Li, and J. W. H. Liu. A supernodal approach to sparse partial pivoting. *SIAM J. Matrix Analysis and Applications*, 20(3):720–755, 1999. pages 53

[161] J. Descloux, N. Nassif, and J. Rappaz. On spectral approximation. I. The problem of convergence. *RAIRO Anal. Numér.*, 12(2):97–112, 1978. pages 279

[162] J. Descloux, N. Nassif, and J. Rappaz. On spectral approximation. II. Error estimates for the Galerkin method. *RAIRO Anal. Numér.*, 12(2):113–119, 1978. pages 279

[163] P. Destuynder and B. Métivet. Explicit error bounds in a conforming finite element method. *Math. Comp.*, 68(228):1379–1396, 1999. pages 333, 335

[164] D. A. Di Pietro and A. Ern. Discrete functional analysis tools for discontinuous Galerkin methods with application to the incompressible Navier-Stokes equations. *Math. Comp.*, 79 (271):1303–1330, 2010. pages 171, 176

[165] D. A. Di Pietro and A. Ern. *Mathematical aspects of discontinuous Galerkin methods*, volume 69 of *Mathématiques & Applications [Mathematics & Applications]*. Springer-Verlag, Berlin, 2012. pages 177, 178

[166] D. A. Di Pietro and A. Ern. A hybrid high-order locking-free method for linear elasticity on general meshes. *Comput. Meth. Appl. Mech. Engrg.*, 283:1–21, 2015. pages 183, 218, 221, 373

[167] D. A. Di Pietro, A. Ern, and J.-L. Guermond. Discontinuous Galerkin methods for anisotropic semi-definite diffusion with advection. *SIAM J. Numer. Anal.*, 46(2):805–831, 2008. pages 200

[168] D. A. Di Pietro, A. Ern, and S. Lemaire. An arbitrary-order and compact-stencil discretization of diffusion on general meshes based on local reconstruction operators. *Comput. Meth. Appl. Math.*, 14(4):461–472, 2014. pages 183, 190

[169] D. A. Di Pietro, A. Ern, A. Linke, and F. Schieweck. A discontinuous skeletal method for the viscosity-dependent Stokes problem. *Comput. Methods Appl. Mech. Engrg.*, 306:175–195, 2016. pages 373, 374

[170] M. Dobrowolski. On the LBB constant on stretched domains. *Math. Nachr.*, 254/255:64–67, 2003. pages 304

[171] W. Dörfler. A convergent adaptive algorithm for Poisson's equation. *SIAM J. Numer. Anal.*, 33(3):1106–1124, 1996. pages 126

[172] J. Droniou, R. Eymard, T. Gallouët, C. Guichard, and R. Herbin. *The gradient discretisation method*, volume 82 of *Mathématiques & Applications [Mathematics & Applications]*. Springer, Cham, Switzerland, 2018. pages 36

[173] M. Dryja. On discontinuous Galerkin methods for elliptic problems with discontinuous co-efficients. *Comput. Methods Appl. Math.*, 3(1):76–85, 2003. pages 200

[174] M. Dryja, J. Galvis, and M. Sarkis. BDDC methods for discontinuous Galerkin discretization of elliptic problems. *J. Complexity*, 23(4-6):715–739, 2007. pages 200

[175] H. Duan, P. Lin, and R. C. E. Tan. C^0 elements for generalized indefinite Maxwell equations. *Numer. Math.*, 122(1):61–99, 2012. pages 252

[176] H. Duan, R. C. E. Tan, S.-Y. Yang, and C.-S. You. A mixed H^1-conforming finite element method for solving Maxwell's equations with non-H^1 solution. *SIAM J. Sci. Comput.*, 40 (1):A224–A250, 2018. pages 252

[177] I. S. Duff, A. M. Erisman, and J. K. Reid. *Direct methods for sparse matrices*. Monographs on Numerical Analysis. The Clarendon Press, Oxford University Press, New York, 1986. pages 53

[178] D. A. Dunavant. High degree efficient symmetrical Gaussian quadrature rules for the triangle. *Internat. J. Numer. Methods Engrg.*, 21(6):1129–1148, 1985. pages 70

[179] N. Dunford and J. T. Schwartz. *Linear operators. I. General theory*, volume 7 of *Pure and Applied Mathematics*. Interscience Publishers, Inc., New York, NY, 1958. With the assistance of W. G. Bade and R. G. Bartle. pages 253

[180] R. Durán, M. A. Muschietti, E. Russ, and P. Tchamitchian. Divergence operator and Poincaré inequalities on arbitrary bounded domains. *Complex Var. Elliptic Equ.*, 55(8-10): 795–816, 2010. pages 341

[181] R. G. Durán and M. A. Muschietti. An explicit right inverse of the divergence operator which is continuous in weighted norms. *Studia Math.*, 148(3):207–219, 2001. pages 341

[182] R. G. Durán, L. Gastaldi, and C. Padra. A posteriori error estimators for mixed approxima-tions of eigenvalue problems. *Math. Models Methods Appl. Sci.*, 9(8):1165–1178, 1999. pages 285

[183] G. Duvaut and J.-L. Lions. *Les inéquations en mécanique et en physique*. Dunod, Paris, 1972. pages 215

[184] H. Egger and C. Waluga. *hp* analysis of a hybrid DG method for Stokes flow. *IMA J. Numer. Anal.*, 33(2):687–721, 2013. pages 373

[185] H. C. Elman, D. J. Silvester, and A. J. Wathen. *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2005. pages 54, 310, 311

[186] A. Ern and J.-L. Guermond. Evaluation of the condition number in linear systems arising in finite element approximations. *M2AN Math. Model. Numer. Anal.*, 40(1):29–48, 2006. pages 51

[187] A. Ern and J.-L. Guermond. A converse to Fortin's lemma in Banach spaces. *C. R. Math. Acad. Sci. Paris*, 354(11):1092–1095, 2016. pages 25, 352

[188] A. Ern and J.-L. Guermond. Analysis of the edge finite element approximation of the Maxwell equations with low regularity solutions. *Comput. Math. Appl.*, 75(3):918–932, 2018. pages 233

[189] A. Ern and M. Vohralík. A posteriori error estimation based on potential and flux reconstruction for the heat equation. *SIAM J. Numer. Anal.*, 48(1):198–223, 2010. pages 333

[190] A. Ern and M. Vohralík. Adaptive inexact Newton methods with a posteriori stopping criteria for nonlinear diffusion PDEs. *SIAM J. Sci. Comput.*, 35(4):A1761–A1791, 2013. pages 333, 334

[191] A. Ern and M. Vohralík. Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations. *SIAM J. Numer. Anal.*, 53(2):1058–1081, 2015. pages 333, 334

[192] A. Ern and M. Vohralík. Stable broken H^1 and $H(\mathrm{div})$ polynomial extensions for polynomial-degree-robust potential and flux reconstruction in three space dimensions. *Math. Comp.*, 89 (322):551–594, 2020. pages 333, 334

[193] A. Ern, S. Nicaise, and M. Vohralík. An accurate $\mathbf{H}(\mathrm{div})$ flux reconstruction for discontinuous Galerkin approximations of elliptic problems. *C. R. Math. Acad. Sci. Paris*, 345(12):709–712, 2007. pages 177

[194] A. Ern, A. F. Stephansen, and P. Zunino. A discontinuous Galerkin method with weighted averages for advection-diffusion equations with locally small and anisotropic diffusivity. *IMA J. Numer. Anal.*, 29(2):235–256, 2009. pages 200

[195] S. Esterhazy and J. M. Melenk. On stability of discretizations of the Helmholtz equation. In *Numerical analysis of multiscale problems*, volume 83 of *Lect. Notes Comput. Sci. Eng.*, pages 285–324. Springer, Heidelberg, 2012. pages 136

[196] L. C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1998. pages 89, 108

[197] R. Eymard, T. Gallouët, and R. Herbin. Convergence of finite volume schemes for semilinear convection diffusion equations. *Numer. Math.*, 82(1):91–116, 1999. pages 119

[198] V. Faber and T. Manteuffel. Necessary and sufficient conditions for the existence of a conjugate gradient method. *SIAM J. Numer. Anal.*, 21(2):352–362, 1984. pages 53

[199] R. S. Falk. Nonconforming finite element methods for the equations of linear elasticity. *Math. Comp.*, 57(196):529–550, 1991. pages 218

[200] X. Feng and H. Wu. Discontinuous Galerkin methods for the Helmholtz equation with large wave number. *SIAM J. Numer. Anal.*, 47(4):2872–2896, 2009. pages 142

[201] M. Fortin. An analysis of the convergence of mixed finite element methods. *RAIRO Anal. Num.*, 11:341–354, 1977. pages 25

[202] M. Fortin. A three-dimensional quadratic nonconforming element. *Numer. Math.*, 46(2): 269–279, 1985. pages 218

[203] M. Fortin and R. Glowinski. *Augmented Lagrangian methods*, volume 15 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, The Netherlands, 1983. Translated from the French by B. Hunt and D. C. Spicer. pages 308

[204] M. Fortin and M. Soulié. A non-conforming piecewise quadratic finite element on triangles. *Internat. J. Numer. Methods Engrg.*, 19:505–520, 1983. pages 157, 218

[205] B. Fraejis de Veubeke. Diffusion des inconnues hyperstatiques dans les voilures à longerons couplés. *Bull. Serv. Technique de l'Aéronautique*, 24:1–56, 1951. pages 216

[206] B. Fraejis de Veubeke. Stress function approach. In *World Congress on the Finite Element Method in Structural Mechanics*, pages J.1–J.51, Bournemouth, UK, 1975. Available at http://orbi.ulg.ac.be/handle/2268/205875. pages 219

[207] B. Fraejis de Veubeke. Displacement and equilibrium models in the finite element method. *Internat. J. Numer. Methods Engrg.*, 52(3):287–342, 2001. pages 327

[208] R. W. Freund. Conjugate gradient-type methods for linear systems with complex symmetric coefficient matrices. *SIAM J. Sci. Statist. Comput.*, 13(1):425–448, 1992. pages 55

[209] G. Fu, B. Cockburn, and H. Stolarski. Analysis of an HDG method for linear elasticity. *Internat. J. Numer. Methods Engrg.*, 102(3-4):551–575, 2015. pages 218

[210] G. P. Galdi. *An Introduction to the mathematical theory of the Navier-Stokes equations. Vol. I*, volume 38 of *Springer Tracts in Natural Philosophy*. Springer-Verlag, New York, NY, 1994. pages 341

[211] L. Gastaldi and R. H. Nochetto. Sharp maximum norm error estimates for general mixed finite element approximations to second order elliptic equations. *RAIRO Modél. Math. Anal. Numér.*, 23(1):103–128, 1989. pages 324

[212] G. N. Gatica. *A simple introduction to the mixed finite element method*. Springer Briefs in Mathematics. Springer, Cham, Switzerland, 2014. pages 321

[213] I. Gelfand. Zur Theorie der Charaktere der Abelschen topologischen Gruppen. *Rec. Math. [Mat. Sbornik] N. S.*, 9 (51):49–50, 1941. pages 254

[214] A. George and J. W. H. Liu. *Computer solution of large sparse positive definite systems*. Prentice-Hall Series in Computational Mathematics. Prentice-Hall Inc., Englewood Cliffs, NJ, 1981. pages 53, 61, 64

[215] A. George, J. R. Gilbert, and J. W. H. Liu, editors. *Graph theory and sparse matrix computation*, volume 56 of *The IMA Volumes in Mathematics and its Applications*. Springer-Verlag, New York, NY, 1993. pages 61, 64

[216] D. Gilbarg and N. S. Trudinger. *Elliptic partial differential equations of second order*. Classics in Mathematics. Springer-Verlag, Berlin, 2001. Reprint of the 1998 edition. pages 108

[217] V. Girault and P.-A. Raviart. *Finite element methods for Navier–Stokes equations. Theory and algorithms*. Springer Series in Computational Mathematics. Springer-Verlag, Berlin, Germany, 1986. pages 25, 340, 347, 351, 358, 366

[218] G. H. Golub and C. F. van Loan. *Matrix computations*. John Hopkins University Press, Baltimore, MD, second edition, 1989. pages 52, 307

[219] J. Gopalakrishnan and W. Qiu. An analysis of the practical DPG method. *Math. Comp.*, 83 (286):537–552, 2014. pages 313

[220] J. Gopalakrishnan, F. Li, N.-C. Nguyen, and J. Peraire. Spectral approximations by the HDG method. *Math. Comp.*, 84(293):1037–1059, 2015. pages 285

[221] A. Greenbaum. *Iterative methods for solving linear systems*, volume 17 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997. pages 310

[222] I. Greff. *Schémas boîte : étude théorique et numérique.* PhD thesis, University of Metz, France, 2003. pages 157

[223] P. Grisvard. *Elliptic problems in nonsmooth domains*, volume 24 of *Monographs and Studies in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1985. pages 89, 90, 165

[224] P. Grisvard. *Singularities in boundary value problems*, volume 22 of *Recherches en Mathématiques Appliquées [Research in Applied Mathematics]*. Masson, Paris, France, 1992. pages 89

[225] C. Grossmann and H.-G. Roos. *Numerical treatment of partial differential equations.* Universitext. Springer, Berlin, 2007. Translated and revised from the 3rd (2005) German edition by Martin Stynes. pages 51

[226] T. Gudi. A new error analysis for discontinuous finite element methods for linear elliptic problems. *Math. Comp.*, 79(272):2169–2189, 2010. pages 41, 206

[227] J.-L. Guermond. A finite element technique for solving first-order PDEs in L^p. *SIAM J. Numer. Anal.*, 42(2):714–737, 2004. pages 6

[228] J.-L. Guermond and B. Popov. An optimal L^1-minimization algorithm for stationary Hamilton-Jacobi equations. *Commun. Math. Sci.*, 7(1):211–238, 2009. pages 6

[229] B. Guo and I. Babuška. Regularity of the solutions for elliptic problems on nonsmooth domains in $\mathbb{R}^3$. I. Countably normed spaces on polyhedral domains. *Proc. Roy. Soc. Edinburgh Sect. A*, 127(1):77–126, 1997. pages 90

[230] B. Guo and I. Babuška. Regularity of the solutions for elliptic problems on nonsmooth domains in $\mathbb{R}^3$. II. Regularity in neighbourhoods of edges. *Proc. Roy. Soc. Edinburgh Sect. A*, 127(3):517–545, 1997. pages 90

[231] K. Gustafson. The Toeplitz-Hausdorff theorem for linear operators. *Proc. Amer. Math. Soc.*, 25:203–204, 1970. pages 264

[232] R. J. Guyan. Reduction of stiffness and mass matrices. *J. Am. Inst. Aeron. and Astro.*, 3: 380, 1965. pages 47

[233] J. Guzmán. Pointwise error estimates for discontinuous Galerkin methods with lifting operators for elliptic problems. *Math. Comp.*, 75(255):1067–1085, 2006. pages 174

[234] J. Guzmán, D. Leykekhman, J. Rossmann, and A. H. Schatz. Hölder estimates for Green's functions on convex polyhedral domains and their applications to finite element methods. *Numer. Math.*, 112(2):221–243, 2009. pages 95

[235] W. Hackbusch. *Multigrid methods and applications*, volume 4 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, Germany, 1985. pages 51, 54

[236] J. Hadamard. *Le problème de Cauchy et les équations aux dérivées partielles linéaires hyperboliques*. Hermann, Paris, France, 1932. pages 12

[237] P. C. Hammer and A. H. Stroud. Numerical integration over simplexes. *Math. Tables Aids Comput.*, 10:137–139, 1956. pages 70

[238] A. Hannukainen, R. Stenberg, and M. Vohralík. A unified framework for a posteriori error estimation for the Stokes problem. *Numer. Math.*, 122(4):725–769, 2012. pages 333

[239] P. Hansbo and M. G. Larson. Discontinuous Galerkin methods for incompressible and nearly incompressible elasticity by Nitsche's method. *Comput. Methods Appl. Mech. Engrg.*, 191 (17-18):1895–1908, 2002. pages 218

[240] P. Hansbo and M. G. Larson. Discontinuous Galerkin and the Crouzeix-Raviart element: application to elasticity. *M2AN Math. Model. Numer. Anal.*, 37(1):63–72, 2003. pages 218

[241] E. Hellinger. Die allgemeinen Ansätze der Mechanik der Kontinua. In F. Klein and C. Müller, editors, *Enzyklopädie der mathematischen Wissenschaften*, volume 4, pages 601–694. Teubner, Leipzig, 1914. pages 216

[242] M. R. Hestenes and E. Stiefel. Method of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Standards*, 49:409–436, 1952. pages 53

[243] U. Hetmaniuk. Stability estimates for a class of Helmholtz problems. *Commun. Math. Sci.*, 5(3):665–678, 2007. pages 133, 136, 138

[244] R. Hiptmair. Finite elements in computational electromagnetism. *Acta Numerica*, 11:237–339, 1 2002. pages 237

[245] I. Hlaváček, J. Haslinger, J. Nečas, and J. Lovíšek. *Solution of variational inequalities in mechanics*, volume 66 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1988. Translated from the Slovak by J. Jarník. pages 335

[246] C. O. Horgan. Korn's inequalities and their applications in continuum mechanics. *SIAM Rev.*, 37:491–511, 1995. pages 223

[247] L. Hörmander. *The analysis of linear partial differential operators. III.* Classics in Mathematics. Springer, Berlin, 2007. Pseudo-differential operators, Reprint of the 1994 edition. pages 80

[248] H.-C. Hu. On some variational methods on the theory of elasticity and the theory of plasticity. *Scientia Sinica*, 4:33–54, 1955. pages 216

[249] N. Hu, X. Guo, and I. N. Katz. Bounds for eigenvalues and condition numbers in the p-version of the finite element method. *Math. Comp.*, 67(224):1423–1450, 1998. pages 51

[250] T. J. R. Hughes, G. Engel, L. Mazzei, and M. G. Larson. The continuous Galerkin method is locally conservative. *J. Comput. Phys.*, 163(2):467–488, 2000. pages 333

[251] F. Ihlenburg and I. Babuška. Finite element solution of the Helmholtz equation with high wave number. I. The h-version of the FEM. *Comput. Math. Appl.*, 30(9):9–37, 1995. pages 137, 138, 142, 143

[252] B. M. Irons. Structural eigenvalue problems – elimination of unwanted variables. *AIAA J.*, 3(5):961–962, 1965. pages 47

[253] R. C. James. Reflexivity and the supremum of linear functionals. *Ann. of Math. (2)*, 66: 159–169, 1957. pages 387

[254] Y. Jeon, E.-J. Park, and D. Sheen. A hybridized finite element method for the Stokes problem. *Computers and Mathematics with Applications*, 68:2222–2232, 2014. pages 373

[255] D. S. Jerison and C. E. Kenig. The Neumann problem in Lipschitz domains. *Bull. Amer. Math. Soc. (N.S.)*, 4(2):203–207, 1981. pages 90

[256] D. S. Jerison and C. E. Kenig. The inhomogeneous Dirichlet problem in Lipschitz domains. *J. Func. Anal.*, 130(1):161–219, 1995. pages 90

[257] X. Jiang and R. H. Nochetto. Effect of numerical integration for elliptic obstacle problems. *Numer. Math.*, 67(4):501–512, 1994. pages 109

[258] F. Jochmann. An H^s-regularity result for the gradient of solutions to elliptic equations with mixed boundary conditions. *J. Math. Anal. Appl.*, 238:429–450, 1999. pages 91

[259] F. Jochmann. Regularity of weak solutions of Maxwell's equations with mixed boundary-conditions. *Math. Methods Appl. Sci.*, 22(14):1255–1274, 1999. pages 235

[260] L. John, M. Neilan, and I. Smears. Stable discontinuous Galerkin FEM without penalty parameters. In *Numerical Mathematics and Advanced Applications ENUMATH 2015*, volume 112 of *Lecture Notes in Computational Science and Engineering*, pages 165–173. Springer, Cham, Switzerland, 2016. pages 176, 191

[261] V. John, A. Linke, C. Merdon, M. Neilan, and L. G. Rebholz. On the divergence constraint in mixed finite element methods for incompressible flows. *SIAM Rev.*, 59(3):492–544, 2017. pages 346

[262] M. Juntunen and R. Stenberg. Nitsche's method for general boundary conditions. *Math. Comp.*, 78(267):1353–1374, 2009. pages 160

[263] H. Kanayama, R. Motoyama, K. Endo, and F. Kikuchi. Three-dimensional magnetostatic analysis using Nédélec's elements. *IEEE T. Magn.*, 26:682–685, 1990. pages 289

[264] G. Kanschat and R. Rannacher. Local error analysis of the interior penalty discontinuous Galerkin method for second order elliptic problems. *J. Numer. Math.*, 10(4):249–274, 2002. pages 174

[265] P. Keast. Moderate-degree tetrahedral quadrature formulas. *Comput. Methods Appl. Mech. Engrg.*, 55(3):339–348, 1986. pages 70

[266] R. B. Kellogg and J. E. Osborn. A regularity result for the Stokes problem in a convex polygon. *J. Functional Analysis*, 21(4):397–431, 1976. pages 342

[267] F. Kikuchi. On a discrete compactness property for the Nédélec finite elements. *J. Fac. Sci. Univ. Tokyo Sect. IA Math.*, 36(3):479–490, 1989. pages 237

[268] K. Y. Kim. A posteriori error estimators for locally conservative methods of nonlinear elliptic problems. *Appl. Numer. Math.*, 57(9):1065–1080, 2007. pages 177

[269] K. Y. Kim. Guaranteed a posteriori error estimator for mixed finite element methods of linear elasticity with weak stress symmetry. *SIAM J. Numer. Anal.*, 48(6):2364–2385, 2011. pages 223

[270] V. A. Kondrat′ev. Boundary value problems for elliptic equations in domains with conical or angular points. *Trudy Moskov. Mat. Obšč.*, 16:209–292, 1967. pages 90

[271] E. Kreyszig. *Introductory functional analysis with applications.* John Wiley & Sons, New York-London-Sydney, 1978. pages 253, 254, 257, 258

[272] Y. A. Kuznetsov. Efficient iterative solvers for elliptic finite element problems on nonmatching grids. *Russian J. Numer. Anal. Math. Modelling*, 10(3):187–211, 1995. pages 311

[273] P. Ladevèze and D. Leguillon. Error estimate procedure in the finite element method and applications. *SIAM J. Numer. Anal.*, 20(3):485–509, 1983. pages 335

[274] M. G. Larson and A. J. Niklasson. Analysis of a nonsymmetric discontinuous Galerkin method for elliptic problems: stability and energy error estimates. *SIAM J. Numer. Anal.*, 42(1):252–264, 2004. pages 174

[275] M. G. Larson and A. J. Niklasson. A conservative flux for the continuous Galerkin method based on discontinuous enrichment. *Calcolo*, 41(2):65–76, 2004. pages 333

[276] J. E. Lavery. Nonoscillatory solution of the steady-state inviscid Burgers' equation by mathematical programming. *J. Comput. Phys.*, 79(2):436–448, 1988. pages 6

[277] J. E. Lavery. Solution of steady-state one-dimensional conservation laws by mathematical programming. *SIAM J. Numer. Anal.*, 26(5):1081–1089, 1989. pages 6

[278] P. D. Lax. *Functional analysis.* Pure and Applied Mathematics. Wiley-Interscience [John Wiley & Sons], New York, NY, 2002. pages 253, 257, 258, 377, 378, 380, 382, 384

[279] P. Lederer, A. Linke, C. Merdon, and J. Schöberl. Divergence-free reconstruction operators for pressure-robust Stokes discretizations with continuous pressure finite elements. *SIAM J. Numer. Anal.*, 55(5):1291–1314, 2017. pages 345, 358

[280] C. Lehrenfeld. *Hybrid Discontinuous Galerkin methods for solving incompressible flow problems.* PhD thesis, Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen, Germany, 2010. pages 183

[281] C. Lehrenfeld and J. Schöberl. High order exactly divergence-free hybrid discontinuous Galerkin methods for unsteady incompressible flows. *Comput. Methods Appl. Mech. Engrg.*, 307:339–361, 2016. pages 183, 373

[282] S. Levy. Structural analysis and influence coefficients for delta wings. *J. Aeronaut. Sci.*, 20, 1953. pages 212

[283] A. Linke. On the role of the Helmholtz decomposition in mixed methods for incompressible flows and a new variational crime. *Comput. Methods Appl. Mech. Engrg.*, 268:782–800, 2014. pages 345, 374

[284] A. Linke, C. Merdon, and W. Wollner. Optimal L^2 velocity error estimate for a modified pressure-robust Crouzeix-Raviart Stokes element. *IMA J. Numer. Anal.*, 37(1):354–374, 2017. pages 304

[285] J.-L. Lions. Problèmes aux limites non homogènes à données irrégulières: Une méthode d'approximation. In *Numerical Analysis of Partial Differential Equations (C.I.M.E. 2 Ciclo, Ispra, Italy, 1967)*, pages 283–292. Edizioni Cremonese, Rome, Italy, 1968. pages 107

[286] J.-L. Lions and E. Magenes. *Non-homogeneous Boundary Value Problems and Applications. Vols. I, II.* Springer-Verlag, New York-Heidelberg, 1972. Translated from the French by P. Kenneth, Die Grundlehren der mathematischen Wissenschaften, Band 181-182. pages 88, 91

[287] X. Liu. A framework of verified eigenvalue bounds for self-adjoint differential operators. *Appl. Math. Comput.*, 267:341–355, 2015. pages 285

[288] J. M.-S. Lubuma and S. Nicaise. Dirichlet problems in polyhedral domains. I. Regularity of the solutions. *Math. Nachr.*, 168:243–261, 1994. pages 90

[289] J. M.-S. Lubuma and S. Nicaise. Dirichlet problems in polyhedral domains. II. Approximation by FEM and BEM. *J. Comput. Appl. Math.*, 61(1):13–27, 1995. pages 90

[290] R. Luce and B. I. Wohlmuth. A local a posteriori error estimator based on equilibrated fluxes. *SIAM J. Numer. Anal.*, 42(4):1394–1414, 2004. pages 333

[291] N. Lüthen, M. Juntunen, and R. Stenberg. An improved a priori error analysis of Nitsche's method for Robin boundary conditions. *Numer. Math.*, 138(4):1011–1026, 2018. pages 206

[292] G. I. Marchuk and Y. A. Kuznetsov. On optimal iteration processes. *Soviet. Math. Dokl.*, 9: 1041–1045, 1968. pages 54

[293] G. I. Marchuk and Y. A. Kuznetsov. Méthodes itératives et fonctionnelles quadratiques. In J.-L. Lions and G. I. Marchuk, editors, *Sur les méthodes numériques en sciences physiques et économiques*, pages 3–131, Paris, France, 1974. Dunod. pages 54

[294] K.-A. Mardal, J. Schöberl, and R. Winther. A uniformly stable Fortin operator for the Taylor-Hood element. *Numer. Math.*, 123(3):537–551, 2013. pages 304, 358

[295] L. D. Marini. An inexpensive method for the evaluation of the solution of the lowest order Raviart-Thomas mixed method. *SIAM J. Numer. Anal.*, 22(3):493–496, 1985. pages 156, 330

[296] V. G. Maz'ja and B. A. Plamenevskiĭ. Weighted spaces with inhomogeneous norms, and boundary value problems in domains with conical points. In *Elliptische Differentialgleichungen (Meeting, Rostock, Germany, 1977)*, pages 161–190. Wilhelm-Pieck-Univ., Rostock, Germany, 1978. pages 90

[297] A. L. Mazzucato and V. Nistor. Well-posedness and regularity for the elasticity equation with mixed boundary conditions on polyhedral domains and domains with cracks. *Arch. Ration. Mech. Anal.*, 195(1):25–73, 2010. pages 216

[298] W. McLean. *Strongly elliptic systems and boundary integral equations.* Cambridge University Press, Cambridge, UK, 2000. pages 215

[299] J. M. Melenk. *On generalized finite-element methods.* ProQuest LLC, Ann Arbor, MI, 1995. Ph.D. thesis, University of Maryland, MD. pages 133

[300] J. M. Melenk and S. Sauter. Wavenumber explicit convergence analysis for Galerkin discretizations of the Helmholtz equation. *SIAM J. Numer. Anal.*, 49(3):1210–1243, 2011. pages 142

[301] B. Mercier, J. E. Osborn, J. Rappaz, and P.-A. Raviart. Eigenvalue approximation by mixed and hybrid methods. *Math. Comp.*, 36(154):427–453, 1981. pages 285

[302] P. Monk. A finite element method for approximating the time-harmonic Maxwell equations. *Numer. Math.*, 63(2):243–261, 1992. pages 241

[303] P. Monk. *Finite element methods for Maxwell's equations*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, NY, 2003. pages 225, 241

[304] P. Monk and L. Demkowicz. Discrete compactness and the approximation of Maxwell's equations in $\mathbb{R}^3$. *Math. Comp.*, 70(234):507–523, 2001. pages 237

[305] P. Morin, R. H. Nochetto, and K. G. Siebert. Data oscillation and convergence of adaptive FEM. *SIAM J. Numer. Anal.*, 38(2):466–488, 2000. pages 126, 127

[306] P. Morin, R. H. Nochetto, and K. G. Siebert. Local problems on stars: a posteriori error estimators, convergence, and performance. *Math. Comp.*, 72(243):1067–1097, 2003. pages 122, 335

[307] M. E. Morley. A family of mixed finite elements for linear elasticity. *Numer. Math.*, 55(6): 633–666, 1989. pages 219

[308] I. Muga and K. G. van der Zee. Discretization of linear problems in Banach spaces: residual minimization, nonlinear Petrov–Galerkin, and monotone mixed methods. arXiv:1511.04400v3 [Math.NA], 2018. pages 25, 313

[309] M. F. Murphy, G. H. Golub, and A. J. Wathen. A note on preconditioning for indefinite linear systems. *SIAM J. Sci. Comput.*, 21(6):1969–1972, 2000. pages 311

[310] J. Nečas. Sur une méthode pour résoudre les équations aux dérivées partielles de type elliptique, voisine de la variationnelle. *Ann. Scuola Norm. Sup. Pisa*, 16:305–326, 1962. pages 14

[311] S. Nicaise. Regularity of the solutions of elliptic systems in polyhedral domains. *Bull. Belg. Math. Soc. Simon Stevin*, 4(3):411–429, 1997. pages 90

[312] L. Nirenberg. Remarks on strongly elliptic partial differential equations. *Comm. Pure Appl. Math.*, 8:649–675, 1955. pages 89

[313] J. Nitsche. Ein Kriterium für die Quasi-Optimalität des Ritzschen Verfahrens. *Numer. Math.*, 11:346–348, 1968. pages 98

[314] J. Nitsche. Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind. *Abh. Math. Sem. Univ. Hamburg*, 36:9–15, 1971. pages 107, 160

[315] R. H. Nochetto and A. Veeser. Primer of adaptive finite element methods. In *Multiscale and adaptivity: modeling, numerics and applications*, volume 2040 of *Lecture Notes in Math.*, pages 125–225. Springer, Heidelberg, 2012. pages 120, 123, 126

[316] R. H. Nochetto, K. G. Siebert, and A. Veeser. Theory of adaptive finite element methods: an introduction. In *Multiscale, nonlinear and adaptive approximation*, pages 409–542. Springer, Berlin, 2009. pages 126

[317] J. T. Oden. *Finite elements: An introduction*, volume II : Finite element methods of *Handbook of Numerical Analysis*, chapter 1, pages 3–15. North Holland, Amsterdam, The Netherlands, 1991. P.G. Ciarlet and J.L. Lions, editors. pages 212

[318] J. T. Oden, I. Babuška, and C. E. Baumann. A discontinuous *hp* finite element method for diffusion problems. *J. Comput. Phys.*, 146(2):491–519, 1998. pages 174

[319] I. Oikawa. A hybridized discontinuous Galerkin method with reduced stabilization. *J. Sci. Comput.*, 65(1):327–340, 2015. pages 183

[320] E. T. Olsen and J. Douglas, Jr. Bounds on spectral condition numbers of matrices arising in the *p*-version of the finite element method. *Numer. Math.*, 69(3):333–352, 1995. pages 51

[321] J. E. Osborn. Spectral approximation for compact operators. *Math. Comp.*, 29:712–725, 1975. pages 278, 279

[322] N. Parés, P. Díez, and A. Huerta. Subdomain-based flux-free a posteriori error estimators. *Comput. Methods Appl. Mech. Engrg.*, 195(4-6):297–323, 2006. pages 335

[323] I. Perugia and D. Schötzau. The *hp*-local discontinuous Galerkin method for low-frequency time-harmonic Maxwell equations. *Math. Comp.*, 72(243):1179–1214, 2003. pages 174

[324] I. Perugia and V. Simoncini. Block-diagonal and indefinite symmetric preconditioners for mixed finite element formulations. *Numer. Linear Algebra Appl.*, 7(7-8):585–616, 2000. pages 311

[325] D. Peterseim. Eliminating the pollution effect in Helmholtz problems by local subscale correction. *Math. Comp.*, 86(305):1005–1036, 2017. pages 142

[326] R. J. Plemmons. *M*-matrix characterizations. I. Nonsingular *M*-matrices. *Linear Algebra and Appl.*, 18(2):175–188, 1977. pages 51

[327] W. Prager and J. L. Synge. Approximations in elasticity based on the concept of function space. *Quart. Appl. Math.*, 5:241–269, 1947. pages 335

[328] J. Qin. *On the convergence of some low order mixed finite elements for incompressible fluids.* ProQuest LLC, Ann Arbor, MI, 1994. Ph.D. thesis, The Pennsylvania State University, PA. pages 366, 370, 371

[329] R. Rannacher and R. L. Scott. Some optimal error estimates for piecewise linear finite element approximations. *Math. Comp.*, 38(158):437–445, 1982. pages 95

[330] R. Rannacher and S. Turek. Simple nonconforming quadrilateral Stokes element. *Numer. Methods Partial Differential Equations*, 8(2):97–111, 1992. pages 157, 373

[331] P.-A. Raviart and J.-M. Thomas. *Introduction à l'analyse numérique des équations aux dérivées partielles.* Masson, Paris, France, 1983. pages 265

[332] M. Reed and B. Simon. *Methods of modern mathematical physics. IV. Analysis of operators.* Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, 1978. pages 80

[333] E. Reissner. On a variational theorem in elasticity. *J. Math. Physics*, 29:90–95, 1950. pages 216

[334] S. I. Repin. Computable majorants of constants in the Poincaré and Friedrichs inequalities. *J. Math. Sci.*, 186(2):307–321, 2012. pages 119

[335] B. Rivière, M. F. Wheeler, and V. Girault. Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems. I. *Comput. Geosci.*, 8:337–360, 1999. pages 174

[336] B. Rivière, M. F. Wheeler, and V. Girault. A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems. *SIAM J. Numer. Anal.*, 39(3):902–931, 2001. pages 174

[337] W. Rudin. *Real and complex analysis.* McGraw-Hill Book Co., New York, NY, third edition, 1987. pages 377, 378, 380, 381, 384

[338] T. Rusten and R. Winther. A preconditioned iterative method for saddlepoint problems. *SIAM J. Matrix Anal. Appl.*, 13(3):887–904, 1992. pages 307

[339] Y. Saad. *Iterative methods for sparse linear systems.* PWS Publishing Company, Boston, MA, 1996. pages 53, 54, 61, 309, 311

[340] Y. Saad and M. H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 7:856–869, 1986. pages 54

[341] G. Savaré. Regularity results for elliptic equations in Lipschitz domains. *J. Func. Anal.*, 152: 176–201, 1998. pages 91

[342] F.-J. Sayas. Aubin-Nitsche estimates are equivalent to compact embeddings. *BIT*, 44(2): 287–290, 2004. pages 97, 98

[343] A. H. Schatz. An observation concerning Ritz-Galerkin methods with indefinite bilinear forms. *Math. Comp.*, 28:959–962, 1974. pages 140

[344] A. H. Schatz and L. B. Wahlbin. On the quasi-optimality in L_∞ of the $\dot{H}^1$-projection into finite element spaces. *Math. Comp.*, 38(157):1–22, 1982. pages 99

[345] L. R. Scott and M. Vogelius. Norm estimates for a maximal right inverse of the divergence operator in spaces of piecewise polynomials. *RAIRO Modél. Math. Anal. Numér.*, 19(1): 111–143, 1985. pages 217, 370

[346] I. Šebestová and T. Vejchodský. Two-sided bounds for eigenvalues of differential operators with applications to Friedrichs, Poincaré, trace, and similar constants. *SIAM J. Numer. Anal.*, 52(1):308–329, 2014. pages 119

[347] D. J. Silvester and A. J. Wathen. Fast iterative solution of stabilised Stokes systems. II. Using general block preconditioners. *SIAM J. Numer. Anal.*, 31(5):1352–1367, 1994. pages 311

[348] I. Smears and E. Süli. Discontinuous Galerkin finite element approximation of Hamilton-Jacobi-Bellman equations with Cordes coefficients. *SIAM J. Numer. Anal.*, 52(2):993–1016, 2014. pages 80

[349] V. A. Solonnikov. L_p-estimates for solutions of the heat equation in a dihedral angle. *Rend. Mat. Appl. (7)*, 21(1-4):1–15, 2001. pages 341

[350] P. Sonneveld. CGS, a fast Lanczos-type solver for nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 10(1):36–52, 1989. pages 54

[351] S.-C. Soon, B. Cockburn, and H. K. Stolarski. A hybridizable discontinuous Galerkin method for linear elasticity. *Internat. J. Numer. Methods Engrg.*, 80(8):1058–1092, 2009. pages 218

[352] R. Stenberg. Analysis of mixed finite elements methods for the Stokes problem: a unified approach. *Math. Comp.*, 42(165):9–23, 1984. pages 358, 366

[353] R. Stenberg. On the construction of optimal mixed finite element methods for the linear elasticity problem. *Numer. Math.*, 48(4):447–462, 1986. pages 219, 366

[354] R. Stenberg. On some three-dimensional finite elements for incompressible media. *Comput. Methods Appl. Mech. Engrg.*, 63(3):261–269, 1987. pages 358, 366

[355] R. Stenberg. A family of mixed finite elements for the elasticity problem. *Numer. Math.*, 53 (5):513–538, 1988. pages 219

[356] R. Stevenson. Optimality of a standard adaptive finite element method. *Found. Comput. Math.*, 7(2):245–269, 2007. pages 126

[357] R. Stevenson. The completion of locally refined simplicial partitions created by bisection. *Math. Comp.*, 77(261):227–241, 2008. pages 126, 127, 128

[358] G. Strang. Variational crimes in the finite element method. In A. Aziz, editor, *The mathematical foundations of the finite element method with applications to partial differential equations*, New York, NY, 1972. Academic Press. pages 39

[359] G. Strang and G. J. Fix. *An analysis of the finite element method*. Prentice-Hall Inc., Englewood Cliffs, NJ, 1973. Prentice-Hall Series in Automatic Computation. pages 51, 279

[360] A. H. Stroud. *Approximate calculation of multiple integrals*. Prentice-Hall Inc., Englewood Cliffs, NJ, 1971. pages 70

[361] F. Tantardini and A. Veeser. The L^2-projection and quasi-optimality of Galerkin methods for parabolic equations. *SIAM J. Numer. Anal.*, 54(1):317–340, 2016. pages 30, 31

[362] L. Tartar. *An introduction to Sobolev spaces and interpolation spaces*, volume 3 of *Lecture Notes of the Unione Matematica Italiana*. Springer, Berlin, Germany; UMI, Bologna, Italy, 2007. pages 88

[363] R. Temam. *Navier–Stokes equations*, volume 2 of *Studies in Mathematics and its Applications*. North-Holland, Amsterdam, The Netherlands, 1977. pages 150

[364] A. Ten Eyck and A. Lew. Discontinuous Galerkin methods for non-linear elasticity. *Internat. J. Numer. Methods Engrg.*, 67(9):1204–1243, 2006. pages 176

[365] U. Trottenberg, C. W. Oosterlee, and A. Schüller. *Multigrid*. Academic Press, Inc., San Diego, CA, 2001. With contributions by A. Brandt, P. Oswald and K. Stüben. pages 54

[366] S. Turek. *Efficient solvers for incompressible flow problems. An algorithmic and computational approach*, volume 6 of *Lecture Notes in Computational Science and Engineering*. Springer-Verlag, Berlin, Germany, 1999. pages 310, 373

[367] M. J. Turner, R. W. Clough, H. C. Martin, and L. J. Topp. Stiffness and deflection analysis of complex structures. *J. Aero. Sci.*, 23:805–823, 1956. pages 212

[368] G. M. Vainikko. Asymptotic error bounds for projection methods in the eigenvalue problem. *Ž. Vyčisl. Mat. i Mat. Fiz.*, 4:405–425, 1964. pages 279

[369] G. M. Vainikko. Rapidity of convergence of approximation methods in eigenvalue problems. *Ž. Vyčisl. Mat. i Mat. Fiz.*, 7:977–987, 1967. pages 279

[370] H. A. van der Vorst. Bi-CGStab: a more stably converging variant of CG-S for the solution of nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 13:631–644, 1992. pages 54

[371] A. Veeser. Approximating gradients with continuous piecewise polynomial functions. *Found. Comput. Math.*, 16(3):723–750, 2016. pages 97

[372] A. Veeser and R. Verfürth. Poincaré constants for finite element stars. *IMA J. Numer. Anal.*, 32(1):30–47, 2012. pages 119, 120, 122

[373] A. Veeser and P. Zanotti. Quasi-optimal nonconforming methods for symmetric elliptic problems. I—Abstract theory. *SIAM J. Numer. Anal.*, 56(3):1621–1642, 2018. pages 36, 149

[374] T. Vejchodský and P. Šolín. Discrete maximum principle for Poisson equation with mixed boundary conditions solved by hp-FEM. *Adv. Appl. Math. Mech.*, 1(2):201–214, 2009. pages 109

[375] R. Verfürth. A combined conjugate gradient-multigrid algorithm for the numerical solution of the Stokes problem. *IMA J. Numer. Anal.*, 4(4):441–455, 1984. pages 307

[376] R. Verfürth. Error estimates for a mixed finite element approximation of the Stokes equation. *RAIRO, Anal. Num.*, 18:175–182, 1984. pages 353

[377] R. Verfürth. Robust a posteriori error estimates for nonstationary convection-diffusion equations. *SIAM J. Numer. Anal.*, 43(4):1783–1802, 2005. pages 118

[378] R. Verfürth. *A posteriori error estimation techniques for finite element methods*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, UK, 2013. pages 118, 124, 126

[379] V. V. Voevodin. The question of non-self-adjoint extension of the conjugate gradients method is closed. *USSR Comput. Maths. Math. Phys.*, 23(2):143–144, 1983. pages 53

[380] M. Vogelius. An analysis of the p-version of the finite element method for nearly incompressible materials. Uniformly valid, optimal error estimates. *Numer. Math.*, 41(1):39–53, 1983. pages 217

[381] M. Vohralík. On the discrete Poincaré-Friedrichs inequalities for nonconforming approximations of the Sobolev space H^1. *Numer. Funct. Anal. Optim.*, 26(7-8):925–952, 2005. pages 119

[382] M. Vohralík. A posteriori error estimates for lowest-order mixed finite element discretizations of convection-diffusion-reaction equations. *SIAM J. Numer. Anal.*, 45(4):1570–1599, 2007. pages 330

[383] M. Vohralík. Unified primal formulation-based a priori and a posteriori error analysis of mixed finite element methods. *Math. Comp.*, 79(272):2001–2032, 2010. pages 331, 336

[384] M. Vohralík. Guaranteed and fully robust a posteriori error estimates for conforming discretizations of diffusion problems with discontinuous coefficients. *J. Sci. Comput.*, 46(3): 397–438, 2011. pages 333

[385] M. Vohralík and B. I. Wohlmuth. Mixed finite element methods: implementation with one unknown per element, local flux expressions, positivity, polygonal meshes, and relations to other methods. *Math. Models Methods Appl. Sci.*, 23(5):803–838, 2013. pages 330

[386] J. Wang and X. Ye. A weak Galerkin mixed finite element method for second order elliptic problems. *Math. Comp.*, 83(289):2101–2126, 2014. pages 183

[387] J. Wang and X. Ye. A weak Galerkin finite element method for the Stokes equations. *Adv. Comput. Math.*, 42(1):155–174, 2016. pages 373

[388] K. Washizu. On the variational principles of elasticity and plasticity. Technical Report 25-18, MIT, Cambridge, MA, 1955. pages 216

[389] A. J. Wathen. Realistic eigenvalue bounds for the Galerkin mass matrix. *IMA J. Numer. Anal.*, 7(4):449–457, 1987. pages 49

[390] A. J. Wathen and D. J. Silvester. Fast iterative solution of stabilised Stokes systems. I. Using simple diagonal preconditioners. *SIAM J. Numer. Anal.*, 30(3):630–649, 1993. pages 307

[391] C. Weber. A local compactness theorem for Maxwell's equations. *Math. Methods Appl. Sci.*, 2(1):12–25, 1980. pages 231

[392] W. L. Wendland. Strongly elliptic boundary integral equations. In *The state of the art in numerical analysis (Birmingham, UK, 1986)*, volume 9 of *The Institute of Mathematics and its Applications Conference Series. New Series*, pages 511–562. Oxford Univ. Press, New York, NY, 1987. pages 32

[393] P. Wesseling. *An introduction to multigrid methods.* Pure and Applied Mathematics. John Wiley & Sons, Ltd., Chichester, UK, 1992. pages 54

[394] M. F. Wheeler. An elliptic collocation-finite element method with interior penalties. *SIAM J. Numer. Anal.*, 15:152–161, 1978. pages 169

[395] T. P. Wihler. Locking-free DGFEM for elasticity problems in polygons. *IMA J. Numer. Anal.*, 24(1):45–75, 2004. pages 218

[396] J. Xu and L. Zikatanov. A monotone finite element scheme for convection-diffusion equations. *Math. Comp.*, 68(228):1429–1446, 1999. pages 110

[397] J. Xu and L. Zikatanov. Some observations on Babuška and Brezzi theories. *Numer. Math.*, 94(1):195–202, 2003. pages 30

[398] K. Yosida. *Functional analysis.* Classics in Mathematics. Springer-Verlag, Berlin, Germany, 1995. Reprint of the sixth (1980) edition. pages 377, 378, 380, 382, 383, 384

[399] A. Younes, P. Ackerer, and G. Chavent. From mixed finite elements to finite volumes for elliptic PDEs in two and three dimensions. *Internat. J. Numer. Methods Engrg.*, 59(3):365–388, 2004. pages 330

[400] E. Zeidler. *Applied functional analysis*, volume 108 of *Applied Mathematical Sciences*. Springer-Verlag, New York, NY, 1995. pages 377

[401] S. Zhang. A new family of stable mixed finite elements for the 3D Stokes equations. *Math. Comp.*, 74(250):543–554, 2005. pages 371

[402] S. Zhang. On the P1 Powell-Sabin divergence-free finite element for the Stokes equations. *J. Comput. Math.*, 26(3):456–470, 2008. pages 371

[403] S. Zhang. Quadratic divergence-free finite elements on Powell-Sabin tetrahedral grids. *Calcolo*, 48(3):211–244, 2011. pages 369, 371

[404] S. Zhang and S. Zhang. C_0P_2-P_0 Stokes finite element pair on sub-hexahedron tetrahedral grids. *Calcolo*, 54(4):1403–1417, 2017. pages 367

[405] L. Zhong, S. Shu, G. Wittum, and J. Xu. Optimal error estimates for Nedelec edge elements for time-harmonic Maxwell's equations. *J. Comput. Math.*, 27(5):563–572, 2009. pages 241

Index