3

Coercive Problems

This chapter deals with problems whose weak formulation is endowed with a coercivity property. The key examples investigated henceforth are scalar elliptic PDEs, spectral problems associated with the Laplacian, and PDE systems derived from continuum mechanics. The goal is twofold: First, to set up a mathematical framework for well-posedness; then, to investigate conforming and non-conforming finite element approximations based on Galerkin methods. Error estimates are derived from the theoretical results of Chapters 1 and 2 and are illustrated numerically. The last section of this chapter is concerned with coercivity loss and is meant to be a transition to Chapters 4 and 5.

3.1 Scalar Elliptic PDEs: Theory

Let Ω be a domain in $\mathbb{R}^d$. Consider a differential operator $\mathcal{L}$ in the form

$$\mathcal{L}u = -\nabla\cdot(\sigma\cdot\nabla u) + \beta\cdot\nabla u + \mu u, \tag{3.1}$$

where σ, β, and μ are functions defined over Ω and taking their values in $\mathbb{R}^{d,d}$, $\mathbb{R}^d$, and $\mathbb{R}$, respectively. Given a function $f : \Omega \to \mathbb{R}$, consider the problem of finding a function $u : \Omega \to \mathbb{R}$ such that

$$\begin{cases} \mathcal{L}u = f & \text{in } \Omega, \\ \mathcal{B}u = g & \text{on } \partial\Omega, \end{cases} \tag{3.2}$$

where the operator $\mathcal{B}$ accounts for boundary conditions. The model problem (3.2) arises in several applications:

(i) *Heat transfer*: u is the temperature, $\sigma = \kappa\mathcal{I}$ where κ is the thermal conductivity, β is the flow field, $\mu = 0$, and f is the externally supplied heat per unit volume.

(ii) *Advection–diffusion*: u is the concentration of a solute transported in a flow field β. The matrix σ models the solute diffusivity resulting from either molecular diffusion or turbulent mixing by the carrier flow. Solute production or destruction by chemical reaction is accounted for by the linear term μu, and the right-hand side f models fixed sources or sinks.

Henceforth, the following assumptions are made on the data: $f \in L^2(\Omega)$, $\sigma \in [L^\infty(\Omega)]^{d,d}$, $\beta \in [L^\infty(\Omega)]^d$, $\nabla\cdot\beta \in L^\infty(\Omega)$, and $\mu \in L^\infty(\Omega)$. Furthermore, the operator $\mathcal{L}$ is assumed to be *elliptic* in the following sense:

Definition 3.1. *The operator $\mathcal{L}$ defined in* (3.1) *is said to be* elliptic *if there exists $\sigma_0 > 0$ such that*

$$\forall \xi \in \mathbb{R}^d, \qquad \sum_{i,j=1}^{d} \sigma_{ij}\xi_i\xi_j \geq \sigma_0 \|\xi\|_d^2 \quad \text{a.e. in } \Omega. \tag{3.3}$$

Equation (3.2) *is then called an* elliptic PDE*.*

Example 3.2. A fundamental example of an elliptic operator is the *Laplacian*, $\mathcal{L} = -\Delta$, which is obtained for $\sigma = \mathcal{I}$, $\beta = 0$, and $\mu = 0$. □

3.1.1 Review of boundary conditions and their weak formulation

We first proceed formally and then specify the mathematical framework for the weak formulation.

Homogeneous Dirichlet boundary condition. We want to enforce $u = 0$ on $\partial\Omega$. Multiplying the PDE $\mathcal{L}u = f$ by a (sufficiently smooth) test function v vanishing at the boundary, integrating over Ω, and using the Green formula

$$\int_\Omega -\nabla\cdot(\sigma\cdot\nabla u)\, v = \int_\Omega \nabla v\cdot\sigma\cdot\nabla u - \int_{\partial\Omega} v\,(n\cdot\sigma\cdot\nabla u), \tag{3.4}$$

yields

$$\int_\Omega \nabla v\cdot\sigma\cdot\nabla u + v(\beta\cdot\nabla u) + \mu uv = \int_\Omega fv.$$

A possible regularity requirement on u and v for the integrals over Ω to be meaningful is

$$u \in H^1(\Omega) \qquad \text{and} \qquad v \in H^1(\Omega).$$

Since $u \in H^1(\Omega)$, Theorem B.52 implies that u has a trace at the boundary. Because of the boundary condition $u_{|\partial\Omega} = 0$, the solution is sought in $H^1_0(\Omega)$. Test functions are also taken in $H^1_0(\Omega)$, leading to the following weak formulation:

$$\begin{cases} \text{Seek } u \in H^1_0(\Omega) \text{ such that} \\ a_{\sigma,\beta,\mu}(u,v) = \int_\Omega fv, \quad \forall v \in H^1_0(\Omega), \end{cases} \tag{3.5}$$

with the bilinear form

$$a_{\sigma,\beta,\mu}(u,v) = \int_\Omega \nabla v\cdot\sigma\cdot\nabla u + v(\beta\cdot\nabla u) + \mu uv. \tag{3.6}$$

Proposition 3.3. *If u solves* (3.5)*, then $\mathcal{L}u = f$ a.e. in Ω and $u = 0$ a.e. on $\partial\Omega$.*

Proof. Let $\varphi \in \mathcal{D}(\Omega)$ and let u be a solution to (3.5). Hence,

$$\begin{aligned} \langle -\nabla\cdot(\sigma\cdot\nabla u), \varphi\rangle_{\mathcal{D}',\mathcal{D}} = \langle \sigma\cdot\nabla u, \nabla\varphi\rangle_{\mathcal{D}',\mathcal{D}} &= \int_\Omega \nabla\varphi\cdot\sigma\cdot\nabla u \\ &= \int_\Omega (f - \beta\cdot\nabla u - \mu u)\,\varphi, \end{aligned}$$

yielding $\langle \mathcal{L}u, \varphi\rangle_{\mathcal{D}',\mathcal{D}} = \int_\Omega f\varphi$. Owing to the density of $\mathcal{D}(\Omega)$ in $L^2(\Omega)$, $\mathcal{L}u = f$ in $L^2(\Omega)$. Therefore, $\mathcal{L}u = f$ a.e. in Ω. Moreover, $u = 0$ a.e. on $\partial\Omega$ by definition of $H^1_0(\Omega)$; see Theorem B.52. ☐

Non-homogeneous Dirichlet boundary condition. We want to enforce $u = g$ on $\partial\Omega$, where $g : \partial\Omega \to \mathbb{R}$ is a given function. We assume that g is sufficiently smooth so that there exists a lifting u_g of g in $H^1(\Omega)$, i.e., a function $u_g \in H^1(\Omega)$ such that $u_g = g$ on $\partial\Omega$; see §2.1.4. We obtain the weak formulation:

$$\begin{cases} \text{Seek } u \in H^1(\Omega) \text{ such that} \\ u = u_g + \phi, \quad \phi \in H^1_0(\Omega), \\ a_{\sigma,\beta,\mu}(\phi, v) = \int_\Omega fv - a_{\sigma,\beta,\mu}(u_g, v), \quad \forall v \in H^1_0(\Omega). \end{cases} \tag{3.7}$$

Proposition 3.4. *Let $g \in H^{\frac{1}{2}}(\partial\Omega)$. If u solves* (3.7)*, then $\mathcal{L}u = f$ a.e. in Ω and $u = g$ a.e. on $\partial\Omega$.*

Proof. Similar to that of Proposition 3.3. ☐

When the operator $\mathcal{L}$ is the Laplacian, (3.7) is called a *Poisson problem*.

Neumann boundary condition. Given a function $g : \partial\Omega \to \mathbb{R}$, we want to enforce $n\cdot\sigma\cdot\nabla u = g$ on $\partial\Omega$. Note that in the case $\sigma = \mathcal{I}$, the Neumann condition specifies the normal derivative of u since $n\cdot\nabla u = \partial_n u$. Proceeding as before and using the Neumann condition in the surface integral in (3.4) yields the weak formulation:

$$\begin{cases} \text{Seek } u \in H^1(\Omega) \text{ such that} \\ a_{\sigma,\beta,\mu}(u, v) = \int_\Omega fv + \int_{\partial\Omega} gv, \quad \forall v \in H^1(\Omega). \end{cases} \tag{3.8}$$

Proposition 3.5. *Let $g \in L^2(\partial\Omega)$. If u solves* (3.8)*, then $\mathcal{L}u = f$ a.e. in Ω and $n\cdot\sigma\cdot\nabla u = g$ a.e. on $\partial\Omega$.*

Proof. Taking test functions in $\mathcal{D}(\Omega)$ readily implies $\mathcal{L}u = f$ a.e. in Ω. Therefore, $-\nabla\cdot(\sigma\cdot\nabla u) \in L^2(\Omega)$. Corollary B.59 implies $n\cdot\sigma\cdot\nabla u \in H^{\frac{1}{2}}(\partial\Omega)' = H^{-\frac{1}{2}}(\partial\Omega)$ since

$$\forall \phi \in H^{\frac{1}{2}}(\partial\Omega), \quad \langle n\cdot\sigma\cdot\nabla u, \phi\rangle_{H^{-\frac{1}{2}},H^{\frac{1}{2}}} = \int_\Omega -\nabla\cdot(\sigma\cdot\nabla u)u_\phi + \int_\Omega \nabla u_\phi\cdot\sigma\cdot\nabla u,$$

where $u_\phi \in H^1(\Omega)$ is a lifting of ϕ in $H^1(\Omega)$. Then, (3.8) yields

$$\forall \phi \in H^{\frac{1}{2}}(\partial\Omega), \quad \langle n\cdot\sigma\cdot\nabla u, \phi\rangle_{H^{-\frac{1}{2}},H^{\frac{1}{2}}} = \int_{\partial\Omega} g\phi,$$

showing that $n\cdot\sigma\cdot\nabla u = g$ in $H^{-\frac{1}{2}}(\partial\Omega)$ and, therefore, in $L^2(\partial\Omega)$ since g belongs to this space. □

Mixed Dirichlet–Neumann boundary conditions. Consider a partition of the boundary in the form $\partial\Omega = \partial\Omega_{\mathrm{D}} \cup \partial\Omega_{\mathrm{N}}$. Impose a Dirichlet condition on $\partial\Omega_{\mathrm{D}}$ and a Neumann condition on $\partial\Omega_{\mathrm{N}}$. If the Dirichlet condition is non-homogeneous, assume that $\partial\Omega_{\mathrm{D}}$ is smooth enough so that, for all $g \in H^{\frac{1}{2}}(\partial\Omega_{\mathrm{D}})$, there exists an extension $\widetilde{g} \in H^{\frac{1}{2}}(\partial\Omega)$ such that $\widetilde{g}_{|\partial\Omega_{\mathrm{D}}} = g$ and $\|\widetilde{g}\|_{H^{\frac{1}{2}}(\partial\Omega)} \leq c\,\|g\|_{H^{\frac{1}{2}}(\partial\Omega_{\mathrm{D}})}$ uniformly in g. Then, using the lifting of $\widetilde{g}$ in $H^1(\Omega)$, one can assume that the Dirichlet condition is homogeneous. The boundary conditions are thus

$$\begin{cases} u = 0 & \text{on } \partial\Omega_{\mathrm{D}}, \\ n\cdot\sigma\cdot\nabla u = g & \text{on } \partial\Omega_{\mathrm{N}}, \end{cases}$$

with a given function $g : \partial\Omega_{\mathrm{N}} \to \mathbb{R}$.

Proceeding as before, we split the boundary integral in (3.4) into its contributions over $\partial\Omega_{\mathrm{D}}$ and $\partial\Omega_{\mathrm{N}}$. Taking the solution and the test function in the functional space

$$H^1_{\partial\Omega_{\mathrm{D}}}(\Omega) = \{u \in H^1(\Omega);\, u = 0 \text{ on } \partial\Omega_{\mathrm{D}}\},$$

the surface integral over $\partial\Omega_{\mathrm{D}}$ vanishes. Furthermore, using the Neumann condition in the surface integral over $\partial\Omega_{\mathrm{N}}$ yields the weak formulation:

$$\begin{cases} \text{Seek } u \in H^1_{\partial\Omega_{\mathrm{D}}}(\Omega) \text{ such that} \\ a_{\sigma,\beta,\mu}(u,v) = \int_\Omega fv + \int_{\partial\Omega_{\mathrm{N}}} gv, \quad \forall v \in H^1_{\partial\Omega_{\mathrm{D}}}(\Omega). \end{cases} \tag{3.9}$$

Proposition 3.6. *Let $\partial\Omega_{\mathrm{D}} \subset \partial\Omega$, assume* $\mathrm{meas}(\partial\Omega_{\mathrm{D}}) > 0$*, and set $\partial\Omega_{\mathrm{N}} = \partial\Omega\backslash\partial\Omega_{\mathrm{D}}$. Let $g \in L^2(\partial\Omega_{\mathrm{N}})$. If u solves* (3.9)*, then $\mathcal{L}u = f$ a.e. in Ω, $u = 0$ a.e. on $\partial\Omega_{\mathrm{D}}$, and $(n\cdot\sigma\cdot\nabla u) = g$ a.e. on $\partial\Omega_{\mathrm{N}}$.*

Proof. Proceed as in the previous proofs. □

Robin boundary condition. Given two functions $g, \gamma : \partial\Omega \to \mathbb{R}$, we want to enforce $\gamma u + n\cdot\sigma\cdot\nabla u = g$ on $\partial\Omega$. Using this condition in the surface integral in (3.4) yields the weak formulation:

$$\begin{cases} \text{Seek } u \in H^1(\Omega) \text{ such that} \\ a_{\sigma,\beta,\mu}(u,v) + \int_{\partial\Omega} \gamma uv = \int_\Omega fv + \int_{\partial\Omega} gv, \quad \forall v \in H^1(\Omega). \end{cases} \tag{3.10}$$

Problem	V	$a(u,v)$	$f(v)$
Homogeneous Dirichlet	$H_0^1(\Omega)$	$a_{\sigma,\beta,\mu}(u,v)$	$\int_\Omega fv$
Neumann	$H^1(\Omega)$	$a_{\sigma,\beta,\mu}(u,v)$	$\int_\Omega fv + \int_{\partial\Omega} gv$
Dirichlet–Neumann	$H^1_{\partial\Omega_{\mathrm{D}}}(\Omega)$	$a_{\sigma,\beta,\mu}(u,v)$	$\int_\Omega fv + \int_{\partial\Omega_{\mathrm{N}}} gv$
Robin	$H^1(\Omega)$	$a_{\sigma,\beta,\mu}(u,v) + \int_{\partial\Omega} \gamma uv$	$\int_\Omega fv + \int_{\partial\Omega} gv$

Table 3.1. Weak formulation corresponding to the various boundary conditions for the second-order PDE (3.2). The bilinear form $a_{\sigma,\beta,\mu}(u,v)$ is defined in (3.6).

Proposition 3.7. *Let $g \in L^2(\partial\Omega)$ and let $\gamma \in L^\infty(\partial\Omega)$. If u solves* (3.10), *then $\mathcal{L}u = f$ a.e. in Ω and $\gamma u + n\cdot\sigma\cdot\nabla u = g$ a.e. on $\partial\Omega$.*

Proof. Proceed as in the previous proofs. □

Summary. Except for the non-homogeneous Dirichlet problem, all the problems considered herein take the generic form:

$$\begin{cases} \text{Seek } u \in V \text{ such that} \\ a(u,v) = f(v), \quad \forall v \in V, \end{cases} \tag{3.11}$$

where V is a Hilbert space satisfying

$$H_0^1(\Omega) \subset V \subset H^1(\Omega).$$

Moreover, a is a bilinear form defined on $V \times V$, and f is a linear form defined on V; see Table 3.1. For the non-homogeneous Dirichlet problem, $u \in H^1(\Omega)$, $u = u_g + \phi$ where u_g is a lifting of the boundary data and ϕ solves a problem of the form (3.11).

Essential and natural boundary conditions. It is important to observe the different treatment between Dirichlet conditions and Neumann or Robin conditions. The former are imposed explicitly in the functional space where the solution is sought, and the test functions vanish on the corresponding part of the boundary. For this reason, Dirichlet conditions are often termed *essential boundary conditions*. Neumann and Robin conditions are not imposed by the functional setting but by the weak formulation itself. The fact that test functions have degrees of freedom on the corresponding part of the boundary is sufficient to enforce the boundary conditions in question. For this reason, these conditions are often termed *natural boundary conditions*. Note that it is also possible to treat Dirichlet conditions as natural boundary conditions by using a penalty method; see §8.4.3.

3.1.2 Coercivity

Theorem 3.8. *Let $f \in L^2(\Omega)$, let $\sigma \in [L^\infty(\Omega)]^{d,d}$ be such that* (3.3) *holds, let $\beta \in [L^\infty(\Omega)]^d$ with $\nabla\cdot\beta \in L^\infty(\Omega)$, and let $\mu \in L^\infty(\Omega)$. Set*

$p = \inf \mathrm{ess}_{x\in\Omega}\left(\mu - \frac{1}{2}\nabla\cdot\beta\right)$ and let c_Ω be the constant in the Poincaré inequality (B.23)*.*

(i) *Both the homogeneous Dirichlet problem* (3.5) *and the non-homogeneous Dirichlet problem* (3.7) *are well-posed if*

$$\sigma_0 + \min\Big(0, \tfrac{p}{c_\Omega}\Big) > 0. \tag{3.12}$$

(ii) *The Neumann problem* (3.8) *is well-posed if*

$$p > 0 \qquad \text{and} \qquad \inf_{x\in\partial\Omega} \mathrm{ess}(\beta\cdot n) \geq 0. \tag{3.13}$$

(iii) *The mixed Dirichlet–Neumann problem* (3.9) *is well-posed if* (3.12) *holds,* $\mathrm{meas}(\partial\Omega_{\mathrm{D}}) > 0$*, and* $\partial\Omega^- = \{x \in \partial\Omega;\, (\beta\cdot n)(x) < 0\} \subset \partial\Omega_{\mathrm{D}}$*.*

(iv) *Set* $q = \inf \mathrm{ess}_{x\in\partial\Omega}(\gamma + \frac{1}{2}\beta\cdot n)$*. The Robin problem* (3.10) *is well-posed if*

$$p \geq 0, \quad q \geq 0, \quad \text{and} \quad pq \neq 0. \tag{3.14}$$

Proof. We prove (i) and (iv) only, leaving the remaining items as an exercise.
(1) Proof of (i). Using the ellipticity of $\mathcal{L}$ and the identity

$$\int_\Omega u(\beta\cdot\nabla u) = -\tfrac{1}{2}\int_\Omega (\nabla\cdot\beta)u^2 + \tfrac{1}{2}\int_{\partial\Omega}(\beta\cdot n)u^2,$$

which is a direct consequence of the divergence formula (B.19), yields

$$\forall u \in H^1_0(\Omega), \quad a_{\sigma,\beta,\mu}(u,u) \geq \sigma_0|u|^2_{1,\Omega} + p\|u\|^2_{0,\Omega}.$$

Setting $\delta = \min(0, \frac{p}{c_\Omega})$ and using the Poincaré inequality (B.23) yields

$$\forall u \in H^1_0(\Omega), \quad a_{\sigma,\beta,\mu}(u,u) \geq \left(\sigma_0 + \frac{\delta}{c_\Omega}\right)|u|^2_{1,\Omega} \geq \alpha\|u\|^2_{1,\Omega},$$

with $\alpha = \frac{c_\Omega(c_\Omega\sigma_0+\delta)}{1+c_\Omega^2}$, showing that the bilinear form $a_{\sigma,\beta,\mu}$ is coercive on $H^1_0(\Omega)$. The well-posedness of the homogeneous Dirichlet problem then results from the Lax–Milgram Lemma, while that of the non-homogeneous Dirichlet problem results from Proposition 2.10.
(2) Proof of (iv). Let $a(u,v) = a_{\sigma,\beta,\mu}(u,v) + \int_{\partial\Omega}\gamma uv$. A straightforward calculation shows that

$$\forall u \in H^1(\Omega), \quad a(u,u) \geq \sigma_0|u|^2_{1,\Omega} + p\|u\|^2_{0,\Omega} + q\|u\|^2_{0,\partial\Omega}.$$

If $p > 0$ and $q \geq 0$, the bilinear form a is clearly coercive on $H^1(\Omega)$ with constant $\alpha = \min(\sigma_0, p)$. If $p \geq 0$ and $q > 0$, the coercivity of a is readily deduced from Lemma B.63. In both cases, well-posedness then results from the Lax–Milgram Lemma. □

Remark 3.9.

(i) For the homogeneous and the non-homogeneous Dirichlet problem, f can be taken in $H^{-1}(\Omega) = (H_0^1(\Omega))'$. In this case, the right-hand side in (3.11) becomes $f(v) = \langle f, v\rangle_{H^{-1},H_0^1}$, and the problem is still well-posed. The stability estimate takes the form $\|u\|_{1,\Omega} \leq c\,\|f\|_{-1,\Omega}$.

(ii) Consider the Laplacian with homogeneous Dirichlet boundary conditions, i.e., given $f \in H^{-1}(\Omega)$, solve $-\Delta u = f$ in Ω with the boundary condition $u_{|\partial\Omega} = 0$. Then, the weak formulation of this problem amounts to seeking $u \in H_0^1(\Omega)$ such that $\int_\Omega \nabla u\cdot\nabla v = \langle f, v\rangle_{H^{-1},H_0^1}$ for all $v \in H_0^1(\Omega)$. Owing to Theorem 3.8(i) with $\beta = 0$, $\sigma = \mathcal{I}$, and $\mu = 0$, this problem is well-posed. This means that the operator $(-\Delta)^{-1} : H^{-1}(\Omega) \to H_0^1(\Omega)$ is an isomorphism.

(iii) Uniqueness is not a trivial property in spaces larger than $H^1(\Omega)$. For instance, one can construct domains in which this property does not hold in L^2 for the Dirichlet problem; see Exercise 3.4.

(iv) Consider problem (3.11). If the advection field β vanishes and if the diffusion matrix σ is symmetric a.e. in Ω, the bilinear form a is symmetric and positive. Therefore, owing to Proposition 2.4, (3.11) can be reformulated into a *variational form*. For the homogeneous Dirichlet problem, the variational form in question is

$$\min_{v\in H_0^1(\Omega)} \left(\tfrac{1}{2}\int_\Omega \nabla v\cdot\sigma\cdot\nabla v + \tfrac{1}{2}\int_\Omega \mu v^2 - \int_\Omega f v \right).$$

The case of other boundary conditions is left as an exercise.

(v) When μ and β vanish, the solution to the Neumann problem (3.8) is defined up to an additive constant. Therefore, we decide to seek a solution with zero-mean over Ω. Accordingly, we introduce the space

$$H^1_{\int=0}(\Omega) = \left\{ v \in H^1(\Omega);\ \int_\Omega v = 0 \right\}.$$

To ensure the existence of a solution, the data f and g must satisfy a compatibility relation. Owing to the fact that $\int_\Omega f = -\int_\Omega \nabla\cdot(\sigma\cdot\nabla u) = -\int_{\partial\Omega} n\cdot\sigma\cdot\nabla u = -\int_{\partial\Omega} g$, the compatibility condition is

$$\int_\Omega f + \int_{\partial\Omega} g = 0. \tag{3.15}$$

Thus, the weak formulation of the purely diffusive Neumann problem is:

$$\begin{cases} \text{Seek } u \in H^1_{\int=0}(\Omega) \text{ such that} \\ \int_\Omega \nabla v\cdot\sigma\cdot\nabla u = \int_\Omega f v + \int_{\partial\Omega} g v, \quad \forall v \in H^1_{\int=0}(\Omega). \end{cases} \tag{3.16}$$

Test functions have also been restricted to the functional space $H^1_{\int=0}(\Omega)$. Indeed, owing to (3.15), a constant test function leads to the trivial equation "0 = 0." Moreover, under the conditions (3.3) and (3.15), assuming that the data satisfy $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$, and using Lemma B.66, one readily verifies that problem (3.16) is well-posed with a stability estimate of the form $\forall f \in L^2(\Omega)$, $\forall g \in L^2(\partial\Omega)$, $\|u\|_{1,\Omega} \leq c\,(\|f\|_{0,\Omega} + \|g\|_{0,\partial\Omega})$. □

3.1.3 Smoothing properties

We have seen that the natural functional space V in which to seek the solution to (3.11) is such that $H_0^1(\Omega) \subset V \subset H^1(\Omega)$. For sufficiently smooth data, stronger regularity results can be derived. The interest of these results stems from the fact that in the framework of finite element methods, the regularity of the exact solution directly controls the convergence rate of the approximate solution; see §3.2.5 for numerical illustrations. In this section, it is implicitly assumed that the hypotheses of Theorem 3.8 hold so that the problems considered henceforth are well-posed. This section is set at an introductory level; see, e.g., [Gri85, Gri92, CoD02] for further insight.

Theorem 3.10 (Domain with smooth boundary). *Let $m \geq 0$, let Ω be a domain of class $\mathcal{C}^{m+2}$, and let $f \in H^m(\Omega)$. Assume that the coefficients σ_{ij} are in $\mathcal{C}^{m+1}(\overline{\Omega})$ and that the coefficients β_i and μ are in $\mathcal{C}^m(\overline{\Omega})$. Then:*

(i) *The solution to the homogeneous Dirichlet problem* (3.5) *is in $H^{m+2}(\Omega)$.*
(ii) *Assuming $g \in H^{m+\frac{3}{2}}(\partial\Omega)$, the solution to the non-homogeneous Dirichlet problem* (3.7) *is in $H^{m+2}(\Omega)$.*
(iii) *Assuming $g \in H^{m+\frac{1}{2}}(\partial\Omega)$, the solution to the Neumann problem* (3.8) *is in $H^{m+2}(\Omega)$.*
(iv) *Assuming $g \in H^{m+\frac{1}{2}}(\partial\Omega)$ and $\gamma \in \mathcal{C}^{m+1}(\partial\Omega)$, the solution to the Robin problem* (3.10) *is in $H^{m+2}(\Omega)$.*

Remark 3.11.

(i) The reader who is not familiar with Sobolev spaces involving fractional exponents may replace an assumption such as $g \in H^{m+\frac{3}{2}}(\partial\Omega)$ by $g \in \mathcal{C}^{m+1}(\partial\Omega)$ and $g^{(m+1)} \in \mathcal{C}^{0,1}(\partial\Omega)$; see Example B.32(ii).

(ii) There is no regularity result for the mixed Dirichlet–Neumann problem. Indeed, even if f, g, and the domain Ω are smooth, the solution u may not necessarily belong to $H^2(\Omega)$. For instance, in two dimensions, the solution to $-\Delta u = 0$ on the upper half-plane $\{x_2 > 0\}$ with the mixed Dirichlet–Neumann conditions

$$\begin{aligned} &\partial_2 u = 0, &&\text{for } x_1 \leq 0 \text{ and } x_2 = 0,\\ &u = r^{\frac{1}{2}}\sin(\tfrac{1}{2}\theta), &&\text{otherwise,} \end{aligned}$$

is $u(x_1, x_2) = r^{\frac{1}{2}}\sin(\frac{1}{2}\theta)$. Clearly, $u \notin H^2$ owing to the singularity at the origin.

(iii) Theorem 3.10 can be extended to more general Sobolev spaces; see, e.g., [GiR86, pp. 12–15]. For instance, let p be a real satisfying $1 < p < \infty$ and let $m \geq 0$. Let $f \in W^{m,p}(\Omega)$ and $g \in W^{m+2-\frac{1}{p},p}(\partial\Omega)$. Then, the solution to the non-homogeneous Dirichlet problem (3.7) is in $W^{m+2,p}(\Omega)$. □

Theorem 3.12 (Convex polyhedron). *Let Ω be a* convex polyhedron *and denote by $\bigcup_{j=1}^{J} \partial\Omega_j$ the set of boundary faces (edges in two dimensions). Assume that the coefficients σ_{ij} are in $\mathcal{C}^1(\overline{\Omega})$ and that the coefficients β_i and μ are in $\mathcal{C}^0(\overline{\Omega})$. Then:*

(i) *The solution to the homogeneous Dirichlet problem* (3.5) *is in $H^2(\Omega)$.*
(ii) *In dimension* 2, *if $g \in H^{\frac{3}{2}}(\partial\Omega)$, the solution to the non-homogeneous Dirichlet problem* (3.7) *is in $H^2(\Omega)$.*
(iii) *In dimension* 2, *if $g_{|\partial\Omega_j} \in H^{\frac{1}{2}}(\partial\Omega_j)$ for $1 \le j \le J$, the solution to the Neumann problem* (3.8) *is in $H^2(\Omega)$. In dimension* 3, *the conclusion still holds if $g = 0$.*

Remark 3.13.

(i) When the polyhedron Ω is not convex, the best regularity result is $u \in H^{\frac{3}{2}}(\Omega)$. In particular, it can be shown (see [Gri85, Gri92]) that in the neighborhood of a vertex S with an interior angle $\omega > \pi$, the solution u to the homogeneous Dirichlet problem can be decomposed into the form

$$u = \Upsilon + \widetilde{u},$$

where $\widetilde{u} \in H^2(\Omega)$ and Υ is a singular function behaving like $r^{\frac{\pi}{\omega}}$ in the neighborhood of S, r being the distance to S.

(ii) Theorem 3.12 can be extended to more general Sobolev spaces. For instance, let p be a real satisfying $1 < p < \infty$, and let $f \in L^p(\Omega)$. Then, the solution to the homogeneous Dirichlet problem (3.5) posed on a convex polyhedron is in $W^{2,p}(\Omega)$.

(iii) The assumption on g in Theorem 3.12(ii) can be weakened as follows: Denote by $\{S_j\}_{1\le j\le J}$ the vertices of $\partial\Omega$ so that $\partial\Omega_j$ is the segment S_jS_{j+1}, and conventionally set $S_{J+1} = S_1$ and $\partial\Omega_{J+1} = \partial\Omega_1$. Then, if $g_{|\partial\Omega_j} \in H^{\frac{3}{2}}(\partial\Omega_j)$ and $g_{|\partial\Omega_j}(S_j) = g_{|\partial\Omega_{j+1}}(S_{j+1})$ for all $1 \le j \le J$, the solution to the non-homogeneous Dirichlet problem (3.7) is in $H^2(\Omega)$.

(iv) A regularity result analogous to Theorem 3.12(iii) is valid for the purely diffusive Neumann problem (3.16). □

Definition 3.14 (Smoothing property). *Problem* (3.11) *is said to have* smoothing properties in Ω *if assumption* (AN1) *in* §2.3.4 *is satisfied with $Z = H^2(\Omega) \cap H_0^1(\Omega)$, $L = L^2(\Omega)$, and $l(\cdot,\cdot) = (\cdot,\cdot)_{0,\Omega}$, i.e., if there exists c_S such that, for all $\varphi \in L^2(\Omega)$, the solution w to the adjoint problem:*

$$\begin{cases} \text{Seek } w \in V \text{ such that} \\ a(v,w) = \int_\Omega \varphi v, \quad \forall v \in V, \end{cases} \tag{3.17}$$

satisfies $\|w\|_{2,\Omega} \le c_S \|\varphi\|_{0,\Omega}$.

Remark 3.15. Because the Laplace operator is self-adjoint, the Laplacian has *smoothing properties in Ω* if the unique solution to the homogeneous Dirichlet problem with $f \in L^2(\Omega)$ is in $H^2(\Omega) \cap H_0^1(\Omega)$, i.e., if the operator $(-\Delta)^{-1} : L^2(\Omega) \to H^2(\Omega) \cap H_0^1(\Omega)$ is an isomorphism. □

3.2 Scalar Elliptic PDEs: Approximation

This section reviews various finite element methods to approximate second-order, scalar, elliptic PDEs. Assume that the well-posedness conditions stated in Theorem 3.8 hold and denote by $u \in V$ the unique solution to (3.11).

3.2.1 H^1-conforming approximation

Let Ω be a polyhedron in $\mathbb{R}^d$, let $\{\mathcal{T}_h\}_{h>0}$ be a family of meshes of Ω, and let $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$ be a reference Lagrange finite element of degree $k \geq 1$. Let $L^k_{\mathrm{c},h}$ be the H^1-conforming approximation space defined by

$$L^k_{\mathrm{c},h} = \{v_h \in \mathcal{C}^0(\overline{\Omega});\, \forall K \in \mathcal{T}_h,\, v_h \circ T_K \in \widehat{P}\}. \tag{3.18}$$

For instance, $L^k_{\mathrm{c},h} = P^k_{\mathrm{c},h}$ or $Q^k_{\mathrm{c},h}$ defined in (1.76) and (1.77), respectively, if a $\mathbb{P}_k$ or $\mathbb{Q}_k$ Lagrange finite element is used. To obtain a V-conforming approximation space, we must account for the boundary conditions, i.e., we set

$$V_h = L^k_{\mathrm{c},h} \cap V. \tag{3.19}$$

This yields $V_h = \{v_h \in L^k_{\mathrm{c},h};\, v_h = 0 \text{ on } \partial\Omega\}$ for the homogeneous Dirichlet problem and $V_h = L^k_{\mathrm{c},h}$ for the Neumann and the Robin problems. For the mixed Dirichlet–Neumann problem, we assume, for the sake of simplicity, that $\partial\Omega_{\mathrm{D}}$ is a union of mesh faces; in this case, a suitable approximation space is $V_h = \{v_h \in L^k_{\mathrm{c},h};\, v_h = 0 \text{ on } \partial\Omega_{\mathrm{D}}\}$.

Consider the approximate problem:

$$\begin{cases} \text{Seek } u_h \in V_h \text{ such that} \\ a(u_h, v_h) = f(v_h), \quad \forall v_h \in V_h. \end{cases} \tag{3.20}$$

Our goal is to estimate the error $u - u_h$, first in the H^1-norm, then in the L^2-norm, and finally in more general norms.

Theorem 3.16 (H^1-estimate). *Let Ω be a polyhedron in $\mathbb{R}^d$ and let $\{\mathcal{T}_h\}_{h>0}$ be a shape-regular family of geometrically conforming meshes of Ω. Let V_h be defined in* (3.19)*. Then, $\lim_{h\to 0} \|u - u_h\|_{1,\Omega} = 0$. Furthermore, if $u \in H^s(\Omega)$ with $\frac{d}{2} < s \leq k+1$, there exists c such that*

$$\forall h, \quad \|u - u_h\|_{1,\Omega} \leq c\, h^{s-1} |u|_{s,\Omega}. \tag{3.21}$$

Proof. Since $s > \frac{d}{2}$, Corollary B.43 implies that u is in the domain of the Lagrange interpolation operator $\mathcal{I}^k_h$ associated with $L^k_{\mathrm{c},h}$. Moreover, $\mathcal{I}^k_h u \in V_h$ since the Lagrange interpolant preserves Dirichlet boundary conditions. As a result, Céa's Lemma yields

$$\|u - u_h\|_{1,\Omega} \leq c \left(\inf_{v_h \in V_h} \|u - v_h\|_{1,\Omega} \right) \leq c\, \|u - \mathcal{I}^k_h u\|_{1,\Omega}.$$

Owing to Corollary 1.110 (with $p = 2$) and since $s \leq k + 1$,

$$\|u - \mathcal{I}_h^k u\|_{1,\Omega} \leq c\, h^{s-1} |u|_{s,\Omega}.$$

Combining the above inequalities yields (3.21). If $u \in H^1(\Omega)$ only, the convergence of u_h results from the density of $H^s(\Omega) \cap V$ in V. □

Remark 3.17. The assumption $s > \frac{d}{2}$ in Theorem 3.16 can be lifted on simplicial meshes by considering the Clément or the Scott–Zhang interpolation operator instead of the Lagrange interpolation operator; details are left as an exercise. □

For the sake of simplicity, we shall henceforth restrict ourselves to homogeneous Dirichlet conditions.

Theorem 3.18 (L^2-estimate). *Along with the hypotheses of Theorem* 3.16*, assume* $V = H_0^1(\Omega)$*,* $V_h = L_{c,h}^k \cap H_0^1(\Omega)$*, and that problem* (3.11) *has smoothing properties. Then, there exists c such that*

$$\forall h, \quad \|u - u_h\|_{0,\Omega} \leq c\, h |u - u_h|_{1,\Omega}. \tag{3.22}$$

Proof. Apply the Aubin–Nitsche Lemma. □

Example 3.19. Consider the homogeneous Dirichlet problem posed on a convex polyhedron, say Ω. Owing to Theorem 3.12, the Laplacian has smoothing properties in Ω. Therefore, using $\mathbb{P}_1$ finite elements yields the estimates

$$\forall h, \quad \|u - u_h\|_{0,\Omega} + h\|u - u_h\|_{1,\Omega} \leq c\, h^2 \|f\|_{0,\Omega}.$$ □

Using again duality techniques, it is possible to derive negative-norm estimates for the error, provided Lagrange finite elements of degree 2 at least are employed. For $s \geq 1$, we define the norm

$$\|v\|_{-s,\Omega} = \sup_{z \in H^s(\Omega) \cap H_0^1(\Omega)} \frac{(v, z)_{0,\Omega}}{\|z\|_{s,\Omega}}.$$

Recall that this is not the norm considered to define the dual space $H^{-s}(\Omega)$, except in the particular case $s = 1$. Here, the norm $\|\cdot\|_{-s,\Omega}$ is simply used as a quantitative measure for functions in $L^2(\Omega)$.

Theorem 3.20 (Negative-norm estimates). *Along with the hypotheses of Theorem* 3.16*, assume* $V_h \subset H_0^1(\Omega)$*. Assume* $k \geq 2$ *and let* $1 \leq s \leq k - 1$*. Assume that there exists a stability constant* $c_S > 0$ *such that, for all* $\varphi \in H^s(\Omega)$*, the solution* w *to the adjoint problem* (3.17) *satisfies* $\|w\|_{s+2,\Omega} \leq c_S \|\varphi\|_{s,\Omega}$*. Then, there exists c such that*

$$\forall h, \quad \|u - u_h\|_{-s,\Omega} \leq c\, h^{s+1} \|u - u_h\|_{1,\Omega}. \tag{3.23}$$

Proof. Let $1 \le s \le k-1$, let $z \in H^s(\Omega) \cap H_0^1(\Omega)$, and let $w \in H^{s+2}$ be the solution to the adjoint problem (3.17) with data z. Then, for any $w_h \in V_h$, Galerkin orthogonality implies

$$\begin{aligned}(u-u_h, z)_{0,\Omega} &= a(u-u_h, w) \\ &= a(u-u_h, w-w_h) \\ &\le \|a\| \, \|u-u_h\|_{1,\Omega} \|w-w_h\|_{1,\Omega}.\end{aligned}$$

Since $w \in H^{s+2} \cap H_0^1(\Omega)$, it is legitimate to take for w_h the Lagrange interpolant of w in V_h (if $s+2 \le \frac{d}{2}$, the Clément or the Scott–Zhang interpolation operator must be considered). Corollary 1.109 implies

$$\|w-w_h\|_{1,\Omega} \le c\, h^{s+1} |w|_{s+2,\Omega},$$

and, therefore, $\|w-w_h\|_{1,\Omega} \le c\, h^{s+1} \|z\|_{s,\Omega}$. Hence,

$$(u-u_h, z)_{0,\Omega} \le c\, h^{s+1} \|u-u_h\|_{1,\Omega} \|z\|_{s,\Omega},$$

and taking the supremum over z yields the desired estimate. □

Error estimates in the Sobolev norms $\|\cdot\|_{1,p,\Omega}$ are useful in the context of nonlinear problems; see [BrS94, p. 188] for an example. For second-order, elliptic PDEs, the main result is a stability property for the discrete problem (3.20) in the $W^{1,p}$-norm. The result requires some technical assumptions on the discretization and some regularity properties for the exact problem. For the sake of brevity, the former are not restated here. These assumptions hold for the Lagrange finite elements introduced in §1.2.3–§1.2.5 and for quasi-uniform families of geometrically conforming meshes.

Theorem 3.21 ($W^{1,p}$-stability). *Let Ω be a polyhedron in $\mathbb{R}^d$ with $d \le 3$. Assume that:*

(i) *The bilinear form a is elliptic and coercive on $H_0^1(\Omega)$.*
(ii) *The assumptions of* [BrS94, p. 170] *on the finite element space V_h hold.*
(iii) *The diffusion coefficients are such that $\sigma \in [W^{1,p}(\Omega)]^{d,d}$ for $p > 2$ if $d = 2$ and for $p \ge \frac{12}{5}$ if $d = 3$.*
(iv) *There exists $\delta > d$ such that for all $q \in \,]1, \delta[$ and for all $f \in L^q(\Omega)$, the unique solution to the exact problem* (3.11) *posed on $H_0^1(\Omega)$ is in $W^{2,q}(\Omega)$. Assume also that the adjoint problem* (3.17) *satisfies the same regularity property.*

Then, there exist c and $h_0 > 0$ such that

$$\forall h \le h_0,\ \forall 1 < p \le \infty, \quad \|u_h\|_{1,p,\Omega} \le c\, \|u\|_{1,p,\Omega}. \tag{3.24}$$

Proof. See [RaS82] and [BrS94, p. 169]. □

Remark 3.22. Owing to assumption (iv) and Corollary B.43, the solution to (3.11) is in $W^{1,\infty}(\Omega)$ whenever $f \in L^q(\Omega)$ with $q > d$. □

Corollary 3.23 ($W^{1,p}$-estimate). *Under the assumptions of Theorem* 3.21*,*

$$\lim_{h\to 0} \|u-u_h\|_{1,p,\Omega} = 0. \tag{3.25}$$

Furthermore, if $u \in W^{s,p}(\Omega)$ for some $s \geq 2$,

$$\forall h, \quad \|u-u_h\|_{1,p,\Omega} \leq c\, h^l |u|_{l+1,p,\Omega}, \tag{3.26}$$

with $l = \min(k, s-1)$ and k is the degree of the finite element.

Proof. Let $v_h \in V_h$ and $1 < p \leq \infty$. Since $a(u_h - v_h, w_h) = a(u - v_h, w_h)$ for all $w_h \in V_h$, Theorem 3.21 implies $\|u_h - v_h\|_{1,p,\Omega} \leq c\, \|u - v_h\|_{1,p,\Omega}$. Using the triangle inequality readily yields the estimate

$$\|u-u_h\|_{1,p,\Omega} \leq c \inf_{v_h\in V_h} \|u-v_h\|_{1,p,\Omega}.$$

Equations (3.25) and (3.26) then result from (1.100) and (1.101). ☐

Using duality techniques, one can obtain an L^p-norm estimate.

Proposition 3.24 (L^p-estimate). *Under the assumptions of Theorem* 3.21*, there exist c and $h_0 > 0$ such that*

$$\forall h \leq h_0,\ \forall \delta' < p < \infty, \quad \|u-u_h\|_{L^p(\Omega)} \leq c\, h \|u-u_h\|_{1,p,\Omega}, \tag{3.27}$$

where $\frac{1}{\delta} + \frac{1}{\delta'} = 1$ and δ is defined in assumption (iv) *of Theorem* 3.21*.*

Proof. The proof uses duality techniques; see Exercise 3.8. ☐

The derivation of L^∞-norm estimates is more technical; see [Nit76, Sco76]. In the framework of the above assumptions, one can show that for finite elements of degree 2 at least,

$$\forall h \leq h_0, \quad \|u-u_h\|_{L^\infty(\Omega)} \leq c\, h \|u-u_h\|_{1,\infty,\Omega}.$$

However, for piecewise linear approximations in two dimensions, the best error estimate in the L^∞-norm is

$$\forall h \leq h_0, \quad \|u-u_h\|_{L^\infty(\Omega)} \leq c\, h |\ln h|\, \|u-u_h\|_{1,\infty,\Omega}.$$

Remark 3.25.

(i) Let x_i be a mesh node, let $\delta_{x=x_i}$ be the Dirac mass at x_i, and assume that the following problem:

$$\begin{cases} \text{Seek } G_i \in V \text{ such that} \\ a(v, G_i) = \langle \delta_{x=x_i}, v\rangle_{\mathcal{D}',\mathcal{D}}, \quad \forall v \in V, \end{cases}$$

is well-posed. Its solution G_i is said to be the *Green function* at point x_i. If it happens that $G_i \in V_h$, Galerkin orthogonality implies

$$0 = a(u - u_h, G_i) = \langle \delta_{x=x_i}, u - u_h \rangle_{\mathcal{D}',\mathcal{D}} = u(x_i) - u_h(x_i),$$

showing that the error vanishes identically at the mesh nodes. This situation occurs when approximating the Laplacian in one dimension with Lagrange finite elements since, in this case, the Green function is continuous and piecewise linear; see also Example 3.90 for the Green function associated with a beam flexion problem.

(ii) When the solution u is not smooth enough, error estimates in weaker norms can be derived. For instance, under the assumptions of Theorem 3.18 and assuming that the family of meshes $\{\mathcal{T}_h\}_{h>0}$ is quasi-uniform, one can show (see, e.g., [QuV97, p. 174]) that there exists c such that

$$\forall h, \quad \|u - u_h\|_{L^\infty(\Omega)} \le c\, h^{l+1-\frac{d}{2}} |u|_{l+1,\Omega},$$

with $l \le k$. For instance, if the solution u is in $H^2(\Omega)$, the convergence in the L^∞-norm is first-order in dimension 2, and of order $\frac{1}{2}$ in dimension 3. It would scale like $h^2|\ln h|$ provided $u \in W^{2,\infty}(\Omega)$ and $\mathbb{P}_1$ finite elements are used.

(iii) Consider the purely diffusive version of problem (3.11). When the diffusion coefficients do not satisfy assumption (iii) of Theorem 3.21, but are only measurable and bounded, it is still possible to prove a stability result in $W^{1,p}(\Omega)$ if $|p-2|$ is small enough. The proof uses the inf-sup condition to express the stability of the exact problem; see [BrS94, p. 184]. □

3.2.2 Non-homogeneous Dirichlet boundary conditions

Given $f \in L^2(\Omega)$ and $g \in H^{\frac{1}{2}}(\partial\Omega)$, the non-homogeneous version of problem (3.11) is:

$$\begin{cases} \text{Seek } u \in H^1(\Omega) \text{ such that} \\ a(u,v) = \int_\Omega fv, \quad \forall v \in H^1_0(\Omega), \\ \gamma_0(u) = g, \qquad\quad \text{in } H^{\frac{1}{2}}(\partial\Omega), \end{cases} \tag{3.28}$$

where γ_0 is the trace operator defined in §B.3.5. We assume that problem (3.28) is well-posed, namely that the bilinear form a satisfies the assumptions of the BNB Theorem on $H^1_0(\Omega) \times H^1_0(\Omega)$; see §2.1.4 for the theoretical background. For instance, a may be coercive on $H^1_0(\Omega)$. Henceforth, the reader unfamiliar with fractional Sobolev spaces may replace the assumption $g \in H^{\frac{1}{2}}(\partial\Omega)$ by $g \in \mathcal{C}^{0,1}(\partial\Omega)$ (since $\mathcal{C}^{0,1}(\partial\Omega) \subset H^{\frac{1}{2}}(\partial\Omega)$ with continuous embedding; see Example B.32(ii)).

We seek an approximate solution to (3.28) in the discrete space $V_h = L^k_{\mathrm{c},h}$ defined in (3.18). Let N be the dimension of V_h. Denote by $\{\varphi_1, \dots, \varphi_N\}$ the nodal basis of V_h and by $\{a_1, \dots, a_N\}$ the associated nodes. Recall that the Lagrange interpolant of a continuous function u on Ω is defined as

$$\mathcal{I}_h u = \sum_{i=1}^{N} u(a_i)\varphi_i.$$

Assuming that g is continuous on $\partial\Omega$, we introduce its Lagrange interpolant

$$\mathcal{I}_h^\partial g = \sum_{a_i \in \partial\Omega} g(a_i)\gamma_0(\varphi_i).$$

Since $\{\varphi_1, \ldots, \varphi_N\}$ is a nodal basis,

$$(a_i \notin \partial\Omega) \implies (\gamma_0(\varphi_i) = 0). \tag{3.29}$$

As a result, for $u \in \mathcal{C}^0(\overline{\Omega}) \cap H^1(\Omega)$,

$$\begin{aligned}\gamma_0(\mathcal{I}_h u) &= \gamma_0\left(\sum_{i=1}^N u(a_i)\varphi_i\right) = \sum_{i=1}^N u(a_i)\gamma_0(\varphi_i)\\ &= \sum_{a_i \in \partial\Omega} u(a_i)\gamma_0(\varphi_i) = \mathcal{I}_h^\partial\big(\gamma_0(u)\big),\end{aligned}$$

so that $\gamma_0 \circ \mathcal{I}_h = \mathcal{I}_h^\partial \circ \gamma_0$, i.e., the trace of the interpolant of a sufficiently smooth function coincides with the interpolant of its trace.

Consider the approximate problem :

$$\begin{cases}\text{Seek } u_h \in V_h \text{ such that}\\ a(u_h, v_h) = \int_\Omega f v_h, \quad \forall v_h \in V_{h0},\\ \gamma_0(u_h) = \mathcal{I}_h^\partial g, \qquad \text{on } \partial\Omega,\end{cases} \tag{3.30}$$

where $V_{h0} = \{v_h \in V_h; \gamma_0(v_h) = 0\} \subset H_0^1(\Omega)$. Assume that the bilinear form a satisfies the condition (BNB1_h) on $V_{h0} \times V_{h0}$.

Proposition 3.26. *If g is smooth enough to have a lifting in $\mathcal{C}^0(\overline{\Omega}) \cap H^1(\Omega)$, problem* (3.30) *is well-posed.*

Proof. Let u_g be a lifting of g in $\mathcal{C}^0(\overline{\Omega}) \cap H^1(\Omega)$. Clearly,

$$\gamma_0(\mathcal{I}_h u_g) = \mathcal{I}_h^\partial\big(\gamma_0(u_g)\big) = \mathcal{I}_h^\partial(g) = \gamma_0(u_h).$$

Therefore, setting $\phi_h = u_h - \mathcal{I}_h u_g$ yields $\phi_h \in V_{h0}$ and $a(\phi_h, v_h) = \int_\Omega f v_h - a(\mathcal{I}_h u_g, v_h)$ for all $v_h \in V_{h0}$. Since the bilinear form a satisfies the condition (BNB1_h) on $V_{h0} \times V_{h0}$, problem (3.30) is well-posed. □

The approximate problem (3.30) being well-posed, our goal is now to estimate the approximation error $u - u_h$ in the H^1- and L^2-norms, where u and u_h solve (3.28) and (3.30), respectively. The results below generalize Céa's and Aubin–Nitsche Lemmas; see Exercises 3.9 and 3.10 for proofs.

Lemma 3.27. *Along with the hypotheses of Proposition* 3.26*, assume that the exact solution u is sufficiently smooth for its Lagrange interpolant $\mathcal{I}_h u$ to be well-defined. Set $\|a\| := \|a\|_{H^1(\Omega),H^1(\Omega)}$. Then,*

$$\|u - u_h\|_{1,\Omega} \leq \left(1 + \tfrac{\|a\|}{\alpha_h}\right) \|u - \mathcal{I}_h u\|_{1,\Omega}.$$

Lemma 3.28. *Along with the hypotheses of Lemma* 3.27*, assume that:*

(i) *Problem* (3.11) *has smoothing properties.*
(ii) *The bilinear form a satisfies the following continuity property: there exists c such that, for all $v \in H^1(\Omega)$ and $w \in H^2(\Omega)$,*

$$|a(v,w)| \leq c\,(\|v\|_{0,\Omega} + \|\gamma_0(v)\|_{0,\partial\Omega})\|w\|_{2,\Omega}.$$

(iii) *There exists an interpolation constant $c > 0$ such that*

$$\forall h,\ \forall \theta \in H^2(\Omega), \quad \|\theta - \mathcal{I}_h\theta\|_{1,\Omega} \leq c\,h\|\theta\|_{2,\Omega}.$$

Then, there exists c such that

$$\forall h, \quad \|u - u_h\|_{0,\Omega} \leq c\,(h\|\mathcal{I}_h u - u\|_{1,\Omega} + \|\mathcal{I}_h u - u\|_{0,\Omega} + \|\mathcal{I}_h g - g\|_{0,\partial\Omega}).$$

Corollary 3.29. *Let Ω be a polyhedron, let $\{\mathcal{T}_h\}_{h>0}$ be a shape-regular family of geometrically conforming meshes of Ω, and let V_h be a H^1-conforming approximation space based on $\mathcal{T}_h$ and a Lagrange finite element of degree $k \geq 1$. Along with the hypotheses of Lemma* 3.28*, assume that the exact solution u is in $H^{k+1}(\Omega)$. Then, there is c such that*

$$\forall h, \quad \|u - u_h\|_{0,\Omega} + h\|u - u_h\|_{1,\Omega} \leq c\,h^{k+1}\|u\|_{k+1,\Omega}. \tag{3.31}$$

Proof. Direct consequence of Lemmas 3.27 and 3.28. □

Example 3.30. Assumptions (i)–(iii) of Lemma 3.28 are satisfied for the Poisson problem posed in dimension 2 or 3 on either a convex polyhedron or a domain of class $\mathcal{C}^2$ and for a Lagrange finite element of degree $k \geq 1$ using a shape-regular family of meshes. More precisely, assumption (i) is stated in §3.1.3. Assumption (ii) results from the identity

$$\forall v \in H^1(\Omega),\ \forall w \in H^2(\Omega), \quad a(v,w) = \int_\Omega \nabla v \cdot \nabla w = -\int_\Omega v\Delta w + \int_{\partial\Omega} v\,\partial_n w,$$

together with the continuity of the normal derivative operator $\gamma_1 : H^2(\Omega) \to L^2(\partial\Omega)$; see Theorem B.54. Assumption (iii) is a direct consequence of Corollary 1.109. □

3.2.3 Crouzeix–Raviart non-conforming approximation

In this section, we present an example of non-conforming approximation for the Laplacian based on the Crouzeix–Raviart finite element. Let Ω be a polyhedron in $\mathbb{R}^d$ and let u be the solution to the homogeneous Dirichlet problem with data $f \in L^2(\Omega)$. Assume that $u \in H^2(\Omega)$. This property holds, for instance, if Ω is convex; see Theorem 3.12.

Let $\{\mathcal{T}_h\}_{h>0}$ be a shape-regular family of geometrically conforming, affine meshes of Ω. Let $P^1_{\mathrm{pt},h}$ be the Crouzeix–Raviart finite element space defined in (1.69). Let

$$P^1_{\mathrm{pt},h,0} = \left\{ v_h \in P^1_{\mathrm{pt},h};\, \forall F \in \mathcal{F}_h^\partial,\ \int_F v_h = 0 \right\}, \tag{3.32}$$

where $\mathcal{F}_h^\partial$ denotes the set of faces of the mesh located at the boundary. Recall that $\dim P^1_{\mathrm{pt},h,0} = N_{\mathrm{f}}^{\mathrm{i}}$, the number of internal faces (edges in two dimensions) in the mesh. Since functions in $P^1_{\mathrm{pt},h,0}$ can be discontinuous, the bilinear form $\int_\Omega \nabla u{\cdot}\nabla v$ must be broken over the elements, yielding:

$$\begin{cases} \text{Seek } u_h \in P^1_{\mathrm{pt},h,0} \text{ such that} \\ a_h(u_h, v_h) = f(v_h), \quad \forall v_h \in P^1_{\mathrm{pt},h,0}, \end{cases} \tag{3.33}$$

with

$$a_h(u_h, v_h) = \sum_{K \in \mathcal{T}_h} \int_K \nabla u_h{\cdot}\nabla v_h \qquad \text{and} \qquad f(v_h) = \int_\Omega f v_h. \tag{3.34}$$

Set $V(h) = P^1_{\mathrm{pt},h,0} + H^1_0(\Omega)$ and for $v_h \in V(h)$ define the broken H^1-seminorm

$$|v_h|_{h,1,\Omega} = \left(\sum_{K \in \mathcal{T}_h} \|\nabla v_h\|^2_{0,K} \right)^{\frac{1}{2}}.$$

Equip the space $V(h)$ with the norm $\|\cdot\|_{V(h)} = \|\cdot\|_{0,\Omega} + |\cdot|_{h,1,\Omega}$.

Our goal is to investigate the convergence of the solution to the approximate problem (3.33) in the norm $\|\cdot\|_{V(h)}$. To this end, we must exhibit stability, continuity, consistency, and approximability properties; see §2.3.1. To obtain a stability property for problem (3.33), we would like to establish the coercivity of a_h on $P^1_{\mathrm{pt},h,0}$. Since $P^1_{\mathrm{pt},h,0} \not\subset H^1_0(\Omega)$, this is a non-trivial result.

Lemma 3.31 (Extended Poincaré inequality). *There exists c depending only on Ω such that, for all $h \le 1$,*

$$\forall u \in V(h), \quad c\,\|u\|_{0,\Omega} \le |u|_{h,1,\Omega}. \tag{3.35}$$

Proof. We restate the proof given in [Tem77, Prop. 4.13]; see also [CrG02]. Let $u \in V(h)$; then

$$\|u\|_{0,\Omega} \le \sup_{v \in L^2(\Omega)} \frac{(u,v)_{0,\Omega}}{\|v\|_{0,\Omega}}.$$

For $v \in L^2(\Omega)$, there exists $p \in [H^1(\Omega)]^d$ such that $\nabla{\cdot}p = v$ and $\|p\|_{1,\Omega} \le c\,\|v\|_{0,\Omega}$, where c depends only on Ω. Integration by parts yields

$$(u,v)_{0,\Omega} = (u, \nabla{\cdot}p)_{0,\Omega} = -\sum_{K \in \mathcal{T}_h} (\nabla u, p)_{0,K} + \sum_{K \in \mathcal{T}_h} \sum_{F \in \partial K} \int_F (p{\cdot}n_K) u,$$

where F is a face of K and n_K is the outward normal to K. Consider the second term in the right-hand side of the above equality. If F is an interface,

$F = K_m \cap K_n$, it appears twice in the sum, and since $\int_F u_{|K_m} = \int_F u_{|K_n}$ for $u \in V(h)$, we can subtract from $p{\cdot}n_K$ a constant function on F that we take equal to $\overline{p}{\cdot}n_K$ with $\overline{p} = \frac{1}{\mathrm{meas}(F)}\int_F p$. The same conclusion is valid for faces located at the boundary since $\int_F u = 0$ on such faces. Therefore,

$$\begin{aligned}\sum_{K\in\mathcal{T}_h}\sum_{F\in\partial K}\int_F (p{\cdot}n_K)u &= \sum_{K\in\mathcal{T}_h}\sum_{F\in\partial K}\int_F (p-\overline{p}){\cdot}n_K u \\ &= \sum_{K\in\mathcal{T}_h}\sum_{F\in\partial K}\int_F (p-\overline{p}){\cdot}n_K(u-\overline{u}),\end{aligned}$$

and using Lemma 3.32 below, this yields

$$\begin{aligned}(u,v)_{0,\Omega} &\le \|p\|_{0,\Omega}|u|_{h,1,\Omega} + \sum_{K\in\mathcal{T}_h} c\,h_K^{\frac{1}{2}}|p|_{1,K}h_K^{\frac{1}{2}}|u|_{1,K} \\ &\le \|p\|_{0,\Omega}|u|_{h,1,\Omega} + c\,h|p|_{1,\Omega}|u|_{h,1,\Omega}.\end{aligned}$$

Since $h \le 1$, $(u,v)_{0,\Omega} \le c\,\|v\|_{0,\Omega}|u|_{h,1,\Omega}$ and, hence, (3.35) holds. □

Lemma 3.32. *Let $\{\mathcal{T}_h\}_{h>0}$ be a shape-regular family of geometrically conforming affine meshes. Let $m \ge 1$ be a fixed integer. For $K \in \mathcal{T}_h$, $\psi \in [H^1(K)]^m$, and a face $F \in \partial K$, set $\overline{\psi} = \frac{1}{\mathrm{meas}(F)}\int_F \psi$. Then, there exists c such that*

$$\forall h,\ \forall K \in \mathcal{T}_h,\ \forall F \in \partial K,\ \forall \psi \in [H^1(K)]^m, \quad \|\psi - \overline{\psi}\|_{0,F} \le c\,h_K^{\frac{1}{2}}|\psi|_{1,K}. \tag{3.36}$$

Proof. Let $K \in \mathcal{T}_h$, let $\psi \in [H^1(K)]^m$, and consider a face $F \in \partial K$. Let $\widehat{K}$ be the reference simplex and let $T_K : \widehat{K} \to K$ be the corresponding affine transformation with Jacobian J_K. Letting $\widehat{F} = T_K^{-1}(F)$, it is clear that

$$\|\psi - \overline{\psi}\|_{0,F} \le \left(\frac{\mathrm{meas}\,F}{\mathrm{meas}\,\widehat{F}}\right)^{\frac{1}{2}} \|\widehat{\psi} - \overline{\widehat{\psi}}\|_{0,\widehat{F}} \le c\left(\frac{\mathrm{meas}\,F}{\mathrm{meas}\,\widehat{F}}\right)^{\frac{1}{2}} \|\widehat{\psi} - \overline{\widehat{\psi}}\|_{1,\widehat{K}},$$

owing to the Trace Theorem B.52. The Deny–Lions Lemma implies

$$\|\widehat{\psi} - \overline{\widehat{\psi}}\|_{1,\widehat{K}} \le c\,|\widehat{\psi}|_{1,\widehat{K}}.$$

Returning to element K and using the shape-regularity of the mesh yields

$$\begin{aligned}\|\psi - \overline{\psi}\|_{0,F} &\le c\left(\frac{\mathrm{meas}\,F}{\mathrm{meas}\,\widehat{F}}\right)^{\frac{1}{2}} \|J_K^{-1}\|_d \left(\frac{\mathrm{meas}\,\widehat{K}}{\mathrm{meas}\,K}\right)^{\frac{1}{2}} |\psi|_{1,K} \\ &\le c\,h_K^{\frac{d-1}{2}} h_K h_K^{-\frac{d}{2}}|\psi|_{1,K} \le c\,h_K^{\frac{1}{2}}|\psi|_{1,K},\end{aligned}$$

thereby completing the proof. □

Corollary 3.33 (Stability). *The bilinear form a_h defined in* (3.34) *is coercive on $P^1_{\mathrm{pt},h,0}$.*

Proof. Direct consequence of the extended Poincaré inequality (3.35). □

Lemma 3.34 (Continuity). *The bilinear form a_h defined in* (3.34) *is uniformly bounded on $V(h) \times V(h)$.*

Proof. Use the fact that, for all $u_h \in V(h)$, $|u_h|_{h,1,\Omega} \le \|u_h\|_{V(h)}$. □

Corollary 3.35 (Well-Posedness). *Problem* (3.33) *is well-posed.*

Proof. Direct consequence of the Lax–Milgram Lemma. □

Lemma 3.36 (Asymptotic consistency). *Let u be the solution to the homogeneous Dirichlet problem with data $f \in L^2(\Omega)$. Assume that $u \in H^2(\Omega)$. Then, there exists c such that*

$$\forall h,\ \forall w_h \in P^1_{\mathrm{pt},h,0},\quad \frac{|f(w_h) - a_h(u, w_h)|}{\|w_h\|_{V(h)}} \le c\,h|u|_{2,\Omega}. \tag{3.37}$$

Proof. Let $w_h \in P^1_{\mathrm{pt},h,0}$. Since $f = -\Delta u$,

$$a_h(u, w_h) - f(w_h) = \sum_{K\in\mathcal{T}_h} \int_K (\nabla u\cdot\nabla w_h - f w_h) = \sum_{K\in\mathcal{T}_h}\sum_{F\in\partial K} \int_F \nabla u\cdot n_K\, w_h.$$

Since each face F of an element K located inside Ω appears twice in the above sum, we can subtract from w_h its mean-value on the face, $\overline{w_h}$. If F is on $\partial\Omega$, it is clear that $\overline{w_h} = 0$. Therefore,

$$a_h(u, w_h) - f(w_h) = \sum_{K\in\mathcal{T}_h}\sum_{F\in\partial K} \int_F \nabla u\cdot n_K (w_h - \overline{w_h}).$$

We can also subtract from ∇u its mean-value on F, $\overline{\nabla u}$, yielding

$$a_h(u, w_h) - f(w_h) = \sum_{K\in\mathcal{T}_h}\sum_{F\in\partial K} \int_F (\nabla u - \overline{\nabla u})\cdot n_K (w_h - \overline{w_h}).$$

The Cauchy–Schwarz inequality implies

$$|a_h(u, w_h) - f(w_h)| \le \sum_{K\in\mathcal{T}_h}\sum_{F\in\partial K} \|\nabla u - \overline{\nabla u}\|_{0,F} \|w_h - \overline{w_h}\|_{0,F}.$$

Lemma 3.32 yields

$$\begin{aligned} |a_h(u, w_h) - f(w_h)| &\le \sum_{K\in\mathcal{T}_h} c\,h_K^{\frac12} |u|_{2,K} h_K^{\frac12} |w_h|_{1,K} \\ &\le c\,h \left(\sum_{K\in\mathcal{T}_h} |u|^2_{2,K} \sum_{K\in\mathcal{T}_h} |w_h|^2_{1,K} \right)^{\frac12} \le c\,h|u|_{2,\Omega}\|w_h\|_{V(h)}, \end{aligned}$$

leading to (3.37). □

Lemma 3.37 (Approximability). *There exists* c *such that*

$$\forall h,\ \forall u \in H^2(\Omega) \cap H_0^1(\Omega), \qquad \inf_{v_h \in P^1_{\mathrm{pt},h,0}} \|u - v_h\|_{V(h)} \le c\,h|u|_{2,\Omega}. \tag{3.38}$$

Proof. Use $P^1_{\mathrm{c},h,0} = P_{\mathrm{c},h} \cap H_0^1(\Omega) \subset P^1_{\mathrm{pt},h,0}$ and Corollary 1.109. □

Theorem 3.38 (Convergence). *Under the assumptions of Lemma* 3.36*, there exists* c *such that*

$$\forall h, \quad \|u - u_h\|_{V(h)} \le c\,h|u|_{2,\Omega}. \tag{3.39}$$

Proof. Direct consequence of Lemma 2.25 and the above results. □

Finally, an error estimate in the L^2-norm can be obtained by generalizing the Aubin–Nitsche Lemma to non-conforming approximation spaces.

Theorem 3.39 (L^2-estimate). *Along with the assumptions of Theorem* 3.38*, assume that the Laplacian has smoothing properties in* Ω*. Then, there exists* c *such that*

$$\forall h, \quad \|u - u_h\|_{0,\Omega} \le c\,h|u - u_h|_{h,1,\Omega}. \tag{3.40}$$

Proof. See [Bra97, p. 108]. □

3.2.4 Discontinuous Galerkin (DG) Approximation

In the previous section, we have investigated a first example of non-conforming method to approximate second-order elliptic PDEs. Because the degrees of freedom in the finite element space were located at the faces of the mesh, the method can be viewed as a *face-centered approximation*. In this section, we continue the investigation of non-conforming methods for elliptic problems by analyzing *cell-centered approximations* in which the degrees of freedom in the finite element space are defined independently on each cell. In the literature, such methods are often termed Discontinuous Galerkin (DG) methods, and this terminology will be employed henceforth.

For the sake of simplicity, we restrict ourselves to the approximation of the Laplacian with homogeneous Dirichlet conditions and data $f \in L^2(\Omega)$. As in the previous section, we assume that the domain Ω is a polyhedron in $\mathbb{R}^d$ in which the Laplacian has smoothing properties; hence, the exact solution u is in $H^2(\Omega)$. The material presented below is adapted from [ArB01].

Mixed formulation. We recast the problem in the form of a mixed system of first-order PDEs

$$\sigma = \nabla u, \quad -\nabla\cdot\sigma = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega. \tag{3.41}$$

From a physical viewpoint, the auxiliary unknown σ plays the role of a flux, and the PDE $-\nabla\cdot\sigma = f$ expresses a conservation property. The unknown u is called the *primal variable*. Multiplying the first and second equations in (3.41) by test functions τ and v, respectively, and integrating formally over a subset K of Ω yields the weak formulation

$$\begin{cases} \int_K \sigma\cdot\tau = -\int_K u\,\nabla\cdot\tau + \int_{\partial K} u\,\tau\cdot n_K, \\ \int_K \sigma\cdot\nabla v = \int_K fv + \int_{\partial K} v\,\sigma\cdot n_K, \end{cases} \tag{3.42}$$

where n_K is the outward normal to ∂K.

Let $\{\mathcal{T}_h\}_{h>0}$ be a shape-regular family of simplicial meshes of the domain Ω, and for $k \geq 1$, consider the finite element spaces

$$\begin{cases} V_h = \{v \in L^1(\Omega);\, \forall K \in \mathcal{T}_h,\, v_{|K} \in \mathbb{P}_k\}, \\ \Sigma_h = \{\tau \in [L^1(\Omega)]^d;\, \forall K \in \mathcal{T}_h,\, \tau_{|K} \in [\mathbb{P}_k]^d\}. \end{cases}$$

Note that V_h coincides with the space $P^k_{\mathrm{td},h}$ introduced in §1.4.3. For $v \in V_h$ and $\tau \in \Sigma_h$, let $\nabla_h v$ and $\nabla_h\cdot\tau$ be the functions whose restriction to each element $K \in \mathcal{T}_h$ is equal to ∇v and $\nabla\cdot\tau$, respectively. Following [CoS98], a discrete mixed formulation is derived by summing (3.42) over the mesh elements:

$$\begin{cases} \text{Seek } u_h \in V_h \text{ and } \sigma_h \in \Sigma_h \text{ such that} \\ \int_\Omega \sigma_h\cdot\tau = -\int_\Omega u_h\,\nabla_h\cdot\tau + \sum_{K\in\mathcal{T}_h}\int_{\partial K}\phi_u\,\tau\cdot n_K, \quad \forall\tau \in \Sigma_h, \\ \int_\Omega \sigma_h\cdot\nabla_h v = \int_\Omega fv + \sum_{K\in\mathcal{T}_h}\int_{\partial K} v\,\phi_\sigma\cdot n_K, \quad \forall v \in V_h, \end{cases} \tag{3.43}$$

where the *numerical fluxes* ϕ_u and ϕ_σ are approximations to the double-valued traces at the mesh interfaces of u_h and σ_h, respectively. The numerical fluxes need not be single-valued at the mesh interfaces.

To specify the numerical fluxes, we introduce an appropriate functional setting. For an integer $l \geq 1$, let $H^l(\mathcal{T}_h)$ be the space of functions on Ω whose restriction to each element $K \in \mathcal{T}_h$ belongs to $H^l(K)$. Recall that $\mathcal{F}_h^{\mathrm{i}}$ denotes the set of interior faces, $\mathcal{F}_h^\partial$ the set of boundary faces, and $\mathcal{F}_h = \mathcal{F}_h^{\mathrm{i}} \cup \mathcal{F}_h^\partial$. The traces on element boundaries of functions in $H^1(\mathcal{T}_h)$ belong to a space denoted by $T(\mathcal{F}_h)$. Functions in $T(\mathcal{F}_h)$ are double-valued on $\mathcal{F}_h^{\mathrm{i}}$ and single-valued on $\mathcal{F}_h^\partial$. Denote by $L^2(\mathcal{F}_h)$ the space of single-valued functions on $\mathcal{F}_h$ whose restriction to each face $F \in \mathcal{F}_h$ is in $L^2(F)$.

Using the above notation, the numerical fluxes are chosen to be linear functions

$$\phi_u : H^1(\mathcal{T}_h) \longrightarrow T(\mathcal{F}_h), \qquad \phi_\sigma : H^2(\mathcal{T}_h) \times [H^1(\mathcal{T}_h)]^d \longrightarrow [T(\mathcal{F}_h)]^d.$$

In the present setting, ϕ_u depends only on u_h, while ϕ_σ depends on both u_h and σ_h; other settings can be considered as well.

Two properties of the numerical fluxes are important in the analysis of DG methods: consistency and conservativity.

Definition 3.40 (Consistency). *The numerical fluxes ϕ_u and ϕ_σ are said to be* consistent *if for any smooth function $v \in H^2(\Omega) \cap H_0^1(\Omega)$,*

$$\phi_u(v) = v_{|\mathcal{F}_h} \qquad \text{and} \qquad \phi_\sigma(v, \nabla v) = \nabla v_{|\mathcal{F}_h}.$$

Proposition 3.41. *If the numerical fluxes ϕ_u and ϕ_σ are consistent, the exact solution u and its gradient ∇u satisfy* (3.43).

Proof. Straightforward verification. □

Definition 3.42 (Conservativity). *The numerical fluxes ϕ_u and ϕ_σ are said to be* conservative *if they are single-valued on $\mathcal{F}_h$.*

Proposition 3.43. *Assume that the numerical fluxes are conservative. Let ω be the union of any collection of elements. Then, if (u_h, σ_h) solves* (3.43),

$$\int_\omega f + \int_{\partial\omega} \phi_\sigma(u_h, \sigma_h) \cdot n_\omega = 0,$$

where n_ω is the outward normal to $\partial\omega$.

Proof. Take v to be the characteristic function of ω. □

Primal formulation. A primal formulation is a discrete problem in which u_h is the only unknown.

To derive a primal formulation, the discrete unknown σ_h must be eliminated through a *flux reconstruction formula*, that is, a formula expressing the discrete flux σ_h in terms of the discrete primal variable u_h only. It is convenient to define averages and jumps across faces. Let F be an interior face shared by elements K_1 and K_2, and let n_1 and n_2 be the normal vectors to F pointing toward the exterior of K_1 and K_2, respectively. For $v \in V_h$, setting $v_i = v_{|F \cap K_i}$, $i = 1, 2$, define the average $\{\cdot\}$ and jump $[\![\cdot]\!]$ operators as

$$\{v\} = \tfrac{1}{2}(v_1 + v_2) \quad \text{and} \quad [\![v]\!] = v_1 n_1 + v_2 n_2 \quad \text{on each } F \in \mathcal{F}_h^{\mathrm{i}}.$$

Using a similar notation for $\tau \in \Sigma_h$, set

$$\{\tau\} = \tfrac{1}{2}(\tau_1 + \tau_2) \quad \text{and} \quad [\![\tau]\!] = \tau_1 \cdot n_1 + \tau_2 \cdot n_2 \quad \text{on each } F \in \mathcal{F}_h^{\mathrm{i}}.$$

Note that the jump of a scalar-valued function is vector-valued, and *vice versa* (to alleviate the notation, we write $[\![\tau]\!]$ instead of $[\![\tau \cdot n]\!]$). For $F \in \mathcal{F}_h^\partial$, set $[\![v]\!] = vn$ and $\{\tau\} = \tau$ where n is the outward normal. Owing to the identity

$$\int_\Omega \nabla_h \cdot \tau \, v + \int_\Omega \tau \cdot \nabla_h v = \sum_{K \in \mathcal{T}_h} \int_{\partial K} v \, \tau \cdot n_K = \int_{\mathcal{F}_h} [\![v]\!] \cdot \{\tau\} + \int_{\mathcal{F}_h^{\mathrm{i}}} \{v\} [\![\tau]\!], \quad (3.44)$$

holding for all $v \in V_h$ and $\tau \in \Sigma_h$, (3.43) is recast into the form

$$\begin{cases} \int_\Omega \sigma_h \cdot \tau = -\int_\Omega u_h \nabla_h \cdot \tau + \int_{\mathcal{F}_h} [\![\phi_u(u_h)]\!] \cdot \{\tau\} + \int_{\mathcal{F}_h^{\mathrm{i}}} \{\phi_u(u_h)\} [\![\tau]\!], \\ \int_\Omega \sigma_h \cdot \nabla_h v - \int_{\mathcal{F}_h} \{\phi_\sigma(u_h, \sigma_h)\} \cdot [\![v]\!] - \int_{\mathcal{F}_h^{\mathrm{i}}} [\![\phi_\sigma(u_h, \sigma_h)]\!] \{v\} = \int_\Omega fv, \end{cases} \tag{3.45}$$

for all $\tau \in \Sigma_h$ and $v \in V_h$. Using (3.44) to eliminate the term $\int_\Omega u_h \nabla_h \cdot \tau$ in the first equation of (3.45) yields

$$\int_\Omega \sigma_h \cdot \tau = \int_\Omega \nabla_h u_h \cdot \tau + \int_{\mathcal{F}_h} [\![\phi_u(u_h) - u_h]\!] \cdot \{\tau\} + \int_{\mathcal{F}_h^{\mathrm{i}}} \{\phi_u(u_h) - u_h\} [\![\tau]\!]. \tag{3.46}$$

Introduce the lifting operators $l_1 : L^2(\mathcal{F}_h^{\mathrm{i}}) \to \Sigma_h$ and $l_2 : [L^2(\mathcal{F}_h)]^d \to \Sigma_h$ such that, for $q \in L^2(\mathcal{F}_h^{\mathrm{i}})$ and $\rho \in [L^2(\mathcal{F}_h)]^d$,

$$\forall \tau \in \Sigma_h, \quad \int_\Omega l_1(q) \cdot \tau = -\int_{\mathcal{F}_h^{\mathrm{i}}} q [\![\tau]\!], \quad \int_\Omega l_2(\rho) \cdot \tau = -\int_{\mathcal{F}_h} \rho \cdot \{\tau\}. \tag{3.47}$$

These lifting operators involve local L^2-projections. For instance, for $F \in \mathcal{F}_h$, define the operator $l_F : [L^1(F)]^d \to \Sigma_h$ such that, for $\rho \in [L^1(F)]^d$,

$$\forall \tau \in \Sigma_h, \quad \int_\Omega l_F(\rho) \cdot \tau = -\int_F \rho \cdot \{\tau\}.$$

Clearly, the support of $l_F(\rho)$ consists of the one or two simplices sharing F as a face. For $\rho \in [L^2(\mathcal{F}_h)]^d$, it is clear that $l_2(\rho) = \sum_{F \in \mathcal{F}_h} l_F(\rho)$. A similar construction is possible for the lifting operator l_1.

Recalling that $\nabla_h V_h \subset \Sigma_h$ and using the above lifting operators, we deduce from (3.46) the flux reconstruction formula

$$\sigma_h = \nabla_h u_h - l_1(\{\phi_u(u_h) - u_h\}) - l_2([\![\phi_u(u_h) - u_h]\!]). \tag{3.48}$$

Taking now $\tau = \nabla_h v$ in (3.46), the second equation in (3.45) yields $a_h(u_h, v) = \int_\Omega fv$, where

$$\begin{aligned} a_h(u_h, v) = & \int_\Omega \nabla_h u_h \cdot \nabla_h v \\ & + \int_{\mathcal{F}_h} [\![\phi_u(u_h) - u_h]\!] \cdot \{\nabla_h v\} - \{\phi_\sigma(u_h, \sigma_h)\} \cdot [\![v]\!] \\ & + \int_{\mathcal{F}_h^{\mathrm{i}}} \{\phi_u(u_h) - u_h\} [\![\nabla_h v]\!] - [\![\phi_\sigma(u_h, \sigma_h)]\!] \{v\}, \end{aligned} \tag{3.49}$$

with σ_h evaluated from (3.48). The bilinear form a_h is defined on $H^2(\mathcal{T}_h) \times H^2(\mathcal{T}_h)$. The primal formulation is thus:

$$\begin{cases} \text{Seek } u_h \in V_h \text{ such that} \\ a_h(u_h, v) = \int_\Omega fv, \quad \forall v \in V_h. \end{cases} \tag{3.50}$$

Clearly, if $(u_h, \sigma_h) \in V_h \times \Sigma_h$ solves (3.45), then u_h solves (3.50) provided the flux σ_h is reconstructed using (3.48).

Remark 3.44. If the fluxes are conservative, (3.49) simplifies into

$$a_h(u_h, v) = \int_\Omega \nabla_h u_h \cdot \nabla_h v - \int_{\mathcal{F}_h} [\![u_h]\!] \cdot \{\nabla_h v\} + \{\phi_\sigma(u_h, \sigma_h)\} \cdot [\![v]\!] \\ + \int_{\mathcal{F}_h^{\mathrm{i}}} (\phi_u(u_h) - \{u_h\}) [\![\nabla_h v]\!]. \qquad \square$$

Error analysis. To estimate the error induced by the approximate problem (3.50), it is convenient to introduce the space $V(h) = V_h + H^2(\Omega) \cap H_0^1(\Omega)$. For $v \in V(h)$, set

$$|v|_{h,1,\Omega}^2 = \sum_{K \in \mathcal{T}_h} |v|_{1,K}^2, \qquad |v|_{\mathrm{j}}^2 = \sum_{F \in \mathcal{F}_h} \|l_F([\![v]\!])\|_{0,\Omega}^2,$$

and let

$$\|v\|_{V(h)}^2 = |v|_{h,1,\Omega}^2 + |v|_{\mathrm{j}}^2 + \sum_{K \in \mathcal{T}_h} h_K^2 |v|_{2,K}^2. \tag{3.51}$$

This choice will appear more clearly in the examples presented below.

Lemma 3.45. *If Ω has smoothing properties, there exists c, independent of h, such that*

$$\forall v \in V(h), \quad c\,\|v\|_{0,\Omega} \le |v|_{h,1,\Omega} + |v|_{\mathrm{j}}.$$

Proof. (1) Using inverse inequalities, one can prove that there exist positive constants c_1 and c_2 such that

$$\forall \rho \in [\mathbb{P}_k(F)]^d, \quad c_1 \|l_F(\rho)\|_{0,\Omega}^2 \le h_F^{-1} \|\rho\|_{0,F}^2 \le c_2 \|l_F(\rho)\|_{0,\Omega}^2.$$

These inequalities can be applied to $\rho = [\![v]\!]$ for $v \in V(h)$, yielding

$$\forall v \in V(h), \quad c_1 |v|_{\mathrm{j}}^2 \le \sum_{F \in \mathcal{F}_h} h_F^{-1} \|[\![v]\!]\|_{0,F}^2 \le c_2 |v|_{\mathrm{j}}^2. \tag{3.52}$$

(2) Let $v \in V(h)$ and let $\psi \in H_0^1(\Omega)$ solve $-\Delta\psi = v$. Since Ω has smoothing properties, there is $c > 0$ such that $\|\psi\|_{2,\Omega} \le c\,\|v\|_{0,\Omega}$. Then,

$$\|v\|_{0,\Omega}^2 = -\textstyle\int_\Omega v \Delta\psi = \int_\Omega \nabla\psi \cdot \nabla_h v - \int_{\mathcal{F}_h} \nabla\psi \cdot [\![v]\!] \\ \le c\,|v|_{1,h,\Omega} \|v\|_{0,\Omega} + \left(\sum_{F \in \mathcal{F}_h} h_F^{-1} \|[\![v]\!]\|_{0,F}^2 \right)^{\frac{1}{2}} \left(\sum_{F \in \mathcal{F}_h} h_F |\psi|_{1,F}^2 \right)^{\frac{1}{2}}.$$

Using a trace theorem and a scaling argument yields

$$h_F |\psi|_{1,F}^2 \le c \left(|\psi|_{1,K}^2 + h_F^2 |\psi|_{2,K}^2 \right) \le c' \|\psi\|_{2,K}^2. \tag{3.53}$$

Hence,

$$\|v\|_{0,\Omega}^2 \le c_1 |v|_{1,h,\Omega} \|v\|_{0,\Omega} + c_2 |v|_j \|v\|_{0,\Omega},$$

and this completes the proof. $\square$

Remark 3.46. Lemma 3.45 is a discrete Poincaré-type inequality. □

Proposition 3.47 (Well-posedness). *Assume that the bilinear form a_h defined in* (3.49) *satisfies the following properties:*

(i) *Uniform boundedness on $V(h)$: there exists $c_b > 0$, independent of h, such that*

$$\forall v, w \in V(h), \quad a_h(w, v) \leq c_b \|w\|_{V(h)} \|v\|_{V(h)}. \tag{3.54}$$

(ii) *Coercivity on V_h: there exists $c_s > 0$, independent of h, such that*

$$\forall v \in V_h, \quad a_h(v, v) \geq c_s \|v\|_{V(h)}^2. \tag{3.55}$$

Then, problem (3.50) *is well-posed.*

Proof. Direct consequence of the Lax–Milgram Lemma. □

Proposition 3.48 (Consistency). *Assume that the numerical fluxes ϕ_u and ϕ_σ are consistent. Then, the exact solution u satisfies*

$$\forall v \in V_h, \quad a_h(u, v) = \int_\Omega f v.$$

Proof. Since $u \in H^2(\Omega)$, taking $\tau = \nabla_h u$ in (3.44) yields, for all $v \in V_h$,

$$\int_\Omega \nabla_h u \cdot \nabla_h v = -\int_\Omega \Delta u \, v + \int_{\mathcal{F}_h} [\![v]\!] \cdot \{\nabla_h u\} + \int_{\mathcal{F}_h^{\mathrm{i}}} \{v\} [\![\nabla_h u]\!].$$

Since $\{u\} = u$, $[\![u]\!] = 0$, $\{\nabla_h u\} = \nabla u$, $[\![\nabla_h u]\!] = 0$, and $-\Delta u = f$,

$$\begin{aligned} a_h(u, v) = & \int_\Omega f v + \int_{\mathcal{F}_h} [\![\phi_u(u)]\!] \cdot \{\nabla_h v\} + (\nabla u - \{\phi_\sigma(u, \sigma_h(u))\}) \cdot [\![v]\!] \\ & + \int_{\mathcal{F}_h^{\mathrm{i}}} \{\phi_u(u) - u\} [\![\nabla_h v]\!] - [\![\phi_\sigma(u, \sigma_h(u))]\!] \{v\}. \end{aligned}$$

Owing to the consistency of the numerical flux ϕ_u, $\phi_u(u) = u$. Moreover, the reconstruction formula (3.48) implies $\sigma_h(u) = \nabla u$. Since the numerical flux ϕ_σ is also consistent, $\{\phi_\sigma(u, \sigma_h(u))\} = \nabla u$ and $[\![\phi_\sigma(u, \sigma_h(u))]\!] = 0$. Therefore, all the face integrals vanish. □

Lemma 3.49 (Approximability). *There exists c such that, for all $1 \leq s \leq k+1$,*

$$\forall h, \ \forall u \in H^s(\Omega) \cap H_0^1(\Omega), \quad \inf_{v \in V_h} \|u - v\|_{V(h)} \leq c \, h^{s-1} |u|_{s,\Omega}.$$

Proof. Let $1 \leq s \leq k+1$. Since V_h contains the H^1-conforming Scott–Zhang interpolant $\mathcal{SZ}_h u$ of u and since the face jumps of $u - \mathcal{SZ}_h u$ vanish,

$$\|u - \mathcal{SZ}_h u\|_{V(h)}^2 = |u - \mathcal{SZ}_h u|_{1,\Omega}^2 + \sum_{K \in \mathcal{T}_h} h_K^2 |u - \mathcal{SZ}_h u|_{2,K}^2 \leq c \, h^{2(s-1)} |u|_{s,\Omega}^2,$$

the last inequality being a direct consequence of Lemma 1.130. □

Theorem 3.50 (Convergence). *Let u be the solution to the homogeneous Dirichlet problem with data $f \in L^2(\Omega)$. Assume that the Laplacian has smoothing properties in Ω and that $u \in H^s(\Omega)$ for some $s \in \{2, \dots, k+1\}$. Let u_h be the solution to* (3.50). *Along with hypotheses* (i)–(ii) *of Proposition* 3.47, *assume that the numerical fluxes ϕ_u and ϕ_σ are consistent. Then, there exists c such that*

$$\forall h, \quad \|u - u_h\|_{V(h)} \le c\, h^{s-1} |u|_{s,\Omega}. \tag{3.56}$$

Proof. Direct consequence of Lemma 2.25 and the above results. □

An L^2-norm error estimate can be obtained using duality techniques.

Definition 3.51 (Adjoint-consistency). *The bilinear form a_h is said to be* adjoint-consistent *if, for all $w \in H^2(\Omega) \cap H_0^1(\Omega)$,*

$$\forall v \in V(h), \quad a_h(v, w) = -\int_\Omega \Delta w\, v. \tag{3.57}$$

Lemma 3.52. *Assume that the numerical fluxes ϕ_u and ϕ_σ are conservative. Then, the bilinear form a_h is adjoint-consistent.*

Proof. Let $w \in H^2(\Omega) \cap H_0^1(\Omega)$ and let $v \in V(h)$. Note that $[\![w]\!] = 0$, $[\![\nabla_h w]\!] = 0$, and $\{\nabla_h w\} = \nabla w$. Using (3.44) yields

$$\int_\Omega \nabla_h v \cdot \nabla_h w = -\int_\Omega \Delta w\, v + \int_{\mathcal{F}_h} [\![v]\!] \cdot \nabla w.$$

Since w is smooth, Remark 3.44 implies $a_h(v, w) = \int_\Omega \nabla_h v \cdot \nabla_h w - \int_{\mathcal{F}_h} [\![v]\!] \cdot \nabla w$. The conclusion follows readily. □

Theorem 3.53 (L^2-convergence). *Under the hypotheses of Theorem* 3.50, *assuming that the numerical fluxes ϕ_u and ϕ_σ are conservative, there exists c such that*

$$\forall h, \quad \|u - u_h\|_{0,\Omega} \le c\, h^s |u|_{s,\Omega}. \tag{3.58}$$

Proof. Let $\psi \in H_0^1(\Omega)$ be such that $-\Delta \psi = u - u_h$. Since the Laplacian has smoothing properties in Ω, $|\psi|_{2,\Omega} \le c\, \|u - u_h\|_{0,\Omega}$. Furthermore, since the approximate fluxes ϕ_u and ϕ_σ are conservative, Lemma 3.52 implies

$$\forall v \in V(h), \quad a_h(v, \psi) = \int_\Omega (u - u_h) v.$$

Since $u - u_h \in V(h)$ and the numerical fluxes are consistent,

$$\begin{aligned}\|u - u_h\|_{0,\Omega}^2 &= a_h(u - u_h, \psi) = a_h(u - u_h, \psi - \mathcal{SZ}_h \psi) \\ &\le c_b \|u - u_h\|_{V(h)} \|\psi - \mathcal{SZ}_h \psi\|_{V(h)} \le c\, h |\psi|_{2,\Omega} \|u - u_h\|_{V(h)},\end{aligned}$$

where $\mathcal{SZ}_h \psi$ is the Scott–Zhang interpolant of ψ. Conclude using (3.56). □

Example 1 (LDG). The so-called Local Discontinuous Galerkin (LDG) method has been introduced by Cockburn and Shu in 1998 [CoS98] to approximate time-dependent convection–diffusion problems. Written within the above framework, it consists of taking the numerical fluxes

$$\phi_u(u_h) = \begin{cases} \{u_h\} - \beta\cdot[\![u_h]\!] & \text{on } \mathcal{F}_h^{\mathrm{i}}, \\ 0 & \text{on } \mathcal{F}_h^{\partial}, \end{cases} \tag{3.59}$$

and

$$\phi_\sigma(u_h, \sigma_h) = \begin{cases} \{\sigma_h\} + \beta\cdot[\![\sigma_h]\!] - \eta_F h_F^{-1}[\![u_h]\!] & \text{on } \mathcal{F}_h^{\mathrm{i}}, \\ \{\sigma_h\} - \eta_F h_F^{-1}[\![u_h]\!] & \text{on } \mathcal{F}_h^{\partial}. \end{cases} \tag{3.60}$$

Here, $\beta \in [L^\infty(\mathcal{F}_h^{\mathrm{i}})]^d$ is a vector-valued function that is constant on each interior face, η_F is a given positive parameter on the face F, and h_F denotes the diameter of F. A straightforward calculation yields the following:

Proposition 3.54. *The numerical fluxes ϕ_u and ϕ_σ defined by* (3.59)–(3.60) *are consistent and conservative.*

In the LDG method, the flux reconstruction formula (3.48) takes the form

$$\sigma_h = \nabla_h u_h + l_1(\beta\cdot[\![u_h]\!]) + l_2([\![u_h]\!]),$$

and the bilinear form a_h is given by

$$\begin{aligned} a_h(u_h, v) = &\int_\Omega \nabla_h u_h\cdot\nabla_h v - \int_{\mathcal{F}_h} [\![u_h]\!]\cdot\{\nabla_h v\} + \{\nabla_h u_h\}[\![v]\!] \\ &+ \int_{\mathcal{F}_h} \eta_F h_F^{-1}[\![u_h]\!]\,[\![v]\!] + \int_{\mathcal{F}_h^{\mathrm{i}}} \beta\cdot[\![u_h]\!]\,[\![v]\!] + [\![\nabla_h u_h]\!]\,\beta\cdot[\![v]\!] \\ &+ \int_\Omega \big(l_1(\beta\cdot[\![u_h]\!]) + l_2([\![u_h]\!])\big)\cdot\big(l_1(\beta\cdot[\![v]\!]) + l_2([\![v]\!])\big). \end{aligned} \tag{3.61}$$

Proposition 3.55. *The bilinear form a_h defined by* (3.61) *is continuous on $V(h)$ and, provided* $\inf_F \eta_F$ *is large enough, it is also coercive on V_h.*

Proof. The proof is only sketched; see [ArB01] and the references therein.
(i) To prove continuity, i.e., property (3.54), the various terms appearing in the right-hand side of (3.61) must be bounded. Let $w, v \in V(h)$. First, it is clear that $\int_\Omega \nabla_h w\cdot\nabla_h v \le |w|_{h,1,\Omega}|v|_{h,1,\Omega}$. Owing to (3.52), $\int_{\mathcal{F}_h} \eta_F h_F^{-1}[\![u_h]\!]\,[\![v]\!] \le c_3|w|_{\mathrm{j}}|v|_{\mathrm{j}}$ with $c_3 = c_2 \sup_F \eta_F$. Next, for $w \in H^2(K)$ and a face F of K, (3.53) implies

$$\|\nabla w\cdot n\|_{0,F}^2 \le c_4\big(h_F^{-1}|w|_{1,K}^2 + h_F|w|_{2,K}^2\big).$$

This in turn implies

$$\begin{aligned} \int_{\mathcal{F}_h} \{\nabla_h w\}\cdot[\![v]\!] &\le c_5 \left(\sum_{K\in\mathcal{T}_h} |w|_{1,K}^2 + h_K^2|w|_{2,K}^2\right)^{\frac{1}{2}} \left(\sum_{F\in\mathcal{F}_h} h_F^{-1}\|[\![v]\!]\|_{0,F}^2\right)^{\frac{1}{2}} \\ &\le c_5\|w\|_{V(h)}|v|_{\mathrm{j}}. \end{aligned}$$

The remaining face integrals in (3.61) are bounded similarly. Finally, one can readily show that

$$\forall v \in V(h), \quad \|l_1(\beta\cdot[\![v]\!])\|_{0,\Omega} \leq c_6 \|\beta\|^{\frac{1}{2}}_{L^\infty(\mathcal{F}_h^{\mathrm{i}})} |v|_{\mathrm{j}},$$

and

$$\forall v \in V(h), \quad \|l_2([\![v]\!])\|_{0,\Omega} \leq c_7 |v|_{\mathrm{j}}.$$

Using the above estimates, one easily bounds the second integral over Ω in the right-hand side of (3.61).
(ii) Let us prove the coercivity of a_h, i.e., property (3.55). Consider $v \in V_h$. It is clear that

$$a_h(v,v) = |v|^2_{h,1,\Omega} + \int_{\mathcal{F}_h} \eta_F h_F^{-1} [\![v]\!]^2 + b(v,v),$$

where the bilinear form b gathers all the remaining terms. It follows from the first part of the proof that

$$\int_{\mathcal{F}_h} \eta_F h_F^{-1} [\![v]\!]^2 \geq c_8 \left(\inf_F \eta_F\right) |v|^2_{\mathrm{j}} \qquad \text{and} \qquad b(v,v) \leq c_9 \|v\|_{V(h)} |v|_{\mathrm{j}}.$$

Therefore,

$$a_h(v,v) \geq |v|^2_{h,1,\Omega} + c_8 \left(\inf_F \eta_F\right) |v|^2_{\mathrm{j}} - c_9 \|v\|_{V(h)} |v|_{\mathrm{j}},$$

and the last term in the right-hand side can be lower bounded in the form $-c_9\|v\|_{V(h)}|v|_{\mathrm{j}} \geq -\epsilon\|v\|^2_{V(h)} - \frac{c_9^2}{4\epsilon}|v|^2_{\mathrm{j}}$ for any positive ϵ. Moreover, using an inverse inequality on V_h yields

$$\|v\|^2_{V(h)} \leq c_{10}(|v|^2_{h,1,\Omega} + |v|^2_{\mathrm{j}}).$$

Coercivity follows by taking ϵ small enough and $\inf_F \eta_F$ large enough. □

The above results show that the LDG method approximates the exact solution to $\mathcal{O}(h^k)$ in the H^1-norm and to $\mathcal{O}(h^{k+1})$ in the L^2-norm.

Example 2 (NIPG). The so-called Non-symmetric Interior Penalty Galerkin (NIPG) method has been derived in [OdB98, BaO99] and further investigated in [RiW99]. Written within the above framework, it consists of taking the numerical fluxes

$$\phi_u(u_h) = \begin{cases} \{u_h\} + n_K\cdot[\![u_h]\!] & \text{on } \mathcal{F}_h^{\mathrm{i}}, \\ 0 & \text{on } \mathcal{F}_h^{\partial}, \end{cases} \tag{3.62}$$

and

$$\phi_\sigma(u_h,\sigma_h) = \{\nabla_h u_h\} - \eta_F h_F^{-1} [\![u_h]\!] \quad \text{on } \mathcal{F}_h. \tag{3.63}$$

Note that ϕ_u is not single-valued on $\mathcal{F}_h^{\mathrm{i}}$. A straightforward calculation yields the following:

Proposition 3.56. *The numerical fluxes ϕ_u and ϕ_σ given by* (3.62)–(3.63) *are consistent, but not conservative.*

In the NIPG method, the bilinear form a_h is given by

$$\begin{aligned} a_h(u_h, v) = \int_\Omega \nabla_h u_h \cdot \nabla_h v + \int_{\mathcal{F}_h} \eta_F h_F^{-1} [\![u_h]\!]\, [\![v]\!] \\ + \int_{\mathcal{F}_h} [\![u_h]\!] \cdot \{\nabla_h v\} - \{\nabla_h u_h\} [\![v]\!]. \end{aligned} \tag{3.64}$$

Proposition 3.57. *The bilinear form a_h given by* (3.64) *is continuous on $V(h)$ and coercive on V_h.*

Proof. Similar to that of Proposition 3.55. □

The above results show that the NIPG method approximates the exact solution to $\mathcal{O}(h^k)$ in the H^1-norm. However, because of the lack of conservativity in the numerical fluxes, an improved error estimate in the L^2-norm cannot be derived in general.

Remark 3.58.
(i) Because of the skew-symmetric form of the face integrals in (3.64), $\inf_F \eta_F$ needs not be large to ensure the coercivity of a_h. However, skew-symmetry is at the origin of the lack of adjoint-consistency, thus preventing optimal convergence order in the L^2-norm.
(ii) For a face $F \in \mathcal{F}_h$, one can choose the penalty parameter η_F to be proportional to a negative power of h_F, leading to the so-called superpenalty procedure. It is then possible to recover optimal convergence order for the error in the L^2-norm. The NIPG method with superpenalty is analyzed in [RiW99]. □

3.2.5 Numerical illustrations

This section presents two examples of finite element approximations to elliptic PDEs. The purpose of the first example is to illustrate the link between the convergence order of the finite element approximation and the regularity of the exact solution. The purpose of the second example is to illustrate qualitatively the behavior of the solution of advection–diffusion equations depending on whether advection effects dominate or not.

Convergence tests. Consider the Laplace equation in the domain $\Omega =]0,1[\times]0,1[$ and a positive parameter α. Choose the right-hand side f and the non-homogeneous Dirichlet conditions so that the exact solution is $u(x_1, x_2) = (x_1^2 + x_2^2)^{\frac{\alpha}{2}}$. Note that $u \in H^1(\Omega)$ if $0 < \alpha \le 1$, $u \in H^2(\Omega)$ if $1 < \alpha \le 2$, and $u \in H^3(\Omega)$ if $2 < \alpha \le 3$. In the numerical experiments, we consider the values $\alpha = 0.25$, 1.25, and 2.25. A H^1-conforming Lagrange

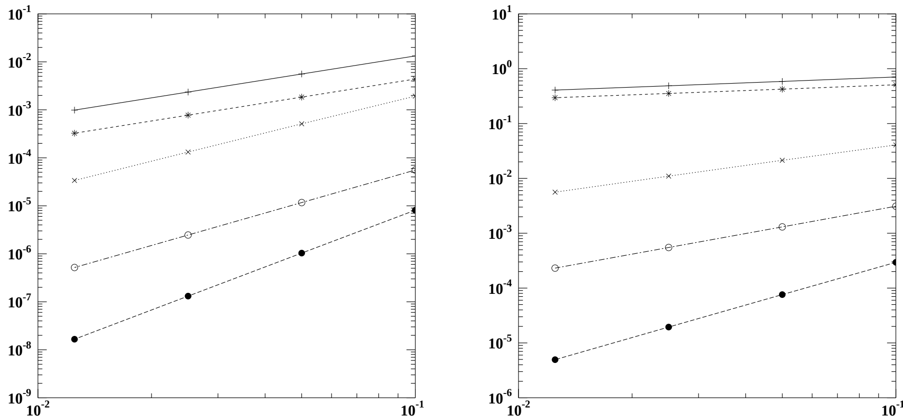


Fig. 3.1. Errors in the L^2-norm (left) and H^1-norm (right) as a function of the mesh step size h: $\mathbb{P}_1$ finite element and $\alpha = 0.25$ ($+$); $\mathbb{P}_2$ finite element and $\alpha = 0.25$ ($*$); $\mathbb{P}_1$ finite element and $\alpha = 1.25$ ($\times$); $\mathbb{P}_2$ finite element and $\alpha = 1.25$ ($\circ$); $\mathbb{P}_2$ finite element and $\alpha = 2.25$ ($\bullet$).

finite element approximation of degree $k = 1$ or 2 is implemented. The triangulation of Ω is uniform with vertices of the triangles given by (ih, jh), $0 \le i, j \le N + 1$, where $h = \frac{1}{N+1}$ and N is a given integer.

Figure 3.1 presents the error in the L^2- and H^1-norms as a function of h. Results are presented in a log-log scale so that the slopes indicate orders of convergence. For $\alpha = 0.25$ and $k = 1$, the error converges "slowly" to zero as $h \to 0$, with a slope lower than 1 in the H^1-norm and lower than 2 in the L^2-norm. For $\alpha = 1.25$ and still $k = 1$, the slope is equal to 1 in the H^1-norm and to 2 in the L^2-norm. Moreover, using a higher-order method ($k = 2$) does not improve the convergence order. Finally, for $\alpha = 2.25$ and with a second-order finite element, the slopes in both the H^1-norm and the L^2-norm are one order higher than those obtained with the first-order finite element method, in agreement with theoretical predictions. As a conclusion, only when the exact solution is smooth enough does it pay off to use a high-order finite element method.

Advection–diffusion equation. Consider a two-dimensional flow through a heated pipe. The flow velocity is assumed to be known, and we want to evaluate the temperature u inside the pipe at steady-state. The temperature is governed by the advection–diffusion equation

$$\beta\cdot\nabla u - \epsilon\Delta u = 0. \tag{3.65}$$

The pipe is modeled by a rectangular domain Ω with sides numbered clockwise from 1 to 4 starting from the left-most side. The flow enters the pipe through $\partial\Omega_1$ and flows out through $\partial\Omega_3$ while the sides $\partial\Omega_2$ and $\partial\Omega_4$ are

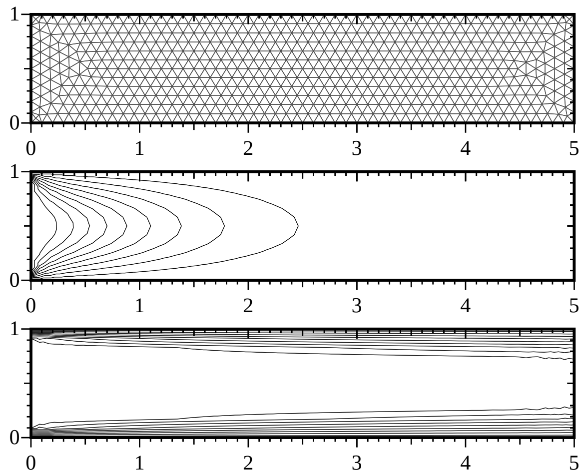


Fig. 3.2. Heat transfer problem through a two-dimensional pipe: computational mesh (top); temperature field for dominant diffusion (center); and temperature field for dominant advection (bottom).

solid boundaries. Spatial coordinates are denoted by (x_1, x_2) with the x_1-axis parallel to the pipe axis. Temperature boundary conditions are $u = 0$ on $\partial\Omega_1$ (cold upstream flow), $u = 1$ on $\partial\Omega_2$ and $\partial\Omega_4$ (heated boundaries), and $\partial_1 u = 0$ on $\partial\Omega_3$ (outflow condition). The flow velocity is taken to be $\beta = (4x_2(1 - x_2), 0)$. The solution to (3.65) is approximated on the mesh shown in the top panel of Figure 3.2 using continuous $\mathbb{P}_1$ finite elements. The central panel of Figure 3.2 presents isotherms for a diffusion-dominated case ($\epsilon = 10^{-1}$); the peak temperature is quickly reached on the symmetry axis. The bottom panel displays isotherms resulting from a moderate diffusion coefficient ($\epsilon = 10^{-3}$). Advection effects are dominant, i.e., the boundary layer in which the temperature undergoes significant variations remains localized near the top and bottom boundaries. If advection effects become even more dominant, the approximation method needs to be stabilized to avoid spurious oscillations in the solution profile; see Chapter 5.

3.3 Spectral Problems

This section contains a brief introduction to spectral problems and their approximation by finite element methods. Spectral problems occur when analyzing the response of buildings, vehicles, or aircrafts to vibrations. Henceforth, we restrict the presentation to a simple model problem: the Laplace operator with homogeneous Dirichlet conditions. Although this problem is somewhat simple, it is representative of a large class of engineering applications. As such, it models membrane and string vibrations.

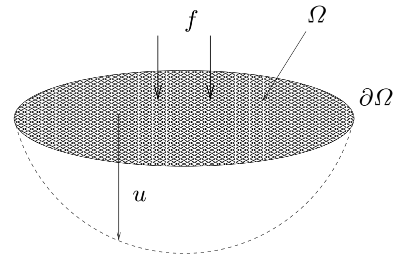


Fig. 3.3. Elastic deformation of a membrane: reference configuration Ω, externally applied load f, and equilibrium displacement u. The boundary $\partial\Omega$ of the membrane is kept fixed.

3.3.1 Modeling a vibrating membrane

Figure 3.3 presents an elastic homogeneous membrane. In the reference configuration, the membrane occupies the domain Ω in $\mathbb{R}^2$ and is tightened according to a two-dimensional stress tensor $\sigma \in \mathbb{R}^{2,2}$. For the sake of simplicity, we assume that σ is uniform and isotropic, i.e., $\sigma = \tau\mathcal{I}$ where τ is the membrane tension. Apply now a transverse load f and assume first that f is *time-independent*. If the strains in the membrane are sufficiently small, the equilibrium configuration is described by a transverse displacement which is a function $u : \Omega \to \mathbb{R}$ governed by the PDE

$$-\tau\Delta u = f \quad \text{in } \Omega. \tag{3.66}$$

We assume that the boundary of the membrane is kept fixed, yielding the homogeneous Dirichlet condition $u = 0$ on $\partial\Omega$.

Consider now the *time-dependent* load $f(x,t) = g(x)\cos(\omega t)$ for $(x,t) \in Q$, where $g : \Omega \to \mathbb{R}$ is a given function, ω a real parameter representing the angular velocity of the excitation, $Q = \Omega \times]0,T[$, and T a given time. Assuming again that the strains in the membrane remain sufficiently small, the (time-dependent) displacement $u : Q \to \mathbb{R}$ is governed by the PDE

$$\rho\partial_{tt}u - \tau\Delta u = g(x)\cos(\omega t) \quad \text{in } Q, \tag{3.67}$$

where ρ is the membrane density. Equation (3.67) is a *wave equation* with celerity $c = (\tau\rho^{-1})^{\frac{1}{2}}$. It has to be supplemented with initial and boundary conditions. The initial data comprises the initial value of the displacement $u_0(x)$ and its time-derivative $u_1(x)$, i.e., the initial membrane velocity. We assume that the membrane boundary is kept fixed at all times., i.e., we enforce a homogeneous Dirichlet boundary condition.

3.3.2 The spectral problem

Consider the *spectral problem*:

$$\begin{cases} \text{Seek } \psi \in H_0^1(\Omega),\ \psi \neq 0, \text{ and } \lambda \in \mathbb{R} \text{ such that} \\ -\Delta\psi = \lambda\psi, \end{cases}$$

for which a weak formulation is

$$\begin{cases} \text{Seek } \psi \in H_0^1(\Omega),\ \psi \neq 0, \text{ and } \lambda \in \mathbb{R} \text{ such that} \\ \int_\Omega \nabla\psi\cdot\nabla v = \lambda \int_\Omega \psi v, \quad \forall v \in H_0^1(\Omega). \end{cases} \tag{3.68}$$

Definition 3.59. *Let* $\{\lambda, \psi\}$ *be a solution to* (3.68)*. The real* λ *is called an* eigenvalue of the Laplacian *(with homogeneous Dirichlet conditions) and the function* ψ *an* eigenfunction*.*

Theorem 3.60 (Spectral decomposition). *Let* Ω *be a domain in* $\mathbb{R}^d$*. Then, the spectral problem* (3.68) *admits infinitely many solutions. These solutions form a sequence* $\{\lambda_n, \psi_n\}_{n>0}$ *such that:*

(i) $\{\lambda_n\}_{n>0}$ *is an increasing sequence of positive numbers, and* $\lambda_n \to \infty$*.*
(ii) $\{\psi_n\}_{n>0}$ *is an orthonormal Hilbert basis of* $L^2(\Omega)$*.*

Proof. This is a consequence of the fact that the injection $H_0^1(\Omega) \subset L^2(\Omega)$ is compact; see Theorem B.46 and [Yos80, p. 284] or [Bre91, p. 192]. □

Example 3.61. For $\Omega =]0, 1[$, the eigenvalues of the Laplacian are $\lambda_n = n^2\pi^2$ with corresponding eigenfunctions $\psi_n(x) = \sin(n\pi x)$. These functions become more and more oscillatory as n grows. □

The solution u to the wave equation (3.67) can be written as a series in terms of the Laplacian eigenfunctions. Indeed, set $\omega_n = (\lambda_n \tau \rho^{-1})^{\frac{1}{2}}$ and assume $\omega \neq \omega_n$. Denote by $g_n = \int_\Omega g\psi_n$ the coordinates of g relative to the orthonormal basis $\{\psi_n\}_{n>0}$ and by α_n and β_n the coordinates of the initial data u_0 and u_1, respectively. A straightforward calculation shows that for $\omega \neq \omega_n$,

$$u(x,t) = \sum_{n=1}^{\infty} \Bigg(\alpha_n \cos(\omega_n t) + \beta_n \sin(\omega_n t) + \frac{g_n}{\rho(\omega + \omega_n)} \frac{\sin\left(\frac{\omega-\omega_n}{2} t\right)}{\frac{\omega-\omega_n}{2}} \sin\left(\tfrac{\omega+\omega_n}{2} t\right) \Bigg) \psi_n(x).$$

As ω draws closer to one of the ω_n's, a resonance phenomenon occurs. In particular, when $\omega = \omega_n$, $u(x,t)$ grows linearly in time.

3.3.3 The Rayleigh quotient

Set $a(u, v) = (\nabla u, \nabla v)_{0,\Omega}$ for all u, v in $H_0^1(\Omega)$. This bilinear form is symmetric, continuous, and coercive on $H_0^1(\Omega)$. The *Rayleigh quotient* of a function $u \in H_0^1(\Omega)$, $u \neq 0$, is defined to be

$$R(u) = \frac{a(u,u)}{\|u\|_{0,\Omega}^2}.$$

Proposition 3.62. *Let λ_1 be the smallest eigenvalue of the spectral problem* (3.68) *and let ψ_1 be a corresponding eigenfunction. Then,*

$$\lambda_1 = R(\psi_1) = \inf_{v \in H_0^1(\Omega)} R(v).$$

Proof. Clearly, $\lambda_1 = R(\psi_1) \geq \inf_{v \in H_0^1(\Omega)} R(v)$. Furthermore, for $v \in H_0^1(\Omega)$, the identity $v = \sum_{n=1}^{\infty} v_n \psi_n$ yields

$$R(v) = \frac{\sum_{n=1}^{\infty} \lambda_n v_n^2}{\sum_{n=1}^{\infty} v_n^2} \geq \lambda_1. \qquad \square$$

Proposition 3.63. *Let λ_m be the m-th eigenvalue of problem* (3.68) *(eigenvalues are counted with their multiplicity and ordered increasingly). Let V_m denote the set of subspaces of $H_0^1(\Omega)$ having dimension m. Then,*

$$\lambda_m = \min_{E_m \in V_m} \max_{v \in E_m} R(v). \tag{3.69}$$

Proof. Let $E_m = \operatorname{span}\{\psi_1, \ldots, \psi_m\}$ be the space spanned by the m first eigenfunctions. For all $v = \sum_{n=1}^{m} v_n \psi_n$ in E_m,

$$R(v) = \frac{\sum_{n=1}^{m} \lambda_n v_n^2}{\sum_{n=1}^{m} v_n^2} \leq \lambda_m,$$

which yields

$$\lambda_m \geq \min_{E_m \in V_m} \max_{v \in E_m} R(v).$$

Consider now $E_m \in V_m$. A simple dimensional argument shows that there exists $v \neq 0$ in $E_m \cap E_{m-1}^{\perp}$. Since v can be written in the form $v = \sum_{n=m}^{\infty} v_n \psi_n$, it is clear that $R(v) \geq \lambda_m$. As a result, $\max_{v \in E_m} R(v) \geq \lambda_m$; hence,

$$\lambda_m \leq \min_{E_m \in V_m} \max_{v \in E_m} R(v). \qquad \square$$

3.3.4 H^1-conforming approximation

The spectral problem (3.68) can be solved analytically only in a limited number of remarkable cases when the domain Ω has a very simple shape. In the general case, eigenvalues and eigenfunctions must be approximated using, for instance, a finite element method.

Let $\{\mathcal{T}_h\}_{h>0}$ be a family of geometrically conforming meshes of Ω and let $\{V_h\}_{h>0}$ be the corresponding family of H^1-conforming approximation spaces. Denote by N the dimension of V_h. The approximate spectral problem we consider is the following:

$$\begin{cases} \text{Seek } \psi_h \in V_h,\ \psi_h \neq 0, \text{ and } \lambda_h \in \mathbb{R} \text{ such that} \\ \int_\Omega \nabla\psi_h \cdot \nabla v_h = \lambda_h \int_\Omega \psi_h v_h, \quad \forall v_h \in V_h. \end{cases} \tag{3.70}$$

Let $\{\varphi_1, \ldots, \varphi_N\}$ be a basis of V_h and let $\Psi_h \in \mathbb{R}^N$ be the coordinate vector of Ψ_h relative to this base. The approximate problem (3.70) is recast in the form:

$$\begin{cases} \text{Seek } \Psi_h \in \mathbb{R}^N, \, \Psi_h \neq 0, \text{ and } \lambda_h \in \mathbb{R} \text{ such that} \\ \mathcal{A}\Psi_h = \lambda_h \mathcal{M}\Psi_h, \end{cases} \tag{3.71}$$

where the *stiffness matrix* $\mathcal{A}$ and the *mass matrix* $\mathcal{M}$ have entries

$$\mathcal{A}_{ij} = \int_\Omega \nabla\varphi_i \cdot \nabla\varphi_j \qquad \text{and} \qquad \mathcal{M}_{ij} = \int_\Omega \varphi_i \varphi_j. \tag{3.72}$$

Because the matrix $\mathcal{M}$ is not the identity matrix, problem (3.71) is often called a generalized eigenvalue problem.

Proposition 3.64. *The matrices $\mathcal{A}$ and $\mathcal{M}$ defined in* (3.72) *are symmetric positive definite. Furthermore, the spectral problem* (3.71) *admits N (positive) eigenvalues (counted with their multiplicity).*

Proof. The symmetry and positive definiteness of the matrices $\mathcal{A}$ and $\mathcal{M}$ directly results from the fact that they are Gram matrices; see also Remark 2.20. Orthogonalizing the quadratic form associated with $\mathcal{A}$ with respect to the scalar product induced by $\mathcal{M}$ yields N positive reals $\{\lambda_{h1}, \ldots, \lambda_{hN}\}$ and a basis $\{\Psi_{h1}, \ldots, \Psi_{hN}\}$ of $\mathbb{R}^N$ such that, for $1 \leq i, j \leq N$,

$$(\Psi_{hi}, \mathcal{A}\Psi_{hj})_N = \lambda_{hi}\delta_{ij}, \qquad (\Psi_{hi}, \mathcal{M}\Psi_{hj})_N = \delta_{ij},$$

where $(\cdot,\cdot)_N$ denotes the Euclidean product in $\mathbb{R}^N$. As a result,

$$\mathcal{A}\Psi_{hi} = \lambda_{hi}\mathcal{M}\Psi_{hi}, \qquad 1 \leq i \leq N,$$

showing that the λ_{hi}'s are the eigenfunctions of the generalized eigenvalue problem (3.71) and that the Ψ_{hi}'s are the corresponding eigenvectors. □

3.3.5 Error analysis

Let $\{\psi_{h1}, \ldots, \psi_{hN}\}$ be an orthonormal basis of eigenvectors in V_h, i.e., $(\psi_{hi}, \psi_{hj})_{0,\Omega} = \delta_{ij}$ for $1 \leq i, j \leq N$, and assume that the enumeration of these vectors is such that $\lambda_{h1} \leq \ldots \leq \lambda_{hN}$.

Henceforth, $m \geq 1$ denotes a fixed number, and we assume that h is small enough so that $m \leq N$. Set $V_m = \text{span}\{\psi_1, \ldots, \psi_m\}$, and define S_m to be the unit sphere of V_m in $L^2(\Omega)$. Introduce the elliptic projector $\Pi_h : H_0^1(\Omega) \to V_h$ such that $a(\Pi_h u - u, v_h) = 0$ for all v_h in V_h, and define

$$\sigma_{hm} = \inf_{v \in S_m} \|\Pi_h v\|_{0,\Omega}. \tag{3.73}$$

Lemma 3.65. *Let $1 \leq m \leq N$. Assume $\sigma_{hm} \neq 0$. Then,*

$$\lambda_m \leq \lambda_{hm} \leq \lambda_m \sigma_{hm}^{-2}. \tag{3.74}$$

Proof. The first inequality is a simple consequence of Proposition 3.63. Furthermore, since $\sigma_{hm} \neq 0$, $\mathrm{Ker}(\Pi_h) \cap V_m = \{0\}$; hence, the Rank Theorem implies $\dim(\Pi_h V_m) = m$. Adapting the proof of Proposition 3.63, one readily infers

$$\lambda_{hm} \leq \max_{v_h \in \Pi_h V_m} \frac{a(v_h, v_h)}{\|v_h\|_{0,\Omega}^2} = \max_{v \in V_m} \frac{a(\Pi_h v, \Pi_h v)}{\|\Pi_h v\|_{0,\Omega}^2}.$$

Hence,

$$\lambda_{hm} \leq \max_{v \in V_m} \frac{a(v,v)}{\|\Pi_h v\|_{0,\Omega}^2} \leq \max_{v \in V_m} R(v) \max_{v \in V_m} \frac{\|v\|_{0,\Omega}^2}{\|\Pi_h v\|_{0,\Omega}^2} = \frac{1}{\sigma_{hm}^2} \max_{v \in S_m} R(v).$$

Then, use $\lambda_m = \max_{v \in S_m} R(v)$ to conclude. □

Remark 3.66. It is remarkable that, independently of the approximation space (provided conformity holds), the N eigenvalues of the approximate problem (3.71) are larger than the corresponding eigenvalues of the exact problem (3.68). Eigenvalues are thus approximated from above. □

Lemma 3.67. *Let $1 \leq m \leq N$. There is $c(m)$, independent of h, such that*

$$\sigma_{hm}^2 \geq 1 - c(m) \max_{v \in S_m} \|v - \Pi_h v\|_{1,\Omega}^2. \tag{3.75}$$

Proof. Let $v \in S_m$. Let $(V_i)_{1 \leq i \leq m}$ be the coordinate vector of v relative to the basis $\{\psi_1, \ldots, \psi_m\}$. It is clear that $\|v\|_{0,\Omega}^2 = \sum_{1 \leq i \leq m} V_i^2 = 1$. In addition, $\|\Pi_h v\|_{0,\Omega}^2$ is bounded from below as follows:

$$\|\Pi_h v\|_{0,\Omega}^2 \geq \|v\|_{0,\Omega}^2 - 2(v, v - \Pi_h v)_{0,\Omega}. \tag{3.76}$$

Using the symmetry of a and the definition of $\Pi_h v$ yields

$$\begin{aligned}
(v, v - \Pi_h v)_{0,\Omega} &= \sum_{1 \leq i \leq m} V_i (\psi_i, v - \Pi_h v)_{0,\Omega} = \sum_{1 \leq i \leq m} \frac{V_i}{\lambda_i} a(\psi_i, v - \Pi_h v) \\
&= \sum_{1 \leq i \leq m} \frac{V_i}{\lambda_i} a(\psi_i - \Pi_h \psi_i, v - \Pi_h v) \\
&\leq \frac{\|a\|}{\lambda_1} \|v - \Pi_h v\|_{1,\Omega} \left(\sum_{1 \leq i \leq m} \|\psi_i - \Pi_h \psi_i\|_{1,\Omega}^2 \right)^{\frac{1}{2}} \\
&\leq \sqrt{m} \frac{\|a\|}{\lambda_1} \|v - \Pi_h v\|_{1,\Omega} \sup_{w \in S_m} \|w - \Pi_h w\|_{1,\Omega} \\
&\leq \sqrt{m} \frac{\|a\|}{\lambda_1} \sup_{w \in S_m} \|w - \Pi_h w\|_{1,\Omega}^2.
\end{aligned}$$

Then, the desired estimate is obtained by inserting this bound into (3.76) and setting $c(m) = 2\sqrt{m}\frac{\|a\|}{\lambda_1}$. □

Lemma 3.68. *Assume that the sequence of approximation spaces* $\{V_h\}_{h>0}$ *is endowed with the following approximability property:*

$$\forall v \in H^1_0(\Omega), \quad \lim_{h\to 0}\left(\inf_{v_h\in V_h} \|v - v_h\|_{1,\Omega}\right) = 0. \tag{3.77}$$

Then, for all $m \geq 1$*, there is* $h_0(m)$ *such that, for all* $h \leq h_0(m)$*,*

$$0 \leq \lambda_{hm} - \lambda_m \leq 2\lambda_m c(m) \max_{v\in S_m} \inf_{v_h\in V_h} \|v - v_h\|^2_{1,\Omega}. \tag{3.78}$$

Proof. Let $m \geq 1$ be a fixed number, and assume that h is small enough so that $m \leq N$. Since S_m is compact, there is v_0 in S_m such that $\sup_{v\in S_m} \|v - \Pi_h v\|^2_{1,\Omega} = \|v_0 - \Pi_h v_0\|^2_{1,\Omega}$. Owing to (2.24),

$$\|v_0 - \Pi_h v_0\|_{1,\Omega} \leq \left(\tfrac{\|a\|}{\alpha}\right)^{\frac{1}{2}} \inf_{v_h\in V_h} \|v_0 - v_h\|_{1,\Omega}.$$

Since m is fixed, (3.77) implies that there is $h_0(m)$ such that, for all $h \leq h_0(m)$, $c(m)\|v_0 - \Pi_h v_0\|^2_{1,\Omega} \leq \frac{1}{2}$. Then, observing that $1+2x \geq \frac{1}{1-x}$ for all $0 \leq x \leq \frac{1}{2}$ and using (3.75) yields

$$1 + 2c(m)\|v_0 - \Pi_h v_0\|^2_{1,\Omega} = 1 + 2c(m) \sup_{v\in S_m} \|v - \Pi_h v\|^2_{1,\Omega} \geq \sigma^{-2}_{hm}.$$

Conclude using (3.74). □

To analyze the approximation error for eigenvectors, we assume, for the sake of simplicity, that the eigenvalues are simple.

Lemma 3.69. *Let* $1 \leq m \leq N$ *and set* $\rho_{hm} = \max_{1\leq i\neq m\leq N} \frac{\lambda_m}{|\lambda_m - \lambda_{hi}|}$. *If* λ_m *is simple, there is* $h_0(m)$ *and a choice of eigenvector such that, for all* $h \leq h_0(m)$*,*

$$\|\psi_m - \psi_{hm}\|_{0,\Omega} \leq 2(1 + \rho_{hm})\|\psi_m - \Pi_h\psi_m\|_{0,\Omega}. \tag{3.79}$$

Proof. (1) Note that owing to Lemma 3.68, $\lambda_{hi} \to \lambda_i$ as $h \to 0$. Hence, since λ_m is simple, ρ_{hm} is uniformly bounded when h is small enough.
(2) Define $v_{hm} = (\Pi_h\psi_m, \psi_{hm})_{0,\Omega}\psi_{hm}$ and let us evaluate $\|\Pi_h\psi_m - v_{hm}\|_{0,\Omega}$. Note first that

$$(\Pi_h\psi_m, \psi_{hi})_{0,\Omega} = \frac{1}{\lambda_{hi}} a(\psi_{hi}, \Pi_h\psi_m) = \frac{1}{\lambda_{hi}} a(\psi_m, \psi_{hi}) = \frac{\lambda_m}{\lambda_{hi}} (\psi_m, \psi_{hi})_{0,\Omega}.$$

Hence, $(\Pi_h\psi_m, \psi_{hi})_{0,\Omega} = \frac{\lambda_m}{\lambda_{hi}-\lambda_m}(\psi_m - \Pi_h\psi_m, \psi_{hi})_{0,\Omega}$. As a result,

$$\|\Pi_h\psi_m - v_{hm}\|^2_{0,\Omega} = \sum_{1\leq i\neq m\leq N} (\Pi_h\psi_m, \psi_{hi})^2_{0,\Omega} \leq \rho^2_{hm}\|\psi_m - \Pi_h\psi_m\|^2_{0,\Omega}. \tag{3.80}$$

(3) Let us now estimate $\|\psi_{hm} - v_{hm}\|_{0,\Omega}$. Since

$$\|\psi_m\|_{0,\Omega} - \|\psi_m - v_{hm}\|_{0,\Omega} \leq \|v_{hm}\|_{0,\Omega} \leq \|\psi_m\|_{0,\Omega} + \|\psi_m - v_{hm}\|_{0,\Omega},$$

and $\|\psi_m\|_{0,\Omega} = 1$, we infer $|\|v_{hm}\|_{0,\Omega} - 1| \leq \|\psi_m - v_{hm}\|_{0,\Omega}$. But,

$$\|\psi_{hm} - v_{hm}\|_{0,\Omega} = |(\Pi_h\psi_m - \psi_{hm}, \psi_{hm})_{0,\Omega}| = |(\Pi_h\psi_m, \psi_{hm})_{0,\Omega} - 1|.$$

Assume that ψ_{hm} is chosen so that $(\Pi_h\psi_m, \psi_{hm})_{0,\Omega} \geq 0$. Then, $\|v_{hm}\|_{0,\Omega} = (\Pi_h\psi_m, \psi_{hm})_{0,\Omega}$, yielding

$$\|\psi_{hm} - v_{hm}\|_{0,\Omega} \leq \|\psi_m - v_{hm}\|_{0,\Omega}. \tag{3.81}$$

(4) To conclude, use the triangle inequality together with (3.80) and (3.81):

$$\begin{aligned}\|\psi_m - \psi_{hm}\|_{0,\Omega} &\leq \|\psi_m - \Pi_h\psi_m\|_{0,\Omega} + \|\Pi_h\psi_m - v_{hm}\|_{0,\Omega} + \|v_{hm} - \psi_{hm}\|_{0,\Omega} \\ &\leq 2(\|\psi_m - \Pi_h\psi_m\|_{0,\Omega} + \|\Pi_h\psi_m - v_{hm}\|_{0,\Omega}).\end{aligned}$$

The conclusion follows from (3.80). □

Theorem 3.70. *Let $1 \leq m \leq N$. If λ_m is simple, there is $h_0(m)$ and a choice of eigenvector such that, for all $h \leq h_0(m)$,*

$$\|\psi_m - \psi_{hm}\|_{0,\Omega} \leq c_2(m)\|\psi_m - \Pi_h\psi_m\|_{0,\Omega}, \tag{3.82}$$

$$\|\psi_m - \psi_{hm}\|_{1,\Omega} \leq c_1(m) \max_{v\in S_m} \inf_{v_h\in V_h} \|v - v_h\|_{1,\Omega}. \tag{3.83}$$

Proof. Estimate (3.82) is a direct consequence of Lemma 3.69. To control $\|\psi_m - \psi_{hm}\|_{1,\Omega}$, use the coercivity of a as follows:

$$\begin{aligned}\alpha\|\psi_m - \psi_{hm}\|_{1,\Omega}^2 &\leq a(\psi_m - \psi_{hm}, \psi_m - \psi_{hm}) \\ &= \lambda_{hm} + \lambda_m - 2\lambda_m(\psi_m, \psi_{hm})_{0,\Omega} \\ &= \lambda_{hm} - \lambda_m + \lambda_m\|\psi_m - \psi_{hm}\|_{0,\Omega}^2.\end{aligned}$$

Then, (3.83) is a consequence of the above equality, together with Lemmas 3.68 and 3.69. □

Corollary 3.71. *Let $1 \leq m \leq N$. Assume that the approximation setting is such that there is $k \geq 1$ and $c_1(m)$ so that $\inf_{v\in S_m} \|\Pi_h v - v\|_{0,\Omega} + h\|\Pi_h v - v\|_{1,\Omega} \leq c_1(m)h^{k+1}$. Then, there are $c_2(m)$, $c_3(m)$, $c_4(m)$, independent of h, such that, if h is sufficiently small, the following estimates hold:*

$$\lambda_m \leq \lambda_{hm} \leq \lambda_m + c_2(m)\, h^{2k}\lambda_m^2. \tag{3.84}$$

Moreover, if the eigenvalue λ_m is simple,

$$\begin{cases} \|\psi_m - \psi_{hm}\|_{0,\Omega} \leq c_3(m)\, h^{k+1}\lambda_m, \\ \|\psi_m - \psi_{hm}\|_{1,\Omega} \leq c_4(m)\, h^k\lambda_m, \end{cases} \tag{3.85}$$

and the constants $c_2(m)$, $c_3(m)$, $c_4(m)$ grow unboundedly as $m \to +\infty$. If λ_m is multiple, ψ_m can be chosen so that (3.85) still holds.

Proof. Simple consequence of Lemma 3.68 and Theorem 3.70. □

Remark 3.72. The above corollary shows that when h is fixed, the accuracy of the approximation decreases as m increases since $c_2(m)$, $c_3(m)$, and $c_4(m)$ grow unboundedly as $m \to +\infty$; see §3.3.6 for an illustration. □

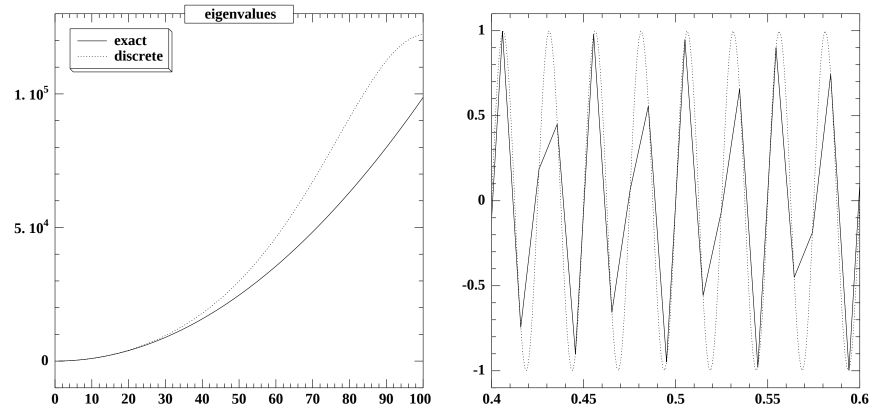


Fig. 3.4. Left: Finite element approximation to the eigenvalues of the Laplacian in one dimension. Right: Eightieth eigenfunction for the exact problem (dashed line) and for the approximate problem (solid line).

3.3.6 Numerical illustrations

In one dimension. Consider the spectral problem for the Laplacian posed in the domain $\Omega =]0, 1[$, whose solutions are the pairs

$$\{\lambda_m, \psi_m\} = \{m^2\pi^2, \sin(m\pi x)\} \qquad \text{for } m \geq 1.$$

Consider now a uniform mesh of Ω with step size $h = \frac{1}{N+1}$ and a $\mathbb{P}_1$ Lagrange finite element approximation. A straightforward calculation shows that the matrices $\mathcal{A}$ and $\mathcal{M}$ are tridiagonal and given by

$$\mathcal{A} = \frac{1}{h}\text{tridiag}(-1, 2, -1), \qquad \mathcal{M} = \frac{h}{6}\text{tridiag}(1, 4, 1).$$

The eigenvalues of the approximate problem (3.71) are easily shown to be

$$\lambda_{hm} = \frac{6}{h^2}\left(\frac{1 - \cos(m\pi h)}{2 + \cos(m\pi h)}\right), \qquad 1 \leq m \leq N.$$

The left panel in Figure 3.4 presents the first 100 eigenvalues of both the exact and the approximate problems, the latter being obtained with a mesh containing $N = 100$ points. The exact eigenvalues are approximated from above, as predicted by the theory. We also observe that only the first eigenvalues are approximated accurately. Eigenfunctions corresponding to large eigenvalues oscillate too much to be represented accurately on the mesh; see the right panel in Figure 3.4. To approximate the m-th eigenvalue with a relative accuracy of ϵ, i.e., $|\lambda_{hm} - \lambda_m| < \epsilon\lambda_m$, a mesh with step size lower than $\frac{\sqrt{\epsilon}}{m}$ must be used. In the present example, only the first 10 eigenvalues are approximated within 1% accuracy.

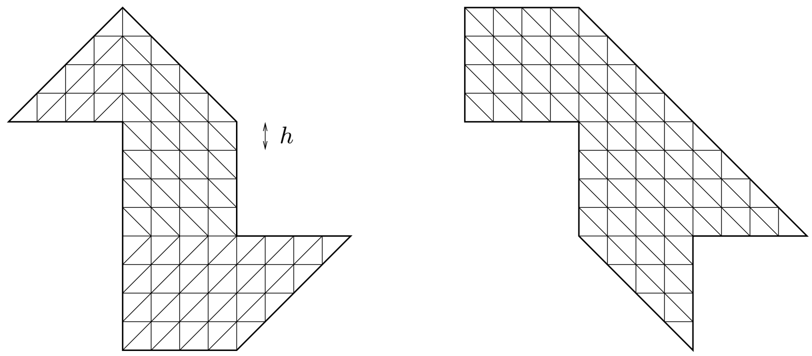


Fig. 3.5. Two domains on which the Laplacian has the same spectrum: the hen-shaped domain (left) and the arrow-shaped domain (right). The coarsest meshes used for the finite element approximation are shown. The length scale is such that the area of the two domains is equal to $\frac{7}{2}$ and that the meshes correspond to $h = \frac{1}{4}$.

Shape	Hen			Arrow		
Mesh size	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$
Eigenvalue 1	11.16	10.44	10.24	11.03	10.42	10.24
Eigenvalue 2	16.37	15.09	14.76	16.19	15.06	14.75
Eigenvalue 3	24.45	21.67	20.98	24.07	21.64	20.98

Table 3.2. First three eigenvalues for the hen- and arrow-shaped domains obtained with a first-order finite element method on three meshes of increasing refinement.

In two dimensions. Relating the spectrum and the shape of a two-dimensional membrane is a nontrivial task. For instance, knowing the spectrum $\{\lambda_m\}_{m\geq 1}$, is it possible to reconstruct the shape of the domain Ω (or, in other words, can we hear the shape of a drum)? The answer is negative, as proven recently by Gordon and Webb [GoW96] who discovered two domains in $\mathbb{R}^2$ having exactly the same spectrum. These domains take on the shape of a "hen" and an "arrow" as depicted in Figure 3.5. We verify numerically that the first eigenvalues of these two domains indeed coincide. Eigenvalues are computed using the $\mathbb{P}_1$ Lagrange finite element on a sequence of three meshes that are successively refined. The coarsest meshes are displayed in Figure 3.5; results are presented in Table 3.2. Both sets of eigenvalues converge to a common limit as $h \to 0$. The first two eigenfunctions are shown in Figure 3.6.

3.4 Continuum Mechanics

This section is concerned with PDE systems endowed with a multicomponent coercivity property. Important examples include those arising in continuum mechanics. Hereafter we restrict ourselves to *linear isotropic elasticity*. The

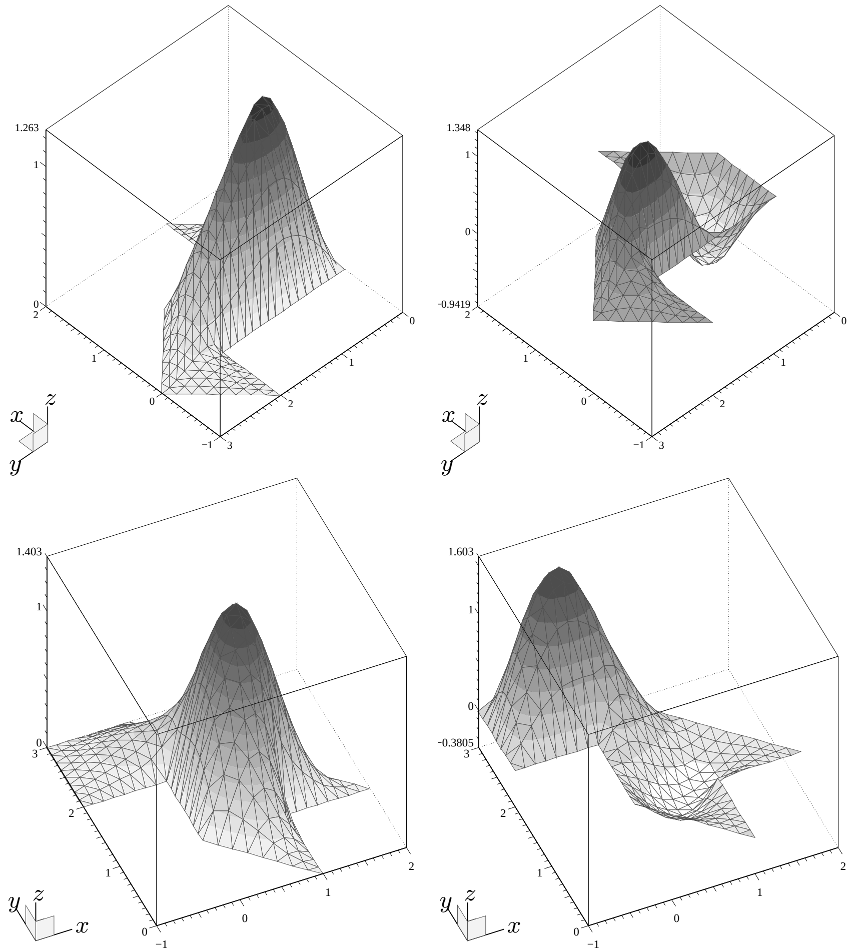


Fig. 3.6. Two first eigenfunctions for the hen-shaped domain (top) and for the arrow-shaped domain (bottom). Courtesy of E. Cancès (ENPC).

first part of this section introduces a setting for the mathematical analysis and the finite element approximation of continuum mechanics problems in this framework. The second part focuses on some problems related to beam flexion.

3.4.1 Model problems and their weak formulation

The physical model. The domain $\Omega \subset \mathbb{R}^3$ represents a deformable medium initially at equilibrium and to which an external load $f : \Omega \to \mathbb{R}^3$ is applied. Our goal is to determine the displacement field $u : \Omega \to \mathbb{R}^3$ induced by f once the system has reached equilibrium again. We assume that the deformations are small enough so that the linear elasticity theory applies.

Let $\sigma : \Omega \to \mathbb{R}^{3,3}$ be the *stress tensor* in the medium. The equilibrium conditions under the external load f can be expressed as

$$\nabla{\cdot}\sigma + f = 0 \quad \text{in } \Omega. \tag{3.86}$$

Let $\varepsilon(u) : \Omega \to \mathbb{R}^{3,3}$ be the (linearized) *strain rate tensor* defined as

$$\varepsilon(u) = \tfrac{1}{2}(\nabla u + \nabla u^T). \tag{3.87}$$

In the framework of linear isotropic elasticity, the stress tensor is related to the strain rate tensor by the relation

$$\sigma(u) = \lambda \operatorname{tr}(\varepsilon(u))\mathcal{I} + 2\mu\varepsilon(u),$$

where λ and μ are the so-called *Lamé coefficients*, and $\mathcal{I}$ is the identity matrix. Using (3.87), the above relation yields

$$\sigma(u) = \lambda(\nabla{\cdot}u)\mathcal{I} + \mu(\nabla u + \nabla u^T). \tag{3.88}$$

The Lamé coefficients λ and μ are phenomenological coefficients. Owing to thermodynamic stability, these coefficients are constrained to be such that $\mu > 0$ and $\lambda + \frac{2}{3}\mu \geq 0$. Moreover, for the sake of simplicity, we shall henceforth assume that λ and μ are constant and that $\lambda \geq 0$. In this case, owing to the identity $\nabla{\cdot}\big(\varepsilon(u)\big) = \frac{1}{2}\big(\Delta u + \nabla(\nabla{\cdot}u)\big)$, (3.86) and (3.88) yield

$$-\mu\Delta u - (\lambda + \mu)\nabla(\nabla{\cdot}u) = f \quad \text{in } \Omega.$$

The model problem (3.86)–(3.88) must be supplemented with boundary conditions. We investigate two cases: a *mixed problem* in which the displacement is imposed on part of the boundary, and a *pure-traction problem* in which the normal component of the stress tensor is imposed on the entire boundary. The *pure-displacement problem* in which the displacement is imposed on the entire boundary can be treated as a special case of the mixed problem.

Remark 3.73.

(i) The coefficient $\lambda + \frac{2}{3}\mu$ describes the compressibility of the medium; very large values correspond to almost incompressible materials.

(ii) Instead of using λ and μ, it is sometimes more convenient to consider the *Young modulus* E and the *Poisson coefficient* ν. These quantities are related to the Lamé coefficients by

$$E = \mu\frac{3\lambda + 2\mu}{\lambda + \mu} \qquad \text{and} \qquad \nu = \frac{1}{2}\frac{\lambda}{\lambda + \mu}.$$

The Poisson coefficient is such that $-1 \leq \nu < \frac{1}{2}$, and owing to the assumption $\lambda \geq 0$, we infer $\nu \geq 0$. An almost incompressible material corresponds to a

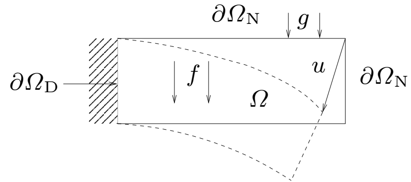


Fig. 3.7. Example of a mixed problem in continuum mechanics.

Poisson coefficient very close to $\frac{1}{2}$.

(iii) 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.

(iv) The finite element method originated in the 1950s when engineers developed it to solve continuum mechanics problems in aeronautics; see, e.g., [Lev53, ArK67] and the references cited in [Ode91]. These problems involved complex geometries that could not be easily handled by classical finite difference techniques. At the same time, theoretical researches on the approximation of linear elasticity equations were carried out [TuC56]. In 1960, Clough coined the terminology "finite elements" in a paper dealing with linear elasticity in two dimensions [Clo60]. □

Mixed problem and its weak formulation. Consider the partition $\partial\Omega = \partial\Omega_{\mathrm{D}} \cup \partial\Omega_{\mathrm{N}}$ illustrated in Figure 3.7. The boundary $\partial\Omega_{\mathrm{D}}$ is clamped, whereas a normal load $g : \partial\Omega_{\mathrm{N}} \to \mathbb{R}^3$ is imposed on $\partial\Omega_{\mathrm{N}}$. The model problem we consider is the following:

$$\begin{cases} \nabla\cdot\sigma(u) + f = 0 & \text{in } \Omega, \\ \sigma(u) = \lambda(\nabla\cdot u)\mathcal{I} + \mu(\nabla u + \nabla u^T) & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega_{\mathrm{D}}, \\ \sigma(u)\cdot n = g & \text{on } \partial\Omega_{\mathrm{N}}. \end{cases} \tag{3.89}$$

To derive a weak formulation for (3.89), take the scalar product of the equilibrium equation with a test function $v : \Omega \to \mathbb{R}^3$. Since $\int_\Omega -\big(\nabla\cdot\sigma(u)\big)\cdot v = \int_\Omega \sigma(u){:}\nabla v - \int_{\partial\Omega} v\cdot\sigma(u)\cdot n$ and $\sigma(u){:}\nabla v = \sigma(u){:}\varepsilon(v)$ owing to the symmetry of $\sigma(u)$,

$$\int_\Omega \sigma(u){:}\varepsilon(v) - \int_{\partial\Omega} v\cdot\sigma(u)\cdot n = \int_\Omega f\cdot v.$$

The displacement u and the test function v are taken in the functional space

$$V_{\mathrm{DN}} = \{v \in [H^1(\Omega)]^3;\ v = 0 \text{ on } \partial\Omega_{\mathrm{D}}\}, \tag{3.90}$$

equipped with the norm $\|v\|_{1,\Omega} = \sum_{i=1}^{3} \|v_i\|_{1,\Omega}$ where $v = (v_1, v_2, v_3)^T$. The weak formulation of (3.89) is thus:

$$\begin{cases} \text{Seek } u \in V_{\mathrm{DN}} \text{ such that} \\ a(u,v) = \int_\Omega f{\cdot}v + \int_{\partial\Omega_{\mathrm{N}}} g{\cdot}v, \quad \forall v \in V_{\mathrm{DN}}, \end{cases} \tag{3.91}$$

with the bilinear form

$$a(u,v) = \int_\Omega \sigma(u){:}\varepsilon(v) = \int_\Omega \lambda\, \nabla{\cdot}u\, \nabla{\cdot}v + \int_\Omega 2\mu\, \varepsilon(u){:}\varepsilon(v). \tag{3.92}$$

In continuum mechanics, the test function v plays the role of a virtual displacement and the weak formulation (3.91) expresses the *principle of virtual work*.

Proposition 3.74. *Let Ω be a domain in $\mathbb{R}^3$, consider the partition $\partial\Omega = \partial\Omega_{\mathrm{D}} \cup \partial\Omega_{\mathrm{N}}$, and assume that the measure of $\partial\Omega_{\mathrm{D}}$ is positive. Let λ and μ be two coefficients satisfying $\mu > 0$ and $\lambda \geq 0$. Let $f \in [L^2(\Omega)]^3$ and $g \in [L^2(\partial\Omega_{\mathrm{N}})]^3$. Then, the solution u to* (3.91) *satisfies*

$$-\mu\Delta u - (\lambda+\mu)\nabla(\nabla{\cdot}u) = f \quad \text{a.e. in } \Omega, \tag{3.93}$$

$u = 0$ a.e. on $\partial\Omega_{\mathrm{D}}$, and $\sigma{\cdot}n = g$ a.e. on $\partial\Omega_{\mathrm{N}}$.

Proof. Straightforward verification. □

Pure-traction problem and its weak formulation. The pure-traction problem consists of the following equations:

$$\begin{cases} \nabla{\cdot}\sigma(u) + f = 0 & \text{in } \Omega, \\ \sigma(u) = \lambda(\nabla{\cdot}u)\mathcal{I} + \mu(\nabla u + \nabla u^T) & \text{in } \Omega, \\ \sigma(u){\cdot}n = g & \text{on } \partial\Omega. \end{cases} \tag{3.94}$$

It is natural to seek the solution and take the test functions in $[H^1(\Omega)]^3$. Proceeding as before yields the problem:

$$\begin{cases} \text{Seek } u \in [H^1(\Omega)]^3 \text{ such that} \\ a(u,v) = \int_\Omega f{\cdot}v + \int_{\partial\Omega} g{\cdot}v, \quad \forall v \in [H^1(\Omega)]^3. \end{cases} \tag{3.95}$$

The bilinear form a is still defined by (3.92). The difficulty is that a becomes singular on $[H^1(\Omega)]^3$. To see this, introduce the set $\mathcal{R} = \{u \in [H^1(\Omega)]^3;\, u(x) = \alpha + \beta \times x\}$, where α and β are vectors in $\mathbb{R}^3$ and where $\times$ denotes the cross-product in $\mathbb{R}^3$. A function in $\mathcal{R}$ is called a *rigid displacement* field since it corresponds to a global motion consisting of a translation and a rotation.

Lemma 3.75. *The following equivalence holds:*

$$(u \in \mathcal{R}) \iff (\forall v \in [H^1(\Omega)]^3,\ a(u,v) = 0).$$

Proof. Let $u \in \mathcal{R}$. Clearly, $\nabla\cdot u = 0$ and $\varepsilon(u) = 0$. Therefore, $a(u,v) = 0$ for all $v \in [H^1(\Omega)]^3$. Conversely, if $a(u,v) = 0$ for all $v \in [H^1(\Omega)]^3$, take $v = u$ to obtain

$$a(u,u) = \int_\Omega \lambda(\nabla\cdot u)^2 + \int_\Omega 2\mu\varepsilon(u){:}\varepsilon(u) = 0,$$

implying that $\varepsilon(u) = 0$. Moreover, the fact that, for all j,k with $1 \le j,k \le 3$,

$$\begin{aligned}\partial_{jk}u_i = \partial_k(\partial_j u_i) &= \partial_k(2\varepsilon_{ij}) - \partial_i\partial_k u_j = \partial_j(2\varepsilon_{ik}) - \partial_i\partial_j u_k \\ &= \partial_k\varepsilon_{ij} + \partial_j\varepsilon_{ik} - \partial_i\varepsilon_{jk} = 0,\end{aligned}$$

implies that all the components u_i of u are first-order polynomials. Hence,

$$u(x) = \alpha + Bx,$$

with $\alpha \in \mathbb{R}^3$ and $B \in \mathbb{R}^{3,3}$. Moreover, $\varepsilon(u) = 0$ implies $B + B^T = 0$, showing that the matrix B is skew-symmetric. Therefore, there exists a vector $\beta \in \mathbb{R}^3$ such that $Bx = \beta \times x$. This shows that $u \in \mathcal{R}$. $\square$

Taking $v \in \mathcal{R}$ in (3.95), Lemma 3.75 shows that a necessary condition for the existence of a solution to (3.94) is that the data f and g satisfy the compatibility relation

$$\forall v \in \mathcal{R}, \quad \int_\Omega f\cdot v + \int_{\partial\Omega} g\cdot v = 0. \tag{3.96}$$

Note that (3.96) expresses that the sum of the externally applied forces and their moments vanish. Furthermore, it is clear that the solution u, if it exists, is defined only up to a rigid displacement. Conventionally, we choose to seek the solution u such that $\int_\Omega u = \int_\Omega \nabla\times u = 0$ (note that both quantities are meaningful if $u \in [H^1(\Omega)]^3$). This leads to the following weak formulation:

$$\begin{cases}\text{Seek } u \in V_{\mathrm{N}} \text{ such that} \\ a(u,v) = \int_\Omega f\cdot v + \int_{\partial\Omega} g\cdot v, \quad \forall v \in V_{\mathrm{N}},\end{cases} \tag{3.97}$$

with

$$V_{\mathrm{N}} = \left\{ u \in [H^1(\Omega)]^3;\ \int_\Omega u = 0;\ \int_\Omega \nabla\times u = 0 \right\}, \tag{3.98}$$

equipped with the norm $\|\cdot\|_{1,\Omega}$.

Proposition 3.76. *Let Ω be a domain in $\mathbb{R}^3$. Let λ and μ be two coefficients satisfying $\mu > 0$ and $\lambda \ge 0$. Let $f \in [L^2(\Omega)]^3$ and let $g \in [L^2(\partial\Omega)]^3$. Assume that the compatibility condition* (3.96) *is satisfied. Then, the solution u to* (3.97) *satisfies* (3.93) *and $\sigma\cdot n = g$ a.e. on $\partial\Omega$.*

Proof. Straightforward verification. $\square$

3.4.2 Well-posedness

The *coercivity* of the bilinear form a defined in (3.92) relies on the following Korn inequalities:

Theorem 3.77 (Korn's first inequality). *Let Ω be a domain in $\mathbb{R}^3$. Set $\|\varepsilon(v)\|_{0,\Omega} = (\int_\Omega \varepsilon(v){:}\varepsilon(v))^{\frac{1}{2}}$. Then, there exists c such that*

$$\forall v \in [H_0^1(\Omega)]^3, \quad c\,\|v\|_{1,\Omega} \leq \|\varepsilon(v)\|_{0,\Omega}. \tag{3.99}$$

Proof. Let $v \in [H_0^1(\Omega)]^3$. Since v vanishes at the boundary,

$$\begin{aligned}\int_\Omega \nabla v{:}\nabla v^T &= \sum_{i,j}\int_\Omega (\partial_i v_j)(\partial_j v_i) = -\sum_{i,j}\int_\Omega (\partial^2_{ij} v_j) v_i \\ &= \sum_{i,j}\int_\Omega (\partial_i v_i)(\partial_j v_j) = \int_\Omega (\nabla{\cdot}v)^2.\end{aligned}$$

A straightforward calculation yields

$$\begin{aligned}\int_\Omega \varepsilon(v){:}\varepsilon(v) &= \tfrac{1}{4}\int_\Omega (\nabla v + \nabla v^T){:}(\nabla v + \nabla v^T) \\ &= \tfrac{1}{2}\int_\Omega \nabla v{:}\nabla v + \tfrac{1}{2}\int_\Omega \nabla v{:}\nabla v^T \\ &= \tfrac{1}{2}\int_\Omega \nabla v{:}\nabla v + \tfrac{1}{2}\int_\Omega (\nabla{\cdot}v)^2 \geq \tfrac{1}{2}\int_\Omega \nabla v{:}\nabla v = \tfrac{1}{2}|v|^2_{1,\Omega}.\end{aligned}$$

Hence, $|v|^2_{1,\Omega} \leq 2\|\varepsilon(v)\|^2_{0,\Omega}$. Inequality (3.99) then results from the Poincaré inequality applied componentwise. □

Theorem 3.78 (Korn's second inequality). *Let Ω be a domain in $\mathbb{R}^3$. Then, there exists c such that*

$$\forall v \in [H^1(\Omega)]^3, \qquad c\,\|v\|_{1,\Omega} \leq \|\varepsilon(v)\|_{0,\Omega} + \|v\|_{0,\Omega}. \tag{3.100}$$

Proof. See [Cia97, p. 11] or [DuL72, p. 110]. □

Proposition 3.79 (Mixed problem). *Let Ω be a domain in $\mathbb{R}^3$ and let $\partial\Omega_{\mathrm{D}} \subset \partial\Omega$ have positive measure. Let $f \in [L^2(\Omega)]^3$ and let $g \in [L^2(\partial\Omega_{\mathrm{N}})]^3$. Then, problem* (3.91) *is well-posed and there exists c such that*

$$\forall f \in [L^2(\Omega)]^3,\ \forall g \in [L^2(\partial\Omega_{\mathrm{N}})]^3, \quad \|u\|_{1,\Omega} \leq c\,(||f||_{0,\Omega} + \|g\|_{0,\partial\Omega_{\mathrm{N}}}).$$

Moreover, (3.91) *is equivalent to the* variational formulation

$$\min_{u\in V_{\mathrm{DN}}} \left(\tfrac{1}{2}\lambda\int_\Omega (\nabla{\cdot}u)^2 + \tfrac{1}{2}\mu\int_\Omega \varepsilon(u){:}\varepsilon(u) - \int_\Omega f{\cdot}u - \int_{\partial\Omega_{\mathrm{N}}} g{\cdot}u \right).$$

Proof. If $\partial\Omega_{\mathrm{D}} = \partial\Omega$, $V_{\mathrm{DN}} = [H_0^1(\Omega)]^3$. Coercivity then results from Korn's first inequality since

$$\forall u \in [H_0^1(\Omega)]^3, \quad a(u,u) \geq 2\mu\int_\Omega \varepsilon(u){:}\varepsilon(u) \geq c\,\|u\|_{1,\Omega}^2.$$

If $\partial\Omega_{\mathrm{D}} \subsetneq \partial\Omega$, coercivity results from Korn's second inequality and a compacity argument; see the proof of Proposition 3.81. Conclude using the Lax–Milgram Lemma and Proposition 2.4. □

Remark 3.80. Given a displacement u, the quantity $J(u)$ represents the total energy of the deformed medium Ω. The quadratic terms correspond to the elastic deformation energy and the linear terms to the potential energy associated with external loads. □

Proposition 3.81 (Pure-traction problem). *Let Ω be a domain in $\mathbb{R}^3$. Assume that $f \in [L^2(\Omega)]^3$ and $g \in [L^2(\partial\Omega)]^3$ satisfy the compatibility condition* (3.96)*. Then, problem* (3.97) *is well-posed and there exists c such that*

$$\forall f \in [L^2(\Omega)]^3,\ \forall g \in [L^2(\partial\Omega)]^3, \quad \|u\|_{1,\Omega} \leq c\,(||f||_{0,\Omega} + \|g\|_{0,\partial\Omega}).$$

Moreover, (3.97) *is equivalent to the* variational formulation

$$\min_{u\in V_{\mathrm{N}}} \left(\tfrac{1}{2}\lambda\int_\Omega (\nabla{\cdot}u)^2 + \tfrac{1}{2}\mu\int_\Omega \varepsilon(u){:}\varepsilon(u) - \int_\Omega f{\cdot}u - \int_{\partial\Omega} g{\cdot}u \right).$$

Proof. Coercivity results from Korn's second inequality and from the Petree–Tartar Lemma. Indeed, set $X = V_{\mathrm{N}}$, $Y = [L^2(\Omega)]^{3,3}$, and $A : X \ni u \mapsto \varepsilon(u) \in Y$. Lemma 3.75 implies that the operator A is injective. Set $Z = [L^2(\Omega)]^3$ and let T be the compact injection from X into Z. Korn's second inequality yields

$$\forall u \in X, \quad \|u\|_X \leq c\,(\|Au\|_Y + \|Tu\|_Z).$$

Applying the Petree–Tartar Lemma yields $\|u\|_X \leq c\,\|Au\|_Y$ for all $u \in X$, i.e.,

$$\forall u \in V_{\mathrm{N}}, \quad \|u\|_{1,\Omega} \leq c\,\|\varepsilon(u)\|_{0,\Omega}.$$

This inequality shows that the bilinear form a is coercive on V_{N}. To complete the proof, use the Lax–Milgram Lemma and Proposition 2.4. □

3.4.3 Finite element approximation

For the sake of simplicity, we assume that Ω is a polyhedron.

H^1-conforming approximation. We consider a H^1-conforming finite element approximation of problems (3.91) and (3.97) based on a family of affine, geometrically conforming meshes $\{\mathcal{T}_h\}_{h>0}$ and a Lagrange finite element of degree $k \geq 1$ denoted by $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$.

To approximate the mixed problem, we assume, for the sake of simplicity, that $\partial\Omega_{\mathrm{D}}$ is a union of mesh faces. Hence, the approximation space

$$V_h^k = \{v_h \in [\mathcal{C}^0(\overline{\Omega})]^3;\, \forall K \in \mathcal{T}_h,\, v_h \circ T_K \in [\widehat{P}]^3;\, v_h = 0 \text{ on } \partial\Omega_{\mathrm{D}}\},$$

is V_{DN}-conforming. Consider the discrete problem:

$$\begin{cases} \text{Seek } u_h \in V_h^k \text{ such that} \\ a(u_h, v_h) = \int_\Omega f{\cdot}v_h + \int_{\partial\Omega_{\mathrm{N}}} g{\cdot}v_h, \quad \forall v_h \in V_h^k. \end{cases} \tag{3.101}$$

Proposition 3.82 (Mixed problem). *Let u solve* (3.91) *and let u_h solve* (3.101). *In the above setting,* $\lim_{h\to 0} \|u - u_h\|_{1,\Omega} = 0$. *Furthermore, if $u \in [H^{l+1}(\Omega)]^3 \cap V_{\mathrm{DN}}$ for some $l \in \{1, \ldots, k\}$, there exists c such that*

$$\forall h, \quad \|u - u_h\|_{1,\Omega} \leq c\, h^l |u|_{l+1,\Omega}.$$

Proof. Direct consequence of Céa's Lemma and Corollary 1.109 applied componentwise. ☐

Remark 3.83. It is not possible to apply the Aubin–Nitsche Lemma to derive an error estimate in the $[L^2(\Omega)]^3$-norm because the mixed problem is not endowed with a suitable smoothing property. ☐

For the pure-traction problem, one possible way to eliminate the arbitrary rigid displacement is the following:

(i) Impose the displacement of a node, say a_0, to be zero.
(ii) Choose three additional nodes a_1, a_2, a_3, and three unit vectors τ_1, τ_2, τ_3 such that the set $\{(a_i - a_0) \times \tau_i\}_{1\leq i\leq 3}$ forms a basis of $\mathbb{R}^3$, and impose the displacement of the node a_i along the direction τ_i to be zero.

This procedure leads to the approximation space

$$\begin{aligned} W_h^k = \{v_h \in [\mathcal{C}^0(\overline{\Omega})]^3;\, \forall K \in \mathcal{T}_h,\, v_h \circ T_K \in [\widehat{P}]^3;& \\ v_h(a_0) = 0;\, v_h(a_i){\cdot}\tau_i = 0,\, i = 1,2,3\},& \end{aligned}$$

and to the discrete problem:

$$\begin{cases} \text{Seek } u_h \in W_h^k \text{ such that} \\ a(u_h, v_h) = \int_\Omega f{\cdot}v_h + \int_{\partial\Omega} g{\cdot}v_h, \quad \forall v_h \in W_h^k. \end{cases} \tag{3.102}$$

Proposition 3.84 (Pure-traction problem). *Let u solve* (3.91) *and let u_h solve* (3.102). *In the above setting,* $\lim_{h\to 0} \|u - u_h\|_{1,\Omega} = 0$. *Furthermore, if $u \in [H^{l+1}(\Omega)]^3 \cap V_{\mathrm{N}}$ for some $l \in \{1, \ldots, k\}$, there exists c such that*

$$\forall h, \quad \|u - u_h\|_{1,\Omega} \le c\, h^l |u|_{l+1,\Omega}.$$

In addition, if Ω is convex and $g = 0$, there is c such that

$$\forall h, \quad \|u - u_h\|_{0,\Omega} \le c\, h^{l+1} |u|_{l+1,\Omega}.$$

Proof. Use Céa's Lemma, together with Corollary 1.109, to obtain the H^1-error estimate. Furthermore, the homogeneous pure-traction problem posed over a convex polyhedron is endowed with a smoothing property [Gri92, p. 135]. The L^2-error estimate then results from the Aubin–Nitsche Lemma. □

Crouzeix–Raviart approximation. Non-conforming finite element approximations to the equations of elasticity can be considered using the Crouzeix–Raviart finite element introduced in §1.2.6. For pure-traction problems, the main difficulty in the analysis is to prove an appropriate version of Korn's second inequality. This result can be established for non-conforming piecewise quadratic or cubic finite elements, but is false for piecewise linear interpolation. For Crouzeix–Raviart interpolation, appropriate modifications of the method are discussed in [Fal91, Rua96].

One important advantage of non-conforming approximations is that they yield optimal-order error estimates that are uniform in the Poisson coefficient ν. Such a property is particularly useful when modeling almost incompressible materials since it is well-known that, in this case, H^1-conforming finite elements suffer from a severe deterioration in the convergence rate; see §3.5.3 for an illustration.

Numerical illustrations. As a first example, consider the horizontal deformations of a two-dimensional, rectangular plate with a circular hole. The triangulation of the plate is depicted in the left panel of Figure 3.8. The left side is clamped, the displacement $(1, 0)$ is imposed on the right side, and zero normal stress is imposed on the three remaining sides. There is no external load, and the Lamé coefficients are such that $\frac{\lambda}{\mu} = 1$. The plate in its equilibrium configuration is shown in the right panel of Figure 3.8. $\mathbb{P}_1$ Lagrange finite elements have been used.

The second example deals with the three-dimensional body illustrated in Figure 3.9. A transverse load is imposed at the forefront of the body. The approximate solution has been obtained using first-order prismatic Lagrange

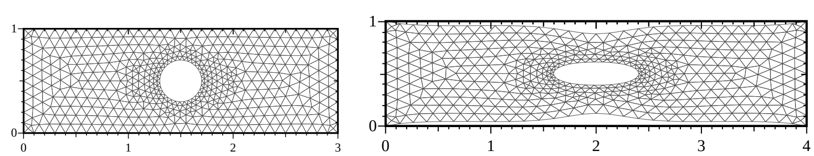


Fig. 3.8. Deformation of an elastic plate with a hole: reference configuration (left); equilibrium configuration (right).

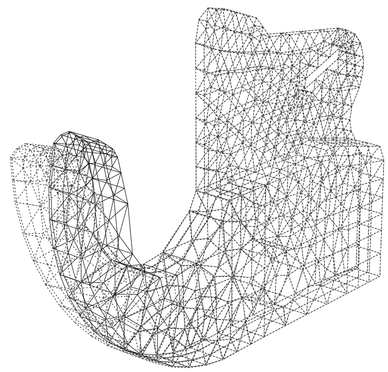


Fig. 3.9. Three-dimensional continuum mechanics problem in which a transverse load is applied to the forefront of the body; reference and equilibrium configurations are presented; approximation with prismatic Lagrange finite elements of degree 1. Courtesy of D. Chapelle (INRIA).

finite elements. Figure 3.9 presents the reference and the equilibrium configurations.

3.4.4 Beam flexion and fourth-order problems

The physical model. We investigate a model for beam flexion due to *Timoshenko*; see, e.g., [Bat96]. Consider the horizontal beam of length L shown in Figure 3.10. The x-coordinate is set so as to coincide with the beam axis. The beam is clamped into a rigid wall at $x = 0$. Impose a distributed load $f = (f_x, f_y)$ in the (x, y)-plane and a distributed momentum m parallel to the z-axis. Impose further a point force $F = (F_x, F_y)$ and a point momentum M at the beam extremity located at $x = L$. Assuming that the axis of the beam remains in the (x, y)-plane, the beam flexion can be described by the displacement $u = (u_x, u_y)$ of the points along the axis and by the rotation angle θ of the corresponding transverse sections.

In the Timoshenko model, the tangential displacement u_x uncouples from the unknowns u_y and θ. Setting $\Omega =]0, L[$, u_x solves $-u_x'' = \frac{1}{ES} f_x$ in Ω with boundary conditions $u_x(0) = 0$ and $u_x'(L) = \frac{1}{ES} F_x$, where E is the Young modulus and S is the area of the beam section. Thus, a one-dimensional second-order PDE with mixed boundary conditions is recovered.

To alleviate the notation, we now write u instead of u_y, f instead of f_y, and F instead of F_y. The displacement u and the rotation angle θ satisfy the PDEs

$$-(u'' - \theta') = \frac{\gamma}{EI} f \quad \text{and} \quad -\gamma\theta'' - (u' - \theta) = \frac{\gamma}{EI} m, \tag{3.103}$$

where I is the inertia moment of the beam, $\gamma = \frac{2(1+\nu)I}{S\kappa}$, and κ is an empirical correction factor (usually set to $\frac{5}{6}$). Boundary conditions for u and θ are

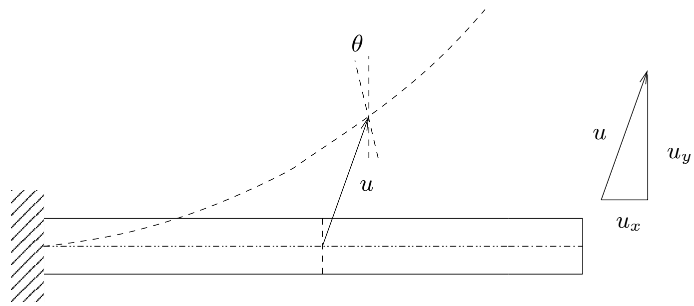


Fig. 3.10. Timoshenko model for beam flexion.

$$u(0) = 0, \quad \theta(0) = 0, \qquad (u' - \theta)(L) = \frac{\gamma}{EI} F, \quad \theta'(L) = \frac{1}{EI} M. \tag{3.104}$$

Weak formulation and coercivity. Let v be a test function for the normal displacement u and let ω be a test function for the rotation angle θ. Multiply the first equation in (3.103) by v, the second by ω, and integrate by parts over Ω to obtain the weak formulation:

$$\begin{cases} \text{Seek } (u, \theta) \in X \times X \text{ such that } \forall (v, \omega) \in X \times X, \\ a\big((u, \theta), (v, \omega)\big) = \frac{\gamma}{EI} [\int_\Omega (fv + m\omega) + Fv(L) + M\omega(L)], \end{cases} \tag{3.105}$$

where

$$a\big((u, \theta), (v, \omega)\big) = \int_\Omega \gamma \theta' \omega' + \int_\Omega (u' - \theta)(v' - \omega), \tag{3.106}$$

and $X = \{v \in H^1(\Omega); v(0) = 0\}$. Equip the product space $X \times X$ with the norm $\|(u, \theta)\|_{X \times X} = \|u\|_{1,\Omega} + \|\theta\|_{1,\Omega}$. One readily verifies the following:

Proposition 3.85. *Let f and $m \in L^2(\Omega)$. If the couple (u, θ) solves (3.105), it satisfies (3.103) a.e. in Ω and the boundary conditions (3.104).*

Theorem 3.86 (Coercivity). *Let $\gamma > 0$, let $f, m \in L^2(\Omega)$, and let $F, M \in \mathbb{R}$. Then, problem (3.105) is well-posed. Moreover, (u, θ) solves (3.105) if and only if it minimizes over $X \times X$ the energy functional*

$$J(u, \theta) = \tfrac{1}{2} \int_\Omega \gamma (\theta')^2 + \tfrac{1}{2} \int_\Omega (u' - \theta)^2 - \tfrac{\gamma}{EI} \left[\int_\Omega (fu + m\theta) + Fu(L) + M\theta(L) \right].$$

Proof. The key point is to verify the coercivity of the bilinear form a defined by (3.106). A straightforward calculation yields

$$a\big((u, \theta), (u, \theta)\big) = \int_\Omega \gamma (\theta')^2 + \int_\Omega (u')^2 + \int_\Omega \theta^2 - 2 \int_\Omega \theta u'.$$

Let $\mu > 0$. Use inequality (A.3) with parameter μ, together with the Poincaré inequality $c_\Omega \|v\|_{0,\Omega} \le \|v'\|_{0,\Omega}$ valid for all $v \in X$, to obtain

$$a\big((u,\theta),(u,\theta)\big) \geq \gamma|\theta|^2_{1,\Omega} + |u|^2_{1,\Omega} + \|\theta\|^2_{0,\Omega} - \mu\|\theta\|^2_{0,\Omega} - \frac{1}{\mu}|u|^2_{1,\Omega}$$
$$\geq \left(1 - \frac{1}{\mu}\right)|u|^2_{1,\Omega} + \frac{\gamma}{2}|\theta|^2_{1,\Omega} + \left(\frac{\gamma}{2}c^2_\Omega + 1 - \mu\right)\|\theta\|^2_{0,\Omega}.$$

Taking $\mu = 1 + \frac{\gamma}{2}c^2_\Omega$ yields

$$a\big((u,\theta),(u,\theta)\big) \geq \frac{\frac{\gamma}{2}c^2_\Omega}{1+\frac{\gamma}{2}c^2_\Omega}|u|^2_{1,\Omega} + \frac{\gamma}{2}|\theta|^2_{1,\Omega} \geq \alpha(\gamma)\|(u,\theta)\|^2_{X\times X},$$

with $\alpha(\gamma) = \frac{\gamma}{4}\frac{c^2_\Omega}{1+c^2_\Omega}\inf\big(1, c^2_\Omega/(1+\frac{\gamma}{2}c^2_\Omega)\big) > 0$; since $\gamma > 0$, a is coercive. Conclude using the Lax–Milgram Lemma and Proposition 2.4. □

Discrete approximation. Let $\mathcal{T}_h$ be a mesh of Ω with vertices $0 = x_0 < x_1 < \ldots < x_N < x_{N+1} = L$ where N is a given integer. Consider a conforming $\mathbb{P}_k$ Lagrange finite element approximation for both u and θ. The approximation space we consider is thus

$$X_h = \{v_h \in \mathcal{C}^0(\overline{\Omega});\, \forall i \in \{0,\ldots,N\},\, v_{h|[x_i,x_{i+1}]} \in \mathbb{P}_k;\, v_h(0) = 0\},$$

yielding the approximate problem:

$$\begin{cases} \text{Seek } (u_h,\theta_h) \in X_h \times X_h \text{ such that, } \forall (v_h,\omega_h) \in X_h \times X_h, \\ a\big((u_h,\theta_h),(v_h,\omega_h)\big) = \frac{\gamma}{EI}[\int_\Omega (f v_h + m\omega_h) + F v_h(L) + M\omega_h(L)]. \end{cases} \quad (3.107)$$

Theorem 3.87. *Let $\mathcal{T}_h$ be a mesh of Ω. Along with the assumptions of Theorem* 3.86*, assume that u and $\theta \in H^s(\Omega)$ for some $s \geq 2$. Then, setting $l = \min(k, s-1)$, there exists c such that, for all h,*

$$|u - u_h|_{1,\Omega} + |\theta - \theta_h|_{1,\Omega} \leq c\, h^l \max(|u|_{l+1,\Omega}, |\theta|_{l+1,\Omega}),$$
$$\|u - u_h\|_{0,\Omega} + \|\theta - \theta_h\|_{0,\Omega} \leq c\, h^{l+1} \max(|u|_{l+1,\Omega}, |\theta|_{l+1,\Omega}).$$

Proof. The estimate in the H^1-norm results from Céa's Lemma and from Proposition 1.12 applied to u and θ. The estimate in the L^2-norm results from the Aubin–Nitsche Lemma. Indeed, one easily checks that the adjoint problem is endowed with the required smoothing property. □

Navier–Bernoulli model and fourth-order problems. A case often encountered in applications arises when the parameter γ becomes extremely small. In the limit $\gamma \to 0$, the Navier–Bernoulli model is recovered

$$u' - \theta = 0 \quad \text{on } \Omega,$$

meaning that the sections of the bended beam remain orthogonal to the axis. Assuming that $m = 0$, $EI = 1$, and that the beam is clamped at its two extremities, the normal displacement u is governed by the fourth-order PDE

$u'''' = f$ in Ω with boundary conditions $u(0) = u(L) = u'(0) = u'(L) = 0$, leading to the weak formulation:

$$\begin{cases} \text{Seek } u \in H_0^2(\Omega) \text{ such that} \\ \int_0^L u''v'' = \int_0^L fv, \quad \forall v \in H_0^2(\Omega). \end{cases} \tag{3.108}$$

Proposition 3.88. *Let $f \in L^2(\Omega)$. Then, problem* (3.108) *is well-posed. Moreover, problem* (3.108) *is equivalent to minimizing over $H_0^2(\Omega)$ the energy functional $J(v) = \frac{1}{2}\int_\Omega (v'')^2 - \int_\Omega fv$.*

Proof. Left as an exercise. □

We consider a H^2-*conforming* approximation to problem (3.108) using a Hermite finite element approximation. Taking the boundary conditions into account leads to the approximation space

$$\begin{aligned} X_{h0}^3 = \{v_h \in \mathcal{C}^1(\overline{\Omega}); \forall i \in \{0, \dots, N\}, v_{h|[x_i, x_{i+1}]} \in \mathbb{P}_3; \\ v_h(0) = v_h'(0) = v_h(L) = v_h'(L) = 0\}, \end{aligned}$$

and the discrete problem:

$$\begin{cases} \text{Seek } u_h \in X_{h0}^3 \text{ such that} \\ \int_0^1 u_h''v_h'' = \int_0^1 fv_h, \quad \forall v_h \in X_{h0}^3. \end{cases} \tag{3.109}$$

Proposition 3.89. *Let $\mathcal{T}_h$ be a mesh of Ω. Let $f \in L^2(\Omega)$, let u solve* (3.108), *and let u_h solve* (3.109). *Then, there exists c such that, for all h,*

$$\|u - u_h\|_{0,\Omega} + h|u - u_h|_{1,\Omega} + h^2|u - u_h|_{2,\Omega} \le c\, h^4 \|f\|_{0,\Omega}.$$

Proof. Left as an exercise. □

Example 3.90. Consider a unit-length beam clamped at its two extremities. Apply a unit load $f \equiv 1$. Approximate problem (3.109) using uniform meshes with step size $h = \frac{1}{10}, \frac{1}{20}, \frac{1}{40}$, and $\frac{1}{80}$. The left panel in Figure 3.11 presents the error along the beam. We observe that the error vanishes at the mesh points. This is because, in this simple one-dimensional problem, the Green function associated with (3.108) belongs to the approximation space X_{h0}^3; see Remark 3.25 for a justification. The right panel in Figure 3.11 presents the error in the L^2-norm, H^1-seminorm, and H^2-seminorm. Convergence orders are 4, 3, and 2, respectively, as predicted by the theory. □

Remark 3.91. The two-dimensional version of problem (3.108) is to seek $u \in H_0^2(\Omega)$ such that

$$\int_\Omega \Delta u\, \Delta v = \int_\Omega fv, \quad \forall v \in H_0^2(\Omega). \tag{3.110}$$

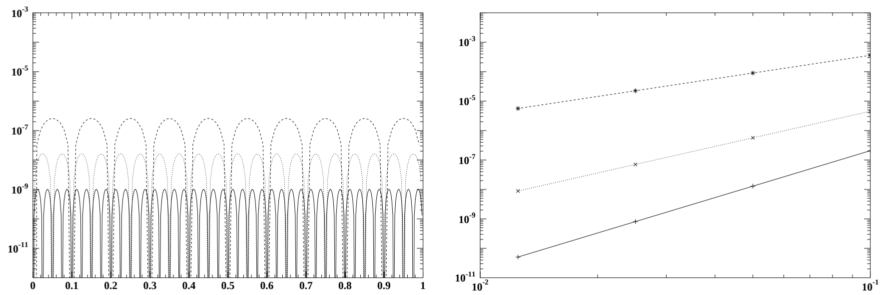


Fig. 3.11. Hermite finite element approximation for a beam flexion problem. Left: Error distribution along the beam for various mesh sizes; $h = \frac{1}{10}$ (dashed), $\frac{1}{20}$ (dotted), and $\frac{1}{40}$ (solid). Right: Error in the L^2-norm (solid), H^1-seminorm (dotted), and H^2-seminorm (dashed) as a function of mesh size.

This problem models, for instance, the bending of a clamped plate submitted to a transverse load; see [Des86, Cia97]. Regularity results for problem (3.110) are found in [GiR86, p. 17], [Cia91, p. 297], and [Gri92, p. 109]. Finite element approximations are discussed, e.g., in [Cia91, p. 273]; see also [GiR86, p. 204] for a related mixed formulation of problem (3.110) in the context of the Stokes equations in dimension 2. □

3.5 Coercivity Loss

Coercivity loss occurs when some model parameters take extreme values. In this case, although the exact problem is well-posed, discrete stability is observed only if very fine meshes are employed. The examples addressed in this section are:

(i) Advection–diffusion problems of the form (3.2) with dominant advection.
(ii) Elastic deformations of a quasi-incompressible material.
(iii) Elastic bending of a very thin Timoshenko beam.

The scope of this section is not to fix the above-mentioned problems, but to highlight the mathematical background related to coercivity loss. We identify the model parameter taking extreme values, and by letting this parameter approach zero, we derive formally a problem with no coercivity, i.e., typically involving a saddle-point or a first-order PDE. Such problems are thoroughly investigated in Chapters 4 and 5.

3.5.1 The setting

Consider the problem:

$$\begin{cases} \text{Seek } u \in V \text{ such that} \\ a_\eta(u,v) = f(v), \quad \forall v \in V, \end{cases} \tag{3.111}$$

where V is a Hilbert space, $f \in V'$, and a_η is a continuous, *coercive*, bilinear form on $V \times V$. The form a_η depends on the phenomenological parameter η that will subsequently take arbitrarily small values. Set $\|a_\eta\| := \|a_\eta\|_{V,V}$ and denote by α_η the coercivity constant of a_η, i.e.,

$$\alpha_\eta = \inf_{u \in V} \frac{a_\eta(u,u)}{\|u\|_V^2}.$$

Definition 3.92. Coercivity loss *occurs in* (3.111) *if*

$$\lim_{\eta \to 0} \frac{\|a_\eta\|}{\alpha_\eta} = \infty.$$

Remark 3.93. By analogy with the terminology adopted for linear systems in §9.1, coercivity loss amounts to the ill-conditioning of the form a. □

Let V_h be a V-conforming approximation space and assume, as is often the case in practice, that V_h is endowed with the optimal interpolation property

$$\forall u \in W, \qquad \inf_{v_h \in V_h} \|u - v_h\|_V \leq c_i h^k \|u\|_W,$$

where W is a dense subspace of V and c_i is an interpolation constant. Let u_h be the solution to the approximate problem:

$$\begin{cases} \text{Seek } u_h \in V_h \text{ such that} \\ a_\eta(u_h, v_h) = f(v_h), \quad \forall v_h \in V_h. \end{cases}$$

Assuming that the exact solution u is in W yields the error estimate

$$\|u - u_h\|_V \leq \frac{\|a_\eta\|}{\alpha_\eta} c_i h^k \|u\|_W.$$

If problem (3.111) suffers from coercivity loss, this estimate does not yield any practical control of the error. Obviously, keeping η fixed and letting $h \to 0$, convergence is achieved. However, the mesh size is limited from below by the available computer resources. Therefore, it is not always possible in practice to compensate coercivity losses by systematic mesh refinement. Some explicit examples where this situation occurs are detailed below.

3.5.2 Advection–diffusion with dominant advection

Let Ω be a domain in $\mathbb{R}^d$. Consider the advection–diffusion equation

$$-\nu \Delta u + \beta \cdot \nabla u = f \quad \text{in } \Omega, \tag{3.112}$$

where $\nu > 0$ is the diffusion coefficient, $\beta : \Omega \to \mathbb{R}^d$ the advection velocity, and $f : \Omega \to \mathbb{R}$ the source term. Following §3.1, we consider the bilinear form

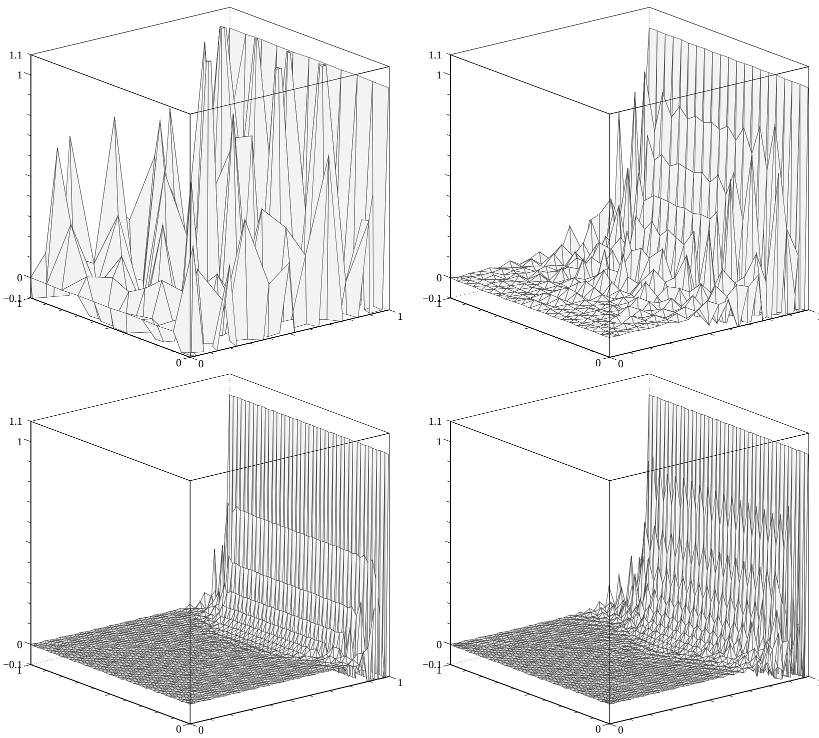


Fig. 3.12. Finite element approximation of an advection–diffusion equation with dominant advection: $h \approx \frac{1}{10}$ and $\mathbb{P}_1$ approximation (top left); $h \approx \frac{1}{20}$ and $\mathbb{P}_1$ approximation (top right); $h \approx \frac{1}{40}$ and $\mathbb{P}_1$ approximation (bottom left); $h \approx \frac{1}{20}$ and $\mathbb{P}_2$ approximation (bottom right).

$$a_\eta(u,v) = \int_\Omega \nu \nabla u \cdot \nabla v + \int_\Omega v(\beta \cdot \nabla u).$$

The parameter $\eta = \frac{\nu}{\|\beta\|_{[L^\infty(\Omega)]^d}}$ measures the relative importance of advective and diffusive effects. Assuming $\eta \ll 1$ implies

$$\frac{\|a_\eta\|}{\alpha_\eta} = O\left(\frac{\|\beta\|_{[L^\infty(\Omega)]^d}}{\nu}\right) = O\left(\frac{1}{\eta}\right) \gg 1,$$

leading to *coercivity loss*.

Figure 3.12 presents various approximate solutions to the advection–diffusion equation (3.112). The domain Ω is the unit square in $\mathbb{R}^2$. We impose $u = 1$ on the right side, $u = 0$ on the left side, and $\partial_{x_2} u = 0$ on the two other sides. The diffusion coefficient is set to $\nu = 0.002$, the advection velocity is constant and equal to $\beta = (1, 0)$, and the source term f is zero. The exact solution is

$$u(x_1, x_2) = \frac{e^{\frac{x_1}{\nu}} - 1}{e^{\frac{1}{\nu}} - 1}.$$

Since the diffusion coefficient ν takes very small values, the exact solution u is almost identically zero in Ω except in a boundary layer of width ν located near the right side where u sharply goes from 0 to 1. Three unstructured triangulations of the domain Ω are considered: a coarse mesh containing 238 triangles (triangle size $h \approx \frac{1}{10}$); an intermediate mesh containing 932 triangles (triangle size $h \approx \frac{1}{20}$); and a fine mesh containing 3694 triangles (triangle size $h \approx \frac{1}{40}$). The $\mathbb{P}_1$ Galerkin solution is computed on the three meshes: $h \approx \frac{1}{10}$, top left panel; $h \approx \frac{1}{20}$, top right panel; and $h \approx \frac{1}{40}$, bottom left panel. The $\mathbb{P}_2$ Galerkin solution computed on the intermediate mesh is shown in the bottom right panel. We observe that *spurious oscillations* pollute the approximate solution in the four cases presented. Oscillations are larger on the two coarser meshes and for the $\mathbb{P}_1$ approximation.

In the limit $\eta \to 0$, the diffusion term is negligible and the solution u is governed by a first-order PDE. Hence, to understand and fix the problems associated with coercivity loss, it is important to analyze the limit first-order PDE; this is the purpose of Chapter 5.

3.5.3 Almost incompressible materials

Almost incompressible materials, such as rubber, are characterized by Lamé coefficients λ and μ with a very large ratio $\frac{\lambda}{\mu}$. Another equivalent characterization is that the Poisson coefficient ν is very close to $\frac{1}{2}$. In §3.4.1 we introduced the bilinear form

$$a_\eta(u,v) = \int_\Omega \lambda\, \nabla\!\cdot\! u\, \nabla\!\cdot\! v + \int_\Omega 2\mu\, \varepsilon(u){:}\varepsilon(v),$$

where $\varepsilon(u)$ is the strain rate tensor. When the ratio $\eta = \frac{\mu}{\lambda}$ is very small, one verifies that

$$\frac{\|a_\eta\|}{\alpha_\eta} = O\left(\frac{\lambda}{\mu}\right) = O\left(\frac{1}{\eta}\right) \gg 1,$$

leading to *coercivity loss.*

Consider a horizontal elastic flat plate with three internal holes; see Figure 3.13. The left side is kept fixed, the displacement $(1,0)$ is imposed on the right side, and zero normal stress is imposed on the remaining external sides as well as on the three internal sides. No internal load is applied, and the

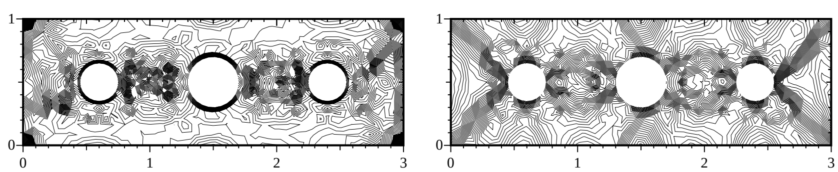


Fig. 3.13. Deformations of a horizontal, flat plate with three holes: maximal stresses (left); Tresca stresses (right).

ratio of the Lamé coefficients is $\frac{\lambda}{\mu} = 100$. Figure 3.13 presents Tresca stresses and maximal stresses obtained with a $\mathbb{P}_1$ Lagrange finite element approximation. We observe that *spurious oscillations* pollute the discrete solution; in the literature, this phenomenon is often referred to as *locking*.

When $\frac{\lambda}{\mu} \gg 1$, one can show that $\nabla\cdot u \to 0$. Introducing a new scalar unknown p in place of the product $-\lambda\nabla\cdot u$ yields

$$\begin{cases} \quad \sigma = -p\mathcal{I} + 2\mu\varepsilon(u), \\ \nabla\cdot u = 0. \end{cases}$$

Since $\Delta u = 2\nabla\cdot\varepsilon(u)$ when $\nabla\cdot u = 0$, the governing equations of an incompressible medium in the framework of linear elasticity become

$$\begin{cases} -\mu\Delta u + \nabla p = f, \\ \qquad\qquad \nabla\cdot u = 0. \end{cases}$$

Formally, we recover the Stokes equations often considered to model steady, incompressible flows of creeping fluids. The new unknown p can be identified with a pressure. The Stokes equations are endowed with a saddle-point structure. The analysis of this class of problems is the purpose of Chapter 4.

3.5.4 Very thin beams

Referring to §3.4.4 for more details, the bilinear form arising in Timoshenko's model of beam flexion is

$$a_\eta\big((u,\theta),(v,\omega)\big) = \int_\Omega \gamma\theta'\omega' + \int_\Omega (u'-\theta)(v'-\omega),$$

where u is the normal displacement of the beam axis and θ the rotation angle of the beam section. The parameter η is simply equal to γ. When $\gamma \ll 1$, the proof of Theorem 3.86 shows

$$\frac{\|a_\eta\|}{\alpha_\eta} = O\left(\frac{1}{\gamma}\right) = O\left(\frac{1}{\eta}\right) \gg 1,$$

leading to *coercivity loss*. Note that $\gamma \ll 1$ when the ratio between the inertia moment and the section of the beam is very small, as for very thin beams. In this case, the beam bends according to the Navier–Bernoulli assumption, meaning that the sections remain almost perpendicular to the beam axis.

Figure 3.14 compares analytical and approximate solutions for a beam of length $L = 1$ and parameter $EI = 1$. The flexion is induced by a force $F = 1$ applied at the extremity $x = L$. Solutions are obtained using the $\mathbb{P}_1$ finite element approximation for both the displacement u and the rotation angle θ on a uniform mesh with step size $h = \frac{1}{20}$. The left column in Figure 3.14 corresponds to the case $\gamma = 0.01$ and the right column to the case $\gamma = 0.0001$.

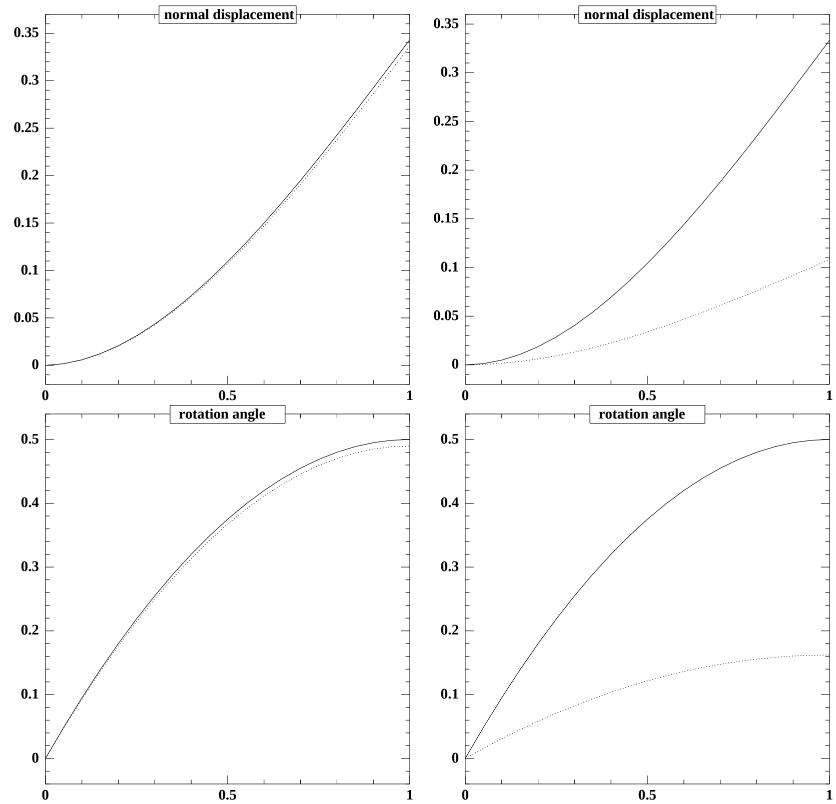


Fig. 3.14. Comparison between the analytical and finite element solutions (solid and dashed lines, respectively) for the bending of a Timoshenko beam clamped at its left extremity: $\gamma = 0.01$ (left column); $\gamma = 0.0001$ (right column).

In the second case, coercivity loss leads to very poor accuracy, indicating a *locking* phenomenon.

To pass to the limit $\gamma \to 0$ in Timoshenko's model (3.103), we introduce the auxiliary unknown

$$v = \frac{1}{\gamma}\left(u - \int_0^x \theta\right).$$

The unknowns (v, θ) satisfy the PDEs $-v'' = \frac{1}{EI} f$ and $-\theta'' - v' = \frac{1}{EI} m$ in $]0, L[$, together with the boundary conditions $v(0) = 0$, $\theta(0) = 0$, $v'(L) = \frac{1}{EI} F$, and $\theta'(L) = \frac{1}{EI} M$. One readily checks that this new problem leads to a coercive bilinear form. Furthermore, the displacement u is recovered from the first-order PDE

$$\begin{cases} u' = \gamma v' + \theta, \\ u(0) = 0. \end{cases}$$

Here, as in §3.5.2, coercivity loss is associated with the presence of a first-order PDE in the limit problem. The finite element approximation of such PDEs is investigated in Chapter 5.

3.6 Exercises

Exercise 3.1. Complete the proof of Theorem 3.8.

Exercise 3.2. Let $\Omega =]0,1[$, let $f \in L^2(\Omega)$, and let $k \in \mathbb{R}$. Consider the problem:

$$\begin{cases} \text{Seek } u \in H_0^1(\Omega) \text{ such that} \\ \int_0^1 u'v' + k\int_0^1 u'v + \int_0^1 uv = \int_0^1 fv, \quad \forall v \in H_0^1(\Omega). \end{cases}$$

(i) Write the corresponding PDE and boundary conditions.
(ii) Prove that the problem is well-posed. (*Hint*: Use the Lax–Milgram Lemma.)

Exercise 3.3. Let Ω be a domain in $\mathbb{R}^2$, let $f \in L^2(\Omega)$, and let $\sigma \in \mathbb{R}$. Show that if $|\sigma| < 1$, the following problem is well-posed:

$$\begin{cases} \text{Seek } u \in H_0^1(\Omega) \text{ such that} \\ \int_\Omega \left[\partial_x u \partial_x v + \sigma(\partial_x u \partial_y v + \partial_y u \partial_x v) + \partial_y u \partial_y v\right] = \int_\Omega fv, \quad \forall v \in H_0^1(\Omega). \end{cases}$$

Exercise 3.4. Consider the domain Ω whose definition in polar coordinates is $\Omega = \{(r,\theta);\, 0 < r < 1,\, \frac{\pi}{\alpha} < \theta < 0\}$ with $\alpha < -\frac{1}{2}$. Let $\partial\Omega_1 = \{(r,\theta);\, r = 1,\, \frac{\pi}{\alpha} < \theta < 0\}$ and $\partial\Omega_2 = \partial\Omega \backslash \partial\Omega_1$. Consider the following problem: $-\Delta u = 0$ in Ω, $u = \sin(\alpha\theta)$ on $\partial\Omega_1$, and $u = 0$ on $\partial\Omega_2$.

(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 $L^2(\Omega)$ if $-1 < \alpha < -\frac{1}{2}$.
(iii) Consider the following problem: Seek $u \in H^1(\Omega)$ such that $u = \sin(\alpha\theta)$ on $\partial\Omega_1$, $u = 0$ on $\partial\Omega_2$, and $\int_\Omega \nabla u \cdot \nabla v = 0$ for all $v \in H_0^1(\Omega)$. Prove that φ_2 solves this problem, but φ_1 does not. Comment.

Exercise 3.5 (Péclet number). Let $\Omega =]0,1[$, let $\nu > 0$, and let $\beta \in \mathbb{R}$. Consider the following problem:

$$\begin{cases} -\nu u'' + \beta u' = 1, \\ u(0) = u(1) = 0. \end{cases}$$

(i) Verify that the exact solution is $u(x) = \frac{1}{\beta}(x - \frac{1-e^{\lambda x}}{1-e^\lambda})$ with $\lambda = \frac{\beta}{\nu}$.
(ii) Plot the solution for $\beta = 1$ and $\nu = 1$, $\nu = 0.1$, and $\nu = 0.01$. Comment.
(iii) Write the problem in weak form and show that it is well-posed.

(iv) Consider a $\mathbb{P}_1$ H^1-conforming finite element approximation on a uniform grid $\mathcal{T}_h = \bigcup_{0\le i\le N}[ih,(i+1)h]$ where $h = \frac{1}{N+1}$. Show that the stiffness matrix is $\mathcal{A} = \frac{\nu}{h}\text{tridiag}(-1-\frac{\gamma}{2}, 2, -1+\frac{\gamma}{2})$, where $\gamma = \frac{\beta h}{\nu}$ is the so-called local *Péclet number*.

(v) Solve the linear system and comment. (*Hint*: If $\gamma \neq 2$, the solution is $U_i = \frac{1}{\beta}(ih - \frac{1-\delta^i}{1-\delta^{N+1}})$ where $\delta = \frac{2+\gamma}{2-\gamma}$.) What happens if $\gamma = 2$ or $\gamma = -2$?

(vi) Plot the approximate solution for $\gamma = 1$ and $\gamma = 10$. Comment.

Exercise 3.6. Let $\nu > 0$ and $b > 0$. Consider the equation $-\nu u'' + bu' = f$ posed on $]0,1[$ with the boundary conditions $u(0) = 0$ and $u'(1) = 0$.

(i) Write the weak formulation of the problem.

(ii) Let $\mathcal{T}_h$ be a mesh of $]0,1[$ and use $\mathbb{P}_1$ finite elements to approximate the problem. Let $[x_{N-1}, x_N]$ be the element such that $x_N = 1$. Let U_{N-1} and U_N be the value of the approximate solution at x_{N-1} and x_N. Write the equation satisfied by U_{N-1} and U_N when testing the weak formulation by the nodal shape function φ_N.

(iii) What is the limit of the equation derived in question (ii) when $|x_N - x_{N-1}| \to 0$? What is the limit equation when $\nu \ll |x_N - x_{N-1}|$. Comment.

Exercise 3.7. Let Ω be a domain in $\mathbb{R}^d$. Let μ be a positive constant, let β be a constant vector field, and let $f \in L^2(\Omega)$. Equip $V = H^1_0(\Omega)$ with the norm $v \mapsto \|v\|_V = \|\nabla v\|_{0,\Omega}$. Consider the problem: Seek $u \in V$ such that, for all $v \in V$, $a(u,v) = \int_\Omega fv$, where $a(u,v) = \int_\Omega \mu\nabla u\cdot\nabla v + (\beta\cdot\nabla u)v$.

(i) Explain why $v \mapsto \|\nabla v\|_{0,\Omega}$ is a norm in V.

(ii) Show that the above problem is well-posed.

(iii) Let V_h be a finite-dimensional subspace of V. Let $\lambda \ge 0$, define the bilinear form $a_h(w_h, v_h) = a(w_h, v_h) + \lambda h \int_\Omega \nabla w_h\cdot\nabla v_h$, and let $u_h \in V_h$ be such that $a_h(u_h, v_h) = \int_\Omega f v_h$ for all $v_h \in V_h$. Set $\mu_h = \mu + \lambda h$. Prove

$$\|u - u_h\|_V \le \inf_{v_h\in V_h}\left\{\frac{\lambda h}{\mu_h}\sup_{w_h\in V_h}\frac{\int_\Omega \nabla v_h\cdot\nabla w_h}{\|w_h\|_V} + \left(1 + \frac{\|a\|}{\mu_h}\right)\|u - v_h\|_V\right\}.$$

(iv) Assume that there is an interpolation operator Π_h and an integer $k > 0$ such that $\|v - \Pi_h v\|_V \le c\,h^{l-1}\|v\|_{l,\Omega}$ for all $1 \le l \le k+1$ and all $v \in H^l(\Omega) \cap V$. Prove and comment the following estimate:

$$\|u - u_h\|_V \le c\left\{\left(1 + \frac{\|a\|}{\mu_h}\right)h^k|u|_{k+1,\Omega} + \frac{\lambda}{\mu_h}h\|\nabla u\|_{0,\Omega}\right\}.$$

Exercise 3.8. The goal of this exercise is to prove estimate (3.27) using duality techniques. Assume $p < \infty$. Let $v = |u - u_h|^{p-1}\text{sgn}(u - u_h)$ and let z be the solution to the adjoint problem (3.17) with data v.

(i) Verify that $v \in L^{p'}(\Omega)$ with $\frac{1}{p} + \frac{1}{p'} = 1$.

(ii) Using assumption (iv) of Theorem 3.21, find a constant δ' such that, for $p' > \delta'$, $z \in W^{2,p'}(\Omega)$.
(iii) Show that, for all $z_h \in V_h$,

$$\|u - u_h\|^p_{L^p(\Omega)} \leq \|a\| \, \|u - u_h\|_{1,p,\Omega} \|z - z_h\|_{1,p',\Omega}.$$

(iv) Conclude.

Exercise 3.9 (Proof of Lemma 3.27).

(i) Explain why $\gamma_0(\mathcal{I}_h u) = \mathcal{I}_h^\partial\big(\gamma_0(u)\big) = \mathcal{I}_h^\partial(g) = \gamma_0(u_h)$.
(ii) Show that $a(\mathcal{I}_h u - u_h, v_h) = a(\mathcal{I}_h u - u, v_h)$, for all $v_h \in V_{h0}$.
(iii) Use (BNB1$_\text{h}$) to prove $\alpha_h \|\mathcal{I}_h u - u_h\|_{1,\Omega} \leq \|a\| \, \|\mathcal{I}_h u - u\|_{1,\Omega}$.
(iv) Conclude.

Exercise 3.10 (Proof of Lemma 3.28).

(i) Prove that there is $\theta \in H^2(\Omega) \cap H^1_0(\Omega)$ such that $a(v, \theta) = \int_\Omega (\mathcal{I}_h u - u_h) v$ for all $v \in H^1_0(\Omega)$. Show that

$$\|\mathcal{I}_h u - u_h\|^2_{0,\Omega} \leq \|a\| \, \|u - u_h\|_{1,\Omega} \|\theta - w_h\|_{1,\Omega} + a(\mathcal{I}_h u - u, \theta).$$

(ii) Using Lemma 3.27 to estimate $\|u - u_h\|_{1,\Omega}$ and using assumption (ii) in Lemma 3.27, show that

$$\begin{aligned} \|\mathcal{I}_h u - u_h\|^2_{0,\Omega} &\leq c \, \|u - \mathcal{I}_h u\|_{1,\Omega} \inf_{w_h \in V_{h0}} \|\theta - w_h\|_{1,\Omega} \\ &\quad + c \, (\|u - \mathcal{I}_h u\|_{0,\Omega} + \|g - \mathcal{I}_h g\|_{0,\partial\Omega}) \|\theta\|_{2,\Omega}. \end{aligned}$$

(iii) Show that $\inf_{w_h \in V_{h0}} \|\theta - w_h\|_{1,\Omega} \leq c \, h \|\theta\|_{2,\Omega}$ and that

$$\|\mathcal{I}_h u - u_h\|_{0,\Omega} \leq c \, (h \|u - \mathcal{I}_h u\|_{1,\Omega} + \|u - \mathcal{I}_h u\|_{0,\Omega} + \|g - \mathcal{I}_h g\|_{0,\partial\Omega}).$$

(iv) Conclude.

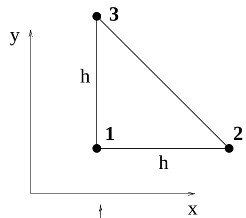
Exercise 3.11. Prove Propositions 3.88 and 3.89.

Exercise 3.12. Assume that Ω is a bounded domain of class $\mathcal{C}^2$ in $\mathbb{R}^2$. Using the notation of Lemma B.69, prove that $\nabla\cdot : [H^1_0(\Omega)]^2 \to L^2_{\int=0}(\Omega)$ is continuous and surjective. (*Hint*: For $g \in L^2_{\int=0}(\Omega)$, construct $[H^1_0(\Omega)]^2 \ni u = \nabla q + \nabla\times\psi$ such that $\nabla\cdot u = g$ and q solves a Poisson problem, ψ solves a biharmonic problem, and $\nabla\times\psi := (\partial_2\psi, -\partial_1\psi)$.)

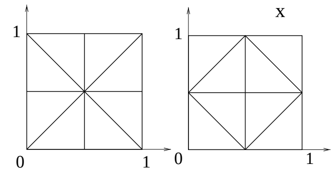
Exercise 3.13. Let Ω be a domain in $\mathbb{R}^d$. Prove that $\mathcal{C}^{0,1}(\partial\Omega) \subset H^{\frac{1}{2}}(\partial\Omega)$ with continuous embedding.

Exercise 3.14. Let $\Omega =]0,1[^2$. Consider the problem $-\Delta u + u = 1$ in Ω and $u_{|\partial\Omega} = 0$. Approximate its solution with $\mathbb{P}_1$ H^1-conforming finite elements.

(i) Let $\{\lambda_0, \lambda_1, \lambda_2\}$ be the barycentric coordinates in the triangle K_h shown in the figure. Compute the entries of the elementary stiffness matrix $\mathcal{A}_{ij} = \int_{K_h} \nabla\lambda_i \cdot \nabla\lambda_j + \int_{K_h} \lambda_i\lambda_j$, and the right-hand side vector $\int_{K_h} \lambda_i$. (*Hint*: Use a quadrature from Table 8.2.)

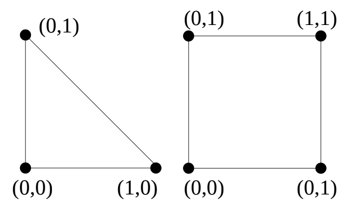


(ii) Consider the two meshes shown in the figure. Assemble the stiffness matrix and the right-hand side in both cases and compute the solution. For a fine mesh composed of 800 elements, $u_h(\frac{1}{2}, \frac{1}{2}) \approx 0,0702$. Comment.

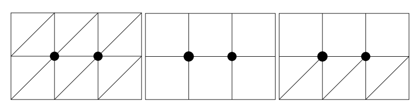


Exercise 3.15. Let $\Omega =]0,3[\times]0,2[$. Consider the problem $-\Delta u = 1$ in Ω and $u_{|\partial\Omega} = 0$. Approximate its solution with $\mathbb{P}_1$ H^1-conforming finite elements.

(i) Consider the reference simplex $\widehat{T}$ and the reference square $\widehat{K}$ shown in the figure. The nodes are numbered anticlockwise from $(0,0)$. Let $\{\widehat{\lambda}_1, \widehat{\lambda}_2, \widehat{\lambda}_3\}$ and $\{\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3, \widehat{\theta}_4\}$ be the local shape functions on $\widehat{T}$ and $\widehat{K}$, respectively. Compute the matrices $\left(\int_{\widehat{T}} \nabla\widehat{\lambda}_i \cdot \nabla\widehat{\lambda}_j\right)_{1\le i,j\le 3}$ and $\left(\int_{\widehat{K}} \nabla\widehat{\theta}_i \cdot \nabla\widehat{\theta}_j\right)_{1\le i,j\le 4}$.



(ii) Consider the meshes shown in the figure. Assemble the stiffness matrix for each of these three meshes.



Exercise 3.16. Let Ω be a two-dimensional domain and let $\{\mathcal{T}_h\}_{h>0}$ be a shape-regular family of meshes composed of affine simplices. Let $P^2_{\mathrm{pt},h}$ be the finite element space defined in (1.71). Let

$$P^2_{\mathrm{pt},h,0} = \left\{ v_h \in P^2_{\mathrm{pt},h};\ \forall F \in \mathcal{F}^\partial_h,\ \int_F v_h = 0 \right\}.$$

Prove that the extended Poincaré inequality (3.35) holds in $P^2_{\mathrm{pt},h,0}$. (*Hint*: Proceed as in the proof of Lemma 3.31.)

Exercise 3.17 (Discrete maximum principle). Let Ω be a polygonal domain in $\mathbb{R}^2$ and let $\mathcal{T}_h$ be an affine simplicial mesh of Ω. Assume that all the angles of the triangles in $\mathcal{T}_h$ are acute. Let $P^1_{\mathrm{c},h}$ be the approximation space constructed on $\mathcal{T}_h$ using continuous, piecewise linears. Let $\{\varphi_1, \dots, \varphi_N\}$ be the global shape functions and let $\mathcal{A}$ be the stiffness matrix associated with the Laplace operator, i.e., $\mathcal{A}_{ij} = \int_\Omega \nabla\varphi_i \cdot \nabla\varphi_j$ for $1 \le i,j \le N$.

(i) Show that $\mathcal{A}$ is an *M-matrix*, i.e., all its off-diagonal entries are non-positive and its row-wise sums are non-negative.

(ii) Prove the following discrete maximum principle: If $f \in L^2(\Omega)$ is such that $f \leq 0$ in Ω, the finite element solution u_h to the homogeneous Dirichlet problem with right-hand side f is such that $u_h \leq 0$ in Ω.