Introduction to the Finite Element method

Gustavo C. Buscaglia

Graduate course

ICMC-USP, São Carlos, Brasil
gustavo.buscaglia@gmail.com

Motivation

A PDE:

$$\mathcal{L}u = f \qquad \text{in } \Omega \subset \mathbb{R}^n, \qquad \mathcal{B}u = g \qquad \text{on } \partial\Omega$$

A numerical approximation:

$$\underline{\underline{A}}\,\underline{U} = \underline{R} \qquad \rightarrow \qquad u_h$$

- **Existence** of u, u_h.
- **Uniqueness** of u, u_h.
- **Well-posedness:** Continuous dependence on the data.
- **Convergence:** A **numerical method** is a **systematic** way of constructing approximations to u, in such a way that the difference $u - u_h$ can be made arbitrarily small (in what sense?).
- **Robustness:** u_h is **not** exact, there is some **error** but... is it an error one can tolerate (qualitatively speaking)?

Motivation

Finite Element Method: When the PDE is **elliptic**, the most popular approximation method is the FEM. It is **general**, **geometrically flexible**, **easy to code**, **robust**, etc. etc.

Understanding PDE's/FEM requires generalizations of the basic tools of linear algebra:

- The spaces are infinite dimensional.
- The "matrices" are now "operators" between such spaces.
- The rank theorem $\dim(\mathrm{Ker}(\underline{\underline{A}}))+\dim(\mathrm{Im}(\underline{\underline{A}})) = n$ no longer makes sense...(existence and uniqueness).
- Linear bijections may not have continuous inverse... (well-posedness).
- Different notions of convergence (norms) make a world of difference.

and of the basic tools of differential calculus:

- Function spaces.
- Derivatives, integrals.
- Boundary values.

Overview

- **Galerkin approximations:** Differential, variational and extremal formulations of a simple 1D boundary value problem. Well-posedness of variational formulations. Functional setting. Strong and weak coercivity. Lax-Milgram lemma. Banach's open mapping theorem. Céa's best-approximation property. Convergence under weak coercivity. (2 lectures)
- **The spaces of FEM:** (3 lectures)
- **Interpolation error and convergence:** (1 lecture)
- **Application to convection-diffusion-reaction problems:** (2 lectures)
- **Application to linear elasticity:** (1 lecture)
- **Mixed problems:** (2 lectures)
- **FEM for parabolic problems:** (2 lectures)

1 Galerkin approximations

1.1 Variational formulation of a simple 1D example

Let u be the solution of

$$\begin{cases} -u'' + u = f & \text{in } (0,1) \\ u(0) = u(1) = 0 \end{cases} \tag{1.1}$$

The **differential formulation** (DF) of the problem requires $-u'' + u$ to be exactly equal to f in **all** points $x \in (0,1)$.
Multiplying the equation by any function v and integrating by parts (recall that

$$\int_0^1 w'\, z\, dx = w(1)z(1) - w(0)z(0) - \int_0^1 w\, z'\, dx \tag{1.2}$$

holds for all w and z that are *regular enough*) one obtains that u satisfies

$$\int_0^1 (u'\, v' + u\, v)\, dx - u'(1)v(1) + u'(0)v(0) = \int_0^1 f\, v\, dx \qquad \forall\, v. \tag{1.3}$$

- The requirement "for all x" of the DF has become "for all functions v".
- Does equation (1.3) fully determine u?
- What happened with the boundary conditions?

Consider the following problem in **variational formulation** (VF): "Determine $u \in W$, such that $u(0) = u(1) = 0$ and that

$$\int_0^1 (u'\, v' + u\, v)\; dx = \int_0^1 f\, v\; dx \tag{1.4}$$

holds for all $v \in W$ satisfying $v(0) = v(1) = 0$."

Prop. 1.1 *The solution u of the DF (eq. 1.1) is also a solution of the VF if W consists of continuous functions of sufficient regularity. As a consequence, problem VF admits at least one solution whenever DF does.*

Proof. Following the steps that lead to the VF, it becomes clear that the only requirement for u to satisfy (1.4) is that the integration by parts formula (1.2) be valid. □

Exo. 1.1 *Show that the solution of*

$$\begin{cases} -u'' + u = f & in\ (0,1) \\ u(0) = 0, \qquad u'(1) = g \in \mathbb{R} \end{cases} \tag{1.5}$$

is a solution to: "Find $u \in W$ such that $u(0) = 0$ and that

$$\int_0^1 (u'\, v' + u\, v)\; dx = \int_0^1 f\, v\; dx \;+\; g\, v(1) \tag{1.6}$$

holds for all $v \in W$ satisfying $v(0) = 0$."

Consider the following problem in **extremal formulation** (EF): "Determine $u \in W$ such that it minimizes the function

$$J(w) = \int_0^1 \left(\frac{1}{2} w'(x)^2 + \frac{1}{2} w(x)^2 - f\, w \right) \, dx \tag{1.7}$$

over the functions $w \in W$ that satisfy $w(0) = w(1) = 0$."

Prop. 1.2 *The unique solution u of (1.1) is also a solution to EF. As a consequence, EF admits at least one solution.*

Proof. We need to show that $J(w) \geq J(u)$ for all $w \in W_0$, where

$$W_0 = \{ w \in W \, , \, w(0) = w(1) = 0 \}$$

Writing $w = u + \alpha v$ and replacing in (1.7) one obtains

$$J(u + \alpha\, v) = J(u) + \alpha \left[\int_0^1 (u'\, v' + u\, v - f\, v) \, dx \right] + \alpha^2 \int_0^1 \left(\frac{1}{2} v'(x)^2 + \frac{1}{2} v(x)^2 \right) \, dx$$

The last term is not negative and the second one is zero. □

Exo. 1.2 *Identify the EF of the previous exercise.*

Prop. 1.3 *Let u be the solution of*

$$\begin{cases} -u'' + u = f & \text{in } (0,1) \\ u(0) = 1, \qquad u'(1) = g \in \mathbb{R} \end{cases} \tag{1.8}$$

then u is also a solution of "Determine $u \in W$ such that $u(0) = 1$ and that

$$\int_0^1 (u'\, v' + u\, v)\; dx = \int_0^1 f\, v\; dx \;+\; g\, v(1) \tag{1.9}$$

holds for all $v \in W$ satisfying $v(0) = 0$."
Further, defining for any $a \in \mathbb{R}$

$$W_a = \{w \in W, w(0) = a\},$$

u minimizes over W_1 the function

$$J(w) = \int_0^1 \left(\frac{1}{2}w'(x)^2 + \frac{1}{2}w(x)^2 - f\, w\right)\; dx \;-\; g\, w(1). \tag{1.10}$$

Exo. 1.3 *Prove the last proposition.*

Let us define the bilinear and linear forms corresponding to problem (1.1):

$$a(v,w) = \int_0^1 (v'w' + vw)\, dx \qquad\qquad \ell(v) = \int_0^1 f\, v\, dx \tag{1.11}$$

and the function $J(v) = \frac{1}{2}a(v,v) - \ell(v)$. Remember that W is a space of functions with some (yet unspecified) regularity and let $W_0 = \{w \in W,\, w(0) = w(1) = 0\}$.

The three formulations that we have presented up to now are, thus:

DF: Find a function u such that

$$-u''(x) + u(x) = f(x) \qquad \forall\, x \in (0,1), \qquad\qquad u(0) = u(1) = 0$$

VF: Find a function $u \in W_0$ such that

$$a(u,v) = \ell(v) \quad \forall\, v \in W_0$$

EF: Find a function $u \in W_0$ such that

$$J(u) \leq J(w) \qquad \forall\, w \in W_0$$

and **we know that the exact solution of DF is also a solution of VF and of EF**.

The logic of the construction is justified by the following

Theorem 1.4 *If W is taken as*

$$W = \{w : (0,1) \to \mathbb{R}, \int_0^1 w(x)^2 \, dx < +\infty, \int_0^1 w'(x)^2 \, dx < +\infty\} \stackrel{\text{def}}{=} H^1(0,1)$$

and if f is such that there exists $C \in \mathbb{R}$ for which

$$\int_0^1 f(x)\, w(x) \, dx \leq C \sqrt{\int_0^1 w'(x)^2 \, dx} \qquad \forall\, w \in W_0 \tag{1.12}$$

then problems (VF) and (EF) have one and only one solution, and their solutions coincide.

The proof will be given later, now let us consider its consequences:

- The differential equation has <u>at most one solution</u> in W.
- If the solution u to (VF)-(EF) is regular enough to be considered a solution to (DF), then u is <u>the solution</u> to (DF).
- If the solution u to (VF)-(EF) is <u>not</u> regular enough to be considered a solution to (DF), then (DF) <u>has no solution</u>.

$\Rightarrow$ (VF) is a <u>generalization</u> of (DF).

Exo. 1.4 *Show that* $W_0 \subset C^0(0,1)$. *Further, compute* $C \in \mathbb{R}$ *such that*

$$\max_{x\in[0,1]} |w(x)| \leq C\sqrt{\int_0^1 w'(x)^2 \, dx} \qquad \forall\, w \,\in\, W_0$$

Hint: You may assume that $\int_0^1 f(x)\, g(x) \, dx \,\leq\, \sqrt{\int_0^1 f(x)^2 \, dx}\,\sqrt{\int_0^1 g(x)^2 \, dx}$ *for any* f *and* g *(Cauchy-Schwarz).*

Exo. 1.5 *Consider* $f(x) = |x - 1/2|^\gamma$. *For which exponents* γ *is* $\int_0^1 f(x)\, w(x) \, dx < +\infty$ *for all* $w \in W_0$*?*

Exo. 1.6 *Consider as* f *the "Dirac delta function" at* $x = 1/2$, *that we will denote by* $\delta_{1/2}$. *It can be considered as a "generalized" function defined by*

$$\int_0^1 \delta_{1/2}(x)\, w(x) \, dx = w(1/2) \qquad \forall\, w \,\in\, C^0(0,1)$$

Prove that $\delta_{1/2}$ *satisfies (1.12) and determine the analytical solution to (VF).*

Exo. 1.7 *Determine the DF and the EF corresponding to the following VF: "Find* $u \in W = H^1(0,1)$, $u(0) = 1$, *such that*

$$\int_0^1 (u'w' + uw) \, dx = w(1/2) \qquad \forall\, w \,\in\, W_0 \tag{1.13}$$

where $W_0 = \{w \,\in\, W, w(0) = 0\}$*."*

1.2 Variational formulations in general

Let V be a Hilbert space with norm $\|\cdot\|_V$. Let $a(\cdot,\cdot)$ and $\ell(\cdot)$ be bilinear and linear forms on V satisfying (continuity), for all $v,\, w \in V$,

$$a(v,w) \le N_a \, \|v\|_V \, \|w\|_V, \qquad\qquad \ell(v) \le N_\ell \, \|v\|_V \tag{1.14}$$

This last inequality means that $\ell \in V'$, the (topological) dual of V. The minimum N_ℓ that satisfies this inequality is called the norm of ℓ in V', i.e.

$$\|\ell\|_{V'} \stackrel{\text{def}}{=} \sup_{0 \neq v \in V} \frac{\ell(v)}{\|v\|_V} \tag{1.15}$$

The abstract VF we consider here is:

$$\text{“Find } u \in V \text{ such that} \qquad a(u,v) = \ell(v) \qquad\qquad \forall\, v \in V\text{”} \tag{1.16}$$

Exo. 1.8 *Assume that V is finite dimensional, of dimension n, and let $\{\phi^1, \phi^2, \ldots, \phi^n\}$ be a basis. Show that (1.16) is then equivalent to the linear system*

$$\underline{\underline{A}}\;\underline{U} = \underline{L} \tag{1.17}$$

where

$$A_{ij} \stackrel{\text{def}}{=} a(\phi^j, \phi^i), \qquad L_i \stackrel{\text{def}}{=} \ell(\phi^i) \tag{1.18}$$

and $\underline{U}$ is the coefficient column vector of the expansion of u, i.e.,

$$u = \sum_{i=1}^{n} U_i \, \phi^i \tag{1.19}$$

Def. 1.5 *The bilinear form $a(\cdot,\cdot)$ is said to be* **strongly coercive** *if there exists $\alpha > 0$ such that*

$$a(v,v) \geq \alpha \|v\|_V^2 \qquad \forall\, v \in V \tag{1.20}$$

Def. 1.6 *The bilinear form $a(\cdot,\cdot)$ is said to be* **weakly coercive** *(or to satisfy an* **inf-sup** *condition) if there exists $\beta > 0$ such that*

$$\sup_{0 \neq w \in V} \frac{a(v,w)}{\|w\|_V} \geq \beta \|v\|_V \qquad \forall\, v \in V \tag{1.21}$$

and

$$\sup_{0 \neq v \in V} \frac{a(v,w)}{\|v\|_V} \geq \beta \|w\|_V \qquad \forall\, w \in V \tag{1.22}$$

Exo. 1.9 *Prove that strong coercivity implies weak coercivity.*

Exo. 1.10 *Prove that, if V is finite dimensional, then* **(i)** *$a(\cdot,\cdot)$ is strongly coercive iff $\underline{\underline{A}}$ is positive definite ($\underline{X}^T\ \underline{\underline{A}}\ \underline{X} > 0\ \forall\, \underline{X} \in \mathbb{R}^n$), and* **(ii)** *$a(\cdot,\cdot)$ is weakly coercive iff $\underline{\underline{A}}$ is invertible.*

Exo. 1.11 *Prove that, if $a(\cdot,\cdot)$ is weakly coercive, then the solution u of (1.16) depends continuously on the forcing $\ell(\cdot)$. Specifically, prove that*

$$\|u\|_V \leq \frac{1}{\beta}\ \|\ell\|_{V'} \tag{1.23}$$

Theorem 1.7 *Assuming V to be a Hilbert space, problem (1.16) is well posed for any $\ell \in V'$ if and only if* **(i)** *$a(\cdot,\cdot)$ is continuous, and* **(ii)** *$a(\cdot,\cdot)$ is weakly coercive.*

A simpler version of this result is known as **Lax-Milgram lemma**:

Theorem 1.8 *Assuming V to be a Hilbert space, if $a(\cdot,\cdot)$ is continuous and strongly coercive then problem (1.16) is well posed for any $\ell \in V'$.*

Proof. This proof uses the so-called "Galerkin method", which will be useful to introduce... the Galerkin method!
Let $\{\phi^i\}$ be a basis of V. Denoting $V_N = \text{span}(\phi^1, \dots, \phi^N)$ we can define $u_N \in V_N$ as the unique solution of $a(u_N, v) = \ell(v)$ for all $v \in V_N$. This generates a sequence $\{u_N\}_{N=1,2,\dots}$ in V. Further, this sequence is bounded, because

$$\|u_N\|_V^2 \leq \frac{1}{\alpha}\, a(u_N, u_N) = \frac{1}{\alpha}\, \ell(u_N) \leq \frac{\|\ell\|_{V'}}{\alpha}\, \|u_N\|_V \quad \Rightarrow \quad \|u_N\|_V \leq \frac{\|\ell\|_{V'}}{\alpha}, \ \ \forall\, N$$

Recalling the weak compactness of bounded sets in Hilbert spaces, there exists $u \in V$ such that a subsequence of $\{u_N\}$ (still denoted by $\{u_N\}$ for simplicity) converges to u weakly. It remains to prove that $a(u, v) = \ell(v)$ for all $v \in V$. To see this, notice that

$$a(u, \phi^i) = a(\lim_N u_N, \phi^i) = \lim_N a(u_N, \phi^i) = \ell(\phi^i)$$

where the last equality holds because $a(u_N, \phi^i) = \ell(\phi^i)$ whenever $N \geq i$. Uniqueness is left as an exercise. □

Exo. 1.12 *Prove uniqueness in the previous theorem (bounded sequences may have several accumulation points).*

1.3 Galerkin approximations

The previous proof suggests a numerical method, the Galerkin method, to approximate the solution of a variational problem and thus of an elliptic PDE. The idea is simply to restrict the variational problem to a subspace of V that we will denote by V_h.

Discrete variational problem (Galerkin): Find $u_h \in V_h$ such that

$$a(u_h, v_h) = \ell(v_h) \qquad \forall\, v_h \in V_h \tag{1.24}$$

When the bilinear form $a(\cdot,\cdot)$ is symmetric and strongly coercive, this discrete probleme is equivalent to

Discrete extremal problem (Galerkin): Find $u_h \in V_h$ which minimizes over V_h the function

$$J(w) = \frac{1}{2}\, a(w, w) \,-\, \ell(w) \tag{1.25}$$

Exo. 1.13 *Prove this last assertion.*

The natural questions that arise are:

- Does u_h exist? Is it unique?
- Does u_h approximate u (the exact solution)?
- How difficult is it to compute u_h?

Does u_h exist? Is it unique?

Case 1) Strong coercivity of the form $a(\cdot,\cdot)$ over V

If $a(\cdot,\cdot)$ is strongly coercive over V, then

$$\inf_{0\neq w\in V} \frac{a(w,w)}{\|w\|_V^2} \;=\; \alpha > 0.$$

If $V_h \subset V$, then $a(\cdot,\cdot)$ is strongly coercive over V_h (because the infimum is taken over a smaller set). Then u_h exists and is unique as a consequence of Exo. 1.10.

Case 2) Weak coercivity of the form $a(\cdot,\cdot)$ over V

If $a(\cdot,\cdot)$ is just weakly coercive over V, then it may or may not be weakly coercive over V_h. Compare the two following conditions

$$(A)\ \inf_{w\in V}\sup_{v\in V} \frac{a(w,v)}{\|w\|_V\,\|v\|_V} = \beta > 0, \qquad\qquad (B)\ \inf_{w\in V_h}\sup_{v\in V_h} \frac{a(w,v)}{\|w\|_V\,\|v\|_V} = \beta_h > 0.$$

It is not true that (A)⇒(B) because the sup in (B) is taken over a smaller set. In this case the weak coercivity of the discrete problem must be proven independently, it is not inherited from the weak coercivity over the whole space V.

Does u_h approximate u?

Case 1) Strong coercivity of the form $a(\cdot,\cdot)$ over V

Lemma 1.9 (J. Céa) *If $a(\cdot,\cdot)$ and $\ell(\cdot)$ are continuous in V and $a(\cdot,\cdot)$ is strongly coercive, then*

$$\|u-u_h\|_V \leq \frac{N_a}{\alpha}\,\|u-v_h\|_V \qquad \forall\, v_h \,\in\, V_h \tag{1.26}$$

Proof. Notice the so-called **Galerkin orthogonality**:

$$a(u-u_h,v_h)=0 \qquad \forall\, v_h \,\in\, V_h \tag{1.27}$$

which implies that $a(u-u_h,u-u_h)=a(u-u_h,u-v_h)$ for all $v_h\in V_h$. Using this,

$$\|u-u_h\|_V^2 \leq \frac{1}{\alpha}a(u-u_h,u-u_h)=\frac{1}{\alpha}\,a(u-u_h,u-v_h)\leq \frac{N_a}{\alpha}\,\|u-u_h\|_V\,\|u-v_h\|_V \qquad \forall v_h\in V_h$$

In other words, $\|u-u_h\|_V \leq C\,\inf_{v_h\in V_h}\|u-v_h\|_V$. $\square$

Let h be a real parameter, typically a "mesh size". We say that a family $\{V_h\}_{h>0}\subset V$ satisfies the **approximability property** if:

$$\lim_{h\to 0}\operatorname{dist}(u,V_h)=\lim_{h\to 0}\;\inf_{v\,\in\,V_h}\;\|u-v\|_V \;=\; 0 \tag{1.28}$$

Corollary 1.10 *If $a(\cdot,\cdot)$ and $\ell(\cdot)$ are continuous in V, $a(\cdot,\cdot)$ is strongly coercive, and the family $\{V_h\}_{h>0}\subset V$ satisfies (1.28), then*

$$\lim_{h\to 0}u_h=u$$

in the sense of the norm $\|\cdot\|_V$.

Case 2) Weak coercivity of the form $a(\cdot,\cdot)$ over V_h
Assume now that the weak coercivity constant β_h is positive for all $h > 0$, so that u_h exists and is unique. Notice that Galerkin orthogonality still holds.

Lemma 1.11 *If $a(\cdot,\cdot)$ and $\ell(\cdot)$ are continuous in V, and $a(\cdot,\cdot)$ is weakly coercive in V_h with constant $\beta_h > 0$, then*

$$\|u - u_h\|_V \leq \left(1 + \frac{N_a}{\beta_h}\right) \|u - v_h\|_V \qquad \forall\, v_h \in V_h \tag{1.29}$$

Proof. One begins by decomposing the error as follows (we omit the subindex V in the norm)

$$\|u - u_h\| \leq \|u - v_h\| + \|u_h - v_h\| \qquad \forall\, v_h \in V_h \tag{1.30}$$

and then using the weak coercivity

$$\|u_h - v_h\| \leq \frac{1}{\beta_h} \sup_{w_h \in V_h} \frac{a(u_h - v_h, w_h)}{\|w_h\|} = \frac{1}{\beta_h} \sup_{w_h \in V_h} \frac{a(u - v_h, w_h)}{\|w_h\|} \leq \frac{N_a}{\beta_h} \|u - v_h\|$$

Substituting this into (1.30) one proves the claim. □

Corollary 1.12 *Under the hypotheses of Lemma 1.11, if there exists $\beta_0 > 0$ such that $\beta_h > \beta_0$ for all h and the family $\{V_h\}_{h>0} \subset V$ satisfies (1.28), then*

$$\lim_{h \to 0} u_h = u$$

in the sense of the norm $\|\cdot\|_V$.

How difficult is it to compute u_h?

Let us go back to our problem $-u''+u = f$ in $(0,1)$ with $u(0) = u(1) = 0$, which in VF requires to compute $u \in H^1(0,1)$ satisfying the boundary conditions and such that

$$\int_0^1 \left[u'(x)\, v'(x) + u(x)\, v(x)\right]\, dx = \int_0^1 f(x)\, v(x)\, dx \tag{1.31}$$

Suitable spaces for the Galerkin approximation are, for example,

- $\mathcal{P}_k$: The polynomials of degree up to k.
- $\mathcal{F}_k$: The space generated by the functions $\phi^m(x) = \sin(m\,\pi\,x)$, $m = 1, 2, \ldots, k$.

Exo. 1.14 *Show that $a(\cdot,\cdot)$ is continuous and strongly coercive over $V = H^1(0,1)$ with the norm*

$$\|w\|_V \stackrel{def}{=} \left[\int_0^1 \left[w'(x)^2 + w(x)^2\right]\, dx\right]^{\frac{1}{2}}$$

Exo. 1.15 *Build a small program in Matlab or Octave (or something else) that solves the Galerkin approximation of problem (1.31) considering $f = \delta_{1/4}$ and the spaces $\mathcal{P}_k$ and/or $\mathcal{F}_k$, for some values of k. Compare the results to the analytical solution building plots of u and u_h. Also, build graphs of $\|u - u_h\|$ vs k.*

In general, however, the construction of spaces of global basis functions, as the ones above, is not practical because it leads to dense matrices. In the next chapter we will introduce the spaces of the FEM, which are characterized by having bases with small support and thus lead to sparse matrices.

Exercises

Reading assignment: Read Chapter 1 of Duran's notes (all of it).

Exo. 1.16 *Carry out the "easy computation" that shows that $\underline{\underline{A}}$ is the tridiagonal matrix such that the diagonal elements are $2/h + 2h/3$ and the extra-diagonal elements are $-1/h + h/6$ (Durán, page 3).*

Exo. 1.17 *Can a symmetric bilinear form be weakly coercive but not strongly coercive?*

Exo. 1.18 *To what variational formulation and what differential formulation corresponds the following extremal formulation?*
Find $u \in V$, V consisting of functions that are smooth in $(0, 1/2)$ and $(1/2, 1)$ but can exhibit a (bounded) discontinuity at $x = 1/2$, that minimizes the function

$$J(w) = \int_0^1 [w'(x)^2 + 2w(x)^2]\, dx + 4\, [w(1/2+) - w(1/2-)]^2 - \int_0^{1/2} 7\, w(x)\, dx - 9w(0) \tag{1.32}$$

where $w(1/2\pm)$ represent the values on each side of the discontinuity. Notice that the space V (is it a vector space really?) has no boundary condition imposed. What are the boundary conditions of the DF at $x = 0$ and $x = 1$?

Exo. 1.19 *Consider the bilinear form*

$$a(u, v) = \int_0^1 u'(x)\, v'(x)\, dx.$$

Prove that this form is not strongly coercive in $H^1(0, 1)$ considering the norm

$$\|w\|_{H^1} \stackrel{def}{=} \left\{ \int_0^1 \left[u'(x)^2 + u(x)^2\right]\, dx \right\}^{\frac{1}{2}}$$

and that it is, with the same norm, in

$$H_0^1(0, 1) \stackrel{def}{=} \{w \in H^1(0, 1)\,,\, w(0) = w(1) = 0\}.$$

1.4 Variational formulations in 2D and 3D

The ideas are similar, but we need another integration by parts formula:

Lemma 1.13 *Let $f : \Omega \to \mathbb{R}$ be an integrable function, with Ω a Lipschitz bounded open set in $\mathbb{R}^d$ and $\partial_i f$ integrable over Ω, then*

$$\int_\Omega \partial_i f \; d\Omega = \int_{\partial\Omega} f \, n_i \; d\Gamma \tag{1.33}$$

Notice that this implies that

$$\int_\Omega \nabla \cdot \mathbf{v} \; d\Omega = \int_{\partial\Omega} \mathbf{v} \cdot \check{\mathbf{n}} \; d\Gamma \tag{1.34}$$

and that

$$\int_\Omega v \, \nabla^2 u \; d\Omega = \int_{\partial\Omega} v \, \nabla u \cdot \check{\mathbf{n}} \; d\Gamma - \int_\Omega \nabla v \cdot \nabla u \; d\Omega \tag{1.35}$$

We will also introduce the notation

Def. 1.14 *The* **Lebesgue space** *$L^p(\Omega)$, where $p \geq 1$, is the set of all functions such that their $L^p(\Omega)$-norm is finite,*

$$\|w\|_{L^p(\Omega)} \stackrel{def}{=} \left[\int_\Omega |w(x)|^p \; dx \right]^{\frac{1}{p}} \tag{1.36}$$

Exa. 1.15 (Poisson equation) *Consider the DF*

$$-\nabla^2 u = f \qquad \text{in } \Omega, \qquad u = 0 \qquad \text{on } \partial\Omega \tag{1.37}$$

where ∇ *is the gradient operator and* $\nabla^2 u = \sum_{i=1}^{d} \partial_{ii}^2 u$.

A suitable variational formulation is: Find $u \in V$ such that

$$a(u, v) = \ell(v) \qquad \forall\, v \in V$$

where

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \; d\Omega, \qquad \ell(v) = \int_\Omega f\, v \, d\Omega \qquad \text{and} \tag{1.38}$$

$$V = H_0^1(\Omega) = \{w \in L^2(\Omega),\ \partial_i w \in L^2(\Omega)\, \forall i = 1, \ldots, d\,,\ w = 0 \text{ on } \partial\Omega$$

which is a Hilbert space with the norm

$$\|w\|_{H^1} = \left(\|w\|_{L^2}^2 + \|\nabla w\|_{L^2}^2\right)^{\frac{1}{2}} \tag{1.39}$$

Exo. 1.20 *Prove that if u is a solution of the DF, then it solves the VF.*

Exo. 1.21 *Prove that $a(\cdot,\cdot)$ is continuous in V. Prove that $\ell(\cdot)$ is continuous in V if $f \in L^2(\Omega)$. Is this last condition necessary?*

Exo. 1.22 *Determine the EF of the Poisson problem.*

Exo. 1.23 *Is $a(\cdot,\cdot)$ strongly coercive?*

Exo. 1.24 *Let Ω be the unit circle. Determine for which exponents γ is the function r^γ in $H^1(\Omega)$.*

Exo. 1.25 *Assume that the domain Ω is divided into subdomains Ω_1 and Ω_2 by a smooth internal boundary Γ. Let V consist of functions such that their restrictions to Ω_i belong to $H^1(\Omega_i)$ and that are continuous across Γ. Determine the VF corresponding to the following EF:*Find $u \in V$ that minimizes

$$J(w) = \int_{\Omega_1} \frac{w^2 + \|\nabla w\|^2}{2}\, d\Omega + \int_{\Omega_2} \frac{3\|\nabla w\|^2}{2}\, d\Omega + \int_\Gamma (5\, w^2 - w)\, d\Gamma$$

over V.

Exo. 1.26 *Determine the DF that corresponds to the previous exercise.*

2 Finite element spaces and interpolation

The basic reference for what follows is Ciarlet [5]. Basically, the idea is to define finite element spaces that are locally polynomial and that contain complete polynomials of degree k in the space variables. With a judicious choice of the nodes (degrees of freedom), these piecewise polynomial functions can be made continuous by construction (if needed).
In the previous chapter it was shown that if there exists $\beta > 0$ such that, for all $w_h \in V_h$ and all $h > 0$,

$$\sup_{v_h \in V_h} \frac{a(w_h, v_h)}{\|v_h\|_V} \geq \beta \|w_h\|_V \tag{2.1}$$

then there exists $C > 0$ such that

$$\|u - u_h\|_V \leq C \inf_{v_h \in V_h} \|u - v_h\|_V \tag{2.2}$$

Notice that (2.1) is automatically satisfied if the bilinear form $a(\cdot, \cdot)$ is strongly coercive.
Denoting by $\mathcal{I}_h u$ the element-wise Lagrange interpolant of $u \in V \cap C^0(\overline{\Omega})$, it is obvious from (2.2) that

$$\|u - u_h\|_V \leq C \|u - \mathcal{I}_h u\|_V \tag{2.3}$$

The goal of this section is to introduce estimates of the interpolation error $\|u - \mathcal{I}_h u\|_V$ for some spaces V that appear in the applications.

2.1 Basic definitions

Def. 2.1 *A finite element in $\mathbb{R}^n$ is a triplet (K, P_K, Σ_K) where*

(i) *K is a closed (bounded) subset of $\mathbb{R}^n$ with a nonempty interior and Lipschitz boundary;*

(ii) *P_K is a finite-dimensional space of functions defined in K, of dimension m;*

(iii) Σ_K *is a set of* m *linear forms* $\{\sigma_i\}_{i=1,\dots,m}$ *which is* P_K*-unisolvent; i.e., if* $p \in P_K$ *then*

$$\sigma(p) = 0 \qquad \forall \sigma \in \Sigma_K \qquad \Rightarrow \qquad p = 0$$

It is implicitly assumed that the finite element is viewed with a larger function space $V(K)$ associated to it, in general a Sobolev space. Each $\sigma_i \in \Sigma_K$ is then assumed to be extended as an element of $V(K)'$.

Exa. 2.2 P_1.

Prop. 2.3 *There exists a basis* $\{\mathcal{N}_i\}$ *such that* $\sigma_i(\mathcal{N}_j) = \delta_{ij}$.

Finite elements are usually built by mapping a unique master element $\widehat{K}$, the following proposition states that if the master element is in itself a finite element, all the others will also be so. We restrict to affine mappings, since isoparametric finite elements fall slightly outside the classical theory, in that the corresponding spaces do not consist of piecewise polynomial functions.

Prop. 2.4 *If* $K, \widehat{K}$ *are affine equivalent,* $K = \phi(\widehat{K})$*, then if* $(\widehat{K}, \widehat{P}, \widehat{\Sigma})$ *is a finite element then we can define* (K, P_K, Σ_K) *and it is a finite element.*

Proof. The suitable definition that works is the one used in the implementations. Let $F_K : \widehat{K} \to K$ be the (affine) mapping which is assumed to exist. Then we define, for v in $V(K)$, the function $\widehat{v} \in V(\widehat{K})$ by $\widehat{v}(x) = v(F_K(x))$. Further,

$$P_K = \{v : K \to \mathbb{R}, \widehat{v} \in \widehat{P}\}$$

and

$$\Sigma_K = \{\sigma : V(K) \to \mathbb{R}, \sigma(v) = \widehat{\sigma}(\widehat{v}),\ \forall \widehat{v} \in \widehat{P},\ \text{with } \widehat{\sigma} \in \widehat{\Sigma}\}$$

□

The popular "master element" is thus a specific triplet $(\widehat{K}, \widehat{P}, \widehat{\Sigma})$ from which all the other finite elements are obtained by suitably composing with the affine mapping F_K.

Def. 2.5 *The* **local interpolation operator** $\mathcal{I}_K : V(K) \to P_K$ *is defined as*

$$\mathcal{I}_K v = \sum_{i=1}^{m} \sigma_i(v)\mathcal{N}_i \qquad \forall\, v \in V(K)$$

This interpolation is indeed a projection:

Prop. 2.6 $\mathcal{I}_K p = p$ *for all* $p \in P_K$.

and is preserved by composition with the affine mapping:

Prop. 2.7 $\widehat{\mathcal{I}_K v} = \mathcal{I}_{\widehat{K}} \widehat{v}$ *for all* $v \in V(K)$.

Notice also that, if $\widehat{P}$ contains all polynomials up to some degree k, then P_K will also contain all polynomials up to degree k whenever K is affine-equivalent to $\widehat{K}$. The local problem of approximating a function in K with functions in P_K is thus in order, and the subject of the next paragraph.

2.2 Local $L^\infty(K)$ estimates for P_1-triangles

We begin by considering the case of P_1-simplices (triangles in 2D, tetrahedra in 3D). It is a good exercise in which the estimates can be derived explicitly. It is also a good excuse to introduce the multi-point Taylor formula.

Theorem 2.8 *Let* K *be a* P_1*-element,* h_K *its diameter and* ρ_K *the radius of the largest ball contained in* K*. Then, for all* $v \in C^\infty(K)$*,*

(a) $\|v - \mathcal{I}_K v\|_{L^\infty(K)} \leq \frac{d^2\, h_K^2}{2} \max_{|\alpha|=2} \|D^\alpha v\|_{L^\infty(K)}$

(b) $\max_{|\alpha|=1} \|D^\alpha (v - \mathcal{I}_K v)\|_{L^\infty(K)} \leq \frac{(d+1)\, d^2\, h_K^2}{2\, \rho_K} \max_{|\alpha|=2} \|D^\alpha v\|_{L^\infty(K)}$

Proof. Let X^j be the position of the j-th node of the element, then

$$\mathcal{I}_K v(x) = \sum_{j=1}^{d+1} v(X^j)\mathcal{N}^j(x) \tag{2.4}$$

We now perform a Taylor expansion *around* x, and evaluate it at X^j, obtaining

$$v(X^j) = v(x) + \sum_{k=1}^{d} \frac{\partial v}{\partial x_k}(x)\left(X_k^j - x_k\right) + \frac{1}{2}\sum_{k,\ell=1}^{d} \frac{\partial^2 v}{\partial x_k \partial x_\ell}(\xi)\left(X_k^j - x_k\right)\ \left(X_\ell^j - x_\ell\right) \tag{2.5}$$

where $\xi = \eta X^j + (1-\eta)x$ for some $\eta \in [0,1]$. Let us denote by $p^j(x)$ the second term in the right-hand side of (2.5), and by $r^j(x)$ the third term. By direct inspection we notice that

$$|r^j(x)| \leq \frac{d^2\, h_K^2}{2} \max_{|\alpha|=2} \|D^\alpha v\|_{L^\infty(K)}$$

Let us now insert $v(X^j)$ from (2.5) into (2.4) to get

$$\mathcal{I}_K v(x) = \sum_{j=1}^{d+1} v(x)\mathcal{N}^j(x) + \sum_{j=1}^{d+1} p^j(x)\mathcal{N}^j(x) + \sum_{j=1}^{d+1} r^j(x)\mathcal{N}^j(x)$$

The first term on the right is equal to $v(x)$ because $\sum_j \mathcal{N}^j = 1$. The second term vanishes, since

$$\sum_{j=1}^{d+1}\sum_{k=1}^{d} \frac{\partial v}{\partial x_k}(x)\left(X_k^j - x_k\right)\mathcal{N}^j(x) = \sum_{k=1}^{d} \frac{\partial v}{\partial x_k}(x)\left\{\sum_{j=1}^{d+1} X_k^j \mathcal{N}^j(x) - x_k \sum_{j=1}^{d+1}\mathcal{N}^j(x)\right\} =$$

$$= \sum_{k=1}^{d} \frac{\partial v}{\partial x_k}(x)\left\{x_k - x_k\right\} = 0$$

As a consequence, $v(x) - \mathcal{I}_K v(x) = \sum_{j=1}^{d+1} r^j(x)\mathcal{N}^j(x)$ and thus

$$|v(x) - \mathcal{I}_K v(x)| \leq \max_j |r^j(x)| \sum_j \mathcal{N}^j(x) = \max_j |r^j(x)| \leq \frac{d^2\, h_K^2}{2} \max_{|\alpha|=2} \|D^\alpha v\|_{L^\infty(K)}$$

implying assertion (a). Now, by differentiating (2.4) and using (2.5) as before, one obtains

$$\frac{\partial \mathcal{I}_K v}{\partial x_m}(x) = \sum_j v(x) \frac{\partial \mathcal{N}^j}{\partial x_m}(x) + \sum_{j,k} \frac{\partial v}{\partial x_k}(x) \left(X_k^j - x_k\right) \frac{\partial \mathcal{N}^j}{\partial x_m}(x) + \sum_{j,k} r^j(x) \frac{\partial \mathcal{N}^j}{\partial x_m}(x)$$

On the right-hand side above, the first term vanishes and the second term happens to be equal to $\frac{\partial v}{\partial x_m}(x)$, since

$$\sum_{j,k} \frac{\partial v}{\partial x_k}(x) \left(X_k^j - x_k\right) \frac{\partial \mathcal{N}^j}{\partial x_m}(x) = \sum_k \frac{\partial v}{\partial x_k}(x) \left[\sum_j X_k^j \frac{\partial \mathcal{N}^j}{\partial x_m}(x) - x_m \sum_j \frac{\partial \mathcal{N}^j}{\partial x_m}(x)\right] =$$

$$= \sum_k \frac{\partial v}{\partial x_k}(x) \frac{\partial}{\partial x_m} \sum_j X_k^j \mathcal{N}^j(x) = \sum_k \frac{\partial v}{\partial x_k}(x) \frac{\partial x_k}{\partial x_m} = \frac{\partial v}{\partial x_m}(x)$$

As a consequence

$$\left| \frac{\partial \mathcal{I}_K v}{\partial x_m}(x) - \frac{\partial v}{\partial x_m}(x) \right| = \left| \sum_{j=1}^{d+1} r^j(x) \frac{\partial \mathcal{N}^j}{\partial x_m}(x) \right| \leq \max_j |r^j(x)| \sum_{j=1}^{d+1} \left| \frac{\partial \mathcal{N}^j}{\partial x_m}(x) \right|$$

The reader can convince himself that the norm of the gradient of a P_1 basis function, which equals one at one node and zero on the opposite side/face, can never be greater than $\frac{1}{\rho_K}$, which immediately leads to assertion (b). $\square$

2.3 Local estimates in Sobolev norms

The previous paragraph provides us with an interpolation estimate in the norm $L^\infty(K)$ for the function and its first derivatives. Most formulations studied so far, however, have $V = H^1(\Omega)$ and we need thus estimates of $u - \mathcal{I}_K u$ in the $H^m(K)$-norm.

2.3.1 First estimates

A simplistic approach to estimate $\|u - \mathcal{I}_K u\|_{L^2(K)}$ for P_1 elements could be

$$\|u - \mathcal{I}_K u\|^2_{L^2(K)} = \int_K (u - \mathcal{I}_K u)^2 \leq |K| \, \|u - \mathcal{I}_K u\|^2_{L^\infty(K)} \leq 4|K| h_K^4 \max_{|\alpha|=2} \|D^\alpha u\|^2_{L^\infty(K)}$$

so that, with simplified notation,

$$\|u - \mathcal{I}_K u\|_{L^2(K)} \leq 2\sqrt{|K|}\, h_K^2 \, \|D^2 u\|_{L^\infty(K)} \tag{2.6}$$

Proceeding analogously, we obtain a first estimate for $\|\nabla u - \nabla(\mathcal{I}_K u)\|_{L^2(K)}$,

$$\|\nabla u - \nabla(\mathcal{I}_K u)\|^2_{L^2(K)} = \int_K \sum_{i=1}^d \left[\frac{\partial(u - \mathcal{I}_K u)}{\partial x_i}\right]^2 \leq |K| \sum_{i=1}^d \left\| \frac{\partial(u - \mathcal{I}_K u)}{\partial x_i} \right\|^2_{L^\infty(K)}$$

which from Th. 2.8 implies

$$\|\nabla u - \nabla(\mathcal{I}_K u)\|_{L^2(K)} \leq \sqrt{|K|}\, \frac{6\, d\, h_K^2}{\rho_K} \, \|D^2 u\|_{L^\infty(K)} \tag{2.7}$$

Notice that these estimates require $u \in W^{2,\infty}(K)$, which is "too much" regularity.

Exo. 2.1 *Consider the function $u(x) = |x|$ and its P_1 interpolant in the 1D simplex $K = (-h/2, h/2)$. Compute $\|u - \mathcal{I}_K u\|_{L^2(K)}$ and $\|u' - (\mathcal{I}_K u)'\|_{L^2(K)}$, compare to the previous estimates, and discuss briefly.*

2.3.2 An L^2-estimate without second derivatives

If the function to be interpolated does not have second derivatives in K, then $\|u - \mathcal{I}_K u\|_{L^2(K)}$ cannot be expected to be of order $\mathcal{O}(\sqrt{|K|}\, h_K^2)$. The following estimate, proved in *Buscaglia & Agouzal* (IMA J. Numer. Anal. 32, 672-686, 2012), has minimal requirements on both P_K and u. Notice in particular that P_K must contain the constants but not necessarily polynomials of degree 1.

Theorem 2.9 *Assume that the basis functions $\{\mathcal{N}^j\}$ ($j = 1, \ldots, d+1$) of an element K satisfy: (H1) $\mathcal{N}^j(X^k) = \delta_{jk}$, (H2) $\sum_j \mathcal{N}^j(x) = 1$, (H3) $0 \le \mathcal{N}^j(x) \le 1$ for all j and for all $x \in K$.*
Then, for all $u \in W^{1,p}(K)$ with $p > d \ge 2$,

$$\|u - \mathcal{I}_K u\|_{L^2(K)} \le \frac{p\,(d+1)}{p-d}\, |K|^{\frac{1}{2}-\frac{1}{p}}\, h_K\, \|\nabla u\|_{L^p(K)} \tag{2.8}$$

If ∇u is bounded we can take $p = +\infty$ to get

$$\|u - \mathcal{I}_K u\|_{L^2(K)} \le (d+1)\, \sqrt{|K|}\, h_K\, \|\nabla u\|_{L^\infty(K)} \tag{2.9}$$

which is of order $\mathcal{O}(\sqrt{|K|}\, h_K)$.

2.3.3 General local interpolation estimates

Theorem 2.10 *Let (K, P_K, Σ_K) be a Lagrange finite element such that (a) P_K contains all polynomials of degree $\le k$, and (b) it is affine-equivalent to the "master element" $(\widehat{K}, \widehat{P}, \widehat{\Sigma})$. Then, the Lagrange*

interpolant $\mathcal{I}_K u(x) = \sum_j u(X^j)\,\mathcal{N}^j(x)$ *satisfies*

$$\|u - \mathcal{I}_K u\|_{L^2(K)} \leq C\, h_K^{\ell+1}\, \|D^{\ell+1} u\|_{L^2(K)} \tag{2.10}$$

for all $\ell \leq k$, *with* C *depending on* ℓ *but not on* h_K *or* u.
Similarly,

$$|u - \mathcal{I}_K u|_{H^1(K)} = \|\nabla u - \nabla(\mathcal{I}_K u)\|_{L^2(K)} \leq C\, \frac{h_K^{\ell+1}}{\rho_K}\, \|D^{\ell+1} u\|_{L^2(K)} \tag{2.11}$$

The proof of this theorem is somewhat involved. The interested reader may refer to Ciarlet [5] or to Ern-Guermond [8].

2.4 Global interpolation error

The obtention of global interpolation estimates is quite straightforward, but needs a few definitions.

2.4.1 Considerations about meshes

A mesh $\mathcal{T}_h$ of a domain Ω in $\mathbb{R}^d$ is a collection of compacts (elements) K_i, $i = 1, \ldots, N_e$, such that

$$\overline{\Omega} = \bigcup_{i=1}^{N_e} K_i, \qquad \dot{K}_i \cap \dot{K}_j = \emptyset \text{ if } i \neq j, \qquad \partial\Omega \subset \bigcup_{i=1}^{N_e} \partial K_i \tag{2.12}$$

Def. 2.11 *The* **global interpolation operator** $\mathcal{I}_h : W \to W_h$, *where*

$$W = \{w \in L^1(\Omega), w|_K \in V(K),\ \forall\, K \in \mathcal{T}_h\}$$

$$W_h = \{w \in L^1(\Omega), w|K \in P_K,\ \forall\, K \in \mathcal{T}_h\}$$

by

$$\mathcal{I}_h v = \sum_{K \in \mathcal{T}_h} \sum_i \sigma_{K,i}(v|K) \mathcal{N}_{K,i} \tag{2.13}$$

The subscript h refers to the mesh size. In fact, in error estimates one has to consider not a single mesh but a family of meshes indexed by h, and study the error as $h \to 0$. The geometrical properties of the mesh refinement enter thus into consideration. Generally, the mesh-size parameter h is defined as

$$h = \max_{K \in \mathcal{T}_h} h_K \tag{2.14}$$

For global estimates in $H^m(\Omega)$ with $m \geq 1$ the ratio $s_K = \frac{h_K}{\rho_K}$ will appear. This motivates the definition of shape-regular (or, simply, regular) meshes:

Def. 2.12 *A family of meshes $\mathcal{T}_h$, parameterized by the parameter $h \in H$ (where H is some subset of $\mathbb{R}$), is said to be* **shape-regular** *if there exists $S \in \mathbb{R}$ such that*

$$s_K = \frac{h_K}{\rho_K} \leq S \qquad \forall\, K \in \mathcal{T}_h, \quad \forall\, h \in H \tag{2.15}$$

A shape-regular mesh (rigorously speaking, family of meshes) cannot contain needle-like elements. If the elements are triangles, no angle can tend to zero, the so-called "minimum angle condition". This condition is known not to be necessary for the convergence of the finite element interpolant in $H^1(\Omega)$, the necessary one being that no angle in the triangulation tend to π (the so-called "maximum angle condition").

2.4.2 From local to global

The local estimates already obtained can be turned global by simply collecting the contributions from all elements in the mesh.

Consider the estimate of Thm. 2.8(a), to begin with. One can build an $L^\infty(\Omega)$ as follows:

$$\|u - \mathcal{I}_h u\|_{L^\infty(\Omega)} = \max_K \|u - \mathcal{I}_K u\|_{L^\infty(K)} \leq \frac{d^2}{2} \max_K \left\{ h_K^2 \|D^2 u\|_{L^\infty(K)} \right\} \leq \frac{d^2}{2} h^2 \|D^2 u\|_{L^\infty(\Omega)}$$

which holds without any assumption on the mesh.
Similar estimates based on local to global reasonings are left as exercises.

Exo. 2.2 *Starting from Thm. 2.8(b), prove that*

$$\|\nabla u - \nabla(\mathcal{I}_h u)\|_{L^\infty(\Omega)} \leq \frac{(d+1)d^2 S}{2} h \|D^2 u\|_{L^\infty(\Omega)}$$

where S is the shape-regularity constant of the mesh.

Exo. 2.3 *Using (2.9) prove that*

$$\|u - \mathcal{I}_h u\|_{L^2(\Omega)} \leq (d+1) \sqrt{|\Omega|}\, h \|\nabla u\|_{L^\infty(\Omega)} \tag{2.16}$$

Exo. 2.4 *Starting from (2.11) prove that, if the family of meshes is shape-regular and the function u smooth, then*

$$|u - \mathcal{I}_h u|_{H^1(\Omega)} \leq C\, S\, h^k \|D^{k+1} u\|_{L^2(\Omega)} \tag{2.17}$$

where S is the shape-regularity constant of the mesh.

Exo. 2.5 *Assume that there exists a straight line Γ (or planar surface in 3D) in the domain Ω, at which there is a sudden change in material properties. As a consequence, $u \in H^2(\Omega \backslash \Gamma) \cap C^0(\Omega)$, but $u \notin H^2(\Omega)$. Discuss the interpolation estimate for such a function u, showing the advantages of using an "interface-fitting mesh"; i.e., a mesh such that Γ coincides with inter-element boundaries and thus does not cut any element.*

2.4.3 Global estimate

Let us state a global estimate more general than the one we have been building up to now.

Theorem 2.13 *Let $\mathcal{T}_h$, $h > 0$, be a family of shape-regular meshes of a domain $\Omega \subset \mathbb{R}^n$. Let $(\widehat{K}, \widehat{P}, \widehat{\Sigma})$ be the reference element of the mesh, all the mappings $F_K : \widehat{K} \to K$ being affine. Let $\mathcal{I}_h$ be the global interpolation operator corresponding to $\mathcal{T}_h$. Assume further that $P_k \subset \widehat{P}$ (i.e.; that the finite elements are "of degree k"). Then, for each $1 \leq p < +\infty$, and for each $0 \leq \ell \leq k$, there exists C such that for all h and all $v \in W^{\ell+1,p}(\Omega)$,*

$$\|v - \mathcal{I}_h v\|_{L^p(\Omega)} + \sum_{m=1}^{\ell+1} h^m \left(\sum_{K \in \mathcal{T}_h} |v - \mathcal{I}_h v|^p_{W^{m,p}(K)} \right)^{\frac{1}{p}} \leq C\, h^{\ell+1} |v|_{W^{\ell+1,p}(\Omega)} \tag{2.18}$$

If $p = +\infty$,

$$\|v - \mathcal{I}_h v\|_{L^\infty(\Omega)} + \sum_{m=1}^{\ell+1} h^m \left(\max_{K \in \mathcal{T}_h} |v - \mathcal{I}_h v|^p_{W^{m,\infty}(K)} \right)^{\frac{1}{p}} \leq C\, h^{\ell+1} |v|_{W^{\ell+1,\infty}(\Omega)} \tag{2.19}$$

Proof. See Ern-Guermond [8], p. 61. □

Notice that the previous theorem holds not just for simplicial elements but also for affine-equivalent quadrilaterals, hexahedra, etc.

Exo. 2.6 *Deduce from the theorem that, for P_1 and Q_1 elements,*

$$\|v - \mathcal{I}_h v\|_{H^1(\Omega)} \leq C\, h, \qquad \|v - \mathcal{I}_h v\|_{L^2(\Omega)} \leq C\, h^2$$

2.5 Inverse inequalities

Inverse inequalities are sometimes useful in the convergence analysis of finite element methods. They provide bounds on operators that are **unbounded** in $H^m(\Omega)$, with $m > 0$, but **bounded** in V_h due to its finite-dimensionality. Intuitively, in a shape-regular mesh for a derivative $\partial u_h/\partial x_i$ to be "very large" the nodal values of the u_h must also be "very large".
Let $(\hat{K}, \widehat{P}, \widehat{\Sigma})$ be the "reference" or "master" element. Let K be an element that is affine-equivalent to $\widehat{K}$, as defined before, with $F_K : \widehat{K} \to K$ the corresponding linear mapping:

$$F_K(x) = A_K\, x + b_K$$

In such a setting, we have

Lemma 2.14

(a)

$$|det\, A_K| = \frac{|K|}{|\widehat{K}|}, \qquad \|A_K\| \leq \frac{h_K}{\rho_{\widehat{K}}}, \qquad \|A_K^{-1}\| \leq \frac{h_{\widehat{K}}}{\rho_K}$$

(b) *There exists C, depending on s and p but independent of K, such that for all $v \in W^{s,p}(K)$,*

$$|\widehat{v}|_{W^{s,p}(\widehat{K})} \leq C\|A_K\|^s\, |det\, A_K|^{-\frac{1}{p}} |v|_{W^{s,p}(K)} \tag{2.20}$$

$$|v|_{W^{s,p}(K)} \leq C\|A_K^{-1}\|^s\, |det\, A_K|^{\frac{1}{p}} |\widehat{v}|_{W^{s,p}(\widehat{K})} \tag{2.21}$$

Proof. See, e.g., Ciarlet [5], p. 122. □

Let us show how to take advantage of this result to prove some simple estimates.

Prop. 2.15 *There exists $C > 0$, independent of K, such that*

$$\|\nabla v_h\|_{L^2(K)} \leq \frac{C}{\rho_K} \|v_h\|_{L^2(K)} \tag{2.22}$$

for any $v_h \in P_K$.

Proof. This proof uses the so-called *scaling* argument. From (2.21) we have, taking $s = 1$ and $p = 2$,

$$\|\nabla v_h\|_{L^2(K)} \leq C \, \|A_K^{-1}\| \, |\det A_K|^{\frac{1}{2}} \|\nabla \widehat{v_h}\|_{L^2(\widehat{K})} \tag{2.23}$$

Now let us show that there exists a constant $\widehat{C}$ such that

$$\|\nabla \widehat{v_h}\|_{L^2(\widehat{K})} \leq \widehat{C} \|\widehat{v_h}\|_{L^2(\widehat{K})} \tag{2.24}$$

For this, consider the set $\mathcal{S} = \{w \in P_K \mid \|\widehat{w}\|_{L^2(\widehat{K})} = 1\}$, which is bounded and closed in the finite-dimensional space P_K. Let $\widehat{C}$ be the **maximum** that the **continuous** function $\|\nabla \widehat{w}\|_{L^2(\widehat{K})}$ attains in $\mathcal{S}$.
Then, denoting by

$$\widehat{z_h} = \frac{1}{\|\widehat{v_h}\|_{L^2(\widehat{K})}} \widehat{v_h}$$

and noticing that $\widehat{z_h} \in \mathcal{S}$, we have that

$$\|\nabla \widehat{z_h}\|_{L^2(\widehat{K})} \leq \widehat{C}$$

and thus (2.24) is proved. Inserting it into (2.23) and using (2.20) one gets

$$\|\nabla v_h\|_{L^2(K)} \leq C \, \widehat{C} \, \|A_K^{-1}\| \, |\det A_K|^{\frac{1}{2}} \|\widehat{v_h}\|_{L^2(\widehat{K})} \leq C^2 \, \widehat{C} \, \|A_K^{-1}\| \, |\det A_K|^{\frac{1}{2}} \, |\det A_K|^{-\frac{1}{2}} \|v_h\|_{L^2(K)} \leq$$

$$\leq \frac{(C^2\,\widehat{C}\,h_{\widehat{K}})}{\rho_K}\|v_h\|_{L^2(K)}$$

and the proof ends noticing that the product inside the parentheses is a constant independent of K and v_h. $\square$

Notice that there does **not** exist a constant C that makes

$$\|\nabla v\|_{L^2(K)} \leq \frac{C}{\rho_K}\|v\|_{L^2(K)} \tag{2.25}$$

in the **infinite dimensional case**, i.e., for any v in $H^1(K)$.

Exo. 2.7 *Let K be the unit interval $(0,1)$ in 1D. Build a sequence $\{\varphi_n\}$ of functions such that $\|\varphi_n\|_{L^2(K)} = 1$ and $\|\nabla\varphi_n\|_{L^2(K)} = n$.*
Argue that the existence of such a sequence is a counterexample to (2.25).

With a scaling argument one can prove the following discrete trace estimate.

Prop. 2.16 *There exists $C > 0$, independent of K, such that*

$$\|v_h\|_{L^2(F)} \leq C\,h_K^{-\frac{1}{2}}\,\|v_h\|_{L^2(K)} \qquad \forall\, v_h \,\in\, P_K \tag{2.26}$$

where F is an edge (face in 3D) of K.

The proof is left as an optional exercise. Notice that, again, there is no chance of (2.26) holding for all v in an infinite-dimensional space, such as $C^\infty(K)$ for example (build a sequence that shows this!).
Several other inverse inequalities can be extracted as particular cases of the following theorem (see, e.g., [8] p. 75).

Theorem 2.17 *Let $\mathcal{T}_h$ be a shape-regular family of meshes in $\Omega \subset \mathbb{R}^d$. Then, for $0 \leq m \leq \ell$ and $1 \leq p, q \leq \infty$, there exists a constant C such that, for all $h > 0$ and all $K \in \mathcal{T}_h$,*

$$\|v\|_{W^{\ell,p}(K)} \leq C\, h_K^{m-\ell+d\left(\frac{1}{p}-\frac{1}{q}\right)} \|v\|_{W^{m,q}(K)} \tag{2.27}$$

for all $v \in P_K$.

This local estimate, to be made global, puts the restriction on the family of meshes that, as $h \to 0$ the diameter ratio between the largest and smaller h_K in $\mathcal{T}_h$ remain bounded.

Def. 2.18 *A family of meshes $\{\mathcal{T}_h\}_{h>0}$ is said to be* **quasi-uniform** *if it is shape-regular and there exists c such that*

$$\forall h, \quad \forall K \in \mathcal{T}_h, \qquad h_K \geq c\, h \tag{2.28}$$

Exo. 2.8 *Does the quasi-uniformity of the mesh imply the existence of $C > 0$ such that*

$$\|\nabla v_h\|_{L^2(\Omega)} \leq C\, h^{-1} \|v_h\|_{L^2(\Omega)} \qquad \forall\, v_h \in V_h \ ? \tag{2.29}$$

Exo. 2.9 *Does the quasi-uniformity of the mesh imply the existence of $C > 0$ such that*

$$\|v_h\|_{L^2(\partial\Omega)} \leq C\, h^{-\frac{1}{2}} \|v_h\|_{L^2(\Omega)} \qquad \forall\, v_h \in V_h \ ? \tag{2.30}$$

3 Galerkin treatment of elliptic second-order problems

3.1 The continuous problem

We consider the following problem:

$$
\begin{aligned}
-\mathrm{div}(K\nabla u) + \beta\cdot\nabla u + \sigma u &= f \qquad \text{in } \Omega && (3.1)\\
u &= g \qquad \text{on } \Gamma_D && (3.2)\\
(K\nabla u)\cdot\mathbf{n} &= H \qquad \text{on } \Gamma_N && (3.3)
\end{aligned}
$$

where Γ_D and Γ_N are disjoint parts of $\partial\Omega$, and $\overline{\Gamma_D \cup \Gamma_N} = \partial\Omega$.
Notice that, since $K(x)$ is a $n\times n$ symmetric matrix and $\beta(x)$ is an n-vector, the problem above is a general second-order partial differential equation.
Integrating formally by parts we get

$$\int_\Omega (\nabla v\cdot(K\nabla u) + v\,\beta\cdot\nabla u + \sigma uv)\ d\Omega = \int_\Omega fv\ d\Omega + \int_{\partial\Omega} v\mathbf{n}\cdot(K\nabla u)\ d\Gamma$$

We thus consider the bilinear form

$$a(u,v) = \int_\Omega (\nabla v\cdot(K\nabla u) + v\,\beta\cdot\nabla u + \sigma uv)\ d\Omega \tag{3.4}$$

Prop. 3.1 *If $K \in (L^\infty(\Omega))^{n\times n}$, $\beta \in (L^\infty(\Omega))^n$ and $\sigma \in L^\infty(\Omega)$, then $a(\cdot,\cdot)$ is continuous on $H^1(\Omega)$.*

Exo. 3.1 *Prove the proposition.*

It is clear that, for the problem to admit a solution, the data g and Γ_D must be regular enough for a function $u_g \in H^1(\Omega)$ to exist satisfying $u_g = g$ on Γ_D. Such a function is called a "lifting" function, and if it exists one says that g belongs to a "trace space".
We now change the unknown to $w = u - u_g$, so that

$$a(w,v) = \int_\Omega f v \; d\Omega + \int_{\partial\Omega} v\mathbf{n} \cdot (K\nabla u) \; d\Gamma - a(u_g, v)$$

and $w = 0$ on Γ_D. This leads us to consider the following problem: *Find* $w \in H^1_{D0}(\Omega)$ *such that*

$$a(w,v) = \int_\Omega f \, v \; d\Omega + \int_{\Gamma_N} H \, v \; d\Gamma - a(u_g, v) \stackrel{def}{=} \ell(v) \tag{3.5}$$

where $H^1_{D0} = \{v \in H^1(\Omega), v = 0 \text{ on } \Gamma_D\}$.

Prop. 3.2 *Assume the data f, g, H, Γ_N and Γ_D are regular enough for the right-hand side of (3.5) to be a continuous linear functional on $H^1_{D0}(\Omega)$. Assume further that the hypotheses of Prop. 3.1 hold, and that*

$$div\,\beta \in L^\infty(\Omega), \qquad \beta(x)\cdot n(x) > 0 \quad a.e. \text{ on } \Gamma_N \tag{3.6}$$

$$\xi \cdot (K(x)\xi) \geq K_0 \, |\xi|^2 \qquad \forall \xi \in \mathbb{R}^n; \; a.e. \text{ in } \Omega \tag{3.7}$$

$$\sigma(x) - \frac{1}{2} div\,\beta(x) \geq s_{\min} \qquad a.e. \text{ in } \Omega \tag{3.8}$$

where K_0 and $s_{\min}$ are strictly positive constants. Then (3.5) is well-posed.

Proof. Notice first that $H^1_{D0}(\Omega)$ is a closed subspace of $H^1(\Omega)$. To see this, consider the applications $\gamma_0 : H^1(\Omega) \to L^2(\partial\Omega)$ (the boundary trace operator, which is continuous as proved for example in Adams, Brenner-Scott, etc.) and $r_D : L^2(\partial\Omega) \to L^2(\Gamma_D)$, the restriction to Γ_D of a function in $L^2(\Omega)$, which is also continuous. The value of any function $f \in H^1(\Omega)$ on Γ_D is, then, $\gamma_{0D}(f) = r_D(\gamma_0(f))$. The subspace $H^1_{D0}(\Omega)$ is the pre-image of zero by γ_{0D}, and is thus closed.
To conclude the proof, it remains to show that $a(\cdot,\cdot)$ is weakly coercive. In fact, a direct calculation shows that $a(\cdot,\cdot)$ is strongly coercive and thus Lax-Milgram lemma guarantees well-posedness. □
The condition

$$\xi \cdot (K(x)\xi) \geq K_0 \, |\xi|^2 > 0, \qquad \forall \xi \in \mathbb{R}^n; \text{ a.e. in } \Omega$$

is essential to the previous well-posedness result, as it applies only for *elliptic* second-order PDEs (not hyperbolic, not parabolic). The condition $s_{\min} > 0$ is not essential, in the sense that if $s_{\min} \leq 0$ what may happen is that the *homogeneous* problem defined by $f = g = H = 0$ admits non-trivial solutions. It may also happen that for certain data the solution does not exist, in much the same way as a linear system

$$\underline{\underline{A}} \, \underline{x} \; = \; \underline{b}$$

with $\det(\underline{\underline{A}}) = 0$ either does not have a solution, or has infinitely many (the solution is determined only up to the addition of an arbitrary element of $\mathrm{Ker}(\underline{\underline{A}})$).

Exa. 3.3 *The simplest and very important case that is not covered by Prop. 3.2 is the* purely diffusive *problem with Neumann data, corresponding to*

$$\beta = 0 \text{ (no convection)}, \qquad \sigma = 0 \text{ (no reaction)}, \qquad \Gamma_N = \partial\Omega \text{ (no Dirichlet boundary)}. \tag{3.9}$$

The differential formulation is

$$-\, div(K\nabla u) \; = \; f \qquad \text{in } \Omega, \qquad (K\nabla u)\cdot \mathbf{n} \; = \; H \qquad \text{on } \partial\Omega \tag{3.10}$$

which only admits a solution if

$$\int_\Omega f + \int_{\partial\Omega} H \;=\; 0$$

and, in this case, the solution is determined up to an additive constant. Notice that the constant functions are indeed solutions of the homogeneous *problem ($f = H = 0$), and in fact the* only *solutions if Ω is connected.*

Exo. 3.2 *Show that under the hypotheses of Prop. 3.2 the bilinear form $a(\cdot,\cdot)$ is indeed strongly coercive (as claimed) and provide an estimate of the coercivity constant α.*

Let now $H^1_{Dg}(\Omega) = \{v \in H^1(\Omega); v = g \text{ a.e. on } \Gamma_D\}$. Setting $u = u_g + w$ it is clear that u solves the following problem: *Find $u \in H^1_{Dg}(\Omega)$ such that*

$$a(u,v) = \int_\Omega f\, v \; d\Omega + \int_{\Gamma_N} H\, v \; d\Gamma \tag{3.11}$$

for all $v \in H^1_{D0}(\Omega)$.
Further, if u belongs to $H^2(\Omega)$ integration by parts shows that the partial differential equation holds almost everywhere in Ω and that the Neumann boundary condition is satisfied on Γ_N.
Notice that the Neumann boundary condition enters the right-hand side of (3.11), it is a *natural* condition for this formulation, while the Dirichlet condition has to be imposed to the space in which the solution is sought, it is an *essential* boundary condition. One could wonder whether the Neumann boundary condition could also be imposed as an essential condition: The answer is that the set of functions in $H^1(\Omega)$ which satisfy $\mathbf{n}\cdot(K\nabla u) = H$ on Γ_N is *not* closed in $H^1(\Omega)$, implying that the tools we use to prove existence (the Banach and Hahn-Banach theorems in the general case, the Lax-Milgram lemma in the strongly coercive, Hilbertian case) do not apply.

Exo. 3.3 *Let $\Omega = (0,1)$. Let $\varphi(x) = x$. Show a sequence $\{\varphi_n\} \subset H^1(\Omega)$ such that $\varphi_n'(0) = 0$ for all n and such that $\varphi_n \to \varphi$ strongly in $H^1(\Omega)$.*
Hint: *For $1/n = \epsilon > 0$ consider the "trimmed" function*

$$T_\epsilon \varphi(x) = \begin{cases} \varphi(\epsilon) & \text{if } x < \epsilon \\ \varphi(x) & \text{if } x \geq \epsilon \end{cases}$$

3.2 Ritz-Galerkin approximation

Let $V_h(\Omega)$ be a finite element space contained in $H^1(\Omega)$, and let $V_{h0}(\Omega)$ be the subspace of $V_h(\Omega)$ obtained by putting to zero all degrees of freedom corresponding to values on Γ_D. Analogously, $V_{hg}(\Omega)$ is defined as the (linear) subset of $V_h(\Omega)$ consisting of functions that coincide with some given interpolation $I_h g$ of g on Γ_D. The Ritz-Galerkin approximation of u in $V_h(\Omega)$ then solves:

Find $u_h \in V_{hg}(\Omega)$ such that

$$a(u_h, v_h) = \int_\Omega f\, v_h \, d\Omega + \int_{\partial\Omega} H\, v_h \, d\Gamma \tag{3.12}$$

for all $v_h \in V_{h0}(\Omega)$.

Applying Lax-Milgram lemma to the discrete problem immediately implies that it is well-posed. By Céa's lemma (Lemma 1.26),

$$\|u - u_h\|_1 \leq \frac{N_a}{\alpha} \inf_{v_h \in V_{hg}(\Omega)} \|u - v_h\|_1 \leq \frac{N_a}{\alpha} \|u - \mathcal{I}_h u\|_1$$

Thus, if the local space P_K on each element K of the mesh $\mathcal{T}_h$ contains all polynomials up to degree k and the solution is smooth enough,

$$\|u - u_h\|_1 \leq C h^k |u|_{k+1}$$

3.3 Aubin-Nitsche's duality argument

The error bound in the $H^1(\Omega)$-norm, as shown before, is naturally obtained in the Ritz-Galerkin formulation of second-order PDEs. A first estimate in the $L^2(\Omega)$-norm follows from the continuous injection of $H^1(\Omega)$ into $L^2(\Omega)$, yielding

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

This estimate, however, is not optimal, since the interpolant of u (with u smooth) approximates u with order h^{k+1} in the $L^2(\Omega)$-norm. It is possible to obtain optimal-order estimates using a duality argument. Let us show how it works in the simpler case $\beta = 0$, $g = 0$, $\Gamma_D = \partial\Omega$. Let

$$\mathcal{L}u = -\mathrm{div}(K\nabla u) + \sigma u$$

and assume that the domain is regular enough for $\mathcal{L}$ to have a *smoothing property*, namely that the continuous problem

$$\mathcal{L}w = \mathcal{F}, \qquad w = 0 \quad \text{on } \partial\Omega$$

satisfies

$$\|w\|_{H^2(\Omega)} \leq C_s \|\mathcal{F}\|_{L^2(\Omega)} \tag{3.13}$$

This latter inequality is sometimes called a *regularity estimate*.

Exo. 3.4 *Prove the smoothing property in 1D. More specifically, consider the problem*

$$-(k\,u')' + \sigma\,u = f \qquad\qquad \text{in } \Omega = (0,1) \tag{3.14}$$

with $u(0) = u(1) = 0$, $k, \sigma \in L^\infty(\Omega)$ *satisfying* $k(x) \geq \gamma > 0$ *for all* x *and* $\sigma(x) \geq 0$ *for all* x. *Further, assume that* $k' \in L^\infty(\Omega)$, $f \in L^2(\Omega)$. *Notice that* $k'(x)$ *must be bounded. Show that then there exists* $C > 0$ *such that* $\|u''\|_{L^2(\Omega)} \leq C\,\|f\|_{L^2(\Omega)}$ *and provide an estimate for* C. *Show how this implies (3.13).*

Remark 3.4 *The smoothing property (3.13) holds in 2D/3D if the boundary is very regular, of class* C^2, *or if it is a convex polygon/polyhedron.*

Prop. 3.5 *Under the above hypotheses, there exists $C > 0$ such that*

$$\|u - u_h\|_0 \leq Ch\|u - u_h\|_1 \tag{3.15}$$

Proof. Let w be the unique solution of

$$\mathcal{L}w = u - u_h, \quad w = 0 \ \text{ on } \partial\Omega$$

where we have used the error $e = u - u_h$ as source term. The corresponding variational formulation is

$$a(w, v) = (e, v)_0 \quad \forall\, v \in H_0^1(\Omega)$$

Taking $v = e$ we see that $a(w, e) = \|e\|_0^2$, but also, since the bilinear form is symmetric (otherwise one needs a smoothing property for the adjoint differential operator, but the proof is essentially the same),

$$a(w, e) = a(e, w) = a(u - u_h, w) = a(u - u_h, w - \mathcal{I}_h w)$$

where we have introduced the interpolant of w and used the "orthogonality" property of the Galerkin approximation ($a(u - u_h, v_h) = 0$ for all v_h). Finally

$$\|u - u_h\|_0^2 = a(e, w - \mathcal{I}_h w) \leq N_a\|e\|_1\|w - \mathcal{I}_h w\|_1 \leq N_a\|e\|_1 h\|w\|_2$$

where the last inequality follows from an interpolation estimate for w. Combining with (3.13),

$$\|u - u_h\|_0^2 \leq C_s N_a h\|e\|_1\|e\|_0$$

□

Exo. 3.5 *Let $F(v) = \int_\Omega \psi(x)\, v(x)\, d\Omega$, where ψ is a function in $L^2(\Omega)$. For example, if $\psi = 1$ then $F(v)$ is simply the integral of v. How does $F(u_h)$ converge to $F(u)$ when V_h contains all piecewise polynomials*

of degree k and $\|u - u_h\|_1 \leq C\, h^k$?
Hint: Use a variant of Nitsche's trick. Let w be the solution of

$$a(w, v) = F(v) \qquad \forall\, v \,\in\, V = H_0^1(\Omega)$$

which is the weak form of

$$\mathcal{L}w = \psi \qquad \text{in } \Omega, \qquad w = 0 \quad \text{on } \partial\Omega$$

so that, from the smoothing property, $\|w\|_{H^2(\Omega)} \leq C\, \|\psi\|_{L^2(\Omega)}$. Then use the following calculation

$$F(u - u_h) = a(w, u - u_h) = a(w - \mathcal{I}_h w, u - u_h) \leq N_a \|w - \mathcal{I}_h w\|_1 \|u - u_h\|_1$$

to prove that, if ψ is smooth (at least as smooth as f), then $|F(u) - F(u_h)| \leq \widetilde{C}\, h^{2k}$.
*Another question: What is the expected order of convergence for $F(u) = \int_\omega u\, d\Omega$, with ω a region of the domain? (*Answer: h^{k+1}, why?*).*

3.4 The case $s_{\min} = 0$. Poincaré inequality.

In the case $s_{\min} = 0$ we have to prove strong coercivity without counting on the reaction term, so that we start from the estimate

$$a(v, v) \geq \int_\Omega \nabla v \cdot (K \nabla v)\, d\Omega \qquad \forall v \,\in\, H_{D0}^1(\Omega)$$

which in turn implies

$$a(v, v) \geq K_0 \int_\Omega |\nabla v|^2\, d\Omega = K_0\, |v|_1^2$$

Essentially, we need an estimate of the form $|v|_1 \geq c\|v\|_1$ for some $c > 0$. This is provided by Poincaré-Friedrichs inequality:

Lemma 3.6 (Poincaré-Friedrichs inequality) *In a connected bounded domain, if meas$(\Gamma_D) > 0$ then there exists a constant $c_P > 0$ such that $\|\nabla v\|_0 \geq c_P \|v\|_0$ for all $v \,\in\, H_{D0}^1(\Omega)$.*

Proof. We will prove it in the case $\Gamma_D = \partial\Omega$. We show it first for $\varphi \in \mathcal{D}(\Omega)$ and then extend it to $H_0^1(\Omega)$ by a density argument. We consider φ extended by zero to $\mathbb{R}^n$ and assume that the domain is contained in the strip $a \leq x_1 \leq b$ (in other words, $a \leq x_1 \leq b$ for all $x \in \Omega$). Then, since

$$\varphi(x_1, x_2, \ldots, x_n) = \int_a^{x_1} \frac{\partial \varphi}{\partial x_1}(t, x_2, \ldots, x_n)\ dt$$

we have, using Cauchy-Schwarz inequality,

$$\varphi^2(x_1, x_2, \ldots, x_n) \leq |x_1 - a|\ \int_a^{x_1} \left| \frac{\partial \varphi}{\partial x_1}(t, x_2, \ldots, x_n) \right|^2\ dt$$

integration over x_2 to x_n gives

$$\int \varphi^2 dx_2 \ldots dx_n \leq |x_1 - a| \int_a^{x_1} \ldots \int \left| \frac{\partial \varphi}{\partial x_1} \right|^2\ dt\, dx_2 \ldots dx_n \leq |x_1 - a|\ \int_\Omega \left| \frac{\partial \varphi}{\partial x_1} \right|^2\ d\Omega$$

A final integration over x_1 yields

$$\int_\Omega \varphi^2\ d\Omega \leq \frac{(b-a)^2}{2}\ \int_\Omega \left| \frac{\partial \varphi}{\partial x_1} \right|^2\ d\Omega$$

proving that $c_P \geq \frac{\sqrt{2}}{b-a}$. Now we consider $v \in H_0^1(\Omega)$ and $\varphi_n \to v$, then

$$\|v\|_0 \leq \|\varphi_n\|_0 + \|v - \varphi_n\|_0 \leq \frac{1}{c_P} \|\nabla \varphi_n\|_0 + \|v - \varphi_n\|_0 \leq$$

$$\frac{1}{c_P} \|\nabla v\|_0 + \|v - \varphi_n\|_0 + + \frac{1}{c_P} \|\nabla v - \nabla \varphi_n\|_0 \leq \frac{1}{c_P} \|\nabla v\|_0 + \min\left\{1, \frac{1}{c_P}\right\}\ \|v - \varphi_n\|_1$$

and since $\|v - \varphi_n\|_1$ can be made arbitrarily small, the claim is proved. $\square$

Remark 3.7 *Using Poincaré-Friedrichs inequality, it is easily shown that the bilinear form*

$$a(v,w) = \int_\Omega \left[\nabla v \cdot (K\nabla w) + v\,\beta \cdot \nabla w + \sigma\, v\, w\right]\, d\Omega \tag{3.16}$$

is strongly coercive in $H^1_{D0}(\Omega)$ *whenever meas*$(\Gamma_D) > 0$, $\beta(x)\cdot \mathbf{n}(x) > 0$ *a.e. on* Γ_N, $K_0 > 0$ *and* $s_{\min} \geq 0$.

Exo. 3.6 *Prove the previous remark in detail.*

4 Finite elements for linear elasticity

4.1 Introduction and differential formulation

We recall the usual notations for the Cauchy stress tensor $\boldsymbol{\sigma}$ and the linearized strain tensor

$$\boldsymbol{\epsilon}(u) = \frac{1}{2}\left(\nabla u + \nabla u^T\right) \tag{4.1}$$

where u in this case is a *vector field* corresponding to the *displacement* of the body. We also recall the elastic constitutive law for small deformations,

$$\boldsymbol{\sigma} = \lambda \operatorname{tr}(\boldsymbol{\epsilon}(u))\,\mathbf{I} + 2\mu\boldsymbol{\epsilon}(u) = \lambda \operatorname{div} u\ \mathbf{I} + \mu\left(\nabla u + \nabla u^T\right) \tag{4.2}$$

where λ and μ are the Lamé coefficients, which in general depend on the point x and by thermodynamic reasons are constrained to satisfy, for almost all x,

$$\mu(x) > 0; \qquad \lambda(x) + \frac{2}{3}\mu(x) \geq 0 \tag{4.3}$$

Differential Formulation: The governing equation follows from the static equilibrium balance, which reads

$$\operatorname{div}\boldsymbol{\sigma} + f = 0 \tag{4.4}$$

where f is a vector field of applied forces. Replacing the expression of $\boldsymbol{\sigma}$ in terms of u one obtains an equation for the displacement field. This problem admits both Dirichlet and Neumann boundary conditions on u:

$$u = g \qquad \text{on } \Gamma_D; \qquad \sigma \cdot \mathbf{n} = \mathcal{F} \qquad \text{on } \Gamma_N \tag{4.5}$$

where $\mathcal{F}$ is a field of surface forces applied on Γ_N, $\Gamma_N \cap \Gamma_D = \emptyset$ and $\overline{\Gamma_N \cup \Gamma_D} = \partial\Omega$. The domain Ω corresponds to the region of space occupied by the body under consideration, both before and after the application of the forces since just problems with *small displacements* are being considered.

Exo. 4.1 *Let u_1, u_2 be the components of u in a planar elasticity case in which the domain is the unit square. The boundary conditions are: zero displacement on the bottom boundary ($x_2 = 0$), and a normal force equal to P on the rest of $\partial\Omega$. Write down the system of two equations and two unknowns for u_1 and u_2 considering λ and μ independent of x_1 and x_2.*
Hint: Equation (4.4), written in Cartesian indices, becomes

$$\sum_{j=1}^{d} \partial_j \, \sigma_{ij} + f_i = 0 \qquad\qquad \forall \, i \, = 1, \ldots, d$$

and (4.2) becomes,

$$\sigma_{ij} = \lambda(\partial_1 u_1 + \partial_2 u_2) \, \delta_{ij} + \mu \;\; (\partial_j u_i + \partial_i u_j) \, .$$

It remains to replace the latter into the former. For the boundary force we have that, if $\mathbf{x} = (x_1, x_2) \in \partial\Omega$ then at $\mathbf{x}$ we have

$$(\boldsymbol{\sigma} \cdot \mathbf{n})_1 = [(\lambda + 2\mu)\partial_1 u_1 + \lambda \partial_2 u_2] \, n_1 + \mu \;\; (\partial_2 u_1 + \partial_1 u_2) \; n_2 = - \, P \, n_1$$

$$(\boldsymbol{\sigma} \cdot \mathbf{n})_2 = [\lambda \partial_1 u_1 + (\lambda + 2\mu)\partial_2 u_2] \, n_2 + \mu \;\; (\partial_2 u_1 + \partial_1 u_2) \; n_1 = - \, P \, n_2$$

As a consequence, along $x_1 = 0$ (left boundary), the boundary conditions are

$$(\lambda + 2\mu)\partial_1 u_1 + \lambda \partial_2 u_2 = -P, \qquad\qquad \partial_2 u_1 + \partial_1 u_2 = 0$$

the conditions at the other boundaries are analogous.

4.2 Variational Formulation

The variational formulation of this problem can be obtained from the corresponding PDE by integration by parts. In Mechanics, however, it is considered a *fundamental principle*: The Principle of Virtual Work (or of Virtual Power)

Principle of Virtual Power: The internal virtual power of the stresses ($\int_\Omega \boldsymbol{\sigma} : \boldsymbol{\epsilon}(v)$) plus the virtual power of the acceleration ($\int_\Omega \rho\, a \cdot v$) equals the virtual power of the applied forces. This holds for all virtual velocity fields, that is, all vector fields v that are kinematically admissible variations of the body motion.

$$\int_\Omega \boldsymbol{\sigma} : \boldsymbol{\epsilon}(v) + \int_\Omega \rho\, a \cdot v = \int_\Omega f \cdot v + \int_{\Gamma_N} \mathcal{F} \cdot v \qquad \forall\, v \in \text{VAR} \tag{4.6}$$

The kinematically admissible motions must belong to

$$\text{KIN} \;=\; V_{Dg} = \{v \in [H^1(\Omega)]^n; v = g \text{ on } \Gamma_D\} \tag{4.7}$$

so that their variations must belong to

$$\text{VAR} \;=\; V_{D0} = \{v \in [H^1(\Omega)]^n; v = 0 \text{ on } \Gamma_D\} \tag{4.8}$$

The variational formulation of **linear elastostatics** then reads:

"Find $u \in V_{Dg}$ such that

$$a(u,v) = \int_\Omega f \cdot v \; d\Omega + \int_{\Gamma_N} \mathcal{F} \cdot v \; d\Gamma =: \ell(v) \tag{4.9}$$

for all $v \in V_{D0}$", where

$$a(u,v) = \int_\Omega \boldsymbol{\sigma}(\boldsymbol{u}) : \boldsymbol{\epsilon}(v) \; d\Omega = \int_\Omega \left[\lambda \operatorname{div} u \; \operatorname{div} v + 2\mu \boldsymbol{\epsilon}(u) : \boldsymbol{\epsilon}(v)\right] \; d\Omega \tag{4.10}$$

4.3 Well-posedness and Galerkin approximation

Theorem 4.1 *(***Korn's inequality***) Let Ω be a domain in $\mathbb{R}^n$. There exists $C_K > 0$ such that*

$$\|v\|_1 \leq C_K \|\boldsymbol{\epsilon}(v)\|_0 \qquad \forall v \in H_0^1(\Omega)^n \tag{4.11}$$

It is not necessary that v be zero on the whole of $\partial\Omega$, the same result holds if $\text{meas}(\Gamma_D) > 0$ (in connected domains), so that we have strong coercivity of the bilinear form on V. This gives the result below.

Theorem 4.2 *Let Ω be a regular domain on which the elasticity problem (4.9) is posed with meas$(\Gamma_D) > 0$, $f \in L^2(\Omega)^n$ and $\mathcal{F} \in L^2(\Gamma_N)^n$. We assume that the Lamé coefficients are bounded and satisfy (4.3). Then there exists a unique solution u, and there exists $c > 0$ such that*

$$\|u\|_1 \leq c \left(\|f\|_0 + \|\mathcal{F}\|_{0,\Gamma_N}\right) \tag{4.12}$$

Proof. $V = V_{D0}$ is a Hilbert space, the bilinear form is continuous with

$$a(u,v) \leq c \max\{\lambda_{\max}, \mu_{\max}\} \quad \|\nabla u\|_0 \; \|\nabla v\|_0$$

From Korn's inequality we also have

$$a(v,v) = \int_\Omega \left[\lambda(\operatorname{div} v)^2 + 2\mu\boldsymbol{\epsilon}(v) : \boldsymbol{\epsilon}(v)\right]\ d\Omega \geq c\mu_{\min}\|v\|_1^2$$

It only remains to apply Lax-Milgram lemma.
□
Let

$$V_h = \{v_h \in C^0(\overline{\Omega})^n, v_h|_K \in (P_K)^n, v_h = 0 \text{ on } \Gamma_D\}. \tag{4.13}$$

Since $V_h \subset V$, we have well-posedness and convergence of the discrete problem.

Prop. 4.3 *The solution $u_h \in V_{hg}$ satisfying*

$$a(u_h, v_h) = \ell(v_h) \qquad \forall\, v_h \in V_{h0} \tag{4.14}$$

exists and is unique. It satisfies $\lim_{h\to 0}\|u - u_h\|_1 = 0$. If $u \in H^{\ell+1}(\Omega)^n$ for some $\ell \leq k$, with k such that $P_k(K) \subset P_K$, then there exists $c > 0$ such that

$$\|u - u_h\|_1 \leq c\, h^\ell |u|_{\ell+1} \tag{4.15}$$

Exo. 4.2 *Build an extremal formulation of the linear elasticity problem.*
Hint: Consider

$$J(w) = \int_\Omega \left[\frac{\lambda}{2}(\operatorname{div} w)^2 + \mu\,\boldsymbol{\epsilon}(w) : \boldsymbol{\epsilon}(w)\right]\ d\Omega - \int_\Omega f\cdot w\ d\Omega - \int_{\Gamma_N} \mathcal{F}\cdot w\ d\Gamma \tag{4.16}$$

where the first integral is the "strain energy" of the body. The solution u is the displacement field that minimizes J over V_{Dg},

$$J(u) = \inf_{w\in V_{Dg}} J(w) \tag{4.17}$$

4.4 Implementation aspects

A significant difference between the elastostatics problem and the convection-diffusion-reaction problem discussed earlier is that the elasticity unknown is a vector field.
Let $\{\mathcal{N}^j\}$ $(j = 1, \ldots, M)$ be the scalar basis functions associated to a mesh $\mathcal{T}_h$. The space V_h is now of dimension $n \times M$, as to each node j correspond n basis functions:

$$\mathbf{N}^{j,1}(x) = \mathcal{N}^j(x)\,\check{\mathbf{e}}^1 = (\mathcal{N}^j(x), 0) \qquad \ldots \qquad \mathbf{N}^{j,n}(x) = \mathcal{N}^j(x)\,\check{\mathbf{e}}^n = (0, \mathcal{N}^j(x)) \tag{4.18}$$

where we have chosen the local basis $\{\check{\mathbf{e}}^\alpha\}$ equal to the canonical basis ($\check{e}^\alpha_\beta = \delta_{\alpha\beta}$), but any other can be chosen and sometimes is.

Exo. 4.3 *Compute the following in terms of the scalar basis* $\{\mathcal{N}^j\}$*:*

- $div\,(\mathbf{N}^{j,\alpha})$ *(Answer:* $= \partial_\alpha \mathcal{N}^j$*)*
- $\boldsymbol{\epsilon}(\mathbf{N}^{j,\alpha})$
- $\int_K div\,(\mathbf{N}^{j,\alpha})\, div\,(\mathbf{N}^{k,\beta})$
- $\int_K \boldsymbol{\epsilon}(\mathbf{N}^{j,\alpha}) : \boldsymbol{\epsilon}(\mathbf{N}^{k,\beta})$

5 Finite elements for mixed problems

5.1 Constraints and Lagrange multipliers

Applications of the FEM usually involve *constraints* on the admissible set of solutions. Let us briefly describe some examples.

5.1.1 Incompressible elasticity

There exist elastic materials which behave as incompressible, in the sense that they preserve their volume in every deformation. Under the hypothesis of small deformations, the preservation of volume is equivalent to the deformation field having zero divergence,

$$\operatorname{div} u = 0 \qquad \text{a.e. in } \Omega \tag{5.1}$$

Considering the elastic energy functional seen in the previous section (where λ is assumed independent of x for simplicity and $\|\boldsymbol{\epsilon}(v)\|^2 = \boldsymbol{\epsilon}(v) : \boldsymbol{\epsilon}(v)$)

$$J(v) = \frac{\lambda}{2} \int_\Omega (\operatorname{div} v)^2 \, d\Omega + \int_\Omega \mu \, \|\boldsymbol{\epsilon}(v)\|^2 \, d\Omega - \int_\Omega f \cdot v \, d\Omega - \int_{\Gamma_N} \mathcal{F} \cdot v \, d\Gamma \tag{5.2}$$

one can view the first term as a penalization (with coefficient λ) of the incompressibility constraint. As a consequence, totally incompressible behavior corresponds to $\lambda \to +\infty$ in theory, and to λ very large, much larger than the shear modulus μ, in practice.
For the *Primal Formulation*, which is the one we have been studying up to now, the divergence-free constraint is treated as an *essential constraint*, just like the Dirichlet constraints, and is incorporated into the set of admissible displacement fields,

$$Z_{Dg} \stackrel{\text{def}}{=} \{v \in V_{Dg} \mid \operatorname{div} v = 0 \text{ a.e. in } \Omega\} \tag{5.3}$$

Inside Z_{Dg} the first term of J becomes irrelevant, so that defining

$$\widetilde{J}(v) = \int_\Omega \mu \, \|\boldsymbol{\epsilon}(v)\|^2 \, d\Omega - \int_\Omega f \cdot v \, d\Omega - \int_{\Gamma_N} \mathcal{F} \cdot v \, d\Gamma, \tag{5.4}$$

we have the *Primal Extremal Formulation of incompressible elasticity.*

Primal Extremal Formulation of incompressible elasticity: *Find* $u \in Z_{Dg}$ *that minimizes* $\widetilde{J}$ *over* Z_{Dg}, i.e.,

$$\widetilde{J}(u) \le \widetilde{J}(v) \qquad \forall\, v \in Z_{Dg} \tag{5.5}$$

Defining now

$$\widetilde{a}(u,v) = \int_\Omega 2\,\mu\, \boldsymbol{\epsilon}(u) : \boldsymbol{\epsilon}(v) \, d\Omega, \qquad \text{and} \quad \ell(v) = \int_\Omega f \cdot v \, d\Omega + \int_{\Gamma_N} \mathcal{F} \cdot v \, d\Gamma \tag{5.6}$$

we have

$$\widetilde{J}(v) = \frac{1}{2}\,\widetilde{a}(v,v) - \ell(v) \tag{5.7}$$

and also the

Primal Variational Formulation of incompressible elasticity: *Find* $u \in Z_{Dg}$ *such that*

$$\widetilde{a}(u,v) = \ell(v) \qquad \forall\, v \in Z_{D0} \tag{5.8}$$

It can be shown that problem (5.8) is indeed well posed, so that a unique solution u exists. However, the imposition of the zero-divergence constraint on the space creates several difficulties for the finite element discretization.
It is thus convenient to replace the Primal Extremal Formulation by the following equivalent one:

Mixed Extremal Formulation of incompressible elasticity: Defining $b(\cdot,\cdot) : H^1(\Omega)^d \times L^2(\Omega) \to \mathbb{R}$ by

$$b(v,q) = \int_\Omega q \,\mathrm{div}\, v \; d\Omega \tag{5.9}$$

and the Lagrangian $\mathcal{L} : H^1(\Omega)^d \times L^2(\Omega) \to \mathbb{R}$ by

$$\mathcal{L}(v,q) = \widetilde{J}(v) - b(v,q) = \frac{1}{2}\widetilde{a}(v,v) - \ell(v) - b(v,q) \;, \tag{5.10}$$

problem (5.5) becomes equivalent to "Find $(u,p) \in V_{Dg} \times L^2(\Omega)$ that is an extremal point (saddle point) of $\mathcal{L}$", or, in other words,

$$\mathcal{L}(u,p) = \widetilde{J}(u) = \inf_{v \in Z_{Dg}} \widetilde{J}(v) = \inf_{v \in V_{Dg}} \sup_{q \in L^2(\Omega)} \mathcal{L}(v,q) \tag{5.11}$$

The extremality conditions for $\mathcal{L}$ are

$$d\mathcal{L}(v,0) = \lim_{t\to 0} \frac{\mathcal{L}(u+tv,p) - \mathcal{L}(u,p)}{t} = 0 \qquad \forall v \in V_{D0} \tag{5.12}$$

$$d\mathcal{L}(0,q) = \lim_{t\to 0} \frac{\mathcal{L}(u,p+tq) - \mathcal{L}(u,p)}{t} = 0 \qquad \forall q \in L^2(\Omega) \tag{5.13}$$

and lead to the mixed variational formulation.

Mixed Variational Formulation of incompressible elasticity: Find $(u,p) \in V_{Dg} \times L^2(\Omega)$ such that

$$\widetilde{a}(u,v) - b(v,p) = \ell(v) \qquad \forall v \in V_{D0} \tag{5.14}$$

$$b(u,q) = 0 \qquad \forall q \in L^2(\Omega) \tag{5.15}$$

The enforcement of incompressibility in this formulation is not built in the space for u, which is V_{Dg} and not Z_{Dg}. Instead, it appears explicitly in equation (5.15), because

$$b(u,q) = \int_\Omega q \ \text{div}\ u \ d\Omega = 0 \quad \forall\, q \in L^2(\Omega) \qquad \Leftrightarrow \qquad \text{div}\ u = 0 \quad \text{a.e. in } \Omega. \tag{5.16}$$

Integrating by parts the left-hand side of (5.14) one arrives at the

Differential Formulation of incompressible elasticity:

$$-\,\text{div}\ \widetilde{\boldsymbol{\sigma}}(u) \ + \ \nabla p = f, \qquad \text{where } \widetilde{\boldsymbol{\sigma}}(u) = 2\,\mu\,\boldsymbol{\epsilon}(u) \tag{5.17}$$

$$\text{div}\ u = 0 \tag{5.18}$$

$$u = g \qquad \text{on } \Gamma_D \tag{5.19}$$

$$(-p\,\mathbf{I} + \widetilde{\boldsymbol{\sigma}})\cdot \mathbf{n} = \mathcal{F} \qquad \text{on } \Gamma_N \tag{5.20}$$

It is important to notice that the incompressibility constraint "materializes" in the equilibrium equation (5.17) as the gradient of the unknown pressure p, and at the force boundary as a normal contribution $-p\,\mathbf{n}$. In mechanical terms, this means that the Cauchy stress tensor of an incompressible elastic material is

$$\boldsymbol{\sigma} = -\,p\,\mathbf{I} \ + \ \widetilde{\boldsymbol{\sigma}} \ = \ -\ p\,\mathbf{I} \ + \ 2\,\mu\ \boldsymbol{\epsilon}(u) \tag{5.21}$$

Exo. 5.1 *Show that the extremality conditions (5.12)-(5.13) are equivalent to the mixed formulation equations (5.14)-(5.15).*

Exo. 5.2 *Show that, with sufficient regularity of u and p, (5.14) implies (5.17) and (5.20).*

5.1.2 Dirichlet conditions as constraints

Up to now we have imposed the Dirichlet condition directly on the formulation space, but this condition can also be seen as a *constraint* and given a treatment similar to that given to incompressibility in the previous paragraph.
Let us illustrate this possibility in the case of a purely diffusive problem, for which the DF, PVF and PEF read

Differential Formulation:

$$-\operatorname{div}(K\nabla u) = f \ \text{ in } \ \Omega, \qquad u = g \ \text{ on } \ \partial\Omega \tag{5.22}$$

Primal Variational Formulation: Find $u \in V_g = \{v \in H^1(\Omega),\ v = g \text{ on } \partial\Omega\}$ such that

$$a(u,v) = \int_\Omega K\nabla u \cdot \nabla v \, d\Omega = \int_\Omega f\, v \, d\Omega = \ell(v) \qquad \forall\, v \in V_0 \tag{5.23}$$

Primal Extremal Formulation: Find u that extremizes J over V_g, where

$$J(v) = \frac{1}{2}\, a(v,v) - \ell(v) \tag{5.24}$$

To remove the Dirichlet constraint from the space, we introduce the Lagrangian

$$\mathcal{L}(v,\zeta) = J(v) - b(v-g,\zeta) \tag{5.25}$$

where

$$b(v-g,\zeta) = \int_{\partial\Omega} \zeta\,(v-g)\; d\Gamma = (\zeta, v-g)_{L^2(\partial\Omega)} \tag{5.26}$$

In fact, the mixed formulation that will soon be introduced is defined on a larger space than $L^2(\partial\Omega)$, denoted by $H^{-1/2}(\partial\Omega)$. The scalar product of $L^2(\partial\Omega)$ can be continuously extended as a "duality pairing" $\langle\zeta, w\rangle$ between $\zeta \in H^{-1/2}(\partial\Omega)$ and w which is in the space of traces of functions belonging to $H^1(\Omega)$, denoted by $H^{1/2}(\partial\Omega)$. Whenever ζ belongs to $L^2(\partial\Omega)$, $b(w,\zeta) = (\zeta, w)_{L^2(\partial\Omega)}$ but, in general,

$$b(w,\zeta) = \langle w, \zeta\rangle \tag{5.27}$$

Now we can write down the mixed extremal formulation of the Dirichlet problem and the mixed variational formulation that results from the corresponding extremality conditions.

Mixed Extremal Formulation: Find (u,λ) that extremizes $\mathcal{L}(\cdot,\cdot)$ over $H^1(\Omega)\times H^{-1/2}(\partial\Omega)$, i.e.,

$$\mathcal{L}(u,\lambda) = J(u) = \inf_{v\,\in\,H^1(\Omega)}\ \sup_{\zeta\,\in\,H^{-1/2}(\partial\Omega)} \mathcal{L}(v,\zeta) \tag{5.28}$$

Mixed Variational Formulation: Find $(u,\lambda) \in V\times Q = H^1(\Omega)\times H^{-1/2}(\partial\Omega)$ such that

$$\begin{aligned} a(u,v) - b(v,\lambda) &= \ell(v) \qquad \forall\, v \in V && (5.29)\\ b(u,\zeta) &= b(g,\zeta) \qquad \forall\, \zeta \in Q && (5.30) \end{aligned}$$

and integrating by parts $a(u,v)$ we arrive at the

Mixed Differential Formulation:

$$-\text{div}\,(K\,\nabla u) = f \qquad \text{in } \Omega \tag{5.31}$$

$$K\,\frac{\partial u}{\partial n} - \lambda = 0 \qquad \text{on } \partial\Omega \tag{5.32}$$

$$u = g \qquad \text{on } \partial\Omega \tag{5.33}$$

which brings as new information that the Lagrange multiplier λ is in fact equal to the diffusive flux $K\partial_n u$ across $\partial\Omega$.

Exo. 5.3 *Show how to derive the mixed VF from the mixed EF.*

Exo. 5.4 *Show how to derive the mixed DF from the mixed VF.*

5.2 Abstract mixed formulation

Generalizing the previous examples, one considers the problem

Abstract Mixed Problem: Find $(u, p) \in V \times Q$ such that

$$a(u, v) - b(v, p) = \ell(v) \qquad \forall\, v \in V \tag{5.34}$$

$$b(u, q) = g(q) \qquad \forall\, q \in Q \tag{5.35}$$

where $a : V \times V \to \mathbb{R}$, $b : V \times Q \to \mathbb{R}$ are continuous bilinear forms, $\ell \in V'$, $g \in Q'$.
When $a(\cdot, \cdot)$ is symmetric, it is equivalent to the extremization of

$$J(v) = \frac{1}{2}\, a(v, v) - \ell(v) \tag{5.36}$$

over the (constrained) set

$$Z_g = \{v \in V \mid b(v, q) = g(q) \;\; \forall\, q \in Q\} \tag{5.37}$$

and to the extremization over $V \times Q$ (i.e., unconstrained) of the Lagrangian

$$\mathcal{L}(v, q) = J(v) - b(v, q) + g(q) \tag{5.38}$$

The first logical question is whether (5.34)-(5.35) is well-posed. We consider both the cases where V and Q are infinite-dimensional (the continuous case) and finite-dimensional (the discrete case).

Theorem 5.1 *If $a(\cdot,\cdot)$ is strongly coercive on Z_0,*

$$a(v,v) \geq \alpha \, \|v\|_V^2 \qquad \forall \, v \, \in \, Z_0 \tag{5.39}$$

with $\alpha > 0$, and if

$$\inf_{q \in Q} \sup_{v \in V} \frac{b(v,q)}{\|q\|_Q \; \|v\|_V} = \gamma > 0 \tag{5.40}$$

then (5.34)-(5.35) is well-posed.

The proof of this result relies on applying Thm. 1.7 to the setting defined by the product space $W = V \times Q$, the bilinear form $B : W \times W \to \mathbb{R}$ defined by

$$B((u,p),(v,q)) = a(u,v) - b(v,p) - b(u,q) \tag{5.41}$$

and the linear form $S \in W'$ defined by

$$S(v,q) = \ell(v) - g(q). \tag{5.42}$$

Exo. 5.5 *The Abstract Mixed Problem (5.34)-(5.35) is equivalent to the problem:* Find $(u,p) \in W$ such that

$$B((u,p),(v,q)) = S(v,q) \qquad \forall \, (v,q) \, \in \, W \tag{5.43}$$

Now it only remains to prove that,

Theorem 5.2 *(Brezzi) Under hypotheses (5.39) and (5.40), the bilinear form $B(\cdot,\cdot)$ is weakly coercive on $V \times Q$.*

Proof. To simplify things, assume that (5.39) holds $\forall\, v \in V$ and that $a(\cdot,\cdot)$ is symmetric. Taking (u,p) arbitrary in $V \times Q$, choose $w \in V$ such that $\|w\|_V = \|p\|_Q$ and $-b(w,p) \geq \gamma \|p\|^2$. Then, taking $\eta = \alpha\gamma/N_a^2$, one gets

$$B((u,p),(u+\eta w,p)) \geq \frac{\alpha}{2}\ \min\left\{1, \frac{\gamma^2}{N_a^2}\right\}\ \|(u,p)\|^2_{V\times Q}$$

Besides,

$$\|(u+\eta w,p)\|_{V\times Q} \leq \left(1 + \frac{\alpha\gamma}{N_a^2}\right) \|(u,p)\|_{V\times Q}$$

so that

$$\inf_{(u,p)} \sup_{(v,q)} \frac{B((u,p),(v,q))}{\|(u,p)\|\ \|(v,q)\|} \geq \inf_{(u,p)} \frac{B((u,p),(u+\eta w,p))}{\|(u,p)\|\ \|(u+\eta w,p)\|} \geq \frac{\frac{\alpha}{2}\min\left\{1, \frac{\gamma^2}{N_a^2}\right\}}{1+\frac{\alpha\gamma}{N_a^2}} > 0$$

and condition (1.21) is satisfied. Since B is symmetric, the proof is complete. As a by-product, we observe that the coercivity constant of $B(\cdot,\cdot)$ can be chosen as

$$\beta = \frac{\frac{\alpha}{2}\min\left\{1, \frac{\gamma^2}{N_a^2}\right\}}{1+\frac{\alpha\gamma}{N_a^2}} \tag{5.44}$$

□

Exo. 5.6 *Prove that, for all (u,p) and (v,q) in $V \times Q$,*

$$B((u,p),(v,q)) \leq (N_a + 2N_b)\ \|(u,p)\|_{V\times Q}\ \|(v,q)\|_{V\times Q} \tag{5.45}$$

5.3 Abstract approximation

Now we consider the following abstract setting, in which $V_h \subset V$ and $Q_h \subset Q$:

H1 Let $(u,p) \in V \times Q$ satisfy

$$B((u,p),(v_h,q_h)) = S(v_h,q_h) \qquad \forall\,(v_h,q_h) \in V_h \times Q_h \tag{5.46}$$

with the definitions (5.41)-(5.42), assuming all linear and bilinear forms involved are bounded. Notice that we do not assume that $B(\cdot,\cdot)$ coincides with that of the exact mixed formulation on $V \times Q$. The analysis thus includes *non-Galerkin* approximations. B could depend on the mesh.

H2 The subspaces V_h and Q_h are such that

$$\inf_{q_h \in Q_h} \sup_{v_h \in V_h} \frac{b(v_h,q_h)}{\|v_h\|_Q \, \|q_h\|_V} = \gamma_h > 0 \tag{5.47}$$

and

$$a(v_h,v_h) \geq \alpha_h \, \|v_h\|_V^2 \qquad \forall\, v_h \in Z_{h0}, \tag{5.48}$$

with $\alpha_h > 0$ and

$$Z_{h0} = \{v_h \in V_h \mid b(v_h,q_h) = 0 \;\; \forall\, q_h \in Q_h\}. \tag{5.49}$$

Theorem 5.3 *Under the hypotheses* **H1** *and* **H2** *above, the approximation* $(u_h,p_h) \in V_h \times Q_h$ *defined by*

$$B((u_h,p_h),(v_h,q_h)) = S(v_h,q_h) \qquad \forall\,(v_h,q_h) \in V_h \times Q_h \tag{5.50}$$

exists and is unique. Further, there exists $C = C(N_a,N_b,\alpha_h,\gamma_h)$ *such that*

$$\|u-u_h\|_V + \|p-p_h\|_Q \leq C \left(\inf_{v_h \in V_h} \|u-v_h\|_V \; + \; \inf_{q_h \in Q_h} \|p-q_h\|_Q \right) \tag{5.51}$$

Exo. 5.7 *Prove the previous theorem.* Hint: Use Lemma 1.11. The hypothesis H2, together with (5.44) applied to the discrete problem and (5.45) allow to estimate

$$C = 1 + \frac{N_a + 2N_b}{\beta_h} = 1 + \frac{2\,(N_a + 2N_b)\left(1 + \frac{\alpha_h \gamma_h}{N_a^2}\right)}{\alpha_h \ \min\left\{1, \frac{\gamma_h^2}{N_a^2}\right\}} \tag{5.52}$$

Exo. 5.8 *Show that u_h that solves (5.50) also solves:* Find $u_h \in Z_{hg}$ such that

$$a(u_h, v_h) = B((u_h, 0), (v_h, 0)) = S(v_h, 0) = \ell(v_h) \qquad \forall\, v_h \in Z_{h0} \tag{5.53}$$

where

$$Z_{hg} = \{v_h \in V_h \mid b(v_h, q_h) = g(q_h) \ \ \forall\, q_h \in Q_h\} \tag{5.54}$$

- Optimal approximation properties are obtained for the mixed problem on the unconstrained space V_h.
- The space Q_h needs to be chosen such that the inf-sup condition is satisfied, and such that $\|p - \mathcal{I}_h p\|_Q$ is sufficiently small to not degrade the approximation of u. The norm $\|\cdot\|_Q$ is usually weaker than $\|\cdot\|_V$, allowing Q_h to be *coarser*, or of *lower order*, than V_h.
- Estimate (5.52) shows that if there exist $\alpha_0 > 0$ and $\gamma_0 > 0$ such that $\alpha_h \geq \alpha_0$ and $\gamma_h \geq \gamma_0$ for all h, then C in (5.51) can be taken independent of h.

5.4 Application to the Dirichlet problem

Let us consider the Mixed Variational Formulation of the Dirichlet problem (5.29)-(5.30). It has the same structure as (5.34)-(5.35) with $V = H^1(\Omega)$ and $Q = H^{-1/2}(\partial\Omega)$. It is interesting to recognize the different actors of the abstract mixed formulation in a concrete example like this one:

$$a(u,v) = \int_\Omega K\,\nabla u\cdot\nabla v\;d\Omega \tag{5.55}$$

$$b(v,\zeta) = \langle\zeta, v\rangle \simeq \int_{\partial\Omega}\zeta\,v\;d\Gamma \tag{5.56}$$

$$Z_0 = \{v\in V \mid b(v,\zeta) = \langle\zeta,v\rangle = 0\ \forall\zeta\in Q\} = \{v\in V \mid v = 0 \text{ a.e. on } \partial\Omega\} = H^1_0(\Omega) \tag{5.57}$$

Notice that $a(\cdot,\cdot)$ is not coercive on V, as $a(z,v) = 0\ \forall\, v\in V$ whenever z is a constant function. However, $a(\cdot,\cdot)$ is indeed (strongly) coercive on Z_0 as a consequence of Poincaré-Friedrichs inequality, and this is what is needed for the problem to be well-posed.

Now let $V_h\subset V$ and $Q_h\subset Q$ be approximation subspaces, and let $(u_h,\lambda_h)\in V_h\times Q_h$ be the Galerkin approximation defined by

$$\int_\Omega K\,\nabla u_h\cdot\nabla v_h\;d\Omega - \int_{\partial\Omega}\lambda_h\,v_h\;d\Gamma = \int_\Omega f\,v_h\;d\Omega \qquad \forall\, v_h\in V_h \tag{5.58}$$

$$\int_{\partial\Omega}\zeta_h\,u_h\;d\Gamma = \int_{\partial\Omega}\zeta_h\,g\;d\Gamma \qquad \forall\,\zeta_h\in Q_h. \tag{5.59}$$

Then, if the discrete spaces satisfy

$$\inf_{\zeta_h\in Q_h}\ \sup_{v_h\in V_h}\ \frac{\int_{\partial\Omega}\zeta_h\,v_h\;d\Gamma}{\|\zeta_h\|_Q\,\|v_h\|_V} = \gamma_h > 0\ , \tag{5.60}$$

the approximate solution satisfies

$$\|u-u_h\|_V + \|\lambda-\lambda_h\|_Q \le C\left(\inf_{v_h\in V_h}\|u-v_h\|_V + \inf_{\zeta_h\in Q_h}\|\lambda-\zeta_h\|_Q\right). \tag{5.61}$$

Comments:

- The primal approximate solution u_h^P, in which the approximating space is $V_{hg} = \{v_h \in V_h \mid v_h = g \text{ on } \partial\Omega\}$, solves

$$a(u_h^P, v_h) = \ell(v_h) \qquad \forall\, v_h \in V_{h0}. \tag{5.62}$$

  By comparing to (5.53)-(5.54), one observes that u_h^P coincides with the solution u_h of the mixed approximate formulation whenever Q_h is large enough for Z_{h0} to coincide with V_{h0}. In other words, whenever the discrete multiplier space enforces $u_h = g$ pointwise on $\partial\Omega$. This would be the case if one took $Q_h = Q$, or $Q_h = L^2(\partial\Omega)$, in which case one recovers the standard Galerkin formulation but with the inf-sup constant γ_h equal to zero. It is easily seen that λ_h is not uniquely defined, and the error estimate boils down to the approximation properties of V_{hg}; i.e.,

$$\|u - u_h\|_V \le C \inf_{v_h \in V_{hg}} \|u - v_h\|_V \ . \tag{5.63}$$

Exo. 5.9 *Verify the previous assertions. Why is $\gamma_h = 0$ if V_h is finite-dimensional and $Q_h = L^2(\Omega)$?*

- Notice that switching from the primal approximation, in which the Dirichlet boundary condition is directly imposed on V_h, to the mixed approximation, we have significantly increased the computational cost:
  - Instead of eliminating the boundary unknowns, we have approximately doubled them. The boundary unknowns are now true unknowns, and we have added the additional unknowns corresponding to the degrees of freedom of Q_h.
  - In the primal formulation the system matrix is positive definite, amenable to the use of the most effective algorithms of computational linear algebra. The mixed formulation, on the other hand, exhibits a matrix that is symmetric but indefinite, with both positive and negative eigenvalues. It is also ill-conditioned, since u and λ have different units and thus arbitrarily different ranges of value.
- The additional cost has not brought increased accuracy in the computation of u_h if the mesh fits the boundary, since in this case V_{hg} provides an approximation that is of the same order as that provided by V_h; i.e.,

$$\|u - u_h^P\|_V \simeq \inf_{v_h \in V_{hg}} \|u - v_h\|_V \simeq \inf_{v_h \in V_h} \|u - v_h\|_V \simeq \|u - u_h\| \tag{5.64}$$

- The discrete solution u_h of the mixed formulation does not satisfy $u_h = g$ pointwise on $\partial\Omega$, which the primal solution u_h^P satisfies (up to an interpolation of g).

- The mixed formulation provides an approximation of $K\,\partial_n u$. Notice that the primal formulation produces a solution u_h^P which satisfies $\|\nabla u_h^P - \nabla u\|_0 = \mathcal{O}(h)$, but $K\,\partial_n u_h$ may well not converge, as a function defined on $\partial\Omega$, to $K\,\partial_n u$.

- Consequences are very interesting when considering *non-fitted meshes* (or *immersed boundaries*), and also when performing *domain decomposition* with *non-matching meshes* (in which case Q_h is a "mortar" space).

- The design and implementation of spaces Q_h satisfying the inf-sup condition (5.60) uniformly in h is quite cumbersome. This is why the *stabilization methods* have become very popular. We will discuss this further later on.

- You should by now be in a position to appreciate R. Stenberg's article "On some techniques for approximating boundary conditions in the finite element method" (J. Comp. Appl. Math. 63, 139-148, 1995), which is a recommended reading of the course.

- For those avid of more material on mixed finite element methods I suggest starting with "A brief excursion into the mathematical theory of mixed finite element methods", by E. Süli, available at Prof. Süli's website and at the courses' website.

5.5 Application to incompressible elasticity and to Stokes flow

The mixed variational formulation of incompressible elasticity is: *Find* $(u,p) \in V_{Dg} \times L^2(\Omega)$ *such that*

$$\int_\Omega 2\,\mu\,\boldsymbol{\epsilon}(u) : \boldsymbol{\epsilon}(v)\; d\Omega - \int_\Omega p\;\mathrm{div}\;v\; d\Omega \;=\; \int_\Omega f\cdot v\; d\Omega + \int_{\Gamma_N} \mathcal{F}\cdot v\; d\Gamma \qquad \forall\, v\,\in\, V_{D0} \tag{5.65}$$

$$\int_\Omega q\;\mathrm{div}\;u\; d\Omega \;=\; 0 \qquad \forall\, q\,\in\, L^2(\Omega) \tag{5.66}$$

which fits nicely in the framework (5.34)-(5.35). This exact same mathematical problem corresponds to Stokes flow, in which u is the velocity field of an incompressible Newtonian fluid of viscosity μ. Stokes flow models fluid flow in conditions in which inertial effects are negligible, as happens for example in *microfluidics*.

We identify the components of the abstract mixed formulation:

$$a(u,v) \;=\; \int_\Omega 2\,\mu\,\boldsymbol{\epsilon}(u) : \boldsymbol{\epsilon}(v)\; d\Omega \tag{5.67}$$

$$b(v,q) \;=\; \int_\Omega q\;\mathrm{div}\;v\; d\Omega \tag{5.68}$$

$$Z_0 \;=\; \{v\,\in\, V_{D0}\mid \int_\Omega q\;\mathrm{div}\;v\; d\Omega = 0\;\;\forall\, q\,\in\, L^2(\Omega)\} = \{v\,\in\, V_{D0}\mid \mathrm{div}\;v\,=\,0\} \tag{5.69}$$

and we observe that $a(\cdot,\cdot)$ is strongly coercive on $V = V_{D0}$ as a consequence of Korn's inequality. The mixed formulation is well-posed because

$$\inf_{q\,\in\, L^2(\Omega)}\;\sup_{v\,\in\, H^1_0(\Omega)}\;\frac{\int_\Omega q\;\mathrm{div}\;v\; d\Omega}{\|v\|_1\,\|q\|_0} > 0, \tag{5.70}$$

an inequality that was proved by Ladyzhenskaya. But notice that our abstract approximation results do not depend on stability estimates such as (5.70), which correspond to the **exact problem**. Only the **boundedness** of the exact problem and the **stability** (coercivity) of the discrete problem matters.

Turning now to the mixed Galerkin approximation, which reads just as (5.65)-(5.66) replacing all exact spaces by V_{hg}, V_{h0} and Q_h, the following comments are in order:

- Whichever Q_h, the mixed Galerkin formulation admits a unique solution u_h belonging to

$$Z_{hg} = \{v_h \in V_{hg} \mid \int_\Omega q_h \text{ div } v_h \, d\Omega = 0 \quad \forall \, q_h \in Q_h\} \tag{5.71}$$

  and satisfying

$$\|u - u_h\|_V \leq C \inf_{v_h \in Z_{hg}} \|u - v_h\|_V \ . \tag{5.72}$$

- If Q_h is too large the approximation ability of Z_{hg} may be much poorer than that of V_{hg}. This lack of approximability is known as "locking". It manifests as largely inaccurate u_h even for very fine meshes.

- If Q_h is "balanced" with V_{h0}, in the sense that

$$\inf_{q_h \in Q_h} \sup_{v_h \in V_{h0}} \frac{\int_\Omega q_h \text{ div } v_h \, d\Omega}{\|q_h\|_Q \, \|v_h\|_V} = \gamma_h > 0 \tag{5.73}$$

  then there exists a unique $p_h \in Q_h$ such that (u_h, p_h) satisfies the mixed Galerkin formulation and

$$\|u - u_h\|_V + \|p - p_h\|_Q \leq \frac{c}{\gamma_h^2} \left(\inf_{v_h \in V_{hg}} \|u - v_h\|_V + \inf_{q_h \in Q_h} \|p - q_h\|_Q \right). \tag{5.74}$$

  with c independent of h for h small.

- If $\gamma_h = 0$, then p_h is not uniquely defined. This implies in particular that the system matrix of the mixed Galerkin formulation is singular.

- Though condition (5.73) is cumbersome to satisfy and check, there exists a vast collection of combinations $V_h - Q_h$ for which (5.73) holds uniformly in h (i.e., with $\gamma_h \geq \gamma_0 > 0$ for all h). These combinations are called **stable mixed elements**. Equal-order elements are not stable.

5.6 Stabilization: The case of the Dirichlet problem and Nitsche's method

The inf-sup condition on the discrete spaces V_h and Q_h, that is required for the mixed Galerkin formulation to be well-posed and optimally accurate, can be circumvented by adopting *stabilized mixed formulations*. If one looks carefully at the hypotheses of the abstract approximation result, Thm. 5.3, one perceives that:

- The exact solution (u, λ) must satisfy $B((u,\lambda),(v_h,\zeta_h)) = S(v_h,\zeta_h)$ for all $(v_h,\zeta_h) \in V_h \times Q_h$. Notice that it is not needed that B be the bilinear form of the exact problem. The following *stabilized* bilinear form, which depends on the mesh, is also acceptable:

$$B((u,\lambda),(v,\zeta)) = \int_\Omega K\,\nabla u\cdot\nabla v\;d\Omega - \langle\lambda, v\rangle - \langle\zeta, u\rangle - \delta\,h\int_{\partial\Omega}(\lambda - K\,\partial_n u)\,(\zeta - K\,\partial_n v)\;d\Gamma \quad (5.75)$$

  The modification from the Galerkin bilinear form is the addition of the last term, which is zero if (u, λ) is the exact solution because, as we have shown, $\lambda = K\partial_n u$.

- The bilinear form must be continuous, but in what space and equipped with what norm? If one looks carefully at the proof one observes that **(a)** B needs to be bounded on the space $(u,\lambda) + (V_h \times Q_h)$, of linear combinations of the exact solution with functions in the discrete space; and **(b)** the norms $\|\cdot\|_V$ and $\|\cdot\|_Q$ need not be the same as those of the exact problem. One shows that the stabilized bilinear form is continuous on $H^1(\Omega)\times L^2(\partial\Omega)$ with the *mesh-dependent* norms

$$\|v\|_V \stackrel{\text{def}}{=} \|v\|_1 + h^{-1/2}\,\|v\|_{L^2(\partial\Omega)} \quad (5.76)$$

$$\|\zeta\|_Q \stackrel{\text{def}}{=} h^{1/2}\,\|\zeta\|_{L^2(\partial\Omega)} \quad (5.77)$$

  Notice that $H^1(\Omega)\times L^2(\partial\Omega)$ contains $(u,\lambda)+(V_h\times Q_h)$ under the regularity assumption $\lambda \in L^2(\partial\Omega)$ and under the restriction $Q_h \in L^2(\partial\Omega)$, which is not really restrictive.

Exo. 5.10 *Read Stenberg's article for next class. Sketch the different steps in the proof and relate them to the theory presented in the course.*

Answer:

1. Let (u, λ) be the exact solution. By exact solution we understand the only elements of $H^1_{Dg}(\Omega)$ and $H^{-1/2}(\partial\Omega)$ satisfying
$$-\nabla \cdot (K \nabla u) = f \qquad \text{and} \qquad \lambda = K\, \partial_n u \ . \tag{5.78}$$
We assume (u, λ) to belong to $H^2(\Omega) \times L^2(\partial\Omega)$.

2. Given a specific mesh $\mathcal{T}_h$, let $V_h \subset H^1(\Omega)$ be a (piecewise polynomial, conforming) finite element space, and let Q_h be an arbitrary closed subspace of $H^{-1/2}(\partial\Omega)$.

3. The variational formulation is set up on the space
$$W = (u, \lambda) + (V_h \times Q_h) = [u + V_h] \times [\lambda + Q_h] = V \times Q \tag{5.79}$$
equiped with the norm
$$\|(v, \zeta)\|_W = \|v\|_V + \|\zeta\|_Q \tag{5.80}$$
with the definitions given in (5.76)-(5.77).

4. The finite element approximation $(u_h, \lambda_h) \in W_h = V_h \times Q_h$ is defined by
$$B((u_h, \lambda_h), (v_h, \zeta_h)) = S(v_h, \zeta_h) \qquad \forall\, (v_h, \zeta_h) \in W_h \tag{5.81}$$

with $B(\cdot,\cdot)$ given in (5.75) or, equivalently,

$$\int_\Omega K\,\nabla u_h\cdot\nabla v_h\;d\Omega-\int_{\partial\Omega}\lambda_h\;v_h\;d\Gamma+\delta\,h\int_{\partial\Omega}(\lambda_h-K\partial_n u_h)\,K\partial_n v_h\;d\Gamma \;=\; \int_\Omega f\;v_h\;d\Omega \qquad \forall\, v_h\in V_h \tag{5.82}$$

$$-\int_{\partial\Omega}\zeta_h\;u_h\;d\Gamma-\delta\,h\int_{\partial\Omega}(\lambda_h-K\partial_n u_h)\,\zeta_h\;d\Gamma \;=\; -\int_{\partial\Omega}\zeta_h\;g\;d\Gamma \qquad \forall\,\zeta_h\in Q_h. \tag{5.83}$$

5. It is verified that (u,λ) satisfy (5.82)-(5.83) for all v_h and ζ_h, so that

$$B((u-u_h,\lambda-\lambda_h),(v_h,\zeta_h)=0 \qquad \forall\,(v_h,\zeta_h)\in V_h\times Q_h\;. \tag{5.84}$$

6. It is verified that B is continuous in the W-norm. For example,

$$\int_{\partial\Omega}\xi\,\zeta\;d\Gamma\le\|\xi\|_{L^2(\partial\Omega)}\|\zeta\|_{L^2(\partial\Omega)}\le h^{1/2}\|\xi\|_{L^2(\partial\Omega)}h^{-1/2}\|\zeta\|_{L^2(\partial\Omega)}\le\|\xi\|_Q\|\zeta\|_V\;.$$

The normal derivative trace operator $\partial_n\cdot$ appears in these calculations, which is continuous from $H^1(\Omega)$ to $H^{-1/2}(\partial\Omega)$.

Exo. 5.11 *Complete the proof of the continuity of B.*

7. It remains to prove the weak coercivity of B in W_h and to find out what is the allowed range for the constant δ. For this we refer to the article. The constant δ ends up having to satisfy

$$0<\delta<C_I$$

where C_I is the constant of the inverse estimate. Hypotheses about the shape quality and quasi-uniformity of the mesh are thus necessary. One arrives at a coercivity constant that is independent of h.

8. As a consequence, *the discrete problem is well-posed for all h, and for any choice of Q_h*!

9. Also, one obtains optimal convergence for u in $H^1(\Omega)$:

$$\|u-u_h\|_1+h^{-1/2}\|u-u_h\|_{L^2(\partial\Omega)}+h^{1/2}\|K\partial_n u-\lambda_h\|_{L^2(\partial\Omega)} \leq\ c\ \left(h^k\,\|u\|_{k+1}+h^{1/2}\,\|\partial_n u-\Pi_h\partial_n u\|_{L^2(\partial\Omega)}\right)\ . \tag{5.85}$$

where Π_h is the L^2 projection onto Q_h.

Exo. 5.12 *Is the convergence of optimal order for $\|u-u_h\|_{L^2(\partial\Omega)}$? And for $\|K\partial_n u-\lambda_h\|_{L^2(\Omega)}$? If both V_h and Q_h consist of piecewise polynomials of degree r and s, respectively, what is a sensible choice for the difference $r-s$?*

10. In particular, if Q_h is rich enough to make the second term on the right-hand side of (5.85) negligible, the error $\|u-u_h\|_1$ is $\mathcal{O}(h^k)$, while $\|u-u_h\|_{L^2(\partial\Omega)}$ is $\mathcal{O}(h^{k+1/2})$.

11. The Lagrange multiplier λ_h converges to $K\partial_n u$, and $\|K\partial_n u-\lambda_h\|_{L^2(\partial\Omega)}$ is $\mathcal{O}(h^{k-1/2})$.

12. **Nitsche's method** corresponds simply to taking $Q_h=L^2(\partial\Omega)$, so that (5.83) can be solved explicitly, yielding

$$\lambda_h = K\,\partial_n u_h-(\delta\,h)^{-1}\,(u_h-g)\ . \tag{5.86}$$

Substituting into (5.82) one arrives at a formulation in the sole variable u_h (the primal variable) that reads

$$\begin{aligned}\int_\Omega K\,\nabla u_h\cdot\nabla v_h\ d\Omega-\int_{\partial\Omega}K\partial_n u_h\ v_h\ d\Gamma-\int_{\partial\Omega}K u_h\ \partial_n v_h\ d\Gamma+(\delta\,h)^{-1}\int_{\partial\Omega}u_h\,v_h\ d\Gamma=\\ =\ \int_\Omega f\,v_h\ d\Omega-\int_{\partial\Omega}K\,g\,\partial_n v_h\ d\Gamma+(\delta\,h)^{-1}\int_{\partial\Omega}g\,v_h\ d\Gamma\ .\end{aligned} \tag{5.87}$$

Contrary to the primal discrete formulation, Nitsche's method provides u_h with provable convergence of $\partial_n u_h$ to $\partial_n u$.

13. The primal formulation will in general approximate the boundary condition with an error $\mathcal{O}(h^{k+1})$ because of the need to interpolate g with traces of functions of V_h. Notice that Nitsche's method is 1/2-short of achieving optimal order.

Exo. 5.13 *Compute the elementary matrix and elementary right-hand side of the element* $(0, h)$ *corresponding to Nitsche's method, assuming that the boundary condition at* $x = 0$ *is* $u(0) = a$.

6 Galerkin method for parabolic problems

6.1 Differential and variational formulations

Consider the transient version of the convection-diffusion-reaction problem of Section 3.

Differential Formulation: Find $u : \Omega\times]0,T[$ satisfying $u(x,0) = u_0(x)$ for $x \in \Omega$ and, for $0 < t < T$,

$$\begin{aligned} \partial_t u - \operatorname{div}(K\nabla u) + \beta\cdot\nabla u + \sigma u &= f \quad \text{in } \Omega && (6.1)\\ u &= g \quad \text{on } \Gamma_D && (6.2)\\ (K\nabla u)\cdot\mathbf{n} &= H \quad \text{on } \Gamma_N && (6.3) \end{aligned}$$

where now all coefficients are continuous functions of time, and Γ_D and Γ_N are disjoint parts of $\partial\Omega$ that do not vary with time, and $\overline{\Gamma_D \cup \Gamma_N} = \partial\Omega$.
To write the problem in weak form, we assume that u satisfies the **DF** so that for $0 < t < T$ if we multiply by $v \in V_0 = H^1_{D0}(\Omega)$, integrate over Ω and apply the integration by parts formula we arrive at

$$\int_\Omega \partial_t u\, v\, d\Omega + \int_\Omega (\nabla v\cdot(K\nabla u) + v\,\beta\cdot\nabla u + \sigma u\, v)\, d\Omega = \int_\Omega f\, v\, d\Omega + \int_{\Gamma_N} H\, v\, d\Gamma \qquad (6.4)$$

which u satisfies for all $v \in V_0$. Denoting by $(\cdot,\cdot)$ the scalar product of $L^2(\Omega)$, by $\|\cdot\|$ its norm, and using the definitions of $a(t;\cdot,\cdot)$ and $\ell(t;\cdot)$ of Section 3, the previous equation can be rewritten as

$$(\partial_t u(t), v) + a(t; u(t), v) = \ell(t; v) \qquad \forall\, v \in V_0. \qquad (6.5)$$

For simplicity, we take $g = 0$ and denote $V_g = V_0 = V$. Knowing that $a(t;u,\cdot)$ and $\ell(t;\cdot)$ are continuous linear forms on V, the same must hold for $\partial_t u$. In other words,

$$\partial_t u(\cdot,t) \in V', \quad 0 < t < T. \qquad (6.6)$$

The weak or variational formulation of the problem considered is:

Variational Formulation: Find $u(x,t)$, $0 < t < T$, such that $u \in L^2(0,T;V)$, $\partial_t u \in L^2(0,T;V')$, and

$$(\partial_t u, v) + a(t;u,v) = \ell(t;v) \qquad \forall\, v \in V \tag{6.7}$$

$$u(x,0) = u_0(x)\ . \tag{6.8}$$

We will assume that there exists a constant $\alpha > 0$ such that

$$a(t;v,v) \geq \alpha \|v\|_V^2\ , \qquad \forall\, v \in V,\ \forall\, t \in [0,T]. \tag{6.9}$$

Remark 6.1 *If $\sigma = 0$ and there is no Dirichlet boundary, then the bilinear form is not strongly coercive on $V = H^1(\Omega)$. However, one may change variable to $w = e^{-\gamma t}u$ for some $\gamma > 0$. Clearly,*

$$\partial_t u = \gamma\, w \,+\, e^{\gamma t}\, \partial_t w, \qquad a(t;u,v) = e^{\gamma t}\, a(t;w,v)$$

so that canceling out the factor $e^{\gamma t}$ we have that w satisfies

$$(\partial_t w, v) + a(t;w,v) + \gamma\,(w,v) = e^{-\gamma t}\,\ell(t;v) \qquad \forall\, v \in V\ . \tag{6.10}$$

The bilinear form for w has additional coercivity on $L^2(\Omega)$, making the Poincaré-Friedrichs inequality unnecessary.

Prop. 6.2 (Uniqueness) *The solution u of the* **VF***, if it exists, is unique.*

Proof. Assume that there are two solutions u and $\tilde{u}$. Then $w = u - \tilde{u}$ satisfies $w(x,0) = 0$ and

$$(\partial_t w, v) + a(t; w, v) = 0 \qquad \forall\, v \in V \ .$$

Choosing $v = w$ at all instants and integrating from 0 to t one gets

$$\frac{1}{2}\|w(\cdot,t)\|^2 + \int_0^t a(s; w, w)\ ds = 0 \tag{6.11}$$

and since each term is non-negative, both must be zero. Notice that we have used

$$(\partial_t u, v) = \frac{d}{dt}(u, v) \ .$$

□

Because any solution of the **VF** is also a solution of the **DF**, we have also proved uniqueness of the differential formulation.

Theorem 6.3 *(**Existence**) Assume that $u_0 \in L^2(\Omega)$, that $\ell(t;\cdot)$ belongs to $L^2(0,T;V')$ and that $a(t;\cdot,\cdot)$ is continuous and strongly coercive on $V \times V$, uniformly in t. Then there exists a solution to the* **VF**.

The proof of existence is longer than that of uniqueness. We refer to Dautray-Lions (Volume 5, page 513) for a detailed presentation. The idea is to use the Galerkin method, which has an interest by itself, to build a sequence of solutions that is then shown to have an accumulation point.

Prop. 6.4 *(**Continuity**, **Stability**) Let (u_0,ℓ) and $(\tilde{u}_0,\tilde{\ell}) \in L^2(\Omega)\times L^2(0,T;V')$. Let u and $\tilde{u}$ be the corresponding solutions of* **VF**. *Then*

$$\|u-\tilde{u}\|_{L^\infty(0,T;L^2(\Omega))} \leq \left[\|u_0-\tilde{u}_0\|^2 + \frac{1}{\alpha}\|\ell-\tilde{\ell}\|^2_{L^2(0,T;V')}\right]^{\frac{1}{2}} \tag{6.12}$$

$$\|u-\tilde{u}\|_{L^2(0,T;V)} \leq \frac{1}{\sqrt{\alpha}}\left[\|u_0-\tilde{u}_0\|^2 + \frac{1}{\alpha}\|\ell-\tilde{\ell}\|^2_{L^2(0,T;V')}\right]^{\frac{1}{2}} \tag{6.13}$$

Exo. 6.1 *Prove the previous proposition.* Hints: The function $w = u - \tilde{u}$ solves problem **VF** with $w(0) = u_0 - \tilde{u}_0$ and right-hand side $g = \ell - \tilde{\ell}$. Choosing $v = w$ and integrating in time as before one arrives at

$$\frac{1}{2}\|w(\cdot,t)\|^2 + \int_0^t a(s;w,w)\ ds = \frac{1}{2}\|w(\cdot,0)\|^2 + \int_0^t g(s;w(\cdot,s))\ ds\ .$$

Now using the coercivity and

$$g(s;w(\cdot,s)) \le \|g(s;\cdot)\|_{V'}\|w(\cdot,s)\|_V \le \frac{1}{2\alpha}\|g(s;\cdot)\|_{V'}^2 + \frac{\alpha}{2}\|w(\cdot,s)\|_V^2$$

one gets

$$\frac{1}{2}\|w(\cdot,t)\|^2 + \frac{\alpha}{2}\|w\|_{L^2(0,t;V)}^2 \le \frac{1}{2}\|w(\cdot,0)\|^2 + \frac{1}{2\alpha}\|g(s;\cdot)\|_{L^2(0,t;V')}^2$$

and the result comes.

6.2 Galerkin approximation

Taking a finite-dimensional subspace $V_h \subset V$, the Galerkin approximation of **VF** reads:

Space-Discretized Variational Formulation: Find $u_h(t)$ (or $u_h(\cdot,t)$) belonging to V_h for $0 < t < T$, such that

$$(\partial_t u_h, v_h) + a(t; u_h, v_h) = \ell(t; v_h) \qquad \forall\, v_h \in V_h,\ \forall\, t \in]0,T[\tag{6.14}$$

$$u_h(x,0) = u_{0h}(x)\ , \tag{6.15}$$

where u_{0h} is some approximation of u_0.

Denoting by $\{\mathcal{N}^j\}$ a basis of V_h, it is clear that for $u_h(t)$ to belong to V_h at all times it must be of the form

$$u_h(x,t) = \sum_j U^j(t)\, \mathcal{N}^j(x)\ . \tag{6.16}$$

It is also clear that (6.14) holds if and only if

$$(\partial_t u_h, \mathcal{N}^i) + a(t; u_h, \mathcal{N}^i) = \ell(t; \mathcal{N}^i) \qquad \forall\, i = 1, \cdot, M,$$

so that the **SDVF** is equivalent to

$$\underline{\underline{M}}\ \underline{U}'(t)\ +\ \underline{\underline{A}}(t)\ \underline{U}(t)\ =\ \underline{L}(t) \tag{6.17}$$

Exo. 6.2 *Use the previous equation to show that* **the space-discretized problem admits a unique solution** *under suitable hypotheses on the data (i.e., on the bilinear form a and the linear form ℓ).*

Exo. 6.3 *Verify that, if u_{0h} is taken as the $L^2(\Omega)$-projection of u_0 onto V_h, then the semi-discrete Galerkin solution u_h satisfies the uniform (independent of h) bounds*

$$\|u_h\|_{L^\infty(0,T;L^2(\Omega))} \leq \left[\|u_0\|^2 + \frac{1}{\alpha}\|\ell\|^2_{L^2(0,T;V')}\right]^{\frac{1}{2}} \tag{6.18}$$

$$\|u_h\|_{L^2(0,T;V)} \leq \frac{1}{\sqrt{\alpha}}\left[\|u_{0h}\|^2 + \frac{1}{\alpha}\|\ell\|^2_{L^2(0,T;V')}\right]^{\frac{1}{2}} . \tag{6.19}$$

These bounds eventually allow to prove that if a sequence of spaces V_h of growing dimension produce a sequence u_h of semi-discrete solutions, then these solutions converge to some $u \in L^2(0,T;V)$ that satisfies the **VF**. *This argument proves thus existence.*

Theorem 6.5 *(***Convergence of the Galerkin approximation***) If the space V_h has an* interpolation operator $\mathcal{I}_h : H^r(\Omega) \cap V \to V_h$ *($r \geq 2$) such that*

$$\|v - \mathcal{I}_h v\| + h\|\nabla(v - \mathcal{I}_h v)\| \leq C\, h^s \|v\|_{H^s(\Omega)} \qquad 1 \leq s \leq r \tag{6.20}$$

and if the norms on the right-hand side of the inequality are finite, then

$$\|u_h(t) - u(t)\| \leq \|u_{0h} - u_0\| + C\, h^r \left(\|u_0\|_r + \int_0^t \|\partial_t u(s)\|_r \, ds\right. \tag{6.21}$$

The proof of this theorem uses the *elliptic projection.*

Def. 6.6 *Given an arbitrary function $w \in V$, its elliptic projection onto V_h at time t, denoted by $P_{ht}w$, is the unique solution of*

$$a(t; P_{ht}w, v_h) = a(t; w, v_h) \qquad \forall\, v_h \in V_h \tag{6.22}$$

Exo. 6.4 *Prove that $P_{ht} : V \to V_h$ is indeed a projection, and that*

$$\|P_{ht}w - w\| + h\,\|\nabla(P_{ht}w - w)\| \le C\,h^s\,\|w\|_s \qquad 1 \le s \le r\ . \tag{6.23}$$

Of course (6.20) is assumed to hold.

Going back to the stategy for proving convergence, it begins by splitting the error as

$$e(t) \doteq u_h(t) - u(t) = \theta(t) + \rho(t) \qquad \text{where } \ \theta = u_h - P_{ht}u, \ \ \rho = P_{ht}u - u\ . \tag{6.24}$$

Clearly the term ρ can be shown to tend to zero with h using (6.23). The difficulty lies in bounding θ.

Exo. 6.5 *Prove that θ satisfies*

$$(\partial_t\theta, v_h) + a(t; \theta, v_h) = -(\partial_t\rho, v_h) \qquad \forall\, v_h \in V_h, \ \ 0 < t < T. \tag{6.25}$$

From this, taking $v_h = \theta \in V_h$, show that $d\|\theta\|/dt \le \|\partial_t\rho\|$, from which the error estimate (6.21) follows. The details can be found in Johnson, page 151, in Thomée, page 9, among others.

6.3 Fully discrete approximation

The SDVF is nothing but a system of ODEs, and as such can be discretized in time by various methods. In general, the time discretization will bring an error of order $O(\delta t^k)$, where k is the formal order of the method. It can also bring *stability* issues. Methods can be *unconditionally stable*, *conditionally stable*, or *unconditionally unstable*. Conditionally stable methods have restrictions on the time step of the form

$$\delta t \leq c\, h^m \ , \tag{6.26}$$

which can be very expensive if $m > 1$.
Let us finish this section with an example of an unconditionally stable method and its analysis.

Def. 6.7 *The* **Backward Euler** *or* **Fully Implicit** *approximation of* **SDVF** *consists of finding* $u_h^n \in V_h$, $n \geq 1$, *such that*

$$\frac{1}{\delta t}(u_h^n - u_h^{n-1}, v_h) + a(t_n; u_h^n, v_h) = \ell(t_n; v_h) \qquad \forall\, v_h \in V_h \tag{6.27}$$

with $u_h^0 = u_{0h}$. *In matrix notation,*

$$\left(\underline{\underline{M}} + \delta t\ \underline{\underline{A}}(t_n)\right)\ \underline{U}^n = \underline{\underline{M}}\ \underline{U}^{n-1} + \delta t\ \underline{L}(t_n) \tag{6.28}$$

Theorem 6.8 *(**Convergence of the Backward Euler Galerkin method**)*

$$\|u_h^n - u(t_n)\| \leq C\, h^r \left(\|u_0\|_r + \int_0^{t_n} \|\partial_t u(s)\|_r \, ds \right) + \delta t \int_0^{t_n} \|\partial_{tt}^2 u(s)\| \, ds \tag{6.29}$$

$$\|\nabla(u_h^n - u(t_n))\| \leq c(u) \left(h^{r-1} + \delta t \right) \tag{6.30}$$

Notice that if one tries to obtain (6.30) directly from (6.29) and an inverse estimate, one arrives at the less sharp bound $\|\nabla(u_h^n - u(t_n))\| \leq c(u)\ (h^{r-1} + h^{-1}\,\delta t)$. The proof of this theorem can be found in Thomée, pag. 15.

Thank you for your attention in class

and your dedication throughout the course.

Happy holidays!!

References

[1] R. Adams. Sobolev spaces. Academic Press. 1975.

[2] S. Brenner and L. R. Scott, The Mathematical Theory of Finite Element Methods. Springer-Verlag, 1994.

[3] H. Brezis. Analyse fonctionnelle. Théorie et applications. Masson. 1983.

[4] F. Brezzi and M. Fortin, Mixed and Hybrid Finite Element Methods. Springer-Verlag, 1991.

[5] P. Ciarlet. Basic error estimates for elliptic problems. Handbook of Numerical Analysis, Vol. II. Finite Element Methods (Part 1). Edited by P. Ciarlet and J.L. Lions. Elsevier. 1991.

[6] R. Dautray and J.-L. Lions. Mathematical analysis and numerical methods for Science and Technology. Springer-Verlag. 1992.

[7] R. Durán. Galerkin approximations and finite element methods. Lecture notes (available at the author's website).

[8] A. Ern and J.-L. Guermond. Theory and practice of finite elements. Applied Mathematical Sciences 159. Springer. 2004.

[9] D. Gilbarg and N. Trudinger. Elliptic partial differential equations of second order. Grundlehren der mathematischen Wissenschaften 224. Second edition. Springer-Verlag. 1983.

[10] O. Ladyzenskaja and N. Uralceva, Equations aux dérivées partielles de type elliptique. Dunod, Paris, 1968.

[11] M. Renardy and R. Rogers. An introduction to partial differential equations. Texts in Applied Mathematics 13. Springer. 1993.

[12] V. Thomée. Galerkin finite element methods for parabolic problems. Springer. 2006.