Lecture Notes of the course

Numerical Approximation of Partial Differential Equations – I

Fabio Nobile

A.Y. 2017-2018

Contents

Chapter 1

Weak formulation of elliptic problems

1.1 The Poisson problem

We start by considering a simple Poisson equation. Let $\Omega \subset \mathbb{R}^d$ be a bounded open domain with Lipschitz boundary $\partial\Omega$. In Ω we set the problem

$$\begin{cases} -\Delta u = f, & \text{in} \quad \Omega \\ u = 0, & \text{on} \quad \partial\Omega. \end{cases} \tag{1.1}$$

The numerical methods that we will study rely on the so called weak formulation of the problem.

Take a function $v : \Omega \to \mathbb{R}$ sufficiently smooth. We multiply equation (1.1) by v, integrate over Ω and integrate by parts the second derivatives,

$$\int_\Omega -\Delta u \cdot v = -\int_\Omega \operatorname{div}(\nabla u)v = \int_\Omega \nabla u \cdot \nabla v - \int_{\partial\Omega} (\underbrace{\nabla u \cdot n}_{\frac{\partial u}{\partial n}})v.$$

Therefore problem (1.1) becomes

$$\int_\Omega \nabla u \cdot \nabla v - \int_{\partial\Omega} \frac{\partial u}{\partial n} v = \int_\Omega fv, \qquad \forall v \text{ sufficiently smooth.}$$

In mechanics, such a procedure leads to the *principle of virtual works*. Thus the "test function" v should be interpreted as a virtual displacement and should preserve the constraints on the solution. Since the value of the solution is prescribed on the boundary, it is reasonable to take test functions that vanish on the boundary. Moreover, for the term $\int_\Omega \nabla u \cdot \nabla v$ to be bounded, it is enough to require that both ∇u and ∇v are square integrable.

1.1.1 Brief review of Sobolev spaces

We recall now the definition of the following functional spaces:

- *Space* L^2:

$$L^2(\Omega) = \{v : \Omega \to \mathbb{R} \ : \quad \int_\Omega v^2 < +\infty\}.$$

It is a Hilbert space with inner product $(f,g) = \int_\Omega fg$ and associated norm $\|f\|_{L^2(\Omega)} = (\int_\Omega f^2)^{1/2}$.

We recall moreover the important Cauchy-Schwarz inequality

$$(f,g) \leq \|f\|\|g\|.$$

- *Space* H^1:

$$H^1(\Omega) = \{v : \Omega \to \mathbb{R} \ : \quad v \in L^2(\Omega), \nabla v \in L^2(\Omega)\}.$$

This is also a Hilbert space with inner product $(f,g)_{H^1} = \int_\Omega fg + \int_\Omega \nabla f \cdot \nabla g$ and norm $\|f\|_{H^1} = \sqrt{\|f\|^2_{L^2} + \int_\Omega |\nabla f|^2} = \left(\|f\|^2_{L^2} + \sum_{i=1}^d \|\frac{\partial f}{\partial x_i}\|^2_{L^2}\right)^{\frac{1}{2}}$.

Functions in $H^1(\Omega)$ are not necessary $C^1(\Omega)$, so the derivatives have to be interpreted in a weak (distributional) sense:

$$< \partial_{x_i} v, \phi >= - < v, \partial_{x_i} \phi > \quad \forall\, \phi \in \mathcal{D}(\Omega),$$

where $\mathcal{D}(\Omega) = C_0^\infty(\Omega)$ is the space of infinitely differentiable functions with compact support in Ω. Similarly, $< \nabla v, \vec{\phi} >= - < v, \operatorname{div} \vec{\phi} > \quad \forall\, \vec{\phi} \in (\mathcal{D}(\Omega))^d$.

- *Space* H^m, with $m \in \mathbb{N}_+$. More generally, let $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ be a multi-index and

$$D^\alpha v = \frac{\partial^{|\alpha|} v}{\partial^{\alpha_1}_{x_1} \ldots \partial^{\alpha_d}_{x_d}} \quad \text{with} \quad |\alpha| = \sum_{i=1}^d \alpha_i$$

The space H^m is then defined as

$$H^m(\Omega) = \{v : \Omega \to \mathbb{R} \ : \quad D^\alpha v \in L^2(\Omega), \ \ \forall \alpha : |\alpha| \leq m\}.$$

It is a Hilbert space with respect to the inner product $(f,g)_{H^m} = \sum_{|\alpha| \leq m} (D^\alpha f, D^\alpha g)_{L^2}$ and norm $\|f\|_{H^m} = \left(\sum_{|\alpha| \leq m} \|D^\alpha f\|^2_{L^2(\Omega)}\right)^{1/2}$. We will also need the semi-norm defined as $|f|_{H^m} = \left(\sum_{|\alpha| = m} \|D^\alpha f\|^2_{L^2(\Omega)}\right)^{1/2}$.

- *Space* H^1_0:

$$H^1_0(\Omega) = \{v \in H^1(\Omega) \ : \quad v|_{\partial\Omega} = 0\}.$$

Here the condition $v|_{\partial\Omega} = 0$ has to be understood in terms of traces. We recall that the space $C^1(\Omega)$ is dense in $H^1(\Omega)$, therefore, $\forall v \in H^1$, we can build a sequence $v^\epsilon \in C^1(\Omega)$ such that $v^\epsilon \xrightarrow{H^1} v$ as $\epsilon \to 0$. We can then define the trace of v on $\partial\Omega$ denoted by $\gamma(v)$ as

$$\gamma(v) = \lim_{\epsilon \to 0} v^\epsilon|_{\partial\Omega}.$$

Here γ is a linear bounded operator from $H^1(\Omega)$ to $L^2(\partial\Omega)$, i.e. there exists $C_T > 0$ such that

$$\|\gamma(v)\|_{L^2(\partial\Omega)} \leq C\|v\|_{H^1(\Omega)}, \qquad \forall v \in H^1(\Omega).$$

The operator γ is not surjective and one can find functions $g \in L^2(\partial\Omega)$ that are not traces of any H^1 function. The image space of γ is called $H^{1/2}(\partial\Omega)$ and can be endowed with the "induced" norm

$$\|g\|_{H^{1/2}(\partial\Omega)} = \inf_{\substack{v\in H^1(\Omega)\\ \gamma(v)=g}} \|v\|_{H^1(\Omega)}.$$

With such a norm, $H^{1/2}(\partial\Omega)$ is a Banach space.

- *Poincaré inequality*: there exists $C_p > 0$, depending only on the domain Ω such that

$$\forall\, v \in H_0^1(\Omega); \quad \|v\|_{L^2} \le C_p \|\nabla v\|_{L^2}. \tag{1.2}$$

Such inequality implies in particular that in $H_0^1(\Omega)$ the full H^1-norm

$$\|v\|_{H^1} = \sqrt{\|v\|_{L^2}^2 + \|\nabla v\|_{L^2}^2}$$

and the H^1-semi norm $|v|_{H^1} = \|\nabla v\|_{L^2}$ are equivalent. Indeed

$$|v|_{H^1}^2 = \|\nabla v\|_{L^2}^2 \le \underbrace{\|v\|_{L^2}^2 + \|\nabla v\|_{L^2}^2}_{=\|v\|_{H^1}^2} \le (1 + C_p^2)\|\nabla v\|_{L^2}^2 = (1 + C_p^2)|v|_{H^1}^2$$

In H_0^1 we can therefore define the alternative norm $\|v\|_{H_0^1} = |v|_{H^1} = \|\nabla v\|_{L^2}$.

1.1.2 Weak formulation and well posedness of the Poisson problem

We now come back to the weak formulation of the Poisson problem. The right space for the test functions and the solution itself is $H_0^1(\Omega)$:

$$\text{Find } u \in H_0^1(\Omega): \quad \int_\Omega \nabla u \cdot \nabla v = \int_\Omega f v \quad \forall\, v \in H_0^1(\Omega). \tag{1.3}$$

Setting $V = H_0^1(\Omega)$, $\quad a(u,v) = \int_\Omega \nabla u \cdot \nabla v$, $\quad F(v) = \int_\Omega f v$, the previous problem (1.3) can be written in abstract form:

$$\text{find} \quad u \in V: \quad a(u,v) = F(v) \quad \forall\, v \in V. \tag{1.4}$$

We will see that many other problems can be set in the abstract form (1.4) with V a Hilbert space.

Formulation (1.3) can also be obtained following another path, from a minimization principle. We define the energy functional

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

and we observe that the functional is well defined for functions $v \in H^1(\Omega)$ and forcing term $f \in L^2(\Omega)$. We then set the minimization problem:

$$\text{find} \quad u = \operatorname*{argmin}_{w\in H_0^1(\Omega)} J(w). \tag{1.5}$$

This problem is indeed equivalent to (1.3). The optimality conditions are given by the Euler-Lagrange equations and correspond to

$$\frac{\partial}{\partial \epsilon} J(u+\epsilon v) = 0 = \int_\Omega \nabla u \cdot \nabla v - \int_\Omega f v \quad \forall\, v \in H_0^1(\Omega).$$

We see from here that the space $H_0^1(\Omega)$ of test functions appears naturally as the space of functions with bounded energy and such that $u + \epsilon v$ satisfies the correct boundary conditions.

We remark, however, that not all problems in weak form can be derived as a minimization of a proper energy functional.

To establish the well posedness of problem (1.3), we recall here the important Lax-Milgram theorem. Let

- V be a Hilbert space.
- $F : V \to \mathbb{R}$ be a linear bounded form (functional), i.e.

$$F(\alpha v_1 + \beta v_2) = \alpha F(v_1) + \beta F(v_2)$$
$$\|F\|_{V'} := \sup_{v \in V} \frac{|F(v)|}{\|v\|_V} < +\infty$$

  where V' is the dual space of V.

- $a : V \times V \to \mathbb{R}$ be a bilinear, continuous, coercive form, i.e.

$$a(\alpha u_1 + \beta u_2, v) = \alpha a(u_1, v) + \beta a(u_2, v)$$
$$\text{similarly} \quad a(u, \alpha v_1 + \beta v_2) = \alpha a(u, v_1) + \beta a(u, v_2)$$
$$\exists\, M > 0 : \quad a(u,v) \le M \|u\|_V \|v\|_V \quad \text{(continuity)}$$
$$\exists\, \alpha > 0 : \quad a(u,u) \ge \alpha \|u\|_V^2 \quad \text{(coercivity)}.$$

Theorem 1.1 (Lax-Milgram theorem)**.** *Given a Hilbert space V, a linear bounded functional F and a bilinear, continuous, coercive form a, the problem*

$$\textit{find} \quad u \in V : a(u,v) = F(v) \quad \forall\, v \in V \tag{1.6}$$

admits a unique solution. Moreover such solution satisfies the stability property

$$\|u\|_V \le \frac{1}{\alpha} \|F\|_{V'}, \tag{1.7}$$

For the proof see e.g. [8]. Inequality (1.7) follows immediately from the coercivity of a and boundedness of F:

$$\alpha \|u\|_V^2 \le a(u,u) = F(u) \le \|F\|_{V'} \|u\|_V.$$

We now apply Theorem 1.1 to problem (1.3). Verifying the hypotheses of the theorem is straightforward:

$$\text{continuity of } a: \quad a(u,v) = \int_\Omega \nabla u \cdot \nabla v \le \|\nabla u\|_{L^2} \|\nabla v\|_{L^2} = \|u\|_{H_0^1} \|v\|_{H_0^1};$$

$$\text{coercivity of } a: \quad a(u,u) = \int_\Omega |\nabla u|^2 = \|u\|_{H_0^1}^2;$$

$$\text{continuity of } F: \quad (\text{assuming } f \in L^2(\Omega)) : F(v) = \int_\Omega f v \le \|f\|_{L^2} \|v\|_{L^2} \le C_p \|f\|_{L^2} \|v\|_{H_0^1}.$$

It follows that problem (1.3) has a unique solution in $H_0^1(\Omega)$ that satisfies $\|u\|_{H_0^1} \leq C_p \|f\|_{L^2}$.

1.1.3 Mixed boundary conditions

We consider now the Poisson equation with mixed boundary conditions

$$\begin{cases} -\Delta u = f & \text{in} \quad \Omega \\ u = g & \text{on} \quad \Gamma_D \\ \partial_n u = d & \text{on} \quad \Gamma_N \end{cases} \tag{1.8}$$

where the boundary $\partial\Omega$ is partitioned in two non overlapping portions, i.e. $\partial\bar{\Omega} = \bar{\Gamma}_D \cup \bar{\Gamma}_N$ and $\Gamma_D \cap \Gamma_N = \emptyset$. We call the condition $u = g$ *essential or Dirichlet* and the condition $\partial_n u = d$ *natural or Neumann.*

It is natural in this case to look for the solution u in the *affine* space $V_g = \{v \in H^1(\Omega) : v|_{\Gamma_D} = g\}$. On the other hand, the "virtual displacements" have to be compatible with the boundary conditions, i.e. $u + v \in V_g$ for any virtual displacement. This implies the condition $v|_{\Gamma_D} = 0$. We define therefore the space

$$V_0 = H^1_{\Gamma_D}(\Omega) = \{v \in H^1(\Omega) : v|_{\Gamma_D} = 0\}.$$

Observe that V_0 is a closed subspace of $H^1(\Omega)$. On the other hand, V_g is not a subspace since

$$\forall\ u_1, u_2 \in V_g \quad \Longrightarrow \quad u_1 + u_2 \in V_{2g}.$$

To derive the weak formulation we proceed in the usual way: multiply the equation by $v \in V_0$, integrate over the domain and use integration by parts for the 2^{nd} derivatives:

$$\Rightarrow \int_\Omega \nabla u \cdot \nabla v - \int_{\partial\Omega} \partial_n u\, v = \int_\Omega f v \quad \forall v \in V_0 (= H^1_{\Gamma_D}(\Omega)).$$

We observe now that

$$\int_{\partial\Omega} \partial_n u\, v = \int_{\Gamma_D} \partial_n u\, v + \int_{\Gamma_N} \partial_n u\, v = 0 + \int_{\Gamma_N} d\, v$$

The term on Γ_D vanishes since we have chosen test functions v such that $v|_{\Gamma_D} = 0$. On the other hand, on Γ_N the normal derivative of the solution is known, so what remains is a known term. The weak formulation reads therefore

$$\text{Find } u \in V_g : \qquad \int_\Omega \nabla u \cdot \nabla v = \int_\Omega f v + \int_{\Gamma_N} d\, v \quad \forall v \in V_0. \tag{1.9}$$

The same weak formulation can be derived from a minimization principle: we define the energy functional

$$J(v) = \frac{1}{2} \int_\Omega |\nabla v|^2 - \int_\Omega f v - \int_{\Gamma_N} d\, v.$$

Observe that, this time, the functional has to include the work done by the boundary forces on Γ_N. Then u satisfies

$$u = \underset{w \in V_g}{\operatorname{argmin}}\, J(w). \tag{1.10}$$

Observe also that the minimization is done on the constrained space V_g of functions that satisfy the non homogeneous Dirichlet boundary condition $w = g$ on Γ_D. Again, we can take $w = u + \epsilon v$ with $u \in V_g$, the solution to (1.10) and $v \in V_0$ so that $u + \epsilon v \in V_g$ for all $\epsilon \in \mathbb{R}$ and write the variations

$$\frac{\partial}{\partial \epsilon} J(u + \epsilon v) = \int_\Omega \nabla u \cdot \nabla v - \int_\Omega f v - \int_{\Gamma_N} d\, v = 0$$

from which we find back formulation (1.9).

The weak form (1.9) can be put in the abstract form

$$\text{find} \quad u \in V_g \quad \text{such that} \quad a(u, v) = F(v) \quad \forall\ v \in V_0$$

with $a(u, v) = \int_\Omega \nabla u \cdot \nabla v$ and $F(v) = \int_\Omega f v + \int_{\Gamma_N} dv$. Here $V = H^1(\Omega)$ is a Hilbert space, $V_0 \subset V$ is a closed subspace of V (and hence it is also a Hilbert space), $V_g \subset V$ is an affine subspace of V.

The Lax-Milgram theorem cannot be applied straightforwardly to show the well posedness of (1.9) since the solution u is sought in an affine space different from that of test functions. However, one can proceed in the following way: assuming that $g \in H^{1/2}(\Gamma_D)$, i.e. it is the trace of some function $G \in H^1(\Omega)$ such that $\|G\|_{H^1(\Omega)} \leq \gamma_\Gamma \|g\|_{H^{1/2}(\Gamma_D)}$, then we can write the problem for the unknown $\mathring{u} = u - G \in V_0$

$$\text{find } \mathring{u} \in V_0 : \quad \underbrace{\int_\Omega \nabla \mathring{u} \nabla v}_{a(\mathring{u}, v)} = \underbrace{\int_\Omega f v + \int_{\Gamma_N} d\, v - \int_\Omega \nabla G \cdot \nabla v}_{\tilde{F}(v)} \quad \forall v \in V_0.$$

Hence, in abstract form:

$$\text{find} \quad \mathring{u} \in V_0 \quad \text{s.t.} \quad a(u, v) = \tilde{F}(v) \quad \forall\ v \in V_0.$$

To apply the Lax-Milgram theorem, we observe that

- the Poincaré inequality holds also in $H^1_{\Gamma_D}$ as long as $|\Gamma_D| > 0$.

  $$\forall\ v \in H^1_{\Gamma_D}(\Omega) \quad \|v\|_{L^2(\Omega)} \leq C_p \|\nabla v\|_{L^2(\Omega)}.$$

  Hence, $\|\nabla u\|_{L^2(\Omega)} = \|u\|_{H^1_0(\Omega)}$ is equivalent to the full norm $\|u\|_{H^1(\Omega)}$ and $a(u, u) = \|u\|^2_{H^1_0(\Omega)}$ is coercive. Continuity is also immediate.

- the functional $\tilde{F}(v)$ is bounded as long as $f \in L^2(\Omega)$ and $d \in L^2(\Gamma_N)$. Indeed

  $$\begin{aligned} |\tilde{F}(v)| \leq & \left| \int_\Omega f v \right| + \left| \int_{\Gamma_N} d\, v \right| + \left| \int_\Omega \nabla G \cdot \nabla v \right| \\ \leq & \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} + \|d\|_{L^2(\Gamma_N)} \|v\|_{L^2(\Gamma_N)} + \|\nabla G\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} \\ \leq & \left(C_p \|f\|_{L^2(\Omega)} + C_T \sqrt{1 + C_p^2}\, \|d\|_{L^2(\Gamma_N)} + \gamma_\Gamma \|g\|_{H^{1/2}(\Gamma_D)} \right) \|\nabla v\|_{L^2(\Omega)} \end{aligned}$$

  where we have used the trace inequality $\|v\|_{L^2(\Gamma_N)} \leq C_T \|v\|_{H^1(\Omega)}$.

We have therefore the following:

Lemma 1.2. *Given $f \in L^2(\Omega),\ d \in L^2(\Gamma_N),\ g \in H^{1/2}(\Gamma_D)$, the problem*

$$\textit{find } u \in V_g \textit{ s.t. } \int_\Omega \nabla u \cdot \nabla v = \int_\Omega f v - \int_{\Gamma_N} d\, v \quad \forall\ v \in V_0$$

admits a unique solution.

The assumptions on the data can actually be weakend. We recall the definition of the dual space. Let V be a Banach space, denote by $V^{'}$ the space of all linear bounded functionals on V, i.e. $F : V \to \mathbb{R}$ s.t.

$$F(\alpha v_1 + \beta v_2) = \alpha F(v_1) + \beta F(v_2), \qquad |F(v)| \leq c\|v\|_V.$$

$V^{'}$ is a Banach space with respect to the norm

$$\|F\|_{V'} = \sup_{v \in V, v \neq 0} \frac{|F(v)|}{\|v\|_V}.$$

With the above definitions, problem (1.9) is well posed for any $f \in V_0'$.

1.1.4 Pure Neumann problem

$$\begin{cases} -\Delta u = f & \text{in} \quad \Omega \\ \partial_n u = d & \text{on} \quad \partial\Omega \end{cases} \tag{1.11}$$

It is easy to realize that in this case if u is a solution to (1.11) then $u + c$ with c an arbitrary constant, is also a solution. Therefore, the solution is not unique. On the other hand, if we integrate the equation and use the Gauss theorem and the boundary conditions we get

$$\int_\Omega f = \int_\Omega -\Delta u = -\int_{\partial\Omega} \partial_n u = -\int_{\partial\Omega} d, \qquad \Longrightarrow \qquad \int_\Omega f + \int_{\partial\Omega} d = 0. \tag{1.12}$$

We see that the data have to satisfy the compatibility condition (1.12). From a mechanical point of view, this condition corresponds to requiring that the resultant of all forces applied to the system is zero.

To prove well posedness one can set the problem in the quotient space $H^1(\Omega)\backslash\mathbb{R} = \{v \in H^1(\Omega) : \int_\Omega v = 0\}$. One can show that in the quotient space a Poincaré inequality

$$\|v\|_{L^2(\Omega)} \leq c\|\nabla v\|_{L^2(\Omega)} \quad \forall\ v \in H^1(\Omega)\backslash\mathbb{R}$$

still holds. Hence the bilinear form is coercive and the problem is well posed under condition (1.12).

1.1.5 Regularity of the solution

We ask now the question whether the solution u has some extra regularity than simply $u \in H^1$. For the pure Dirichlet and pure Neumann problems, the following result holds.

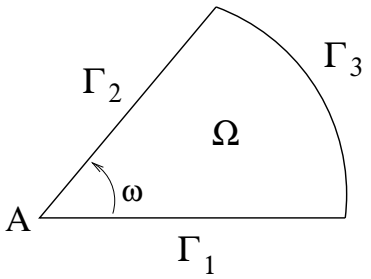
Theorem 1.3 (Shift theorem)**.** *For $m \geq 0$, assume the domain has smooth boundary $\partial\Omega \in C^{m+2}$, $f \in H^m(\Omega)$, $g \in H^{m+3/2}(\partial\Omega)$ (for the pure Dirichlet problem) or $d \in H^{m+1/2}(\partial\Omega)$ (for the pure Neumann problem). Then, $u \in H^{m+2}(\Omega)$.*

Two remarks are in order:

- If the data are smooth (where smoothness is measured in a proper Sobolev norm), the solution is smooth. This result is however true only if the domain has also suitable smoothness. Problems defined in domains with corners are problematic.
- The result holds for pure Dirichlet and pure Neumann problems. The mixed case is more problematic and the solution might not be smooth even with smooth data and domain.

2D Domains with corners

Consider the domain in the figure, having a corner of angle ω, with $\frac{\pi}{\omega} \notin \mathbb{N}$.



The solution of the Poisson equation $-\Delta u = f$ in Ω will develop a singularity at the point A (i.e. some derivatives will go to infinity in A). Locally, around the corner A the solution behaves as $u(r,\theta) = r^\alpha f(\theta)$ for some $\alpha > 0$, where (r,θ) are the polar coordinates centered in A.

Dirichlet problem: $u = 0$ on $\partial\Omega$.

The corner singularities have the form

$$\Phi_k(r,\theta) = r^{k\pi/w} \sin\left(\frac{k\pi\theta}{w}\right), \quad k \in \mathbb{N} \quad \Longrightarrow \quad u \in H^s \text{ with } s < 1 + \frac{\pi}{w}$$

In particular for a re-entrant corner $\omega > \pi$, the solution $u \notin H^2(\Omega)$!

Neumann problem: $\partial_n u = 0$ on $\partial\Omega$.

The corner singularities have the form

$$\Phi_k(r,\theta) = r^{k\pi/w} \cos\left(\frac{k\pi\theta}{w}\right), \quad k \in \mathbb{N} \quad \Longrightarrow \quad u \in H^s \text{ with } s < 1 + \frac{\pi}{w}.$$

Also in this case, for a re-entrant corner $\omega > \pi$, the solution $u \notin H^2(\Omega)$!

Mixed problem: $u = 0$ on Γ_1, $\partial_n u = 0$ on Γ_2 and $\frac{\pi}{2\omega} \notin \mathbb{N}$.

The corner singularities have the form

$$\Phi_k(r,\theta) = r^{(k+\frac{1}{2})\pi/w} \sin((k+\frac{1}{2})\pi\theta/w) \quad \Longrightarrow \quad u \in H^s \text{ with } s < 1 + \frac{\pi}{2w}$$

Even for a flat boundary $w = \pi$, the solution $u \notin H^2$!

1.2 Advection-diffusion reaction equation

We consider now the general second order linear elliptic equation $Lu = f$ with

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

with $a_{ij} = a_{ji}$. We can introduce the matrix field $A : \Omega \to \mathbb{R}^{d\times d}$ as $(A(x))_{ij} = a_{ij}(x)$ and the vector field $\vec{b} : \Omega \to \mathbb{R}^d$; $\vec{b}(x)_i = b_i(x)$.

The operator L can then be written in compact form as

$$Lu = -\mathrm{div}(A\nabla u) + \vec{b}\cdot\nabla u + c\,u. \tag{1.13}$$

The operator L is said to be elliptic if the matrix A is positive definite, i.e. there exists $\alpha_0 > 0$ such that

$$\vec{\xi}^T A(x)\vec{\xi} \geq \alpha_0 |\vec{\xi}|^2 \quad \forall\ \vec{\xi} \in \mathbb{R}^d,\ \forall\ x \in \Omega.$$

When applying the divergence theorem to the first term

$$\int_\Omega -\mathrm{div}(A\nabla u) = \int_{\partial\Omega} (A\nabla u)\cdot n$$

we see that the natural "flux" appearing on the boundary is $(A\nabla u)\cdot n$. Therefore the natural boundary conditions (Neumann b.cs) will be of the type $(A\nabla u)\cdot n = d$.

The advection-diffusion-reaction problem with mixed boundary conditions reads therefore

$$\begin{cases} -\mathrm{div}(A\nabla u) + \vec{b}\cdot\nabla u + c\,u = f & \text{in } \Omega, \\ (A\nabla u)\cdot n = d & \text{on } \Gamma_N, \\ u = g & \text{on } \Gamma_D. \end{cases} \tag{1.14}$$

The weak formulation can be obtained with the usual procedure. Since only second order derivatives appear in the operator, the natural functional setting is again $H^1(\Omega)$. Observe that whenever $\vec{b} \neq 0$, the weak formulation cannot be derived from a minimization principle (the resulting bilinear form is not symmetric).

Weak formulation: find $u \in V_g$ such that

$$\underbrace{\int_\Omega A\nabla u\cdot\nabla v + \int_\Omega \vec{b}\cdot\nabla u\, v + \int_\Omega c\,u\,v}_{a(u,v)} = \underbrace{\int_\Omega f v + \int_{\Gamma_N} d\,v}_{F(v)} \quad \forall\ v \in V_0. \tag{1.15}$$

The continuity of $a(\cdot,\cdot)$ is guaranteed for any $A, b, c \in L^\infty(\Omega)$. Indeed, denoting $\sigma_A(x) = |A(x)| = \sup_{\xi\in\mathbb{R}^d} \frac{|A(x)\vec{\xi}|}{|\vec{\xi}|}$ and $|\vec{b}(x)| = \sqrt{\sum_i b_i(x)^2}$,

$$\begin{aligned} a(u,v) &\le \int_\Omega |(A\nabla u)\cdot\nabla v| + \int_\Omega |\vec{b}\cdot\nabla u||v| + \int_\Omega |c\,u\,v| \\ &\le \int_\Omega \sigma_A(x)|\nabla u|\,|\nabla v| + \int_\Omega |\vec{b}(x)|\,|\nabla u|\,|v| + \int_\Omega |c(x)|\,|u|\,|v| \\ &\le \sup_{x\in\Omega}\sigma_A(x)\|\nabla u\|_{L^2(\Omega)}\|\nabla v\|_{L^2(\Omega)} + \sup_{x\in\Omega}|\vec{b}(x)|\|\nabla u\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \\ &+ \sup_{x\in\Omega}|c(x)|\|u\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \\ &\le \Big(\|A\|_{L^\infty(\Omega)} + \|\vec{b}\|_{L^\infty(\Omega)} + \|c\|_{L^\infty(\Omega)}\Big)\,\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)} \end{aligned}$$

where $\|c\|_{L^\infty(\Omega)} = \operatorname{supess}_{x\in\Omega}|c(x)|$, $\|b\|_{L^\infty(\Omega)} = \operatorname{supess}_{x\in\Omega}|\vec{b}(x)|$ and $\|A\|_{L^\infty(\Omega)} = \operatorname{supess}_{x\in\Omega}\sigma_A(x)$.

To show coerciveness we need further assumptions on A, b, c. Observe that the form $a_s(u,v) = \int_\Omega A\nabla u\cdot\nabla v$ is coercive if $|\Gamma_D| > 0$

$$a_s(u,u) = \int_\Omega (A\nabla u)\cdot\nabla u = \int_\Omega (\nabla u)^T A\nabla u \ge \alpha_0\int_\Omega |\nabla u|^2 = \alpha_0\|\nabla u\|^2_{L^2(\Omega)}.$$

Sufficient conditions for the well posedness of the problem are the following (verify as an exercise):

- $f \in L^2(\Omega)$ (or V_0'), $d \in L^2(\Gamma_N)$, $g \in H^{1/2}(\Gamma_D)$
- $A \in [L^\infty(\Omega)]^{d\times d}$, $\vec{\xi}^T A(x)\vec{\xi} \ge \alpha_0|\vec{\xi}|^2 \quad \forall\, \vec{\xi}\in\mathbb{R}^d,\ x\in\Omega$
- $b \in [L^\infty(\Omega)]^d$, $\operatorname{div}\vec{b} \in L^\infty(\Omega)$, $c \in L^\infty(\Omega)$
- $\operatorname{infess}_{x\in\Omega}(c(x) - \frac{1}{2}\operatorname{div}\vec{b}(x)) \ge 0$ if $|\Gamma_D| > 0$ (strict inequality if $|\Gamma_D| = 0$)
- $\partial\Omega^- = \{x\in\partial\Omega : \vec{b}(x)\cdot\vec{n}(x) < 0\} \subset \Gamma_D$, where $\vec{n}(x)$ is the unit outward normal vector in $x\in\partial\Omega$.

1.3 Linear infinitesimal elasticity

Consider a body occupying the domain $\Omega\subset\mathbb{R}^d$ that undergoes a small (infinitesimal) deformation under the action of a force field $\vec{f}:\Omega\to\mathbb{R}^d$. Denote by $\vec{u}:\Omega\to\mathbb{R}^d$ the deformation field of each material point (see Figure1.1).

Strain measure

Given two material points close to each other $\vec{X}_1 = \vec{X}$, $\vec{X}_2 = \vec{X} + d\vec{X}$, after deformation they will occupy the positions $\vec{x}_1 = \vec{X}_1 + \vec{u}(\vec{X}_1)$ and $\vec{x}_2 = \vec{X}_2 + \vec{u}(\vec{X}_2)$, respectively. The change of

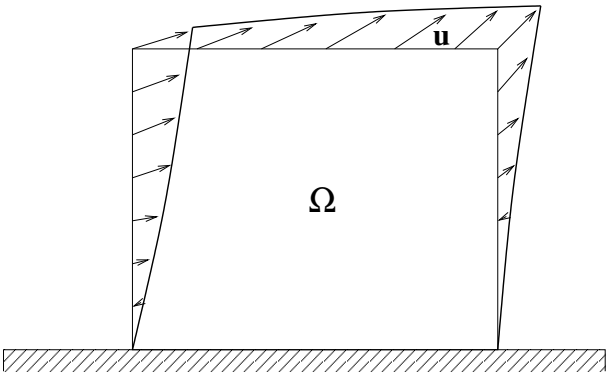


Figure 1.1: Deformation of a bi-dimensional material body. $\vec{u} : \Omega \to \mathbb{R}^2$ represents the deformation field

length is given by

$$\begin{aligned}\frac{|\vec{x}_2 - \vec{x}_1|^2}{|\vec{X}_2 - \vec{X}_1|^2} =& \frac{|\vec{X}_2 + \vec{u}(\vec{X}_2) - \vec{X}_1 - \vec{u}(\vec{X}_1)|^2}{|d\vec{X}|^2} \approx \frac{|d\vec{X} + \nabla\vec{u}(\vec{X}_1) \cdot d\vec{X}|^2}{|dX|^2} = \frac{|(I + \nabla\vec{u})d\vec{X}|^2}{|dX|^2} \\ =& \vec{v}^T(I + \nabla\vec{u})^T(I + \nabla\vec{u})\vec{v} \quad \text{having set} \quad \vec{v} = \frac{d\vec{X}}{|d\vec{X}|} \\ =& \vec{v}^T(I + \nabla\vec{u} + (\nabla\vec{u})^T + (\nabla\vec{u})^T\nabla\vec{u})\vec{v}.\end{aligned}$$

For infinitesimal displacements we neglect the quadratic term $(\nabla\vec{u})^T\nabla\vec{u}$. The relative elongation is thus given by

$$\frac{|\vec{x}_2 - \vec{x}_1|^2 - |d\vec{X}|^2}{|d\vec{X}|^2} \approx \vec{v}^T(\nabla\vec{u} + (\nabla\vec{u})^T)\vec{v}$$

In infinitesimal elasticity it is customary to take as a measure of strain the tensor

$$\varepsilon(\vec{u}) = \frac{\nabla\vec{u} + (\nabla\vec{u})^T}{2}.$$

Observe that if $\vec{u}(\vec{X})$ is a pure rotation or translation, $\vec{u}(\vec{X}) = \vec{c} + \vec{w} \times (\vec{X} - \vec{X}_0)$ then the strain is zero.

Stress tensor

$\sigma : \Omega \to \mathbb{R}^{d\times d}$ represents the internal stresses due to the deformation. In linear infinitesimal elasticity the stress tensor is related linearly to the strain tensor,

$$\sigma_{ij} = \sum_{kl} C_{ijkl}\varepsilon_{kl}.$$

This is the tensorial equivalent of the well-known Hooke's law for the elongation ΔL of a spring under a force F, $F = -k\Delta L$.

If a material is isotropic, i.e. its elastic properties are the same in each directions, many of the coefficients C_{ijkl} will be equal, and it is possible to simplify the previous strain-stress relation. It can actually be shown that the most general linear relation that can be written for an isotropic material has the form

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

where μ and λ are known as the Lamé constants and $I \in \mathbb{R}^{d\times d}$ is the identity matrix.

Balance equations

The balance of volumetric forces translates into the equation

$$-\operatorname{div}\sigma = \vec{f}.$$

Indeed, on any volume $\omega \subset \Omega$, $\int_\omega -\operatorname{div}\sigma = \int_{\partial\omega} -\sigma\cdot\vec{n}$ represents the resultant of the internal forces acting on the volume ω, which has to balance the volume forces $\int_\omega \vec{f}$ acting on it.

Infinitesimal linear elasticity equations with mixed boundary conditions

$$\begin{cases} -\operatorname{div}\sigma(\vec{u}) = \vec{f}, & \text{in } \Omega \\ \sigma(\vec{u})\cdot\vec{n} = \vec{d}, & \text{on } \Gamma_N \\ \vec{u} = \vec{g}, & \text{on } \Gamma_D \end{cases} \tag{1.16}$$

$$\text{with } \sigma(\vec{u}) = 2\mu\varepsilon(\vec{u}) + \lambda \operatorname{tr}\varepsilon(\vec{u}) I, \quad \text{and } \varepsilon(\vec{u}) = \frac{\nabla\vec{u} + (\nabla\vec{u})^T}{2}. \tag{1.17}$$

Here $\vec{f}$ represents the volumetric forces (e.g. the weight), $\vec{d}$ represents the boundary traction and $\vec{g}$ represents the imposed deformation on the boundary (typically $\vec{g} = 0$ for a clamped structure).

Weak formulation

To write the weak formulation, we need some additional notation. Given two equal sized matrices A and B, we define $A : B = \sum_{ij} A_{ij}B_{ij} = \operatorname{tr}(A^T B) = \operatorname{tr}(B^T A)$.

We can now take a smooth test function $\vec{v} : \Omega \to \mathbb{R}^d$ with $\vec{v}|_{\Gamma_D} = 0$ and proceed as usual

$$\begin{aligned} \int_\Omega -\operatorname{div}\sigma(\vec{u})\cdot\vec{v} &= \int_\Omega \sigma(\vec{u}) : \nabla\vec{v} - \int_{\partial\Omega} (\sigma(\vec{u})\cdot\vec{n})\cdot\vec{v} \\ &= \int_\Omega (2\mu\varepsilon(\vec{u}) + \lambda \operatorname{tr}\varepsilon(\vec{u}) I) : \nabla\vec{v} - \int_{\Gamma_N} \vec{d}\cdot\vec{v} \\ &= \int_\Omega 2\mu\varepsilon(\vec{u}) : \nabla\vec{v} + \int_\Omega \lambda \operatorname{div}\vec{u} \operatorname{div}\vec{v} - \int_{\Gamma_N} \vec{d}\cdot\vec{v} \\ &= \int_\Omega 2\mu\varepsilon(\vec{u}) : \varepsilon(\vec{v}) + \int_\Omega \lambda \operatorname{div}\vec{u} \operatorname{div}\vec{v} - \int_{\Gamma_N} \vec{d}\cdot\vec{v} \end{aligned}$$

where we have used the following observations:

- $\varepsilon(\vec{u}) : \nabla\vec{v} = \varepsilon(\vec{u}) : \varepsilon(\vec{v})$. In fact, since $\varepsilon(\vec{u})$ is a symmetric tensor we can take only the symmetric part of $\nabla\vec{v}$, i.e. $\frac{\nabla\vec{v}+(\nabla\vec{v})^T}{2} = \varepsilon(\vec{v})$

- $\operatorname{tr}\varepsilon(\vec{u}) = \operatorname{div}\vec{u}$
- $I : \nabla\vec{v} = \operatorname{div}\vec{v}$.

Again, after integration by parts only first derivatives appear, so that a natural functional space is $[H^1(\Omega)]^d$, i.e.

$$[H^1(\Omega)]^d = \{\vec{v} : \Omega \to \mathbb{R}^d : v_i \in H^1(\Omega) \quad \text{for } i = 1, \dots, d\}.$$

Each component of the vector is an H^1 function, so we can define traces and in particular the space $V_g = \{\vec{v} \in [H^1(\Omega)]^d : \ \vec{v}|_{\Gamma_D} = \vec{g}\}$. We are thus lead to the following weak form of (1.16):

$$\text{Find } \vec{u} \in V_g \text{ such that} \quad a(\vec{u}, \vec{v}) = F(\vec{v}) \quad \forall \vec{v} \in V_0 \tag{1.18}$$

where

$$a(\vec{u}, \vec{v}) = \int_\Omega (2\mu\varepsilon(\vec{u}) : \varepsilon(\vec{v}) + \lambda \operatorname{div}\vec{u} \operatorname{div}\vec{v})$$
$$F(\vec{v}) = \int_\Omega \vec{f}\cdot\vec{v} + \int_{\Gamma_N} \vec{d}\cdot\vec{v}$$

The continuity of $a(\cdot,\cdot)$ is quite straightforward and left as an exercise. The coerciveness of $a(\cdot,\cdot)$ is not obvious a priori. Observe in particular that $\varepsilon(\vec{u})$ (and $\operatorname{div}(\vec{u})$ as well) vanishes on the space of roto-translatory motions

$$V_{\mathrm{RT}} = \{\vec{v} : \Omega \to \mathbb{R}^d : \ \vec{v}(\vec{x}) = \vec{c} + \vec{w} \times \vec{x}, \ \text{ with } \vec{c}, \vec{w} \in \mathbb{R}^d\}$$

therefore $a(u, u) = 0 \quad \forall\ u \in V_{\mathrm{RT}}$. We should therefore not expect coerciveness $\alpha\|\vec{u}\|^2_{H^1(\Omega)} \leq a(\vec{u}, \vec{u})$ unless $\vec{u}$ is orthogonal to V_{RT}. The following result asserts that it is actually enough to remove all possible functions in V_{RT} from the space where we look for the solution to ensure coerciveness. In particular, rototranslations are excluded if we enforce no deformation on a portion of $\partial\Omega$, i.e. if we work in $[H^1_{\Gamma_D}(\Omega)]^d$:

Theorem 1.4 (Korn inequality)**.** *There exists $\kappa > 0$ such that*

$$\|\nabla\vec{v}\|^2_{[L^2(\Omega)]^d} \leq \kappa \int_\Omega \varepsilon(\vec{v}) : \varepsilon(\vec{v}), \quad \forall \vec{v} \in [H^1_{\Gamma_D}(\Omega)]^d \tag{1.19}$$

Thanks to the Korn and Poincaré inequalities, one can show that all assumptions of the Lax-Milgram theorem are satisfied and therefore problem (1.18) is well posed.

Finally, since the bilinear form $a(\cdot,\cdot)$ is symmetric, the weak formulation can be derived from a minimization principle

$$\vec{u} = \operatorname*{argmin}_{\vec{w} \in V_g} J(\vec{w})$$

with energy functional

$$J(\vec{w}) = \int_\Omega \mu\varepsilon(\vec{w}) : \varepsilon(\vec{w}) + \int_\Omega \frac{\lambda}{2}(\operatorname{div}\vec{w})^2 - \int_\Omega \vec{f}\cdot\vec{w} - \int_{\Gamma_N} \vec{d}\cdot\vec{w}. \tag{1.20}$$

Chapter 2

Approximation of variational problems – Galerkin method

All the problems that we have introduced so far can be recast (after eventually a suitable lifting of the Dirichlet boundary datum) in the following abstract form

$$\text{find} \quad u \in V \quad \text{such that} \quad a(u,v) = F(v) \quad \forall\, v \in V \tag{2.1}$$

with V a Hilbert space. We assume here that the assumptions of the Lax-Milgram theorem are satisfied so that problem (2.1) admits a unique solution. To approximate problem (2.1) we proceed as follows:

1. We introduce a sequence of finite dimensional spaces V_h with $N_h = \dim V_h$ such that

$$V_h \subset V \quad \forall\, h \quad \text{(conformity)} \tag{2.2}$$

$$\forall\, w \in V \quad \lim_{h\to 0} \left(\inf_{w_h \in V_h} \|w - w_h\|_V \right) = 0 \quad \text{(approximability)}. \tag{2.3}$$

   The second assumption is essential and says that in the limit $h \to 0$ the space V_h becomes dense in V, i.e. any element in V can be approximated arbitrary well by a sequence $w_h \in V_h$ for $h \to 0$.

2. We set the problem (*Galerkin approximation*):

$$\text{find } u_h \in V_h \text{ s.t.} \quad a(u_h, v_h) = F(v_h) \quad \forall\, v_h \in V_h. \tag{2.4}$$

More generally, we could think of a discrete problem where also the bilinear form $a(\cdot,\cdot)$ is approximated by $a_h(\cdot,\cdot)$, the right hand side $F(\cdot)$ by $F_h(\cdot)$. This will typically be the case when using quadrature formulas to approximate the integrals appearing in the bilinear form $a(\cdot,\cdot)$ and right hand side $F(\cdot)$.

Even further, we could also remove the conformity assumption $V_h \subset V$ and introduce a *Generalized Galerkin approximation*

$$\text{find } u_h \in V_h \text{ s.t.} \quad a_h(u_h, v_h) = F_h(v_h) \quad \forall\, v_h \in V_h \tag{2.5}$$

Some nomenclature

A (generalized) Galerkin approximation is said to be

- **Conforming** if $V_h \subset V$ and **non-conforming** if $V_h \not\subseteq V$.
- **Strongly consistent** if the exact solution satisfies the discrete problem, i.e.

$$a_h(u, v_h) = F_h(v_h) \quad \forall\ v_h \in V_h$$

- **Asymptotically consistent** if the exact solution satisfies the discrete problem only in the limit $h \to 0$, i.e.

$$\lim_{h\to 0} \sup_{v_h \in V_h} \frac{F_h(v_h) - a_h(u, v_h)}{\|v_h\|_V} = 0.$$

  This definition and the previous one are valid if one can extend the bilinear form $a_h(\cdot, \cdot)$ as a continuous form on the whole space V.

2.1 Properties of the conforming Galerkin problem

We consider here the standard (conforming) Galerkin approximation (2.4)

$$\text{find } u_h \in V_h \subset V \text{ s.t.} \quad a(u_h, v_h) = F(v_h) \quad \forall\ v_h \in V_h \subset V.$$

Since $V_h \subset V$ is finite dimensional, it is a closed subspace of V and therefore it is a Hilbert space with respect to the same norm defined in V.

By assumption, $a(\cdot, \cdot)$ is continuous and coercive in V. A fortiori, it will be continuous and coercive in V_h, and the same argument holds for the continuity of F. We conclude that problem (2.4) satisfies all the hypotheses of the Lax-Milgram theorem and therefore admits a unique solution. Moreover, the solution u_h satisfies the stability inequality

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

2.1.1 Reduction to an algebraic system

We now introduce a basis $\{\varphi_j\}_{j=1}^{\mathcal{N}_h}$ of V_h, so that every element $v \in V_h$ can be expanded as $v = \sum_{i=1}^{\mathcal{N}_h} v_i \varphi_i$. If we define the vector $\vec{v} = (v_1, \ldots, v_{\mathcal{N}_h}) \in \mathbb{R}^{\mathcal{N}_h}$, we can establish a bijection between V_h and $\mathbb{R}^{\mathcal{N}_h}$ as

$$v \in V_h \iff \vec{v} = (v_1, \ldots, v_{\mathcal{N}_h}) \in \mathbb{R}^{\mathcal{N}_h} \quad \text{with } v = \sum_{i=1}^{\mathcal{N}_h} v_i \varphi_i.$$

We now expand the solution u_h of (2.4) on the basis $u = \sum_j u_j \varphi_j$ and test the equation (2.4) for all basis functions $\varphi_i, i = 1, \ldots, \mathcal{N}_h$, i.e. we take $v_h = \varphi_i$. It is actually enough to test (2.4) only on the basis functions as all other functions $v \in V_h$ can be obtained by linear combination of the $\{\varphi_i\}$:

$$a\Big(\sum_{j=1}^{\mathcal{N}_h} u_j \varphi_j, \varphi_i\Big) = F(\varphi_i) \quad i = 1, \ldots, \mathcal{N}_h.$$

By the linearity of $a(\cdot,\cdot)$ this is equivalent to

$$\sum_{j=1}^{\mathcal{N}_h} u_j a(\varphi_j,\varphi_i) = F(\varphi_i) \quad i = 1,\dots,\mathcal{N}_h. \tag{2.6}$$

Defining now the matrix $A \in \mathbb{R}^{\mathcal{N}_h\times\mathcal{N}_h}$, $A_{ij} = a(\varphi_j,\varphi_i)$ and the vector $\vec{f} \in \mathbb{R}^{\mathcal{N}_h}$, $f_i = F(\varphi_i)$, then (2.6) is equivalent to

$$A\vec{u} = \vec{f}. \tag{2.7}$$

For the Poisson/linear elasticity problem, the matrix A is typically called the stiffness matrix.

2.1.2 Positivity of the stiffness matrix

The matrix A is *positive definite*. This follows immediately from the coerciveness of $a(\cdot,\cdot)$

$$\begin{aligned}\vec{u}^T A\vec{u} = \sum_{ij} u_j A_{ij} u_i = &\sum_{ij} u_j a(\varphi_j,\varphi_i) u_i \\ =&a\Big(\underbrace{\sum_j u_j\varphi_j}_{u_h}, \sum_i u_i\varphi_i\Big) = a(u_h,u_h) \geq \alpha\|u_h\|_V^2 > 0\end{aligned}$$

Moreover, if the bilinear form a is symmetric, i.e. $a(u,v) = a(v,u)$ for all $u,v \in V$, it follows immediately that the stiffness matrix A is symmetric.

2.2 Convergence analysis of the conforming Galerkin method

We now aim at comparing the exact solution u of (2.1) with the approximate solution u_h of (2.4). We first observe that

$$\begin{aligned} a(u,v) &= F(v) \quad \forall\ v \in V \\ a(u_h,v_h) &= F(v_h) \quad \forall\ v_h \in V_h. \end{aligned}$$

If we take only test functions in V_h in the exact problem and subtract the two, we have

$$a(u-u_h,v_h) = 0 \quad \forall\ v_h \in V_h. \tag{2.8}$$

This relation is called *Galerkin orthogonality*. If $a(\cdot,\cdot)$ is symmetric (and continuous and coercive), it actually defines an inner product equivalent to the standard one defined on V (the proof of this statement is left as an exercise), and the corresponding norm $\|u-u_h\|_a$ is often called "a-norm" or "energy norm". Then equation (2.8) is actually an orthogonality relation, i.e. the function $u-u_h$ is orthogonal to the subspace V_h, with respect to the energy inner product. Consequently, the approximate solution $u_h \in V_h$ is the one for which the distance $\|u-u_h\|_a$ is minimal. Concerning the approximation error, the following result holds:

Lemma 2.1 (Cea's Lemma)**.**

$$\|u-u_h\|_V \leq \frac{M}{\alpha} \inf_{v_h\in V} \|u-v_h\|_V \tag{2.9}$$

Proof.

$$\begin{aligned}\|u-u_h\|_V^2 \leq & \frac{1}{\alpha}a(u-u_h,u-u_h) = \frac{1}{\alpha}a(u-u_h,u-v_h+v_h-u_h) \quad (v_h \in V_h)\\ = & \frac{1}{\alpha}a(u-u_h,u-v_h) + \underbrace{\frac{1}{\alpha}a(u-u_h,v_h-u_h)}_{=0 \text{ (by Galerkin orthogonality)}}\\ \leq & \frac{M}{\alpha}\|u-u_h\|_V\|u-v_h\|_V\end{aligned}$$

from which the thesis follows given the arbitrariness of $v_h \in V_h$. □

The Cea's lemma relates the actual approximation error $\|u-u_h\|_V$ with the

$$\text{best approximation error} \quad \inf_{v_h \in V_h} \|u-v_h\|_V \quad \text{(BAE)}$$

i.e. the best approximation of u that can be achieved in the subspace V_h. This quantity is not related to the differential problem that we are solving but only to the properties of the solution u. Proving approximation rates for the (BAE) for a given class of functions and approximating subspaces V_h is a classical topic of approximation theory. Results for finite element approximation spaces will be given in Chapter 5.

Chapter 3

Finite element spaces

A finite element space is a space of functions that are piecewise polynomials over a partition of the domain Ω into non-overlapping polyhedra, called a *mesh*. Finite element spaces may differ for the polynomial degree used, the type of polyhedra in the mesh and the overall continuity properties between the elements of the partition. In this chapter we assume that the domain $\Omega \subset \mathbb{R}^d$ is a polygon in 2D or a polyhedron in 3D.

3.1 The mesh

Definition 3.1. *A **polyhedral mesh** $\mathcal{T}_h$ is the union of a finite number of polyhedra K_j such that*

- $\bar{\Omega} = \bigcup_{K_j \in \mathcal{T}_h} K_j$
- $\mathring{K}_i \cap \mathring{K}_j = \emptyset$ *if* $i \neq j$.

The polyhedra K_i are called the elements *of the mesh.*

The most used polyhedra are triangles or quadrilaterals in 2D and tetrahedra, hexahedra and sometimes prisms in 3D.

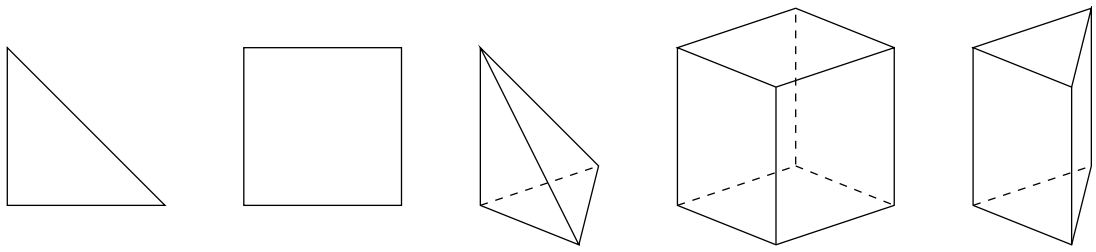


Figure 3.1: Examples of polyhedra typically used in 2D and 3D problems

Definition 3.2. *A **geometrical conformal mesh** is a mesh for which if $E = K_i \cap K_j \neq \emptyset$ then E is a common vertex or a common edge (or a common face in 3D).*

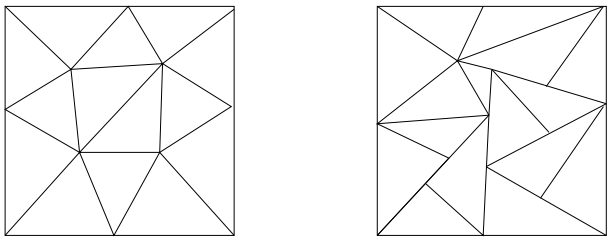


Figure 3.2: Example of a conformal mesh (left) and a non-conformal mesh (right)

Figure 3.2 shows an example of a conformal mesh (left) and a non-conformal mesh (right). We introduce now three important parameters that characterize the elements of a mesh:

- **Element (outer) diameter**:

$$h_K = \text{diam}(K) = \max_{x,y \in K} |x - y|, \qquad K \in \mathcal{T}_h$$

- **Element inner diameter** (also called **chunkiness** or **sphericity**):

$$\rho_K = \text{diameter of the largest ball contained in K}, \qquad K \in \mathcal{T}_h$$

  The **aspect ratio** $\gamma_K = h_K/\rho_K$ is a measure of how much the element K is stretched.

- **Mesh size**:

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

  The parameter h controls the overall size of the elements.

We consider now a family of meshes $\{\mathcal{T}_h\}_{h \searrow 0}$ with smaller and smaller mesh size. The following definitions characterize different types of sequences of meshes.

Definition 3.3. ***Family of regular meshes*** $\{\mathcal{T}_h\}_{h \searrow 0}$*: is a family of meshes for which* $\exists\ \gamma > 1$ *such that*

$$h_K \leq \gamma \rho_K \quad \forall K \in \mathcal{T}_h$$

with γ *independent of* h*.*

In other words, for a sequence of regular meshes, the aspect ratio of each element is bounded by γ uniformly in the family with respect to h.

Definition 3.4. ***Family of quasi-uniform meshes*** $\{\mathcal{T}_h\}_{h \searrow 0}$*: is a family of* regular *meshes for which* $\exists\ 0 < \delta < 1$ *such that*

$$h_K \geq \delta h \quad \forall K \in \mathcal{T}_h$$

with δ *independent of* h*.*

For quasi-uniform meshes, the diameter of the smallest element compares with the diameter of the largest one, i.e. all elements have more or less the same diameter. This is not necessary a nice feature as it prevents from having local mesh refinements. On the contrary, regular meshes do not have this restriction and allow for local mesh refinement, however with a control on the aspect ratio of the elements.

As we will see, approximation properties of a finite element space hold, in general, for regular meshes, i.e. they apply also to the case of highly refined meshes, provided the elements are not too stretched.

Anisotropic meshes, for which the aspect ratio h_K/ρ_K can be large and might go to infinity as $h \to 0$, are also sometimes used, typically to describe boundary layers or sharp gradients of the solution only in certain directions. The theory, however, is more difficult and will not be addressed in these notes. Figure 3.3 shows an example of a quasi-uniform mesh, a regular mesh and an anisotropic mesh.

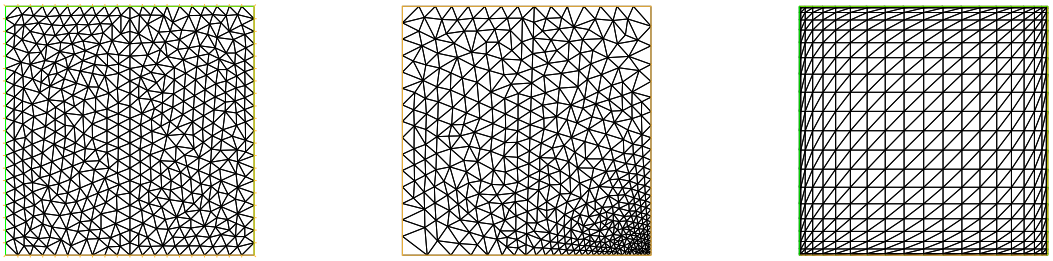


Figure 3.3: Example of a quasi-uniform mesh (left), regular mesh (center) and anisotropic mesh (right).

It is useful to introduce the concept of **reference element** $\hat{K}$. This will be

- for triangular elements: $\hat{K}$ is the triangle of vertices $(0,0)$, $(1,0)$ and $(0,1)$.
- for quadrilateral elements: $\hat{K}$ is the unit square $[0,1]^2$.
- for tetrahedral elements: $\hat{K}$ is the tetrahedron of vertices $(0,0,0)$, $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$.
- for hexahedral elements: $\hat{K}$ is the unit cube $[0,1]^3$.

Definition 3.5. *An **affine mesh** is a mesh for which each element K can be mapped onto the reference element $\hat{K}$ by an affine transformation,*

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

for some matrix $B_K \in \mathbb{R}^{d\times d}$ and vector $b_K \in \mathbb{R}^d$ such that $K = F_K(\hat{K})$.

Some remarks are in order:

- A triangle (tetrahedron in 3D) with straight edges can always be mapped by an affine transformation onto the reference triangle $\hat{K}$.
- A quadrilateral can not be mapped, in general, onto the square $\hat{K} = [0,1]^2$ by an affine transformation, unless it is a parallelogram.

  For a general quadrilateral (with straight edges), the transformation $K = F_K(\hat{K})$ will be linear in each variable but could have quadratic terms xy in 2D or cubic terms xyz in 3D.

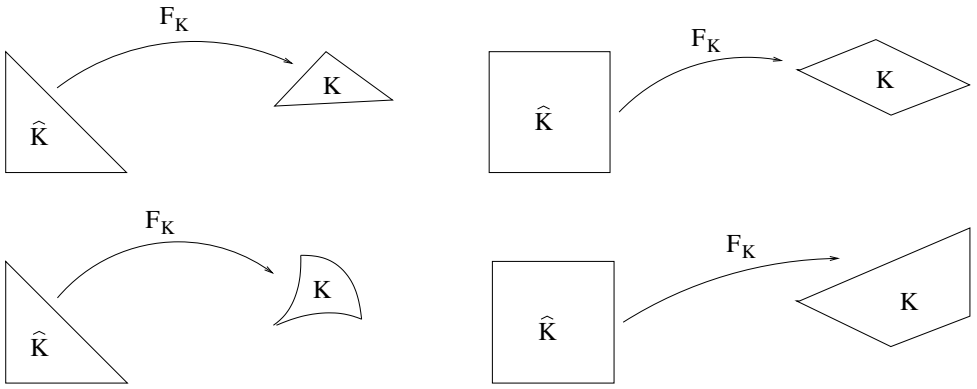


Figure 3.4: Examples of affine transformations (top row) and non-affine transformations (bottom row).

- In some cases also for triangular meshes one is interested in using non affine transformations. This is the case, for instance, to generate "triangles with curved boundaries" to better fit the boundary of the domain. In such a case the map is typically polynomial of degree greater than one.

The figure 3.4 shows examples of affine and non affine transformations from the reference to the current element.

3.1.1 Map to the reference element

We detail here the construction and main properties of the transformation from the reference to the current element in the case of an affine triangular mesh. Let $\hat{K}$ be the reference triangle

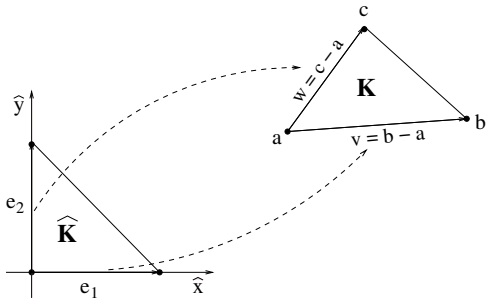


Figure 3.5: Map from reference to current element

of vertices $\{(0,0),(1,0),(0,1)\}$ and $K \in \mathcal{T}_h$ a triangle of the mesh of vertices $\{\vec{a} = (a_1, a_2), \vec{b} = (b_1, b_2), \vec{c} = (c_1, c_2)\}$. If we denote $\vec{v} = \vec{b} - \vec{a}$ and $\vec{w} = \vec{c} - \vec{a}$, we can construct the map $F_K(\hat{x}) = B_K(\hat{x}) + b_K$ in such a way that the canonical vector $e_1 = (1,0)$ is mapped into $\vec{a} + \vec{v}$ and $e_2 = (0,1)$ is mapped into $\vec{a} + \vec{w}$ (see Figure 3.5). Such a map is given by

$$b_K = \vec{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \qquad B_K = \begin{bmatrix} \vec{v} & \vec{w} \end{bmatrix} = \begin{bmatrix} b_1 - a_1 & c_1 - a_1 \\ b_2 - a_2 & c_2 - a_2 \end{bmatrix}$$

The Jacobian matrix satisfies the following properties

Lemma 3.6 (Properties of B_K)**.** *Let h_K and ρ_K be the outer and inner diameter of K and $\hat{h}$ and $\hat{\rho}$ the diameters of $\hat{K}$. Then*

$$\|B_K\| \leq \frac{h_K}{\hat{\rho}}, \qquad \|B_K^{-1}\| \leq \frac{\hat{h}}{\rho_K}, \qquad \det B_K = \frac{|K|}{|\hat{K}|} \tag{3.1}$$

where $\|B_K\| = \sup_{\xi\in\mathbb{R}^2,\xi\neq 0} \frac{|B_K\xi|}{|\xi|}$ is the spectral norm of B_K.

Proof. It is easy to verify that

$$\det B_K = |\vec{v}\times\vec{w}| = \frac{1}{2}|K| = \frac{|K|}{|\hat{K}|}.$$

Moreover, for any $\xi \in \mathbb{R}^2$, $|\xi| = \hat{\rho}$, we can find two points $\hat{x}, \hat{y} \in \hat{K}$ such that $\xi = \hat{x} - \hat{y}$. Letting now $x = F_K(\hat{x})$ and $y = F_K(\hat{y})$, we have

$$|B_K(\hat{x} - \hat{y})| = |x - y| \leq h_K$$

and

$$\|B_K\| = \sup_{\xi\in\mathbb{R}^2,\xi\neq 0} \frac{|B_K\xi|}{|\xi|} = \sup_{\xi\in\mathbb{R}^2,|\xi|=\hat{\rho}} \frac{|B_K\xi|}{|\xi|} \leq \frac{h_K}{\hat{\rho}}.$$

The bound on $\|B_k^{-1}\|$ can be obtained in a similar way. □

3.2 Continuous Finite Elements on triangular affine meshes

A finite element space is a space of piecewise polynomial functions over the elements of a mesh $\mathcal{T}_h$. Let us denote by $\mathbb{P}_r(K)$ the space of polynomial functions in $K \subset \mathbb{R}^d$, of degree less or equal to r:

$$\mathbb{P}_r(K) = \text{span}\{x_1^{k_1}x_2^{k_2}\dots x_d^{k_d}, \quad \sum_{j=1}^{d} k_j \leq r, \quad k_j \geq 0, \quad (x_1,\dots,x_d) \in K\} \tag{3.2}$$

For example:

in 2D	$\mathbb{P}_1(\mathbb{R}^2) = \text{span}\{1, x, y\}$,	$\mathbb{P}_2(\mathbb{R}^2) = \text{span}\{1, x, y, x^2, xy, y^2\}$,
in 3D	$\mathbb{P}_1(\mathbb{R}^3) = \text{span}\{1, x, y, z\}$,	$\mathbb{P}_2(\mathbb{R}^3) = \text{span}\{1, x, y, z, x^2, y^2, z^2, xy, xz, yz\}$.

One can show by combinatorial arguments that the dimension of $\mathbb{P}_r(\mathbb{R}^d)$ is

$$\dim(\mathbb{P}_r(\mathbb{R}^d)) = \binom{r+d}{d}. \tag{3.3}$$

Definition 3.7. *The space of **continuous Finite Elements of degree r over a triangular affine mesh** $\mathcal{T}_h$, hereafter called X_h^r or, sometimes, for brevity simply $\mathbb{P}_r$, is defined for $r \geq 1$ as*

$$X_h^r = \{v \in C^0(\Omega) : \ v|_K \in \mathbb{P}_r(K) \quad \forall K \in \mathcal{T}_h\} \tag{3.4}$$

We remark that the finite element space defined above satisfies the property

$$X_h^r \subset H^1(\Omega).$$

Indeed, a function $v_h \in X_h^r$ is continuous by definition and have bounded (distributional) derivatives since v_h is a polynomial in each element of the mesh. It follows that $v_h, \nabla v_h \in L^2(\Omega)$.

3.2.1 Degrees of freedom, basis functions and interpolation operator

Let us start our study of continuous finite elements by the simplest case of piecewise linear finite elements (often called $\mathbb{P}_1$) in 2D.

$\mathbb{P}_1$ finite elements in 2D

The dimension of $\mathbb{P}_1(\mathbb{R}^2)$ is $\dim(\mathbb{P}_1(\mathbb{R}^2)) = 3$, see (3.3). Therefore, to uniquely identify a linear polynomial in each triangle $K \in \mathcal{T}_h$, we have to provide three values; these could be the three coefficients of the polynomial or its value on three non aligned points. The second option is more common and leads to the so called *Lagrangian finite elements*. The corresponding values of the polynomial in these points are called *nodal degrees of freedom* (dofs).

Let us follow the second approach: as we have just discussed, by prescribing arbitrary values on 3 non-aligned points per triangle, we can uniquely identify a linear function on each triangle. However, by doing so, we have no guarantee that the overall function thus defined is globally continuous over the whole mesh .

There is, however, a particularly clever choice of points that automatically enforces the continuity of the function. This choice corresponds to taking the values of the function on the vertices of each triangle as degrees of freedom (see Figure 3.6). Indeed, on each edge the function is linear and to identify uniquely a linear function on an edge (1D domain) one only needs two point values, e.g. the vertices. Hence, the set of nodal values on the vertices of the mesh uniquely identifies a globally continuous piecewise linear function. Let N_v denote the number of vertices

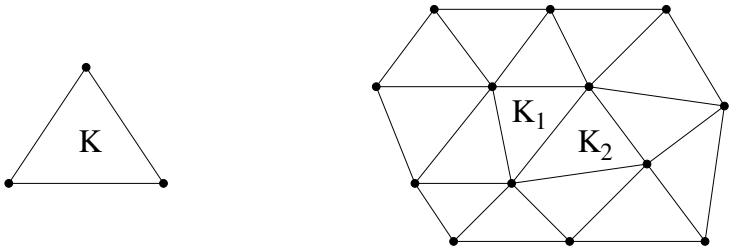


Figure 3.6: Set of nodal degrees of freedom for $\mathbb{P}_1$ finite elements in 2D. The choice of vertices as dofs guarantees automatically the inter-element continuity.

of the mesh $\mathcal{T}_h$ and $\{a_j\}_{j=1}^{N_v}$ the set of vertices. By prescribing arbitrary values $(v_1, \dots, v_{N_v})$ on those vertices, we can construct a unique function $v_h \in X_h^1$ that matches those nodal values, i.e. such that $v_h(a_j) = v_j$. On the other hand, given any function $v_h \in X_h^1$, we can always evaluate it on the vertices and associate to it a unique set of degrees of freedom $v_j = v_h(a_j)$. This shows that there is a one-to-one correspondence between X_h^1 and $\mathbb{R}^{N_v}$ and, in particular,

$$\dim(X_h^1) = N_v.$$

We now aim at constructing a basis for X_h^1. Exploiting the one-to-one correspondence $X_h^1 \leftrightarrow \mathbb{R}^{N_v}$, we could take as basis of X_h^1 the image of the canonical basis of $\mathbb{R}^{N_v}$. We can therefore introduce the following basis of X_h^1, that is usually called *Lagrangian basis*

$$\textbf{basis of } X_h^1: \qquad \{\varphi_j \in X_h^1,\ j = 1, \dots, N_v\}: \ \ \varphi_j(a_k) = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

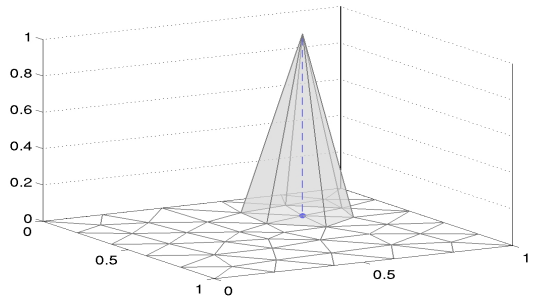


Figure 3.7: A basis function for the space X_h^1.

i.e. the basis function φ_j takes the value 1 on the vertex a_j and 0 on all other vertices (and, of course, it is a globally continuous piecewise linear function on the mesh). Figure 3.7 shows one such basis function.

Any function v_h can be expanded on the basis as

$$v_h(x) = \sum_{i=1}^{N_v} v_j \varphi_j(x), \qquad \text{with } v_j = v_h(a_j).$$

Finally, once we have a set of degrees of freedom and a basis for the space X_h^1, it is easy to introduce an *interpolation operator*, denoted I_h^1, that, given a function $u \in C^0(\Omega)$, associates a function $u_h \in X_h^1$. It is indeed enough to evaluate u on the vertices and reconstruct a continuous piecewise linear interpolation.

$$\textbf{interpolant operator}: \qquad I_h^1 : C^0(\Omega) \to X_h^1, \qquad u_h = I_h^1(u) = \sum_{j=1}^{N_v} u(a_j)\varphi_j. \tag{3.6}$$

$\mathbb{P}_r$ finite elements

We now generalize the above construction to finite elements of arbitrary degree r. Let us consider the case $r = 2$ in 2D. In this case we have

$$\dim(\mathbb{P}_2(\mathbb{R}^2)) = 6, \qquad \dim(\mathbb{P}_2(\mathbb{R}^1)) = 3.$$

Therefore, we need 6 nodal values per triangle to identify uniquely a quadratic function in 2D and 3 nodal values per edge to identify uniquely a quadratic function on a line. Hence, the set of vertices and mid points is unisolvent for $\mathbb{P}_2(K)$ (i.e. identifies uniquely a polynomial of degree 2 on K) and guarantees the global continuity of the function. Denoting by $\{a_j\}_{j=1}^{N_v}$ the set of vertices and $\{c_k\}_{k=1}^{N_e}$ the set of midpoints of each edge, we have

$$\dim(X_h^2) = N_v + N_e.$$

Proceeding as for $\mathbb{P}_1$ finite elements, we can introduce a basis

$$\{\varphi_j \in X_h^2\}_{j=1}^{N_v+N_e} = \{\varphi_j^{(v)} \in X_h^2\}_{j=1}^{N_v} \cup \{\varphi_j^{(e)} \in X_h^2\}_{j=1}^{N_e}$$

with

$$\text{vertex basis function } \begin{cases} \varphi_j^{(v)}(a_j) = 1, & \\ \varphi_j^{(v)}(a_l) = 0 & l \neq j, \\ \varphi_j^{(v)}(c_k) = 0 & \forall k, \end{cases} \qquad \text{edge basis function } \begin{cases} \varphi_j^{(e)}(c_j) = 1, & \\ \varphi_j^{(e)}(c_k) = 0 & l \neq j, \\ \varphi_j^{(e)}(a_l) = 0 & \forall l. \end{cases}$$

In the case $r = 3$ in 2D we have $\dim(\mathbb{P}_3(\mathbb{R}^2)) = 10$ and $\dim(\mathbb{P}_3(\mathbb{R}^1)) = 4$. Therefore we need 10 nodes per triangle and 4 per edge to identify a globally continuous piecewise cubic function. Figure 3.8-(top-right) shows a possible choice of dofs.

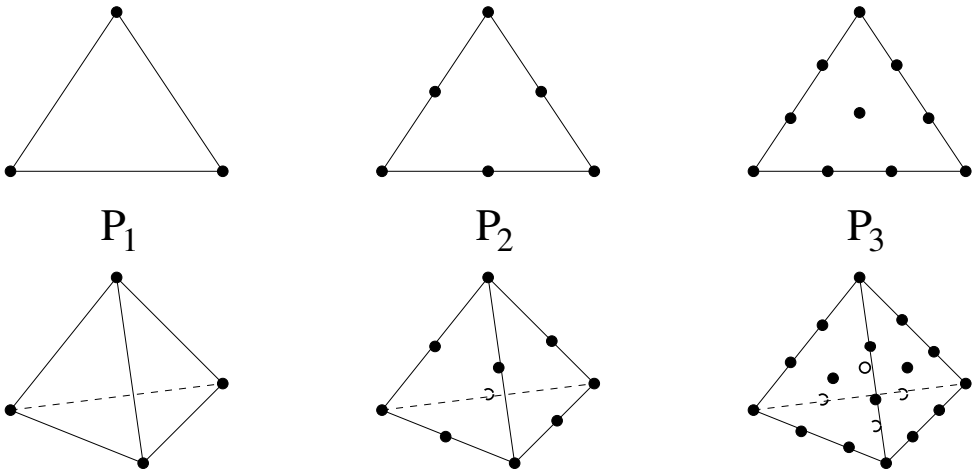


Figure 3.8: Choice of nodal dofs for finite elements of degree $r = 1, 2, 3$ in 2D on triangles (top row) and 3D on tetrahedra (bottom row).

The construction that we have presented in 2D can be extended without difficulty also in 3D on tetrahedra. The second row of Figure 3.8 shows a proper choice of dofs for Finite Elements of degree $r = 1, 2, 3$ in 3D.

3.3 Discontinuous finite elements on triangular affine meshes

In some cases one might want to remove the continuity requirement in the definition of the finite element space.

Definition 3.8. *The space of* ***discontinuous Finite Elements of degree*** r ***over a triangular affine mesh*** $\mathcal{T}_h$*, hereafter called* $X_h^{r,dc}$ *or, sometimes, for brevity simply* $\mathbb{P}_r^{dc}$*, is defined as*

$$X_h^{r,dc} = \{v \in L^2(\Omega) : \ v|_K \in \mathbb{P}_r(K) \quad \forall K \in \mathcal{T}_h\} \tag{3.7}$$

The construction of a set of dofs and a basis for this space is easier than in the continuous case. Indeed, we do not need to care about inter-element continuity and it is enough to choose $\binom{r+d}{d}$ points inside each element K to uniquely identify a polynomial of degree r.

In particular, one may choose the same points as in the continuous case (see Figure 3.8). Observe, however, that since the function is globally discontinuous, we will have to prescribe multiple values on each point, one for each element sharing that point. The dimension of the space $X^r_{h,dc}$ is

$$\dim(X^r_{h,dc}) = \binom{r+d}{d} N_{el}$$

where N_{el} is the number of elements in the mesh.

3.4 Non affine meshes and isoparametric finite elements

In the case of a non affine mesh, if $x = F_K(\hat{x})$ denotes the mapping from the reference element (triangle in 2D, tetrahedron in 3D) to the current *curved* element (see Figure 3.4), the definition of continuous finite elements (3.4) changes as

$$X^r_h = \{v \in C^0(\bar{\Omega}) : v|_K \circ F_K \in \mathbb{P}_r(\hat{K}) \quad \forall K \in \mathcal{T}_h\} \tag{3.8}$$

i.e. the *mapped* basis functions are polynomial on the reference element.

In practice, one has to build an invertible mapping for each curved element K of the mesh. A common way to describe such a mapping is to use again a polynomial space: $F_K \in \mathbb{P}_s(\hat{K})^d$. If s is taken equal to r (i.e. the degree of the mapping is equal to the degree of the finite elements), the resulting finite element space is called *isoparametric*.

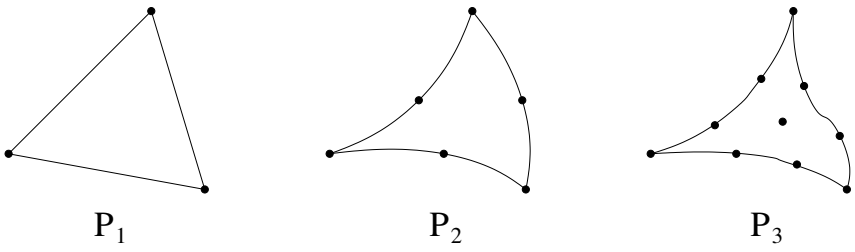
Definition 3.9. *A **isoparametric finite element space** over a triangulation $\mathcal{T}_h$ is the space*

$$X^{r,iso}_h = \{v \in C^0(\Omega) : \ v|_K \circ F_K \in \mathbb{P}_r(\hat{K}), \ \ F_K \in \mathbb{P}_r(\hat{K}), \quad \forall K \in \mathcal{T}_h\}.$$

The advantage of using isoparametric finite elements is that the mapping can be expanded on the same basis functions used for the finite element functions. Let $\hat{\varphi}_j$, $j = 1 \ldots, \dim(\mathbb{P}_r)$ be the set of Lagrangian basis functions constructed on the reference element $\hat{K}$. Then

$$F_K(\hat{x}) = \sum_{j=1}^{\dim(\mathbb{P}_r)} x_j \hat{\varphi}_j(\hat{x}),$$

where $x_j \in \mathbb{R}^d$ are the coordinates of the points (*geometric degrees of freedom*) defining the curved element K as the figure below shows



Clearly, the isoparametric $\mathbb{P}_1$ finite element space coincides with the standard $\mathbb{P}_1$ space. This is not the case anymore for a degree $r > 1$.

3.5 Continuous finite elements on quadrilateral meshes

We now turn to quadrilateral meshes in 2D, hexahedral in 3D. The definition we gave in (3.4) can not be extended directly to this case. Let us consider for instance linear finite elements in 2D. In this case we should define 3 degrees of freedom per element. However, being the function linear on each edge, we should provide 2 values per edge to guarantee overall continuity. Even if we chose the vertices of the quadrilateral, we will have a mismatch as we would have to prescribe 4 nodal values, which are too many to define a linear function in 2D.

For this type of finite elements, one has to work with richer polynomial spaces than the usual $\mathbb{P}_r$ ones. Let $\hat{K} = [0,1]^d$ be the unit cube in $\mathbb{R}^d$ and let us denote by $\mathbb{Q}_r$ the *tensor product polynomial space* of degree less than or equal to r:

$$\mathbb{Q}_r(\hat{K}) = \text{span}\{x_1^{k_1} x_2^{k_2} \dots x_d^{k_d}, \quad k_j \leq r, \ \ \forall j = 1, \dots, d, \quad (x_1, \dots, x_d) \in \hat{K}\} \tag{3.9}$$

It is easy to see that

$$\dim(\mathbb{Q}_r) = (r+1)^d.$$

The set of nodal degrees of freedom, on the reference element, which allows us to enforce automatically the overall continuity of the function is depicted in Figure 3.9 for the spaces $\mathbb{Q}_1$, $\mathbb{Q}_2$ and $\mathbb{Q}_3$ in 2 and 3 dimensions. Notice that, in general, a quadrilateral (in 2D) can not be

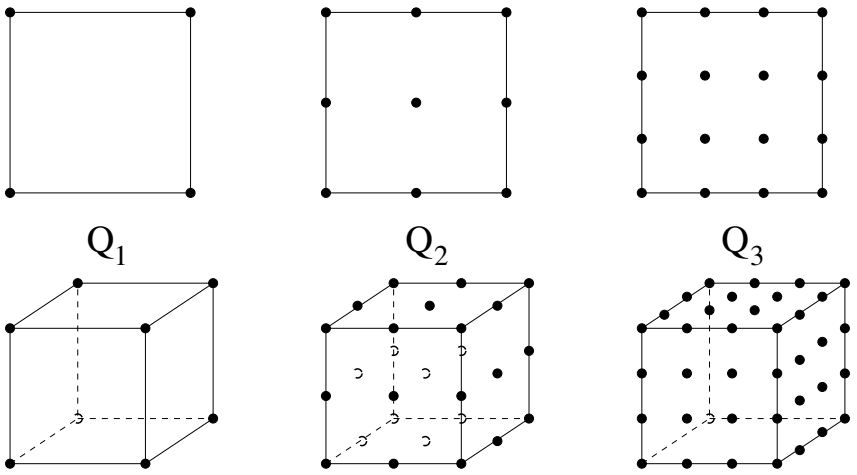


Figure 3.9: Choice of nodal dofs for finite elements of degree $r = 1, 2, 3$ in 2D on quadrilateral (top row) and 3D on hexahedra (bottom row).

mapped into a cube by an affine transformation unless it is a parallelogram. We have therefore to use the more general definition of finite element spaces for non affine meshes.

Definition 3.10. *The space of* ***continuous Finite Elements of degree*** r ***over a quadrilateral mesh*** $\mathcal{T}_h$*, hereafter called* Y_h^r *or, for brevity simply* $\mathbb{Q}_r$*, is defined for* $r \geq 1$ *as*

$$Y_h^r = \{v \in C^0(\Omega) : \ \ v|_K \circ F_K \in \mathbb{Q}_r(\hat{K}) \quad \forall K \in \mathcal{T}_h\}. \tag{3.10}$$

The Lagrangian basis and interpolant operator are defined as for triangular meshes.

Chapter 4

Finite element approximation of elliptic problems – implementation aspects

In this chapter we consider again the simple Poisson equation with mixed boundary conditions

$$\begin{cases} -\Delta u = f & \text{in} \quad \Omega \subset \mathbb{R}^d \\ \partial_n u = d & \text{on} \quad \Gamma_N \\ u = g & \text{on} \quad \Gamma_D \end{cases} \tag{4.1}$$

and will detail its finite element approximation, the practical construction of the algebraic system and the main properties of the system matrix (stiffness matrix). We recall the weak formulation of (4.1):

$$\text{find } u \in V_g \text{ s.t.} \qquad a(u,v) = F(v) \quad \forall\ v \in V_0 \tag{4.2}$$

with

$$\begin{aligned} &V_0 = H^1_{\Gamma_D}(\Omega) = \{v \in H^1(\Omega),\ \ v|_{\Gamma_D} = 0\} \\ &V_g = \{v \in H^1(\Omega),\ \ v|_{\Gamma_D} = g\} \\ &a(u,v) = \int_\Omega \nabla u \cdot \nabla v, \qquad F(v) = \int_\Omega fv + \int_{\Gamma_D} dv. \end{aligned}$$

Let $\mathcal{T}_h$ be a conforming regular triangulation of Ω. We assume hereafter that the domain Ω is polygonal and that both Γ_D and Γ_N are reproduced exactly as the union of straight edges of triangles in $\mathcal{T}_h$ (resp. faces of tetrahedra in 3D).

4.1 The full Neumann problem

We start by considering the full Neumann problem $\Gamma_D = \emptyset$. The solution is defined only up to a constant and one should set the problem in the space $V = H^1(\Omega)\backslash\mathbb{R} = \{v \in H^1(\Omega), \int_\Omega v = 0\}$. However, we forget this issue for the moment and we set the problem in $V = H^1(\Omega)$:

$$\text{find } u \in H^1(\Omega) \text{ s.t.} \qquad \int_\Omega \nabla u \cdot \nabla v = \int_\Omega fv + \int_{\Gamma_N} dv \quad \forall\ v \in H^1(\Omega). \tag{4.3}$$

We consider the finite element space of continuous piecewise polynomials of degree r:

$$X_h^r = \{v \in C^0(\Omega) : v|_K = \mathbb{P}_r(K),\ \forall K \in \mathcal{T}_h\}.$$

We have seen that $X_h^r \subset H^1(\Omega)$, so we can use it to construct a conforming Galerkin approximation to (4.3):

$$\text{find } u_h \in X_h^r \text{ s.t.} \qquad \int_\Omega \nabla u_h \cdot \nabla v_h = \int_\Omega f v_h + \int_{\Gamma_N} dv_h \quad \forall\ v_h \in X_h^r. \tag{4.4}$$

Let $\{\varphi_i,\ \ i = 1, \ldots, \mathcal{N}_h\}$ be the Lagrangian basis of X_h^r, with $\mathcal{N}_h = \dim(X_h^r)$. Then the solution u_h can be expanded on the basis as

$$u_h(x) = \sum_{i=1}^{\mathcal{N}_h} u_i \varphi_i(x)$$

with $u_i = u_h(x_i)$ the nodal values (degrees of freedom) and x_i the corresponding nodes in the mesh (set of vertices for $\mathbb{P}_1$ finite elements; vertices + mid points of edges for $\mathbb{P}_2$ finite elements, etc.). Having introduced such a basis, (4.4) is equivalent to the

$$\textbf{Algebraic system}: \qquad A\vec{u} = \vec{f}$$

with

$$\begin{aligned}
&\vec{u} \in \mathbb{R}^{\mathcal{N}_h}, && \vec{u} = (u_1, \ldots, u_{\mathcal{N}_h})^T \quad \text{(nodal values)}\\
&A \in \mathbb{R}^{\mathcal{N}_h \times \mathcal{N}_h}, && A_{ij} = \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i \quad \text{(stiffness matrix)}\\
&\vec{f} \in \mathbb{R}^{\mathcal{N}_h}, && f_i = \int_\Omega f\varphi_i + \int_{\Gamma_N} d\varphi_i \quad \text{(load vector).}
\end{aligned}$$

4.1.1 Construction of the stiffness matrix

Each element of the stiffness matrix is given by

$$A_{ij} = \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i = \sum_{K \in \mathcal{T}_h} \int_K \nabla\varphi_j \cdot \nabla\varphi_i.$$

The computation of A_{ij} reduces to the computation of $\int_K \nabla\varphi_j \cdot \nabla\varphi_i$ on each triangle $K \in \mathcal{T}_h$. The actual way this is done in many finite element codes is the following. (For simplicity we limit to $\mathbb{P}_1$ finite elements on triangles, but the argument generalizes to many more finite element spaces). Consider an element $K \in \mathcal{T}_h$. The triangle K has 3 vertices, which we denote by $a_{1,K}$, $a_{2,K}$ and $a_{3,K}$. Let $\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3$ be the global numbering of those vertices in $\mathcal{T}_h$ (see Figure 4.1). One can thus establish a *local to global map* for each K between the local numbering of the vertices and the corresponding global numbering (see Table 4.1).

On the triangle K, we can compute the *local stiffness matrix* $A_K \in \mathbb{R}^{3\times 3}$,

$$(A_K)_{ij} = \int_K \nabla\varphi_{a_{j,K}} \cdot \nabla\varphi_{a_{i,K}}.$$

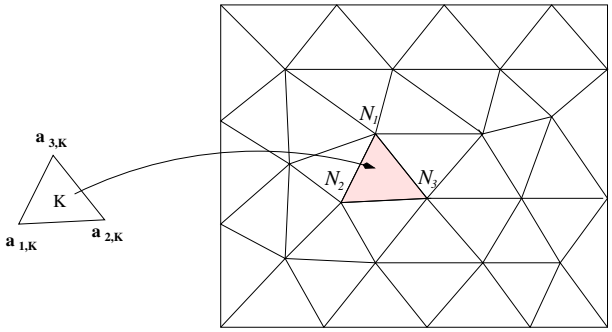


Figure 4.1: Local to global map

triangle K	local numbering		global numbering
	1 (vertex $a_{1,K}$)	$\longrightarrow$	$\mathcal{N}_1$
	2 (vertex $a_{2,K}$)	$\longrightarrow$	$\mathcal{N}_2$
	3 (vertex $a_{3,K}$)	$\longrightarrow$	$\mathcal{N}_3$

Table 4.1: Illustration of the local to global map for an element $K \in \mathcal{T}_h$.

Then, the term $(A_K)_{ij}$ will contribute to the global entry $A_{\mathcal{N}_i\mathcal{N}_j}$ of the stiffness matrix A

$$A_{\mathcal{N}_i\mathcal{N}_j} \leftarrow A_{\mathcal{N}_i\mathcal{N}_j} + (A_K)_{ij}$$

Similar considerations hold also for the right hand side.

A general implementation of a finite element solver consists of:

loop over the elements $K \in \mathcal{T}_h$

- compute the local stiffness matrix A_K and right hand side f_K
- Assemble the local matrix into the global matrix and the local r.h.s. into the global one:

  for $i = 1, \ldots, 3$

  for $j = 1, \ldots, 3$

  $A_{\mathcal{N}_i\mathcal{N}_j} = A_{\mathcal{N}_i\mathcal{N}_j} + (A_K)_{ij}$

  end

  $f_{\mathcal{N}_i} = f_{\mathcal{N}_i} + (f_K)_i$

  end

end

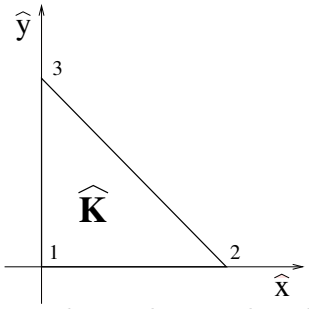
4.1.2 Computation of the local matrix

In the example we have considered, each entry of the local stiffness matrix is $(A_K)_{ij} = \int_K \nabla\varphi_j \cdot \nabla\varphi_i$. For $\mathbb{P}_1$ finite elements on an affine mesh, $\nabla\varphi_i$ is constant over K and the computation of $(A_K)_{ij}$ can be done directly starting from the coordinates of the vertices of K.

In more general cases of higher order finite elements and/or non affine meshes, the construction and evaluation of $\nabla\varphi_i$ on K might not be straightforward. In this case, one can first recast the integral onto the reference element $\hat{K}$, by introducing the map $x = F_K(\hat{x}) = B_K\hat{x} + b_K$, $K = F_K(\hat{K})$. Then

$$(A_K)_{ij} = \int_K \nabla\varphi_j \cdot \nabla\varphi_i = \int_{\hat{K}} B_K^{-T}\hat{\nabla}\hat{\varphi}_j \cdot B_K^{-T}\hat{\nabla}\hat{\varphi}_i \,|\det B_K|\, d\hat{x}. \tag{4.5}$$

Note that on the reference element $\hat{K}$ the expression of the basis functions $\hat{\varphi}_i$ is known analytically and all derivatives can be calculated easily. For instance, for $\mathbb{P}_1$ finite elements we have:



$$\begin{aligned}
&\hat{\varphi}_1(\hat{x},\hat{y}) = 1-\hat{x}-\hat{y}, && \nabla\hat{\varphi}_1 = [-1,-1]^T\\
&\hat{\varphi}_2(\hat{x},\hat{y}) = \hat{x}, && \nabla\hat{\varphi}_2 = [1,0]^T\\
&\hat{\varphi}_3(\hat{x},\hat{y}) = \hat{y}, && \nabla\hat{\varphi}_3 = [0,1]^T
\end{aligned}$$

Moreover, for triangles with straight edges, the map $x = F_K(\hat{x})$ can also be easily constructed starting from the coordinates of the vertices $a_{1,K}$, $a_{2,K}$ and $a_{3,K}$ as we have seen in Section 3.1.1.

Then, a quadrature formula can be used to compute the integral on the right hand side of (4.5). A quadrature formula on the reference element $\hat{K}$ will have the form:

$$Q_{\hat{K}}(f) = \sum_{l=1}^{nqp} f(\hat{x}_l)\omega_l$$

where $Q_{\hat{K}}(f) \approx \int_{\hat{K}} f(\hat{x})d\hat{x}$ and nqp is the number of quadrature points used by the quadrature formula. One typically chooses a quadrature formula that is exact in computing the stiffness matrix $\int_{\hat{K}} \hat{\nabla}\hat{\varphi}_j \cdot \hat{\nabla}\hat{\varphi}_i d\hat{x}$ or the mass matrix $\int_{\hat{K}} \hat{\varphi}_j\hat{\varphi}_i d\hat{x}$. Then, the approximation of the local stiffness matrix will be

$$(A_K)_{ij} \approx \sum_{l=1}^{nqp} B_K^{-T}\hat{\nabla}\hat{\varphi}_j(\hat{x}_l) \cdot B_K^{-T}\hat{\nabla}\hat{\varphi}_i(\hat{x}_l)\,|\det B_K|\omega_l.$$

One has therefore to compute on $\hat{K}$, once and for all, the matrix

$$D\Phi_{lj} = \hat{\nabla}\hat{\varphi}_j(\hat{x}_l)$$

which will then be used in the computation of the local stiffness matrix on each element $K \in \mathcal{T}_h$.

In the case of the mass matrix, one will have to compute and store also the matrix

$$\Phi_{lj} = \hat{\varphi}_j(\hat{x}_l).$$

An example of a quadrature formula on the reference triangle $\hat{K}$ is the following

$$Q_{\hat{K}}(f) = |\hat{K}|\left(\frac{9}{20}f(b) + \frac{2}{15}(f(m_1)+f(m_2)+f(m_3)) + \frac{1}{20}(f(a_1)+f(a_2)+f(a_3))\right) \tag{4.6}$$

where b is the barycenter, m_i, $i = 1, 2, 3$, the mid-points of the edges and a_i, $i = 1, 2, 3$, the vertices of the triangle. This formula has degree of exactness 3 hence it will integrate exactly the local stiffness matrix (in the case of constant coefficients and affine meshes) when using $\mathbb{P}_2$ finite elements.

If the map $x = F_K(\hat{x})$ is non-affine, then the Jacobian matrix $J_K = \nabla F_K$ will not be a constant matrix. Therefore the contribution

$$(A_K)_{ij} = \int_{\hat{K}} J_K^{-T} \hat{\nabla} \hat{\varphi}_j \cdot J_K^{-T} \hat{\nabla} \hat{\varphi}_i \, |\det J_K| \, d\hat{x} \tag{4.7}$$

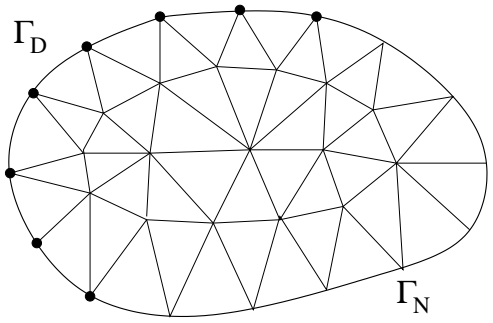
will have to be computed a fortiori with a quadrature formula, as for instance the formula (4.6). Remember that in the non-affine case, the space X_h^r is defined as

$$X_h^r = \{v \in C^0(\Omega) : v|_K \circ F_K \in \mathbb{P}_r(\hat{K}) \quad \forall K \in \mathcal{T}_h\}.$$

and the mapped basis functions $\hat{\varphi}_i = \varphi_i|_K \circ F_K$ are the usual Lagrangian basis functions on the reference element $\hat{K}$ and can be easily evaluated at any point.

4.2 Treatment of non-homogeneous Dirichlet boundary conditions

We consider again problem (4.1) and its weak formulation (4.2). Referring to the figure below,



let us denote by x_i^B, $i = 1, \ldots, \mathcal{N}_h^B$ the nodes of the mesh that fall on the Dirichlet boundary and by x_j^I, $j = 1, \ldots, N_h^I$ the nodes that fall inside the domain or on the Neumann boundary. On the Dirichlet nodes x_i^B, we would like to impose the condition $u_h(x_i^B) = g(x_i^B)$. We can therefore define the two spaces

$$X_{h,0}^r = \{v_h \in X_h^r : \ v_h(x_i^B) = 0\}$$
$$X_{h,g}^r = \{v_h \in X_h^r : \ v_h(x_i^B) = g(x_i^B)\}$$

Observe that, if the domain is polygonal and the Dirichlet boundary is exactly represented by edges of the mesh, then the functions in $X_{h,0}^r$ will vanish on Γ_D and $X_{h,0}^r \subset V_0$. On the other hand, in general, $X_{h,g}^r \not\subseteq V_g$ since $v_h \in X_{h,g}^r$ on the Dirichlet boundary will be equal to the interpolation of the Dirichlet datum g in the finite element space $X_h^r(\Gamma_D)$ and not to g itself.

The finite element formulation of problem (4.2) is

$$\text{find } u_h \in X_{h,g}^r \text{ s.t.} \qquad a(u_h, v_h) = F(v_h) \quad \forall v_h \in X_{h,0}^r. \tag{4.8}$$

Strictly speaking, this is a non conforming approximation since $u_h \in X^r_{h,g} \not\subseteq V_g$. In practice, to impose the Dirichlet boundary conditions and derive the algebraic system, we can follow three alternative approaches.

1. *Eliminate the boundary Dirichlet nodes and reduce the system*

   Let $\{\varphi^I_j\}_{j=1}^{\mathcal{N}^I_h}$ be the set of basis functions corresponding to the interior and Neumann nodes x^I_j, which is clearly a basis of $X^r_{h,0}$. Let moreover $\{\varphi^B_j\}_{j=1}^{\mathcal{N}^B_h}$ be the set of basis functions corresponding to Dirichlet nodes x^B_j. We have

$$\begin{aligned} u_h(x) &= \sum_{j=1}^{\mathcal{N}_h} u_j \varphi_j(x) = \sum_{j=1}^{\mathcal{N}^I_h} u_j \varphi^I_j(x) + \sum_{l=1}^{\mathcal{N}^B_h} u_l \varphi^B_l(x) \\ &= \underbrace{\sum_{j=1}^{\mathcal{N}^I_h} u_j \varphi^I_j(x)}_{\mathring{u}_h \in X^r_{h,0},\ \text{unknown}} + \underbrace{\sum_{l=1}^{\mathcal{N}^B_h} g(x_l)\varphi^B_l(x)}_{G_h \in X^r_{h,g},\ \text{known}} \end{aligned}$$

   With this splitting, we can rewrite the problem (4.8) as

$$\text{find } \mathring{u}_h \in X^r_{h,0} \text{ s.t.} \qquad a(\mathring{u}_h, v_h) = F(v_h) - a(G_h, v_h) \quad \forall\ v_h \in X^r_{h,0}. \tag{4.9}$$

   Notice that the function G_h plays the same role as the *lifting of the Dirichlet datum* that we have already used in Chapter 1 to analyze the well posedness of the Poisson equation with non-homogeneous Dirichlet boundary conditions. However, in this case, the lifting G_h is confined only in the layer of elements with an edge on Γ_D (so it will not be bounded in H^1 as $h \to 0$).

   At the algebraic level, problem (4.9) leads to the linear system

$$A^{II}\vec{\mathring{u}} = \vec{f}^I - A^{IB}\vec{g}$$

   where

$$\begin{aligned} &A^{II} \in \mathbb{R}^{\mathcal{N}^I_h \times \mathcal{N}^I_h}, \quad A^{II}_{ij} = a(\varphi^I_j, \varphi^I_i) \\ &A^{IB} \in \mathbb{R}^{\mathcal{N}^B_h \times \mathcal{N}^I_h}, \quad A^{IB}_{ij} = a(\varphi^B_j, \varphi^I_i) \\ &\vec{f}^I \in \mathbb{R}^{\mathcal{N}^I_h}, \quad f^I_j = F(\varphi^I_j) \\ &\vec{\mathring{u}} \in \mathbb{R}^{\mathcal{N}^I_h}, \quad \mathring{u}_j = \mathring{u}_h(x^I_j) \quad \text{(vector of dofs on interior and Neumann nodes)} \\ &\vec{g} \in \mathbb{R}^{\mathcal{N}^B_h}, \quad g_l = g(x^B_l), \quad \text{(boundary Dirichlet values).} \end{aligned}$$

   This strategy has the disadvantage that one has to introduce two numbering of nodes. The global numbering (of all nodes) and the reduced numbering of only the interior+Neumann nodes.

   Also, in the assembling of the local matrices into the global one, one has to check on each triangle K if it has Dirichlet nodes and assemble this contribution in the right hand side instead of the matrix. This implies adding conditional statements in the loop of elements with corresponding slow down of performances.

2. *Ignore the Dirichlet conditions when assembling the matrix and enforce them afterward.*

   We assemble the matrix $A \in \mathbb{R}^{\mathcal{N}_h \times \mathcal{N}_h}$ without Dirichlet boundary conditions, so no conditional statements are inserted in the loop over the elements.

   Afterward, say that x_i is a Dirichlet node. Then, we want to replace row i in the matrix with the simple equation $u_i = g(x_i)$.

   We can therefore modify the system matrix and the right hand side in the following way. Let $\tilde{A} = A$ and $\vec{\tilde{f}} = \vec{f}$:

   for all Dirichlet nodes x_i set
   - $\tilde{A}_{ii} = 1$, $\tilde{A}_{ij} = 0$, $\forall j \neq i$, and $\tilde{f}_i = g(x_i)$

   end

   This corresponds to "zeroing" the $i-$th row, putting 1 on the diagonal and changing the corresponding term on the right hand side to the Dirichlet value $g(x_i)$. Hence the $i-$th equation becomes $u_i = g(x_i)$.

   Once all the rows corresponding to Dirichlet nodes have been modified, one solves the linear system

   $$\tilde{A}\vec{u} = \vec{\tilde{f}}$$

   where $\vec{u}$ is the unknown vector containing *all degrees of freedom*, including those on the Dirichlet boundary.

   In case the original matrix A is symmetric, the disadvantage of this technique is that the symmetry is lost in the matrix $\tilde{A}$ since few rows have been zeroed but not the corresponding columns. Hence, we can not use a symmetric factorization method as a Cholesky factorization. On the other hand, the loss of symmetry might not be a problem if we use an iterative solver (even Conjugate Gradient) provided that the initial solution satisfies exactly the boundary values.

3. *Zero also the columns to recover symmetry of the matrix.*

   Let x_i be a Dirichlet node. In the modified system $\tilde{A}\vec{u} = \vec{\tilde{f}}$ constructed before, the i-th equation is trivial and u_i is not really an unknown.

   Consider now the row j of the system, not corresponding to a Dirichlet node: $\sum_{j=1}^{\mathcal{N}_h} \tilde{A}_{jk} u_k = \tilde{f}_j$ (notice that in this row, the entries of the matrix and right hand side have not been changed so $\tilde{A}_{jk} = A_{jk}$ for all k and $\tilde{f}_j = f_j$).

   If we isolate the element $\tilde{A}_{ji}$, we have

   $$\sum_{j=1, j\neq i}^{\mathcal{N}_h} \tilde{A}_{jk} u_k + \tilde{A}_{ji} u_i = \tilde{f}_j$$

   Since u_i is known, this can be written as

   $$\sum_{j=1, j\neq i}^{\mathcal{N}_h} \tilde{A}_{jk} u_k = \tilde{f}_j - \tilde{A}_{ji} g(x_i)$$

which can be further rewritten as

$$\sum_{j=1}^{\mathcal{N}_h} \hat{A}_{jk} u_k = \hat{f}_j$$

having set $\hat{A}_{jk} = \tilde{A}_{jk}$ for all $k \neq i$, $\hat{A}_{ji} = 0$ and $\hat{f}_j = \tilde{f}_j - A_{ji} g(x_i)$.

This step corresponds actually to zeroing the i-th column of the matrix (except for the diagonal term) and correcting accordingly the right hand side. Observe that, in general, many entries of the i-th column are already zero (the stiffness matrix is sparse, see next section), so only the rows corresponding to nodes in the neighborhood of x_i have to be modified.

After doing this procedure for all Dirichlet nodes x_i we obtain a modified matrix $\hat{A}$ where both rows and columns corresponding to Dirichlet nodes have been zeroed and a modified right hand side $\vec{\hat{f}}$ to account for the non homogeneous Dirichlet data. Finally, the system to solve is

$$\hat{A}\vec{u} = \vec{\hat{f}}.$$

Observe that, if A is symmetric, so will be the matrix $\hat{A}$.

This procedure together with the previous one can be written in algorithmic form as follows. Set $\hat{A} = A$ and $\vec{\hat{f}} = \vec{f}$.

for all Dirichlet nodes x_i

- zero the $i-th$ row: $\hat{A}_{ii} = 1, \quad \hat{A}_{ij} = 0, \ \forall j \neq i, \quad \hat{f}_i = g(x_i)$
- for all interior or Neumann nodes x_j neighboring x_i, set
  - $\hat{f}_j = \hat{f}_j - \hat{A}_{ji} g(x_i)$
  - $\hat{A}_{ji} = 0$

  end

end

4.3 Some properties of the stiffness matrix

We discuss now some properties of the stiffness matrix $A_{ij} = a(\varphi_j, \varphi_i)$ corresponding to the bilinear form $a(u, v) = \int_\Omega \nabla u \cdot \nabla v$. We focus, in particular, on the submatrix A^{II} related to interior or Neumann nodes, since the extra rows corresponding to Dirichlet nodes, that are added to the matrix in the approaches 2) and 3) above are trivial ones.

For convenience, we rename the submatrix A^{II} simply as A, i.e.

$$A_{ij} = a(\varphi_j^I, \varphi_i^I), \quad \forall i, j = 1, \ldots \mathcal{N}_h^I \tag{4.10}$$

- (**symmetry**) The matrix A is symmetric. This follows immediately from the symmetry of the bilinear form.

- (**positivity**) The matrix A is positive definite. This is a consequence of the coercivity of the bilinear form. Indeed given $\vec{v} = (v_1, \ldots, v_{\mathcal{N}_h^I}) \in \mathbb{R}^{\mathcal{N}_h^I}$, form the function $v_h = \sum_{i=1}^{\mathcal{N}_h^I} v_i \varphi_i \in H^1_{\Gamma_D}(\Omega)$. Then,

$$\vec{v}^T A \vec{v} = a(v_h, v_h) = \int_\Omega |\nabla v_h|^2 \geq 0,$$

  and $\vec{v}^T A \vec{v} = 0$ iff $\vec{v} = 0$ thanks to the Dirichlet boundary conditions imposed.

- (**Sparsity**) The matrix A is sparse, i.e. the number of non-zero entries in each row is $\mathcal{O}(1)$. Indeed $A_{ij} = \int_\Omega \nabla \varphi_j^I \cdot \nabla \varphi_i^I = 0$ whenever the supports of φ_j^I and φ_i^I have empty intersection.

  If we consider, to fix ideas, $\mathbb{P}_1$ continuous finite elements and a Lagrangian basis, an entry A_{ij} can be non zero only if the vertices i and j are connected by an edge. Therefore, the number of non-zero entries in row i is smaller or equal to the number of vertices j connected to i by an edge. For a *sequence of regular meshes* this number is bounded uniformly with respect to the mesh size h, and is usually small. Referring to the meshes in Figure 3.3 in Chapter 3, the number of non-zero entries per row is about 6.

  Since the matrix is sparse, one will typically use a *sparse representation* of the matrix, i.e. only the non-zero entries are stored and the memory occupation is only $\mathcal{O}(\mathcal{N}_h)$ instead of $\mathcal{O}(\mathcal{N}_h^2)$ for a full representation.

  Moreover, since every row contains only $\mathcal{O}(1)$ non-zero entries, a matrix-vector multiplication will entail $\mathcal{O}(\mathcal{N}_h)$ floating point operations (instead of $\mathcal{O}(\mathcal{N}_h^2)$ for a full representation).

  This makes iterative methods particularly attractive to solve finite element problems, especially for low order approximations.

Notice that the first two properties (symmetry and positivity) are only related to the properties of the bilinear form and not to the choice of the discretization space. On the other hand, the *sparsity* property is a consequence of the particular choice of the discretization space (finite elements) and the use of a Lagrangian basis. Other discretizations (as for instance spectral methods) might not have this property.

4.4 Condition number of the stiffness matrix

We consider again the stiffness matrix (4.10) corresponding to interior and Neumann nodes only. Let $\{\mathcal{T}_h\}_{h \searrow 0}$ be a family of *regular* and *quasi-uniform* affine meshes such that $h_K \leq \gamma \rho_K$ and $h_K \geq \delta h$ for all $K \in \mathcal{T}_h$ and all $h > 0$, where ρ_K is the diameter of the largest ball contained in K, h_K the diameter of K and $h = \max_{K \in \mathcal{T}_h} h_K$ (see Chapter 3 for the exact definitions). We set moreover $\rho_{min} = \min_{K \in \mathcal{T}_h} \rho_K$. Finally, for a basis function φ_i^I of $X^r_{h,0}$, let ζ_i be the number of elements in the mesh with non zero intersection with the support of φ_i^I, namely $\zeta_i = \#\{K \in \mathcal{T}_h : \quad K \cap \mathrm{supp}(\varphi_i^I) \neq \emptyset\}$, and set $\zeta = \max_{i=1,\ldots,N_h^I} \zeta_i$. We prove the following result:

Theorem 4.1. *The condition number of the stiffness matrix A in* (4.10) *can be bounded by*

$$\kappa(A) \leq C \zeta \left(\frac{h}{\rho_{min}} \right)^d \rho_{min}^{-2}, \tag{4.11}$$

where $C > 0$ does not depend on the mesh $\mathcal{T}_h$. Moreover, for a quasi uniform family of meshes it holds

$$\kappa(A) \le \tilde{C} h^{-2}, \tag{4.12}$$

with $\tilde{C} = C\zeta\,(\gamma/\delta)^{d+2}$ and ζ, γ, δ independent of h.

Before proving this theorem, we need some preliminary observations and lemmas. Consider the reference element $\hat{K}$ and the space $\mathbb{P}_r(\hat{K})$ of polynomials of degree at most r. Let $\{\hat{\varphi}_i,\ i = 1, \dots, \mathcal{N}_r\}$ be a Lagrangian basis of $\mathbb{P}_r(\hat{K})$. Here $\mathcal{N}_r$ is the local number of degrees of freedom. Then, any polynomial $\hat{v} \in \mathbb{P}_r(\hat{K})$ can be expanded on the basis $\hat{v}(\hat{x}) = \sum_{i=1}^{\mathcal{N}_r} v_i \hat{\varphi}_i(\hat{x})$ and $\mathbb{P}_r(\hat{K})$ is in 1-to-1 correspondence with $\mathbb{R}^{\mathcal{N}_r}$

$$\hat{v} \in \mathbb{P}_r(\hat{K}) \quad \Leftrightarrow \quad \vec{v} = (v_1, \dots, v_{\mathcal{N}_r}) \in \mathbb{R}^{\mathcal{N}_r}.$$

Being $\mathbb{P}_r(\hat{K})$ (and $\mathbb{R}^{\mathcal{N}_r}$) finite dimensional, all norms are equivalent. Therefore, there exist $C_1, C_m, C_M > 0$ such that

$$\|\hat{v}\|_{L^2(\hat{K})} \le \|\hat{v}\|_{H^1(\hat{K})} \le C_1 \|\hat{v}\|_{L^2(\hat{K})} \quad \text{and} \quad C_m |\vec{v}| \le \|\hat{v}\|_{L^2(\hat{K})} \le C_M |\vec{v}|, \quad \forall\ \hat{v} \in \mathbb{P}_r(\hat{K}), \tag{4.13}$$

the last being the euclidean norm of the vector $\vec{v}$.

Consider now an element $K \in \mathcal{T}_h$ and the mapping $x = B_K \hat{x} + b_K$ from $\hat{K}$ to K. We recall the following properties of the matrix B_K (see Lemma 3.6):

$$|\det B_K| = \frac{|K|}{|\hat{K}|}, \qquad \|B_K\| \le \frac{h_K}{\rho_{\hat{K}}}, \qquad \|B_K^{-1}\| \le \frac{h_{\hat{K}}}{\rho_K}$$

which imply

$$\frac{\rho_K}{h_{\hat{K}}} \le \|B_K\| \le \frac{h_K}{\rho_{\hat{K}}}, \qquad \frac{\rho_{\hat{K}}}{h_K} \le \|B_K^{-1}\| \le \frac{h_{\hat{K}}}{\rho_K}, \tag{4.14}$$

and, for any $v \in \mathbb{P}_r(K)$

$$\|v\|^2_{L^2(K)} = \int_{\hat{K}} \hat{v}^2\, |\det B_K| = \frac{|K|}{|\hat{K}|} \|\hat{v}\|^2_{L^2(\hat{K})}. \tag{4.15}$$

The following lemma is necessary to analyze the condition number of the stiffness matrix.

Lemma 4.2 (**Local inverse inequality**). *For all $K \in \mathcal{T}_h$ and $v \in \mathbb{P}_r(K)$*

$$\|\nabla v\|_{L^2(K)} \le C \rho_K^{-1} \|v\|_{L^2(K)} \tag{4.16}$$

where $C = C_1 h_{\hat{K}}$ and C_1 is defined in (4.13).

Proof. It holds

$$\begin{aligned}
\|\nabla v\|^2_{L^2(K)} &= \int_{\hat{K}} |B_K^{-T} \hat{\nabla} \hat{v}|^2\, |\det B_K| \le \frac{|K|}{|\hat{K}|} \frac{h^2_{\hat{K}}}{\rho^2_K} \|\hat{\nabla} \hat{v}\|^2_{L^2(\hat{K})} \\
&\le C_1^2 \frac{|K|}{|\hat{K}|} \frac{h^2_{\hat{K}}}{\rho^2_K} \|\hat{v}\|^2_{L^2(\hat{K})} = C \rho_K^{-2} \|v\|^2_{L^2(K)}.
\end{aligned}$$

□

As an immediate consequence we have

Lemma 4.3 (**Global inverse inequality**). *For any $v_h \in X_h^r$ it holds*

$$\|\nabla v_h\|_{L^2(\Omega)} \leq C\rho_{min}^{-1}\|v_h\|_{L^2(\Omega)}. \tag{4.17}$$

Moreover, if the family of meshes $\{\mathcal{T}_h\}_h$ is quasi-uniform (hence regular), i.e. $\rho_K \geq \frac{1}{\gamma}h_K \geq \frac{\delta}{\gamma}h$ for all $K \in \mathcal{T}_h$ and all $h > 0$, then

$$\|\nabla v_h\|_{L^2(\Omega)} \leq C\left(\frac{\gamma}{\delta}\right)h^{-1}\|v_h\|_{L^2(\Omega)}, \qquad \forall v_h \in X_h^r. \tag{4.18}$$

Observe that the inequality will not be true if we replace the (finite dimensional) space X_h^r with the (infinite dimensional) space $H^1_{\Gamma_D}(\Omega)$. This type of inequalities are called *inverse inequalities* and it has to be expected that the bound in (4.18) degenerates as $h \to 0$.

We are now ready to prove Theorem 4.1.

Proof. (*of Theorem 4.1*)

Being A symmetric and positive definite, its eigenvalues are all real and positive and can be estimated via the Rayleigh quotient

$$\lambda_{\min}(A) \leq \frac{\vec{v}^T A\vec{v}}{|\vec{v}|^2} \leq \lambda_{\max}(A) \qquad \forall \vec{v} \in \mathbb{R}^{\mathcal{N}_h^I}$$

We split the Rayleigh quotient in two factors

$$R(\vec{v}) = \frac{\vec{v}^T A\vec{v}}{|\vec{v}|^2} = \underbrace{\frac{\vec{v}^T A\vec{v}}{\vec{v}^T M\vec{v}}}_{R_A(\vec{v})} \cdot \underbrace{\frac{\vec{v}^T M\vec{v}}{|\vec{v}|^2}}_{R_M(\vec{v})}$$

where M is the mass matrix $M_{ij} = \int_\Omega \varphi_i\varphi_j$. Observe that M is also symmetric, positive definite and

$$\forall u_h \in X_h^r, \quad u_h(x) = \sum_{i=1}^{\mathcal{N}_h^I} u_i\varphi_i(x), \quad \vec{u}^T M\vec{u} = \int_\Omega u_h^2 = \|u_h\|^2_{L^2(\Omega)}.$$

Estimate for $R_A(\vec{v})$.

For any $\vec{v} = (v_1, \ldots, v_{\mathcal{N}_h^I})$ and associated finite element function $v_h(x) = \sum_{i=1}^{\mathcal{N}_h^I} v_i\varphi_i(x)$, we easily see that $R_A(\vec{v})$ has the following characterization:

$$R_A(\vec{v}) = \frac{\|\nabla v_h\|^2_{L^2(\Omega)}}{\|v_h\|^2_{L^2(\Omega)}}.$$

A lower bound for $R_A(\vec{v})$ is therefore given by the Poincaré inequality$\|v_h\|_{L^2(\Omega)} \leq C_p\|\nabla v_h\|_{L^2(\Omega)}$. Therefore

$$R_A(\vec{v}) \geq C_p^{-2}.$$

On the other hand, an upper bound follows from the global inverse inequality (4.17):

$$R_A(\vec{v}) \leq C^2\rho_{\min}^{-2}$$

Therefore

$$C_p^{-2} \leq R_A(\vec{v}) \leq C^2 \rho_{\min}^{-2} \tag{4.19}$$

Estimate for $R_M(\vec{v})$.

Recall that

$$\vec{v}^T M \vec{v} = \|v_h\|^2_{L^2(\Omega)} = \sum_K \|v_h\|^2_{L^2(K)} = \sum_K \frac{|K|}{|\hat{K}|} \|\hat{v}_h\|^2_{L^2(\hat{K})}.$$

Let now $\Theta_K = \{i \in \{1, \ldots, N_h^I\} : \quad K \cap \mathrm{supp}(\varphi_i^I) \neq \emptyset\}$. From the norm equivalence (4.13) we have

$$C_m^2 \sum_{i \in \Theta_K} v_i^2 \leq \|\hat{v}_h\|^2_{L^2(\hat{K})} \leq C_M^2 \sum_{i \in \Theta_K} v_i^2,$$

and

$$|\vec{v}|^2 \leq \sum_{K \in \mathcal{T}_h} \sum_{i \in \Theta_K} v_i^2 \leq \sum_{i=1}^{N_h^I} \zeta_i v_i^2 \leq \zeta |\vec{v}|^2.$$

Therefore, recalling that $\hat{C} \rho_K^d \leq |K| \leq h_K^d$, we have

$$C_m^2 \, \frac{\hat{C} \, \rho_{\min}^d}{|\hat{K}|} \leq R_M(\vec{v}) = \frac{\|v_h\|^2_{L^2(\Omega)}}{|\vec{v}|^2} \leq C_M^2 \zeta \frac{h^d}{|\hat{K}|}. \tag{4.20}$$

Finally

$$\frac{C_m^2 \, \hat{C}}{|\hat{K}| C_p^2} \, \rho_{\min}^d \leq R(\vec{v}) \leq \frac{C_M^2 C^2}{|\hat{K}|} \, \rho_{\min}^{-2} h^d$$

and

$$\kappa(A) \leq C \rho_{min}^{-(d+2)} h^d. \tag{4.21}$$

with $C = \frac{C_M^2 C^2 C_p^2}{C_m^2 \hat{C}}$. Finally, for a family of quasi uniform meshes with $\rho_K \geq \frac{1}{\gamma} h_K \geq \frac{\delta}{\gamma} h$, $\forall K \in \mathcal{T}_h$ and $h > 0$, the bound (4.12) follows. □

Remark 4.4. *The result* (4.19) *provides a bound on $\kappa(M^{-1}A)$, useful when the mass matrix is used as a preconditioner for the stiffness matrix.*

Chapter 5

Approximation results for Finite Elements spaces

We look at the case of continuous triangular finite elements on affine meshes

$$X_h^r = \{v \in C^0(\Omega), v|_K \in \mathbb{P}_r(K) \quad \forall\ K \in \mathcal{T}_h\}.$$

although many of the arguments generalize to other finite elements as well. We are interested in studying the approximability properties of this space for smooth functions $v \in H^s(\Omega)$ with $s > 1$, namely we would like to quantify the *best approximation error*

$$H^1\text{-BAE} = \inf_{w_h \in X_h^r} \|v - w_h\|_{H^1(\Omega)},$$
$$L^2\text{-BEA} = \inf_{w_h \in X_h^r} \|v - w_h\|_{L^2(\Omega)}$$

for v in $H^s(\Omega)$, with $s > 1$. Observe that since $X_h^r \subset H^1$ but $X_h^r \nsubseteq H^r, r \geq 2$, it does not make sense to measure the best approximation error in a norm higher than H^1. However, on each element $K \in \mathcal{T}_h$, a function $v_h \in X_h^r$ is polynomial and hence infinitely differentiable. We could, therefore, measure the BEA in higher norms *element-wise*. Let us introduce the so called *broken* H^m *space*:

$$H^m_{bro}(\Omega) = \{v \in L^2(\Omega) :\ \|v_{|K}\|_{H^m(K)} < +\infty,\ \forall K \in \mathcal{T}_h\}$$

endowed with the norm $\|v\|^2_{H^m_{bro}} = \sum_{K \in \mathcal{T}_h} \|v_{|K}\|^2_{H^m(K)}$.

For $m > 0$, H^m is strictly contained in H^m_{bro}, whereas for $m = 0$ (L^2 space) the two spaces coincide. Moreover, it holds

$$\|v\|_{H^m(\Omega)} = \|v\|_{H^m_{bro}(\Omega)}, \quad \forall v \in H^m(\Omega).$$

With this definition, we could also try to estimate the best approximation error in H^m-broken norms

$$H^m_{bro}\text{-BAE} = \inf_{w_h \in X_h^r} \|v - w_h\|_{H^m_{bro}(\Omega)}, \qquad m = 0, \ldots, r.$$

Observe that it is not worth going beyond $m = r$. Indeed, for any $w_h \in X_h^r$, we have $D^{\boldsymbol{\alpha}} w_{h|K} = 0$ for all $|\boldsymbol{\alpha}| > r$ since w_h is a polynomial of degree r in K and the function w_h is not capable of providing any approximation of derivatives of v higher than r.

The typical procedure to obtain estimates on the best approximation error for a given function $v \in H^s$ consists in building a particular function $v_h \in X_h^r$ starting from v, that is a good approximation of it. Then, the best approximation error in, say, the H^m_{bro}-norm will be bounded by $\|v - v_h\|_{H^m_{bro}(\Omega)}$.

A natural candidate for the approximating function v_h is to use the *finite element interpolant*: $v_h = I_h^r v$ that we have introduced in Chapter 3. This is however not the only choice (although it is the only one we will discuss). Notice that since the interpolant operator involves point evaluations of the function v, this procedure is well suited only if the function v is at least continuous. This is guaranteed if $v \in H^s(\Omega)$ for $s > d/2$, d being the physical dimension. For 2D and 3D applications, we can take $s \geq 2$. To obtain best approximation error estimates for functions $v \in H^s(\Omega)$ with $s < 2$, other reconstructions v_h than the interpolant will have to be used, as e.g. the Clément or the Scott and Zhang interpolants (see e.g. [5, 2]).

We anticipate here the main result concerning the interpolation error, whose proof will be the subject of the next sections.

Theorem 5.1 (Intepolation error for smooth functions)**.** *Given a family of regular triangulations $\{\mathcal{T}_h\}_{h>0}$ of a polygonal domain $\Omega \subset \mathbb{R}^d$, $d \leq 3$, and the space X_h^r of continuous finite elements of degree $r \geq 1$, there exist $C_m > 0$, $m = 0, \ldots, r$, such that for any function $v \in H^s$, $s \geq r+1$*

$$\|v - I_h^r v\|_{L^2(\Omega)} \leq C_0 h^{r+1} |v|_{H^{r+1}(\Omega)} \tag{5.1}$$

$$\|v - I_h^r v\|_{H^1(\Omega)} \leq C_1 h^r |v|_{H^{r+1}(\Omega)} \tag{5.2}$$

$$\|v - I_h^r v\|_{H^m_{bro}(\Omega)} \leq C_m h^{r+1-m} |v|_{H^{r+1}(\Omega)}, \quad 2 \leq m \leq r, \tag{5.3}$$

with C_m that depend on $\gamma = \max_{K \in \mathcal{T}_h} \frac{h_K}{\rho_K}$, r and m, but are otherwise independent of h.

Observe that we could have just written (5.3) for $m = 0, \ldots, r$, thanks to the fact that under the hypotheses of the theorem, $\|v - I_h^r v\|_{H^m} = \|v - I_h^r v\|_{H^m_{bro}}$ for $m = 0, 1$.

In the case where the function v does not have the required regularity to achieve the maximum convergence rate, the previous result generalizes as

Theorem 5.2 (Interpolation error for possibly non-smooth functions)**.** *Given a family of regular triangulations $\{\mathcal{T}_h\}_{h>0}$ of a polygonal domain $\Omega \subset \mathbb{R}^d$, $d \leq 3$, and the space X_h^r of continuous finite elements of degree $r \geq 1$, for $s \geq 2$, and setting $\eta = \min\{s, r+1\}$, there exist $C_m > 0$, $m = 0, \ldots, \eta$, such that for any function $v \in H^s$*

$$\|v - I_h^r v\|_{H^m_{bro}(\Omega)} \leq C_m h^{\eta - m} |v|_{H^\eta(\Omega)}, \quad 0 \leq m \leq \eta,$$

with C_m that depend on $\gamma = \max_{K \in \mathcal{T}_h} \frac{h_K}{\rho_K}$, r, s and m, but are otherwise independent of h.

5.1 Local approximation estimates

The first step to prove Theorems 5.1 and 5.2 is to understand what are the *local approximation properties* of the space X_h^r on a single element $K \in \mathcal{T}_h$. Remember that on each element the space X_h^r is made of polynomials of degree r. Therefore, the question we ask is how well we can approximate a given function $v \in H^s(K)$ by a polynomial in $\mathbb{P}_r$ in a domain K.

The following result is valid on any bounded convex Lipschitz domain $K \subset \mathbb{R}^d$ with outer diameter h_K and inner diameter ρ_K.

Lemma 5.3 (Deny-Lions)**.** *Given any bounded convex Lipschitz domain $K \subset \mathbb{R}^d$ and $s \geq 0$, setting $\eta = \min\{s, r+1\}$, there exists $C_{DL} > 0$ such that*

$$\forall v \in H^s(K), \quad \inf_{p \in \mathbb{P}_r(K)} \|v - p\|_{H^m(K)} \leq C_{DL} |v|_{H^\eta}, \quad m = 0, 1, \ldots, \eta, \tag{5.4}$$

with constant $C_{DL} = C_{DL}(h_K, \rho_K, m, s, d)$. Moreover, asymptotically as $h_K \to 0$, the constant C_{DL} scales as $C_{DL} \sim h_K^{\eta-m} \left(\frac{h_K}{\rho_K}\right)^m$.

Constructive proof in 1D. The proof can be done by contradiction (see e.g. [8, Proposition 3.4.4]). We propose here a constructive proof in the 1D case, i.e. K is an interval in $\mathbb{R}$.

The statement is obviously true for $m = s = 0$ as we can just take $p = 0$. We focus then on the case $s \geq 1$. Let $x_0 \in K$ and notice that in 1D a function $v \in H^s(K)$ has at least $\eta - 1$ continuous derivatives. We consider the Taylor expansion of v in x_0, of degree $\eta - 1$:

$$v(x) = T^\eta_{x_0} v(x) + R^\eta(x), \qquad T^\eta_{x_0} v(x) = \sum_{k=0}^{\eta-1} \frac{v^{(k)}(x_0)}{k!}(x - x_0)^k, \qquad R^\eta(x) = \int_{x_0}^x \frac{v^\eta(t)}{(\eta-1)!}(x-t)^{\eta-1} dt.$$

If we differentiate the previous formula $m < \eta$ times we have

$$v^{(m)}(x) = \frac{d^m}{dx^m} T^\eta_{x_0} v(x) + \frac{d^m}{dx^m} R^\eta(x), \quad \text{with } \frac{d^m}{dx^m} R^\eta(x) = \int_{x_0}^x \frac{v^\eta(t)}{(\eta-1-m)!}(x-t)^{\eta-1-m} dt.$$

Notice that

$$\left|\frac{d^m}{dx^m} R^\eta(x)\right|^2 \leq \int_{x_0}^x (v^\eta(t))^2 dt \int_{x_0}^x \frac{(x-t)^{2(\eta-1-m)}}{((\eta-1-m)!)^2} dt \leq \|v^\eta\|_{L^2(K)} \frac{|K|^{2(\eta-m)-1}}{(2(\eta-m)-1)((\eta-1-m)!)^2}$$

so that

$$|v - T^\eta_{x_0} v|_{H^m(K)} = \|v^{(m)} - \frac{d^m}{dx^m} T^\eta_{x_0} v\|_{L^2(K)} = \|\frac{d^m}{dx^m} R^\eta(x)\|_{L^2(K)} \leq C_m |v|_{H^\eta(K)}$$

with $C_m = \frac{|K|^{\eta-m}}{\sqrt{2(\eta-m)-1}(\eta-1-m)!}$. Finally

$$\inf_{p \in \mathbb{P}_r(K)} \|v - p\|_{H^m(K)} \leq \|v - T^\eta_{x_0} v\|_{H^m(K)} \leq (m+1) C_m |v|_{H^\eta(K)}.$$

□

The key point of the Lemma is that the *best approximation error* $\inf_{p \in \mathbb{P}_r} \|v - p\|_{H^m(K)}$ is related to the derivatives of v of order $r+1$ (seminorm $|v|^2_{H^{r+1}(K)} = \sum_{|\boldsymbol{\alpha}|=r+1} \|D^{\boldsymbol{\alpha}} v\|^2_{L^2(K)}$) if v is smooth enough. On the other hand, if v is not smooth enough, that is $v \in H^s$ with $s < r+1$, the best approximation error is related to the highest derivatives of v.

The constructive proof of Lemma 5.3 that we have given in 1D, can be generalized with some care to the multidimensional case. Let $K \subset \mathbb{R}^d$ and consider a Taylor expansion up to degree $\eta - 1$, with $\eta = \min\{s, r+1\}$ in a point $x_0 \in K$

$$T^\eta_{x_0} v(\vec{x}) = \sum_{|\boldsymbol{\alpha}| \leq \eta-1} \frac{1}{\boldsymbol{\alpha}!} D^{\boldsymbol{\alpha}} v(\vec{x}_0)(\vec{x} - \vec{x}_0)^{\boldsymbol{\alpha}} \tag{5.5}$$

where we have used the multi-index notation: $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d)$, $\boldsymbol{\alpha}! = \prod_{i=1}^d \alpha_i!$, $(\vec{x} - \vec{x}_0)^{\boldsymbol{\alpha}} = \prod_{i=1}^d (\vec{x}_i - \vec{x}_{0,i})^{\alpha_i}$, and $D^{\boldsymbol{\alpha}} v = \frac{\partial^{\alpha_1 + \dots + \alpha_d}}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}} v$. Using the integral formula for the residual of the Taylor expansion, we have

$$\begin{aligned} R^\eta v(\vec{x}) =& v(\vec{x}) - T^\eta_{x_0} v(\vec{x}) = \int_0^1 \frac{(1-s)^{\eta-1}}{(\eta-1)!} \frac{d^\eta}{ds^\eta} v(\vec{x}_0 + s(\vec{x} - \vec{x}_0)) ds \\ =& \eta \sum_{|\boldsymbol{\alpha}|=\eta} (\vec{x} - \vec{x}_0)^{\boldsymbol{\alpha}} \int_0^1 \frac{(1-s)^{\eta-1}}{\boldsymbol{\alpha}!} D^{\boldsymbol{\alpha}} v(\vec{x}_0 + s(\vec{x} - \vec{x}_0)) ds \end{aligned}$$

provided $v \in C^\eta(K)$. From this we can easily prove a version of Lemma 5.3 in the spaces $C^m(K)$, $m = 0, \dots, \eta$, defined as

$$C^m(K) = \{ v : K \to \mathbb{R}^d, \quad \max_{|\boldsymbol{\alpha}| \le m} \|D^{\boldsymbol{\alpha}} v(\vec{x})\|_{C^0(K)} < +\infty \}$$

endowed with the norm $\|v\|_{C^m(K)} = \max_{|\boldsymbol{\alpha}| \le m} \|D^{\boldsymbol{\alpha}} v\|_{C^0(K)}$ and semi-norm $|v|_{C^m(K)} = \max_{|\boldsymbol{\alpha}| = m} \|D^{\alpha} v\|_{C^0(K)}$.

Lemma 5.4 (Deny-Lions – version in C^m-spaces)**.** *Given any bounded convex Lipschitz domain $K \subset \mathbb{R}^d$, and $s \ge 0$, setting $\eta = \min\{s, r+1\}$, there exists $C_{DL} > 0$ such that*

$$\forall v \in C^s(K), \qquad \inf_{p \in \mathbb{P}_r(K)} \|v - p\|_{C^m(K)} \le C_{DL} |v|_{C^\eta}; \quad m = 0, 1, \dots, \eta \tag{5.6}$$

with $C_{DL} \sim h_K^{\eta - m}$ as $h_K \to 0$.

Proof. Take $p = T^\eta_{x_0} v$ and estimate

$$\begin{aligned} \|v - T^\eta_{x_0} v\|_{C^0(K)} =& \|R^\eta v\|_{C^0(K)} \\ =& \max_{\vec{x} \in K} \left| \eta \sum_{|\boldsymbol{\alpha}|=\eta} (\vec{x} - \vec{x}_0)^{\boldsymbol{\alpha}} \int_0^1 \frac{(1-s)^{\eta-1}}{\boldsymbol{\alpha}!} D^{\boldsymbol{\alpha}} v(\vec{x}_0 + s(\vec{x} - \vec{x}_0)) ds \right| \\ \le& \sum_{|\boldsymbol{\alpha}|=\eta} \|D^{\boldsymbol{\alpha}} v\|_{C^0(K)} \max_{\vec{x} \in K} \left| \eta (\vec{x} - \vec{x}_0)^{\boldsymbol{\alpha}} \int_0^1 \frac{(1-s)^{\eta-1}}{\boldsymbol{\alpha}!} ds \right| \\ \le& C(\eta)\, h_K^\eta |v|_{C^\eta(K)} \end{aligned}$$

Since $D^{\boldsymbol{\beta}} T^\eta_{x_0} v = T^{\eta - |\boldsymbol{\beta}|}_{x_0} D^{\boldsymbol{\beta}} v$, for any $|\boldsymbol{\beta}| < \eta$

$$\|D^{\boldsymbol{\beta}}(v - T^\eta_{x_0} v)\|_{C^0(K)} = \|D^{\boldsymbol{\beta}} v - T^{\eta - |\boldsymbol{\beta}|}_{x_0} D^{\boldsymbol{\beta}} v\|_{C^0(K)} \le C(\eta - |\boldsymbol{\beta}|) h_K^{\eta - |\boldsymbol{\beta}|} |v|_{C^\eta(K)},$$

hence, for $m = 0, \dots, \eta$,

$$\begin{aligned} |v - T^\eta_{x_0} v|_{C^m(K)} =& \max_{|\boldsymbol{\alpha}| = m} \|D^{\boldsymbol{\alpha}}(v - T^\eta_{x_0} v)\|_{C^0(K)} \\ \le& \max_{|\boldsymbol{\alpha}| = m} C(\eta - |\boldsymbol{\alpha}|)\, h_K^{\eta - |\boldsymbol{\alpha}|}\, |v|_{C^\eta(K)} \le C h_K^{\eta - m} |v|_{C^\eta} \end{aligned}$$

and therefore

$$\inf_{p \in \mathbb{P}_r(K)} \|v - p\|_{C^m(K)} \le \|v - T^\eta_{x_0} v\|_{C^m(K)} \le C h_K^{\eta - m} |v|_{C^\eta}.$$

□

Estimates in $H^m(K)$: The argument based on Taylor expansion can be extended with some care also to prove Lemma 5.3 in the Sobolev spaces $H^m(K)$. Indeed, since the pointwise $(\eta-1)$-th derivative is not well defined for functions in H^η, the idea is to build an averaged Taylor expansion. Let B_K be the largest ball contained in K, with radius ρ_K, and $\forall \vec{y} \in B_K$ consider the Taylor expansion of degree $\eta-1$ centered in $\vec{y}$:

$$T^\eta_{\vec{y}} v(\vec{x}) = \sum_{|\boldsymbol{\alpha}|\leq \eta-1} \frac{1}{\boldsymbol{\alpha}!} D^{\boldsymbol{\alpha}} v(\vec{y})(\vec{x}-\vec{y})^{\boldsymbol{\alpha}}.$$

Then, we can define an averaged Taylor expansion over B_K as

$$T^\eta_{\text{av}} v(\vec{x}) = \frac{1}{|B_K|}\int_{B_K} T^\eta_{\vec{y}} v(\vec{x}) d\vec{y} = \frac{1}{|B_K|}\sum_{|\boldsymbol{\alpha}|\leq \eta-1} \frac{1}{\boldsymbol{\alpha}!}\int_{B_K} D^{\boldsymbol{\alpha}} v(\vec{y})(\vec{x}-\vec{y})^{\boldsymbol{\alpha}} d\vec{y}.$$

which is well defined now for functions $v \in H^\eta$. Observe that it still holds

$$D^{\boldsymbol{\beta}} T^\eta_{\text{av}}\, v = T^{\eta-|\boldsymbol{\beta}|}_{\text{av}}\, D^{\boldsymbol{\beta}}\, v$$

Moreover, the reminder is given by

$$\begin{aligned} R^\eta_{\text{av}} v(\vec{x}) =& v(\vec{x}) - T^\eta_{\text{av}} v(\vec{x}) = \frac{1}{|B_K|}\int_{B_K} \left(v(\vec{x}) - T^\eta_{\vec{y}} v(\vec{x}) \right) d\vec{y} \\ =& \frac{\eta}{|B_K|}\sum_{|\boldsymbol{\alpha}|=\eta} \int_{B_K} (\vec{x}-\vec{y})^{\boldsymbol{\alpha}} \int_0^1 \frac{(1-s)^{\eta-1}}{\boldsymbol{\alpha}!} D^{\boldsymbol{\alpha}} v(\vec{y}+s(\vec{x}-\vec{y})) ds\, d\vec{y} \end{aligned}$$

which is also well defined $\forall v \in H^\eta(K)$. Proceeding in a similar way as for the C^m case (but with integral norms), one can prove the result

$$|v - T^\eta_{\text{av}} v|_{H^m(K)} \leq C |v|_{H^\eta}, \qquad \text{for } m = 0, \ldots, \eta$$

with constant $C \sim h_K^{\eta-m} \left(\frac{h_K}{\rho_K}\right)^m$. For details, see [2, Chapter 4].

5.2 Local interpolation estimates

Let us consider now a function $v \in H^s$, $s \geq 2$ and its interpolant $I^r_h v \in X^r_h$. In this section we focus on quantifying the error $v - I^r_h v$ on a single element $K \in \mathcal{T}_h$. Let us denote by v_K and $I^r_{h,K} v$ the restrictions of v and $I^r_h v$ on K, respectively. We aim at estimating $|v_K - I^r_{h,K} v|_{H^m(K)}$ with $m \leq \eta := \min\{s, r+1\}$.

The path that we follow is to map the quantity $|v_K - I^r_{h,K} v|_{H^m}$ onto the reference element $\hat{K}$ using the affine map $x = F_K(\hat{x}) = B_K \hat{x} + b_K$ introduced in Section 3.1.1. We denote $\hat{v}_K = v_K \circ F_K$ and $\widehat{I^r_{h,K} v} = I^r_{h,K} v \circ F_K$. Notice that

$$\widehat{I^r_{h,K} v} = I^r_{h,K} v \circ F_K = \sum_{i=1}^{N_r} v(a_{i,K}) \varphi_i|_K \circ F_K = \sum_{i=1}^{N_r} \hat{v}_K(\hat{a}_i)\hat{\varphi}_i = I^r_{\hat{K}} \hat{v}_K$$

where $\{a_{i,K}, i = 1, \ldots, N_r\}$ is the set of nodes defining the degrees of freedom on K (vertices for $\mathbb{P}_1$ elements, vertices and midpoints for $\mathbb{P}_2$ elements, etc.) and $\varphi_i|_K$ the corresponding Lagrangian basis functions restricted to K. Similarly, $\{\hat{a}_i\}$ and $\hat{\varphi}_i$ denote the nodes and basis functions on the reference element $\hat{K}$.

The first result we need is how the H^m-seminorm transforms through the mapping F_K.

Lemma 5.5 (Seminorm transformation)**.** *For any $v \in H^m(K)$, $m \geq 0$, let $\hat{v} = v \circ F_K$. Then $\hat{v} \in H^m(\hat{K})$ and there exists $C_{sn} = C_{sn}(m) > 0$ such that*

$$|v|_{H^m(K)} \leq C_{sn} \|B_K^{-1}\|^m |\det B_K|^{\frac{1}{2}} |\hat{v}|_{H^m(\hat{K})} \tag{5.7}$$

$$|\hat{v}|_{H^m(\hat{K})} \leq \hat{C}_{sn} \|B_K\|^m |\det B_K|^{-\frac{1}{2}} |v|_{H^m(K)} \tag{5.8}$$

where $\|B_K\|$ is the spectral norm of the matrix B_K. Moreover, $C_{sn} = \hat{C}_{sn} = 1$ for $m = 0, 1$.

Proof for $m = 0, 1$ only. we give here the proof only for the cases $m = 0, 1$. For $m = 0$ we have

$$\|v\|^2_{L^2(K)} = \int_K v^2(x)dx = \int_{\hat{K}} \hat{v}^2(\hat{x}) |\det B_K| d\hat{x} = |\det B_K| \|\hat{v}\|^2_{L^2(\hat{K})}$$

which proves (5.7) for $m = 0$ with constant $C_{sn} = 1$. Inequality (5.8) can be proved analogously. For $m = 1$ observe first that

$$\partial_{\hat{x}_i} v = \sum_{j=1}^{d} \partial_{x_j} v \frac{\partial x_j}{\partial \hat{x}_i}, \quad \Longrightarrow \quad \hat{\nabla} v = B_K^T \nabla v$$

where $\hat{\nabla}$ denotes the gradient with respect to the variables $\hat{x}$. Hence, we have

$$\begin{aligned} |v|^2_{H^1(K)} &= \int_K |\nabla v(x)|^2 dx = \int_{\hat{K}} |B_K^{-T} \hat{\nabla} \hat{v}(\hat{x})|^2 |\det B_K| d\hat{x} \\ &\leq \int_{\hat{K}} (\|B_K^{-1}\| |\hat{\nabla} \hat{v}(\hat{x})|)^2 |\det B_K| d\hat{x} \leq \|B_K^{-1}\|^2 |\det B_K| |\hat{v}|^2_{H^1(\hat{K})} \end{aligned}$$

which proves (5.7) for $m = 1$, again with constant $C_{sn} = 1$. Inequality (5.8) can be proved analogously. □

Bounds on the spectral norms of B_K and B_K^{-1} have been given in Lemma 3.6. The second result we need concerns the continuity of the interpolant operator $I^r_{\hat{K}}$ on the reference element.

Lemma 5.6 (Continuity of interpolant operator)**.** *Let $I^r_{\hat{K}} : C^0(\hat{K}) \to \mathbb{P}_r(\hat{K})$ be the finite element interpolant operator on the reference element $\hat{K} \subset \mathbb{R}^d$. Then, for $d \leq 3$, $I^r_{\hat{K}}$ is a linear bounded operator from $H^2(\hat{K})$ to any $H^m(\hat{K})$ with $0 \leq m \leq r + 1$, i.e. there exists $C_{I,m} > 0$ such that*

$$\|I^r_{\hat{K}} \hat{v}\|_{H^m(\hat{K})} \leq C_{I,m} \|v\|_{H^2(\hat{K})}, \qquad \forall \hat{v} \in H^2(\hat{K})$$

Proof. We have

$$\|I^r_{\hat{K}} \hat{v}\|_{H^m(\hat{K})} = \|\sum_{i=1}^{N_r} \hat{v}(\hat{a}_i) \hat{\varphi}_i\|_{H^m(\hat{K})} \leq \sum_{i=1}^{N_r} |\hat{v}(\hat{a}_i)| \|\hat{\varphi}_i\|_{H^m(\hat{K})} \leq \|\hat{v}\|_{C^0(\hat{K})} \left(\sum_{i=1}^{N_r} \|\hat{\varphi}_i\|_{H^m(\hat{K})} \right)$$

Since the embedding $H^2(\hat{K}) \hookrightarrow C^0(\hat{K})$ is continuous for $d \leq 3$, there exists $C_I > 0$ such that $\|\hat{v}\|_{C^0(\hat{K})} \leq C_I \|\hat{v}\|_{H^2(\hat{K})}$. Moreover the functions $\hat{\varphi}_i$ are polynomials, hence infinitely differentiable, and the quantity $C_{\varphi,m} = \sum_{i=1}^{N_r} \|\hat{\varphi}_i\|_{H^m(\hat{K})}$ is bounded. The thesis then follows with $C_{I,m} = C_I C_{\varphi,m}$. □

We finally need the important observation that

Lemma 5.7 (Exactness of $I^r_{\hat{K}}$ on $\mathbb{P}_r(\hat{K})$)**.** *The interpolant operator $I^r_{\hat{K}} : C^0(\hat{K}) \to \mathbb{P}_r(\hat{K})$ is exact on $\mathbb{P}_r(\hat{K})$, i.e.*

$$I^r_{\hat{K}}\hat{p} = \hat{p}, \qquad \forall \hat{p} \in \mathbb{P}_r(\hat{K}).$$

Proof. This comes directly from the unisolvency of the set of degrees of freedom on $\hat{K}$, i.e. if a polynomial $\hat{p}$ is such that $\hat{p}(\hat{a}_i) = 0$ for all $i = 1, \dots, N_r$, then $\hat{p} = 0$. □

We have now all the ingredients to prove the following local error estimate

Lemma 5.8 (Local error estimate)**.** *Let $K \in \mathcal{T}_h$ be an element of the mesh with outer diameter h_K and inner diameter ρ_K. Then for $s \geq 2$ and any $0 \leq m \leq \eta := \min\{s, r+1\}$ there exists $C_l = C_l(s, r, m, \hat{K}) > 0$ such that*

$$|v - I^r_{h,K} v|_{H^m(K)} \leq C_l \left(\frac{h_K}{\rho_K}\right)^m h_K^{\eta - m} |v|_{H^\eta(K)}, \qquad v \in H^s(K).$$

Proof. We use in sequence: the seminorm transformation (5.7) in Lemma 5.5, the exactness of $I^r_{\hat{K}}$ on $\mathbb{P}_r(\hat{K})$ and the boundedness of $I^r_{\hat{K}} : H^2(\hat{K}) \to H^m(\hat{K})$. In what follows $\hat{p}$ is an arbitrary polynomial in $\mathbb{P}_r(\hat{K})$. We have

$$\begin{aligned}
|v - I^r_{h,K} v|_{H^m(K)} &\leq C_{sn} \|B_K^{-1}\|^m |\det B_K|^{\frac{1}{2}} |\hat{v} - I^r_{\hat{K}} \hat{v}|_{H^m(\hat{K})} \\
&\leq C_{sn} \|B_K^{-1}\|^m |\det B_K|^{\frac{1}{2}} (|\hat{v} - \hat{p}|_{H^m(\hat{K})} + |\hat{p} - I^r_{\hat{K}} \hat{v}|_{H^m(\hat{K})}) \\
&\leq C_{sn} \|B_K^{-1}\|^m |\det B_K|^{\frac{1}{2}} (|\hat{v} - \hat{p}|_{H^m(\hat{K})} + |I^r_{\hat{K}}(\hat{p} - \hat{v})|_{H^m(\hat{K})}) \\
&\leq C_{sn} \|B_K^{-1}\|^m |\det B_K|^{\frac{1}{2}} (|\hat{v} - \hat{p}|_{H^m(\hat{K})} + C_{I,m} \|\hat{p} - \hat{v}\|_{H^2(\hat{K})}) \\
&\leq C_{sn}(1 + C_{I,m}) \|B_K^{-1}\|^m |\det B_K|^{\frac{1}{2}} \|\hat{v} - \hat{p}\|_{H^{\max\{m,2\}}(\hat{K})}
\end{aligned}$$

Since $\hat{p}$ is arbitrary, we deduce

$$|v - I^r_{h,K} v|_{H^m(K)} \leq C_{sn}(1 + C_{I,m}) \|B_K^{-1}\|^m |\det B_K|^{\frac{1}{2}} \inf_{\hat{p} \in \mathbb{P}_r(\hat{K})} \|\hat{v} - \hat{p}\|_{H^{\max\{m,2\}}(\hat{K})}.$$

Using now the local approximation estimate in Lemma 5.3 (Deny-Lions) and the seminorm transformation (5.8) in Lemma 5.5 we obtain

$$\begin{aligned}
|v - I^r_{h,K} v|_{H^m(K)} &\leq C_{sn}(1 + C_{I,m}) C_{DL} \|B_K^{-1}\|^m |\det B_K|^{\frac{1}{2}} |\hat{v}|_{H^\eta(\hat{K})} \\
&\leq C_{sn} \hat{C}_{sn}(1 + C_{I,m}) C_{DL} \|B_K^{-1}\|^m \|B_K\|^\eta |v|_{H^\eta(K)} \\
&\leq C_{sn} \hat{C}_{sn}(1 + C_{I,m}) C_{DL} \left(\frac{\hat{h}}{\rho_K}\right)^m \left(\frac{h_K}{\hat{\rho}}\right)^\eta |v|_{H^\eta(K)} \\
&\leq C_l \left(\frac{h_K}{\rho_K}\right)^m h_K^{\eta - m} |v|_{H^\eta(K)}
\end{aligned}$$

with $C_l = C_{sn} \hat{C}_{sn}(1 + C_{I,m}) C_{DL} \hat{h}^m \hat{\rho}^{-\eta}$. □

5.3 Global interpolation estimates

We finally derive estimates for the global error $(v - I_h^r v)$. The following result holds, that generalizes Theorem 5.2.

Theorem 5.9. *Given a family of regular triangulations $\{\mathcal{T}_h\}_{h>0}$ of a polygonal domain $\Omega \subset \mathbb{R}^d$, $d \leq 3$ and the space X_h^r of continuous finite elements of degree r, for $s \geq 2$ and $0 \leq m \leq \eta := \min\{s, r+1\}$ it holds*

$$\|v - I_h^r v\|_{H^m_{bro}(\Omega)} \leq C_l \gamma^m \left(\sum_{K \in \mathcal{T}_h} h_K^{2(\eta-m)} |v|^2_{H^\eta(K)} \right)^{\frac{1}{2}}, \qquad \forall v \in H^s(\Omega), \tag{5.9}$$

where C_l is the constant appearing in Lemma 5.8 and $\gamma = \max_{K \in \mathcal{T}_h} h_K/\rho_K$.

Proof. Exploiting the fact that the triangulation is regular, hence $h_K/\rho_K \leq \gamma$ for all $K \in \mathcal{T}_h$ and $h > 0$, we have

$$\begin{aligned}
\|v - I_h^r v\|^2_{H^m_{bro}(\Omega)} &= \sum_{K \in \mathcal{T}_h} \|v - I_h^r v\|^2_{H^m(K)} \\
&\leq \sum_{K \in \mathcal{T}_h} C_l^2 \left(\frac{h_K}{\rho_K} \right)^{2m} h_K^{2(\eta-m)} |v|^2_{H^\eta(K)} \\
&\leq C_l^2 \gamma^{2m} \sum_{K \in \mathcal{T}_h} h_K^{2(\eta-m)} |v|^2_{H^\eta(K)}.
\end{aligned}$$

□

By introducing the global mesh size parameter $h = \max_K h_K$ in (5.9) one easily proves Theorems 5.1 and 5.2, stated at the beginning of the Chapter.

We remark that the result of Theorem 5.9 is stronger than that of Theorem 5.2. The advantage of Theorem 5.9 is that it provides a representation of the interpolation error as the sum of local contributions from each element of the mesh. This is a starting point for *mesh adaptivity*. One could indeed try to drive an adaptive algorithm based on local estimates of the $H^{\min\{s,r+1\}}$-seminorm of the solution on each element K and refine the mesh in those elements for which the estimated seminorm, weighted by the corresponding factor $h_K^{(\eta-m)}$, is large.

Chapter 6

Finite element approximation of elliptic problems – Convergence analysis

We consider again the model problem

$$\begin{cases} -\Delta u = f \quad \text{in} \quad \Omega \\ u = g \quad \text{on} \quad \Gamma_D \\ \partial_n u = d \quad \text{on} \quad \Gamma_N \end{cases} \tag{6.1}$$

and its weak formulation:

$$\text{find } u \in V_g \text{ s.t.} \quad a(u,v) = F(v) \quad \forall v \in V_0$$

with $V_g = \{v \in H^1(\Omega) : v|_{\Gamma_D} = g\}$, $a(u,v) = \int_\Omega \nabla u \cdot \nabla v$, and $F(v) = \int_\Omega fv + \int_{\Gamma_N} dv$. We recall moreover the coercivity and continuity estimates

$$a(u,u) \geq \alpha \|u\|^2_{H^1}, \ \ \forall u \in V_0, \qquad a(u,v) \leq M\|u\|_{H^1}\|v\|_{H^1}, \ \ \forall u,v \in V.$$

with $\alpha = \frac{1}{1+C_p^2}$ and $M = 1$.

Assume Ω polygonal and $\mathcal{T}_h$ a suitable triangulation of Ω which reproduces exactly the boundary $\partial\Omega$ as well as Γ_D and Γ_N. Let $X_h^r = \{v_h \in C^0(\bar{\Omega}), v_h|_K = \mathbb{P}_r(K) \quad \forall\, K \in \mathcal{T}_h\}$ be the space of continuous piecewise polynomials of degree r. Let $V_{h,0} = X_h^r \cap V_0$ and $V_{h,g} = \{v_h \in X_h^r,\ v_h|_{\Gamma_D} = I_h^r g\}$ where $I_h^r g$ is a suitable interpolation of the Dirichlet boundary datum. We recall the *finite element formulation*

$$\text{Find } u_h \in V_{h,g} \text{ s.t.} \quad a(u_h, v_h) = F(v_h) \quad \forall\ v_h \in V_{h,0}. \tag{6.2}$$

6.1 Case of homogeneous Dirichlet boundary conditions

We start by considering the case of homogeneous Dirichlet boundary conditions

$$\text{Find } u_h \in V_{h,0} \text{ s.t.} \quad a(u_h, v_h) = F(v_h) \quad \forall\ v_h \in V_{h,0}.$$

6.1.1 Error estimate in H^1

In this case, since $V_{h,0} \subset V_0$, we can apply Ceá's Lemma 2.1

$$\|u - u_h\|_{H^1} \leq \frac{M}{\alpha} \inf_{v_h \in V_{h,0}} \|u - v_h\|_{H^1}.$$

The best approximation error can further be bounded by the *interpolation error* for which estimates have been derived in Chapter 5:

$$\inf_{v_h \in V_{h,0}} \|u - v_h\|_{H^1} \leq \|u - I_h^r u\|_{H^1} \leq Ch^{\eta-1}|u|_{H^\eta}, \qquad \eta = \min\{r+1, s\}$$

Putting everything together we have the estimate for the error in the H^1-norm

$$\|u - u_h\|_{H^1} \leq Ch^{\eta-1}|u|_{H^\eta}, \qquad \eta = \min\{r+1, s\}.$$

6.1.2 Error estimate in L^2 (Aubin-Nitsche trick)

Let $e = u - u_h$. We aim at estimating $\|e\|_{L^2}$. Define the dual problem:

$$\text{find } \phi \in V_0 \text{ s.t.} \quad a(v, \phi) = \int_\Omega e\, v \quad \forall\, v \in V_0. \tag{6.3}$$

Since $e \in L^2(\Omega)$ and problem (6.2) has smoothing properties (at least in the case of Ω convex and either a full Dirichlet or a full Neumann problem), one has $\phi \in H^2(\Omega)$ and $\|\phi\|_{H^2} \leq C\|e\|_{L^2}$. Then

$$\begin{aligned}
\|e\|_{L^2}^2 =& a(e, \phi) = a(u - u_h, \phi) \\
=& a(u - u_h, \phi - w_h) \quad \text{by Galerkin orthogonality} \\
\leq& M\|u - u_h\|_{H^1}\|\phi - w_h\|_{H^1}
\end{aligned}$$

Hence

$$\|e\|_{L^2}^2 \leq M\|u - u_h\|_{H^1} \inf_{w_h \in V_{h,0}} \|\phi - w_h\|_{H^1}$$

Since $\phi \in H^2$, we have $\inf_{w_h \in V_{h,0}} \|\phi - w_h\|_{H^1} \leq Ch|\phi|_{H^2}$ and

$$\|e\|_{L^2}^2 \leq M\|u - u_h\|_{H^1} Ch|\phi|_{H^2} \leq Ch\|u - u_h\|_{H^1}\|e\|_{L^2}.$$

We have proven the following result:

$$\|u - u_h\|_{L^2} \leq Ch\|u - u_h\|_{H^1} \leq Ch^\eta|u|_{H^\eta}, \qquad \eta = \min\{r+1, s\}. \tag{6.4}$$

Observe that this improved convergence for the L^2 error has been obtained thanks to the smoothing properties of the operator $-\Delta^{-1}$ for which

$$f \in L^2(\Omega) \qquad \longrightarrow \qquad u = (-\Delta^{-1})f \in H^2(\Omega)$$

which holds if Ω is either a convex polygonal domain or a domain with C^2 boundary and for the full Dirichlet or full Neumann problem. One should be careful when mixed boundary conditions are employed or when the domain is a non convex polygon (reentry corners) as the H^2 regularity might not hold.

6.1.3 Error estimate on functionals of the solution

Many times one is interested in computing some specific quantities of interest associated to the solution. Examples are:

$$Q(u) = \int_\Omega u, \qquad Q(u) = \int_\Omega \frac{\partial u}{\partial x_i}, \qquad Q(u) = \int_{\partial\Omega} u.$$

Let $Q : V_0 \to \mathbb{R}$ be a linear functional on V_0. We aim at estimating $|Q(u) - Q(u_h)|$. As for the estimate in the L^2 norm, we introduce the dual (or adjoint) problem

$$\text{find } \phi \in V_0 \text{ s.t.} \quad a(v, \phi) = Q(v) \quad \forall\, v \in V_0$$

Then

$$Q(u) - Q(u_h) = Q(u - u_h) = a(u - u_h, \phi) = a(u - u_h, \phi - w_h), \qquad \forall w_h \in V_{h,0}.$$

The last step follows from the *Galerkin orthogonality*. Therefore

$$|Q(u) - Q(u_h)| \le M\|u - u_h\| \inf_{w_h \in V_{h,0}} \|\phi - w_h\|. \tag{6.5}$$

Assume now that both the primal solution u and the dual solution ϕ are smooth and in particular $u, \phi \in H^{r+1}(\Omega)$. Then,

$$|Q(u) - Q(u_h)| \le Ch^{2r}|u|_{H^{r+1}}|\phi|_{H^{r+1}} \tag{6.6}$$

i.e. *the quantity of interest converges twice as fast as the H^1 norm of the error.*

In the case of possibly non smooth solutions, if $u \in H^s$ and $\phi \in H^{s'}$ the previous result generalizes as

$$|Q(u) - Q(u_h)| \le Ch^{\eta+\eta'-2}|u|_{H^\eta}|\phi|_{H^{\eta'}}$$

with $\eta = \min\{r + 1, s\}$ and $\eta' = \min\{r + 1, s'\}$.

6.1.4 Error estimate in negative norms

Another way to read the previous result is the following. Assume $Q(u) = \int_\Omega \psi u$ with $\psi \in H^m(\Omega)$, then the solution to the adjoint problem is

$$\begin{cases} -\Delta\phi = \psi \\ \phi = 0 \quad \text{on} \quad \Gamma_D \\ \partial_n\phi = 0 \quad \text{on} \quad \Gamma_N. \end{cases}$$

Assume a shift theorem holds, i.e. $\partial\Omega$ is sufficiently smooth and the boundary conditions are such that

$$\psi \in H^m(\Omega) \qquad \longrightarrow \qquad \phi = (-\Delta^{-1})\psi \in H^{m+2}.$$

Then,

$$\begin{aligned} Q(u) - Q(u_h) = \int_\Omega \psi(u - u_h) &\le C\|u - u_h\|_{H^1} \inf_{w_h \in V_{h,0}} \|\phi - w_h\|_{H^1} \\ &\le Ch^{\min\{r,m+1\}}\|u - u_h\|_{H^1}|\phi|_{H^{m+2}} \le Ch^{\min\{r,m+1\}}\|u - u_h\|_{H^1}\|\psi\|_{H^m}. \end{aligned}$$

It follows that

$$\|u-u_h\|_{H^{-m}} = \sup_{\psi\in H^m} \frac{\int_\Omega \psi(u-u_h)}{\|\psi\|_{H^m}} \leq Ch^{\min\{r,m+1\}}\|u-u_h\|_{H^1}. \tag{6.7}$$

The convergence in negative norms is faster than the convergence in H^1 and the more negative the norm is, the faster the convergence provided a shift theorem holds. However, there is a limit in the gain, as we can not gain more than a factor h^r. If we detail the result in (6.7) we have

$$\begin{aligned} &\|u-u_h\|_{H^{-m}} \leq Ch^{m+1}\|u-u_h\|_{H^1}, && 0\leq m\leq r-1\\ &\|u-u_h\|_{H^{-m}} \leq Ch^{r}\|u-u_h\|_{H^1}, && m\geq r. \end{aligned}$$

6.2 Case of non-homogeneous Dirichlet boundary conditions

In this case, the finite element approximation (6.2) is non conforming since $V_{h,g} \nsubseteq V_g$ and

$$u_h|_{\Gamma_D} = I_h^r g \neq u|_{\Gamma_D}.$$

However, a Galerkin orthogonality still holds

$$a(u-u_h, v_h) = 0 \quad \forall\ v_h \in V_{h,0} \subset V_0.$$

6.2.1 Error estimates in H^1

We can not apply Ceá's Lemma straightforwardly but we have to proceed in a slightly different way. Denote by $I_h^r u$ the finite element interpolant of the exact solution u. Then

$$|u-u_h|_{H^1}^2 \leq a(u-u_h, u-u_h) = a(u-u_h, u-I_h^r u) + a(u-u_h, I_h^r u - u_h)$$

Observe now that both u_h and $I_h^r u$ are in the space $V_{h,g}$ so that $I_h^r u - u_h \in V_{h,0}$ and by Galerkin orthogonality $a(u-u_h, I_h^r u - u_h) = 0$. Therefore

$$|u-u_h|_{H^1}^2 \leq a(u-u_h, u-I_h^r u) \leq |u-u_h|_{H^1}|u-I_h^r u|_{H^1}$$

and finally

$$|u-u_h|_{H^1} \leq |u-I_h^r u|_{H^1} \leq Ch^{\eta-1}|u|_{H^\eta}, \quad \eta = \min\{r, s-1\},$$

which is the same result as in the case of homogeneous Dirichlet boundary conditions. A result in the full H^1 norm can be recovered using the following Poincaré type inequality (see e.g. [5])

$$\|u\|_{H^1(\Omega)} \leq C_p(|u|_{H^1(\Omega)} + \|u\|_{L^2(\Gamma_D)}), \qquad \forall u\in H^1(\Omega), \quad \text{if } |\Gamma_D|>0.$$

We have then

$$\begin{aligned} \|u-u_h\|_{H^1(\Omega)} &\leq C_p(|u-u_h|_{H^1(\Omega)} + \|u-u_h\|_{L^2(\Gamma_D)})\\ &\leq C_p(|u-I_h^r u|_{H^1(\Omega)} + \|g-I_h^r g\|_{L^2(\Gamma_D)})\\ &\leq C(h^{\min\{r,s-1\}}|u|_{H^s} + h^{\min\{r+1,s'\}}|g|_{H^{s'}}). \end{aligned}$$

6.2.2 Error estimate in L^2

For a general problem it is not obvious that one can obtain an improved convergence rate when looking at the L^2 norm. However, this is true for the specific problem (6.1) as the following Lemma shows.

Lemma 6.1. *Assuming that the exact solution of problem* (6.1) *satisfies* $u \in H^s(\Omega)$ *and* $g \in H^{s'}(\Gamma_D)$, $s' \geq s - 1/2$, *the finite element solution* u_h *satisfies the estimate*

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^{\min\{r+1,s,s'\}} \left(|u|_{H^s(\Omega)} + |g|_{H^{s'}(\Gamma_D)} \right)$$

provided that problem (6.1) *has smoothing properties, i.e.* $u \in H^2(\Omega)$ *whenever* $f \in L^2(\Omega)$ *and* $d, g = 0$.

Proof. Let $e_h = I_h^r u - u_h \in V_{h,0}$. We first estimate $\|e_h\|_{L^2}$ and conclude by triangular inequality $\|u - u_h\|_{L^2} \leq \|u - I_h^r u\|_{L^2} + \|e_h\|_{L^2}$. Define the dual problem:

$$\text{find } \phi \in V_0 \text{ s.t.} \quad a(v, \phi) = \int_\Omega e_h \, v \quad \forall v \in V_0. \tag{6.8}$$

Since $e_h \in L^2(\Omega)$, if problem (6.8) has smoothing properties, then $\phi \in H^2(\Omega)$ and $\|\phi\|_{H^2} \leq C\|e_h\|_{L^2}$. Observe that the following inequalities hold:

$$\begin{aligned}
&\|\Delta\phi\|_{L^2(\Omega)} \leq C\|\phi\|_{H^2(\Omega)} \\
&\|\partial_n\phi\|_{L^2(\partial\Omega)} \leq C\|\nabla\phi\|_{H^1(\Omega)} \leq C\|\phi\|_{H^2(\Omega)}
\end{aligned}$$

Then

$$\begin{aligned}
\|e_h\|^2_{L^2(\Omega)} &= a(e_h, \phi) = a(I_h^r u - u, \phi) + a(u - u_h, \phi) \\
&\leq a(I_h^r u - u, \phi) + \underbrace{a(u - u_h, \phi - w_h)}_{\text{by Galerkin orth.}} \\
&\leq a(I_h^r u - u, \phi) + M\|u - u_h\|_{H^1} \inf_{w_h \in V_{h,0}} \|\phi - w_h\|_{H^1} \\
&\leq a(I_h^r u - u, \phi) + Ch\|u - u_h\|_{H^1}|\phi|_{H^2}
\end{aligned}$$

We focus now on the term $a(I_h^r u - u, \phi)$:

$$\begin{aligned}
a(I_h^r u - u, \phi) = \int_\Omega \nabla(I_h^r u - u) \cdot \nabla\phi &= -\int_\Omega (I_h^r u - u)\Delta\phi + \underbrace{\int_{\partial\Omega} (I_h^r u - u)\partial_n\phi}_{\partial_n\phi = 0 \text{ on } \Gamma_N} \\
&\leq \|I_h^r u - u\|_{L^2(\Omega)}\|\Delta\phi\|_{L^2(\Omega)} + \|I_h^r u - u\|_{L^2(\Gamma_D)}\|\partial_n\phi\|_{L^2(\Gamma_D)} \\
&\leq C(\|u - I_h^r u\|_{L^2(\Omega)} + \|g - I_h^r g\|_{L^2(\Gamma_D)})\|\phi\|_{H^2(\Omega)}.
\end{aligned}$$

Putting this estimate in the previous one and recalling that $\|\phi\|_{H^2} \leq C\|e_h\|_{L^2}$ we have

$$\|e_h\|^2_{L^2(\Omega)} \leq C(h\|u - u_h\|_{H^1(\Omega)} + \|u - I_h^r u\|_{L^2(\Omega)} + \|g - I_h^r g\|_{L^2(\Gamma_D)})\|e_h\|_{L^2(\Omega)}$$

and by triangular inequality

$$\|u - u_h\|_{L^2(\Omega)} \leq C(h\|u - u_h\|_{H^1(\Omega)} + \|u - I_h^r u\|_{L^2(\Omega)} + \|g - I_h^r g\|_{L^2(\Gamma_D)}). \tag{6.9}$$

Assuming now $u \in H^{r+1}(\Omega)$ and $g \in H^{r+1}(\Gamma_D)$, we have

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^{r+1} \left(|u|_{H^{r+1}(\Omega)} + |g|_{H^{r+1}(\Gamma_D)} \right) \tag{6.10}$$

and for possibly non smooth functions $u \in H^s(\Omega)$, $g \in H^{s'}(\Gamma_D)$

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^{\min\{r+1,s,s'\}} \left(|u|_{H^s(\Omega)} + |g|_{H^{s'}(\Gamma_D)} \right) \tag{6.11}$$

which again is a similar estimate as in the case of homogeneous data. □

6.3 Variational crimes: numerical integration

As we have already discussed, in many cases the integrals appearing in the weak / finite element formulation can not be computed exactly and one often uses quadrature formulas to approximate them. A typical situation is the case of a Poisson problem with non constant coefficients

$$\begin{cases} -\operatorname{div}(\mu \nabla u) = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{6.12}$$

The weak formulation of this problem in the functional space $V = H_0^1(\Omega)$ reads

$$\text{find } u \in V \text{ s.t.} \quad a(u, v) = F(v), \qquad \forall v \in V$$

with $a(u, v) = \int_\Omega \mu(x) \nabla u(x) \cdot \nabla v(x) dx$ and $F(v) = \int_\Omega f(x) v(x) dx$, and its finite element formulation in the space $V_h = X_h^r \cap V = \{v_h \in X_h^r : \ v_{h|_{\partial\Omega}} = 0\}$ reads

$$\text{find } u_h \in V_h \text{ s.t.} \quad a(u_h, v_h) = F(v_h) \quad \forall v_h \in V_h.$$

Here we assume again that no approximation of the domain Ω is induced by the triangulation $\mathcal{T}_h$.

The practical computation of the stiffness matrix $A_{ij} = a(\varphi_j, \varphi_i) = \int_\Omega \mu(x) \nabla \varphi_j(x) \nabla \varphi_i(x) dx$ and the right hand side $F_i = F(\varphi_i) = \int_\Omega f(x) \varphi_i(x) dx$, often require the use of quadrature formulas. Let us consider a quadrature formula on an element $K \in \mathcal{T}_h$

$$Q_K(f) = \sum_{l=1}^{nqp} \omega_{l,K} \, f(z_{l,K}) \ \approx \int_K f(x) dx \tag{6.13}$$

where $z_{l,K}$ are the quadrature knots and $\omega_{l,K}$ the corresponding weights, and the composite formula

$$Q_h(f) = \sum_{K \in \mathcal{T}_h} Q_K(f).$$

Typically, the elementary quadrature formula $Q_K(f)$ is first defined on the reference element $\hat{K}$

$$Q_{\hat{K}}(\hat{f}) = \sum_{l=1}^{nqp} \hat{\omega}_l \, \hat{f}(\hat{z}_l) \approx \int_{\hat{K}} f(\hat{x}) d\hat{x}.$$

Then, introducing the mapping $x = B_K \hat{x} + b_K$ from $\hat{K}$ to K,

$$\int_K f(x) dx = |\det B_K| \int_{\hat{K}} \hat{f}(\hat{x}) d\hat{x} \approx |\det B_K| \, Q_{\hat{K}}(\hat{f})$$

and formula (6.13) will have $\omega_{l,K} = |\det B_K| \, \hat{\omega}_l$ and $z_{l,K} = B_K \hat{z}_l + b_K$. We give here three examples of quadrature formulas on triangles. Let us denote by c_K the barycenter of the triangle K, by $m_{i,K}$, $i = 1, 2, 3$ the mid-points of the edges of K and by $a_{i,K}$, $i = 1, 2, 3$ the vertices of K.

- Formula exact on $\mathbb{P}_1$
$$Q_K(f) = |K| f(c_K),$$
- formula exact on $\mathbb{P}_2$
$$Q_K(f) = \frac{|K|}{3} \sum_{i=1}^{3} f(m_{i,K}),$$
- formula exact on $\mathbb{P}_3$
$$Q_K(f) = |K| \left[\frac{1}{20} \sum_{i=1}^{3} f(a_{i,K}) + \frac{2}{15} \sum_{j=1}^{3} f(m_{j,K}) + \frac{9}{20} f(c_K) \right].$$

We define now an approximate bilinear form $a_h(\cdot,\cdot)$ and forcing term $F_h(\cdot)$ as

$$a_h(u_h, v_h) = Q_h(\mu \nabla u_h \cdot \nabla v_h) = \sum_{K \in \mathcal{T}_h} \sum_{l=1}^{nqp} \omega_{l,K}\, \mu(z_{l,K}) \nabla u_h(z_{l,K}) \cdot \nabla v_h(z_{l,K})$$
$$F_h(v_h) = Q_h(f v_h) = \sum_{K \in \mathcal{T}_h} \sum_{l=1}^{nqp} \omega_{l,K}\, f(z_{l,K}) v_h(z_{l,K})$$

and introduce the **generalized Galerkin formulation**

$$\text{find } u_h^* \in V_h \text{ s.t.} \quad a_h(u_h^*, v_h) = F_h(v_h) \quad \forall v_h \in V_h. \tag{6.14}$$

Observe that, in general, $a_h(\cdot,\cdot)$ is well defined only in the discrete space V_h (which contains only continuous functions) but not in the continuous space $H_0^1(\Omega)$. Indeed $H_0^1(\Omega) \not\subseteq C^0(\Omega)$ for $d > 1$, so we are not allowed to take point values for H^1 functions and the bilinear form $a_h(\cdot,\cdot)$ is not continuous in $H^1(\Omega)$.

A general result on the generalized Galerkin formulation is the following:

Lemma 6.2 (Strang's Lemma)**.** *Assume that*

- $a_h : V_h \times V_h \to \mathbb{R}$ *is continuous and uniformly coercive in* V_h*, i.e.* $\exists\ \alpha^* > 0$*:*
$$a_h(v_h, v_h) \geq \alpha^* \|v_h\|_V^2 \quad \forall\ h > 0, \quad \forall\ v_h \in V_h \tag{6.15}$$
- $F_h : V_h \to \mathbb{R}$ *is bounded.*

Then

1. *there exists a unique solution* $u_h^* \in V_h$ *to problem* (6.14) *that satisfies*
$$\|u_h^*\|_V \leq \frac{1}{\alpha^*} \sup_{v_h \in V_h} \frac{F_h(v_h)}{\|v_h\|_V}$$
2. *it holds*
$$\|u - u_h^*\|_V \leq \inf_{w_h \in V_h} \left\{ (1 + \frac{M}{\alpha^*}) \|u - w_h\|_V + \frac{1}{\alpha^*} \sup_{v_h \in V_h} \frac{a(w_h, v_h) - a_h(w_h, v_h)}{\|v_h\|_V} \right\} + \frac{1}{\alpha^*} \sup_{v_h \in V_h} \frac{F(v_h) - F_h(v_h)}{\|v_h\|_V} \tag{6.16}$$

Proof. The first part is just an application of Lax-Milgram's Lemma in V_h. For the second part, let $w_h \in V_h$ be arbitrary, then

$$\begin{aligned}
\alpha^* \|u_h^* - w_h\|_V^2 \leq & a_h(u_h^* - w_h, u_h^* - w_h) = F_h(u_h^* - w_h) - a_h(w_h, u_h^* - w_h) \\
= & a(u, u_h^* - w_h) - F(u_h^* - w_h) + F_h(u_h^* - w_h) - a_h(w_h, u_h^* - w_h) \pm a(w_h, u_h^* - w_h) \\
= & a(u - w_h, u_h^* - w_h) + [a(w_h, u_h^* - w_h) - a_h(w_h, u_h^* - w_h)] \\
& + [F_h(u_h^* - w_h) - F(u_h^* - w_h)] \\
\leq & M\|u - w_h\|_V \|u_h^* - w_h\|_V + \|u_h^* - w_h\|_V \sup_{v_h \in V_h} \frac{a(w_h, v_h) - a_h(w_h, v_h)}{\|v_h\|_V} \\
& + \|u_h^* - w_h\|_V \sup_{v_h \in V_h} \frac{F_h(v_h) - F(v_h)}{\|v_h\|_V}
\end{aligned}$$

from which the thesis follows by triangular inequality. □

Estimate (6.16) contains 3 terms. The first one $\inf_{w_h \in V_h}(1 + M/\alpha^*)\|u - w_h\|_V$ is the standard best approximation error in V_h. The other two

$$\sup_{v_h \in V_h} \frac{a(w_h, v_h) - a_h(w_h, v_h)}{\|v_h\|_V}, \quad \sup_{v_h \in V_h} \frac{F(v_h) - F_h(v_h)}{\|v_h\|}$$

represent consistency errors due to the quadrature formula.

The important questions are: whether these two errors are of the same order of the best approximation error; and whether the quadrature formula leads to a uniformly coercive discrete bilinear form $a_h(\cdot, \cdot)$. The following result holds (see e.g. [4]):

Lemma 6.3. *Assume that the quadrature formula $Q_{\hat{K}}$ is exact on $\mathbb{P}_p(\hat{K})$, $p \geq r$, and the mesh is regular with $\gamma \rho_K \geq h_K$, $\forall K \in \mathcal{T}_h$. Then, for any $u \in H^{r+1}$ and its finite element interpolation $I_h^r u \in X_h^r$, it holds*

$$\sup_{v_h \in V_h} \frac{a(I_h^r u, v_h) - a_h(I_h^r u, v_h)}{\|v_h\|_V} \leq C\gamma^2 h^{p-r+2} \|\mu\|_{W^{p-r+2,\infty}(\Omega)} \|u\|_{H^{r+1}(\Omega)} \tag{6.17}$$

$$\sup_{v_h \in V_h} \frac{F(v_h) - F_h(v_h)}{\|v_h\|_V} \leq C h^{p-r+2} \|f\|_{H^{p-r+2}(\Omega)} \tag{6.18}$$

Proof. For any $\hat{g} \in C^0(\hat{K})$, let us define $E_{\hat{K}}(\hat{g}) = \int_{\hat{K}} \hat{g}(\hat{x}) d\hat{x} - Q_{\hat{K}}(\hat{g})$. By assumption $E_{\hat{K}}(\hat{g}) = 0$,

$\forall \hat{g} \in \mathbb{P}_p(\hat{K})$. Let us consider first the term involving the bilinear form $a(\cdot,\cdot)$.

$$
\begin{aligned}
a(w_h, v_h) - a_h(w_h, v_h) &= \sum_{K\in\mathcal{T}_h} \int_K \mu \nabla w_h \cdot \nabla v_h - Q_K(\mu \nabla w_h \cdot \nabla v_h) \\
&= \sum_{K\in\mathcal{T}_h} |\det B_K|\, E_{\hat{K}}(\hat{\mu}_K \underbrace{B_K^{-T}\hat{\nabla}\hat{w}_{h,K}}_{\hat{p}_K} \cdot \underbrace{B_K^{-T}\hat{\nabla}\hat{v}_{h,K}}_{\hat{q}_K}) \qquad [\text{with } \hat{p}_K, \hat{q}_K \in \mathbb{P}_{r-1}(\hat{K})] \\
&= \sum_{K\in\mathcal{T}_h} |\det B_K|\, E_{\hat{K}}((\hat{\mu}_K \hat{p}_K - I_{\hat{K}}^{p-r+1}(\hat{\mu}_K \hat{p}_K)) \cdot \hat{q}_K) \\
&\le \sum_{K\in\mathcal{T}_h} |\det B_K|\, C |\hat{\mu}_K \hat{p}_K|_{W^{p-r+2,\infty}(\hat{K})} \|\hat{q}_K\|_{L^\infty(\hat{K})} \\
&\le C \sum_{K\in\mathcal{T}_h} |\det B_K| \left(\sum_{j=0}^{p-r+2} |\hat{\mu}_K|_{W^{p-r+2-j,\infty}(\hat{K})} |\hat{p}_K|_{W^{j,\infty}(\hat{K})} \right) \|\hat{q}_K\|_{L^\infty(\hat{K})} \\
&\le C \sum_{K\in\mathcal{T}_h} |\det B_K| \left(\sum_{j=0}^{p-r+2} |\hat{\mu}_K|_{W^{p-r+2-j,\infty}(\hat{K})} |\hat{p}_K|_{H^{j,\infty}(\hat{K})} \right) \|\hat{q}_K\|_{L^2(\hat{K})} \\
&\qquad [\text{by equivalence of norms in finite dimensional spaces}] \\
&\le C \sum_{K\in\mathcal{T}_h} |\det B_K|\, \|B_K^{-1}\|^2 \sum_{j=0}^{r-1} |\hat{\mu}_K|_{W^{p-r+2-j,\infty}(\hat{K})} |\hat{\nabla}\hat{w}_{h,K}|_{H^j(\hat{K})} \|\hat{\nabla}\hat{v}_{h,K}\|_{L^2(\hat{K})} \\
&\le C \sum_{K\in\mathcal{T}_h} |\det B_K|\, \|B_K^{-1}\|^2 \sum_{j=0}^{r-1} |\hat{\mu}_K|_{W^{p-r+2-j,\infty}(\hat{K})} |\hat{w}_{h,K}|_{H^{j+1}(\hat{K})} |\hat{v}_{h,K}|_{H^1(\hat{K})}
\end{aligned}
$$

Then take $w_h = I_h^r u$ and note that for $i = 0, \dots, r$,

$$
\begin{aligned}
|\hat{I}_{\hat{K}}^r \hat{u}_K|_{H^i(\hat{K})} &\le |\hat{u}_K|_{H^i(\hat{K})} + |\hat{u}_K - \hat{I}_{\hat{K}}^r \hat{u}_K|_{H^i(\hat{K})} \le |\hat{u}_K|_{H^i(\hat{K})} + C|\hat{u}_K|_{H^{r+1}(\hat{K})} \\
&\le C |\det B_K|^{-\frac{1}{2}} \|B_K\|^i (|u|_{H^i(K)} + \|B_K\|^{r+1-i} |u|_{H^{r+1}(K)}) \\
&\le C |\det B_K|^{-\frac{1}{2}} \|B_K\|^i \|u\|_{H^{r+1}(K)}.
\end{aligned}
$$

Hence

$$
a(I_h^r u, v_h) - a_h(I_h^r u, v_h) \le C \sum_{K\in\mathcal{T}_h} \|B_K^{-1}\|^2 \|B_K\|^{p-r+4} \|\mu\|_{W^{p-r+2,\infty}(K)} \|u\|_{H^{r+1}(K)} |v_h|_{H^1(K)}
$$

which leads to (6.17).

Let us consider now the consistency error on the right hand side.

$$
\begin{aligned}
F(v_h) - F_h(v_h) &= \sum_{K\in\mathcal{T}_h} \int_K f v_h - Q_K(f v_h) \\
&= \sum_{K\in\mathcal{T}_h} |\det B_K| E_{\hat{K}}(\hat{f}_K \hat{v}_{h,K}), \qquad [\text{setting } \bar{v}_K = \frac{1}{|\hat{K}|} \int_{\hat{K}} \hat{v}_{h,K}] \\
&= \sum_{K\in\mathcal{T}_h} |\det B_K| \left[\underbrace{E_{\hat{K}}(\hat{f}_K(\hat{v}_{h,K} - \bar{v}_K))}_{(A)} + \underbrace{E_{\hat{K}}(\hat{f}_K)\bar{v}_K}_{(B)} \right]
\end{aligned}
$$

For a quadrature formula exact on $\mathbb{P}_p(\hat{K})$ one has for any $1 \le s \le p$ and dimension $d \le 3$,

$$E_{\hat{K}}(\hat{f}_K) = E_{\hat{K}}(\hat{f}_K - I^s_{\hat{K}}\hat{f}_K) \le C\|\hat{f}_K - I^s_{\hat{K}}\hat{f}_K\|_{L^\infty(\hat{K})} \le C\|\hat{f}_K - I^s_{\hat{K}}\hat{f}_K\|_{H^2(\hat{K})} \le C|\hat{f}_K|_{H^{s+1}(\hat{K})}.$$

We take in particular $s = p - r + 1$. Moreover, $\bar{v}_K \le \|\hat{v}_{h,K}\|_{L^2(\hat{K})}$, so that

$$(B) \le C|\hat{f}_K|_{H^{p-r+2}(\hat{K})}\|\hat{v}_{h,K}\|_{L^2(\hat{K})}.$$

On the other hand,

$$\begin{aligned}
(A) &\le E_{\hat{K}}(\hat{f}_K(\hat{v}_{h,K} - \bar{v}_K)) \le E_{\hat{K}}((\hat{f}_K - I^{p-r}_{\hat{K}}\hat{f}_K)(\hat{v}_{h,K} - \bar{v}_K)) \\
&\le C\|\hat{f}_K - I^{p-r}_{\hat{K}}\hat{f}_K\|_{L^\infty(\hat{K})}\|\hat{v}_{h,K} - \bar{v}_K\|_{L^\infty(\hat{K})} \\
&\le C\|\hat{f}_K - I^{p-r}_{\hat{K}}\hat{f}_K\|_{H^2(\hat{K})}\|\hat{v}_{h,K} - \bar{v}_K\|_{L^2(\hat{K})}, \qquad \text{[by norm equiv. on } \mathbb{P}_r(\hat{K})] \\
&\le C|\hat{f}_K|_{H^{p-r+1}(\hat{K})}|\hat{v}_{h,K}|_{H^1(\hat{K})}
\end{aligned}$$

and finally

$$\begin{aligned}
F(v_h) - F_h(v_h) &\le C \sum_{K\in\mathcal{T}_h} |\det B_K| \left[|\hat{f}_K|_{H^{p-r+1}(\hat{K})}|\hat{v}_{h,K}|_{H^1(\hat{K})} + |\hat{f}_K|_{H^{p-r+2}(\hat{K})}\|\hat{v}_{h,K}\|_{L^2(\hat{K})}\right] \\
&\le C \sum_{K\in\mathcal{T}_h} \|B_K\|^{p-r+2} \left[|f|_{H^{p-r+1}(K)}|v_h|_{H^1(K)} + |f|_{H^{p-r+2}(K)}\|v_h\|_{L^2(K)}\right]
\end{aligned}$$

hence

$$\sup_{v_h\in V_h} \frac{F(v_h) - F_h(v_h)}{\|v_h\|_V} \le Ch^{p-r+2}\|f\|_{H^{p-r+2}(\Omega)}$$

and this concludes the proof. □

From the previous Lemma, we see that if we wish to have a consistency error of the same order of the finite element approximation error, we have to take $p - r + 2 = r$ which implies $p = 2r - 2$. The following result follows easily from Lemmas 6.3 and 6.2:

Lemma 6.4. *Assume that the quadrature formula $Q_{\hat{K}}$ is exact on $\mathbb{P}_{2r-2}(\hat{K})$ and has positive weights. Then problem* (6.14) *is well posed and*

$$\|u - u^*_h\|_{H^1} \le Ch^r(|u|_{H^{r+1}} + \|u\|_{H^{r+1}}\|\mu\|_{W^{r,\infty}} + \|f\|_{H^r}) \tag{6.19}$$

Proof. We first show that if $Q_{\hat{K}}$ is exact on $\mathbb{P}_{2r-2}(\hat{K})$ and has positive weights, then the discrete bilinear form $a_h(u_h, v_h)$ is uniformly coercive. Indeed

$$\begin{aligned}
a_h(v_h, v_h) &= \sum_{K\in\mathcal{T}_h}\sum_{l=1}^{nqp} \omega_{l,K}\,\mu(z_{l,K})|\nabla v_h(z_{l,K})|^2 \\
&\ge \min_{x\in\Omega}\mu(x) \underbrace{\sum_{K\in\mathcal{T}_h}\sum_{l=1}^{nqp} \omega_{l,K}\,|\nabla v_h(z_{l,K})|^2}_{=\|\nabla v_h\|^2_{L^2(\Omega)} \text{ since } Q_K \text{ is exact on } \mathbb{P}_{2r-2}} .
\end{aligned}$$

By Strang's Lemma 6.2, it follows that the problem (6.14) is well posed. Still from Strang's Lemma and Lemma 6.3, it holds

$$\begin{aligned}\|u-u_h^*\|_V &\le \inf_{w_h\in V_h}\left\{(1+\frac{M}{\alpha^*})\|u-w_h\|_V+\frac{1}{\alpha^*}\sup_{v_h\in V_h}\frac{a(w_h,v_h)-a_h(w_h,v_h)}{\|v_h\|_V}\right\}+\frac{1}{\alpha^*}\sup_{v_h\in V_h}\frac{F(v_h)-F_h(v_h)}{\|v_h\|_V}\\ &\le (1+\frac{M}{\alpha^*})\|u-I_h^r u\|_V+\frac{1}{\alpha^*}\sup_{v_h\in V_h}\frac{a(I_h^r u,v_h)-a_h(I_h^r u,v_h)}{\|v_h\|_V}+\frac{1}{\alpha^*}\sup_{v_h\in V_h}\frac{F(v_h)-F_h(v_h)}{\|v_h\|_V}\\ &\le Ch^r\left(|u|_{H^{r+1}}+\|\mu\|_{W^{r,\infty}}\|u\|_{H^{r+1}}+\|f\|_{H^r}\right).\end{aligned}$$

□

Error estimate on functionals:

We consider now the approximation of a quantity of interest $Q(u)$ by $Q(u_h^*)$, where $Q: V\to\mathbb{R}$ is a linear functional. Observe that in the generalized Galerkin framework, the Galerkin orthogonality does not hold any more. However, we can derive the following generalization of the Galerkin orthogonality:

$$a(u-u_h^*,v_h)=[F(v_h)-F_h(v_h)]-[a(u_h^*,v_h)-a_h(u_h^*,v_h)],\qquad \forall v_h\in V_h \tag{6.20}$$

We define now the (continuous) dual problem

$$\text{find } \phi\in V \text{ s.t.}\quad a(v,\phi)=Q(v)\quad \forall v\in V \tag{6.21}$$

and the finite element interpolant $I_h^r\phi\in V_h$. We thus have the following characterization of the error on the quantity of interest

$$\begin{aligned}Q(u)-Q(u_h^*)&=a(u-u_h^*,\phi)\\ &=a(u-u_h^*,\phi-I_h^r\phi)+[F(I_h^r\phi)-F_h(I_h^r\phi)]-[a(u_h^*,I_h^r\phi)-a_h(u_h^*,I_h^r\phi)]\end{aligned} \tag{6.22}$$

For the consistency errors $F(I_h^r\phi)-F_h(I_h^r\phi)$ and $a(u_h^*,I_h^r\phi)-a_h(u_h^*,I_h^r\phi)$ a similar result as in Lemma 6.3 can be established (see [1]):

Lemma 6.5. *Assume that the quadrature formula $Q_{\hat K}$ is exact on $\mathbb{P}_p(\hat K)$, $p\ge r$, and the mesh is regular with $h_K\le\gamma\rho_K$, $\forall K\in\mathcal{T}_h$. Then, for $u,\phi\in H^{r+1}$ and u_h^* solution of (6.14), it holds*

$$a(u_h^*,I_h^r\phi_h)-a_h(u_h^*,I_h^r\phi_h)\le C\gamma^2\|\mu\|_{W^{p+1,\infty}(\Omega)}\|\phi\|_{H^{r+1}(\Omega)}\left(h^{p+1}\|u\|_{H^{r+1}(\Omega)}+h^{p+1-r}\|u-u_h^*\|_{H^1(K)}\right) \tag{6.23}$$

$$F(I_h^r\phi)-F_h(I_h^r\phi)\le Ch^{p+1}\|f\|_{H^{p+1}(\Omega)}\|\phi\|_{H^{r+1}(\Omega)} \tag{6.24}$$

Proof. We start with the consistency error on the right hand side.

$$F(v_h)-F_h(v_h)=\sum_{K\in\mathcal{T}_h}\int_K fv_h-Q_K(fv_h)=\sum_{K\in\mathcal{T}_h}|\det B_K|E_{\hat K}(\hat f_K\hat v_{h,K})$$

Since $E_{\hat{K}}$ is exact on $\mathbb{P}_p(\hat{K})$, $p \geq 1$, we can further write

$$\begin{aligned}
E_{\hat{K}}(\hat{f}_K \hat{v}_{h,K}) &= E_{\hat{K}}(\hat{f}_K \hat{v}_{h,K} - I_h^p(\hat{f}_K \hat{v}_{h,K})) \leq C|\hat{f}_K \hat{v}_{h,K}|_{H^{p+1}(\hat{K})} \\
&\leq C \sum_{j=0}^{p+1} |\hat{f}_K|_{H^{p+1-j}(\hat{K})} |\hat{v}_{h,K}|_{W^{j,\infty}(\hat{K})} \\
&\leq C \sum_{j=0}^{r} |\hat{f}_K|_{H^{p+1-j}(\hat{K})} |\hat{v}_{h,K}|_{H^j(\hat{K})}, \qquad \text{[by norm equiv. on } \mathbb{P}_r(\hat{K})]
\end{aligned}$$

Taking now $v_h = I_h^r \phi$ and proceeding as in the proof of Lemma 6.3 we have for any $i = 0, \ldots, r$, $|\hat{I}_{\hat{K}}^r \hat{\phi}_K|_{H^i(\hat{K})} \leq C|\det B_K|^{-\frac{1}{2}} \|B_K\|^i \|\phi\|_{H^{r+1}(K)}$ and

$$F(I_h^r \phi) - F_h(I_h^r \phi) \leq C \sum_{K \in \mathcal{T}_h} \|B_K\|^{p+1} \|f\|_{H^{p+1}(K)} \|\phi\|_{H^{r+1}(K)} \leq C h^{p+1} \|f\|_{H^{p+1}(\Omega)} \|\phi\|_{H^{r+1}(\Omega)}.$$

Concerning the consistency error for the bilinear form, we have

$$a(w_h, v_h) - a_h(w_h, v_h) = \sum_{K \in \mathcal{T}_h} \int_K \mu \nabla w_h \cdot \nabla v_h - Q_K(\mu \nabla w_h \cdot \nabla v_h) = \sum_{K \in \mathcal{T}_h} |\det B_K| E_{\hat{K}}(\hat{\mu}_K \hat{p}_K \hat{q}_K)$$

with $\hat{p}_K = B_K^{-T} \hat{\nabla} \hat{w}_{h,K}$ and $\hat{q}_K = B_K^{-T} \hat{\nabla} \hat{v}_{h,K}$. Since $E_{\hat{K}}$ is exact on $\mathbb{P}_p(\hat{K})$, $p \geq 1$, we can further write

$$\begin{aligned}
E_{\hat{K}}(\hat{\mu}_K \hat{p}_K \hat{q}_K) &= E_{\hat{K}}(\hat{\mu}_K \hat{p}_K \hat{q}_K - I_h^p(\hat{\mu}_K \hat{p}_K \hat{q}_K)) \leq C|\hat{\mu}_K \hat{p}_K \hat{q}_K|_{H^{p+1}(\hat{K})} \\
&\leq C \sum_{j=0}^{r-1} |\hat{\mu}_K \hat{p}_K|_{W^{p+1-j,\infty}(\hat{K})} |\hat{q}_K|_{H^j(\hat{K})} \\
&\leq C \sum_{j=0}^{r-1} \sum_{s=0}^{p+1-j} |\hat{\mu}_K|_{W^{p+1-j-s,\infty}(\hat{K})} |\hat{p}_K|_{W^{s,\infty}(\hat{K})} |\hat{q}_K|_{H^j(\hat{K})} \\
&\leq C\|B_K^{-1}\|^2 \sum_{j=0}^{r-1} \sum_{s=0}^{p+1-j} |\hat{\mu}_K|_{W^{p+1-j-s,\infty}(\hat{K})} |\hat{\nabla} \hat{w}_{h,K}|_{H^s(\hat{K})} |\hat{\nabla} \hat{v}_{h,K}|_{H^j(\hat{K})} \\
&\leq C\|B_K^{-1}\|^2 \sum_{j=0}^{r-1} \sum_{s=0}^{\eta_j} |\hat{\mu}_K|_{W^{p+1-j-s,\infty}(\hat{K})} |\hat{w}_{h,K}|_{H^{s+1}(\hat{K})} |\hat{v}_{h,K}|_{H^{j+1}(\hat{K})}, \quad \eta_j = \min\{p+1-j, r-1\}
\end{aligned}$$

We now take $v_h = I_h^r \phi$ for which we have $|\hat{I}_{\hat{K}}^r \hat{\phi}_K|_{H^i(\hat{K})} \leq C|\det B_K|^{-\frac{1}{2}} \|B_K\|^i \|\phi\|_{H^{r+1}(K)}$ and $w_h = u_h^*$, for which we have

$$\begin{aligned}
|\hat{u}_{h,K}^*|_{H^i(\hat{K})} &\leq |\hat{u}_K|_{H^i(\hat{K})} + |\hat{u}_K - \hat{I}_{\hat{K}}^r \hat{u}_K|_{H^i(\hat{K})} + |\hat{I}_{\hat{K}}^r \hat{u}_K - \hat{u}_{h,K}^*|_{H^i(\hat{K})} \\
&\leq C(|\hat{u}_K|_{H^i(\hat{K})} + |\hat{u}_K|_{H^{r+1}(\hat{K})} + \|\hat{I}_{\hat{K}}^r \hat{u}_K - \hat{u}_{h,K}^*\|_{H^1(\hat{K})}), \qquad \text{[by norm equiv. in } \mathbb{P}_r(\hat{K})] \\
&\leq C(|\hat{u}_K|_{H^i(\hat{K})} + |\hat{u}_K|_{H^{r+1}(\hat{K})} + \|\hat{I}_{\hat{K}}^r \hat{u}_K - \hat{u}_K\|_{H^1(\hat{K})} + \|\hat{u}_K - \hat{u}_{h,K}^*\|_{H^1(\hat{K})}) \\
&\leq C(|\hat{u}_K|_{H^i(\hat{K})} + |\hat{u}_K|_{H^{r+1}(\hat{K})} + \|\hat{u}_K - \hat{u}_{h,K}^*\|_{H^1(\hat{K})}) \\
&\leq C|\det B_k|^{-\frac{1}{2}} (\|B_K\|^i \|u\|_{H^{r+1}(K)} + \|u - u_h^*\|_{H^1(K)})
\end{aligned}$$

Finally

$$\begin{aligned}a(u_h^*, I_h^r\phi) - a_h(u_h^*, I_h^r\phi) &\le C\sum_{K\in\mathcal{T}_h} \|B_K^{-1}\|^2\Big(\|B_K\|^{p+3}\|\mu\|_{W^{p+1,\infty}(K)}\|u\|_{H^{r+1}(K)}\|\phi\|_{H^{r+1}(K)} \\ &\quad + \sum_{j=0}^{r-1}\sum_{s=0}^{\eta_j}\|B_K\|^{p+2-s}\|\mu\|_{W^{p+1,\infty}(K)}\|u-u_h^*\|_{H^1(K)}\|\phi\|_{H^{r+1}(K)}\Big) \\ &\le C\gamma^2\|\mu\|_{W^{p+1,\infty}(\Omega)}\|\phi\|_{H^{r+1}(\Omega)}\left(h^{p+1}\|u\|_{H^{r+1}(\Omega)} + h^{p+1-r}\|u-u_h^*\|_{H^1(K)}\right)\end{aligned}$$

□

From the previous Lemma we see that for the consistency error to be of order h^{2r} we need $p+1 \ge 2r$, hence $p \ge 2r-1$ a slighly stronger condition than the one of Lemma 6.4:

Lemma 6.6. *Assume that the quadrature formula $Q_{\hat{K}}$ is exact on $\mathbb{P}_{2r-1}(\hat{K})$ and has positive weights. If the solution u of (6.12) as well as the dual solution ϕ of (6.21) satisfy $u, \phi \in H^{r+1}(\Omega)$, then*

$$Q(u) - Q(u_h^*) \le C(\mu,\gamma)h^{2r}(\|u\|_{H^{r+1}(\Omega)} + \|f\|_{H^{2r}(\Omega)})\|\phi\|_{H^{r+1}(\Omega)}$$

where the constant C depends on $\|\mu\|_{W^{2r,\infty}(\Omega)}$ and γ but is otherwise independent of h.

Proof. Using the characterization (6.22) of the error on the Quantity of Interest we have

$$\begin{aligned}Q(u) - Q(u_h^*) &= a(u-u_h^*, \phi - I_h^r\phi) + [F(I_h^r\phi) - F_h(I_h^r\phi)] - [a(u_h^*, I_h^r\phi) - a_h(u_h^*, I_h^r\phi)] \\ &\le M\|u-u_h^*\|_{H^1(\Omega)}\|\phi - I_h^r\phi\|_{H^1(\Omega)} + Ch^{2r}\|f\|_{H^{2r}(\Omega)}\|\phi\|_{H^{r+1}(\Omega)} \\ &\quad + C\gamma^2\|\mu\|_{W^{2r,\infty}(\Omega)}\|\phi\|_{H^{r+1}(\Omega)}\left(h^{2r}\|u\|_{H^{r+1}(\Omega)} + h^r\|u-u_h^*\|_{H^1(K)}\right)\end{aligned}$$

Replacing the result of Lemma 6.4 and the interpolation error estimate $\|\phi - I_h^r\phi\|_{H^1(\Omega)} \le Ch^r|\phi|_{H^{r+1}(\Omega)}$ leads to the desired result. □

Bibliography

[1] I. Babuška, U. Banerjee, and H. Li. The effect of numerical integration on the finite element approximation of linear functionals. *Numer. Math,*, 117:65–88, 2011.

[2] S.C. Brenner and R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer, 2008.

[3] Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.

[4] P.G. Ciarlet. *The Finite Element Method for Elliptic Problems*. Society for Industrial and Applied Mathematics, 2002.

[5] A. Ern and J.L. Guermond. *Theory and Practice of Finite Elements*. Springer, 2004.

[6] V. Girault and P-A. Raviart. *Finite element methods for Navier-Stokes equations*. Springer-Verlag, Berlin, 1986. Theory and algorithms.

[7] A. Quarteroni. *Numerical Models for Differential Problems*. Springer, 2010.

[8] A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations*. Springer, 2008.

[9] S. Salsa. *Partial Differential Equations in Action: From Modelling to Theory*. Springer, 2008.