Partial Differential Equations and the Finite Element Method

Frédéric LEGOLL

January 2019

Contents

General introduction

Many phenomena in the engineering sciences (in solid and fluid mechanics, in finance, in traffic flow, ...) can be modelled using Partial Differential Equations (PDEs).

The aim of this course is to introduce the students to some modern mathematical tools to address these types of equations. In these lecture notes, we first focus on the Poisson equation in a bounded domain, which reads

$$\left\{\begin{array}{l} \text{Find a function } u \,:\, \Omega \longrightarrow \mathbb{R} \text{ such that} \\ -\Delta u = f \quad \text{in } \Omega \\ u = 0 \quad \text{on } \partial\Omega \end{array}\right. \tag{1}$$

where Ω is an open bounded subset of $\mathbb{R}^d$, $\partial\Omega$ is its boundary and f is a given function, defined on Ω and valued in $\mathbb{R}$. Problem (1) arises for instance in mechanics: the solution u to (1) is the vertical displacement of an elastic membrane submitted to the load f, and clamped at its boundary.

Along this course, we will also consider several variants of (1), in different directions:

- non-symmetric problems;
- boundary conditions different from homogeneous Dirichlet boundary conditions.

We will also introduce the students to non-coercive problems (through the example of the Stokes problem) and to the inf-sup theory.

This course is devoted to the *mathematical analysis* and the *numerical analysis* of Problem (1) and its variants.

From the *mathematical analysis* standpoint, the questions we address are the following:

- Does the problem (1) have a solution? Is it unique?
- If this is the case, is the map $f \mapsto u$ continuous?

Problem (1) is a *boundary value problem* in the sense that it is composed of a Partial Differential Equation (PDE) and a boundary condition ($u = 0$ on $\partial\Omega$). Its mathematical analysis allows to emphasize a generic trend in modern analysis: the use of *Geometry* tools in infinite dimensional vector spaces allows to obtain results for *Analysis* problems such as (1) in a simple and elegant manner.

In the second part of these lecture notes, we turn to *numerical analysis* questions. Indeed, in practice, the exact solution to (1) (just as for any PDE) is not known in closed form. One often resorts to numerical approaches (such as the Finite Element Method) to compute an approximation u_h of the exact solution u. In this context, we describe general Galerkin approximations of (1), that yield finite dimensional problems, and address the following questions:

- Is the finite dimensional problem well-posed?

- Denoting u_h its solution, how the error $\|u - u_h\|$ can be estimated?
- In practice, how is such an approximation implemented on a computer? For that latter question, we will focus on the Finite Element Method, an extremely popular technique in the engineering community.

The overall objective of this course is to give to the students the tools needed for the mathematical analysis of PDEs, and for the detailed understanding of the Finite Element Method: estimation of the error as a function of the discretization parameters, practical implementation.

These lecture notes are organized as follows. The first part is devoted to the construction of the appropriate functional spaces in which the solution to (1) should be looked for. We start by an introduction to the distribution theory (in Chapter 1), followed by the construction of the Sobolev spaces (in Chapter 2), which are the relevant spaces of functions when dealing with PDEs.

The mathematical analysis of (1) and its variants (in the coercive case) is carried out in Chapter 3. To that aim, we introduce the notion of *variational formulation* of a boundary value problem, the analysis of which is performed using the Lax-Milgram theorem. The analysis of non-coercive problems is discussed in Chapter 4.

The numerical analysis of (1) and its variants is carried out in Chapter 5 in the coercive case, and in Chapter 6 in the non-coercive case.

This course is based on notions introduced in two ENPC courses, *Outils Mathématiques pour l'Ingénieur* and *Analyse et Calcul Scientifique.* However, the present lecture notes are, as much as possible, self-contained. Nevertheless, the reader is supposed to be familiar with Banach and Hilbert spaces. Some basic facts concerning the Lebesgue integral and L^p spaces have been collected in Appendix A. For historical reasons, these lectures notes are written in English. A glossary of the main mathematical terms can be found in Appendix B.

I am indebted to Eric Cancès and Alexandre Ern, who have been teaching an *Analyse* course at ENPC for many years. The structure of this document is very much inspired from their lectures notes [2, 5] and from [3].

I am grateful to the colleagues who gave me opportunities to teach, at ENPC and elsewhere. It has always been a wonderful experience, and I am very pleased to thank them. I would also like to thank Thomas Hudson and Antoine Levitt for their remarks on a draft version of these lecture notes.

As always, there are certainly still several typos in these notes, despite our best efforts. Any suggestions on these notes are therefore warmly welcome.

Frédéric Legoll, Champs sur Marne, January 2019

Chapter 1

Introduction to distribution theory

The distribution theory, introduced by Laurent Schwartz in 1946, is a fundamental piece of Analysis. Distributions generalize the notion of functions, and satisfy the following two striking properties, that make them particularly easy to manipulate:

- any distribution is differentiable and its derivative is again a distribution;
- if a sequence of distributions T_n converges to some distribution T, then the derivatives of T_n converge to the derivatives of T.

In our context, distributions will be useful to build the appropriate functional spaces for the study of Partial Differential Equations (see Chapter 2).

In what follows, Ω is an open subset (which may be either bounded or unbounded) of $\mathbb{R}^d$, with $d \geq 1$. Following the international convention, we denote by (a, b) the one-dimensional open interval $\{x \in \mathbb{R},\ a < x < b\}$.

1.1 Definitions

Roughly speaking, a distribution is a linear and continuous form on the vector space $\mathcal{D}(\Omega)$ of the functions which are in $C^\infty(\Omega)$ and which have a compact support in Ω. We start by recalling the definition of the support of a continuous function.

Definition 1.1. *Let Ω be an open subset of $\mathbb{R}^d$ and $\phi : \Omega \longrightarrow \mathbb{R}$ be a continuous function. The* support *of ϕ is*

$$Supp(\phi) = \overline{\{x \in \Omega,\ \phi(x) \neq 0\}}^{\Omega}.$$

We recall that the notation $\overline{A}^{\Omega}$ means "closure of the set A in Ω". The set $\overline{A}^{\Omega}$ is hence the set of points in Ω that are limits of sequences of points of A. Here are some examples of support computations:

1. $\Omega = \mathbb{R}$ and $\phi(x) = \sin x$: then

$$\{x \in \mathbb{R},\ \phi(x) \neq 0\} = \mathbb{R} \setminus \pi\mathbb{Z}, \qquad \text{Supp}(\phi) = \mathbb{R}.$$

2. $\Omega = (-1, 2)$ and $\phi(x) = \begin{cases} x+1 & \text{if } x \in (-1, 0], \\ 1-x & \text{if } x \in [0, 1], \\ 0 & \text{otherwise.} \end{cases}$

   Then $\{x \in \Omega,\ \phi(x) \neq 0\} = (-1, 1)$ and $\text{Supp}(\phi) = (-1, 1]$.

3. $\Omega = (-2,2)$ and ϕ is the same function as above, that is $\phi(x) = \begin{cases} x+1 & \text{if } x \in (-1,0], \\ 1-x & \text{if } x \in [0,1], \\ 0 & \text{otherwise.} \end{cases}$

   Then $\{x \in \Omega,\ \phi(x) \neq 0\} = (-1,1)$ and $\mathrm{Supp}(\phi) = [-1,1]$.

The support of ϕ is compact only in the last example (we recall that the compact sets of $\mathbb{R}^d$ are the sets which are bounded and closed).

Exercise 1.2. *Let f and g be continuous functions from $\mathbb{R}^d$ to $\mathbb{R}$.*

- *Show that $Supp(fg) \subset Supp(f) \cap Supp(g)$ and that the inclusion may be strict.*
- *Show that $Supp(f+g) \subset Supp(f) \cup Supp(g)$ and that the inclusion may be strict.*

Definition 1.3. *Let $\mathcal{D}(\Omega)$ be the vector space of functions defined on Ω, which are infinitely differentiable (hence, $\mathcal{D}(\Omega) \subset C^\infty(\Omega)$), and which have a compact support. The functions in $\mathcal{D}(\Omega)$ are called* test functions*. For any compact set $K \subset \Omega$, we denote by $\mathcal{D}_K(\Omega)$ the set of test functions with support contained in K.*

We recall the following notation: for any $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d$, we set $|\alpha| = \sum_{i=1}^{d} \alpha_i$ and

$$\partial^\alpha \phi = \frac{\partial^{|\alpha|}\phi}{\partial_{x_1}^{\alpha_1} \dots \partial_{x_d}^{\alpha_d}} = \frac{\partial^{\alpha_1 + \dots + \alpha_d}\phi}{\partial_{x_1}^{\alpha_1} \dots \partial_{x_d}^{\alpha_d}}. \tag{1.1}$$

Definition 1.4. *A* distribution *T in the open set Ω is a linear form on $\mathcal{D}(\Omega)$ which satisfies the following "continuity property": for any compact set $K \subset \Omega$, there exists an integer p and a constant C such that*

$$\forall \phi \in \mathcal{D}_K(\Omega), \qquad |\langle T, \phi \rangle| \leq C \sup_{x \in K,\ |\alpha| \leq p} |\partial^\alpha \phi(x)|. \tag{1.2}$$

The vector space of distributions in Ω is denoted $\mathcal{D}'(\Omega)$.

Remark 1.5. *On a vector space E endowed with a norm $\|\cdot\|$, the continuity property of a linear form T simply reads*

$$\forall x \in E, \qquad |\langle T, x \rangle| \leq C\|x\|.$$

The more complex form of the "continuity property" (1.2) stems from the fact that the topology with which the vector space $\mathcal{D}(\Omega)$ should be endowed such that its dual $\mathcal{D}'(\Omega)$ has good properties is not *a topology arising from a norm.*

Definition 1.6. *When the integer p can be chosen independently of K, one says that the distribution T has a* finite order*. The smallest possible value of p is the* order *of T.*

Remark 1.7. *Heuristically, the larger the order of a distribution T is, the more singular T is.*

1.2 First examples of distributions

We now give some classical examples. First, to any function $f \in L^1_{\text{loc}}(\Omega)$, one can associate the distribution T_f (also denoted f) defined by

$$\forall \phi \in \mathcal{D}(\Omega), \qquad \langle T_f, \phi \rangle = \int_\Omega f\, \phi. \tag{1.3}$$

Next, for any $a \in \Omega \subset \mathbb{R}$, we denote by δ_a the distribution defined by

$$\forall \phi \in \mathcal{D}(\Omega), \qquad \langle \delta_a, \phi \rangle = \phi(a) \tag{1.4}$$

and by δ'_a the distribution defined by

$$\forall \phi \in \mathcal{D}(\Omega), \qquad \langle \delta'_a, \phi \rangle = -\phi'(a). \tag{1.5}$$

Last, consider a locally bounded measure μ, i.e. a measure μ defined on the set $\mathcal{B}(\Omega)$ of the borelian subsets of $\Omega \subset \mathbb{R}^d$ and such that $\mu(K) < \infty$ for any compact subset $K \subset \Omega$. We denote T_μ (or simply μ) the distribution defined by

$$\forall \phi \in \mathcal{D}(\Omega), \qquad \langle T_\mu, \phi \rangle = \int_\Omega \phi\, d\mu. \tag{1.6}$$

Exercise 1.8. *The aim of this exercise is to compute the order of some particular distributions.*

- *Show that the linear forms T_f defined by* (1.3)*, δ_a defined by* (1.4) *and T_μ defined by* (1.6) *are distributions of order 0 on Ω.*
- *Show that the linear form δ'_a defined by* (1.5) *is a distribution of order 1 on Ω (Hint: consider the simple case $\Omega = \mathbb{R}$, and consider the action of δ'_a on the sequence $\phi_j(x) = \phi_0(x)\, atan(jx)$, where $\phi_0 \in \mathcal{D}(\mathbb{R})$ is an even function that is equal to* 1 *in the neighborhood of* 0*; such a test function indeed exists in view of Lemma 1.18 below).*

Remark 1.9. *There does not exist any function $f \in L^1_{\text{loc}}(\mathbb{R})$ such that $T_f = \delta_0$. This shows that the space of distributions is (much) larger than the space of functions. The proof of this statement is not easy. However, it is easy to see that there exists no function $f \in C^0(\mathbb{R})$ such that $T_f = \delta_0$.*

1.3 Derivation in the sense of distributions

Let $f \in C^1(\mathbb{R})$. The integration by parts formula reads

$$\forall \phi \in \mathcal{D}(\mathbb{R}), \qquad \int_{\mathbb{R}} \frac{df}{dx}\, \phi = -\int_{\mathbb{R}} f\, \frac{d\phi}{dx}.$$

Likewise, for any $f \in C^1(\mathbb{R}^d)$, we have, for any $1 \le i \le d$,

$$\forall \phi \in \mathcal{D}(\mathbb{R}^d), \qquad \int_{\mathbb{R}^d} \frac{\partial f}{\partial x_i}\, \phi = -\int_{\mathbb{R}^d} f\, \frac{\partial \phi}{\partial x_i}.$$

The definition of the derivative of a distribution is inspired by this formula. Note that, in contrast to the case of functions, *all distributions are differentiable*!

Definition 1.10. *Let $T \in \mathcal{D}'(\Omega)$. The derivative of T with respect to the variable x_i, which is denoted by $\dfrac{\partial T}{\partial x_i}$, is defined by*

$$\forall \phi \in \mathcal{D}(\Omega), \qquad \langle \frac{\partial T}{\partial x_i}, \phi \rangle = -\langle T, \frac{\partial \phi}{\partial x_i} \rangle.$$

It is an easy exercise to check that the linear form $\dfrac{\partial T}{\partial x_i}$ defined above is indeed a distribution (i.e. that it satisfies the continuity property (1.2)).

Exercise 1.11. *Compute the first and second derivative (in the sense of distributions) of the "hat" function (see Figure 1.1) defined on $\mathbb{R}$ by*

$$f(x) = \begin{cases} x+1 & \text{if } \; -1 < x < 0, \\ 1-x & \text{if } \; 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

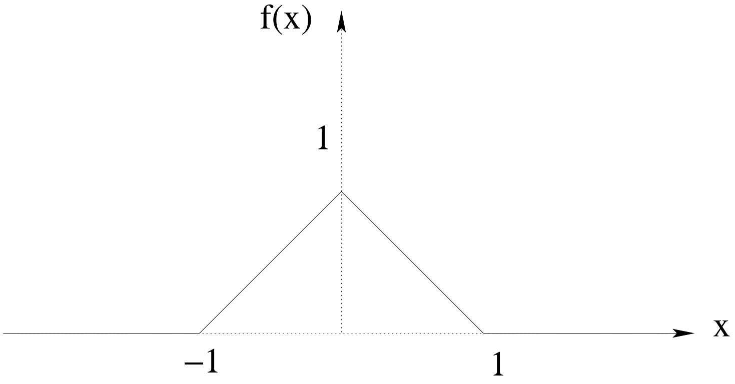

Figure 1.1: Hat function

Exercise 1.12. *Check that the distribution δ'_a defined by* (1.5) *is the derivative of δ_a in the sense of distributions.*

Exercise 1.13. *Consider the Heaviside function H defined on $\mathbb{R}$ by $H(x) = 1$ if $x > 0$, $H(x) = 0$ otherwise. Why can H be considered to be a distribution on $\mathbb{R}$? Compute its derivative in the sense of distributions.*

Remark 1.14. *Any function $f \in L^1_{\text{loc}}(\mathbb{R})$ can be considered to be a distribution. It is thus possible to define its derivative f' in the sense of distributions. In general, f' is* not *a function, but a distribution only.*

1.4 The space of test functions

The space of test functions plays an important role in the construction and the manipulation of distributions. For instance, to show that the distribution δ'_a is of order 1 (see the exercise 1.8), we use a test function $\phi_0 \in \mathcal{D}(\mathbb{R})$ that is even and is equal to 1 in the neighborhood of 0. This section collects some results ensuring the existence of such test functions. It is important to know that there exist test functions satisfying the properties discussed below. In contrast, the proof of these results are often technical. They will all be skipped here.

We start with the following important result:

Theorem 1.15. *The space $\mathcal{D}(\Omega)$ is dense in $L^p(\Omega)$ for any $1 \leq p < +\infty$.*

Remark 1.16. *Note that the case $p = \infty$ is not allowed in the above theorem. It turns out that $\mathcal{D}(\Omega)$ is not dense in $L^\infty(\Omega)$. Consider indeed the following example: on $\Omega = (0,1)$, let $f \in L^\infty(0,1)$ be defined by $f(x) = 1$ almost everywhere. Then, for any $\phi \in \mathcal{D}(0,1)$, we have $\|f - \phi\|_{L^\infty} \geq 1$.*

Corollary 1.17. *For any integer $k \geq 0$ and any $1 \leq p < \infty$, the space $C^k(\Omega) \cap L^p(\Omega)$ is dense in $L^p(\Omega)$.*

Proof. This is a direct consequence of Theorem 1.15 as $\mathcal{D}(\Omega) \subset C^k(\Omega) \cap L^p(\Omega) \subset L^p(\Omega)$. □

We now list several technical results.

Lemma 1.18. *Let Ω be an open subset of $\mathbb{R}^d$ and let $B_a(r)$ be the ball of center $a \in \Omega$ and of radius $r > 0$. If $B_a(r) \subset \Omega$, there exists $\phi \in \mathcal{D}(\Omega)$, with support in $B_a(r)$, and satisfying $\phi \geq 0$ and $\displaystyle\int_\Omega \phi = 1$.*

Lemma 1.19. *Let $K \subset \Omega$ be compact. There exists a test function $\phi \in \mathcal{D}(\Omega)$ such that $0 \leq \phi \leq 1$ on Ω and $\phi \equiv 1$ on K (see Figure 1.2).*

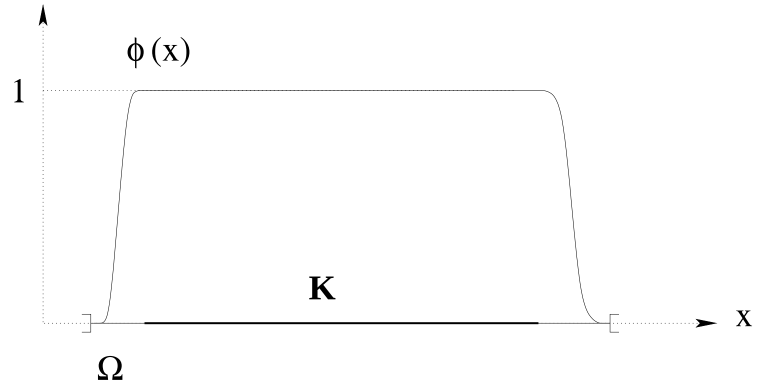

Figure 1.2: A specific test function

Lemma 1.20 (Partition of unity)**.** *Let $\Omega_1, \ldots, \Omega_n$ be open subsets of $\mathbb{R}^d$ and let K be a compact set of $\mathbb{R}^d$ such that*

$$K \subset \bigcup_{k=1}^{n} \Omega_k.$$

There exist test functions $\alpha_1, \ldots, \alpha_n$ such that

- *Supp$(\alpha_k) \subset \Omega_i$,*
- *$0 \leq \alpha_k \leq 1$,*
- *$\displaystyle\sum_{k=1}^{n} \alpha_k = 1$ on K.*

Lemma 1.21 (Borel lemma). *For any sequence $(a_\alpha)_{\alpha\in\mathbb{N}^d}$, there exists $\phi \in \mathcal{D}(\mathbb{R}^d)$ such that*

$$\forall \alpha \in \mathbb{N}^d, \qquad \partial^\alpha \phi(0) = a_\alpha.$$

Lemma 1.22 (Hadamard lemma). *Let $\phi \in \mathcal{D}(\mathbb{R})$ be such that $\phi^{(k)}(0) = 0$ for any integer $0 \leq k \leq n$. Then there exists $\psi \in \mathcal{D}(\mathbb{R})$ such that, for any $x \in \mathbb{R}$, $\phi(x) = x^n \psi(x)$.*

Proof. This follows by using the Taylor formula with a remainder in integral form. □

Remark 1.23. *A similar result holds in dimension higher than $d = 1$.*

1.5 Distributions and locally integrable functions

We have seen above that we can associate a distribution T_f to any function $f \in L^1_{\rm loc}(\Omega)$ (see formula (1.3)). It turns out that the space of test functions is sufficiently large as to ensure that the map $f \mapsto T_f$ from $L^1_{\rm loc}(\Omega)$ to $\mathcal{D}'(\Omega)$ is actually injective. This result, which is very important, is formalized in the theorem below.

Theorem 1.24. *Let f and g be functions in $L^1_{\rm loc}(\Omega)$. Then*

$$f = g \text{ almost everywhere} \quad \Leftrightarrow \quad T_f = T_g \text{ in } \mathcal{D}'(\Omega).$$

Proof. The implication $\Rightarrow$ is straightforward. We show the converse implication in the simplified case when $\Omega = \mathbb{R}^d$ and when f and g are in $L^1(\mathbb{R}^d)$.

Consider f and g in $L^1(\mathbb{R}^d)$ such that $T_f = T_g$ and set $h = f - g$. We hence have $h \in L^1(\mathbb{R}^d)$ such that

$$\forall \phi \in \mathcal{D}(\mathbb{R}^d), \qquad \int_\Omega h\,\phi = 0. \tag{1.7}$$

Our aim is to show that $h = 0$ almost everywhere. The proof falls in two steps.

Step 1 (approximation of identity): Consider a non-negative function $\chi \in \mathcal{D}(\mathbb{R}^d)$, with support in the unit ball and with integral equal to 1. Define the sequence $(\chi_n)_{n\in\mathbb{N}^\star}$ by

$$\chi_n(x) = n^{-d}\chi(x/n).$$

Such a sequence is an approximation of the identity (see Section 1.7). We recall that, for any n, $\|\chi_n\|_{L^1(\mathbb{R}^d)} = \|\chi\|_{L^1(\mathbb{R}^d)} = 1$. Let $\psi \in \mathcal{D}(\mathbb{R}^d)$. We claim that

$$\|\psi - \psi \star \chi_n\|_{L^1(\mathbb{R}^d)} \underset{n\to+\infty}{\longrightarrow} 0, \tag{1.8}$$

where $\psi \star \chi_n$ is the convolution product of ψ with χ_n (see Definition A.18). To prove (1.8), we first note, using the change of variables $z = y/n$, that

$$\psi \star \chi_n(x) = n^{-d} \int_{\mathbb{R}^d} \psi(x-y)\,\chi(y/n)\,dy = \int_{\mathbb{R}^d} \psi(x - z/n)\,\chi(z)\,dz.$$

Using the dominated convergence theorem (see Theorem A.1), we obtain that, for any $x \in \mathbb{R}^d$, $\psi \star \chi_n(x)$ converges to $\psi(x)$ when $n \to \infty$. We next note that

$$|\psi \star \chi_n(x)| = \left| \int_{\mathbb{R}^d} \psi(x-y)\,\chi_n(y)\,dy \right| \leq \|\psi\|_{L^\infty(\mathbb{R}^d)} \, \|\chi_n\|_{L^1(\mathbb{R}^d)} = \|\psi\|_{L^\infty(\mathbb{R}^d)},$$

and that

$$\text{Supp}(\psi \star \chi_n) \subset \text{Supp}(\psi) + \text{Supp}(\chi_n) \subset \text{Supp}(\psi) + \text{Supp}(\chi).$$

This implies that

$$\begin{cases} |\psi - \psi \star \chi_n| \leq 2 \, \|\psi\|_{L^\infty(\mathbb{R}^d)} \, 1_{\mathrm{Supp}(\psi)+\mathrm{Supp}(\chi)} & \text{for any } n, \\ \psi(x) - \psi \star \chi_n(x) \underset{n\to+\infty}{\longrightarrow} 0 \quad \text{for any } x \in \mathbb{R}^d. \end{cases}$$

We can hence use the dominated convergence theorem again, and obtain the convergence (1.8).

Step 2: Let $\epsilon > 0$ and $\psi \in \mathcal{D}(\mathbb{R}^d)$ such that

$$\|h - \psi\|_{L^1(\mathbb{R}^d)} \leq \epsilon/3. \tag{1.9}$$

The existence of such a function ψ stems from the density of $\mathcal{D}(\mathbb{R}^d)$ in $L^1(\mathbb{R}^d)$ (see Theorem 1.15). Let $(h_n)_{n\in\mathbb{N}^\star}$ be defined by

$$h_n(x) = (h \star \chi_n)(x) = \int_{\mathbb{R}^d} h(y) \, \chi_n(x-y) \, dy,$$

where χ_n was defined in Step 1. Since, for any $n \in \mathbb{N}^\star$, $\chi_n \in \mathcal{D}(\mathbb{R}^d)$, we observe that h_n vanishes on $\mathbb{R}^d$ for any n, in view of (1.7). We hence write

$$\|h\|_{L^1(\mathbb{R}^d)} = \|h - h \star \chi_n\|_{L^1(\mathbb{R}^d)} \leq \|h - \psi\|_{L^1(\mathbb{R}^d)} + \|\psi - \psi \star \chi_n\|_{L^1(\mathbb{R}^d)} + \|(h-\psi) \star \chi_n\|_{L^1(\mathbb{R}^d)} \, . \tag{1.10}$$

By definition of the convolution product (see the Definition A.18), we have

$$\|(h-\psi) \star \chi_n\|_{L^1(\mathbb{R}^d)} \leq \|h - \psi\|_{L^1(\mathbb{R}^d)} \, \|\chi_n\|_{L^1(\mathbb{R}^d)} = \|h - \psi\|_{L^1(\mathbb{R}^d)} \leq \epsilon/3. \tag{1.11}$$

Using (1.8), we know that we can choose N such that, for any $n \geq N$, we have

$$\|\psi - \psi \star \chi_n\|_{L^1(\mathbb{R}^d)} \leq \epsilon/3. \tag{1.12}$$

Collecting (1.10), (1.9), (1.12) and (1.11), we get that

$$\|h\|_{L^1(\mathbb{R}^d)} \leq \epsilon.$$

This upper bound holds for any $\epsilon > 0$. This implies that $h = 0$ in $L^1(\mathbb{R}^d)$. □

It is thus possible to identify any function of $L^1_{\mathrm{loc}}(\Omega)$ with its associated distribution, which can be denoted f instead of T_f. The previous result can be recast as follows.

Theorem 1.25. *Let f and g in $L^1_{\mathrm{loc}}(\Omega)$. Then*

$$f = g \text{ a.e.} \quad \Leftrightarrow \quad f = g \text{ in } \mathcal{D}'(\Omega).$$

Remark 1.26. *We therefore write $L^1_{\mathrm{loc}}(\Omega) \subset \mathcal{D}'(\Omega)$ just like we write $\mathbb{N} \subset \mathbb{R} \subset \mathbb{C}$. The notation $A \subset B$ here means that there exists a canonical injection from A to B.*

Remark 1.27. *The notion of distributions generalizes the notion of functions, but it does not mean that any function is a distribution. Only functions in L^1_{loc} are distributions. More singular functions are not necessarily distributions. For instance, the function $f(x) = 1/|x|$ is a distribution in $\mathbb{R}^d$ for any $d \geq 2$ but not in $\mathbb{R}$. See Section 1.10 for more details in that direction.*

1.6 Multiplication by C^∞ functions

Definition 1.28. *Let $T \in \mathcal{D}'(\Omega)$ and $g \in C^\infty(\Omega)$. The distribution $g\,T$ is defined by*

$$\forall \phi \in \mathcal{D}(\Omega), \qquad \langle g\,T, \phi\rangle = \langle T, g\,\phi\rangle.$$

It is an easy exercise to check that the linear form $g\,T$ is indeed a distribution.

Remark 1.29. *The multiplication of two distributions is* not defined*. Consider for instance the function $f(x) = \dfrac{1}{\sqrt{|x|}}$, which belongs to $L^1_{\mathrm{loc}}(\mathbb{R})$ and hence defines a distribution. But the function f^2 does not define a distribution (recall that $f^2(x) = 1/|x|$ does not belong to $L^1_{\mathrm{loc}}(\mathbb{R})$). More generally, the product $T_1\,T_2$ with $T_1 \in \mathcal{D}'(\Omega)$ and $T_2 \in \mathcal{D}'(\Omega)$ is not defined (see for instance the exercise 1.56).*

Exercise 1.30. *Show that $x\,\delta_0 = 0$.*

1.7 Convergence of distributions

Definition 1.31. *Let $(T_n)_{n\in\mathbb{N}}$ be a sequence of distributions in Ω. This sequence is said to* converge to some *$T \in \mathcal{D}'(\Omega)$ when $n \to \infty$ if*

$$\forall \phi \in \mathcal{D}(\Omega), \qquad \langle T_n, \phi\rangle \underset{n\to+\infty}{\longrightarrow} \langle T, \phi\rangle.$$

The example of the exercise below is often useful.

Exercise 1.32. *Let $\chi \in \mathcal{D}(\mathbb{R}^d)$ be a non-negative function, with support in the unit ball and of integral equal to 1. Set*

$$\chi_n(x) = n^d \chi(n\,x).$$

Show that

$$\chi_n \underset{n\to\infty}{\longrightarrow} \delta_0 \ \text{ in } \mathcal{D}'(\mathbb{R}^d).$$

Definition 1.33. *A sequence $(\chi_n)_{n\in\mathbb{N}}$ of test functions in $\mathcal{D}(\mathbb{R}^d)$ satisfying*

$$\chi_n \underset{n\to+\infty}{\longrightarrow} \delta_0 \quad \text{ in } \mathcal{D}'(\mathbb{R}^d)$$

is called an approximation of the identity.

Remark 1.34. *This name comes from the fact that the distribution δ_0 is the identity for the convolution operation (this operation is defined in Section A.5 for functions in $L^1(\mathbb{R}^d)$ and it can be extended to some distributions).*

Proposition 1.35. *Let $1 \le p \le +\infty$. Convergence in $L^p_{\mathrm{loc}}(\Omega)$ implies convergence in $\mathcal{D}'(\Omega)$.*

Proof. Consider $(f_n)_{n\in\mathbb{N}}$ that converges to some f in $L^p_{\text{loc}}(\Omega)$ when $n\to\infty$. Let $\phi\in\mathcal{D}(\Omega)$. Using the Hölder inequality, we have, for any $1<p<+\infty$,

$$\begin{aligned}
|\langle f_n,\phi\rangle-\langle f,\phi\rangle| &= \left|\int_\Omega (f_n-f)\phi\right| \\
&\leq \int_\Omega |f_n-f|\,|\phi| \\
&\leq \|\phi\|_{L^\infty}\int_{\text{Supp}(\phi)}|f_n-f| \\
&\leq \|\phi\|_{L^\infty}\left(\int_{\text{Supp}(\phi)}|f_n-f|^p\right)^{1/p}\left(\int_{\text{Supp}(\phi)}1^{p'}\right)^{1/p'} \\
&\leq \|\phi\|_{L^\infty}\,\|f_n-f\|_{L^p(\text{Supp}(\phi))}\,|\text{Supp}(\phi)|^{1/p'}.
\end{aligned}$$

We hence get that $\lim\limits_{n\to\infty}|\langle f_n,\phi\rangle-\langle f,\phi\rangle|=0$. If $p=1$ or $p=+\infty$, then the proof, which is even simpler, is left to the reader. □

Remark 1.36. *The fact that a sequence of functions f_n converges almost everywhere to some function f (i.e. $\lim\limits_{n\to\infty}f_n(x)=f(x)$ a.e. on Ω) does not imply convergence in $\mathcal{D}'$. The converse is also false. The three situations below are possible:*

1. *a sequence converges in $\mathcal{D}'$ but not almost everywhere;*
2. *a sequence converges almost everywhere but not in $\mathcal{D}'$;*
3. *a sequence converges almost everywhere and in $\mathcal{D}'$ but the limits are different.*

We refer to the exercise 1.73 for some examples.

With the above definitions, differentiation is a continuous operation in $\mathcal{D}'(\Omega)$, in the following sense.

Proposition 1.37. *Consider a sequence (T_n) that converges to T in $\mathcal{D}'(\Omega)$ when $n\to\infty$. Then, for any $\alpha\in\mathbb{N}^d$, $\partial^\alpha T_n$ converges to $\partial^\alpha T$ in $\mathcal{D}'(\Omega)$ when $n\to\infty$.*

Proof. Let $\phi\in\mathcal{D}(\Omega)$. We have

$$\langle\partial^\alpha T_n,\phi\rangle=(-1)^{|\alpha|}\langle T_n,\partial^\alpha\phi\rangle\longrightarrow(-1)^{|\alpha|}\langle T,\partial^\alpha\phi\rangle=\langle\partial^\alpha T,\phi\rangle.$$

We hence get $\partial^\alpha T_n\longrightarrow_{n\to\infty}\partial^\alpha T$. □

The following proposition is a direct corollary of the above result.

Proposition 1.38. *Let $(T_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{D}'(\Omega)$. Assume that the series $\sum\limits_{n\in\mathbb{N}}T_n$ converges in $\mathcal{D}'(\Omega)$ to a distribution T. Then, for any $\alpha\in\mathbb{N}^d$, the series $\sum\limits_{n\in\mathbb{N}}\partial^\alpha T_n$ converges in $\mathcal{D}'(\Omega)$ and we have*

$$\partial^\alpha T=\sum_{n\in\mathbb{N}}\partial^\alpha T_n.$$

In $\mathcal{D}'$, it is thus possible to differentiate a series as soon as

$$T = \sum_{n \in \mathbb{N}} T_n,$$

this equality being an equality between distributions. This result is of course in sharp contrast with the results when manipulating functions.

The following result is very useful, and make the manipulation of distributions easy. We omit its proof.

Theorem 1.39. *Consider a sequence $(T_n)_{n \in \mathbb{N}}$ of distributions in Ω. Assume that, for any $\phi \in \mathcal{D}(\Omega)$, the sequence $(\langle T_n, \phi \rangle)_{n \in \mathbb{N}}$ converges to some limit ℓ_ϕ. Then, the map T defined by*

$$\langle T, \phi \rangle = \ell_\phi$$

is a distribution on Ω.

Note that T is obviously a linear form on $\mathcal{D}(\Omega)$. To show that it is a distribution, we are left with showing the continuity property (1.2), which is actually not easy. The above theorem states that this property indeed holds.

Exercise 1.40. *Let $a \in \mathbb{R}^d$ and $e \in \mathbb{R}^d$ such that $|e| = 1$. Identify the limit in $\mathcal{D}'(\mathbb{R}^d)$, as ϵ tends to 0, of the family of distributions*

$$T_\epsilon = \frac{1}{\epsilon} \delta_{a+\epsilon e} - \frac{1}{\epsilon} \delta_a.$$

Exercise 1.41. *Let $h \in L^1_{\rm loc}(\mathbb{R}^2)$ and $x_0 \in \mathbb{R}$. We define the* single layer *distribution $SL(h, x_0) \in \mathcal{D}'(\mathbb{R}^3)$ by*

$$\forall \phi \in \mathcal{D}(\mathbb{R}^3), \qquad \langle SL(h, x_0), \phi \rangle = \int_{\mathbb{R}^2} h(y,z)\, \phi(x_0, y, z)\, dy\, dz.$$

Show that $SL(h, x_0)$ is indeed a distribution on $\mathbb{R}^3$.

Let $g \in L^1_{\rm loc}(\mathbb{R}^2)$. Identify the limit in $\mathcal{D}'(\mathbb{R}^3)$, as ϵ tends to 0, of the family of distributions

$$T_\epsilon = SL(g/\epsilon, \epsilon) - SL(g/\epsilon, 0).$$

Exercise 1.42. *Show that $\displaystyle\sum_{n \in \mathbb{Z}} \delta_n$ converges in $\mathcal{D}'(\mathbb{R})$.*

Exercise 1.43. *Does the series*

$$\sum_{n=1}^{+\infty} \delta_{1/n}$$

converge in $\mathcal{D}'(\mathbb{R})$? What about in $\mathcal{D}'(0, +\infty)$?

1.8 More on differentiation: the one-dimensional case

1.8.1 The case of C^1 functions

Consider $f \in C^1(a,b)$. The functions f and $\dfrac{df}{dx}$ are in $L^1_{\text{loc}}(a,b)$, and hence in $\mathcal{D}'(a,b)$. We have

$$\left\langle \frac{d}{dx}T_f, \phi \right\rangle = -\int_a^b f\,\frac{d\phi}{dx} = \int_a^b \frac{df}{dx}\,\phi = \left\langle T_{\frac{df}{dx}}, \phi \right\rangle.$$

This means that

$$\frac{d}{dx}T_f = T_{\frac{df}{dx}},$$

namely that the derivation in the sense of distributions coincides with the standard derivation for functions of class C^1.

1.8.2 The case of functions that are piecewise C^1

Definition 1.44. *A function f is* piecewise C^k *in (a,b) if, for any compact interval $[\alpha,\beta]$ contained in (a,b), there exists a finite number of points $\alpha = a_0 < a_1 < \ldots < a_{N+1} = \beta$ such that*

- *in each interval (a_i, a_{i+1}), f is C^k;*
- *f and its derivatives up to the order k can be extended by continuity on each side of the points $a_1, \ldots, a_N$ (note that the left and right values of f – or its derivatives – at points a_i are not necessarily equal).*

Note that a piecewise C^1 function has at most a countable number of discontinuity points and that the possible accumulation points of the set of discontinuity points are only a and b.

Here are some examples and counter-examples:

- the Heaviside function is piecewise C^1;
- the function $x \mapsto |x|$ is piecewise C^1;
- the function $x \mapsto \tan x$ is not piecewise C^1;
- the function $x \mapsto \sqrt{|x|}$ is not piecewise C^1.

Theorem 1.45 (Jump formula). *Let f be a piecewise C^1 function on (a,b). With the above notation, we have*

$$f' = f'_{\text{reg}} + \sum_{i\in\mathcal{I}} [f(c_i+0) - f(c_i-0)]\,\delta_{c_i},$$

where f' is the derivative of f in the sense of distributions, $\{c_i\}_{i\in\mathcal{I}}$ are the points where f is discontinuous (recall that $\mathcal{I}$ is countable and its only possible accumulation points are a and b), f'_{reg} is the piecewise continuous function defined (except at the points c_i) as the standard derivative of f, $f(c_i+0)$ and $f(c_i-0)$ are the (right and left) limits of f in c_i, and δ_{c_i} is the Dirac mass in c_i.

Proof. The proof is based on the integration by part formula. □

Exercise 1.46. *Let f be a piecewise C^2 function on (a,b). Check that*

$$f'' = f''_{\text{reg}} + \sum_{i\in\mathcal{I}} [f'(c_i+0) - f'(c_i-0)]\,\delta_{c_i} + \sum_{i\in\mathcal{I}} [f(c_i+0) - f(c_i-0)]\,\delta'_{c_i}$$

for some countable set $\mathcal{I}$ and some points $\{c_i\}_{i\in\mathcal{I}}$.

1.8.3 Linear differential equations in $\mathcal{D}'(a,b)$

Theorem 1.47. *Consider the open interval (a,b).*

1. *The distributions T on (a,b) satisfying $T'=0$ in $\mathcal{D}'(a,b)$ are the constant functions.*
2. *For any $S\in\mathcal{D}'(a,b)$, there exists $T\in\mathcal{D}'(a,b)$ such that $T'=S$.*

The first statement means that, if $T'=0$ in $\mathcal{D}'(a,b)$, then there exists a constant C such that $T=C$, which means that, for any $\phi\in\mathcal{D}(a,b)$, we have $\langle T,\phi\rangle = C\int_a^b \phi$.

Proof. We first note that a test function $\phi\in\mathcal{D}(a,b)$ admits a primitive in $\mathcal{D}(a,b)$ if and only if $\int_a^b \phi = 0$ (in this case, the primitive is $\int_a^x \phi$). Let $\rho\in\mathcal{D}(a,b)$ such that $\int_a^b \rho = 1$. Since the integral of $\phi - \left(\int_a^b \phi\right)\rho$ vanishes, there exists $\psi\in\mathcal{D}(a,b)$ such that

$$\psi'(x) = \phi(x) - \left(\int_a^b \phi\right)\rho(x). \tag{1.13}$$

Consider now $T\in\mathcal{D}'(a,b)$ such that $T'=0$. We write

$$\langle T,\phi\rangle = \left(\int_a^b \phi\right)\langle T,\rho\rangle + \langle T,\psi'\rangle = \left(\int_a^b \phi\right)\langle T,\rho\rangle - \langle T',\psi\rangle = \left(\int_a^b \phi\right)\langle T,\rho\rangle.$$

Denoting C the constant $\langle T,\rho\rangle$, we have hence shown that $T=C$.

Consider now $S\in\mathcal{D}'(a,b)$. For any $\phi\in\mathcal{D}(a,b)$, we set

$$\langle T,\phi\rangle = -\langle S,\psi\rangle,$$

where $\psi\in\mathcal{D}(a,b)$ is uniquely defined by (1.13). It is then easy to show that T is a distribution and that $T'=S$. □

Theorem 1.47 can be used to show existence and uniqueness results for differential equations in $\mathcal{D}'$. The case of the equation $xT'+T=0$ is considered in the exercise 1.76.

1.8.4 Link between the standard derivative and the derivative in the sense of distributions

For a function $f\in L^1_{\text{loc}}$, the relation between the standard derivative and the derivative in the sense of distributions is not straightforward:

1. for a C^1 function, the two concepts lead to the same object;
2. for a piecewise C^1 function, the standard derivative is defined almost everywhere, but does not fully represent the variations of f since it ignores jumps; the derivative in the sense of distributions is the correct concept;
3. when f is everywhere differentiable without being piecewise C^1, the situation is more complex. When f' is locally integrable, the two concepts coincide (see Exercise 1.77). Otherwise, the derivative in the sense of distributions is somehow the finite part of the standard derivative (see the exercises 1.54 and 1.58).

1.9 More on differentiation: the multi-dimensional case

1.9.1 Schwarz Theorem

Theorem 1.48 (Schwarz Theorem)**.** *Let $T \in \mathcal{D}'(\Omega)$. We have*

$$\partial^\alpha \partial^\beta T = \partial^\beta \partial^\alpha T = \partial^{\alpha+\beta} T.$$

Proof. The proof is straightforward. □

1.9.2 The case of C^1 functions

As in the one-dimensional case, we have the following result. Consider Ω an open subset of $\mathbb{R}^d$ and $f \in C^1(\Omega)$. Then, for any $1 \leq i \leq d$,

$$\frac{\partial}{\partial x_i} T_f = T_{\frac{\partial f}{\partial x_i}}.$$

1.9.3 Stokes formula and applications

Definition 1.49. *An bounded open set $\Omega \subset \mathbb{R}^d$ is* smooth *(of class C^1) if, at any $x \in \partial\Omega$, it is possible to define an exterior normal vector, denoted $n(x)$, and if the function $x \mapsto n(x)$ from $\partial\Omega$ to $\mathbb{R}^d$ is continuous.*

Examples: the circle (in dimension $d = 2$) or the sphere (when $d = 3$) are smooth. The square (when $d = 2$) or the cube (when $d = 3$) are not smooth.

Theorem 1.50 (Stokes formula)**.** *Let Ω be a bounded smooth open subset of $\mathbb{R}^d$ and let X be a field defined in $\overline{\Omega}$, which is vector valued, and such that all components belong to $C^1(\overline{\Omega})$. Then*

$$\int_\Omega \text{div}\,(X)\,dx = \int_{\partial\Omega} X \cdot n\,d\sigma.$$

Corollary 1.51 (Integration by part formula)**.** *Let Ω be a bounded smooth open subset of $\mathbb{R}^d$ and let f and g in $C^1(\overline{\Omega})$. Then*

$$\int_\Omega \frac{\partial f}{\partial x_i}\,g\,dx = \int_{\partial\Omega} f\,g\,(n \cdot e_i)\,d\sigma - \int_\Omega f\,\frac{\partial g}{\partial x_i}\,dx.$$

Proof. Use the Stokes formula for the field $X(x) = f(x)\,g(x)\,e_i$. □

Corollary 1.52 (Green formula)**.** *Let Ω be a bounded smooth open subset of $\mathbb{R}^d$ and let f and g in $C^2(\overline{\Omega})$. Then*

$$\int_\Omega f\,\Delta g = -\int_\Omega \nabla f \cdot \nabla g + \int_{\partial\Omega} f\,\frac{\partial g}{\partial n}.$$

Proof. Use the Stokes formula for the field $X(x) = f(x)\,\nabla g(x)$. □

1.10 Principal values and finite parts

Principal values and finite parts of functions are distributions that naturally appear when one wants to associate a distribution with some singular functions (i.e. functions not in L^1_{loc}). Here, we only consider two specific examples.

1.10.1 Principal value of $1/x$

We define the principal value of the function $x \in \mathbb{R} \mapsto 1/x$ by

$$\forall \phi \in \mathcal{D}(\mathbb{R}), \qquad \left\langle \mathrm{PV}\left(\frac{1}{x}\right), \phi \right\rangle = \lim_{\epsilon \to 0,\ \epsilon > 0} \int_{|x|>\epsilon} \frac{\phi(x)}{x}\, dx. \tag{1.14}$$

Exercise 1.53. *Show that $PV\left(\frac{1}{x}\right)$ is indeed a distribution, of order 1. Hint: show that the distribution is of order at most 1, and then that its order is exactly 1. To do this, compute its action on the sequence $\phi_j(x) = \phi_0(x)\, atan(jx)$, where $\phi_0 \in \mathcal{D}(\mathbb{R})$ is even and equal to 1 in the neighborhood of 0.*

Exercise 1.54. *Show that the function $x \in \mathbb{R} \mapsto \ln|x|$ defines a distribution in $\mathcal{D}'(\mathbb{R})$ and compute its derivative.*

Exercise 1.55. *Show that $x\, PV\left(\frac{1}{x}\right) = 1$.*

Exercise 1.56. *Explain why the quantities*

$$\left(x\, PV\left(\frac{1}{x}\right)\right) \delta_0 \qquad \textit{and} \qquad (x\, \delta_0)\, PV\left(\frac{1}{x}\right).$$

are well defined in $\mathcal{D}'(\mathbb{R})$ and compute them. In view of this computation, does it seem possible to define the product of any two distributions?

1.10.2 Finite part of $H(x)x^\alpha$

Let $\phi \in \mathcal{D}(\mathbb{R})$ and H be the Heaviside function. The integral

$$\int_{\mathbb{R}} H(x)x^\alpha \phi(x)\, dx = \int_0^{+\infty} x^\alpha \phi(x)\, dx$$

is well-defined for $\alpha > -1$, but not for any $\alpha \leq -1$ if $\phi(0) \neq 0$. In the second case, $H(x)x^\alpha$ is therefore not a distribution. It is nevertheless still possible to associate a distribution to the function $H(x)x^\alpha$, as we explain now.

Using the Hadamard lemma 1.22, one can show that, for any $\epsilon > 0$,

$$\int_\epsilon^{+\infty} \phi(x)x^\alpha\, dx = P_\phi(\epsilon) + R_\phi(\epsilon),$$

where P_ϕ is a linear combination of negative powers of ϵ (and of $\ln \epsilon$ when α is a negative integer) and where $R_\phi(\epsilon)$ converges to a finite limit as ϵ tends to 0. One then defines the finite part of the function $H(x)x^\alpha$, which is denoted $\mathrm{FP}(H(x)x^\alpha)$, by the formula

$$\forall \phi \in \mathcal{D}(\mathbb{R}), \qquad \langle \mathrm{FP}(H(x)x^\alpha), \phi \rangle = \lim_{\epsilon \to 0,\ \epsilon > 0} R_\phi(\epsilon).$$

Exercise 1.57. *Show that the finite part of the function $H(x)x^\alpha$ is a distribution of order the integer part of α.*

Exercise 1.58. *Let $-1 < \alpha < 0$. Show that the derivative (in the sense of distributions) of the function $H(x)x^\alpha$ is $\alpha FP\Big(H(x)x^{\alpha-1}\Big)$.*

1.11 Distributions with compact support

1.11.1 Definitions and first properties

Definition 1.59. *Let $T \in \mathcal{D}'(\Omega)$.*

1. *Let ω be an open subset of Ω. The distribution T is said to vanish on ω if, for any $\phi \in \mathcal{D}(\Omega)$ such that Supp$(\phi) \subset \omega$, we have $\langle T, \phi\rangle = 0$.*
2. *The* support *of T is the complement in Ω of the union of the open subsets of Ω on which T vanishes.*

Note that, by construction, T vanishes on $\Omega \setminus \mathrm{Supp}(T)$. The notion of support of a distribution generalizes the notion of support of a function, as shown by Exercise 1.60 below.

Exercise 1.60. *Consider $f \in C^0(\mathbb{R})$. Since $C^0(\mathbb{R}) \subset L^1_{\text{loc}}(\mathbb{R})$, a distribution T_f can naturally be associated to the function f (see* (1.3)*). Show that Supp$(T_f) =$ Supp(f).*

Exercise 1.61. *Show that the support of $\delta_a \in \mathcal{D}'(\mathbb{R})$ is $\{a\}$.*

Definition 1.62. *The vector space of distributions on Ω with compact support is denoted $\mathcal{E}'(\Omega)$.*

Theorem 1.63. *If a distribution $T \in \mathcal{D}'(\Omega)$ has a compact support, then it is of finite order.*

Proof. Let K be the support of T and $\alpha = d(K, \mathbb{R}^d \setminus \Omega)$ be the distance between K and $\mathbb{R}^d \setminus \Omega$ (if $\Omega = \mathbb{R}^d$, then we take $\alpha = +\infty$). We set $\beta = \inf(1, \alpha)$ and consider the sets

$$K' = \left\{ x \in \mathbb{R}^d, \quad d(x, K) \le \frac{\beta}{3} \right\}, \qquad \Omega' = \left\{ x \in \mathbb{R}^d, \quad d(x, K) < \frac{2\beta}{3} \right\}.$$

The set K' is compact, the set Ω' is open and with compact closure and we have

$$K \subset K' \subset \Omega' \subset \overline{\Omega'} \subset \Omega.$$

Let $p \in \mathbb{N}$ and $C \in \mathbb{R}$ such that

$$\forall \phi \in \mathcal{D}_{\overline{\Omega'}}(\Omega), \qquad |\langle T, \phi\rangle| \le C \sup_{x \in \overline{\Omega'},\, |\alpha| \le p} |\partial^\alpha \phi(x)|.$$

Consider now $\rho \in \mathcal{D}(\Omega)$ which is equal to 1 on K' and with support in Ω'. For any $\phi \in \mathcal{D}(\Omega)$, we have

$$\langle T, \phi\rangle = \langle T, \rho\phi\rangle + \langle T, (1-\rho)\phi\rangle$$

and $\langle T, (1-\rho)\phi\rangle = 0$ since the supports of T and of $(1-\rho)\phi$ are disjoint. Moreover, since $\mathrm{Supp}(\rho\phi) \subset \overline{\Omega'}$, we have

$$\forall \phi \in \mathcal{D}(\Omega), \qquad |\langle T, \phi\rangle| \le C \sup_{x \in \Omega,\, |\alpha| \le p} |\partial^\alpha (\rho\phi)(x)|.$$

Using Leibniz formula, we get

$$\forall \phi \in \mathcal{D}(\Omega), \qquad |\langle T, \phi\rangle| \le C' \sup_{x \in \Omega,\, |\alpha| \le p} |\partial^\alpha \phi(x)|$$

with $C' = C \displaystyle\sup_{x \in \Omega,\, |\alpha| \le p,\, \beta \le \alpha} \frac{\alpha!}{\beta!\,(\alpha-\beta)!} |\partial^\beta \rho(x)|$. We thus obtain that T has a finite order, lower than or equal to p. □

Exercise 1.64. *Let $T \in \mathcal{D}'(\Omega)$ with $\Omega \subset \mathbb{R}^d$. Then, for any $1 \leq i \leq d$, we have Supp$\left(\dfrac{\partial T}{\partial x_i}\right) \subset$ Supp(T).*

Remark 1.65. *Let $T \in \mathcal{D}'(\Omega)$ and consider $\phi \in \mathcal{D}(\Omega)$ such that ϕ vanishes on Supp(T). This does not imply that $\langle T, \phi\rangle = 0$.*

Consider indeed the case of $T = \delta_0'$, the support of which is $\{0\}$ (in view of Exercises 1.61 and 1.64, we have Supp$(\delta_0') \subset \{0\}$, and the support cannot be empty as δ_0' is not equal to 0). Consider $\phi(x) = x\rho(x)$, with $\rho \in \mathcal{D}(\mathbb{R})$ equal to 1 on a neighborhood of 0. We observe that ϕ vanishes on Supp(δ_0') while $\langle \delta_0', \phi\rangle \neq 0$.

However, if ϕ vanishes on an open neighborhood of Supp(T), then $\langle T, \phi\rangle = 0$ (see the proof of Proposition 1.66 below where this argument is used).

1.11.2 Distributions with support restricted to a point

Proposition 1.66. *Let $T \in \mathcal{D}'(\Omega)$ of order 0 such that Supp$(T) = \{a\}$. Then there exists a constant c such that*

$$T = c\,\delta_a.$$

Proof. Let $\epsilon > 0$ such that $B_a(\epsilon) \subset \Omega$. Let $K = \overline{B_a(\epsilon/2)}$ and C be a constant such that

$$\forall \phi \in \mathcal{D}_K(\Omega), \qquad |\langle T, \phi\rangle| \leq C \sup_{x\in K} |\phi(x)|.$$

Let $0 < r < \epsilon$ and $\rho \in \mathcal{D}(\Omega)$, with compact support in $B_a(r)$ and such that $0 \leq \rho \leq 1$ and $\rho = 1$ in K. For any $\phi \in \mathcal{D}(\Omega)$, we have

$$\langle T, \phi\rangle = \langle T, \rho\phi\rangle + \langle T, (1-\rho)\phi\rangle.$$

The function $(1-\rho)\phi$ vanishes in a neighborhood of $\{a\} = \text{Supp}(T)$, hence $\langle T, (1-\rho)\phi\rangle = 0$. We also observe that Supp$(\rho\phi) \subset K$, hence

$$|\langle T, \phi\rangle| = |\langle T, \rho\phi\rangle| \leq C \sup_{\Omega} |\rho\phi| \leq C \sup_{B_a(r)} |\phi|.$$

The constant C does not depend on r. Taking the limit $r \to 0$, we thus deduce that

$$|\langle T, \phi\rangle| \leq C|\phi(a)|. \tag{1.15}$$

Now consider an arbitrary function $\psi \in \mathcal{D}(\Omega)$. We write

$$\psi = \psi(a)\rho + (\psi - \psi(a)\rho)$$

and see that the function $\phi = \psi - \psi(a)\rho$ is such that $\phi(a) = 0$. Using (1.15), we obtain that $\langle T, \psi - \psi(a)\rho\rangle = 0$. Thus

$$\langle T, \psi\rangle = \langle T, \psi(a)\rho\rangle = \langle T, \rho\rangle\, \psi(a).$$

We hence get that $T = c\delta_a$, where $c = \langle T, \rho\rangle$. □

The next proposition (the proof of which is omitted) generalizes Proposition 1.66 by characterizing the distributions (of any order) that have a support equal to the point $\{a\}$.

Proposition 1.67. *Let $T \in \mathcal{D}'(\Omega)$ such that Supp$(T) = \{a\}$. Then T is of the form*

$$T = \sum_{|\alpha|\leq p} c_\alpha \partial^\alpha \delta_a,$$

where p is the order of T and where c_α, $|\alpha| \leq p$, are constant.

Exercise 1.68. *Show that δ_a and $PV\left(\dfrac{1}{x}\right)$ have compact support and compute their support.*

1.12 Exercises

We collect here some exercises that are, in general, less straightforward that the previous ones.

Exercise 1.69. *For the following maps $T : \mathcal{D}(\mathbb{R}) \to \mathbb{R}$, identify which ones are distributions:*

- $\langle T, \phi \rangle = \displaystyle\int_{\mathbb{R}} |\phi(t)|\, dt.$
- $\langle T, \phi \rangle = \displaystyle\sum_{n=0}^{\infty} \phi^{(n)}(n).$

Exercise 1.70. *Show that the map $T : \mathcal{D}(\mathbb{R}) \to \mathbb{R}$ defined by $\langle T, \phi \rangle = \displaystyle\sum_{n=1}^{\infty} \frac{1}{n} \left(\phi\left(\frac{1}{n}\right) - \phi(0) \right)$ is a distribution of order lower or equal to 1.*

Exercise 1.71. *Consider $T : \mathcal{D}(\mathbb{R}^2) \to \mathbb{R}$ defined by*

$$\langle T, \phi \rangle = \int_0^{\infty} \phi(z, 2z)\, dz.$$

Show that T is a distribution and that $\dfrac{\partial T}{\partial x} + 2\dfrac{\partial T}{\partial y} = \delta_{(0,0)}$ in $\mathcal{D}'(\mathbb{R}^2)$.

Exercise 1.72. *Consider $v_n(x) = \dfrac{n}{1 + n^2\, x^2}$ and $w_n(x) = atan(nx)$.*

1. *Show that $v_n(x)$ converges to some $v(x)$ almost everywhere on $\mathbb{R}$. Identify v.*
2. *Show that, for any $\alpha > 0$, v_n converges in $L^1((-\infty, -\alpha) \cup (\alpha, \infty))$.*
3. *Show that w_n converges in $\mathcal{D}'(\mathbb{R})$ to $\dfrac{\pi}{2} sgn(x)$.*
4. *Compute the limit of v_n in $\mathcal{D}'(\mathbb{R})$ by two different methods.*

Exercise 1.73. *The aim of this exercise is to study the links between convergence almost everywhere and convergence in $\mathcal{D}'$.*

1. *Consider $f \in \mathcal{D}(\mathbb{R})$ and the sequence $f_n(x) = f(x - n)$. Study the almost everywhere convergence of f_n, the convergence in $\mathcal{D}'(\mathbb{R})$ and the convergence in $L^1(\mathbb{R})$.*
2. *Consider $f_n(x) = e^{inx}$. Show that f_n does not converge almost everywhere but that f_n converges to 0 in $\mathcal{D}'(\mathbb{R})$.*
3. *Consider $f_n(x) = \displaystyle\sum_{k=1}^{n} \chi_{1/n^2}\left(x - \frac{1}{k}\right)$, where χ_ϵ is an approximation of identity, that is $\chi_\epsilon(x) = \epsilon\chi(\epsilon\, x)$ where $\chi \in \mathcal{D}(\mathbb{R})$, with support in the unit ball and with integral equal to 1. Show that $f_n \to 0$ almost everywhere but that f_n does not converge in $\mathcal{D}'(\mathbb{R})$.*
4. *Consider $f_n(x) = \dfrac{1}{\sigma_n\sqrt{2\pi}}\, e^{-x^2/2\sigma_n^2}$ with $\sigma_n \to 0$. Show that $f_n \to 0$ almost everywhere and that $f_n \to \delta_0$ in $\mathcal{D}'(\mathbb{R})$.*

Exercise 1.74. *Identify all the distributions $T \in \mathcal{D}'(\mathbb{R})$ such that $x\,T = 0$.*

Exercise 1.75. *For any $\phi \in \mathcal{D}(\mathbb{R})$, we define*

$$\left\langle \frac{1}{x+i0}, \phi \right\rangle = \lim_{\epsilon \to 0,\ \epsilon > 0} \int_{\mathbb{R}} \frac{\phi(x)}{x+i\epsilon}\, dx \quad \textit{and} \quad \left\langle \frac{1}{x-i0}, \phi \right\rangle = \lim_{\epsilon \to 0,\ \epsilon > 0} \int_{\mathbb{R}} \frac{\phi(x)}{x-i\epsilon}\, dx.$$

Show that $\dfrac{1}{x+i0}$ and $\dfrac{1}{x-i0}$ define distributions of order 1. Show that

$$\frac{1}{x+i0} - \frac{1}{x-i0} = -2i\pi\delta_0 \quad \textit{and} \quad \frac{1}{x+i0} + \frac{1}{x-i0} = 2\,PV\left(\frac{1}{x}\right).$$

Exercise 1.76. *Identify all distributions $T \in \mathcal{D}'(\mathbb{R})$ such that $x\,T' + T = 0$. (Hint: consider $S = x\,T$).*

Exercise 1.77. *Let $f \in L^1_{\mathrm{loc}}(\mathbb{R})$ and $F(x) = \displaystyle\int_0^x f$. Show that the derivative (in the sense of distributions) of F is equal to f.*

Chapter 2

Sobolev spaces

Using the distribution theory exposed in Chapter 1, we are now in position to build the appropriate spaces of functions for the study of Partial Differential Equations.

2.1 The spaces $H^k(\Omega)$

Definition 2.1. *Let Ω be an open subset of $\mathbb{R}^d$ and let $k \in \mathbb{N}$. The set of functions of $L^2(\Omega)$, the derivative (in the sense of distributions) of which, up to order k, are in $L^2(\Omega)$, is denoted $H^k(\Omega)$:*

$$H^k(\Omega) = \left\{ f \in L^2(\Omega), \quad \forall \alpha \in \mathbb{N}^d, \ |\alpha| \leq k, \quad \partial^\alpha f \in L^2(\Omega) \right\}.$$

We recall that we can associate to any $f \in L^2(\Omega)$ a distribution, which is again denoted f. Since f is a distribution, it can be differentiated, and $\partial^\alpha f$ is hence a distribution (we recall that, for $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$, we denote $\partial^\alpha f = \dfrac{\partial^{\alpha_1 + \cdots + \alpha_d} f}{\partial^{\alpha_1} x_1 \ldots \partial^{\alpha_d} x_d}$, see (1.1)). In the above definition, we wrote that $\partial^\alpha f \in L^2(\Omega)$: this means that there exists a function $F_\alpha \in L^2(\Omega)$ such that

$$\forall \phi \in \mathcal{D}(\Omega), \quad \langle \partial^\alpha f, \phi \rangle = \int_\Omega F_\alpha \phi.$$

Theorem 2.2. *$H^k(\Omega)$ is a vector space. Endowed with the scalar product*

$$(f,g)_{H^k} = \sum_{|\alpha| \leq k} \int_\Omega \partial^\alpha f(x) \, \partial^\alpha g(x) \, dx,$$

it is an Hilbert space. Its norm is denoted $\| \cdot \|_{H^k}$.

Proof. Obviously, $H^k(\Omega)$ is a vector space and $(\cdot, \cdot)_{H_k}$ is a scalar product on $H^k(\Omega)$. We are left with showing that $H^k(\Omega)$ is a complete space for the norm $\| \cdot \|_{H^k}$. Let $(f_n)_{n \in \mathbb{N}}$ be a Cauchy sequence of $H^k(\Omega)$. We first observe that

$$\|f_p - f_q\|_{L^2} \leq \|f_p - f_q\|_{H^k},$$

hence (f_n) is a Cauchy sequence in L^2, hence it converges in L^2 to some $f \in L^2$. Likewise, for any $|\alpha| \leq k$,

$$\|\partial^\alpha f_p - \partial^\alpha f_q\|_{L^2} \leq \|f_p - f_q\|_{H^k},$$

hence $\partial^\alpha f_n$ converges in L^2 to some function g_α. In addition, convergence in L^2 implies convergence in $\mathcal{D}'$. Hence

$$f_n \longrightarrow f \qquad \text{in } \mathcal{D}'$$

and

$$\partial^\alpha f_n \longrightarrow g_\alpha \qquad \text{in } \mathcal{D}'.$$

From the first assertion, we deduce that

$$\partial^\alpha f_n \longrightarrow \partial^\alpha f \qquad \text{in } \mathcal{D}'.$$

By uniqueness of the limit in $\mathcal{D}'(\Omega)$, we find that $\partial^\alpha f = g_\alpha \in L^2$. Hence $f \in H^k$. Since $\partial^\alpha f_n \longrightarrow \partial^\alpha f$ in L^2, we get that $f_n \longrightarrow f$ in H^k. □

2.2 The space $H^1_0(\Omega)$

2.2.1 Definition

We start by an important density result.

Proposition 2.3.

1. *If $\Omega = \mathbb{R}^d$, then $\mathcal{D}(\mathbb{R}^d)$ is dense in $H^1(\mathbb{R}^d)$.*
2. *If $\Omega \subset \mathbb{R}^d$, $\Omega \neq \mathbb{R}^d$, then $\mathcal{D}(\Omega)$ is not dense in $H^1(\Omega)$.*

Proof. The first assertion can be shown following a technical proof using truncation and regularization. We omit it here and focus on the second assertion, in dimension 1, when $\Omega = (0,1)$. Let $\phi \in \mathcal{D}(0,1)$. We have

$$\forall x \in (0,1), \qquad \phi(x) = \int_0^x \phi'(t)\,dt,$$

hence

$$\forall x \in (0,1),\ |\phi(x)| \leq \int_0^x |\phi'(t)|\,dt \leq \left(\int_0^x |\phi'(t)|^2\,dt\right)^{1/2} \left(\int_0^x 1^2\,dt\right)^{1/2} \leq \|\phi'\|_{L^2(0,1)}.$$

We therefore obtain

$$\|\phi\|_{L^2(0,1)} = \left(\int_0^1 |\phi(x)|^2\,dx\right)^{1/2} \leq \|\phi'\|_{L^2(0,1)},$$

an estimate which is valid for any $\phi \in \mathcal{D}(0,1)$.

If $\mathcal{D}(0,1)$ were dense in $H^1(0,1)$, then this estimate would remain true for any $\phi \in H^1(0,1)$. However, consider the function $f = 1$ on $(0,1)$. We see that $f \in H^1(0,1)$, $\|f\|_{L^2(0,1)} = 1$ and $\|f'\|_{L^2(0,1)} = 0$, hence f does not satisfy the above estimate. □

Definition 2.4. *The closure of $\mathcal{D}(\Omega)$ in $H^1(\Omega)$ is denoted $H^1_0(\Omega)$.*

Remark 2.5. *In view of Proposition 2.3, we have $H^1_0(\mathbb{R}^d) = H^1(\mathbb{R}^d)$. In contrast, $H^1_0(\Omega)$ is a strict subspace of $H^1(\Omega)$ when $\Omega \subset \mathbb{R}^d$ with $\Omega \neq \mathbb{R}^d$.*

Proposition 2.6. *The space $H^1_0(\Omega)$ is a vector space. It is an Hilbert space for the scalar product $(\cdot,\cdot)_{H^1}$.*

Proof. It is obvious that $H^1_0(\Omega)$ is a vector space (it is the closure of a vector space for a given norm). The scalar product of $H^1(\Omega)$ restricted to $H^1_0(\Omega)$ of course defines a scalar product on $H^1_0(\Omega)$. Since $H^1_0(\Omega)$ is closed in a complete space, it is itself complete. □

2.2.2 Poincaré inequality

We now state a very important result, that we will often need in these lecture notes (see e.g. Section 3.3).

Theorem 2.7 (Poincaré inequality)**.** *Let Ω be a bounded open subset of $\mathbb{R}^d$. There exists a constant C_Ω such that*

$$\forall u \in H_0^1(\Omega), \qquad \|u\|_{L^2(\Omega)} \le C_\Omega \, \|\nabla u\|_{L^2(\Omega)} .$$

In the language of functional analysis, it is said that the L^2 norm of u is *controlled* by the L^2 norm of its gradient.

Proof of Theorem 2.7. Since the open set Ω is bounded, there exists $L > 0$ such that $\Omega \subset [-L, L]^d$. Let $\phi \in \mathcal{D}(\Omega)$ and ψ the extension of ϕ by 0 to the whole space $\mathbb{R}^d$. We have that $\psi \in C^\infty(\mathbb{R}^d)$ and that

$$\forall x \in [-L, L]^d, \qquad \psi(x) = \int_{-L}^{x_1} \frac{\partial \psi}{\partial x_1}(t, x_2, \cdots, x_d)\, dt.$$

We hence have

$$\begin{aligned}
|\psi(x)|^2 &= \left(\int_{-L}^{x_1} \frac{\partial \psi}{\partial x_1}(t, x_2, \cdots, x_d)\, dt \right)^2 \\
&\le \left(\int_{-L}^{x_1} \left| \frac{\partial \psi}{\partial x_1}(t, x_2, \cdots, x_d) \right|^2 dt \right) \left(\int_{-L}^{x_1} 1^2\, dt \right) \\
&\le 2L \int_{-L}^{L} \left| \frac{\partial \psi}{\partial x_1}(t, x_2, \cdots, x_d) \right|^2 dt.
\end{aligned}$$

Integrating the above inequality on $[-L, L]^d$, we deduce that

$$\int_{[-L,L]^d} |\psi|^2 \le 4L^2 \int_{[-L,L]^d} \left| \frac{\partial \psi}{\partial x_1} \right|^2 \le 4L^2 \int_{[-L,L]^d} |\nabla \psi|^2.$$

Since Supp (ψ) = Supp $(\phi) \subset \Omega$, Supp $(\nabla\psi)$ = Supp $(\nabla\phi) \subset \Omega$, and since $\phi = \psi$ and $\nabla\phi = \nabla\psi$ on Ω, we deduce that

$$\forall \phi \in \mathcal{D}(\Omega), \qquad \|\phi\|_{L^2(\Omega)} \le 2L \, \|\nabla \phi\|_{L^2(\Omega)} . \tag{2.1}$$

In addition, the mappings

$$\begin{array}{rcl} H^1(\Omega) & \longrightarrow & \mathbb{R} \\ u & \mapsto & \|u\|_{L^2(\Omega)} \end{array} \qquad \text{and} \qquad \begin{array}{rcl} H^1(\Omega) & \longrightarrow & \mathbb{R} \\ u & \mapsto & \|\nabla u\|_{L^2(\Omega)} \end{array}$$

are continuous. Since $\mathcal{D}(\Omega)$ is dense in $H_0^1(\Omega)$ under the H^1 norm, we get that the estimate (2.1) remains valid for any function in $H_0^1(\Omega)$. □

2.2.3 On an alternative definition of $H_0^1(\Omega)$ via the use of traces

We close this section by making some remarks on the definition of $H_0^1(\Omega)$ *via* the notion of traces. It is often said that $H_0^1(\Omega)$ is the space of functions in $H^1(\Omega)$ that "vanish on the boundary". To formalize this, we first introduce the notion of trace. For any function $u \in C^0(\overline{\Omega})$, the trace of u on $\partial\Omega$ is defined as

$$\begin{array}{rcl} \gamma(u) \, : \, \partial\Omega & \longrightarrow & \mathbb{R} \\ x & \mapsto & u(x). \end{array}$$

Put differently, we have $\gamma(u) = u|_{\partial\Omega}$. Note that the trace mapping

$$\begin{array}{rcl} \gamma \, : \, C^0(\overline{\Omega}) & \longrightarrow & C^0(\partial\Omega) \\ u & \mapsto & \gamma(u) \end{array}$$

is linear and continuous. The question is now to extend the notion of trace to functions that are less regular than $C^0(\overline{\Omega})$. This is not always possible. Here are some answers:

1. It is not possible to define the trace of a function $u \in L^2(\Omega)$.

   Consider indeed the function $u \, : \, x \mapsto \sin(1/x)$ on $\Omega = (0,1)$, which belongs to $L^\infty(\Omega) \subset L^2(\Omega)$. The boundary $\partial\Omega$ is the union of two points, 0 and 1. The function u is continuous at 1 and we can hence define its trace in 1 (it is the real number $\sin 1$). In contrast, the set of limit points of u at 0 is the whole interval $[-1,1]$. Hence, there is no natural way to define the trace of u (the value of u) at 0.

2. In dimension $d = 1$, a function in $H^1(a,b)$ admits a continuous representation. The proof of this statement is carried out in Exercise 2.14. Hence, one can define the trace in a and in b of a function belonging to $H^1(a,b)$.

3. In dimension $d \geq 2$, a function in $H^1(\Omega)$ is not necessarily continuous (see Exercise 2.15). However, one can still define its trace on $\partial\Omega$. More precisely, there exists a linear and continuous mapping

   $$\begin{array}{rcl} \gamma \, : \, H^1(\Omega) & \longrightarrow & L^2(\partial\Omega) \\ u & \mapsto & \gamma(u) \end{array}$$

   such that, for any $u \in H^1(\Omega) \cap C^0(\overline{\Omega})$, we have $\gamma(u) = u|_{\partial\Omega}$. Actually, the function $\gamma(u)$ is more regular than simply $L^2(\partial\Omega)$, it belongs to the space $H^{1/2}(\partial\Omega)$ (we refer to [1] for more details).

We have the following characterization, which provides a definition of $H^1_0(\Omega)$ which is equivalent to that given in Definition 2.4.

Proposition 2.8.

$$H^1_0(\Omega) = \left\{ u \in H^1(\Omega), \quad \gamma(u) = 0 \right\}.$$

2.3 The space $H^{-1}(\Omega)$

Definition 2.9. *Let Ω be an open subset of $\mathbb{R}^d$. We denote by $H^{-1}(\Omega)$ the vector space of distributions $T \in \mathcal{D}'(\Omega)$ for which there exists a constant C such that*

$$\forall \phi \in \mathcal{D}(\Omega), \qquad |\langle T, \phi \rangle| \leq C \|\phi\|_{H^1(\Omega)}.$$

Remark 2.10. *It is clear that $L^2(\Omega) \subset H^{-1}(\Omega)$. Indeed, if $f \in L^2(\Omega)$, then*

$$\forall \phi \in \mathcal{D}(\Omega), \qquad |\langle f, \phi \rangle| = \left| \int_\Omega f\phi \right| \leq \|f\|_{L^2(\Omega)} \, \|\phi\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)} \, \|\phi\|_{H^1(\Omega)}.$$

Theorem 2.11. *It is possible to identify $H^{-1}(\Omega)$ with the topological dual of $H^1_0(\Omega)$ (i.e. with the vector space of linear and continuous forms on $H^1_0(\Omega)$).*

The proof of Theorem 2.11 is given below. We note here that, as a consequence of the above results, we have the following inclusions:

$$\mathcal{D}(\Omega) \subset H_0^1(\Omega) \subset L^2(\Omega) \subset H^{-1}(\Omega) \subset \mathcal{D}'(\Omega)$$

and, for any $T \in L^2(\Omega)$ and $\phi \in \mathcal{D}(\Omega)$, we have

$$\langle T, \phi \rangle = \langle T, \phi \rangle_{H^{-1}, H_0^1} = \langle T, \phi \rangle_{L^2} = \int_\Omega T\, \phi.$$

The space L^2 is the so-called "pivot" space for all of these dualities: if $T \in L^2(\Omega)$ and $\phi \in \mathcal{D}(\Omega)$, then $\langle T, \phi \rangle$ can be understood in $(\mathcal{D}', \mathcal{D})$, (H^{-1}, H_0^1) or (L^2, L^2).

Proof of Theorem 2.11. Let $T \in H^{-1}(\Omega)$. The linear map

$$\phi \mapsto \langle T, \phi \rangle$$

is continuous on the space $\mathcal{D}(\Omega)$ *endowed with the H^1 norm.* Since $\mathcal{D}(\Omega)$ is dense in $H_0^1(\Omega)$ for that norm, we take as given the fact that the above mapping can be extended (in a unique way) to some linear and continuous mapping on $H_0^1(\Omega)$, that we denote

$$\phi \mapsto \langle T, \phi \rangle_{H^{-1}, H_0^1},$$

and which satisfies

$$\forall \phi \in \mathcal{D}(\Omega), \qquad \langle T, \phi \rangle_{H^{-1}, H_0^1} = \langle T, \phi \rangle.$$

Hence, to any $T \in H^{-1}(\Omega)$, we can associate an element of the topological dual of $H_0^1(\Omega)$ (i.e. the vector space of linear and continuous forms on $H_0^1(\Omega)$). Hence, we define

$$\begin{array}{rcl} \alpha \,:\, H^{-1}(\Omega) & \longrightarrow & (H_0^1(\Omega))' \\ T & \mapsto & \langle T, \cdot \rangle_{H^{-1}, H_0^1}. \end{array}$$

Conversely, let $L \in (H_0^1(\Omega))'$. There exists a constant C such that

$$\forall \phi \in H_0^1(\Omega), \qquad |L(\phi)| \le C \|\phi\|_{H^1}.$$

We restrict L to $\mathcal{D}(\Omega) \subset H_0^1(\Omega)$, and we hence get a linear form on $\mathcal{D}(\Omega)$ that satisfies

$$\forall \phi \in \mathcal{D}(\Omega), \qquad |L(\phi)| \le C \|\phi\|_{H^1}.$$

We are left to show that L is a distribution, i.e. that L satisfies the "continuity property" (1.2). Let K be a compact set contained in Ω and let $\phi \in \mathcal{D}_K(\Omega)$. We see that

$$|L(\phi)| \le C \|\phi\|_{H^1} \le C \left(\|\phi\|_{L^2}^2 + \|\nabla \phi\|_{L^2}^2 \right)^{1/2}.$$

Using that $\|\phi\|_{L^2} \le \sqrt{|K|} \sup |\phi|$ and $\|\nabla \phi\|_{L^2} \le \sqrt{|K|} \sup |\nabla \phi|$, we obtain that

$$|L(\phi)| \le C' \sup_{|\alpha| \le 1,\, x \in K} |\partial^\alpha \phi|.$$

Hence L defines a distribution (of order lower or equal to 1). We hence define

$$\begin{array}{rcl} \beta \,:\, (H_0^1(\Omega))' & \longrightarrow & H^{-1}(\Omega) \\ L & \mapsto & L_{\mathcal{D}(\Omega)}. \end{array}$$

It is easy to check that $\alpha \circ \beta = I_{(H_0^1(\Omega))'}$ and that $\beta \circ \alpha = I_{H^{-1}(\Omega)}$. □

Proposition 2.12 (Characterization of the distributions in H^{-1})**.** *Let Ω be an open subset of $\mathbb{R}^d$. A distribution T belongs to $H^{-1}(\Omega)$ if and only if there exist, for any $|\alpha| \leq 1$, a function $g_\alpha \in L^2(\Omega)$ such that*

$$T = \sum_{|\alpha|\leq 1} \partial^\alpha g_\alpha.$$

Proof. It is obvious that, if T is of the form

$$T = \sum_{|\alpha|\leq 1} \partial^\alpha g_\alpha$$

with $g_\alpha \in L^2(\Omega)$, we have

$$\begin{aligned} \forall \phi \in \mathcal{D}(\Omega), \qquad |\langle T, \phi\rangle| &= \left| \langle \sum_{|\alpha|\leq 1} \partial^\alpha g_\alpha, \phi \rangle \right| \\ &\leq \left| \sum_{|\alpha|\leq 1} (-1)^{|\alpha|} \langle g_\alpha, \partial^\alpha \phi \rangle \right| \\ &\leq \sum_{|\alpha|\leq 1} \|g_\alpha\|_{L^2} \|\partial^\alpha \phi\|_{L^2} \\ &\leq \left(\sum_{|\alpha|\leq 1} \|g_\alpha\|_{L^2} \right) \|\phi\|_{H^1}. \end{aligned}$$

Hence $T \in H^{-1}(\Omega)$. The converse statement is more challenging to prove. We omit its proof. □

2.4 Exercises

Exercise 2.13.

1. *For which values of k do the following functions belong to $H^k(-1,1)$: $x \mapsto \sin x$, $x \mapsto |x|$ and $x \mapsto sgn(x)$?*
2. *For which values of k, α and d does the function $x \mapsto \dfrac{1}{|x|^\alpha}$ belong to $H^k(B_0(1))$, where $B_0(1)$ is the unit ball in $\mathbb{R}^d$?*
3. *For which values of k, α and d does the function $x \mapsto \dfrac{1}{|x|^\alpha}$ belong to $H^k\left(\mathbb{R}^d \setminus \overline{B_0(1)}\right)$?*

Exercise 2.14. *Consider two real numbers a and b with $a < b$. Show that any function in $H^1(a,b)$ has a continuous representation.*

Exercise 2.15. *Let Ω be the disk centered at 0 of radius $1/e$ in $\mathbb{R}^2$. Show that the function*

$$u(x,y) = \ln\left[-\ln\left(\sqrt{x^2+y^2}\right)\right]$$

belongs to $H_0^1(\Omega)$ but that it is not continuous at 0.

Exercise 2.16. *Consider the vector space*

$$V = \left\{ v \in L^2(\Omega),\ \Delta v \in L^2(\Omega) \right\}$$

endowed with the scalar product $(v, w)_V = (v, w)_{L^2} + (\Delta v, \Delta w)_{L^2}$. *Show that* V *is a Hilbert space.*

Exercise 2.17. *Let* p *be a real number with* $1 \leq p \leq +\infty$. *Let* $k \in \mathbb{N}$. *We set*

$$W^{k,p}(\Omega) = \left\{ u \in L^p(\Omega),\ \partial^\alpha u \in L^p(\Omega) \text{ for any } \alpha \in \mathbb{N}^d,\ |\alpha| \leq k \right\}.$$

This vector space is endowed with the norm

$$\|u\|_{W^{k,p}} = \sum_{\alpha \in \mathbb{N}^d,\ |\alpha| \leq k} \|\partial^\alpha u\|_{L^p}.$$

Explain why we can consider the derivative (in the sense of distributions) of u, *and show that the space* $W^{k,p}(\Omega)$ *is a Banach space.*

Exercise 2.18. *Let* $u \in H^1(\mathbb{R})$. *Show that*

$$\|u\|^2_{L^\infty(\mathbb{R})} \leq \|u\|_{L^2(\mathbb{R})}\ \|u'\|_{L^2(\mathbb{R})}.$$

Hint: use the density of $\mathcal{D}(\mathbb{R})$ *in* $H^1(\mathbb{R})$.

Exercise 2.19. *Consider two real numbers* a *and* b *with* $a < b$. *Let* $f \in H^1(a, b)$. *Show that* $|f| \in H^1(a, b)$. *It will be useful to show that*

$$\frac{d|f|}{dx}(x) = \begin{cases} \dfrac{f(x)}{|f(x)|}\ \dfrac{df}{dx}(x) & \text{if } f(x) \neq 0, \\ 0 \quad \text{otherwise.} \end{cases}$$

Deduce that, if f *and* g *belong to* $H^1(a, b)$, *then* $\max(f, g)$ *and* $\min(f, g)$ *belong to* $H^1(a, b)$.

Exercise 2.20. *Let* $u \in H^1(\mathbb{R}^d)$. *We denote* $(e_1, \ldots, e_d)$ *the canonical basis of* $\mathbb{R}^d$. *Show that, for any* $1 \leq i \leq d$,

$$\int_{\mathbb{R}^d} \left| \frac{\partial u}{\partial x_i} \right|^2 = \lim_{t \to 0}\ \frac{1}{t^2}\ \int_{\mathbb{R}^d} |u(x + t\, e_i) - u(x)|^2\ dx.$$

Exercise 2.21. *Let* Ω *be a bounded open subset of* $\mathbb{R}^d$ *and* $f \in L^2(\Omega)$. *On* $H^1_0(\Omega)$, *consider the functional* J *defined by*

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

The quantity $J(u)$ *is the energy that naturally arises in thermal or elasticity problems.*

Show that J *is infinite at infinity, namely that* $J(u) \to \infty$ *when* $\|u\|_{H^1} \to \infty$.

Exercise 2.22. *The aim of this exercise is to show the Hardy inequality in* $\mathcal{D}(\mathbb{R}^3)$.

1. *Let $f \in C^\infty([0,\infty))$ and such that there exists $A > 0$ such that $f(r) = 0$ for any $r \geq A$. Show that*
$$\int_0^\infty f^2(r)\,dr = -2\int_0^\infty r\,f'(r)\,f(r)\,dr$$
and deduce that
$$\int_0^\infty f^2(r)\,dr \leq 4\int_0^\infty r^2\,(f'(r))^2\,dr.$$

2. *Deduce that*
$$\forall \psi \in \mathcal{D}(\mathbb{R}^3), \qquad \int_{\mathbb{R}^3} \frac{\psi^2(x)}{|x|^2}\,dx \leq 4\int_{\mathbb{R}^3} |\nabla\psi(x)|^2\,dx. \tag{2.2}$$
Hint: note that, in spherical coordinates,
$$\int_{\mathbb{R}^3} \frac{\psi^2(x)}{|x|^2}\,dx = \int_0^\pi \int_0^{2\pi} \left(\int_0^\infty \psi^2(r,\theta,\varphi)\,dr\right)\sin(\theta)\,d\theta d\varphi$$
and that $\nabla\psi = \dfrac{\partial\psi}{\partial r}\,e_r + \dfrac{1}{r}\dfrac{\partial\psi}{\partial\theta}\,e_\theta + \dfrac{1}{r\,\sin(\theta)}\,\dfrac{\partial\psi}{\partial\varphi}\,e_\varphi$, where the spherical basis $(e_r, e_\theta, e_\varphi)$ is orthonormal.

Exercise 2.23. *Consider the space*
$$\mathcal{A} = \left\{u \in H^1(\mathbb{R}^3),\ \|u\|_{L^2(\mathbb{R}^3)} = 1\right\}$$
and the function
$$\mathcal{E}(u) = \frac{1}{2}\int_{\mathbb{R}^3} |\nabla u|^2 - \int_{\mathbb{R}^3} \frac{|u(x)|^2}{|x|}\,dx.$$
The quantity $\mathcal{E}$ represents the electronic energy of the Hydrogen atom in quantum physics.

1. *Show that $\mathcal{E}(\psi)$ is well-defined whenever $\psi \in \mathcal{D}(\mathbb{R}^3)$.*

2. *Using the inequality* (2.2)*, show that, for any $\psi \in \mathcal{D}(\mathbb{R}^3)$, we have*
$$\int_{\mathbb{R}^3} \frac{|\psi(x)|^2}{|x|}\,dx \leq 2\|\psi\|_{L^2(\mathbb{R}^3)}\,\|\nabla\psi\|_{L^2(\mathbb{R}^3)}. \tag{2.3}$$

3. *Let $u \in H^1(\mathbb{R}^3)$. The aim of this question is to show that*
$$\int_{\mathbb{R}^3} \frac{|u(x)|^2}{|x|}\,dx \leq 2\|u\|_{L^2(\mathbb{R}^3)}\,\|\nabla u\|_{L^2(\mathbb{R}^3)}. \tag{2.4}$$
   (a) *Recall the reason for which there exists a sequence of functions $\psi_n \in \mathcal{D}(\mathbb{R}^3)$ that converges to u in $H^1(\mathbb{R}^3)$.*

   (b) *Let $\phi \in \mathcal{D}(\mathbb{R}^3)$. Show that $\dfrac{\phi(x)}{|x|} \in L^2(\mathbb{R}^3)$ and that*
$$\lim_{n\to\infty}\int_{\mathbb{R}^3} \frac{\psi_n(x)\phi(x)}{|x|}\,dx = \int_{\mathbb{R}^3} \frac{u(x)\phi(x)}{|x|}\,dx.$$
   (c) *Using the inequality* (2.2)*, show that the sequence of functions $\dfrac{\psi_n(x)}{|x|}$ is a Cauchy sequence in $L^2(\mathbb{R}^3)$.*

(d) *Deduce that there exists* $v \in L^2(\mathbb{R}^3)$ *such that* $\dfrac{\psi_n(x)}{|x|}$ *converges to* v *in* $L^2(\mathbb{R}^3)$*.*

(e) *Using the questions* (3a)–(3d)*, show that* $\dfrac{u(x)}{|x|} = v(x)$ *in* $L^2(\mathbb{R}^3)$*.*

(f) *Show that*

$$\left| \int_{\mathbb{R}^3} \frac{\psi_n(x)^2}{|x|} \, dx - \int_{\mathbb{R}^3} \frac{u(x)^2}{|x|} \, dx \right| \leq \left\| \frac{\psi_n - u}{|x|} \right\|_{L^2(\mathbb{R}^3)} \|\psi_n + u\|_{L^2(\mathbb{R}^3)}$$

and deduce that

$$\lim_{n \to \infty} \int_{\mathbb{R}^3} \frac{\psi_n(x)^2}{|x|} \, dx = \int_{\mathbb{R}^3} \frac{u(x)^2}{|x|} \, dx.$$

(g) *Using inequality* (2.3) *on* ψ_n *and taking the limit* $n \to \infty$*, show that* (2.4) *holds.*

4. *Consider the optimization problem*

$$I = \inf \left\{ \mathcal{E}(u); \ u \in \mathcal{A} \right\}.$$

Using (2.4)*, show that* $I > -\infty$*.*

Exercise 2.24 (Hardy inequality)**.** *Consider the vector space*

$$W^1(\mathbb{R}^3) = \left\{ u \in L^2_{\text{loc}}(\mathbb{R}^3), \ \frac{u(x)}{|x|} \in L^2(\mathbb{R}^3), \ \nabla u \in \left(L^2(\mathbb{R}^3) \right)^3 \right\}.$$

We recall that $L^2_{\text{loc}}(\mathbb{R}^3)$ *is the vector space of functions* $u : \mathbb{R}^3 \mapsto \mathbb{R}$ *that are measurable and such that* $\displaystyle\int_K |u|^2 < \infty$ *for any compact set* $K \subset \mathbb{R}^3$*. Since* $L^2_{\text{loc}}(\mathbb{R}^3) \subset L^1_{\text{loc}}(\mathbb{R}^3)$*, we have that* ∇u *is well-defined in the sense of distributions.*

For any u *and* v *in* $W^1(\mathbb{R}^3)$*, we set*

$$(u, v)_{W^1} = \int_{\mathbb{R}^3} \frac{u(x)\, v(x)}{|x|^2} \, dx + \int_{\mathbb{R}^3} \nabla u \cdot \nabla v. \tag{2.5}$$

1. *Show that, for any* u *and* v *in* $W^1(\mathbb{R}^3)$*, the right-hand side of* (2.5) *is well-defined, and that* $(\cdot, \cdot)_{W^1}$ *defines a scalar product on* $W^1(\mathbb{R}^3)$*.*

2. *Show that* $\mathcal{D}(\mathbb{R}^3) \subset W^1(\mathbb{R}^3)$*.*

3. *Using the question* (3e) *of Exercise 2.23, show that* $H^1(\mathbb{R}^3) \subset W^1(\mathbb{R}^3)$*.*

4. *We admit that* $W^1(\mathbb{R}^3)$*, endowed with the above scalar product, is a Hilbert space and that* $\mathcal{D}(\mathbb{R}^3)$ *is dense in* $W^1(\mathbb{R}^3)$ *(when the space* $W^1(\mathbb{R}^3)$ *is endowed with the norm* $\| \cdot \|_{W^1}$ *associated to the scalar product* $(\cdot, \cdot)_{W^1}$*).*

   We recall (see Exercise 2.22) that, for any $\psi \in \mathcal{D}(\mathbb{R}^3)$*,*

$$\int_{\mathbb{R}^3} \frac{\psi(x)^2}{|x|^2} \, dx \leq 4 \int_{\mathbb{R}^3} |\nabla \psi|^2.$$

   Show that, for any $u \in W^1(\mathbb{R}^3)$*,*

$$\left\| \frac{u}{|x|} \right\|_{L^2} \leq 2 \, \|\nabla u\|_{L^2}.$$

Chapter 3

Linear elliptic boundary value problems

This chapter is devoted to the mathematical analysis of linear elliptic boundary value problems. A boundary value problem is usually composed of

- one (or several) Partial Differential Equation (PDE),
- boundary conditions (or asymptotic conditions if the domain on which the problem is posed is the whole space $\mathbb{R}^d$).

For time-dependent problems (which will not be considered here), one has to additionally consider initial conditions.

We restrict ourselves to linear problems, where the solution depends linearly (or in an affine manner) on the right-hand side of the PDE and on the imposed boundary conditions.

The term *elliptic* designates a class of partial differential equations, a typical example in that class being the Poisson equation $-\Delta u = f$, which is studied in Section 3.3. Other classes of PDEs, that will not be considered here, are *parabolic* problems, such as the heat equation

$$\frac{\partial u}{\partial t} - \Delta u = f \qquad \text{(complemented by appropriate boundary and initial conditions)},$$

or *hyperbolic* problems (such as the wave equation)

$$\frac{\partial^2 u}{\partial t^2} - \Delta u = f \qquad \text{(complemented by appropriate boundary and initial conditions)}.$$

In this chapter, we show that the Poisson equation

$$\left\{ \begin{array}{ll} -\Delta u = f & \text{in } \mathcal{D}'(\Omega), \\ u = 0 & \text{on } \partial\Omega, \end{array} \right. \tag{3.1}$$

is well-posed in $H^1(\Omega)$, in the sense that it has a unique solution, and that this solution depends in a continuous manner on the right-hand side f. Problem (3.1) appears e.g. in electrostatic and is often written in the form $-\Delta V = \rho$, where V is the electric potential and ρ is the distribution of electric charges. Problem (3.1) is also encountered in mechanics (it describes the vertical displacement u of a membrane submitted to some forces f and clamped at its boundary). Besides its own interest, Problem (3.1) is also interesting because many boundary value problems in the engineering sciences are (more or less elaborated) variants of (3.1).

To show the well-posedness of (3.1), we proceed as follows:

1. First, we show that Problem (3.1) is equivalent to another problem, written in the form of a variational formulation of the type

$$\left\{\begin{array}{l}\text{Find } u \in V \text{ such that} \\ \forall v \in V, \quad a(u,v) = b(v)\end{array}\right. \tag{3.2}$$

for some appropriate choices of a Hilbert space V, a bilinear form a and a linear form b (see Section 3.3 for details in the particular case of Problem (3.1)). Problems (3.1) and (3.2) are equivalent in the sense that any solution in $H^1(\Omega)$ of Problem (3.1) is a solution to Problem (3.2), and the converse is also true.

2. Second, using the *Lax-Milgram theorem*, we show that Problem (3.2) is well-posed (i.e. has a unique solution).

Besides the specific problem (3.1), other problems are also studied in this chapter. Several exercises guide the reader through more and more complex examples.

As pointed out above, this chapter is devoted to the *mathematical analysis* of some boundary value problems, establishing that these problems are well-posed. Questions related to the approximation of the exact solutions, to the estimation of the error, and more generally to the *numerical analysis* of these problems, will be addressed in Chapter 5.

We also underline that we restrict ourselves, in this chapter and in Chapter 5, to *coercive problems*. Non-coercive problems will be studied in Chapters 4 and 6, from the mathematical and numerical analysis standpoints, respectively.

3.1 Lax-Milgram theorem (symmetric version)

Let H be a vector space endowed with a norm $\|\cdot\|$. We recall that a linear form b on H is continuous if and only if there exists c such that

$$\forall v \in H, \quad |b(v)| \leq c\|v\|.$$

Likewise, a bilinear form a on $H \times H$ is continuous if and only if there exists M such that

$$\forall (u,v) \in H \times H, \quad |a(u,v)| \leq M\|u\| \, \|v\|.$$

Definition 3.1. *Let H be a Hilbert space and a be a bilinear form on H. We say that a is* coercive *on H if there exists a real number $\alpha > 0$ such that*

$$\forall u \in H, \qquad a(u,u) \geq \alpha \, \|u\|^2.$$

Theorem 3.2 (Lax-Milgram theorem, symmetric version)**.** *Let H be a Hilbert space, a be a bilinear form on H, symmetric, continuous and coercive. Let b be a continuous linear form on H. Then the problem*

$$\left\{\begin{array}{l}\text{Find } u \in H \text{ such that} \\ \forall v \in H, \qquad a(u,v) = b(v)\end{array}\right. \tag{3.3}$$

has a unique solution. This unique solution is also the unique solution to the minimization problem

$$\left\{\begin{array}{l}\text{Find } u \in H \text{ such that} \\ J(u) = \inf\limits_{v \in H} J(v),\end{array}\right. \tag{3.4}$$

where the functional $J(v)$ (the so-called energy functional) is defined by

$$J(v) = \frac{1}{2}a(v,v) - b(v).$$

Proof. The proof falls in two steps.

Step 1 (well-posedness of (3.3)**):** Introduce

$$(u, v)_a = a(u, v).$$

It is clear that $(\cdot, \cdot)_a$ is a scalar product on H. In addition, the norms $\|\cdot\|_H$ and $\|\cdot\|_a$ are equivalent. Indeed, using the coercivity and the continuity of the bilinear form a, we obtain that there exists $\alpha > 0$ and M such that

$$\forall v \in H, \qquad \alpha\|v\|^2 \leq \|v\|_a^2 = a(v, v) \leq M\|v\|^2.$$

The space H, which is complete for the norm $\|\cdot\|$, is thus also complete for the norm $\|\cdot\|_a$. The space H, endowed with the scalar product $(\cdot, \cdot)_a$, is hence a Hilbert space. The bilinear form b is continuous for the norm $\|\cdot\|$, thus also for the equivalent norm $\|\cdot\|_a$. Using the Riesz theorem, we thus obtain that there exists a unique $u \in H$ such that

$$\forall v \in H, \qquad a(u, v) = (u, v)_a = b(v).$$

Step 2 (well-posedness of (3.6)**):** Consider now the solution u to (3.3) and let $v \in H$. We set $h = v - u$ and compute, using the symmetry of a, that

$$\begin{aligned} J(v) &= J(u + h) \\ &= \frac{1}{2}a(u + h, u + h) - b(u + h) \\ &= J(u) + a(u, h) - b(h) + \frac{1}{2}a(h, h) \\ &= J(u) + \frac{1}{2}a(h, h) \\ &\geq J(u). \end{aligned}$$

Hence u is a solution to (3.4).

Conversely, consider a solution u to (3.4). Then, for any $v \in H$ and any $\lambda \in \mathbb{R}$, we have

$$J(u) \leq J(u + \lambda v) = J(u) + \lambda\left(a(u, v) - b(v)\right) + \frac{1}{2}\lambda^2\, a(v, v).$$

Hence, for any $\lambda \in \mathbb{R}$,

$$\lambda(a(u, v) - b(v)) + \frac{1}{2}\lambda^2 a(v, v) \geq 0.$$

This implies that $a(u, v) = b(v)$, an equality that holds for any $v \in H$. Hence u is a solution to (3.3). □

3.2 A first symmetric linear elliptic boundary value problem

Let Ω be a subset of $\mathbb{R}^d$, λ be a positive real number ($\lambda > 0$) and $f \in L^2(\Omega)$. We look for $u \in H^1(\Omega)$ solution to

$$\left\{ \begin{array}{ll} -\Delta u + \lambda u = f & \text{in } \mathcal{D}'(\Omega), \\ u = 0 & \text{on } \partial\Omega. \end{array} \right. \tag{3.5}$$

Remark 3.3. *All what follows can be extended to the case $f \in H^{-1}(\Omega)$. We have chosen $f \in L^2(\Omega)$ only for the sake of simplicity.*

3.2.1 Variational formulation of (3.5)

We set

- $H = H_0^1(\Omega)$,
- $a(u,v) = \displaystyle\int_\Omega \nabla u \cdot \nabla v + \lambda \int_\Omega u\, v$,
- $b(v) = \displaystyle\int_\Omega f\, v$,

and consider the problem

$$\begin{cases} \text{Find } u \in H \text{ such that} \\ \forall v \in H, \qquad a(u,v) = b(v). \end{cases} \tag{3.6}$$

Proposition 3.4. *Problems* (3.5) *and* (3.6) *are equivalent.*

Proof. We first show that a is continuous on $H \times H$ and that b is continuous on H:

$$\forall v \in H = H_0^1(\Omega), \qquad |b(v)| = \left| \int_\Omega f\, v \right| \leq \|f\|_{L^2}\, \|v\|_{H^1},$$

which means that b is continuous on H. Furthermore, we have

$$\begin{aligned} \forall (u,v) \in H \times H, \qquad |a(u,v)| &= \left| \int_\Omega \nabla u \cdot \nabla v + \lambda \int_\Omega u\, v \right| \\ &\leq \left| \int_\Omega \nabla u \cdot \nabla v \right| + \lambda \left| \int_\Omega u\, v \right| \\ &\leq \left(\int_\Omega |\nabla u|^2 \right)^{1/2} \left(\int_\Omega |\nabla v|^2 \right)^{1/2} + \lambda \left(\int_\Omega u^2 \right)^{1/2} \left(\int_\Omega v^2 \right)^{1/2} \\ &\leq (1+\lambda) \|u\|_{H^1}\, \|v\|_{H^1}, \end{aligned}$$

which means that a is continuous on $H \times H$.

We now show that (3.6) implies (3.5). Let u be a solution to (3.6). We already have that $u \in H_0^1(\Omega)$, hence $u \in H^1(\Omega)$ and $u = 0$ on $\partial\Omega$. Let now $\phi \in \mathcal{D}(\Omega)$. We have $\phi \in H_0^1(\Omega)$, thus

$$\begin{aligned} \langle f, \phi \rangle &= \int_\Omega f\, \phi \\ &= b(\phi) \\ &= a(u, \phi) \\ &= \int_\Omega \nabla u \cdot \nabla \phi + \lambda \int_\Omega u\, \phi \\ &= \langle \nabla u, \nabla \phi \rangle + \lambda \langle u, \phi \rangle \\ &= \langle -\Delta u + \lambda u, \phi \rangle. \end{aligned}$$

Therefore $-\Delta u + \lambda u = f$ in $\mathcal{D}'(\Omega)$. Thus u is a solution to (3.5).

Conversely, let u be a solution to (3.5). We already have that $u \in H^1(\Omega)$ and $u = 0$ on $\partial\Omega$, thus $u \in H^1_0(\Omega)$. Let $\phi \in \mathcal{D}(\Omega)$. We have

$$\begin{aligned} b(\phi) &= \int_\Omega f\,\phi \\ &= \langle f, \phi\rangle \\ &= \langle -\Delta u + \lambda u, \phi\rangle \\ &= \langle \nabla u, \nabla\phi\rangle + \lambda\langle u, \phi\rangle \\ &= a(u, \phi). \end{aligned}$$

Hence,

$$\forall \phi \in \mathcal{D}(\Omega), \qquad a(u,\phi) = b(\phi).$$

Since $a(u,\cdot)$ and b are continuous on $H = H^1_0(\Omega)$ and since $\mathcal{D}(\Omega)$ is dense in $H^1_0(\Omega)$, we conclude that

$$\forall v \in H, \qquad a(u,v) = b(v).$$

Therefore, u is a solution to (3.6). □

3.2.2 Checking the assumptions of the Lax-Milgram theorem

We now check that Problem (3.6) falls within the assumptions of the Lax-Milgram theorem. We verify that

- $H = H^1_0(\Omega)$ is a Hilbert space;
- b is linear and continuous on H;
- a is bilinear, symmetric and continuous on $H \times H$. In addition, a is coercive: for any $v \in H$,

$$\begin{aligned} a(v,v) &= \int_\Omega |\nabla v|^2 + \lambda \int_\Omega v^2 \\ &\geq \min(1,\lambda)\left(\int_\Omega |\nabla v|^2 + \int_\Omega v^2\right) \\ &= \min(1,\lambda)\|v\|^2_{H^1}. \end{aligned}$$

We are thus in position to state that Problem (3.6) has a unique solution, in view of the Lax-Milgram theorem 3.2. The problems (3.5) and (3.6) are equivalent. The problem (3.5) thus has a unique solution.

We note that the energy functional associated to Problem (3.5) is

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

3.3 Poisson equation on a bounded open set

Let Ω be a bounded subset of $\mathbb{R}^d$ and $f \in L^2(\Omega)$. We look for $u \in H^1(\Omega)$ solution to

$$\left\{ \begin{array}{ll} -\Delta u = f & \text{in } \mathcal{D}'(\Omega), \\ u = 0 & \text{on } \partial\Omega. \end{array} \right. \tag{3.7}$$

Remark 3.5. *Again, as in Section 3.2, it is possible to take $f \in H^{-1}(\Omega)$ rather than $f \in L^2(\Omega)$.*

3.3.1 Variational formulation of (3.7)

We set

- $H = H_0^1(\Omega)$,
- $a(u,v) = \displaystyle\int_\Omega \nabla u \cdot \nabla v$,
- $b(v) = \displaystyle\int_\Omega f\, v$,

and consider the problem

$$\left\{ \begin{array}{l} \text{Find } u \in H \text{ such that} \\ \forall v \in H, \qquad a(u,v) = b(v). \end{array} \right. \tag{3.8}$$

Proposition 3.6. *Problems* (3.7) *and* (3.8) *are equivalent.*

Proof. The proof follows the same steps as the proof of Proposition 3.4. □

3.3.2 Checking the assumptions of the Lax-Milgram theorem

We now check that Problem (3.8) falls within the assumptions of the Lax-Milgram theorem. We verify that

- $H = H_0^1(\Omega)$ is a Hilbert space;
- b is linear and continuous on H;
- a is bilinear, symmetric and continuous on $H \times H$.

We are left with showing that a is coercive. Using the Poincaré inequality (see Theorem 2.7), we see that, for any $v \in H_0^1(\Omega)$,

$$\|v\|_{H^1}^2 = \|v\|_{L^2}^2 + \|\nabla v\|_{L^2}^2 \leq (1 + C_\Omega^2)\ \|\nabla v\|_{L^2}^2 = (1 + C_\Omega^2)\ a(v,v).$$

Hence

$$a(v,v) \geq \frac{1}{1 + C_\Omega^2}\ \|v\|_{H^1}^2.$$

In view of Lax-Milgram theorem 3.2, we deduce that Problem (3.8) has a unique solution. The problems (3.7) and (3.8) are equivalent. The problem (3.7) thus has a unique solution.

We note that the energy functional associated to Problem (3.7) is

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

3.4 Lax-Milgram theorem

The symmetry assumption on the bilinear form a is restrictive as several boundary value problems are not symmetric. This is in particular the case when some *advection* (i.e. transport) terms are present in the problem. It turns out that the symmetry assumption can be removed.

Theorem 3.7 (Lax-Milgram theorem)**.** *Let H be a Hilbert space, a be a bilinear form on H, continuous and coercive. Let b be a continuous linear form on H. Then the problem*

$$\left\{\begin{array}{ll}\text{Find } u \in H \text{ such that} & \\ \forall v \in H, & a(u,v) = b(v)\end{array}\right. \tag{3.9}$$

has a unique solution.

Of course, the symmetric version (Theorem 3.2) is a particular case of this more general version. Note also that there is no minimization problem associated to the variational formulation (3.9), in contrast to the symmetric case.

Remark 3.8. *The proof of Theorem 3.7 in the finite dimensional case is easy. Consider the Hilbert space $H = \mathbb{R}^n$ and the bilinear form $a(u,v) = v^T Au$ for any u and v in H, where A is a $n \times n$ matrix.*

We first show that Ker $A = \{0\}$. Let $u \in H$ such that $Au = 0$. Then $a(u,u) = u^T Au = 0$. Thanks to the coercivity assumption on a, this implies that $u = 0$.

Using the fact that dim Ker A + dim Im A = dim H, we get that dim Im A = dim H, and hence the matrix A is invertible.

Proof of Theorem 3.7. Consider the mapping

$$\begin{array}{rcl}\Phi : H & \longrightarrow & H \\ u & \mapsto & b \text{ such that } a(u,\cdot) = \langle b, \cdot \rangle_H.\end{array}$$

Note that we have used the Riesz theorem to represent the continuous and linear form $v \in H \mapsto a(u,v) \in \mathbb{R}$ by an element of H, denoted $b = \Phi(u)$. The mapping Φ is of course linear, and it is also continuous:

$$\|b\|^2 = \langle b, b \rangle_H = a(u,b) \leq M\|u\| \, \|b\|$$

hence $\|\Phi(u)\| = \|b\| \leq M\|u\|$. We wish to show that Φ is bijective. The proof falls in three steps.

Step 1: Φ is injective: Let $u \in H$ such that $\Phi(u) = 0$. For any $v \in H$, we have $a(u,v) = 0$, thus $a(u,u) = 0$. Using the coercivity of a, we deduce that

$$0 = a(u,u) \geq \alpha\|u\|^2.$$

Hence $u = 0$.

Step 2: Φ is surjective: Let $V = \mathrm{Im}(\Phi) \subset H$. We show that $V = H$.

(a) Let us first show that $X = \mathrm{Im}(\Phi)$ is closed. Let $u \in H$, $u \neq 0$. We have

$$\sup_{v \in H,\, v \neq 0} \frac{\langle \Phi(u), v \rangle_H}{\|v\|} = \sup_{v \in H,\, v \neq 0} \frac{a(u,v)}{\|v\|} \geq \frac{a(u,u)}{\|u\|} \geq \alpha\|u\|.$$

In addition, for any $v \neq 0$, we have

$$\frac{\langle \Phi(u), v \rangle_H}{\|v\|} \leq \|\Phi(u)\|.$$

We thus obtain that

$$\|\Phi(u)\| \geq \alpha\|u\|$$

and this estimate also holds if $u = 0$. Consider now a sequence $(b_n)_{n\in\mathbb{N}}$ in X that converges in H to some $b \in H$. We want to show that $b \in X$. Let $u_n \in H$ such that $\Phi(u_n) = b_n$. Since $(b_n)_{n\in\mathbb{N}}$ is a Cauchy sequence, we see that $(u_n)_{n\in\mathbb{N}}$ is also a Cauchy sequence. Indeed,

$$\|b_p - b_q\| = \|\Phi(u_p) - \Phi(u_q)\| = \|\Phi(u_p - u_q)\| \geq \alpha\|u_p - u_q\|.$$

Therefore, $(u_n)_{n\in\mathbb{N}}$ converges in H to some $u \in H$. In addition, Φ is continuous, so

$$b_n = \Phi(u_n) \longrightarrow \Phi(u) \quad \text{in } H.$$

Hence $b = \Phi(u) \in X$, by uniqueness of the limit in H. The vector space X is hence closed.

(b) Let $b_0 \in H$. Since X is closed in H, we can consider the orthogonal projection of b_0 on X, that we denote $b_1 \in X$. For any $v \in X$, we have $\langle v, b_0 - b_1\rangle_H = 0$. Hence, for any $w \in H$, we have

$$0 = \langle \Phi(w), b_0 - b_1\rangle_H = a(w, b_0 - b_1).$$

Taking $w = b_0 - b_1$ and using the coercivity of a, we deduce that $b_0 = b_1 \in X$, which implies that $H = X$.

Step 3: Conclusion: Since b is a continuous linear form on H, it can be represented by some $f \in H$. The problem (3.9) can thus be recast as finding $u \in H$ such that $a(u,\cdot) = \langle f, \cdot\rangle_H$, that is finding $u \in H$ such that $\Phi(u) = f$. In view of the above two steps, the mapping Φ is bijective, so this problem has a unique solution. □

3.5 A non-symmetric boundary value problem

Let Ω be a bounded subset of $\mathbb{R}^d$ and $f \in L^2(\Omega)$. Let $c : \Omega \to \mathbb{R}^d$ be a vector field of class $C^1(\overline{\Omega})$ which is divergence-free:

$$\operatorname{div} c = \sum_{i=1}^d \frac{\partial c_i}{\partial x_i} = 0 \quad \text{in } \Omega.$$

We look for $u \in H^1(\Omega)$ solution to

$$\begin{cases} -\Delta u + c \cdot \nabla u = f & \text{in } \mathcal{D}'(\Omega), \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{3.10}$$

Note that the product $c \cdot \nabla u$ is well-defined in $\mathcal{D}'(\Omega)$. This comes from the fact that (i) ∇u is a function in $L^2(\Omega)$, hence in $L^1(\Omega)$ since Ω is bounded, and (ii) c is a bounded function (it is a continuous function on a compact set). The product $c \cdot \nabla u$ is thus in $L^1(\Omega)$, and therefore in $\mathcal{D}'(\Omega)$.

3.5.1 Variational formulation of (3.10)

We set

- $H = H^1_0(\Omega)$,
- $a(u,v) = \displaystyle\int_\Omega \nabla u \cdot \nabla v + \int_\Omega (c \cdot \nabla u)\, v$,
- $b(v) = \displaystyle\int_\Omega f\, v$,

and consider the problem

$$\begin{cases} \text{Find } u \in H \text{ such that} \\ \forall v \in H, \qquad a(u,v) = b(v). \end{cases} \tag{3.11}$$

Proposition 3.9. *Problems* (3.10) *and* (3.11) *are equivalent.*

Proof. In the proof of Proposition 3.4, we have already shown the continuity of b. We now show the continuity of a. For any $(u, v) \in H \times H$, we compute

$$\begin{aligned} |a(u,v)| &= \left| \int_\Omega \nabla u \cdot \nabla v + \int_\Omega (c \cdot \nabla u)\, v \right| \\ &\leq \left| \int_\Omega \nabla u \cdot \nabla v \right| + \left| \int_\Omega (c \cdot \nabla u)\, v \right| \\ &\leq \|\nabla u\|_{L^2} \|\nabla v\|_{L^2} + \sup_{\overline{\Omega}} |c| \; \|\nabla u\|_{L^2} \|v\|_{L^2} \\ &\leq (1 + \sup_{\overline{\Omega}} |c|) \; \|u\|_{H^1} \|v\|_{H^1}. \end{aligned}$$

Thus a is continuous in $H \times H$.

The sequel of the proof follows the arguments of the proof of Proposition 3.4. □

3.5.2 Checking the assumptions of the Lax-Milgram theorem

We now check that Problem (3.11) falls within the assumptions of the Lax-Milgram theorem. We verify that

- $H = H^1_0(\Omega)$ is a Hilbert space;
- b is linear and continuous on H;
- a is bilinear and continuous.

We are left with showing that a is coercive. We note that, for any $\phi \in \mathcal{D}(\Omega)$,

$$\int_\Omega (c \cdot \nabla \phi)\, \phi = \int_\Omega c \cdot \nabla \left(\frac{\phi^2}{2} \right) = \int_{\partial\Omega} (c \cdot n) \left(\frac{\phi^2}{2} \right) - \int_\Omega \left(\frac{\phi^2}{2} \right) (\text{div } c)$$

and the first term in the right-hand side vanishes as ϕ vanishes in the neighborhood of $\partial\Omega$. We thus have, using that c is divergence-free and the Poincaré inequality,

$$\begin{aligned} a(\phi,\phi) &= \int_\Omega |\nabla \phi|^2 + \int_\Omega (c \cdot \nabla \phi)\phi \\ &= \int_\Omega |\nabla \phi|^2 - \frac{1}{2} \int_\Omega \left(\frac{\phi^2}{2} \right) (\text{div } c) \\ &= \int_\Omega |\nabla \phi|^2 \\ &\geq \frac{1}{1 + C_\Omega^2} \|\phi\|_{H^1}. \end{aligned}$$

By continuity of a and density of $\mathcal{D}(\Omega)$ in $H = H^1_0(\Omega)$, we deduce that a is coercive on H.

In view of Lax-Milgram theorem 3.7, we deduce that Problem (3.11) has a unique solution. The problems (3.10) and (3.11) are equivalent. The problem (3.10) thus has a unique solution.

3.6 Exercises

Exercise 3.10. *Let Ω be a bounded open subset of $\mathbb{R}^d$. We consider*

- *a matrix field $A : \Omega \longrightarrow \mathbb{R}^{d \times d}$ that belongs to $(L^\infty(\Omega))^{d \times d}$,*
- *two vector fields b and c in $(L^\infty(\Omega))^d$,*

- *a scalar-valued field $d \in L^\infty(\Omega)$.*

Write the variational formulation of the following boundary value problem: Find $u \in H^1_0(\Omega)$ such that

$$-\mathrm{div}\ (A\nabla u) + \mathrm{div}\ (bu) + c \cdot \nabla u + du = f \qquad \textit{in } \mathcal{D}'(\Omega).$$

We now assume that the matrix field is uniformly coercive on Ω, that is that there exists $\alpha > 0$ such that

$$\forall \xi \in \mathbb{R}^d, \qquad \xi^T A(x)\xi \geq \alpha \xi^T \xi \quad \textit{almost everywhere on } \Omega.$$

Identify conditions on A, b, c and d under which the above boundary value problem has a unique solution in $H^1_0(\Omega)$.

Exercise 3.11 (Regularization)**.** *Let Ω be an open subset of $\mathbb{R}^d$ and $f \in L^2(\Omega)$.*

1. *Let $\varepsilon > 0$. We look for $u_\varepsilon \in H^1(\Omega)$ solution to*

$$\begin{cases} -\varepsilon\, \Delta u_\varepsilon + u_\varepsilon = f & \textit{in } \mathcal{D}'(\Omega), \\ u_\varepsilon = 0 & \textit{on } \partial\Omega. \end{cases}$$

Show that this problem is well-posed and that

$$\forall \varepsilon, \quad \|u_\varepsilon\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)}.$$

2. *We denote u_n the solution in $H^1(\Omega)$ to*

$$\begin{cases} -\dfrac{1}{n}\, \Delta u_n + u_n = f & \textit{in } \mathcal{D}'(\Omega), \\ u_n = 0 & \textit{on } \partial\Omega. \end{cases} \tag{3.12}$$

   (a) *Let $\phi \in \mathcal{D}(\Omega)$, $q \in \mathbb{N}^\star$ and $b_n = \langle\, u_n, \phi\, \rangle$. Show that*

$$|b_{n+q} - b_n| \ \leq \ \frac{1}{n+q}\, \|u_{n+q}\|_{L^2(\Omega)} \|\Delta\phi\|_{L^2(\Omega)} + \frac{1}{n}\, \|u_n\|_{L^2(\Omega)} \|\Delta\phi\|_{L^2(\Omega)}$$

   and that

$$|b_{n+q} - b_n| \ \leq \ \frac{2}{n}\, \|f\|_{L^2(\Omega)} \|\Delta\phi\|_{L^2(\Omega)}.$$

   (b) *Deduce that u_n converges in $\mathcal{D}'(\Omega)$ to some distribution that we denote T.*

   (c) *Using* (3.12)*, show that $T = f$.*

3. *We now assume that $f \in \mathcal{D}(\Omega)$. Let $v_n = u_n - f$.*

   (a) *Write the problem that v_n satisfies.*

   (b) *Using its variational formulation, show that*

$$\|\nabla v_n\|_{L^2(\Omega)} \ \leq \ \|\nabla f\|_{L^2(\Omega)}.$$

   (c) *Deduce that*

$$\lim_{n\to\infty} \|v_n\|_{L^2(\Omega)} = 0$$

   (d) *What does this imply on the convergence of u_n?*

Exercise 3.12. *Let $u_0 \in L^2(\mathbb{R}^d)$ and $\tau > 0$. Consider the sequence $u_n \in H^1(\mathbb{R}^d)$ recursively defined by: for all $n \geq 0$,*

$$\frac{u_{n+1} - u_n}{\tau} = \Delta u_{n+1}. \tag{3.13}$$

1. *Show that the sequence $\{u_n\}_{n\geq 1}$ is well-defined.*
2. *In the case $d = 1$, consider u defined by*

$$u(x) = \begin{cases} \cos(\omega x) & \text{if } x \leq -2\pi/\omega, \\ 1 \quad \text{if } -2\pi/\omega \leq x \leq 2\pi/\omega, \\ \cos(\omega x) & \text{if } x \geq 2\pi/\omega, \end{cases}$$

   where $\omega = 1/\sqrt{\tau}$. Compute the second derivative (in the sense of distributions) of u.
3. *Using the previous question, build a function $u_0 \in L^2(\mathbb{R})$ such that the problem: Find $u_1 \in H^1(\mathbb{R})$ such that*

$$\frac{u_1 - u_0}{\tau} = u_0''$$

   does not have any solution.
4. *In which context does* (3.13) *appear? What conclusion can you draw from the above question?*

Exercise 3.13. *Let $f \in L^2(\mathbb{R}^2)$ and $\lambda \in \mathbb{R}$. Consider the problem:*

$$\begin{cases} \text{Find } u \in V \text{ such that} \\ \forall v \in V, \qquad a_\lambda(u, v) = b(v) \end{cases} \tag{3.14}$$

where $V = H^2(\mathbb{R}^2)$, $b(v) = \displaystyle\int_{\mathbb{R}^2} f\, v$ and

$$a_\lambda(u, v) = \int_{\mathbb{R}^2} \Delta u\, \Delta v + \lambda \int_{\mathbb{R}^2} \nabla u \cdot \nabla v + \int_{\mathbb{R}^2} u\, v.$$

1. *Show that, for any u and v in $H^2(\mathbb{R}^2)$, we have*

$$\int_{\mathbb{R}^2} \Delta u\, \Delta v = \sum_{i,j=1}^{2} \int_{\mathbb{R}^2} \frac{\partial^2 u}{\partial x_i \partial x_j} \frac{\partial^2 v}{\partial x_i \partial x_j}.$$

2. *Show that, if $\lambda > 0$, then Problem* (3.14) *is well-posed.*
3. *Show that, for any $\phi \in \mathcal{D}(\mathbb{R}^2)$,*

$$\int_{\mathbb{R}^2} |\nabla \phi|^2 = - \int_{\mathbb{R}^2} \phi\, \Delta \phi,$$

   and deduce from this that, for any u in $H^2(\mathbb{R}^2)$,

$$\int_{\mathbb{R}^2} |\nabla u|^2 \leq \frac{1}{2} \left(\int_{\mathbb{R}^2} u^2 + \int_{\mathbb{R}^2} |\Delta u|^2 \right).$$

   Deduce that, if $\lambda > -2$, then Problem (3.14) *is well-posed.*

Exercise 3.14 (Non homogeneous Dirichlet boundary conditions)**.** *Let Ω be a bounded open subset of $\mathbb{R}^d$, $f \in L^2(\Omega)$, $u_0 \in H^1(\Omega)$ (this assumption will be made stronger below). We consider the following problem:*

$$\left\{\begin{array}{l}\text{Find } u \in H^1(\Omega) \text{ such that}\\ -\Delta u = f \text{ in } \mathcal{D}'(\Omega),\\ u = u_0 \text{ on } \partial\Omega.\end{array}\right. \tag{3.15}$$

The functions u and u_0 are in $H^1(\Omega)$, thus they both have a trace on $\partial\Omega$. The last equation above simply means that the two traces are equal.

Note also that the function u_0 is defined on purpose on Ω, although only its trace on the boundary $\partial\Omega$ appears in (3.15)*.*

1. *Introduce $w = u - u_0$, and show that u is a solution to* (3.15) *if and only if w is a solution to*

$$\left\{\begin{array}{l}\text{Find } w \in H^1(\Omega) \text{ such that}\\ -\Delta w = g \text{ in } \mathcal{D}'(\Omega),\\ w = 0 \text{ in } \partial\Omega,\end{array}\right. \tag{3.16}$$

   for a distribution g that will be made precise. The interest of (3.16) *is that it is a problem with* homogeneous *boundary conditions.*

2. *We assume for now that $u_0 \in H^2(\Omega)$. Deduce that $g \in L^2(\Omega)$. Write a variational formulation (with unknown w) equivalent to* (3.16)*, and show the existence and uniqueness of a solution to* (3.15)*.*

3. *Using the Green formula, show that the solution u to* (3.15) *satisfies*

$$\left\{\begin{array}{l}u \in H^1(\Omega),\ u = u_0 \text{ on } \partial\Omega, \text{ and}\\ \forall \varphi \in H^1_0(\Omega), \quad \displaystyle\int_\Omega \nabla u \cdot \nabla \varphi = \int_\Omega f\varphi.\end{array}\right.$$

   Note that the solution u and the test function φ do not belong to the same space.

The following exercise aims at proving the Theorem 3.15 below, which is a functional inequality that belongs to the same family as the Poincaré inequality of Theorem 2.7:

Theorem 3.15. *Let $\Omega \subset \mathbb{R}^d$ be a bounded, open and connected subset of $\mathbb{R}^d$, and let β be a real number $\beta > 0$. There exists $C > 0$ such that*

$$\forall v \in H^1(\Omega), \quad \|\nabla v\|^2_{L^2(\Omega)} + \beta\|v\|^2_{L^2(\partial\Omega)} \geq C\|v\|^2_{H^1(\Omega)}. \tag{3.17}$$

Exercise 3.16. *Our aim is to prove Theorem 3.15. We proceed by contradiction and assume that there exists no constant C such that* (3.17) *holds.*

1. *Show that there exists a sequence $\{u_n\}_{n\in\mathbb{N}^\star}$ such that $\|u_n\|_{H^1(\Omega)} = 1$ and*

$$\|\nabla u_n\|^2_{L^2(\Omega)} + \beta\|u_n\|^2_{L^2(\partial\Omega)} \leq \frac{1}{n}. \tag{3.18}$$

   What can be said of the limits (when $n \to \infty$) of $\|\nabla u_n\|_{L^2(\Omega)}$ and of $\|u_n\|_{L^2(\partial\Omega)}$?

2. *We admit that the bound $\|u_n\|_{H^1(\Omega)} = 1$ implies that there exists $u_\star \in H^1(\Omega)$ and a subsequence $u_{n'}$ such that $u_{n'}$ converges to $u_\star$ in $L^2(\Omega)$.*

   Using (3.18)*, show that $\nabla u_\star = 0$. We then deduce from the fact that Ω is connected that $u_\star$ is a constant.*

3. *Show that $u_{n'}$ converges to $u_\star$ in $H^1(\Omega)$, and next that the constant $u_\star$ is different from 0.*

4. *Using that the trace mapping is continuous from $H^1(\Omega)$ to $L^2(\partial\Omega)$, deduce that $u_{n'}$ converges to $u_\star$ in $L^2(\partial\Omega)$.*

5. *Reach a contradiction.*

Exercise 3.17 (Robin boundary conditions)**.** *Let Ω be a bounded open subset of $\mathbb{R}^d$, $f \in L^2(\Omega)$, $g \in C^\infty(\partial\Omega)$ and a real number $\beta > 0$. We look for $u \in H^1(\Omega)$ solution to*

$$\left\{ \begin{array}{ll} -\Delta u = f & \text{in } \mathcal{D}'(\Omega) \\ \dfrac{\partial u}{\partial n} + \beta u = g & \text{on } \partial\Omega, \end{array} \right. \tag{3.19}$$

where $\dfrac{\partial u}{\partial n} = \nabla u \cdot n$ (n is the outward normal vector to $\partial\Omega$). The precise interpretation of the boundary condition will be done below.

1. *As a first step, we wish to write a variational formulation associated to* (3.19)*. We thus assume that all functions are smooth, that all integrations by parts are allowed, ...*

   Show that, if u is a sufficiently regular solution to (3.19)*, then, for any smooth function $v : \Omega \to \mathbb{R}$, we have*
   $$\int_\Omega \nabla u \cdot \nabla v + \beta \int_{\partial\Omega} u\, v = \int_\Omega f\, v + \int_{\partial\Omega} g\, v.$$

2. *We now establish and study a variational formulation of the problem. We recall that, for any function $u \in H^1(\Omega)$, the trace of u on the boundary $\partial\Omega$ is well-defined, that this is a function of class $L^2(\partial\Omega)$, and that there exists C (that only depends on Ω) such that*
   $$\forall u \in H^1(\Omega), \quad \|u\|_{L^2(\partial\Omega)} \leq C\|u\|_{H^1(\Omega)}.$$

   The above question leads to consider the bilinear form
   $$a(u,v) = \int_\Omega \nabla u \cdot \nabla v + \beta \int_{\partial\Omega} u\, v$$

   and the linear form
   $$b(v) = \int_\Omega f\, v + \int_{\partial\Omega} g\, v.$$

   (a) *Show that a and b are well-defined on $H^1(\Omega)$ and that they are continuous.*

   (b) *Show that a is coercive on $H^1(\Omega)$ (use Theorem 3.15). Check also that, if $\beta \leq 0$, then a is not coercive on $H^1(\Omega)$.*

   (c) *Deduce from the above questions that the problem*
   $$\left\{ \begin{array}{l} \text{Find } u \in H^1(\Omega) \text{ such that} \\ \forall v \in H^1(\Omega), \quad a(u,v) = b(v) \end{array} \right. \tag{3.20}$$
   is well-posed.

3. *We now go back to the boundary value problem, and establish in which sense the solution u to* (3.20) *satisfies the boundary value problem* (3.19)*.*

   (a) *Show that the solution u to* (3.20) *satisfies*
   $$-\Delta u = f \text{ in } \mathcal{D}'(\Omega). \tag{3.21}$$

(b) *We admit that, if g is sufficiently smooth, then the solution u to (3.20) is in $H^2(\Omega)$. We keep this assumption for the sequel of the exercise. Show that $\dfrac{\partial u}{\partial n} \in L^2(\partial\Omega)$.*

(c) *Multiplying (3.21) by $v \in H^1(\Omega)$ and using a Green formula, show that u is such that, for any $v \in H^1(\Omega)$, we have*

$$\int_{\partial\Omega} v \left(\frac{\partial u}{\partial n} + \beta u \right) = \int_{\partial\Omega} g\, v.$$

Deduce from this equality that u satisfies the boundary condition $\dfrac{\partial u}{\partial n} + \beta u = g$ (to that aim, admit that the image of the trace mapping $\gamma : H^1(\Omega) \to L^2(\partial\Omega)$ is dense in $L^2(\partial\Omega)$).

Exercise 3.18 (Poincaré-Wirtinger inequality)**.** *We show here a functional inequality that belongs to the same family as the Poincaré inequality of Theorem 2.7.*

1. *Show that there exists $C > 0$ such that, for any $\varphi \in C^\infty([0,1])$, we have*

$$\|\varphi - \langle \varphi \rangle\|_{L^2(0,1)} \leq C \|\varphi'\|_{L^2(0,1)},$$

where $\langle \varphi \rangle = \displaystyle\int_0^1 \varphi$ is the mean of φ. Hint: start by showing that there exists $x_0 \in [0,1]$ such that $\langle \varphi \rangle = \varphi(x_0)$.

2. *We admit that $C^\infty([0,1])$ is dense in $H^1(0,1)$. Show that there exists $C > 0$ such that*

$$\forall v \in H^1(0,1), \quad \|v - \langle v \rangle\|_{L^2(0,1)} \leq C \|v'\|_{L^2(0,1)}. \tag{3.22}$$

The result of Exercise 3.18, shown in dimension $d = 1$, can actually be generalized to the following statement.

Theorem 3.19. *Let $\Omega \subset \mathbb{R}^d$ be a bounded, open and connected subset of $\mathbb{R}^d$. There exists $C > 0$ such that*

$$\forall v \in H^1(\Omega), \quad \|v - \langle v \rangle\|_{L^2(\Omega)} \leq C \|\nabla v\|_{L^2(\Omega)}, \tag{3.23}$$

with $\langle v \rangle = \dfrac{1}{|\Omega|} \displaystyle\int_\Omega v$.

Exercise 3.20 (Neumann boundary conditions)**.** *Let Ω be a regular, bounded, open and connected subset of $\mathbb{R}^d$, $f \in L^2(\Omega)$ and $g \in C^\infty(\partial\Omega)$. We look for $u \in H^1(\Omega)$ solution to*

$$\left\{ \begin{array}{ll} -\Delta u = f & \text{in } \mathcal{D}'(\Omega) \\ \dfrac{\partial u}{\partial n} = g & \text{on } \partial\Omega, \end{array} \right. \tag{3.24}$$

where $\dfrac{\partial u}{\partial n} = \nabla u \cdot n$ (n is the outward normal vector to $\partial\Omega$). The precise interpretation of the boundary condition will be done below.

Verify that Problem (3.24) is ill-posed, in the sense that, if u is solution to (3.24), then $u + c$ is also solution, for any constant c.

We thus introduce the set

$$H^1_m = \left\{ u \in H^1(\Omega), \quad \frac{1}{|\Omega|} \int_\Omega u = 0 \right\}$$

of the H^1 functions with vanishing mean, and look for $u \in H^1_m(\Omega)$ that satisfies (3.24).

1. *As a first step, we wish to write a variational formulation associated to* (3.24)*. We thus assume that all functions are smooth, that all integrations by parts are allowed, ...*

   Show that, if u *is a sufficiently regular solution to* (3.24)*, then, for any smooth function* $v : \Omega \to \mathbb{R}$*, we have*
$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega fv + \int_{\partial\Omega} gv.$$
   Deduce from this computation that, if (3.24) *admits a smooth solution, then*
$$\int_\Omega f + \int_{\partial\Omega} g = 0. \tag{3.25}$$
   We will keep this assumption for the sequel of the exercise.

2. *We now establish and study a variational formulation of the problem. We recall that, for any function* $u \in H^1(\Omega)$*, the trace of* u *on the boundary* $\partial\Omega$ *is well-defined, that this is a function of class* $L^2(\partial\Omega)$*, and that there exists* C *(that only depends on* Ω*) such that*
$$\forall u \in H^1(\Omega), \quad \|u\|_{L^2(\partial\Omega)} \leq C\|u\|_{H^1(\Omega)}.$$
   The above question leads to consider the bilinear form
$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v$$
   and the linear form
$$b(v) = \int_\Omega fv + \int_{\partial\Omega} gv.$$
   (a) *Show that* $H^1_m(\Omega)$ *is a Hilbert space for the* H^1 *scalar product.*

   (b) *Show that* a *and* b *are well-defined on* $H^1_m(\Omega)$ *and that they are continuous.*

   (c) *Using Theorem 3.19, show that* a *is coercive on* $H^1_m(\Omega)$*.*

   (d) *Deduce from the above questions that the problem*
$$\begin{cases} \text{Find } u \in H^1_m(\Omega) \text{ such that} \\ \forall v \in H^1_m(\Omega), \quad a(u,v) = b(v) \end{cases} \tag{3.26}$$
   is well-posed.

3. *We now go back to the boundary value problem, and establish in which sense the solution* u *to* (3.26) *satisfies the boundary value problem* (3.24)*. We recall that we assume that* f *and* g *satisfy* (3.25)*.*

   (a) *Show that the solution* u *to* (3.26) *satisfies*
$$\forall v \in H^1(\Omega), \quad a(u,v) = b(v),$$
   where the functions v *can now be chosen in* $H^1(\Omega)$ *and not only in* $H^1_m(\Omega)$*.*

   (b) *Show that the solution* u *to* (3.26) *satisfies*
$$-\Delta u = f \text{ in } \mathcal{D}'(\Omega). \tag{3.27}$$
   (c) *We admit that, if* g *is sufficiently smooth, then the solution* u *to* (3.26) *is in* $H^2(\Omega)$*. We keep this assumption for the sequel of the exercise. Show that* $\dfrac{\partial u}{\partial n} \in L^2(\partial\Omega)$*.*

(d) *Multiplying* (3.27) *by* $v \in H^1(\Omega)$ *and using a Green formula, show that* u *is such that, for any* $v \in H^1(\Omega)$*, we have*

$$\int_{\partial\Omega} v \, \frac{\partial u}{\partial n} = \int_{\partial\Omega} g\, v.$$

Deduce from this equality that u *satisfies the boundary condition* $\dfrac{\partial u}{\partial n} = g$ *(to that aim, admit that the image of the trace mapping* $\gamma : H^1(\Omega) \to L^2(\partial\Omega)$ *is dense in* $L^2(\partial\Omega)$*).*

Exercise 3.21 (Periodic boundary conditions)**.** *Consider the open set* $Q = \left(-\dfrac{1}{2}, \dfrac{1}{2}\right)^d$*, a function* $f \in L^2(Q)$ *and the problem*

$$\left\{ \begin{array}{l} \textit{Find } u \in H^1(Q) \textit{ such that} \\ -\Delta u = f \textit{ in } \mathcal{D}'(Q), \\ u \textit{ is periodic at the boundary of } Q. \end{array} \right. \tag{3.28}$$

This problem is ill-posed. Of course, if u *is a solution, then* $u + c$ *is also a solution for any constant* c*. But the situation is actually worse. Consider the particular case* $d = 2$*,* $f = 0$ *and* $u(x, y) = x^2 - y^2$*: this function is a solution to* (3.28)*, just as* $u(x, y) = 0$*. The sequel of this exercise aims at properly setting the problem.*

1. ***First variational formulation.***

   We recall that

$$L^2_{\rm per}(Q) = \left\{ u \in L^2_{\rm loc}(\mathbb{R}^d), \quad \forall k \in \mathbb{Z}^d, \ u(x + k) = u(x) \textit{ almost everywhere} \right\}$$

   and that

$$H^1_{\rm per}(Q) = \left\{ u \in L^2_{\rm per}(Q), \quad \frac{\partial u}{\partial x_i} \in L^2_{\rm per}(Q) \textit{ for any } 1 \le i \le d \right\}.$$

   To remove the indetermination of a solution by the addition of a constant, we consider the space

$$H^1_{\rm per,m}(Q) = \left\{ u \in H^1_{\rm per}(Q), \quad \int_Q u = 0 \right\}.$$

   (a) *We recall that* $H^1_{\rm per}(Q)$ *is a Hilbert space for the scalar product*

$$(u, w) = \int_Q u\, w + \int_Q \nabla u \cdot \nabla w.$$

   Show that $H^1_{\rm per,m}(Q)$ *is also a Hilbert space for the same scalar product.*

   (b) *On the space* $H^1_{\rm per,m}(Q)$*, we define the bilinear form*

$$a(u, v) = \int_Q \nabla u \cdot \nabla v$$

   and the linear form

$$b(v) = \int_Q f\, v.$$

   Show that a *and* b *are continuous, and that* a *is coercive on* $H^1_{\rm per,m}(Q)$ *(to that aim, use Theorem 3.19).*

(c) *Deduce from the above questions that the problem*

$$\left\{\begin{array}{l}\textit{Find } u\in H^1_{\mathrm{per,m}}(Q) \textit{ such that}\\ \forall v\in H^1_{\mathrm{per,m}}(Q),\quad a(u,v)=b(v)\end{array}\right. \tag{3.29}$$

is well-posed.

(d) *Let* $\varphi\in\mathcal{D}(Q)$. *Is it possible to choose* φ *as test function in* (3.29)*?*

(e) *Let* $\varphi\in\mathcal{D}(Q)$ *and set* $\psi=\varphi-\langle\varphi\rangle$. *Is it possible to choose* ψ *as test function in* (3.29)*? Deduce that the solution* u *to* (3.29) *satisfies*

$$\forall\varphi\in\mathcal{D}(Q),\qquad \int_Q\nabla u\cdot\nabla\varphi=\int_Q f\varphi-\langle\varphi\rangle\int_Q f.$$

Check that, if f *satisfies*

$$\int_Q f=0, \tag{3.30}$$

then the unique solution u *to* (3.29) *is solution to* (3.28).

Remark 3.22. *This function* u *actually satisfies more boundary conditions than those written in* (3.28) *(we have seen that this problem is ill-posed). Actually, the normal derivative of* u *is also periodic, in the following sense. Consider the restriction* v *of* u *to* Q *(by definition,* $u\in H^1_{\mathrm{per,m}}(Q)$*, hence* u *is defined on* $\mathbb{R}^d$*). Then the trace of* $\dfrac{\partial v}{\partial n}$ *is skew-periodic. In the case of the function* $u(x,y)=x^2-y^2$ *that we mentioned above when* $Q=(-1/2,1/2)^2$ *and* $f=0$ *(i.e. the* Q*-periodic function* u *that is equal to* $u(x,y)=x^2-y^2$ *on* Q*), we have, on the boundary edge corresponding to* $x=1/2$*, that* $\dfrac{\partial v}{\partial n}=2x=1$ *and, on the boundary edge corresponding to* $x=-1/2$*, we have* $\dfrac{\partial v}{\partial n}=-2x=1$*. Hence the trace of* $\dfrac{\partial v}{\partial n}$ *is not skew-periodic.*

2. ***Second variational formulation.***

We have worked with $H^1_{\mathrm{per,m}}(Q)$ *to remove the indetermination by the addition of a constant. There are several ways to remove this indetermination, and we now show another one.*

Define on $H^1_{\mathrm{per}}(Q)$ *the bilinear form*

$$\widetilde{a}(u,v)=\int_Q\nabla u\cdot\nabla v+\langle u\rangle\ \langle v\rangle$$

and the linear form

$$b(v)=\int_Q f\,v,$$

where we recall that $\langle v\rangle=\dfrac{1}{|\Omega|}\displaystyle\int_\Omega v.$

(a) *Show that* $\widetilde{a}$ *and* b *are continuous, and that* $\widetilde{a}$ *is coercive on* $H^1_{\mathrm{per}}(Q)$.

(b) *Deduce that the problem*

$$\left\{\begin{array}{l}\textit{Find } u\in H^1_{\mathrm{per}}(Q) \textit{ such that}\\ \forall v\in H^1_{\mathrm{per}}(Q),\quad \widetilde{a}(u,v)=b(v)\end{array}\right. \tag{3.31}$$

is well-posed.

(c) *Again assuming* (3.30)*, and using an appropriate choice of v in* (3.31)*, show that the solution u to* (3.31) *satisfies $\langle u\rangle = 0$. Deduce that the unique solution u to* (3.31) *is also the unique solution to* (3.29)*.*

3. *Show that, if* (3.28) *has a smooth solution (so that all integrations by parts are correct), then f satisfies* (3.30)*.*

Chapter 4

Inf-sup theory

The previous chapter relies on the Lax-Milgram theorem 3.7. The problem

$$\left\{\begin{array}{ll} \text{Find } u \in V \text{ such that} & \\ \forall w \in V, \qquad a(u,w) = b(w) & \end{array}\right. \tag{4.1}$$

is studied there under a coercivity assumption on the bilinear form a (and in a Hilbert space V).

The coercivity assumption is restrictive. In the finite dimensional setting (that is when $V = \mathbb{R}^n$), consider the bilinear form $a(u,w) = w^T A u$ for some matrix $A \in \mathbb{R}^{n\times n}$. The well-posedness of (4.1) is equivalent to the invertibility of A. We then note that the matrices

$$A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad \text{and} \quad A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

are invertible, but that the associated bilinear form is not cercive.

In the infinite dimensional setting, consider the advection-diffusion problem (3.10), that is

$$\left\{\begin{array}{ll} -\Delta u + c\cdot\nabla u = f & \text{in } \mathcal{D}'(\Omega), \\ u = 0 & \text{on } \partial\Omega, \end{array}\right.$$

for smooth functions f and c, without any smallness assumptions on c or on its divergence. The bilinear form associated to this problem is

$$a(u,w) = \int_\Omega \nabla u \cdot \nabla w + \int_\Omega (c\cdot\nabla u)\, w,$$

and it is possible to find vector fields c such that $\displaystyle\inf_{u\in H^1_0(\Omega)} \frac{a(u,u)}{\|u\|^2_{H^1(\Omega)}} < 0$. In this case, a is not coercive and one cannot use the Lax-Milgram theorem to study the above PDE. It turns out that, in some cases, the above problem is well-posed.

This chapter is devoted to the introduction of a more general theory, the inf-sup theory, to handle problems such as (4.1) without any coercivity assumptions on a.

4.1 The finite dimensional case

We first establish a necessary and sufficient condition for the "invertibility" of a square matrix A, namely the well-posedness of a problem of the type $Au = b$ (see Lemma 4.7 below). The interest of that condition is that, in the infinite dimensional case, the same condition can be used to ensure the well-posedness of a problem of the type $Au = b$, where u belongs to some Banach space V and b belongs to the dual of some Banach space W (see Theorem 4.13 below).

4.1.1 Square matrices

Let us start by considering Problem (4.1) in a finite dimensional setting. We thus set $V = \mathbb{R}^n$, and $a(u,w) = w^T A u$ for some matrix $A \in \mathbb{R}^{n \times n}$. We now consider a vector $b \in \mathbb{R}^n$ and look for $u \in \mathbb{R}^n$ such that (4.1) holds, that is

$$\forall w \in \mathbb{R}^n, \qquad w^T A u = w^T b,$$

which is of course equivalent to $Au = b$. This problem is well-posed if and only if A is invertible.

The coercivity assumption amounts to requesting that there exists $\alpha > 0$ such that, for any $u \in \mathbb{R}^n$, we have $u^T A u \geq \alpha u^T u$. Since $u^T A u = \dfrac{1}{2} u^T \left(A + A^T\right) u$, this amounts to requesting that the eigenvalues of the symmetric matrix $A + A^T$ are all positive. This condition is sufficient for A to be invertible (as the Lax-Milgram theorem shows), but it is not necessary. Consider indeed the particular case of a symmetric matrix A. This matrix is invertible if and only if its eigenvalues do not vanish. But they do not have to be positive.

Consider now the following assumption on A: we assume that there exists $\alpha > 0$ such that

$$\inf_{u \in \mathbb{R}^n,\, u \neq 0} \ \sup_{w \in \mathbb{R}^n,\, w \neq 0} \frac{w^T A u}{\|u\| \, \|w\|} \geq \alpha, \tag{4.2}$$

which can be rephrased as

$$\forall u \in \mathbb{R}^n, \qquad \sup_{w \in \mathbb{R}^n,\, w \neq 0} \frac{w^T A u}{\|w\|} \geq \alpha \|u\|. \tag{4.3}$$

We note that the above supremum is attained for $w = Au$ (equality case in the Cauchy-Schwarz inequality), that is

$$\sup_{w \in \mathbb{R}^n,\, w \neq 0} \frac{w^T A u}{\|w\|} = \sqrt{u^T A^T A u}. \tag{4.4}$$

The eigenvectors of $A^T A$ form an orthonormal basis of $\mathbb{R}^n$. Denoting λ_i^2 the non-negative eigenvalues of $A^T A$, we see that the condition (4.3) is equivalent to

$$\forall u \in \mathbb{R}^n, \qquad \sqrt{\sum_{i=1}^n \lambda_i^2 u_i^2} \geq \alpha \sqrt{\sum_{i=1}^n u_i^2},$$

which is equivalent to

$$\forall 1 \leq i \leq n, \qquad |\lambda_i| \geq \alpha. \tag{4.5}$$

Remark 4.1. *The λ_i introduced above are called the singular values of A.*

Remark 4.2. *Observing that $\sqrt{u^T A^T A u} = \|Au\|$, we deduce from* (4.3) *and* (4.4) *that the inf-sup condition* (4.2) *can be written as $\|Au\| \geq \alpha \|u\|$. This is to be compared with a coercivity assumption on A, which would read $u^T A u \geq \alpha \|u\|^2$.*

Suppose for a moment that A is symmetric. Then the condition (4.5) means that the eigenvalues of A do not vanish, which is exactly a necessary and sufficient condition for A to be invertible.

Suppose now that A is not necessarily symmetric. Let us show that (4.3) implies that A is invertible. Consider first $u \in \text{Ker } A$. Then (4.3) implies that

$$\sup_{w \in \mathbb{R}^n,\, w \neq 0} \frac{w^T A u}{\|w\|} = 0 \geq \alpha \|u\|,$$

thus $u = 0$, hence A is injective. Owing to the fact that dim Ker A + dim Im A = dim V, we get that dim Im A = dim V, hence A is surjective. The condition (4.3) hence implies that A is bijective.

Conversely, if A is invertible, then (4.3) holds. Consider indeed the symetric matrix A^TA, which has non-negative eigenvalues, and let $\lambda = \inf \ \mathrm{Sp} \ A^TA$. If $\lambda = 0$, then there exists $e \in \mathbb{R}^n$ of unit norm such that $A^TAe = 0$. We thus have $e^TA^TAe = 0$, hence $Ae = 0$. Since A is invertible, we get $e = 0$, which is in contradiction with $\|e\| = 1$. We have hence shown that $\lambda = \inf \ \mathrm{Sp} \ A^TA > 0$, and hence that condition (4.5) holds.

We have thus shown the following result:

Lemma 4.3. *Let A be a square matrix. Then the condition* (4.3) *is a necessary and sufficient condition for A to be invertible.*

Remark 4.4. *Assume that A is coercive, namely that there exists $\alpha > 0$ such that $u^TAu \geq \alpha u^Tu$ for any $u \in \mathbb{R}^n$. Then the condition* (4.3) *is obviously satisfied.*

4.1.2 Rectangular matrices

We now go one step further and consider the following generalization of Problem (4.1):

$$\left\{ \begin{array}{l} \text{Find } u \in V \text{ such that} \\ \forall w \in W, \qquad a(u,w) = b(w) \end{array} \right.$$

where V and W are two Banach spaces. Note that the space in which we look for the solution u is a priori different from the space to which the test functions w belong. In addition, V and W are now Banach spaces.

Again, we consider the finite dimensional case, where $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$. We thus consider a matrix $A \in \mathbb{R}^{m \times n}$, a vector $b \in \mathbb{R}^m$, and look for $u \in \mathbb{R}^n$ such that

$$\forall w \in \mathbb{R}^m, \qquad w^TAu = w^Tb,$$

which is of course equivalent to $Au = b$.

Remark 4.5. *Of course, a necessary condition for the well-posedness of the problem $Au = b$ is that u and b belong to spaces of identical dimension, and hence that n = dim V = dim W = m.*

We however do not wish to write this assumption, since there is no equivalent of it in a infinite dimensional context. We rather wish to write necessary and sufficient conditions for the invertibility of A which can easily be translated in a infinite dimensional context. This will be the case of the conditions (4.6) *and* (4.7) *below. We will then check (see Remark 4.9 below) that these conditions imply, in a finite dimensional context, that dim V = dim W.*

Consider now the following assumptions on A. We assume that there exists $\alpha > 0$ such that

$$\inf_{u \in \mathbb{R}^n, \, u \neq 0} \ \sup_{w \in \mathbb{R}^m, \, w \neq 0} \frac{w^TAu}{\|u\| \, \|w\|} \geq \alpha \tag{4.6}$$

and we assume that

$$\text{If } w \in \mathbb{R}^m \text{ is such that } w^TAu = 0 \text{ for any } u \in \mathbb{R}^n, \text{ then } w = 0. \tag{4.7}$$

Note that (4.6) can be rephrased as

$$\forall u \in \mathbb{R}^n, \qquad \sup_{w \in \mathbb{R}^m, \, w \neq 0} \frac{w^TAu}{\|w\|} \geq \alpha \|u\|.$$

The supremum being again attained for $w = Au$, we get

$$\sup_{w \in \mathbb{R}^m,\, w \neq 0} \frac{w^T A u}{\|w\|} = \sqrt{u^T A^T A u}.$$

The eigenvectors of $A^T A$ form an orthonormal basis of $\mathbb{R}^n$. Denoting λ_i^2 the non-negative eigenvalues of $A^T A$, we see that the condition (4.6) is equivalent to

$$\forall u \in \mathbb{R}^n, \qquad \sqrt{\sum_{i=1}^n \lambda_i^2 u_i^2} \geq \alpha \sqrt{\sum_{i=1}^n u_i^2},$$

which is equivalent to

$$\forall 1 \leq i \leq n, \qquad |\lambda_i| \geq \alpha.$$

We have the following result:

Lemma 4.6. *(i) The condition* (4.6) *is equivalent to the fact that Ker* $A = \{0\}$.
(ii) The condition (4.7) *is equivalent to the fact that dim Im* A = *dim* W.

We recall that, in finite dimension, A is surjective if and only if A^T is injective.

Proof of Lemma 4.6. Let us prove the assertion (i). Assume that condition (4.6) holds and consider $u \in \text{Ker } A$. We have

$$\sup_{w \in \mathbb{R}^m,\, w \neq 0} \frac{w^T A u}{\|w\|} = 0 \geq \alpha \|u\|,$$

thus $u = 0$. Conversely, assume that Ker $A = \{0\}$ and that condition (4.6) does not hold. Since we have that $\displaystyle\sup_{w \in \mathbb{R}^m,\, w \neq 0} \frac{w^T A u}{\|w\|} \geq 0$ for any $u \in \mathbb{R}^n$, this means that, for any integer p, there exists $u_p \in \mathbb{R}^n$ with $\|u_p\| = 1$ such that

$$\sup_{w \in \mathbb{R}^m,\, w \neq 0} \frac{w^T A u_p}{\|w\|} \leq \frac{1}{p}.$$

The sequence u_p is bounded in $\mathbb{R}^n$. It thus converges to some $u^\star$, up to the extraction of a subsequence, which satisfies $\|u^\star\| = 1$. For any $w \in \mathbb{R}^m$, we have

$$\frac{w^T A u_p}{\|w\|} \leq \sup_{w \in \mathbb{R}^m,\, w \neq 0} \frac{w^T A u_p}{\|w\|} \leq \frac{1}{p},$$

which yields, passing to the limit $p \to \infty$, that

$$\frac{w^T A u^\star}{\|w\|} \leq 0.$$

Taking $w = Au^\star$, we get $Au^\star = 0$. Since Ker $A = \{0\}$, this yields $u^\star = 0$, which is in contradiction with $\|u^\star\| = 1$. We have thus shown that condition (4.6) holds.

We now prove the assertion (ii). Assume that condition (4.7) holds and consider $w \in \text{Ker } A^T$. Then, for any $u \in \mathbb{R}^n$, we have $w^T A u = 0$, hence $w = 0$. We thus get that Ker $A^T = \{0\}$. Since A^T is injective, we get, thanks to the finite dimension setting, that A is surjective, and hence that dim Im A = dim W.

Assume conversely that dim Im A = dim W. Then A is surjective, hence A^T is injective. This directly shows the condition (4.7). □

We then have the following result:

Lemma 4.7. *Assume that V and W are finite dimensional spaces. Let $b \in W$. The conditions* (4.6)–(4.7) *are necessary and sufficient conditions for the problem $Au = b$ to be well-posed in V.*

Proof. Assume that the conditions (4.6)–(4.7) hold. Then, in view of Lemma 4.6, we see that A is injective and that Im $A = W$. The problem $Au = b$ is thus well-posed in V.

Assume now that the problem $Au = b$ is thus well-posed in V. The existence of a solution to that problem implies that Im $A = W$, hence that condition (4.7) holds, in view of Lemma 4.6(ii). The uniqueness of a solution to that problem implies that Ker $A = \{0\}$, hence that condition (4.6) holds, in view of Lemma 4.6(i). □

Remark 4.8. *Assume that A is a square matrix, that is dim V = dim W, as in Section 4.1.1. The condition* (4.6) *implies that Ker $A = \{0\}$. Since dim Ker A + dim Im A = dim V, we deduce that dim Im A = dim V = dim W, hence the condition* (4.7)*. The converse is also true. We thus see that, if dim V = dim W, then conditions* (4.6) *and* (4.7) *are equivalent. We thus recover Lemma 4.3 as a consequence of Lemma 4.7.*

Remark 4.9. *Assuming that conditions* (4.6)–(4.7) *hold, let us show that dim V = dim W. We recall that $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$.*

In view of condition (4.6) *and Lemma 4.6(i), we get that Ker $A = \{0\}$, and hence that $m \geq n$ (there are more equations in the system $Au = 0$ than unknowns).*

In view of condition (4.7) *and Lemma 4.6(ii), we get that Im $A = W$. In addition, we always have dim Im $A \leq$ dim V=n. We thus get that $m \leq n$.*

4.2 The infinite dimensional case

This section is devoted to the study of the problem

$$\left\{ \begin{array}{l} \text{Find } u \in V \text{ such that} \\ \forall w \in W, \qquad a(u,w) = b(w) \end{array} \right. \tag{4.8}$$

where V and W are two Banach spaces, a is a bilinear continuous form on $V \times W$ and b is linear continuous form on W.

4.2.1 The Hilbert case

We start by the simple case when W is a Hilbert space. We then have the following result:

Theorem 4.10. *Assume that W is a Hilbert space. We assume that the two following conditions are satisfied:*

- *there exists $\alpha > 0$ such that*

$$\inf_{u \in V,\, u \neq 0} \sup_{w \in W,\, w \neq 0} \frac{a(u,w)}{\|u\|_V \|w\|_W} \geq \alpha; \tag{4.9}$$

- *the bilinear form a is such that*

$$\textit{If } w \in W \textit{ is such that } a(u,w) = 0 \textit{ for any } u \in V, \textit{ then } w = 0. \tag{4.10}$$

Then the problem (4.8) *is well-posed (i.e. admits one and only one solution).*

Assume that the conditions (4.9) and (4.10) hold. Then the problem (4.8) has a unique solution $u \in V$. Furthermore, we have the a priori estimate

$$\alpha \|u\|_V \leq \sup_{w\in W,\, w\neq 0} \frac{a(u,w)}{\|w\|_W} = \sup_{w\in W,\, w\neq 0} \frac{b(w)}{\|w\|_W} = \|b\|_{W'}. \tag{4.11}$$

Remark 4.11. *As we will see below (see Theorem 4.13), the conditions* (4.9) *and* (4.10) *are not only sufficient but also necessary conditions for* (4.8) *to be well-posed.*

Remark 4.12. *Consider the particular case when $V = W$ is a Hilbert space, and a is a bilinear continuous, coercive form on V. Then a satisfies the two properties* (4.9) *and* (4.10). *Indeed, the coercivity of a implies that*

$$\sup_{w\in V,\, w\neq 0} \frac{a(u,w)}{\|u\|_V\|w\|_V} \geq \frac{a(u,u)}{\|u\|_V^2} \geq \alpha,$$

from which we deduce (4.9). *Consider now $w \in V$ such that $a(u,w) = 0$ for any $u \in V$. Choosing $u = w$, we get $a(w,w) = 0$, which implies, again using the coercivity of a, that $w = 0$. This proves* (4.10).

Proof of Theorem 4.10. The proof follows the same arguments as the proof of the Lax-Milgram theorem (in the general case), see Theorem 3.7. We assume that W is a Hilbert space and denote $\langle \cdot,\cdot\rangle_W$ its scalar product. The norm of V is denoted $\|\cdot\|_V$.

Consider the mapping

$$\begin{array}{rcl} \Phi : V & \longrightarrow & W \\ u & \mapsto & b \text{ such that } a(u,\cdot) = \langle b,\cdot\rangle_W. \end{array}$$

Note that we have used the Riesz theorem in W to represent the continuous and linear form $w \in W \mapsto a(u,w) \in \mathbb{R}$ by an element of W. The mapping Φ is of course linear, and it is also continuous:

$$\|b\|_W^2 = \langle b,b\rangle_W = a(u,b) \leq M\|u\|_V\,\|b\|_W$$

hence $\|\Phi(u)\|_W = \|b\|_W \leq M\|u\|_V$. We wish to show that, under assumptions (4.9)–(4.10), the map Φ is bijective. The proof falls in three steps.

Step 1: Φ is injective: Let $u \in V$ such that $\Phi(u) = 0$, that is such that $a(u,w) = 0$ for any $w \in W$. In view of (4.9), we deduce that $u = 0$.

Step 2: Φ is surjective: Let $X = \mathrm{Im}(\Phi) \subset W$. We show that $X = W$.

(a) Let us first show that $X = \mathrm{Im}(\Phi)$ is closed. Let $u \in V$, $u \neq 0$. Using (4.9), we have

$$\sup_{w\in W,\, w\neq 0} \frac{\langle \Phi(u), w\rangle_W}{\|w\|_W} = \sup_{w\in W,\, v\neq 0} \frac{a(u,w)}{\|w\|_W} \geq \alpha\|u\|_V.$$

In addition, for any $w \in W$, $w \neq 0$, we have

$$\frac{\langle \Phi(u), w\rangle_W}{\|w\|_W} \leq \|\Phi(u)\|_W.$$

We thus obtain that

$$\|\Phi(u)\|_W \geq \alpha\|u\|_V$$

and this estimate also holds if $u = 0$. Consider now a sequence $(b_n)_{n\in\mathbb{N}}$ in X that converges in W to some $b \in W$. We want to show that $b \in X$. Let $u_n \in V$ such that $\Phi(u_n) = b_n$.

Since $(b_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in W, we see that $(u_n)_{n\in\mathbb{N}}$ is also a Cauchy sequence in V. Indeed,

$$\|b_p - b_q\|_W = \|\Phi(u_p) - \Phi(u_q)\|_W = \|\Phi(u_p - u_q)\|_W \geq \alpha\|u_p - u_q\|_V.$$

Therefore, since V is a complete space, we get that $(u_n)_{n\in\mathbb{N}}$ converges in V to some $u \in V$. In addition, Φ is continuous, so

$$b_n = \Phi(u_n) \longrightarrow \Phi(u) \quad \text{in } W.$$

Hence $b = \Phi(u) \in X$, by uniqueness of the limit in W. The vector space X is hence closed.

(b) Let $b_0 \in W$. Since X is closed in W, we can consider the orthogonal projection of b_0 on X, that we denote $b_1 \in X$. For any $v \in X$, we have $\langle b_0 - b_1, v\rangle_W = 0$. Hence, for any $u \in V$, we have

$$0 = \langle b_0 - b_1, \Phi(u)\rangle_W = a(u, b_0 - b_1).$$

In view of the condition (4.10), we deduce that $b_0 = b_1 \in X$, which implies that $W = X$.

Step 3: Conclusion: Since b is a continuous linear form on W, it can be represented by some $f \in W$. The problem (4.8) can thus be recast as finding $u \in V$ such that $a(u, \cdot) = \langle f, \cdot\rangle_W$, that is finding $u \in V$ such that $\Phi(u) = f$. In view of the above two steps, the mapping Φ is bijective from V to W, so this problem has a unique solution. □

4.2.2 The Banach case

We now turn to the general case when W is a Banach space. We admit the following result, the proof of which is difficult (see e.g. [3, 4]). This result provides a statement similar to Lemma 4.7, but for infinite dimensional spaces.

Theorem 4.13 (Banach–Necas–Babuska)**.** *Assume that W is* reflexive. *The problem* (4.8) *is well-posed (i.e. admits one and only one solution) if and only if the two following conditions are satisfied:*

- *there exists $\alpha > 0$ such that*

$$\inf_{u\in V,\, u\neq 0} \ \sup_{w\in W,\, w\neq 0} \frac{a(u,w)}{\|u\|_V \|w\|_W} \geq \alpha; \tag{4.12}$$

- *the bilinear form a is such that*

$$\text{If } w \in W \text{ is such that } a(u,w) = 0 \text{ for any } u \in V, \text{ then } w = 0. \tag{4.13}$$

This theorem will be henceforth called the BNB theorem.

Assume that the conditions (4.12) and (4.13) hold. Then the problem (4.8) has a unique solution $u \in V$, which again satisfies the a priori estimate (4.11).

Remark 4.14. *The notion of reflexivity goes beyond the scope of these lecture notes. It is defined as follows. Let V be a Banach space. Let $J : V \to V''$ be defined by: for any $v \in V$, we set*

$$\forall v' \in V', \qquad \langle Jv, v'\rangle_{V'',V'} = \langle v', v\rangle_{V',V},$$

where we recall that $V' = \mathcal{L}(V, \mathbb{R})$ is the dual of V. It can be shown that J is always injective. If J is surjective, then the space V is said to be reflexive.

We point out the following useful facts:

- *If the dimension of V is finite, then V is reflexive.*
- *If V is a Hilbert space, then it is reflexive.*
- *For any $p \in (1, +\infty)$ (note that we exclude the cases $p = 1$ and $p = \infty$), the Banach space $L^p(\Omega)$ is reflexive.*

4.3 Stokes problem

Consider a viscous and incompressible fluid. The flow of that fluid is described by the Navier-Stokes equations

$$\frac{\partial u}{\partial t} + (u \cdot \nabla) u + \nabla p - \Delta u = f, \tag{4.14}$$

$$\text{div } u = 0, \tag{4.15}$$

where $u : [0, +\infty[\times \mathbb{R}^d \to \mathbb{R}^d$ is the velocity field and $p : [0, +\infty[\times \mathbb{R}^d \to \mathbb{R}$ is the pressure field. All physical parameters have been set to 1 in (4.14). Of course, these equations should be complemented by appropriate boundary conditions and initial conditions. The function $f : [0, +\infty[\times \mathbb{R}^d \to \mathbb{R}^d$ represents the volumic forces applied on the fluid (for instance, gravity). The equation (4.14) encodes the conservation of the momentum, while the equation (4.15) encodes the conservation of the mass.

When the velocity is small, the nonlinear term $(u \cdot \nabla) u$ may be neglected. We consider in what follows stationary problems, hence the term $\dfrac{\partial u}{\partial t}$ vanishes.

In what follows, we assume that Ω is an open connected bounded subset of $\mathbb{R}^d$. Our aim is to study the Stokes problem, complemented (for simplicity) by homogeneous Dirichlet boundary conditions for the velocity. We hence look for a velocity field $u : \Omega \to \mathbb{R}^d$ and a pressure field $p : \Omega \to \mathbb{R}$ such that

$$-\Delta u + \nabla p = f \quad \text{in } [\mathcal{D}'(\Omega)]^d, \qquad \text{div } u = 0 \quad \text{in } \mathcal{D}'(\Omega), \qquad u = 0 \quad \text{on } \partial\Omega.$$

Of course, the pressure is only determined up to the addition of a constant. We thus restrict ourselves to pressures of vanishing mean, i.e. satisfying $\displaystyle\int_\Omega p = 0$.

More precisely, we introduce the space

$$L^2_0(\Omega) = \left\{ q \in L^2(\Omega), \quad \int_\Omega q = 0 \right\},$$

and we consider the following problem:

$$\begin{cases} \text{Find } (u,p) \in (H^1_0(\Omega))^d \times L^2_0(\Omega) \text{ such that} \\ -\Delta u + \nabla p = f \quad \text{in } [\mathcal{D}'(\Omega)]^d, \\ \text{div } u = 0 \quad \text{in } \mathcal{D}'(\Omega), \end{cases} \tag{4.16}$$

for some $f \in (L^2(\Omega))^d$. When $u \in (H^1_0(\Omega))^d$, we recall that Δu is a vector in $\mathbb{R}^d$, the i-th component of which is equal to Δu_i, where u_i is the i-th component of u.

4.3.1 Variational formulation

The variational formulation of (4.16) reads as follows. We introduce

- the Hilbert space $V = (H^1_0(\Omega))^d \times L^2_0(\Omega)$, endowed with the scalar product

  $$\langle x, y \rangle_V = \langle p, q \rangle_{L^2(\Omega)} + \sum_{i=1}^d \langle u_i, v_i \rangle_{H^1(\Omega)}$$

  for any $x = (u,p) \in V$ and $y = (v,q) \in V$.

- the bilinear form a defined on $V \times V$ by: for any $x = (u,p) \in V$ and $y = (v,q) \in V$, we set
$$a(x,y) = \int_\Omega \nabla u \cdot \nabla v - p \,\mathrm{div}\, v + q \,\mathrm{div}\, u,$$
where $\nabla u \cdot \nabla v$ is defined by $\nabla u \cdot \nabla v = \sum_{i=1}^{d} \nabla u_i \cdot \nabla v_i = \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\partial u_i}{\partial x_j}\frac{\partial v_i}{\partial x_j}$.

- the linear form b defined on V by: for any $y = (v,q) \in V$, we set
$$b(y) = \int_\Omega f \cdot v = \sum_{i=1}^{d} \int_\Omega f_i \, v_i.$$

Consider the problem:
$$\left\{ \begin{array}{ll} \text{Find } x \in V \text{ such that} & \\ \forall y \in V, & a(x,y) = b(y). \end{array} \right. \tag{4.17}$$

Proposition 4.15. *Problems* (4.16) *and* (4.17) *are equivalent.*

We note that Problem (4.17) falls within the scope of Problem (4.8). In addition, the space V is a Hilbert space. As will be shown below, the bilinear form a is continuous on $V \times V$ and the linear form b is continuous on V. However, Problem (4.17) cannot be studied using the Lax-Milgram theorem. Indeed, for any $x = (u,p) \in V$, we have, thanks to the Poincaré inequality, that
$$a(x,x) = \sum_{i=1}^{d} \|\nabla u_i\|^2_{L^2(\Omega)} \geq \frac{1}{1 + C_\Omega^2} \sum_{i=1}^{d} \|u_i\|^2_{H^1(\Omega)}.$$

However, the pressure p does not appear in the quantity $a(x,x)$, and therefore $a(x,x)$ cannot be a upper-bound (up to a multiplicative constant) of $\|x\|_V^2$. We will have to resort to the BNB theorem 4.13 to study the well-posedness of (4.17).

Proof of Proposition 4.15. The proof falls in three steps.

Step 1. We first show that a is continuous on $V \times V$ and that b is continuous on V. For any $y = (v,q) \in V$, we have
$$|b(y)| \leq \sum_{i=1}^{d} \left| \int_\Omega f_i \, v_i \right| \leq \sum_{i=1}^{d} \|f_i\|_{L^2(\Omega} \|v_i\|_{H^1(\Omega)} \leq C \|y\|_V,$$
which means that b is continuous on V. Furthermore, for any $x = (u,p) \in V$ and $y = (v,q) \in V$, we have
$$\begin{aligned} |a(x,y)| &\leq \sum_{i=1}^{d} \left| \int_\Omega \nabla u_i \cdot \nabla v_i \right| + \left| \int_\Omega p \,\mathrm{div}\, v \right| + \left| \int_\Omega q \,\mathrm{div}\, u \right| \\ &\leq \sum_{i=1}^{d} \|u_i\|_{H^1(\Omega)} \|v_i\|_{H^1(\Omega)} + \|p\|_{L^2(\Omega)} \|\mathrm{div}\, v\|_{L^2(\Omega)} + \|q\|_{L^2(\Omega)} \|\mathrm{div}\, u\|_{L^2(\Omega)} \\ &\leq C \|x\|_V \, \|y\|_V, \end{aligned}$$
which means that a is continuous on $V \times V$.

Step 2. We now show that (4.17) implies (4.16). Let $x = (u, p) \in V$ be a solution to (4.17). Let $\phi \in (\mathcal{D}(\Omega))^d$. We see that $y = (\phi, 0) \in V$, thus

$$\begin{aligned}
\sum_{i=1}^{d} \langle f_i, \phi_i \rangle_{\mathcal{D}',\mathcal{D}} &= \sum_{i=1}^{d} \int_\Omega f_i\, \phi_i \\
&= b(y) \\
&= a(x, y) \qquad \text{[Equation (4.17)]} \\
&= \int_\Omega \nabla u \cdot \nabla \phi - p \operatorname{div} \phi \\
&= \sum_{i=1}^{d} \langle \nabla u_i, \nabla \phi_i \rangle_{\mathcal{D}',\mathcal{D}} - \sum_{i=1}^{d} \langle p, \frac{\partial \phi_i}{\partial x_i} \rangle_{\mathcal{D}',\mathcal{D}} \\
&= \sum_{i=1}^{d} \langle -\Delta u_i, \phi_i \rangle_{\mathcal{D}',\mathcal{D}} + \langle \frac{\partial p}{\partial x_i}, \phi_i \rangle_{\mathcal{D}',\mathcal{D}}.
\end{aligned}$$

The functions ϕ_i being independent, we obtain, for any $1 \leq i \leq d$, that $-\Delta u_i + \dfrac{\partial p}{\partial x_i} = f_i$ in $\mathcal{D}'(\Omega)$. This yields the first equation in (4.16).

Let now $\psi \in \mathcal{D}(\Omega)$ and $\overline{\psi} = \psi - m$ with $m = \dfrac{1}{|\Omega|} \displaystyle\int_\Omega \psi$. We see that $y = (0, \overline{\psi}) \in V$, thus

$$\begin{aligned}
0 &= b(y) \\
&= a(x, y) \qquad \text{[Equation (4.17)]} \\
&= \int_\Omega \overline{\psi} \operatorname{div} u \\
&= \int_\Omega \psi \operatorname{div} u - m \int_\Omega \operatorname{div} u.
\end{aligned}$$

By integration by part and using that u vanishes on $\partial\Omega$, we see that $\displaystyle\int_\Omega \operatorname{div} u = \int_{\partial\Omega} u \cdot n = 0$. The last term above thus vanishes, and we get that $\displaystyle\int_\Omega \psi \operatorname{div} u = 0$ for any $\psi \in \mathcal{D}(\Omega)$, which is exactly the second equation in (4.16).

Step 3. Conversely, let $x = (u, p)$ be a solution to (4.16). Let $\phi \in (\mathcal{D}(\Omega))^d$, $\psi \in \mathcal{D}(\Omega)$ and $\overline{\psi} = \psi - m$ with $m = \dfrac{1}{|\Omega|} \displaystyle\int_\Omega \psi$. We see that $y = (\phi, \overline{\psi}) \in V$, and that

$$\begin{aligned}
b(y) &= \int_\Omega f \cdot \phi \\
&= \sum_{i=1}^{d} \langle f_i, \phi_i \rangle_{\mathcal{D}',\mathcal{D}} \\
&= \sum_{i=1}^{d} \langle -\Delta u_i + \frac{\partial p}{\partial x_i}, \phi_i \rangle_{\mathcal{D}',\mathcal{D}} \qquad \text{[First equation of (4.16)]} \\
&= \sum_{i=1}^{d} \langle \nabla u_i, \nabla \phi_i \rangle_{\mathcal{D}',\mathcal{D}} - \sum_{i=1}^{d} \langle p, \frac{\partial \phi_i}{\partial x_i} \rangle_{\mathcal{D}',\mathcal{D}} \\
&= a(x, y) - \int_\Omega \overline{\psi} \operatorname{div} u \\
&= a(x, y) + m \int_\Omega \operatorname{div} u - \int_\Omega \psi \operatorname{div} u.
\end{aligned}$$

The term $\int_\Omega \operatorname{div} u$ vanishes by integration by part and using that u vanishes on $\partial\Omega$. The term $\int_\Omega \psi \operatorname{div} u$ vanishes in view of the second equation of (4.16). We hence get that

$$\forall \phi \in (\mathcal{D}(\Omega))^d, \quad \forall \psi \in \mathcal{D}(\Omega), \qquad a(x,y) = b(y). \tag{4.18}$$

We eventually proceed by density and continuity. Let $y = (v,q) \in V$. By density of $(\mathcal{D}(\Omega))^d$ in $(H^1_0(\Omega))^d$, there exists $\phi_n \in (\mathcal{D}(\Omega))^d$ such that $\lim_{n\to\infty} \|v - \phi_n\|_{(H^1(\Omega))^d} = 0$. By density of $\mathcal{D}(\Omega)$ in $L^2(\Omega)$, there exists $\psi_n \in \mathcal{D}(\Omega)$ such that $\lim_{n\to\infty} \|q - \psi_n\|_{L^2(\Omega)} = 0$. We then have

$$\left| \frac{1}{|\Omega|} \int_\Omega \psi_n \right| = \left| \frac{1}{|\Omega|} \int_\Omega (\psi_n - q) \right| \leq C \|q - \psi_n\|_{L^2(\Omega)}.$$

Setting $\overline{\psi}_n = \psi_n - m_n$ with $m_n = \frac{1}{|\Omega|} \int_\Omega \psi_n$, we hence write

$$\|q - \overline{\psi}_n\|_{L^2(\Omega)} \leq \|q - \psi_n\|_{L^2(\Omega)} + |m_n| \leq C \|q - \psi_n\|_{L^2(\Omega)},$$

and thus $\lim_{n\to\infty} \|q - \overline{\psi}_n\|_{L^2(\Omega)} = 0$. Sstting $y_n = (\phi_n, \overline{\psi}_n)$, we thus have $y_n \in V$ and $\lim_{n\to\infty} \|y - y_n\|_V = 0$. Furthermore, writing (4.18) for (ϕ_n, ψ_n), we have $a(x, y_n) = b(y_n)$. We then pass to the limit $n \to \infty$ and get that $a(x,y) = b(y)$. Therefore, x is a solution to (4.17). □

4.3.2 Recasting the variational formulation of the Stokes problem

The Stokes problem (4.17) has a specific structure, which we exhibit here. In Section 4.3.3, we will state a specific version of the BNB Theorem 4.13 well adapted to that specific structure. This specific structure (namely that of a saddle point problem) will be further discussed in Section 4.4.

We set

1. $X = (H^1_0(\Omega))^d$ and $M = L^2_0(\Omega)$;
2. let a_0 be the bilinear form defined on $X \times X$ by: for any u and v in X,
$$a_0(u,v) = \int_\Omega \nabla u \cdot \nabla v = \sum_{i=1}^d \int_\Omega \nabla u_i \cdot \nabla v_i.$$
3. let b be the bilinear form defined on $X \times M$ by: for any $v \in X$ and any $p \in M$,
$$b(v,p) = -\int_\Omega p \operatorname{div} v.$$
4. for any $v \in X$, we set
$$g_1(v) = \int_\Omega f \cdot v = \sum_{i=1}^d \int_\Omega f_i \, v_i.$$
5. for any $q \in M$, we set $g_2(q) = 0$.

Then the variational formulation (4.17) of the Stokes problem can be recast in the following form:

$$\left\{ \begin{array}{ll} \text{Find } (u,p) \in X \times M \text{ such that} & \\ \forall v \in X, & a_0(u,v) + b(v,p) = g_1(v), \\ \forall q \in M, & b(u,q) = g_2(q). \end{array} \right. \tag{4.19}$$

The problem (4.19) has a specific structure in the sense that

- the unknown function p does not appear in the second equation;
- the unknown functions u and p are coupled through the *same* bilinear form in the two equations in (4.19);
- the space in which the solution is searched is the same as the space in which test functions are considered.

4.3.3 BNB theorem for problems of the type (4.19)

We now state a specific version of the BNB Theorem 4.13 well adapted to the structure of Problem (4.19). Note that we do not assume a_0 to be symmetric.

Theorem 4.16. *Let X and M be two Hilbert spaces. We assume that the bilinear (resp. linear) forms a_0 and b (resp. g_1 and g_2) are continuous. We also assume that the bilinear form a_0 is coercive on X. Then Problem* (4.19) *is well-posed if and only if*

$$\exists \beta > 0, \quad \forall q \in M, \quad \sup_{v \in X,\, v \neq 0} \frac{b(v,q)}{\|v\|_X} \geq \beta \|q\|_M. \tag{4.20}$$

Proof. Let α_0 be the coercivity constant of a_0 on X:

$$\forall v \in X, \quad a_0(v,v) \geq \alpha_0 \|v\|_X^2.$$

We set $V = X \times M$ that we endow with the norm $\|(v,q)\|_V = \|v\|_X + \|q\|_M$. We introduce the bilinear form a, defined on $V \times V$ by

$$\forall x = (u,p) \in V, \quad \forall y = (v,q) \in V, \qquad a(x,y) = a_0(u,v) + b(v,p) + b(u,q).$$

Let $g \in V'$ be defined by

$$\forall y = (v,q) \in V, \qquad g(y) = g_1(v) + g_2(q).$$

It is obvious that Problem (4.19) is equivalent to the following problem:

$$\begin{cases} \text{Find } x \in V \text{ such that} \\ \forall y \in V, \qquad a(x,y) = g(y). \end{cases}$$

Problem (4.19) is hence well-posed if and only if the bilinear form a satisfies the conditions (4.12) and (4.13).

Assume that the conditions (4.12) and (4.13) are satisfied. Let $q \in M$. From (4.12), we deduce that

$$\begin{aligned} \alpha \|q\|_M &\leq \sup_{(\overline{v},\overline{q}) \in V,\, (\overline{v},\overline{q}) \neq 0} \frac{a\Big((0,q),(\overline{v},\overline{q})\Big)}{\|\overline{v}\|_X + \|\overline{q}\|_M} \\ &= \sup_{(\overline{v},\overline{q}) \in V,\, (\overline{v},\overline{q}) \neq 0} \frac{b(\overline{v},q)}{\|\overline{v}\|_X + \|\overline{q}\|_M} \\ &= \sup_{\overline{v} \in X,\, \overline{v} \neq 0} \frac{b(\overline{v},q)}{\|\overline{v}\|_X}, \end{aligned}$$

thus the condition (4.20). Note that we have not used the condition (4.13).

Conversely, assume that the condition (4.20) holds and let us show that the conditions (4.12) and (4.13) are satisfied.

1. The proof of (4.12) relies on tools that go beyond the scope of these lecture notes (it actually relies on the tools that one uses to show Theorem 4.13, that we have admitted). We here admit that the condition (4.20) implies (4.12).

2. We now show (4.13). Let $(v,q) \in V$ such that, for any $(u,p) \in V$, we have $a\Big((u,p),(v,q)\Big) = 0$. Taking $(u,p) = (0,q)$, we get that $b(v,q) = 0$. Taking $(u,p) = (v,0)$, we get that $a_0(v,v) + b(v,q) = 0$, hence $a_0(v,v) = 0$, and thus $v = 0$ thanks to the coercivity of a_0 on X. The condition $a\Big((u,p),(v,q)\Big) = 0$ for any $(u,p) \in V$ hence yields that $b(u,q) = 0$ for any $u \in X$. Using (4.20), we deduce that $q = 0$. This proves (4.13).

This completes the proof. □

Remark 4.17. *A complete proof of Theorem 4.16 can be given in the finite dimensional setting. We then have that $a_0(u,v) = v^T A u$ and $b(v,q) = v^T B q$ for some square matrix A and some matrix B. We introduce $G_1 \in X$ and $G_2 \in M$ such that $g_1(v) = v^T G_1$ and $g_2(q) = q^T G_2$. Problem* (4.19) *can then be written as: find $u \in X$ and $p \in M$ such that, for any $v \in X$ and any $q \in M$,*

$$v^T A u + v^T B p = v^T G_1, \qquad u^T B q = q^T G_2,$$

that is

$$A u + B p = G_1, \qquad B^T u = G_2.$$

The bilinear form a_0 is coercive, hence, thanks to the Lax-Milgram theorem, the matrix A is invertible. We then write that $u = A^{-1}(G_1 - Bp)$, and thus look for $p \in M$ such that $B^T A^{-1}(G_1 - Bp) = G_2$, that is

$$B^T A^{-1} B p = B^T A^{-1} G_1 - G_2.$$

Problem (4.19) *is well-posed if and only if the square matrix $B^T A^{-1} B$ is invertible.*

Assume that (4.20) *holds, that is*

$$\forall q \in M, \qquad \sup_{v \in X,\, v \neq 0} \frac{v^T B q}{\|v\|} \geq \beta \|q\|.$$

Let $q \in$ Ker $B^T A^{-1} B$, that is $B^T A^{-1} B q = 0$. Introduce $u = A^{-1} B q$. We then see that

$$a_0(u,u) = u^T A u = q^T B^T A^{-T} A A^{-1} B q = q^T B^T A^{-T} B q = \Big(q^T B^T A^{-T} B q\Big)^T = q^T B^T A^{-1} B q = 0.$$

Using the coercivity of a_0, we deduce that $u = 0$, hence $Bq = Au = 0$. We infer from (4.20) *that $q = 0$. We hence have shown that Ker $B^T A^{-1} B = \{0\}$. The square matrix $B^T A^{-1} B$ is hence invertible.*

Assume conversely that the matrix $B^T A^{-1} B$ is invertible and that (4.20) *does not hold. This means that, for any integer n, there exists $q_n \in M$ with $\|q_n\| = 1$ such that*

$$\sup_{v \in X,\, v \neq 0} \frac{v^T B q_n}{\|v\|} \leq \frac{1}{n}.$$

The sequence q_n is bounded in the finite dimensional space M. It thus converges to some $q^\star$, up to the extraction of a subsequence, which satisfies $\|q^\star\| = 1$. For any $v \in X$, we have

$$\frac{v^T B q_n}{\|v\|} \leq \sup_{v \in X,\, v \neq 0} \frac{v^T B q_n}{\|v\|} \leq \frac{1}{n},$$

which yields, passing to the limit $n \to \infty$, that

$$\frac{v^T B q^\star}{\|v\|} \leq 0.$$

Taking $v = Bq^\star$, we get $Bq^\star = 0$, and thus $B^T A^{-1} B q^\star = 0$, which implies that $q^\star = 0$ since $B^T A^{-1} B$ is invertible. This is in contradiction with $\|q^\star\| = 1$. We have thus shown that condition (4.20) *holds.*

4.3.4 Application to the Stokes problem

The mathematical analysis of the Stokes problem relies on the following result, the proof of which goes beyond the scope of these lecture notes (see e.g. [6, Lemma 3.1]).

Theorem 4.18. *Let Ω be an open, bounded and connected subset of $\mathbb{R}^d$. Then the mapping*

$$\text{div } : (H^1_0(\Omega))^d \longrightarrow L^2_0(\Omega)$$

is surjective.

The above theorem means that, for any $q \in L^2_0(\Omega)$, there exists $v_q \in (H^1_0(\Omega))^d$ such that $\text{div } v_q = q$. Using the Open Mapping theorem (see e.g. [4]), we obtain the existence of a constant $\rho > 0$ such that $\rho \|v_q\|_{H^1(\Omega)} \leq \|q\|_{L^2(\Omega)}$.

Corollary 4.19. *Under the assumptions of Theorem 4.18, the condition* (4.20) *is satisfied for the Stokes problem, which corresponds to the choices $X = (H^1_0(\Omega))^d$, $M = L^2_0(\Omega)$ and $b(v,p) = -\displaystyle\int_\Omega p \,\text{div}\, v$.*

Proof. For any $q \in L^2_0(\Omega)$, we have

$$\begin{aligned}
\sup_{v \in X,\, v \neq 0} \frac{b(v,q)}{\|v\|_X} &= \sup_{v \in X,\, v \neq 0} \frac{\displaystyle\int_\Omega q \,\text{div}\, v}{\|v\|_{H^1(\Omega)}} \\
&\geq \frac{\displaystyle\int_\Omega q \,\text{div}\, v_q}{\|v_q\|_{H^1(\Omega)}} \\
&\geq \rho \frac{\displaystyle\int_\Omega q^2}{\|q\|_{L^2(\Omega)}} \\
&= \rho \|q\|_M.
\end{aligned}$$

We thus obtain (4.20). □

We are now in position to conclude:

Theorem 4.20. *Under the assumptions of Theorem 4.18, the Stokes problem* (4.16) *is well-posed. In addition, there exists $c > 0$ such that*

$$\forall f \in (L^2(\Omega))^d, \qquad \|u\|_{H^1(\Omega)} + \|p\|_{L^2(\Omega)} \leq c \, \|f\|_{L^2(\Omega)}. \tag{4.21}$$

Proof. We recall that the PDE (4.16) is equivalent to the variational formulation (4.17) (see Proposition 4.15), which we have recast in the form (4.19).

We now verify the assumptions of Theorem 4.16. We have that $X = (H^1_0(\Omega))^d$ and $M = L^2_0(\Omega)$ are two Hilbert spaces. In view of the Poincaré inequality, the bilinear form a_0 is coercive on X. The condition (4.20) is satisfied thanks to Corollary 4.19. We thus deduce that (4.19), and hence (4.16), is well-posed. The estimate (4.21) is a rewriting of (4.11). □

4.4 Saddle-point problems

We discuss here in more generality problems of the type (4.19), where we assume that X and M are two Hilbert spaces, that a_0 is a continuous bilinear form on $X \times X$, b is a continuous bilinear form on $X \times M$, and g_1 (resp. g_2) is a continuous form on X (resp. M).

Definition 4.21. *If the bilinear form a_0 is symmetric and positive on $X \times X$, in the sense that $a_0(v,v) \geq 0$ for any $v \in X$, then the problem* (4.19) *is said to be a saddle-point problem.*

To understand this definition, introduce the so-called lagrangian $L : X \times M \to \mathbb{R}$ defined by

$$L(v,q) = \frac{1}{2} a_0(v,v) + b(v,q) - g_1(v) - g_2(q).$$

Definition 4.22. *The point $(u,p) \in X \times M$ is called a* saddle point *of L if*

$$\forall (v,q) \in X \times M, \qquad L(u,q) \leq L(u,p) \leq L(v,p). \tag{4.22}$$

Proposition 4.23. *Let X and M be two Hilbert spaces. Assume that the bilinear form a_0 is symmetric and positive on $X \times X$. Then, $(u,p) \in X \times M$ is a saddle point of L in the sense of* (4.22) *if and only if (u,p) is a solution to* (4.19).

Proof. We first observe that

- for any $u \in X$, the function $q \in M \mapsto L(u,q)$ is affine;
- for any $p \in M$, the function $v \in X \mapsto L(v,p)$ is convex: for any v_1 and v_2 in X and $\alpha \in (0,1)$, we indeed have, using the symmetry of a_0, that

$$\begin{aligned}
& 2\Big(L(\alpha v_1 + (1-\alpha) v_2, p) - \alpha L(v_1,p) - (1-\alpha) L(v_2,p)\Big) \\
= \;& a_0(\alpha v_1 + (1-\alpha) v_2, \alpha v_1 + (1-\alpha) v_2) - \alpha a_0(v_1,v_1) - (1-\alpha) a_0(v_2,v_2) \\
= \;& \alpha^2 a_0(v_1,v_1) + (1-\alpha)^2 a_0(v_2,v_2) + 2\alpha(1-\alpha) a_0(v_1,v_2) - \alpha a_0(v_1,v_1) - (1-\alpha) a_0(v_2,v_2) \\
= \;& 2\alpha(1-\alpha) a_0(v_1,v_2) - \alpha(1-\alpha) a_0(v_1,v_1) - \alpha(1-\alpha) a_0(v_2,v_2) \\
= \;& -\alpha(1-\alpha) a_0(v_1 - v_2, v_1 - v_2).
\end{aligned}$$

  Since a_0 is positive, we get that

$$L(\alpha v_1 + (1-\alpha) v_2, p) \leq \alpha L(v_1,p) + (1-\alpha) L(v_2,p),$$

  which means that $v \in X \mapsto L(v,p)$ is convex.

- Since a_0 is symmetric, we compute that

$$d_{(v,q)} L(u,p)(\overline{u}, \overline{p}) = a_0(u, \overline{u}) + b(u, \overline{p}) + b(\overline{u}, p) - g_1(\overline{u}) - g_2(\overline{p}).$$

  We thus see that (u,p) is such that $d_{(v,q)} L(u,p) = 0$ if and only if

$$\forall \overline{u} \in X, \qquad a_0(u, \overline{u}) + b(\overline{u}, p) = g_1(\overline{u})$$

  and

$$\forall \overline{p} \in M, \qquad b(u, \overline{p}) = g_2(\overline{p}).$$

  Hence, (u,p) is such that $d_{(v,q)} L(u,p) = 0$ if and only if (u,p) is a solution to (4.19).

We now prove the proposition. Assume first that (u,p) is a solution to (4.19). Then, thanks to the third observaiton above, we have $d_{(v,q)}L(u,p) = 0$. Define J_1 on X by $J_1(v) = L(v,p)$. We then get that $d_v J_1(u) = 0$. Using the convexity of J_1 (second observation above), we deduce that $J_1(u) \leq J_1(v)$ for any $v \in X$, which we write as $L(u,p) \leq L(v,p)$ for any $v \in X$. Define now J_2 on M by $J_2(q) = L(u,q)$. We also have that $d_q J_2(p) = 0$. Since J_2 is affine (first observation above), we deduce that J_2 is constant on M, which we write as $L(u,p) = L(u,q)$ for any $q \in M$.

Assume conversely that $(u,p) \in X \times M$ is a saddle point of L in the sense of (4.22), and define J_1 and J_2 as above. We hence have that $J_1(u) \leq J_1(v)$ for any $v \in X$ and that $J_2(q) \leq J_2(p)$ for any $q \in M$, and hence $d_v J_1(u) = 0$ and $d_q J_2(p) = 0$. We then compute that

$$\begin{aligned} d_{(v,q)}L(u,p)(\overline{u},\overline{p}) &= a_0(u,\overline{u}) + b(u,\overline{p}) + b(\overline{u},p) - g_1(\overline{u}) - g_2(\overline{p}) \\ &= d_v J_1(u)(\overline{u}) + d_q J_2(p)(\overline{p}) \\ &= 0. \end{aligned}$$

Using the third observation above, we deduce that (u,p) is a solution to (4.19). □

4.5 Exercises

Exercise 4.24 (Advection-diffusion problem in a non-coercive setting)**.** *We consider the advection-diffusion problem studied in Section 3.5, that is: Find $u \in H^1_0(\Omega)$ such that*

$$-\Delta u + c \cdot \nabla u = f \qquad \text{in } \mathcal{D}'(\Omega), \tag{4.23}$$

where Ω is a bounded open subset of $\mathbb{R}^d$, $c : \Omega \mapsto \mathbb{R}^d$ is a smooth vector field and $f \in L^2(\Omega)$.

In contrast to Section 3.5, we make no assumptions on c (in particular, its divergence is not small, it is not non-negative, and c is not small). We wish to establish that (4.23) *is well-posed.*

1. *Write the variational formulation of* (4.23)*.*
2. *We assume that there exists a function $\sigma \in C^1(\overline{\Omega})$ such that*

$$-\text{div }(\nabla\sigma + c\,\sigma) = 0$$

and such that $\sigma(x) \geq c_\sigma > 0$ almost everywhere on Ω.

   (a) *In the case when $c = \nabla V$ for a smooth function V, show that such a function σ exists by giving its expression in terms of V.*

   (b) *We admit that, as soon as Ω and c are sufficiently smooth, there exists σ satisfying the above assumptions. Using Theorem 4.13, show that* (4.23) *is well-posed.*

Exercise 4.25 (Another viewpoint on the Stokes equation)**.** *Rather than considering the variational formulation* (4.17) *of the Stokes problem* (4.16)*, which takes into account the equation* div $u = 0$*, it can be tempting to include this equation in the space in which we work. We thus introduce the space*

$$V = \left\{ u \in (H^1_0(\Omega))^d, \quad \text{div } u = 0 \right\}$$

and the variational formulation:

$$\left\{ \begin{array}{l} \text{Find } u \in V \text{ such that} \\ \forall w \in V, \qquad a(u,w) = g(w) \end{array} \right. \tag{4.24}$$

where

$$a(u,w) = \int_\Omega \nabla u \cdot \nabla w, \qquad g(w) = \int_\Omega f \cdot w.$$

1. *We first establish the well-posedness of* (4.24)*.*

   (a) *Recall why the space V is a Hilbert space when endowed with the H^1 scalar product.*

   (b) *Recall why the bilinear form a is continuous on $V \times V$ and why the linear form g is continuous on V.*

   (c) *Show that* (4.24) *is well-posed.*

   (d) *Show that* (4.24) *is equivalent to a minimization problem of the form* $\inf \{J(v), \quad v \in V\}$ *for some functional* $J : V \to \mathbb{R}$*.*

2. *Show that, if $x = (u, p)$ is a solution to* (4.17)*, then u is a solution to* (4.24)*.*

   Since we know that (4.17) *and* (4.24) *are well-posed, this implies that, if u is a solution to* (4.24)*, then there exists $p \in L^2_0(\Omega)$ such that $x = (u, p)$ is a solution to* (4.17)*. We explain below how to obtain this p in practice.*

3. *Let u be a solution to* (4.24)*. Our aim here is to compute $p \in L^2_0(\Omega)$ such that $x = (u, p)$ is a solution to* (4.17)*. Let* $L(w) = g(w) - a(u, w) = \displaystyle\int_\Omega f \cdot w - \int_\Omega \nabla u \cdot \nabla w$*. We know that* $L(w) = 0$ *for any* $w \in V \subset (H^1_0(\Omega))^d$*, and we wish to*

   $$\text{Find } p \in L^2_0(\Omega) \text{ such that, for all } w \in (H^1_0(\Omega))^d, \text{ we have } L(w) = b(w, p), \tag{4.25}$$

   where we recall that $b(w, p) = -\displaystyle\int_\Omega p \operatorname{div} w$*.*

   (a) *Using Corollary 4.19, show that there exists at most one solution to* (4.25)*.*

   (b) *One could think of using the general BNB theorem 4.13 to solve* (4.25)*. The Corollary 4.19 show that Condition* (4.12) *is satisfied. Show that Condition* (4.13) *is not satisfied. This is why we have to argue in a different manner.*

   (c) *On the space $L^2_0(\Omega)$, we introduce the bilinear form* $A(p, q) = \displaystyle\int_\Omega p\, q$ *and the linear form* $G(q) = -L(v_q)$*, where $v_q \in (H^1_0(\Omega))^d$ is the velocity field defined by Theorem 4.18:* $\operatorname{div} v_q = q$ *and* $\|v_q\|_{H^1_0(\Omega)} \leq C \|q\|_{L^2(\Omega)}$*.*
   Show that G is independent of the choice of v_q, as soon as $\operatorname{div} v_q = q$*.*

   (d) *Show that A is continuous and coercive on $L^2_0(\Omega)$, and that G is continuous on $L^2_0(\Omega)$. We note that G is only defined on $L^2_0(\Omega)$, and not on $L^2(\Omega)$.*

   (e) *Show that the problem*

   $$\text{Find } p \in L^2_0(\Omega) \text{ such that, for all } q \in L^2_0(\Omega), \text{ we have } A(p, q) = G(q)$$

   is well posed.

   (f) *Let $w \in (H^1_0(\Omega))^d$. We set* $q = \operatorname{div} w \in L^2_0(\Omega)$*. Check that*

   $$b(w, p) = -A(p, q) = -G(q) = L(v_q)$$

   and that $L(v_q) = L(w)$. Conclude.

We remark that the formulation (4.24) *directly provides the velocity u, but that the pressure p should be computed in a second stage, in a non-trivial manner. In practice, it is often the case that the quantity of interest in the Stokes problem is the pressure. This is why alternatives to* (4.24) *(such as* (4.17)*) have been proposed, despite the fact that their analysis needs more advanced tools than the Lax-Milgram lemma.*

The formulation (4.24) *is also challenging when it comes to its discretization. A natural choice is to introduce a conformal discretization $V_h \subset V$. However, it is not simple to manipulate (globally)* *divergence-free functions that are piecewise polynomials (especially in 3D). Furthermore, to recover the pressure, a possibility is to solve* (4.25) *in a discrete setting. To that aim, one needs the equivalent of Theorem 4.18 in a discrete setting. These are other reasons to prefer* (4.17), *the discretization of which is somewhat simpler.*

Chapter 5

Numerical approximation of boundary value problems (coercive case)

This chapter is devoted to the numerical approximation of linear elliptic boundary value problems, such as those considered in Chapter 3. We focus here on a simple case, namely the Poisson problem with homogeneous Dirichlet boundary conditions: we look for $u \in H^1(\Omega)$ solution to

$$\left\{ \begin{array}{ll} -\Delta u = f & \text{in } \mathcal{D}'(\Omega), \\ u = 0 & \text{on } \partial\Omega, \end{array} \right. \tag{5.1}$$

where Ω is a bounded subset of $\mathbb{R}^d$ and $f \in L^2(\Omega)$. We have proved in Section 3.3 the existence and uniqueness of a solution to (5.1).

Several types of numerical strategies can be used to approximate this solution. In what follows, we focus on the Finite Element Method. It is a very popular method, often used by engineers when addressing continuum mechanics problems, fluid mechanics problems, ... Although we focus here on a simple linear, elliptic, stationary (i.e. time independent) boundary value problem, we underline that the Finite Element Method can be used for many boundary value problems, not necessarily elliptic, and that may be time-dependent and/or nonlinear. It is indeed a very robust approach, that can be used for domains Ω of arbitrary geometry.

In all what follows, we will argue on the variational formulation associated to (5.1), which reads

$$\left\{ \begin{array}{l} \text{Find } u \in V \text{ such that} \\ \forall v \in V, \qquad a(u,v) = b(v), \end{array} \right. \tag{5.2}$$

where

- $V = H^1_0(\Omega)$ is a Hilbert space;
- the bilinear form a, which is defined by

  $$a(u,v) = \int_\Omega \nabla u \cdot \nabla v,$$

  is continuous on $V \times V$ and coercive;

- the linear form b, which is defined by

  $$b(v) = \int_\Omega f\, v,$$

  is continuous on V.

We recall that the problems (5.1) and (5.2) are equivalent, and that Problem (5.2) is well-posed in view of the Lax-Milgram theorem.

Remark 5.1. *In the present case, the bilinear form a is also symmetric. We do not make this assumption in general and will clearly underline the arguments that are based on this additional property.*

5.1 The Galerkin approach

Many approximation methods for (5.1) belong to the class of Galerkin methods. This is the case of Spectral Methods, and of the Finite Element Method on which we focus here. In contrast, finite difference approaches are not Galerkin methods.

In this section, we detail the general principle of Galerkin approaches. We consider the abstract problem (5.2), assuming that

$$V \text{ is a Hilbert space,} \tag{5.3}$$

that the linear form b is continuous on V, namely that there exists C such that

$$\forall v \in V, \quad |b(v)| \leq C\|v\|_V, \tag{5.4}$$

and that the bilinear form a is continuous on $V \times V$ and coercive: there exist M and $\alpha > 0$ such that

$$\forall u, v \in V, \quad |a(u,v)| \leq M\, \|u\|_V\, \|v\|_V \quad \text{and} \quad a(v,v) \geq \alpha \|v\|_V^2. \tag{5.5}$$

We recall that we do *not* assume the bilinear form a to be symmetric.

5.1.1 Principle of the method

The Galerkin[1] approach consists in replacing, in (5.2), the space V (which is usually of infinite dimension) by a space V_h of finite dimension. The numerical approximation u_h of $u \in V$ will be searched in V_h. The space V_h is thus called the *approximation space*. The presence of the subscript h encodes the fact that V_h is of finite dimension. For instance, in the Finite Element Method, it directly refers to the size of the mesh which is used to build V_h (see Section 5.2.1 below). In Spectral Methods, it relates in a more implicit manner to some characteristic of the approximation space V_h.

In what follows, we assume that

$$V_h \subset V.$$

In that case, the approximation method is said to be *conformal*. In the sequel, we approximate Problem (5.2) by the following problem:

$$\left\{ \begin{array}{l} \text{Find } u_h \in V_h \text{ such that} \\ \forall v_h \in V_h, \qquad a(u_h, v_h) = b(v_h). \end{array} \right. \tag{5.6}$$

[1] After the name of the russian mathematician Boris Grigorievich Galerkin, 1871–1945.

Remark 5.2. *It is also possible to introduce* non-conformal *methods. When $V = H^1(\Omega)$, a typical example is to introduce a partition $\cup K$ of Ω (as is usually done in a Finite Element Method), and to consider $V_h = \oplus H^1(K)$, i.e. $v_h \in V_h$ if and only if $v_h = \sum_K v_h^K$ for some $v_h^K \in H^1(K)$.*

Consider an edge Γ shared by two elements K_1 and K_2. Since no contraints are imposed on the values of $v_h^{K_1}$ and $v_h^{K_2}$ on that edge, the function v_h does not belong to $H^1(K_1 \cup K_2)$ in general. As a consequence, in general, $V_h \not\subset V$.

Remark 5.3. *In some cases, it is difficult to work in practice with the exact forms a and b, because they e.g. involve integrals that are difficult to exactly compute. In such cases, it is possible to further approximate* (5.6) *by considering the problem*

$$\left\{\begin{array}{l} \textit{Find } u_h \in V_h \textit{ such that} \\ \forall v_h \in V_h, \qquad a_h(u_h, v_h) = b_h(v_h), \end{array}\right.$$

where a_h (resp. b_h) is an approximation of a (resp. b). For instance, integrals arising in the definition of the bilinear form a may be approximated by numerical quadrature rules. We refer to Exercise 5.31 for more details.

In some other cases, it is actually advantageous to modify the forms a and b, in order to improve the numerical approximation. We refer to Exercise 5.34 for an example.

Remark 5.4. *In the problem* (5.2)*, the space V is both the space to which the solution u belongs, and the space in which the test functions v are. When introducing a finite dimensional problem, it is possible to use two different approximations of this space, and to define the discrete problem as*

$$\left\{\begin{array}{l} \textit{Find } u_h \in V_h \textit{ such that} \\ \forall v_h \in W_h, \qquad a(u_h, v_h) = b(v_h), \end{array}\right.$$

where, for instance, V_h and W_h are two finite dimensional subspaces of V. Such approaches are called Petrov-Galerkin approaches. Their analysis goes beyond the scope of these lecture notes.

Proposition 5.5. *Under assumptions* (5.3), (5.4) *and* (5.5)*, the problem* (5.6) *has one and only one solution.*

Proof. We consider the space $V_h \subset V$ endowed with the scalar product $\langle \cdot, \cdot \rangle_V$ of the Hilbert space V. Since V_h is of finite dimension, it is closed in V, and thus V_h is a Hilbert space. The forms a and b are continuous on V_h. The form a is coercive on V, and thus on V_h. We are thus in position to use the Lax-Milgram theorem on (5.6) and conclude that this problem is well-posed. □

5.1.2 Error estimation

We wish to bound from above the error $e = u - u_h$ in the ambiant norm $\| \cdot \|_V$. We first observe an important property:

Lemma 5.6 (Galerkin orthogonality)**.** *The error $e = u - u_h$ satisfies*

$$\forall v_h \in V_h, \quad a(e, v_h) = 0. \tag{5.7}$$

Proof. The exact solution u satisfies $a(u, v) = b(v)$ for any $v \in V$. Since $V_h \subset V$, we can take a test function in V_h and thus deduce that $a(u, v_h) = b(v_h)$ for any $v_h \in V_h$. Using that $a(u_h, v_h) = b(v_h)$ for any $v_h \in V_h$, we immediately get the result. □

We are now in position to establish the celebrated Céa lemma, which states that the error $u-u_h$ is bounded from above (up to an explicit multiplicative constant) by the best approximation error:

Lemma 5.7 (Céa lemma). *We suppose that assumptions* (5.3), (5.4) *and* (5.5) *hold. Let u be the solution to* (5.2) *and u_h be the solution to* (5.6). *Then*

$$\|u - u_h\|_V \leq \frac{M}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_V. \tag{5.8}$$

Proof. For any $v_h \in V_h$, we deduce from the Galerkin orthogonality (5.7) that $a(u-u_h, u_h-v_h) = 0$. Therefore, we have

$$\alpha \|u - u_h\|_V^2 \leq a(u - u_h, u - u_h) = a(u - u_h, u - v_h) \leq M \, \|u - u_h\|_V \, \|u - v_h\|_V.$$

We can now divide by $\|u - u_h\|_V$ to obtain

$$\|u - u_h\|_V \leq \frac{M}{\alpha} \, \|u - v_h\|_V.$$

As the function v_h is arbitrary in V_h, we deduce (5.8). □

The quantity $\inf\limits_{v_h \in V_h} \|u - v_h\|_V$ is usually called the best approximation error. It is an upper bound (up to a multiplicative constant) of the error $u - u_h$. Obviously, it is also a lower bound of the error, since $\inf\limits_{v_h \in V_h} \|u - v_h\|_V \leq \|u - u_h\|_V$. If the best approximation error converges to 0, then so does the error $u - u_h$, at the same rate.

We also note that the best approximation error quantifies how well the approximation space V_h is suited to approximate u. It is natural that this best approximation error bounds from above the error. An important consequence of the Céa lemma is the fact that, if the approximation space is well-chosen, namely if the best approximation error is small, then the numerical approach is accurate.

Remark 5.8 (Geometric interpretation of the Galerkin orthogonality). *In the specific case when a is symmetric, it induces a scalar product on V. The norm associated to this scalar product is equivalent to the norm $\|\cdot\|_V$, as a consequence of the continuity and coercivity of a: for any $v \in V$,*

$$\alpha \|v\|_V^2 \leq a(v, v) \leq M \|v\|_V^2.$$

In that case, one can understand (5.7) *as the fact that u_h is the orthogonal projection on V_h of the exact solution $u \in V$ (orthogonal with respect to the scalar product $a(\cdot, \cdot)$).*

Remark 5.9 (Error estimate, the symmetric case). *When a is symmetric, it is possible to get a better error estimate than* (5.8). *Indeed, we have, for any $v_h \in V_h$,*

$$a(u - v_h, u - v_h) = a(u - u_h + w_h, u - u_h + w_h)$$

with $w_h = u_h - v_h$, hence

$$a(u - v_h, u - v_h) = a(u - u_h, u - u_h) + a(w_h, w_h) + a(u - u_h, w_h) + a(w_h, u - u_h).$$

Using the symmetry of a, the last two terms are equal. Using the Galerkin orthogonality (5.7)*, they both vanish, thus*

$$a(u-v_h,u-v_h)=a(u-u_h,u-u_h)+a(w_h,w_h)\geq a(u-u_h,u-u_h),$$

where we used the coercivity of a in the last bound. We thus get that

$$\alpha\|u-u_h\|_V^2\leq a(u-u_h,u-u_h)\leq a(u-v_h,u-v_h)\leq M\,\|u-v_h\|_V^2,$$

hence

$$\|u-u_h\|_V\leq\sqrt{\frac{M}{\alpha}}\inf_{v_h\in V_h}\|u-v_h\|_V.$$

This bound is better than (5.8) *as* $M/\alpha\geq 1$.

5.1.3 The linear system

The space V_h is of finite dimension, therefore the problem (5.6) can actually be recasted as a linear system, as we now show. Denote N the dimension of V_h and let $(\varphi_1,\ldots,\varphi_N)$ be a basis of V_h. We expand $u_h\in V_h$ on the basis of V_h as

$$u_h=\sum_{j=1}^N u_j\,\varphi_j.$$

By considering φ_i as test function, we see that the problem (5.6) is equivalent to finding $U=(u_1,\ldots,u_N)\in\mathbb{R}^N$ such that

$$\forall 1\leq i\leq N,\qquad \sum_{j=1}^N a(\varphi_j,\varphi_i)\,u_j=b(\varphi_i).$$

We thus introduce the matrix $A\in\mathbb{R}^{N\times N}$ defined by

$$A_{ij}=a(\varphi_j,\varphi_i)$$

and the vector $B\in\mathbb{R}^N$ defined by

$$B_i=b(\varphi_i),$$

and recast the above problem as the linear system

$$AU=B. \tag{5.9}$$

The matrix A is called the *stiffness matrix*, by reference to problems in mechanics where it has first been introduced. This matrix has properties that directly come from the properties satisfied by the bilinear form a.

Lemma 5.10. *If the bilinear form a is symmetric, then the matrix A is symmetric. If the bilinear form a is coercive, then the matrix A is positive definite.*

Proof. The first assertion is obvious. We only show the second. Let $\xi = (\xi_1, \dots, \xi_N) \in \mathbb{R}^N$ and set $u = \sum_{i=1}^{N} \xi_i \, \varphi_i$ (note that $u \in V_h$). Denoting by $(\cdot, \cdot)_{\mathbb{R}^N}$ the scalar product in $\mathbb{R}^N$, we see that

$$\begin{aligned}
(\xi, A\xi)_{\mathbb{R}^N} &= \sum_{i,j=1}^{N} \xi_i \, A_{ij} \, \xi_j \\
&= \sum_{i,j=1}^{N} \xi_i \, \xi_j a(\varphi_j, \varphi_i) \\
&= a\left(\sum_{j=1}^{N} \xi_j \, \varphi_j, \sum_{i=1}^{N} \xi_i \, \varphi_i \right) \\
&= a(u, u).
\end{aligned}$$

Assume that a is coercive. Then, for any $\xi \in \mathbb{R}^N$, we have $(\xi, A\xi)_{\mathbb{R}^N} \geq 0$. Furthermore, if $(\xi, A\xi)_{\mathbb{R}^N} = 0$, we get $a(u, u) = 0$, hence $u = 0$ by coercivity, hence $\xi = 0$. The matrix A is indeed positive definite. □

5.2 The Lagrange $\mathbb{P}_1$ Finite Element

We have described above the Galerkin approach in general terms. We now detail it in a specific case of approximation space V_h. More precisely, we now describe the simplest Finite Element Method, based on the so-called Lagrange $\mathbb{P}_1$ Finite Element. We assume that the domain Ω is a polygon (see Remark 5.21 below).

For simplicity, we assume that the dimension d is equal to 2. The approach is however not restricted to this case.

In the sequel, we will use the notation $| \cdot |_{H^2(\Omega)}$ that we define here. We have previously introduced the H^2 norm $\| \cdot \|_{H^2(\Omega)}$: for any $v \in H^2(\Omega)$,

$$\|v\|_{H^2(\Omega)} = \sqrt{\|v\|^2_{L^2(\Omega)} + \|\nabla v\|^2_{L^2(\Omega)} + \|\nabla^2 v\|^2_{L^2(\Omega)}}.$$

For any $v \in H^2(\Omega)$, we define

$$|v|_{H^2(\Omega)} = \|\nabla^2 v\|_{L^2(\Omega)}$$

so that

$$|v|^2_{H^2(\Omega)} = \sum_{i,j=1}^{d} \left\| \frac{\partial^2 v}{\partial x_i \partial x_j} \right\|^2_{L^2(\Omega)}.$$

5.2.1 Meshing the domain

Definition 5.11. *A mesh of Ω is a covering of the polygon Ω by triangles (by convention, these triangles are considered as closed sets, whereas Ω is an open set). Denoting $\mathcal{T}_h = \{K_1, \dots, K_{N_e}\}$ these triangles (where N_e is the number of triangles in the mesh), we thus have*

$$\overline{\Omega} = \cup_{i=1}^{N_e} K_i.$$

The mesh is said to be admissible *if, for any $i \neq j$, the set $K_i \cap K_j$ is either empty, or restricted to a point which is a vertex of K_i and K_j, or equal to a segment which is an edge of K_i and K_j.*

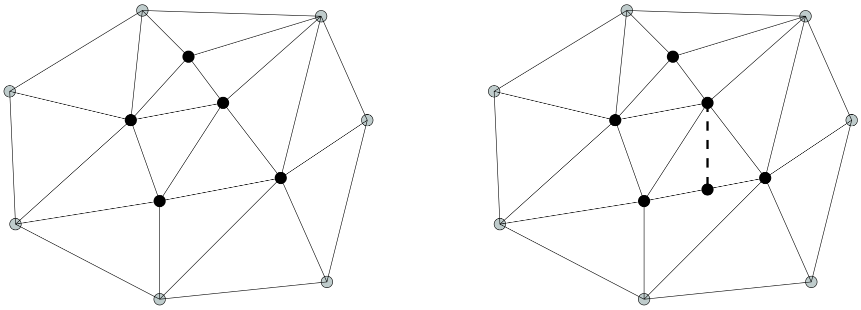

Figure 5.1: Example (left) and counter-example (right) of an admissible mesh

An example and a counter-example of an admissible mesh are shown on Figure 5.1.

For any $1 \leq i \leq N_e$, we denote h_i the diameter of the triangle K_i (defined as the longuest edge of K_i) and we set

$$h = \max_{1 \leq i \leq N_e} h_i.$$

This parameter quantifies how fine the mesh is. We denote $\{s_{i,1}, s_{i,2}, s_{i,3}\}$ the three vertices of the triangle K_i. Collecting these vertices over the different triangles, we obtain the set $\{s_1, \dots, s_{N_s}\}$ of the mesh vertices. It is useful to distinguish the vertices inside the open set Ω (there are $N_s^{\rm int}$ of them) and the vertices on $\partial\Omega$ (there are $N_s^{\rm bound}$ of them). By construction, $N_s = N_s^{\rm int} + N_s^{\rm bound}$.

5.2.2 The space of $\mathbb{P}_1$ polynomial functions

In dimension $d = 2$, we set

$$\mathbb{P}_1 = \left\{ p : \mathbb{R}^2 \to \mathbb{R}, \ \ p(x,y) = \alpha + \beta x + \gamma y \ \text{ for some real numbers } \alpha, \ \beta \text{ and } \gamma \right\}.$$

This is a vector space of dimension 3. The following result is of paramount importance when building the $\mathbb{P}_1$ Lagrange Finite Element:

Proposition 5.12. *A function $p \in \mathbb{P}_1$ is completely determined by its value at three non-aligned points. Furthermore, its value on a segment (not restricted to a point) is completely determined by its value at the two end points of this segment.*

Consider a triangle K in $\mathbb{R}^2$, that we assumed to be non-degenerate (i.e. its three vertices are not aligned). Let a^1, a^2 and a^3 be its three vertices. Thanks to the above proposition, there exists a unique function $\lambda_1 \in \mathbb{P}_1$ such that

$$\lambda_1(a^1) = 1, \quad \lambda_1(a^2) = 0, \quad \lambda_1(a^3) = 0.$$

Likewise, there exists a unique function $\lambda_2 \in \mathbb{P}_1$ such that $\lambda_2(a^1) = 0$, $\lambda_2(a^2) = 1$ and $\lambda_2(a^3) = 0$, and a unique function $\lambda_3 \in \mathbb{P}_1$ such that $\lambda_3(a^1) = 0$, $\lambda_3(a^2) = 0$ and $\lambda_3(a^3) = 1$. The functions λ_1, λ_2 and λ_3 are called the *barycentric coordinates* of the triangle K. They satisfy the following (straightforward) properties:

- $\lambda_1 + \lambda_2 + \lambda_3 = 1$;
- for any $i \in \{1,2,3\}$, λ_i is vanishing on the edge of K which is opposite to the vertex a^i;
- for any $x \in K$ and any $i \in \{1,2,3\}$, we have $0 \leq \lambda_i(x) \leq 1$;
- denoting G the barycenter of K, we have $\lambda_i(G) = 1/3$ for any $i \in \{1,2,3\}$.

5.2.3 Approximation space

Set

$$V_h^{(1)} = \left\{ v_h \in C^0(\overline{\Omega}), \quad v_h|_K \in \mathbb{P}_1 \text{ for any } K \in \mathcal{T}_h, \quad v_h = 0 \text{ on } \partial\Omega \right\}.$$

The functions in $V_h^{(1)}$ are globally continuous over $\overline{\Omega}$, and piecewise C^1. Using the jump formula, we get that $V_h^{(1)} \subset H^1(\Omega)$, the derivative (in the sense of distributions) being equal (on each triangle) as the standard derivative. Note that the gradient of $v_h \in V_h^{(1)}$ is constant on each triangle. Since the functions in $V_h^{(1)}$ vanish on $\partial\Omega$, we see that this approximation is conformal:

Lemma 5.13. *We have $V_h^{(1)} \subset H_0^1(\Omega)$.*

We now identify a basis of the space $V_h^{(1)}$. Let $s \in \Omega$ be a vertex of the mesh (note that s is not on the boundary $\partial\Omega$). Let $\mathcal{K}(s) \subset \mathcal{T}_h$ be the set of elements which have s as a vertex. For any element $K \in \mathcal{K}(s)$, we note $\lambda_{K,s}$ the barycentric coordinate associated to the vertex s in the element K. We then set

$$\varphi_s(x,y) = \begin{cases} \lambda_{K,s}(x,y) \text{ if } (x,y) \in K \text{ for } K \in \mathcal{K}(s), \\ 0 \text{ otherwise.} \end{cases}$$

An exemple of such function φ_s is shown on Figure 5.2. On that example, the support of φ_s consists of 6 triangles. By construction, φ_s is equal to 1 at the vertex s and vanishes at all the other vertices of the mesh. Furthermore, the value of a function in $\mathbb{P}_1$ on an edge is fully determined by its value at the end-points of the edge. We thus see that $\varphi_s \in C^0(\overline{\Omega})$. In addition, we have that $\varphi_s|_{\partial\Omega} = 0$ and that, for any element $K \in \mathcal{T}_h$, $\varphi_s|_K \in \mathbb{P}_1$. This implies that

$$\varphi_s \in V_h^{(1)}.$$

Enumerating the internal vertices of the mesh as $\left\{s_1, \dots, s_{N_s^{\text{int}}}\right\}$, we order the functions built above as $\left\{\varphi_1, \dots, \varphi_{N_s^{\text{int}}}\right\}$, where φ_j is the function that is equal to 1 at vertex s_j and vanishes at all other vertices.

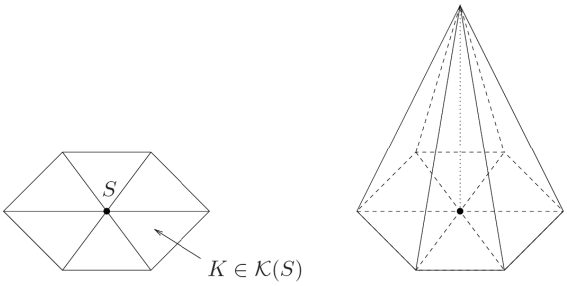

Figure 5.2: Example of function φ_s in dimension 2

The functions φ_j built above form a basis of $V_h^{(1)}$:

Proposition 5.14. *The set $\left\{\varphi_1, \dots, \varphi_{N_s^{\text{int}}}\right\}$ is a basis of the vector space $V_h^{(1)}$, which is of dimension N_s^{int}.*

Proof. The family $\{\varphi_1, \dots, \varphi_{N_s^{\rm int}}\}$ is free: if we have

$$\sum_j a_j \, \varphi_j = 0$$

for some real numbers a_j, then by evaluating this expression on the node s_i of the mesh, we get $a_i = 0$ since $\varphi_j(s_i) = \delta_{ij}$.

Consider now some $v_h \in V_h^{(1)}$. We consider the function w_h defined by

$$w_h = \sum_j v_h(s_j) \, \varphi_j.$$

On any element K, the functions v_h and w_h are affine and they are equal at the three vertices of K. They thus are equal on K. We thus have $v_h = w_h$ on $\overline{\Omega}$, which implies that any function in $V_h^{(1)}$ can be written as a linear combination of the functions φ_j. This concludes the proof. □

Our aim is now to estimate the best approximation error in $V_h^{(1)}$ of any function $u \in H_0^1(\Omega)$, namely to understand how the quantity $\inf\limits_{v_h \in V_h^{(1)}} \|u - v_h\|_{H^1(\Omega)}$ scales with respect to the mesh size h. To that aim, we introduce the interpolation operator

$$\begin{array}{rcl} I_h^{(1)} : C^0(\overline{\Omega}) & \to & V_h^{(1)} \\ v & \mapsto & v_h = \displaystyle\sum_{i=1}^{N_s^{\rm int}} v(s_i) \, \varphi_i. \end{array}$$

We see that $I_h^{(1)} v$ is the unique function in $V_h^{(1)}$ that takes the same values as v on all the internal nodes of the mesh. In the sequel, we estimate the interpolation errors $\left\| v - I_h^{(1)} v \right\|_{H^1(\Omega)}$ and $\left\| v - I_h^{(1)} v \right\|_{L^2(\Omega)}$ for any $v \in H^2(\Omega)$ (we recall that, in dimension $d \leq 3$, any function $v \in H^2(\Omega)$ has a continuous representation, hence $I_h^{(1)}$ is well-defined on $H^2(\Omega)$; see [1] for more details).

In dimension $d \geq 2$, the interpolation error depends on two parameters:

- the diameter of the elements $K \in \mathcal{T}_h$ (as in the one-dimensional case); we denoted h_i the diameter of K_i.
- the radius of the largest cercle included in K_i, that we denote ρ_i.

We set

$$\sigma_{\mathcal{T}_h} = \max_{1 \leq i \leq N_e} \frac{h_i}{\rho_i},$$

which is larger than 1, by construction. The ratio h_i/ρ_i increases when the smallest angle in the triangle K_i decreases (and the ratio diverges when this angle tends to 0). We admit the following result:

Theorem 5.15. *Let $d \leq 3$. There exists a constant $c_{\rm inter}$, independent of the mesh, such that, for any $v \in H^2(\Omega) \cap H_0^1(\Omega)$, we have*

$$\left\| v - I_h^{(1)} v \right\|_{H^1(\Omega)} \leq c_{\rm inter} \, \sigma_{\mathcal{T}_h} \, h \, |v|_{H^2(\Omega)} \quad \text{and} \quad \left\| v - I_h^{(1)} v \right\|_{L^2(\Omega)} \leq c_{\rm inter} \, h^2 \, |v|_{H^2(\Omega)}. \qquad (5.10)$$

Remark 5.16. *The fact that $v = I_h^{(1)} v = 0$ on $\partial\Omega$ plays no role to obtain (5.10). The same estimate holds for any $v \in H^2(\Omega)$, where $I_h^{(1)} v$ is now defined by $I_h^{(1)} v = \sum_{i=1}^{N_s} v(s_i)\, \varphi_i$ (note that the sum in i goes over all nodes of the mesh, and not only over the internal nodes).*

Remark 5.17. *In dimension $d \geq 4$, a function in $H^2(\Omega)$ is not necessarily continuous, and hence $I_h^{(1)} v$ is not defined. However, the following result, which is practice sufficient to perform the numerical analysis, holds: there exists a constant C such that, for any $v \in H^2(\Omega) \cap H_0^1(\Omega)$, we have*

$$\inf_{v_h \in V_h^{(1)}} \|v - v_h\|_{H^1(\Omega)} \leq C\, h\, |v|_{H^2(\Omega)}, \tag{5.11}$$

where C is independent of v and h, but may depend on the mesh geometry.

Remark 5.18. *In (5.10), the H^1 bound may be attained in some simple situations (and thus $\left\| v - I_h^{(1)} v \right\|_{H^1(\Omega)}$ indeed depends on $\sigma_{\mathcal{T}_h}$). Consider the case when the mesh consists of a unique triangle, with vertices $(0,0)$, $(1,0)$ and $(-1,\varepsilon)$. The diameter h is of the order of 1 whereas the parameter ρ is of the order of ε. Let $v(x,y) = x^2$. We then check that $\left[I_h^{(1)} v \right](x,y) = x + 2y/\varepsilon$, so that $\dfrac{\left\| v - I_h^{(1)} v \right\|_{H^1(\Omega)}}{|v|_{H^2(\Omega)}} \sim \varepsilon^{-1}$ when $\varepsilon \to 0$. The same situation may occur in a general mesh (i.e. not restricted to a single triangle), for a thin triangle with two angles close to 0 and the third angle close to π.*

Remark 5.19. *The Céa lemma 5.7 is written in the H^1 norm, which is the one for which the bilinear form a is coercive. It cannot be written in the L^2 norm, in the sense that the inequality $\|u - u_h\|_{L^2(\Omega)} \leq C \inf_{v_h \in V_h} \|u - v_h\|_{L^2(\Omega)}$ does not hold in general. An estimate on the best approximation error in L^2 norm (such as the estimate $\left\| v - I_h^{(1)} v \right\|_{L^2(\Omega)} \leq c_{\rm inter}\, h^2\, |v|_{H^2(\Omega)}$ of (5.10)) hence does not directly imply that $\|u - u_h\|_{L^2(\Omega)} \leq C h^2$. We refer to Exercise 5.37 for L^2 estimates on $u - u_h$.*

In practice, we do not work with a single mesh $\mathcal{T}_h$, but with a sequence of meshes obtained one from the other by sequential refinements (e.g., each triangle of a coarse mesh is subdivided into several triangles to obtain a finer mesh). We denote $\left\{ \mathcal{T}_{h_j} \right\}_{j \geq 1}$ this sequence of meshes.

Definition 5.20. *Let $\sigma_0 \in \mathbb{R}$ be fixed. The sequence of meshes $\left\{ \mathcal{T}_{h_j} \right\}_{j \geq 1}$ is said to be* regular *of parameter σ_0 if, for any j, we have*

$$\sigma_{\mathcal{T}_{h_j}} \leq \sigma_0.$$

Considering a regular sequence of meshes, we directly see that, for any $v \in H^2(\Omega) \cap H_0^1(\Omega)$ and any h,

$$\left\| v - I_h^{(1)} v \right\|_{H^1(\Omega)} \leq c_{\rm inter}\, \sigma_0\, h\, |v|_{H^2(\Omega)} \quad \text{and} \quad \left\| v - I_h^{(1)} v \right\|_{L^2(\Omega)} \leq c_{\rm inter}\, h^2\, |v|_{H^2(\Omega)}.$$

The interpolation error in norm H^1 (resp. in norm L^2) is therefore of the order of h (resp. of the order of h^2). This estimate is sharp with respect to h, in the sense that there exists a function $v \in H^2(\Omega) \cap H_0^1(\Omega)$ such that the interpolation error $v - I_h^{(1)} v$ is of the order of h (resp. h^2) in the H^1 norm (resp. in the L^2 norm).

Remark 5.21. *If Ω is not a polygon, then it cannot be meshed by triangles K_i such that $\overline{\Omega} = \cup_{i=1}^{N_e} K_i$. One possibility is to consider a polygon Ω_{poly} included in Ω, to mesh Ω_{poly} by the triangulation $\mathcal{T}_h$, and to consider the discrete space*

$$\widetilde{V}_h^{(1)} = \left\{ v_h \in C^0(\overline{\Omega}), \quad v_h|_K \in \mathbb{P}_1 \text{ for any } K \in \mathcal{T}_h, \quad v_h = 0 \text{ on } \partial\Omega_{\text{poly}}, \quad v_h = 0 \text{ on } \Omega \setminus \Omega_{\text{poly}} \right\},$$

which satisfies $\widetilde{V}_h^{(1)} \subset H_0^1(\Omega)$. The interpolation error between $v \in H_0^1(\Omega)$ and some $v_h \in \widetilde{V}_h^{(1)}$ has two contributions: the first one comes from the error in Ω_{poly} and can be estimated as above, while the second one comes from the fact that $v_h = 0$ in $\Omega \setminus \Omega_{\text{poly}}$ whereas v does not necessarily vanish there.

Another possibility is to use an approximation space V_h which is not *included in $H_0^1(\Omega)$, in the spirit of Remark 5.2.*

We do not pursue in these directions in these lecture notes.

5.3 Approximation of the Poisson problem by the $\mathbb{P}_1$ Finite Element Method

Let $f \in L^2(\Omega)$. We wish to approximate the unique solution $u \in H^1(\Omega)$ to the Poisson problem (5.1) (which is also the unique solution to the variational formulation (5.2)) by a Galerkin approach on the finite dimensional space $V_h^{(1)}$ built in Section 5.2.

5.3.1 Discrete problem and error estimation

In view of (5.6), the discrete problem is:

$$\left\{ \begin{array}{l} \text{Find } u_h \in V_h^{(1)} \text{ such that} \\ \forall v_h \in V_h^{(1)}, \qquad a(u_h, v_h) = b(v_h). \end{array} \right. \tag{5.12}$$

We have seen in Section 5.1.3 that solving this discrete problem amounts to solving a linear system of the form $AU = B$ (see (5.9)), where the stiffness matrix A is of size $N_s^{\text{int}} \times N_s^{\text{int}}$, where $N_s^{\text{int}} = \dim V_h^{(1)}$ is the number of internal vertices in the mesh. Its generic term, for any $1 \leq i, j \leq N_s^{\text{int}}$, is given by

$$A_{ij} = a(\varphi_j, \varphi_i) = \int_\Omega \nabla\varphi_i \cdot \nabla\varphi_j, \tag{5.13}$$

where the basis functions φ_j of $V_h^{(1)}$ have been built in Section 5.2.3. The right-hand side is given, for any $1 \leq i \leq N_s^{\text{int}}$, by

$$B_i = b(\varphi_i) = \int_\Omega f\,\varphi_i. \tag{5.14}$$

In view of Lemma 5.10, the matrix A is symmetric definite positive, hence the linear system (5.9) has a unique solution U. It is related to the unique solution u_h of (5.12) by $u_h = \sum_{i=1}^{N_s^{\text{int}}} U_i\,\varphi_i$.

We now state an error estimate:

Theorem 5.22. *Let $u \in H^1(\Omega)$ be the solution to (5.2) and $u_h \in V_h^{(1)}$ be the solution to (5.12). We assume that the sequence of meshes is regular of parameter σ_0 and that the domain Ω is convex. Then there exists a constant c, that may depend on Ω and σ_0, but that is independent of h and f, such that*

$$\|u - u_h\|_{H^1(\Omega)} \leq c\,h\,\|f\|_{L^2(\Omega)}.$$

The estimate stated in the above theorem is sharp: there exist Poisson problems such that $\|u-u_h\|_{H^1(\Omega)}$ is indeed of the order of h. Another way to prove this sharpness is to recall that (i) the interpolation error estimate provided in Theorem 5.15 is sharp, and (ii) the error $\|u-u_h\|_{H^1(\Omega)}$ is bounded from above and from below (up to multiplicative constants that are independent of h) by the interpolation error.

Proof. Using the Céa lemma 5.7, we get

$$\|u-u_h\|_{H^1(\Omega)} \leq \frac{M}{\alpha} \inf_{v_h \in V_h^{(1)}} \|u-v_h\|_{H^1(\Omega)}, \tag{5.15}$$

where, for the problem of interest here, the continuity constant is equal to $M=1$ and the coercivity constant is such that $\alpha^{-1} = 1 + C_\Omega^2$, where C_Ω is the Poincaré constant of Ω (see Theorem 2.7).

We next wish to bound from above the best approximation error by the interpolation error (that is, to take $v_h = I_h^{(1)} u$ in (5.15)). For this v_h to be well-defined, it is sufficient (in dimension $d \leq 3$) that $u \in H^2(\Omega)$. This is indeed the case, due to the following regularity result: if Ω is convex and $f \in L^2(\Omega)$, then the unique solution u to (5.2) belongs to $H^2(\Omega)$ and satisfies $|u|_{H^2(\Omega)} \leq \mathcal{C}_\Omega \|f\|_{L^2(\Omega)}$, where the constant $\mathcal{C}_\Omega$ only depends on Ω.

The remainder of the proof is straightforward: taking $v_h = I_h^{(1)} u$, we deduce from (5.15) that

$$\|u-u_h\|_{H^1(\Omega)} \leq (1+C_\Omega^2) \inf_{v_h \in V_h^{(1)}} \|u-v_h\|_{H^1(\Omega)} \leq (1+C_\Omega^2)\ \|u - I_h^{(1)} u\|_{H^1(\Omega)}.$$

We are now in position to use the estimate (5.10) of the interpolation error:

$$\|u-u_h\|_{H^1(\Omega)} \leq (1+C_\Omega^2)\ c_{\text{inter}}\, \sigma_0\, h\, |u|_{H^2(\Omega)} \leq (1+C_\Omega^2)\ c_{\text{inter}}\, \sigma_0\, \mathcal{C}_\Omega\, h\, \|f\|_{L^2(\Omega)}.$$

This concludes the proof. □

Remark 5.23. *In dimension $d \geq 4$, the fact that $u \in H^2(\Omega)$ does not imply that $I_h^{(1)} u$ is well-defined. However, the proof can be performed by inserting* (5.11) *rather than* (5.10) *in* (5.15).

Remark 5.24 (Non-convex domains)**.** *In the case when the domain Ω is not convex, the solution u to* (5.2) *generally does not belong to $H^2(\Omega)$. It instead belongs to $H^{3/2+\varepsilon}(\Omega)$ for some ε satisfying $0 < \varepsilon \leq 1/2$. The parameter ε depends on how much Ω is not convex. More precisely, since Ω is not convex, there are angles between two consecutive edges of $\partial\Omega$ that are larger than π (in the extreme case of a crack, there is an angle of 2π between two consecutive edges). The parameter ε depends on the value of the largest of these (larger than π) angles.*

The interpolation $I_h^{(1)} u$ is not defined, but it is still possible to introduce a variant of this interpolation and estimate its distance to u in terms of h. The Finite Element approximation (5.12) *turns out to be again converging when $h \to 0$, but with a smaller rate. More precisely, we have $\|u-u_h\|_{H^1(\Omega)} \leq c\, h^{1/2+\varepsilon}$ where c is independent of h.*

Remark 5.25. *It is of course possible to deduce from Theorem 5.22 a bound on the error in the L^2 norm: $\|u-u_h\|_{L^2(\Omega)} \leq c\, h\, \|f\|_{L^2(\Omega)}$. However, this estimate is not optimal. We show in Exercise 5.37 how to obtain a better (and actually optimal) bound.*

5.3.2 Assembling the stiffness matrix

The structure of the stiffness matrix, namely the position of non-zero coefficients, is very different between the one-dimensional case and the multi-dimensional case. We recall that, in the former case, the stiffness matrix is tridiagonal. In the multi-dimensional case, there are two different situations:

- a first possibility is to consider a *structured* mesh, obtained as the tensorial product of a one-dimensional mesh in the x direction by a one-dimensional mesh in the y direction. The vertices are hence located on a tensorial grid. These meshes are well-adapted for rectangular domains Ω. If the same meshsize is used in both directions, one says that the mesh is *uniform.*

- another possibility (the so-called *non-structured* meshes) consists in considering (for the mesh vertices) a cloud of points without any geometrical structure. In practice, such meshes are more often used, as they can be used for domains Ω of arbitrary shape. It is also possible to locally refine the mesh, in order to locally improve the approximation. When all the elements are roughly of the same shape (i.e. when the parameters h_i and ρ_i introduced in Section 5.2.3 are essentially the same over all elements), one says that the mesh is *quasi-uniform.*

On Figure 5.3, we show an example of non-structured mesh (which is quasi-uniform) and two examples of structured, uniform meshes for the domain $\Omega = (0, 1) \times (0, 1)$. The internal nodes are shown by black circles. Both structured meshes have been built from a one-dimensional mesh of the segment $(0, 1)$ with the nodes $\{0,\ 1/4,\ 1/2,\ 3/4,\ 1\}$. The domain Ω is thus paved with 16 squares. Each of them is next cut in two (resp. four) triangles to obtain the mesh shown on the center (resp. the right-hand side) of Figure 5.3.

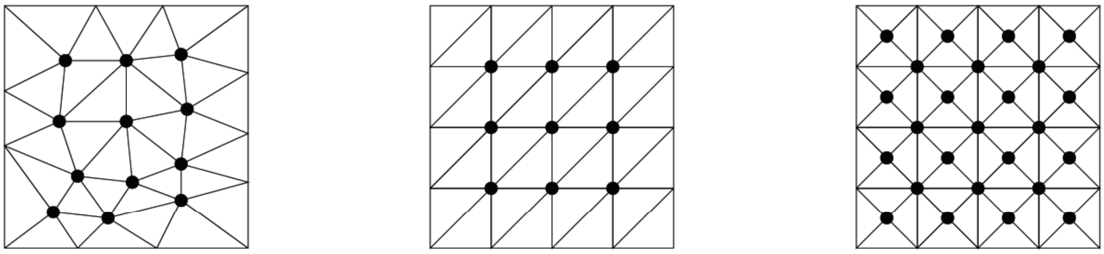

Figure 5.3: An example of a non-structured, quasi-uniform mesh (left) and two examples of structured, uniform meshes (center and right)

On Figure 5.4, we show a non-structured mesh of $\Omega = (0, 1) \times (0, 1)$, with a local refinement around the center of the square. This mesh has been built to approximate with a good accuracy the solution to a problem where the exact solution is singular at the center of the square.

In dimension $d \geq 2$, the stiffness matrix A defined by (5.13) is not tridiagonal. When using structured meshes, this matrix has nevertheless some structure. For instance, for the mesh shown at the center of Figure 5.3, the stiffness matrix is block-tridiagonal, and its coefficients only take a few different values (as in the one-dimensional case on a uniform mesh, where all diagonal coefficients – except possibly the first and the last – are equal, and where all off-diagonal coefficients are equal).

For general meshes, the stiffness matrix still has some particular property, related to the Laplacian operator: it is sparse. This property is interesting in terms of storage and resolution of the linear system (5.9). We first introduce the notion of sparse matrices.

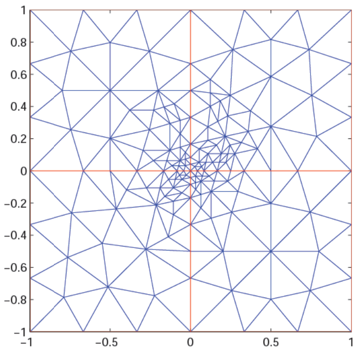

Figure 5.4: An example of non-structured mesh with local refinement

Definition 5.26. *Let A be a matrix of size $N \times N$. Let $N_\star$ be the number of coefficients in A that do not vanish. The matrix A is sparse if $N_\star \ll N^2$.*

We first observe that, for the matrix A defined by (5.13), if $A_{ij} \neq 0$, then there exists an element $K \in \mathcal{T}_h$ such that the two vertices s_i and s_j belong to K. Thus, if $i \neq j$, a necessary condition for $A_{ij} \neq 0$ is that the vertices s_i and s_j form an edge in the mesh. Thus, on a given line i of the matrix A, there are only a few coefficients that do not vanish, namely as many coefficients as edges starting from i. Furthermore, the diagonal coefficients of A do not vanish. We therefore have that

$$N_\star \leq N_s^{\rm int} + 2\, N_f,$$

where N_f is the number of edges in the mesh. The number of edges is roughly proportional to the number of vertices, hence

$$\frac{N_\star}{(N_s^{\rm int})^2} \leq \frac{c}{N_s^{\rm int}}$$

where c is a universal constant. The matrix A is hence sparse as soon as the mesh is sufficiently fine (i.e. $N_s^{\rm int}$ sufficiently large).

5.3.3 Solving the linear system

We now briefly discuss how to solve in practice the linear system (5.9), that reads $A\,U = B$, that we obtained in Section 5.1.3. We denote by N the size of this linear system.

When the stiffness matrix A is symmetric definite positive, a very efficient method to solve the linear system (5.9) is to use the Conjugate Gradient algorithm, which is an *iterative algorithm* based on the computation of quantities of the form $v^T A w$ for given vectors v and w. Its convergence is often fast. We recall that this algorithm is based on the fact that the solution to the linear system $A\,U = B$ is also the unique minimizer of the quadratic functional

$$J(v) = \frac{1}{2} v^T A v - v^T B, \qquad v \in \mathbb{R}^N.$$

In the cases when the matrix A is small (e.g. in one-dimensional cases, or when the two-dimensional mesh to solve the problem is coarse), it is possible to use a *direct method* to solve the linear system. The most general method is based on the so-called LU factorization of A, which consists in finding a lower triangular matrix $L^{\rm low}$ ($L^{\rm low}_{ij} = 0$ as soon as $j > i$) and a upper triangular matrix $U^{\rm up}$ such that $A = L^{\rm low}\, U^{\rm up}$. The resolution of the linear system then amounts to the two following steps:

- find $v \in \mathbb{R}^N$ such that $L^{\rm low}\, v = B$.
- find $U \in \mathbb{R}^N$ such that $U^{\rm up}\, U = v$.

These two steps are easy to implement, due to the fact that the matrices $L^{\rm low}$ and $U^{\rm up}$ are triangular. For instance, to find v, we first identify v_1 by $L^{\rm low}_{11} v_1 = B_1$, next v_2 by $L^{\rm low}_{21} v_1 + L^{\rm low}_{22} v_2 = B_2, \dots$

When the matrix A is symmetric, it is also possible to use the so-called Choleski decomposition of A in the form $A = L^{\rm low}\,(L^{\rm low})^T$ for a lower triangular matrix $L^{\rm low}$. Once this decomposition has been computed, solving the linear system $A\,U = B$ is again easy.

These various approaches can be compared in terms of cost, which is here measured in terms of the number of elementary operations (additions, multiplications, ...) that are needed. In the regime when $N \gg 1$ (which is the relevant regime for our applications), the cost to compute the matrices $L^{\rm low}$ and $U^{\rm up}$ such that $A = L^{\rm low}\, U^{\rm up}$ scales as $N^3/3$. If A is symmetric, the cost to compute its Choleski decomposition $A = L^{\rm low}\,(L^{\rm low})^T$ scales as $N^3/6$. The cost for solving the lower or upper triangular linear systems scales as N^2, and is therefore negligible. Thus, the cost to compute U solution to $A\,U = B$ by a direct method is of the order of N^3. Note also that, if several right-hand sides B have to be considered, the cost for the second, third, ..., resolutions is negligible, as the matrix A has already been factorized as $A = L^{\rm low}\, U^{\rm up}$ or $A = L^{\rm low}\,(L^{\rm low})^T$.

We now turn to the cost of the Conjugate Gradient algorithm. The cost of each iteration is of the order of N^2 (it may even be smaller if A is sparse). The number of iterations is related to the condition number ρ of the matrix A, which is equal to $\rho = \lambda_N/\lambda_1$, where $\{\lambda_j\}_{1 \leq j \leq N}$ are the eigenvalues of A (recall that we use the Conjugate Gradient algorithm in the case when A is symmetric definite positive) that we sorted in the increasing order: $0 < \lambda_1 \leq \cdots \leq \lambda_N$. The number of iterations increases when ρ increases. If the number of iterations is smaller than N, then the Conjugate Gradient algorithm is less expensive than a direct approach.

In practice, stiffness matrices stemming from a Finite Element discretization of the Laplacian operator have a poor condition number. If used directly, the Conjugate Gradient algorithm may need many iterations to converge, and be less efficient than a direct method.

To circumvent this difficulty, it is possible to use a *preconditionner*. Consider an invertible matrix D. Instead of considering the linear system $A\,U = B$, we consider the equivalent system $D\,A\,D^T\,(D^{-T}\,U) = D\,B$. We thus wish to find V such that $\overline{A}\,V = \overline{B}$, with $\overline{A} = D\,A\,D^T$ and $\overline{B} = D\,B$, and next compute U by $U = D^T\,V$. Note that the matrix $\overline{A}$ is again symmetric definite positive. The Conjugate Gradient algorithm may thus be used to solve the system $\overline{A}\,V = \overline{B}$. If D is sufficiently close to $A^{-1/2}$ is the sense that the condition number of $\overline{A}$ is small, then the procedure is efficient. We are thus left with finding a good matrix D, which is called here a precondionner. This question is problem-dependent. One can consider it from a linear algebra viewpoint (forgetting that the matrix A comes from a variationnal formulation) or from a PDE viewpoint (arguing on the variational formulation rather than on the linear system). A basic choice is to pick a diagonal matrix D with $D_{ii} = 1/\sqrt{A_{ii}}$. Anyhow, this question has been thoroughly studied and there exists nowadays very efficient preconditionners, for many types of matrices A.

5.3.4 Approximations by quadrature formulas

In order to compute the right-hand side B of the linear system to solve, which is given by (5.14), it is often needed to use a quadrature formula, as the integral cannot be exactly computed (the same difficulty also arises for the computation of the stiffness matrix, when the operator is more complex than the Laplacian operator, or when the basis functions are not as simple as $\mathbb{P}_1$ basis functions).

Such quadrature formulas are defined as follows. We first give ourselves an integer parameter ℓ_g, which is the number of points that are going to be used on an element K. We next give ourselves

ℓ_g real numbers $\{\omega_1, \dots, \omega_g\}$ which are the *weights* and ℓ_g points $\{\xi_1, \dots, \xi_g\}$ in K which are called the *nodes* of the quadrature formula. The integral of any function χ on K is then approximated as

$$\int_K \chi \approx \sum_{\ell=1}^{\ell_g} \omega_\ell \, \chi(\xi_\ell).$$

To quantify the accuracy of such quadrature formulas, we introduce the vector space of polynomial functions in x and y of total degree lower or equal than k:

$$\mathbb{P}_k = \left\{ p : \mathbb{R}^2 \to \mathbb{R}, \quad p(x,y) = \sum_{0 \le m,n \le k, \ m+n \le k} \alpha_{mn} \, x^m \, y^n \quad \text{for some real numbers } \alpha_{mn} \right\}.$$

The vector space $\mathbb{P}_k$ is of dimension $(k+1)(k+2)/2$. The degree of the quadrature formula, that we denote k_g, is defined as the largest integer k such that the quadrature formula is exact on $\mathbb{P}_k$, that is

$$\forall p \in \mathbb{P}_k, \qquad \int_K p = \sum_{\ell=1}^{\ell_g} \omega_\ell \, p(\xi_\ell).$$

The degree k_g is directly related to the accuracy of the quadrature formula as explained by the following result, the proof of which is based on a Taylor expansion.

Proposition 5.27. *On a triangle $K \subset \mathbb{R}^2$ of diameter h, consider a quadrature formula (of nodes ξ_ℓ and of weights ω_ℓ, with $1 \le \ell \le \ell_g$) of degree k_g. Then, there exists a constant c_q, independent of the element K, such that*

$$\forall \chi \in C^{1+k_g}(K), \qquad \left| \int_K \chi - \sum_{\ell=1}^{\ell_g} \omega_\ell \, \chi(\xi_\ell) \right| \le c_q h^{3+k_g} \|\chi\|_{C^{1+k_g}(K)}.$$

As a consequence, for $\chi \in C^{1+k_g}(\Omega)$, we have

$$\left| \int_\Omega \chi - I(\chi) \right| \le c_q h^{1+k_g} \|\chi\|_{C^{1+k_g}(\Omega)},$$

where $I(\chi)$ is the quadrature formula.

We eventually note that the cost of a quadrature formula is directly related to ℓ_g, as the cost is dominated by the cost of the evaluations of the function χ.

In dimension $d = 2$, it is convenient to locate the nodes of the quadrature by their barycentric coordinates. We collect in Table 5.1 some examples of quadrature formulas on a triangle K. The multiplicity is the number of circular permutations that one should do on the node that is given in the table to obtain all the nodes of the quadrature formula. For instance, the quadrature formula of degree 2 (shown in the third line) is based on 3 nodes, with the same weight $|K|/3$, the barycentric coordinates of which are $\left(\frac{1}{2}, \frac{1}{2}, 0\right)$, $\left(0, \frac{1}{2}, \frac{1}{2}\right)$ and $\left(\frac{1}{2}, 0, \frac{1}{2}\right)$.

Using the first quadrature formula with 3 nodes (the one shown in the second line of the table), we obtain the following approximation:

$$B_i = \int_\Omega f \, \varphi_i = \sum_{K \in \mathcal{K}(s_i)} \int_K f \, \varphi_i \approx \sum_{K \in \mathcal{K}(s_i)} \sum_{\ell=1}^{3} \omega_\ell \, f(\xi_\ell^K) \, \varphi_i(\xi_\ell^K),$$

ℓ_g	weights ω_ℓ	nodes ξ_ℓ	multiplicity	degree k_g of the quadrature formula
1	$\|K\|$	$\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)$	1	1
3	$\|K\|/3$	$(1,0,0)$	3	1
3	$\|K\|/3$	$\left(\frac{1}{2},\frac{1}{2},0\right)$	3	2
4	$-9\|K\|/16$	$\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)$	1	4
	$25\|K\|/48$	$\left(\frac{1}{5},\frac{1}{5},\frac{3}{5}\right)$	3	

Table 5.1: Quadrature formulas on a triangle K of surface $|K|$

where $\{\xi_\ell^K\}_{\ell=1}^3$ are the 3 quadrature nodes inside K (we recall that $\mathcal{K}(s_i)$ is the set of elements having s_i as a vertex). For this quadrature formula, the quadrature nodes are the element vertices, thus $\varphi_i(\xi_\ell^K)$ vanishes on two nodes and is equal to 1 on the third node:

$$B_i \approx \sum_{K\in\mathcal{K}(s_i)} \frac{1}{3}|K|\, f(s_i) = \frac{1}{3}|\mathcal{K}(s_i)|\, f(s_i)$$

where $|\mathcal{K}(s_i)|$ is the surface of the polygon $\mathcal{K}(s_i)$.

It is possible to show that the rate of convergence of the $\mathbb{P}_1$ Finite Element method is preserved as soon as, to approximate the right-hand side B, a quadrature formula of degree $k_g \geq 0$ is used. More precisely, consider the approximation B^q of B obtained by a quadrature formula of degree $k_g \geq 0$. Let U^q be the solution to the linear system $A\,U^q = B^q$ and let $u_h^q = \sum_j U_j^q\,\varphi_j$ be the associated discrete solution (note that the stiffness matrix A is assumed to be exactly computed). Then, there exists a constant c, which is independent of h, such that $\|u - u_h^q\|_{H^1(\Omega)} \leq c\,h$. Comparing this estimate with the one obtained in Theorem 5.22, we observe that the error converges to 0 at the same rate as if the right-hand side B was exactly computed.

5.4 The $\mathbb{P}_2$ Finite Element Method

We now briefly present a variant of the $\mathbb{P}_1$ Finite Element Method presented above, namely the $\mathbb{P}_2$ Finite Element Method. This variant can easily be generalized to the $\mathbb{P}_k$ Finite Element Method for any integer k.

Remark 5.28. *There exists other Finite Element Methods, besides the ones based on the vector space $\mathbb{P}_k$. We refer to the Exercises 5.36 and 5.39 for some examples.*

We again assume (for simplicity) that the dimension d is equal to 2, and that the domain Ω is a polygon. We again mesh Ω by triangles as in Section 5.2.1:

$$\overline{\Omega} = \cup_{i=1}^{N_e} K_i.$$

5.4.1 Approximation space

We introduce the vector space of functions that are piecewise equal to polynomial functions of total degree lower or equal to 2:

$$\mathbb{P}_2 = \left\{p : \mathbb{R}^2 \to \mathbb{R},\ \ p(x,y) = \alpha_1 + \alpha_2 x + \alpha_3 y + \alpha_4 x^2 + \alpha_5 y^2 + \alpha_6 xy \ \text{ for some real numbers } \{\alpha_j\}_{j=1}^6\right\}.$$

We next set

$$V_h^{(2)} = \left\{ v_h \in C^0(\overline{\Omega}), \quad v_h|_K \in \mathbb{P}_2 \text{ for any } K \in \mathcal{T}_h, \quad v_h = 0 \text{ on } \partial\Omega \right\},$$

and we see that this leads to a conformal approximation:

$$V_h^{(2)} \subset H_0^1(\Omega).$$

A function $v_h \in V_h^{(2)}$ is fully determined on an element K by its value at the three vertices and its value at the middle of the three edges. The following set of functions thus forms a basis of $V_h^{(2)}$:

- functions in $V_h^{(2)}$ that are equal to 1 on one vertex of the mesh, and that vanish on all other vertices and at all middle points of all edges;
- functions in $V_h^{(2)}$ that vanish on all vertices of the mesh, that are equal to 1 at the middle point of one edge, and that vanish at all other middle points.

The best approximation error in $V_h^{(2)}$ is again quantified by introducing a suitable interpolation operator $I_h^{(2)} : C^0(\overline{\Omega}) \to V_h^{(2)}$. We admit the following result:

Theorem 5.29. *Consider a regular sequence of meshes $\left\{ \mathcal{T}_{h_j} \right\}_{j \geq 1}$, in the sense of Definition 5.20. There exists a constant c_{inter}, independent of the mesh, such that, for any $v \in H^3(\Omega) \cap H_0^1(\Omega)$, we have*

$$\left\| v - I_h^{(2)} v \right\|_{H^1(\Omega)} \leq c_{\text{inter}}\, h^2\, |v|_{H^3(\Omega)} \quad \textit{and} \quad \left\| v - I_h^{(2)} v \right\|_{L^2(\Omega)} \leq c_{\text{inter}}\, h^3\, |v|_{H^3(\Omega)}.$$

Note that the $\mathbb{P}_2$ interpolation error is bounded for functions that are more regular than in Theorem 5.15 where we consider the $\mathbb{P}_1$ interpolation error: we now ask that $v \in H^3(\Omega)$ instead of $v \in H^2(\Omega)$. For these more regular functions, a smaller interpolation error is obtained, as it is of order h^2 in the H^1 norm (in contrast to the $\mathbb{P}_1$ case, where it is of order h in the H^1 norm).

We thus note that using a $\mathbb{P}_2$ approach instead of a $\mathbb{P}_1$ approach may be useful only if the solution is sufficiently regular. Note also that, for the same mesh size h, the number of degrees of freedom in a $\mathbb{P}_2$ Finite Element space is larger than in a $\mathbb{P}_1$ Finite Element space. It is thus important to compare accuracies not in terms of h, but in terms of the number of degrees of freedom, which is directly related to the cost of the approach.

5.4.2 Approximation of the Poisson problem

Let $f \in L^2(\Omega)$. The Galerkin approximation of the Poisson problem (5.2) on the finite dimensional space $V_h^{(2)}$ is:

$$\left\{ \begin{array}{l} \text{Find } u_h \in V_h^{(2)} \text{ such that} \\ \forall v_h \in V_h^{(2)}, \qquad a(u_h, v_h) = b(v_h). \end{array} \right. \tag{5.16}$$

As for the $\mathbb{P}_1$ approach, this lead to a linear system of the form $AU = B$ (see (5.9)), where the stiffness matrix A is of size $\dim V_h^{(2)} \times \dim V_h^{(2)}$.

We now state an error estimate:

Theorem 5.30. *Let $u \in H^1(\Omega)$ be the solution to* (5.2) *and $u_h \in V_h^{(2)}$ be the solution to* (5.16)*. We assume that the sequence of meshes is regular of parameter σ_0 and that the solution u belongs to $H^3(\Omega)$. Then there exists a constant c, independent of h, such that*

$$\|u - u_h\|_{H^1(\Omega)} \leq c\, h^2.$$

Note that, if u is only in $H^2(\Omega)$, then the above error bound does not hold. Since $V_h^{(1)} \subset V_h^{(2)}$, we get in this case that $\|u - u_h\|_{H^1(\Omega)} \leq c\, h$. For a given mesh size h, we thus obtain the same accuracy as with the $\mathbb{P}_1$ method, for a larger cost. The $\mathbb{P}_2$ method is thus not interesting in this case.

If the solution u is in $H^3(\Omega)$, then the $\mathbb{P}_2$ method is actually a better approach than the $\mathbb{P}_1$ method. For a $\mathbb{P}_1$ method, we indeed have $\|u - u_h^{(1)}\|_{H^1(\Omega)} \leq c_1\, h$ and the number of degrees of freedom (say on a structured mesh of $\Omega = (0,1)^2$ made of squares divided in two triangles) is $N = 1/h^2$. We therefore get

$$\|u - u_h^{(1)}\|_{H^1(\Omega)} \leq \frac{c_1}{\sqrt{N}}.$$

For a $\mathbb{P}_2$ method, we have $\|u - u_h^{(2)}\|_{H^1(\Omega)} \leq c_2\, h^2$ (note that the constant c_2 is in general different from c_1) and the number of degrees of freedom (on the same structured mesh of $\Omega = (0,1)^2$) is $N = 1/(h/2)^2$. We therefore get

$$\|u - u_h^{(2)}\|_{H^1(\Omega)} \leq \frac{4c_2}{N}.$$

Therefore, if N is large enough (i.e. the mesh sufficiently fine), the $\mathbb{P}_2$ method provides a solution with a better accuracy than the $\mathbb{P}_1$ method, for an equal number of degrees of freedom.

5.5 Exercises

Exercise 5.31 (Variational crimes)**.** *Let V be a Hilbert space, b be a continuous linear form on V, and a be a continuous bilinear form on $V \times V$ that is coercive. The problem*

$$\begin{cases} \text{Find } u \in V \text{ such that} \\ \forall v \in V, \qquad a(u,v) = b(v) \end{cases}$$

has a unique solution. In the light of Remark 5.3, we approximate this problem by the following variational formulation:

$$\begin{cases} \text{Find } u_h \in V_h \text{ such that} \\ \forall v_h \in V_h, \qquad a_h(u_h,v_h) = b_h(v_h), \end{cases}$$

where a_h (resp. b_h) is an approximation of a (resp. b), and where $V_h \subset V$. We assume that b_h is a continuous linear form on V_h, that a_h is a continuous bilinear form on $V_h \times V_h$, that is coercive uniformly in h: there exists $\alpha > 0$ independent of h such that, for any h,

$$\forall v \in V_h, \qquad a_h(v,v) \geq \alpha \|v\|^2.$$

The above discrete problem is thus well-posed and we wish to estimate $\|u - u_h\|$.

1. *Let $v_h \in V_h$. Show that*

$$a_h(u_h - v_h, u_h - v_h) = b_h(u_h - v_h) - a_h(v_h, u_h - v_h),$$

then

$$a_h(u_h - v_h, u_h - v_h) = b_h(u_h - v_h) - b(u_h - v_h) + a(u, u_h - v_h) - a_h(v_h, u_h - v_h),$$

and then

$$a_h(u_h - v_h, u_h - v_h) \leq a(u - v_h, u_h - v_h) + E_b\, \|u_h - v_h\| + E_a\, \|v_h\|\, \|u_h - v_h\|$$

where

$$E_b = \sup_{w_h \in V_h,\, w_h \neq 0} \frac{|b_h(w_h) - b(w_h)|}{\|w_h\|}$$

and

$$E_a = \sup_{w_h\in V_h,\, w_h\neq 0} \ \sup_{s_h\in V_h,\, s_h\neq 0} \frac{|a_h(w_h,s_h)-a(w_h,s_h)|}{\|w_h\|\,\|s_h\|}.$$

2. *Denoting M the continuity constant of the bilinear form a on $V\times V$, deduce that, for any $v_h\in V_h$, we have*
$$\alpha\|u_h-v_h\|\leq M\|u-v_h\|+E_b+E_a\,\|v_h\|.$$

3. *Show that*
$$\|u-u_h\|\leq \frac{E_b}{\alpha}+\inf_{v_h\in V_h}\left[\left(1+\frac{M}{\alpha}\right)\|u-v_h\|+\frac{E_a}{\alpha}\|v_h\|\right].$$

Exercise 5.32 (Discrete maximum principle)**.** *Let A be the stiffness matrix of the discrete problem* (5.12)*, in the one-dimensional case, for $\Omega=(a,b)$, when using $\mathbb{P}_1$ finite elements on a uniform mesh of size $h=(b-a)/(N+1)$. We recall that A is of size $N\times N$ and is given (using* (5.13)*) by*

$$A=\frac{1}{h}\,\textit{tridiag}\,(-1,2,-1).$$

Let $B\in\mathbb{R}^N$ be the right-hand side of the linear system equivalent to (5.12)*. The components of B are given by* (5.14)*.*

1. *For a vector $V\in\mathbb{R}^N$, we say that $V\leq 0$ if $V_i\leq 0$ for any $1\leq i\leq N$. Let $V\in\mathbb{R}^N$ such that $AV\leq 0$. Show that $V\leq 0$.*

2. *Deduce that, if the function f in* (5.12)–(5.13)–(5.14) *is such that $f\leq 0$ on Ω, then the discrete solution u_h to* (5.12) *satisfies $u_h\leq 0$ on Ω. This property is called the* discrete maximum principle.

3. *We denote α_{ij}, with $1\leq i,j\leq N$ the coefficients of the matrix A^{-1}. Show that $\alpha_{ij}>0$ for any $1\leq i,j\leq N$.*

4. *Show that, for any $1\leq i\leq N$,*
$$\sum_{j=1}^{N}\alpha_{ij}\leq\frac{1}{8h}.$$
Hint: consider the function $w(x)=\dfrac{1}{2}\,x\,(1-x)$.

5. *Deduce that $\|u_h\|_{L^\infty(\Omega)}\leq\dfrac{1}{8}\|f\|_{L^\infty(\Omega)}$.*

Exercise 5.33 (Interpolation error)**.** *The aim of this exercise is to prove the interpolation error estimate* (5.10)*, in the* one-dimensional case. *Let $\Omega=(a,b)$, that we mesh by the nodes $a=x_0<x_1<\cdots<x_i<\cdots<x_{n+1}=b$. For any $0\leq i\leq n$, we set $K_i=[x_i,x_{i+1}]$ and $h_i=x_{i+1}-x_i$. Let $v\in H^2(\Omega)$.*

1. *We consider the element K_i and we define $w_i : K_i \to \mathbb{R}$ by*

$$\forall s \in K_i, \qquad w_i(s) = v'(s) - \frac{v(x_{i+1}) - v(x_i)}{x_{i+1} - x_i}.$$

Check that this function is such that $v(x) - (I_h^{(1)} v)(x) = \displaystyle\int_{x_i}^{x} w_i(s)\, ds$ *for any* $x \in K_i$*. Show that*

$$\|v - I_h^{(1)} v\|_{L^2(K_i)} \leq h_i \|w_i\|_{L^2(K_i)}.$$

2. *Show that there exists some $y_i \in K_i$ such that $w_i(y_i) = 0$. Deduce that*

$$\|w_i\|_{L^2(K_i)} \leq h_i \|w_i'\|_{L^2(K_i)}.$$

3. *Show that there exists a constant $c_{\rm inter}$, independent of the mesh, such that, for any $v \in H^2(\Omega) \cap H^1_0(\Omega)$, we have*

$$\left\| v - I_h^{(1)} v \right\|_{H^1(\Omega)} \leq c_{\rm inter}\, h \, |v|_{H^2(\Omega)} \quad \text{and} \quad \left\| v - I_h^{(1)} v \right\|_{L^2(\Omega)} \leq c_{\rm inter}\, h^2 \, |v|_{H^2(\Omega)}.$$

Exercise 5.34 (Convection-diffusion problems)**.** *We consider the convection-diffusion problem* (3.10)*, that reads*

$$\begin{cases} -\Delta u + c \cdot \nabla u = f & \text{in } \mathcal{D}'(\Omega), \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

where Ω is a bounded subset of $\mathbb{R}^d$ and $f \in L^2(\Omega)$. We assume that $c : \Omega \to \mathbb{R}^d$ is a vector field of class $C^1(\overline{\Omega})$ which is divergence-free:

$$\operatorname{div} c = \sum_{i=1}^{d} \frac{\partial c_i}{\partial x_i} = 0 \quad \text{in } \Omega.$$

Under these assumptions, we have shown that there exists a unique $u \in H^1(\Omega)$ solution to that problem. To do so, we have considered the variational formulation

$$\begin{cases} \text{Find } u \in H^1_0(\Omega) \text{ such that} \\ \forall v \in H^1_0(\Omega), \qquad a(u,v) = b(v). \end{cases}$$

where

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v + \int_\Omega (c \cdot \nabla u)\, v, \qquad b(v) = \int_\Omega f\, v.$$

The aim of this exercise is to study the discretization of that problem, in particular in the case when the convection field c is large.

For the sake of simplicity, we consider here the one-dimensional case: we set $\Omega = (0,1)$ and we look for $u \in H^1(\Omega)$ such that

$$\begin{cases} -\nu u'' + \beta u' = f & \text{in } \mathcal{D}'(\Omega), \\ u(0) = 0, \quad u(1) = 0, \end{cases} \tag{5.17}$$

where ν and β are two positive real numbers (note that the convection field is here constant, and thus indeed divergence-free). We take $f \in L^2(\Omega)$.

1. *Recall the variational formulation of* (5.17) *and recall why the problem is well-posed. Show also that* $u \in H^2(\Omega)$.

2. *We introduce a discretization of* (5.17) *based on* $\mathbb{P}_1$ *finite elements and denote* u_h *the discrete solution. Give the expression of the stiffness matrix* A *in terms of the ratio* ν/h *and of the so-called Péclet number*
$$\gamma = \frac{h\,\beta}{\nu}.$$

3. *Compute the continuity constant* M *and the coercivity constant* α *of the bilinear form* a*, in terms of* β*,* ν *and the Poincaré constant* C_Ω *of* Ω*. Using the Céa estimate* (5.8)*, write an error bound on* $\|u - u_h\|_{H^1(\Omega)}$.

4. *This estimate can actually be improved by going over the proof of* (5.8) *and carefully bounding the bilinear form. Show that*
$$\nu |u - u_h|^2_{H^1(\Omega)} = a\left(u - u_h, u - I_h^{(1)} u\right)$$
and deduce that there exists a constant c*, independent of* u*,* ν*,* β *and* h*, such that*
$$|u - u_h|_{H^1(\Omega)} \le c \left(1 + \frac{h\,\beta}{\nu}\right) h\, |u|_{H^2(\Omega)}. \tag{5.18}$$
Which estimate on $\|u - u_h\|_{H^1(\Omega)}$ *do we obtain? Show that this estimate is better than the one obtained in the previous question.*

5. *What happens in* (5.18) *when* β *increases or when* ν *decreases?*

   Check that, when $\gamma > 2$*, the coefficient* $A_{i,i+1}$ *becomes positive. It turns out that the discrete solution is polluted by spurious oscillations (see Figure 5.5 below).*

   A first possibility to prevent this phenomenon is to use a sufficiently small mesh size h*, so that* $\gamma < 2$*. However, if* β *increases (or if* ν *decreases), this may lead to very small values for* h *and thus a too large computational cost.*

   Another solution consists in working with a modified bilinear form, that is given by
$$a_\star(u, v) = a(u, v) + a_h(u, v), \qquad a_h(u, v) = \frac{h\,\beta}{2} \int_\Omega u'\, v'.$$
   Along with that modification, we also modify the linear form b *and work with*
$$b_\star(v) = b(v) + b_h(v), \qquad b_h(v) = \frac{h}{2} \int_\Omega f\, v'.$$
   We consider the discrete problem
$$\left\{ \begin{array}{ll} \text{Find } u_h^\star \in V_h^{(1)} \text{ such that} & \\ \forall v \in V_h^{(1)}, & a_\star(u_h^\star, v) = b_\star(v). \end{array} \right. \tag{5.19}$$
   Give the expression of the new stiffness matrix and check that its off-diagonal coefficients always remain negative. In contrast to the previous approximation u_h*, it turns out that* $u_h^\star$ *has no spurious oscillations (see Figure 5.5 below).*

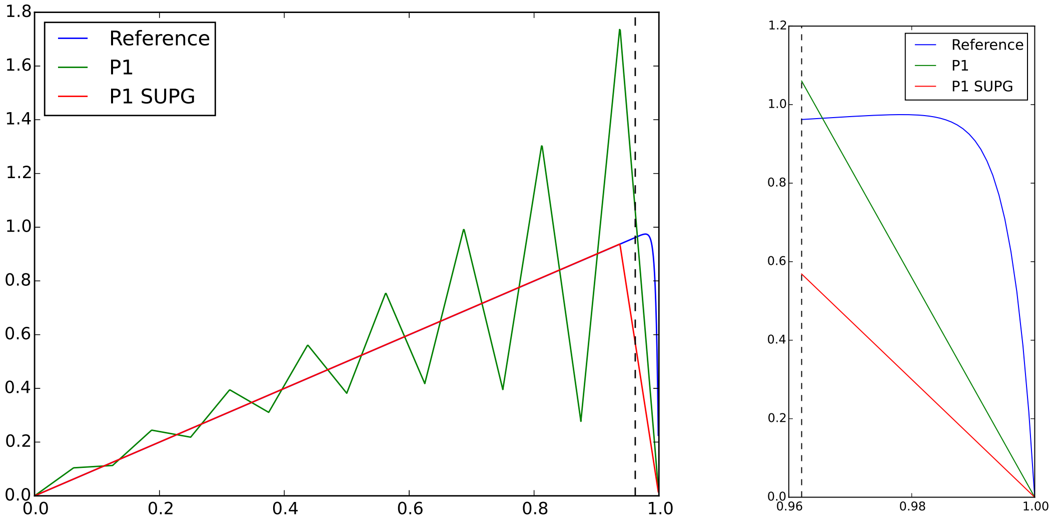

Figure 5.5: Exact and numerical solutions to (5.17) for $\nu = 1/256$, $\beta = 1$, $f = 1$ and $h = 1/16$. Left: Plot on the whole domain. Right: Close-up on the boundary layer at the right-end of the domain. The 'Reference' solution (in blue) is the exact solution to (5.17). The 'P1' solution (in green) is the $\mathbb{P}_1$ approximation using the original bilinear form a (see Question 2). The 'P1 SUPG' solution (in red) is the $\mathbb{P}_1$ approximation using the modified variational formulation (5.19). In this case, $\gamma = 16$, and we indeed observe spurious oscillations on the 'P1' approximation. The 'P1 SUPG' approximation has no such oscillations.

Note: the method that we have introduced here is not restricted to the one-dimensional case. It is called the SUPG method, which stands for Streamline-Upwind/Petrov-Galerkin. We refer e.g. to [7] for more details.

Exercise 5.35 (Approximation errors at the mesh nodes)**.** *We consider the one-dimensional problem*

$$\begin{cases} \text{Find } u \in H_0^1(\Omega) \text{ such that} \\ \forall v \in H_0^1(\Omega), \qquad \displaystyle\int_\Omega u'\,v' = \int_\Omega f\,v, \end{cases}$$

where $\Omega = (a,b)$ and $f \in L^2(\Omega)$. We approximate this problem by a $\mathbb{P}_1$ finite element method on a mesh based on the nodes x_i, $1 \leq i \leq n$. We denote u_h the discrete solution. Our aim is to show that $u_h(x_i) = u(x_i)$ for any i: the discrete solution is equal *to the exact solution at the mesh nodes. We already underline that this miracle only happens in the one-dimensional case.*

For any $1 \leq i \leq n$, we introduce the function G_i defined by

$$G_i(x) = \begin{cases} \dfrac{b-x_i}{b-a}\,(x-a) & \text{if } a \leq x \leq x_i, \\ \dfrac{x_i-a}{b-a}\,(b-x) & \text{if } x_i \leq x \leq b. \end{cases}$$

1. *For any $v \in H_0^1(\Omega)$ and any $1 \leq i \leq n$, show that*

$$\int_\Omega G_i'\,v' = v(x_i).$$

2. *Deduce that $u_h(x_i) = u(x_i)$ for any $1 \leq i \leq n$.*

Exercise 5.36 (Mixed finite elements)**.** *Let Ω be a bounded domain in $\mathbb{R}^d$ and $f \in L^2(\Omega)$. We consider the problem: find $(u,p) \in H_0^1(\Omega) \times (L^2(\Omega))^d$ solution to*

$$\begin{cases} -div p = f & \text{in } \mathcal{D}'(\Omega), \\ \nabla u = p & \text{in } \mathcal{D}'(\Omega). \end{cases} \tag{5.20}$$

Of course, if (u,p) is a solution to (5.20)*, then $u \in H_0^1(\Omega)$ is solution to*

$$-\Delta u = f \qquad \text{in } \mathcal{D}'(\Omega), \tag{5.21}$$

and we have seen various ways to solve that problem. However, if the quantity of interest is p rather than u, it may be interesting to keep p as part of the unknowns and to address (5.20) *rather than* (5.21)*. Our aim here is to directly discretize* (5.20)*.*

1. *Show that* (5.20) *is equivalent to the following variational formulation:*

$$\begin{cases} \text{Find } (u,p) \in H_0^1(\Omega) \times (L^2(\Omega))^d \text{ such that} \\ \forall q \in (L^2(\Omega))^d, \qquad a(p,q) + b(q,u) = 0, \\ \forall v \in H_0^1(\Omega), \qquad b(p,v) = c(v), \end{cases} \tag{5.22}$$

   where a and b are bilinear forms and c is a linear form that will be made precise.

2. *We consider a Galerkin approximation of* (5.22)*, using an admissible mesh $\mathcal{T}_h$ of Ω. To approximate p, we use $\mathbb{P}_0$ finite elements, and thus work with the approximation space $V_h^{(0)}$ (piecewise constant functions). To approximate u, we use $\mathbb{P}_1$ finite elements, and thus work with the approximation space $V_h^{(1)}$ (piecewise affine functions). Write the discrete variational formulation and show that it is equivalent to solving a linear system of the form*

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} P \\ U \end{pmatrix} = \begin{pmatrix} 0 \\ F \end{pmatrix}.$$

Give the expressions of the matrices A and B and of the vector F using the basis functions of $V_h^{(0)}$ and $V_h^{(1)}$.

3. *In the one-dimensional case, compute explicitly the matrices A and B.*

4. *Check that the matrix B is injective. Deduce that the matrix $B^T\,A^{-1}\,B$ is positive definite and next that the matrix*

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix}$$

is invertible.

Exercise 5.37 (Error estimation in the L^2 norm)**.** *Our aim here is to establish an optimal error estimate in the L^2 norm for the Poisson problem discretized using the $\mathbb{P}_1$ Finite Element Method (see Remark 5.25).*

We set $f \in L^2(\Omega)$ and assume that the polygon Ω is convex. We recall that, under this assumption, the unique solution u to (5.2) *belongs to $H^2(\Omega)$ and satisfies $|u|_{H^2(\Omega)} \leq \mathcal{C}_\Omega \|f\|_{L^2(\Omega)}$, where the constant $\mathcal{C}_\Omega$ only depends on Ω.*

Let $u_h \in V_h^{(1)}$ be the solution to (5.12)*. We assume that the sequence of meshes is regular of parameter σ_0. We recall (see Theorem 5.22) that there exists a constant c, that may depend on Ω and σ_0, but that is independent of h and f, such that*

$$\|u - u_h\|_{H^1(\Omega)} \leq c\,h\,\|f\|_{L^2(\Omega)}.$$

1. *Let ξ be the unique solution to the problem*

$$\begin{cases} \text{Find } \xi \in H_0^1(\Omega) \text{ such that} \\ \forall v \in H_0^1(\Omega), \qquad \displaystyle\int_\Omega \nabla\xi \cdot \nabla v = \int_\Omega (u - u_h)\,v. \end{cases}$$

Show that $\xi \in H^2(\Omega)$ and that

$$\|u - u_h\|_{L^2(\Omega)}^2 = \int_\Omega \nabla(u - u_h) \cdot \nabla\left(\xi - I_h^{(1)}\xi\right),$$

where $I_h^{(1)}\xi$ is the interpolation of ξ on $V_h^{(1)}$.

2. *Using the fact that $|\xi|_{H^2(\Omega)} \leq \mathcal{C}_\Omega \|u - u_h\|_{L^2(\Omega)}$, show that*

$$\|u - u_h\|_{L^2(\Omega)} \leq \widehat{c}\,h^2,$$

where $\widehat{c}$ may depend on Ω, σ_0, f but is independent of h. This result is the so-called Aubin-Nitsche lemma. This estimate is sharp in the sense that there exist Poisson problems such that $\|u - u_h\|_{L^2(\Omega)}$ is indeed of the order of h^2.

Exercise 5.38 (L^2 projection). *We consider a polygon $\Omega \subset \mathbb{R}^2$ and an admissible mesh $\mathcal{T}_h$ of Ω with N internal vertices. Let $V_h^{(1)}$ be the $\mathbb{P}_1$ finite element space associated with this mesh. We denote $(\varphi_1, \ldots, \varphi_N)$ the basis of $V_h^{(1)}$.*

We consider here the L^2 projection on $V_h^{(1)}$, namely the operator $\Pi_h : L^2(\Omega) \to V_h^{(1)}$ such that, for any $v \in L^2(\Omega)$, we have

$$\|v - \Pi_h v\|_{L^2(\Omega)} = \inf_{v_h \in V_h^{(1)}} \|v - v_h\|_{L^2(\Omega)}.$$

The existence and uniqueness of $\Pi_h v$ comes from the orthogonal projection theorem in the Hilbert space $L^2(\Omega)$. As usual, $\Pi_h v$ is the function in $V_h^{(1)}$ which is the closest (in the L^2 norm) to v. It also satisfies

$$\forall 1 \leq i \leq N, \qquad \int_\Omega (v - \Pi_h v)\, \varphi_i = 0.$$

1. *Let $M \in \mathbb{R}^{N \times N}$ be the matrix defined by*

   $$\forall 1 \leq i, j \leq N, \qquad M_{ij} = \int_\Omega \varphi_i\, \varphi_j.$$

   Show that M is symmetric definite positive.

2. *Show that $\Pi_h v$ may be computed by solving a linear system of the form $M\, X = V$ for a right-hand side $V \in \mathbb{R}^N$ that will be made precise. The relation between $\Pi_h v$ and $X \in \mathbb{R}^N$ will also be explicited.*

Exercise 5.39 ($\mathbb{Q}_1$ finite element in dimension 2). *We present here the $\mathbb{Q}_1$ finite element method. We assume that the dimension d is equal to 2, although our construction carries over to any dimension. The main idea is to mesh the domain with quadrangles, in contrast to the $\mathbb{P}_k$ Finite Element Methods where the domain is meshed with triangles.*

Let $\Omega = (\alpha_1, \beta_1) \times (\alpha_2, \beta_2)$. We consider two one-dimensional meshes $\{x_i\}_{0 \leq i \leq N+1}$ and $\{y_j\}_{0 \leq j \leq M+1}$ such that

$$\alpha_1 = x_0 < x_1 < \cdots < x_i < \cdots < x_{N+1} = \beta_1 \quad \text{and} \quad \alpha_2 = y_0 < y_1 < \cdots < y_j < \cdots < y_{M+1} = \beta_2$$

and we mesh Ω by the rectangles K_{ij} defined by $K_{ij} = [x_i, x_{i+1}] \times [y_j, y_{j+1}]$. More generally, Ω could be meshed by quadrangles that are not necessarily rectangles. We denote by $\mathcal{T}_h$ this mesh. Note that the mesh is again admissible in the sense that the intersection of any two elements is either empty, or restricted to a point which is a vertex of both elements, or equal to a segment which is an edge of the two elements. The mesh size is defined as

$$h = \max\left(\max_{0 \leq i \leq N} x_{i+1} - x_i, \max_{0 \leq j \leq M} y_{j+1} - y_j \right).$$

We introduce the vector space of polynomial functions of partial degree lower or equal to 1:

$$\mathbb{Q}_1 = \left\{ p : \mathbb{R}^2 \to \mathbb{R}, \quad p(x,y) = \alpha + \beta x + \gamma y + \delta xy \ \text{ for some real numbers } \alpha,\ \beta,\ \gamma \text{ and } \delta \right\}.$$

Note that, in contrast to $\mathbb{P}_1$, the function $(x,y) \in \mathbb{R}^2 \mapsto xy$ belongs to $\mathbb{Q}_1$. A function in $\mathbb{Q}_1$ is fully determined by its value at the four vertices of the non-degenerate quadrangle.

1. *We define the space*

$$W_h^{(1)} = \left\{ v_h \in C^0(\overline{\Omega}), \quad v_h|_K \in \mathbb{Q}_1 \text{ for any } K \in \mathcal{T}_h, \quad v_h = 0 \text{ on } \partial\Omega \right\}.$$

For any (i,j) *with* $1 \leq i \leq N$ *and* $1 \leq j \leq M$*, we consider the function* $\varphi_{ij} \in W_h^{(1)}$ *such that, for any* $1 \leq \ell \leq N$, $1 \leq m \leq M$,

$$\varphi_{ij}(x_\ell, y_m) = \begin{cases} 1 & \text{if } (\ell, m) = (i,j), \\ 0 & \text{otherwise.} \end{cases}$$

Show that the set $\{\varphi_{ij}, \quad 1 \leq i \leq N, \quad 1 \leq j \leq M\}$ *is a basis of the vector space* $W_h^{(1)}$*. Is the space* $W_h^{(1)}$ *is vector subspace of* $H_0^1(\Omega)$*?*

2. *Let* $f \in L^2(\Omega)$*. The Galerkin approximation of the Poisson problem* (5.2) *on the finite dimensional space* $W_h^{(1)}$ *is:*

$$\begin{cases} \text{Find } u_h \in W_h^{(1)} \text{ such that} \\ \forall v_h \in W_h^{(1)}, \qquad a(u_h, v_h) = b(v_h). \end{cases} \tag{5.23}$$

Let $\{\theta_i\}_{1 \leq i \leq N}$ *(resp.* $\{\xi_j\}_{1 \leq j \leq M}$*) be the basis functions of the one-dimensional* $\mathbb{P}_1$ *discrete space based on the mesh* $\{x_i\}_{0 \leq i \leq N+1}$ *(resp.* $\{y_j\}_{0 \leq j \leq M+1}$*) of the segment* (α_1, β_1) *(resp.* (α_2, β_2)*). Let* $\mathcal{M}^\theta$ *and* $\mathcal{A}^\theta$ *be the mass and the stiffness matrices associated to the functions* $\{\theta_i\}_{1 \leq i \leq N}$*:*

$$\forall 1 \leq i_1, i_2 \leq N, \quad \left[\mathcal{M}^\theta\right]_{i_1\, i_2} = \int_{\alpha_1}^{\alpha_2} \theta_{i_1}\, \theta_{i_2} \quad \text{and} \quad \left[\mathcal{A}^\theta\right]_{i_1\, i_2} = \int_{\alpha_1}^{\alpha_2} \theta'_{i_1}\, \theta'_{i_2}.$$

We proceed likewise with the functions $\{\xi_j\}_{1 \leq j \leq M}$ *and introduce the mass matrix* $\mathcal{M}^\xi$ *and the stiffness matrix* $\mathcal{A}^\xi$*.*

Show that the stiffness matrix of (5.23) *can be written in terms of* $\mathcal{M}^\theta$*,* $\mathcal{A}^\theta$*,* $\mathcal{M}^\xi$ *and* $\mathcal{A}^\xi$*.*

3. *The approximation properties in* $W_h^{(1)}$ *are similar as in the* $\mathbb{P}_1$ *space* $V_h^{(1)}$*. We introduce the interpolation operator*

$$\begin{array}{rcl} I_h^{(1)} : C^0(\overline{\Omega}) & \to & W_h^{(1)} \\ v & \mapsto & v_h = \displaystyle\sum_{i=1}^{N} \sum_{j=1}^{M} v(x_i, y_j)\, \varphi_{ij}. \end{array}$$

We see that $I_h^{(1)} v$ *is the unique function in* $W_h^{(1)}$ *that takes the same values as* v *on all the internal nodes of the mesh. We admit the following result: there exists* c_{inter} *such that, for any* h,

$$\forall v \in H^2(\Omega) \cap H_0^1(\Omega), \quad \left\| v - I_h^{(1)} v \right\|_{H^1(\Omega)} \leq c_{\text{inter}}\, h\, |v|_{H^2(\Omega)}.$$

Prove that the solution u_h *to* (5.23) *converges to the solution* u *to* (5.2) *and state an error estimate.*

Note: the interest of using a $\mathbb{Q}_1$ *approach (rather than a* $\mathbb{P}_1$ *approach) is that it is sometimes easier to mesh* Ω *using quadrangles rather than triangles.*

Chapter 6

Numerical approximation of boundary value problems (the non-coercive case)

This chapter is devoted to the numerical approximation of linear elliptic boundary value problems, in a non-coercive setting. We consider here problems such as those studied in Chapter 4, the well-posedness of which has been established using the inf-sup theory.

6.1 Galerkin approximation

In this section, we consider the variational formulation (4.8), which reads

$$\left\{\begin{array}{l} \text{Find } u \in V \text{ such that} \\ \forall w \in W, \qquad a(u,w) = b(w), \end{array}\right. \tag{6.1}$$

where V and W are two Banach spaces, a is a bilinear continuous form on $V \times W$ and b is linear continuous form on W. We assume that a satisfies the inf-sup conditions (4.9) and (4.10). In addition, for the sake of simplicity, we assume that W is a Hilbert space. In view of Theorem 4.10, the problem (6.1) is well-posed.

6.1.1 Well-posedness

Our aim here is to approximate (6.1) in a finite dimensional setting. For the sake of simplicity, we restrict ourselves to a conformal approximation. We thus introduce the finite dimensional spaces $V_h \subset V$ and $W_h \subset W$, and consider the problem

$$\left\{\begin{array}{l} \text{Find } u_h \in V_h \text{ such that} \\ \forall w_h \in W_h, \qquad a(u_h,w_h) = b(w_h). \end{array}\right. \tag{6.2}$$

In view of Lemma 4.7, this problem is well-posed if and only if the two following conditions are satisfied:

$$\text{There exists } \alpha_h > 0 \text{ such that} \quad \inf_{u_h \in V_h,\, u_h \neq 0} \ \sup_{w_h \in W_h,\, w_h \neq 0} \frac{a(u_h,w_h)}{\|u_h\|_V \, \|w_h\|_W} \geq \alpha_h \tag{6.3}$$

and

$$\text{If } w_h \in W_h \text{ is such that } a(u_h,w_h) = 0 \text{ for any } u_h \in V_h, \text{ then } w_h = 0. \tag{6.4}$$

Note that α_h may depend on h (but should be positive).

We first observe that the inf-sup conditions (4.9)–(4.10) on $V \times W$ do not imply the inf-sup conditions (6.3)–(6.4) on $V_h \times W_h$. This is in sharp contrast with the coercivity assumption: if there exists $\alpha > 0$ such that $a(u,u) \geq \alpha \|u\|_V^2$ for any $u \in V$, then, for any $V_h \subset V$ and any $u_h \in V_h$, we have $a(u_h, u_h) \geq \alpha \|u_h\|_V^2$, thus the coercivity of a on V_h (with the same coercivity constant).

It is actually easy to build an example of a bilinear form a such that (4.9)–(4.10) hold and such that (6.3)–(6.4) do not hold. Consider for instance the case $V = W = \mathbb{R}^2$ and $a(u,v) = v^T A u$ with

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

The matrix A is symmetric, its two eigenvalues are 1 and -1, and it is invertible. The conditions (4.9)–(4.10) are thus satisfied. Consider now the subspace

$$V_h = W_h = \left\{ \begin{pmatrix} s \\ 0 \end{pmatrix}, \quad s \in \mathbb{R} \right\} \subset V.$$

For any $u_h \in V_h$ and $w_h \in W_h$, we have $a(u_h, w_h) = 0$. The condition (6.3) is hence not satisfied.

In general, there is no alternative to checking that (6.3)–(6.4) indeed hold on $V_h \times W_h$. We recall that (6.3)–(6.4) imply that $\dim V_h = \dim W_h$ (see Remark 4.9). We also recall that, if $\dim V_h = \dim W_h$, then the conditions (6.3) and (6.4) are equivalent (see Remark 4.8). In that case, it is thus sufficient to check that either (6.3), or (6.4), holds for (6.2) to be well-posed.

6.1.2 Error estimation

We wish to bound from above the error $e = u - u_h$. The equivalent of the Céa lemma in the inf-sup setting is the following result:

Lemma 6.1. *We suppose that assumptions* (4.9)–(4.10) *hold. Consider the subspaces* $V_h \subset V$ *and* $W_h \subset W$*, and assume that* (6.3)–(6.4) *hold as well.*

Let u *be the solution to* (6.1) *and* u_h *be the solution to* (6.2)*. Then*

$$\|u - u_h\|_V \leq \left(1 + \frac{C_a}{\alpha_h}\right) \inf_{v_h \in V_h} \|u - v_h\|_V, \tag{6.5}$$

where α_h *is the constant appearing in* (6.3) *and* C_a *is the continuity constant of* a *on* $V \times W$*: for any* $v \in V$ *and* $w \in W$*, we have* $|a(v,w)| \leq C_a \|v\|_V \|w\|_W$*.*

Note that the assumptions (4.9), (4.10) and (6.4) are only mentioned to ensure the existence and uniqueness of u and u_h. Given some u (resp. u_h) satisfying (6.1) (resp. (6.2)), the only assumption used to prove (6.5) is (6.3).

Proof. Let $v_h \in V_h$. We deduce from (6.3) that

$$\alpha_h \|u_h - v_h\|_V \leq \sup_{w_h \in W_h,\, w_h \neq 0} \frac{a(u_h - v_h, w_h)}{\|w_h\|_W}.$$

Since $a(u_h, w_h) = a(u, w_h)$ for any $w_h \in W_h$, we get that

$$\alpha_h \|u_h - v_h\|_V \leq \sup_{w_h \in W_h,\, w_h \neq 0} \frac{a(u - v_h, w_h)}{\|w_h\|_W} \leq C_a \|u - v_h\|_V.$$

We now use the triangle inequality:

$$\|u - u_h\|_V \leq \|u - v_h\|_V + \|u_h - v_h\|_V \leq \left(1 + \frac{C_a}{\alpha_h}\right) \|u - v_h\|_V.$$

Since $v_h \in V_h$ is arbitrary, we deduce that

$$\|u - u_h\|_V \leq \left(1 + \frac{C_a}{\alpha_h}\right) \inf_{v_h \in V_h} \|u - v_h\|_V,$$

which is (6.5). □

As in the coercive setting (see the Céa lemma 5.7), the best approximation error $\inf_{v_h \in V_h} \|u-v_h\|_V$ is an upper bound (up to a multiplicative constant) of the error $u-u_h$. Obviously, it is also a lower bound of the error, since $\inf_{v_h \in V_h} \|u - v_h\|_V \leq \|u - u_h\|_V$. If the best approximation error converges to 0, and provided that there exists $\alpha > 0$ such that $\alpha_h \geq \alpha$, then so does the error $u - u_h$, at the same rate.

As in the coercive case, we see that the error estimation is based on two ingredients:

- the quality of the approximation space, which is quantified by the best approximation error $\inf_{v_h \in V_h} \|u - v_h\|_V$, and which is independent of the equation which is considered (and of the fact that the equation is, or is not, coercive). This best approximation error can be estimated in terms of h in view of Theorem 5.15 (for $\mathbb{P}_1$ finite elements), Theorem 5.29 (for $\mathbb{P}_2$ finite elements), ...
- the stability of the problem, which is quantified by the constant $1 + C_a/\alpha_h$, where C_a is the continuity constant of a and α_h is the inf-sup constant appearing in (6.3) (in the coercive setting, this stability is quantified by the constant C_a/α, where C_a is again the continuity constant of a, and α is its coercivity constant). We note that α_h depends on the problem which is studied (namely the bilinear form a) and also on the discretization spaces V_h and W_h.

6.2 The Stokes problem

We now consider the Stokes problem (4.16), which reads

$$\begin{cases} \text{Find } (u,p) \in (H^1_0(\Omega))^d \times L^2_0(\Omega) \text{ such that} \\ -\Delta u + \nabla p = f \quad \text{in } [\mathcal{D}'(\Omega)]^d, \\ \text{div } u = 0 \quad \text{in } \mathcal{D}'(\Omega). \end{cases} \tag{6.6}$$

We recall that a possible variational formulation is (4.17), which falls within the general framework (6.1). Another variational formulation, more specific to the Stokes problem, is (4.19), which reads, we recall

$$\begin{cases} \text{Find } (u,p) \in X \times M \text{ such that} \\ \forall v \in X, \qquad a_0(u,v) + b(v,p) = g_1(v), \\ \forall q \in M, \qquad b(u,q) = g_2(q), \end{cases} \tag{6.7}$$

where $X = (H^1_0(\Omega))^d$, $M = L^2_0(\Omega) = \left\{ q \in L^2(\Omega), \quad \int_\Omega q = 0 \right\}$, a_0 is the bilinear form defined on $X \times X$ by

$$a_0(u,v) = \int_\Omega \nabla u \cdot \nabla v = \sum_{i=1}^d \int_\Omega \nabla u_i \cdot \nabla v_i,$$

b is the bilinear form defined on $X \times M$ by

$$\forall v \in X, \quad \forall p \in M, \qquad b(v,p) = -\int_\Omega p \, \text{div } v,$$

g_1 is the linear form defined on X by

$$g_1(v) = \int_\Omega f \cdot v = \sum_{i=1}^d \int_\Omega f_i \, v_i,$$

and g_2 is the linear form on M defined by $g_2(q) = 0$.

We assume in this section that X and M are two Hilbert spaces, that the bilinear (resp. linear) forms a_0 and b (resp. g_1 and g_2) are continuous, and that the bilinear form a_0 is coercive on X.

The well-posedness of (6.7) is established by Theorem 4.16. We have seen there that (6.7) is well-posed if and only if b satisfies the inf-sup condition (4.20).

6.2.1 Galerkin approximation of the abstract problem (6.7)

Our aim here is to approximate (6.7) in a finite dimensional setting. For the sake of simplicity, we again restrict ourselves to a conformal approximation. We thus introduce $X_h \subset X$ and $M_h \subset M$ and consider the problem

$$\left\{ \begin{array}{ll} \text{Find } (u_h, p_h) \in X_h \times M_h \text{ such that} & \\ \forall v_h \in X_h, & a_0(u_h, v_h) + b(v_h, p_h) = g_1(v_h), \\ \forall q_h \in M_h, & b(u_h, q_h) = g_2(q_h). \end{array} \right. \tag{6.8}$$

In view of Theorem 4.16 (this time applied in a finite dimensional setting), this problem is well-posed if and only b satisfies an inf-sup condition in $X_h \times M_h$, that is

$$\exists \beta_h > 0, \quad \forall q_h \in M_h, \qquad \sup_{v_h \in X_h, \, v_h \neq 0} \frac{b(v_h, q_h)}{\|v_h\|_X} \geq \beta_h \|q_h\|_M. \tag{6.9}$$

Note that β_h may depend on h (but should be positive).

We again observe that the inf-sup condition (4.20) on $X \times M$ does not imply the inf-sup condition (6.9) on $X_h \times M_h$. In general, there is thus no alternative to checking that (6.9) indeed holds on $X_h \times M_h$. We will come back to this in the subsequent sections.

We now turn to estimating the error between (u, p) and (u_h, p_h). We have the following result:

Lemma 6.2. *We suppose that the assumption* (4.20) *holds. Consider the subspaces $X_h \subset X$ and $M_h \subset M$, and assume that* (6.9) *holds as well for some β_h satisfying $\beta_h \geq \beta > 0$ where β is independent of the mesh size h.*

Let (u, p) be the solution to (6.7) *and (u_h, p_h) be the solution to* (6.8). *Then there exists a constant C independent of h such that*

$$\|u - u_h\|_X + \|p - p_h\|_M \leq C \left(\inf_{v_h \in X_h} \|u - v_h\|_X + \inf_{q_h \in M_h} \|p - q_h\|_M \right). \tag{6.10}$$

Note that the assumption (4.20) is only mentioned to ensure the existence and uniqueness of (u, p). Given some (u, p) (resp. (u_h, p_h)) satisfying (6.7) (resp. (6.8)), the only assumption used to prove (6.10) is (6.9).

Proof. Let $v_h \in X_h$ and $q_h \in M_h$. We write

$$\begin{aligned} \|p - p_h\|_M &\leq \|p - q_h\|_M + \|q_h - p_h\|_M \\ &\leq \|p - q_h\|_M + \beta_h^{-1} \sup_{v_h \in X_h, \, v_h \neq 0} \frac{b(v_h, q_h - p_h)}{\|v_h\|_X} \qquad \text{[using (6.9)]} \\ &= \|p - q_h\|_M + \beta_h^{-1} \sup_{v_h \in X_h, \, v_h \neq 0} \frac{b(v_h, q_h - p) + b(v_h, p - p_h)}{\|v_h\|_X} \\ &\leq \|p - q_h\|_M + \beta_h^{-1} \sup_{v_h \in X_h, \, v_h \neq 0} \frac{b(v_h, q_h - p)}{\|v_h\|_X} + \beta_h^{-1} \sup_{v_h \in X_h, \, v_h \neq 0} \frac{b(v_h, p - p_h)}{\|v_h\|_X}. \end{aligned} \tag{6.11}$$

We infer from the first line of (6.7) and the first line of (6.8) that

$$\forall v_h \in X_h, \qquad a_0(u-u_h,v_h)+b(v_h,p-p_h)=0. \tag{6.12}$$

Likewise, the second line of (6.7) and the second line of (6.8) yield that

$$\forall q_h \in M_h, \qquad b(u-u_h,q_h)=0. \tag{6.13}$$

Collecting (6.11) and (6.12), we obtain

$$\begin{aligned}\|p-p_h\|_M &\le \|p-q_h\|_M + \beta_h^{-1} \sup_{v_h\in X_h,\, v_h\neq 0} \frac{b(v_h,q_h-p)}{\|v_h\|_X} + \beta_h^{-1} \sup_{v_h\in X_h,\, v_h\neq 0} \frac{a_0(u-u_h,v_h)}{\|v_h\|_X} \\ &\le \left(1+\frac{C_b}{\beta_h}\right)\|p-q_h\|_M + \frac{C_a}{\beta_h}\|u-u_h\|_X, \end{aligned} \tag{6.14}$$

where C_b (resp. C_a) is the continuity constant of the bilinear form b (resp. the bilinear form a_0).

Now using the coercivity of a_0 (with the constant α), we write

$$\begin{aligned}\alpha\|u-u_h\|_X^2 &\le a_0(u-u_h,u-u_h) \\ &= a_0(u-u_h,u-v_h)+a_0(u-u_h,v_h-u_h) \\ &= a_0(u-u_h,u-v_h)-b(v_h-u_h,p-p_h) \qquad \text{[using (6.12)]} \\ &= a_0(u-u_h,u-v_h)-b(v_h-u,p-p_h)-b(u-u_h,p-p_h) \\ &= a_0(u-u_h,u-v_h)-b(v_h-u,p-p_h)-b(u-u_h,p-q_h) \qquad \text{[using (6.13)]} \\ &\le C_a\|u-u_h\|_X\|u-v_h\|_X + C_b\|u-v_h\|_X\|p-p_h\|_M + C_b\|u-u_h\|_X\|p-q_h\|_M. \end{aligned}$$

Collecting the first and the third term, we get

$$\begin{aligned}\alpha\|u-u_h\|_X^2 &\le (C_a+C_b)\|u-u_h\|_X\left(\|u-v_h\|_X+\|p-q_h\|_M\right) + C_b\|u-v_h\|_X\|p-p_h\|_M \\ &\le (C_a+C_b)\|u-u_h\|_X\left(\|u-v_h\|_X+\|p-q_h\|_M\right) \\ &\quad + C_b\left(1+\frac{C_b}{\beta_h}\right)\|u-v_h\|_X\|p-q_h\|_M + C_b\frac{C_a}{\beta_h}\|u-v_h\|_X\|u-u_h\|_X \\ &\le \left(C_a+C_b+\frac{C_a\,C_b}{\beta_h}\right)\|u-u_h\|_X\left(\|u-v_h\|_X+\|p-q_h\|_M\right) \\ &\quad + C_b\left(1+\frac{C_b}{\beta_h}\right)\|u-v_h\|_X\|p-q_h\|_M, \end{aligned} \tag{6.15}$$

where we have used (6.14) at the second line. Let

$$\tau_h = \inf_{v_h\in X_h}\|u-v_h\|_X + \inf_{q_h\in M_h}\|p-q_h\|_M$$

be the best approximation error. Since $v_h \in X_h$ and $q_h \in M_h$ are arbirarty, we deduce from (6.15) that

$$\|u-u_h\|_X^2 \le C\|u-u_h\|_X\ \tau_h + C\tau_h^2$$

where C is a constant depending on C_a, C_b, α and β_h. This implies that

$$\|u-u_h\|_X \le C\tau_h.$$

Inserting this estimate in (6.14), we obtain

$$\|p-p_h\|_M \le C\tau_h.$$

This concludes the proof of (6.10). □

6.2.2 Checkerboard instability for the Stokes problem

In this section, we consider a specific choice of $X_h \subset X$ and $M_h \subset M$ such that the inf-sup condition (6.9) does not hold. The choice we describe here is the most well-known bad choice for X_h and M_h.

We consider the 2D situation, with $\Omega = (0,1)^2$, and we mesh Ω by quadrangles. We then set $X_h = \mathbb{Q}_1$ (i.e. each component of the velocity u_h is assumed to be a $\mathbb{Q}_1$ polynomial, see Exercice 5.39) and

$$M_h = \left\{ q_h \in L^2_0(\Omega), \quad q_h \text{ is constant on each element} \right\}.$$

For any $v \in X_h$ and $q \in M_h$, we denote by $v_{i,j}$ the value of the velocity at node (i,j), and $q_{K_{i,j}}$ the value of the pressure in the element $K_{i,j}$ (for the sake of simplicity, we drop the subscript h for v and q). We then compute

$$b(v,q) = -\int_\Omega q \operatorname{div} v = -\sum_K \int_K q \operatorname{div} v = -\sum_{i,j} q_{K_{i,j}} \int_{K_{i,j}} \operatorname{div} v = -\sum_{i,j} q_{K_{i,j}} \int_{\partial K_{i,j}} v \cdot n.$$

On each edge of $K_{i,j}$, the field $v = (v^x, v^y)$ is affine. We hence have

$$\begin{aligned} & b(v,q) \\ & = -\frac{1}{2} h \sum_{i,j} q_{K_{i,j}} \Big((v^x_{i+1,j} + v^x_{i+1,j+1}) + (v^y_{i,j+1} + v^y_{i+1,j+1}) - (v^x_{i,j} + v^x_{i,j+1}) - (v^y_{i,j} + v^y_{i+1,j}) \Big) \\ & = \frac{1}{2} h \sum_{i,j} (q_{K_{i,j}} - q_{K_{i-1,j}})(v^x_{i,j} + v^x_{i,j+1}) + \frac{1}{2} h \sum_{i,j} (q_{K_{i,j}} - q_{K_{i,j-1}})(v^y_{i,j} + v^y_{i+1,j}) \\ & = h^2 \sum_{i,j} \nabla_x q_{ij} \, v^x_{i,j} + h^2 \sum_{i,j} \nabla_y q_{ij} \, v^y_{i,j}, \end{aligned} \tag{6.16}$$

with the notations

$$\begin{aligned} 2h \nabla_x q_{ij} &= (q_{K_{i,j}} - q_{K_{i-1,j}}) + (q_{K_{i,j-1}} - q_{K_{i-1,j-1}}), \\ 2h \nabla_y q_{ij} &= (q_{K_{i,j}} - q_{K_{i,j-1}}) + (q_{K_{i-1,j}} - q_{K_{i-1,j-1}}). \end{aligned}$$

Note that we have used above, when reordering the terms, the fact that v vanishes on the boundary of Ω.

We wish to build a non-zero field $q \in M_h$ such that $b(v,q) = 0$ for all $v \in X_h$. Since $v^x_{i,j}$ and $v^y_{i,j}$ in (6.16) are arbitrary, the field q should satisfy $\nabla_x q_{ij} = 0$ and $\nabla_y q_{ij} = 0$ for any i and j.

Subtracting the two equations, we get $q_{K_{i,j-1}} = q_{K_{i-1,j}}$. Adding the two equations, we get $q_{K_{ij}} = q_{K_{i-1,j-1}}$. This means that, when moving along each "diagonal" of the cartesian mesh, the value of q should remain the same. The set of solutions is thus completely determined by two contants, say $q_{K_{0,0}} = \alpha$ and $q_{K_{1,0}} = \beta$.

In addition, q should satisfy the constraint of zero-average. This imposes that $\alpha = -\beta$. The pressure field q has thus the structure of a checkerboard, alternating two opposite values. For such a field (called *spurious mode*), we have $b(v,q) = 0$ for any $v \in X_h$.

Remark 6.3. *It could be tempting to remove this spurious mode of the discretization space M_h, by considering*

$$M_h = \left\{ q_h \in L^2_0(\Omega), \quad q_h \textit{ is constant on each element}, \quad \int_\Omega q_h \, q_{\text{spurious}} = 0 \right\}$$

where q_{spurious} is the spurious pressure mode we have identified above. This does not properly fix the issue, as this leads to an inf-sup property (6.9) with some positive β_h, which unfortunately scales as $O(h)$ when $h \to 0$ (see [4, Section 4.2.3] for details).

In order to fix the problem and obtain approximation spaces that satisfy the inf-sup condition (6.9), we have either to enrich X_h or to consider a smaller space M_h.

6.2.3 Numerical locking for the Stokes problem

The fact that the inf-sup condition (6.9) does not hold for $X_h = \mathbb{Q}_1$ and piecewise constant pressure fields is not related to the fact that we use quadrangles. A similar issue arises if we use a mesh made of triangles, as we now show.

We again consider the 2D situation, and we mesh Ω by triangles. We then set $X_h = \mathbb{P}_1$ (i.e. each component of the velocity u_h is assumed to be a $\mathbb{P}_1$ polynomial) and

$$M_h = \left\{ q_h \in L^2_0(\Omega), \quad q_h \text{ is constant on each element} \right\}.$$

For simplicity, we consider a mesh made of isocele triangles (but the argument carries over to general meshes). We denote N^2 the number of internal nodes (so that each "side" of Ω has been cut into $N+1$ elements). We see that dim $X_h = 2N^2$ (two degrees of freedom per internal node) and dim $M_h = 2(N+1)^2 - 1$ (there are $2(N+1)^2$ elements, one degree of freedom per element and we take into account the mean-free constraint), that is dim $M_h = 2N^2 + 4N + 1$.

Consider the operator

$$D : u_h \in X_h \longrightarrow \text{div } u_h \in M_h.$$

If it is surjective, then the inf-sup condition (6.9) is satisfied for the Stokes problem (we just have to follow the arguments of the proof of Theorem 4.18). We have dim Ker D + dim Im D = dim X_h. If D is surjective, then dim Ker D + dim M_h = dim X_h. But this is not possible since dim $M_h >$ dim X_h.

Another (equivalent) way to conclude is to introduce the matrix $B : \mathbb{R}^x \to \mathbb{R}^m$ such that $b(v_h, q_h) = Q_h^T B V_h$, where Q_h (resp. V_h) is the vector collecting the degrees of freedom of $q_h \in M_h$ (resp. of $v_h \in X_h$), and $m = \text{dim } M_h$, $x = \text{dim } X_h$. The inf-sup condition (6.9) reads

$$\exists \beta_h > 0, \quad \forall Q_h \in \mathbb{R}^m, \quad \sup_{V_h \in \mathbb{R}^x,\, V_h \neq 0} \frac{Q_h^T B V_h}{\|V_h\|} \geq \beta_h \|Q_h\|,$$

which is equivalent to the fact that B^T is injective (see Lemma 4.6(i)). We have dim Ker B^T + dim Im B^T = dim M_h, hence

$$\text{dim Ker } B^T = \text{dim } M_h - \text{dim Im } B^T \geq \text{dim } M_h - \text{dim } X_h = 4N + 1 > 0.$$

The matrix B^T thus cannot be injective.

We note that the second line of (6.8) can be recast as $Q_h^T B U_h = 0$ for any $Q_h \in \mathbb{R}^m$. In some cases, the space M_h is so rich (and its dimension so large compared to that of X_h) that the only solution to that equation is $U_h = 0$ (i.e. the matrix B is injective). This motivates the name of *locking* for such a choice of X_h and M_h.

6.2.4 A possible choice of $X_h \times M_h$ for the Stokes problem

We now build an example of spaces $X_h \subset X$ and $M_h \subset M$ such that the inf-sup condition (6.9) holds on $X_h \times M_h$. With the aim to obtain approximation spaces that are easy to manipulate, we wish to only slightly modify the choice of Section 6.2.3 (piecewise affine velocities, piecewise constant pressures).

Consider first the choice $X_h = \mathbb{P}_1$ and

$$M_h = \left\{ q_h \in L^2_0(\Omega), \quad q_h \in \mathbb{P}_1 \right\}. \tag{6.17}$$

For simplicity, we again consider a mesh made of isocele triangles. We denote N^2 the number of internal nodes (so that each "side" of Ω has been cut into $N+1$ elements). The space X_h is the same as in Section 6.2.3. The space M_h is now of dimension $(N+2)^2 - 1 = N^2 + 4N + 3$, which is smaller (as soon as $N \geq 2$) than the dimension of the space M_h in Section 6.2.3. It however turns out that this couple of spaces still does not satisfy the inf-sup condition (6.9) (see [4, Section 4.2.3] for details).

Keeping the pressure space M_h as defined by (6.17), we now enrich the velocity space X_h. In each element K, and for each component j of $v_h \in \mathbb{R}^d$, we assume that

$$v_h^{(j)} \in \text{Span}\left\{\lambda_0^K, \lambda_1^K, \ldots, \lambda_{d+1}^K\right\}, \qquad 1 \leq j \leq d,$$

where, for any $1 \leq i \leq d+1$, we have $\lambda_i^K \in \mathbb{P}_1$ and $\lambda_i^K(x_j) = \delta_{ij}$ (where x_j are the nodes of K) and where $\lambda_0^K \in H_0^1(K)$, $0 \leq \lambda_0^K \leq 1$ on K and $\lambda_0^K(x_K) = 1$, where x_K is the barycenter of K. The functions λ_i^K $(1 \leq i \leq d+1)$ are the usual piecewise affine functions on K, which are equal to 1 at one node of K and vanish at the other nodes of K. The function λ_0^K is a so-called bubble function: it vanishes on ∂K, is non-negative on K and is equal to 1 somewhere (e.g. at the barycenter) in K. For instance, up to a multiplicative constant, we can take $\lambda_0^K = \Pi_{i=1}^{d+1} \lambda_i^K$.

Theorem 6.4. *Let Ω be an open, bounded and connected subset of $\mathbb{R}^d$. Assume that the pressure space M_h is defined by* (6.17) *and that X_h is defined as above. Then the inf-sup condition* (6.9) *holds for some constant $\beta > 0$ independent of h.*

This choice of spaces X_h and M_h is sometimes called the *mini-element*. Other choices can be made (with polynomial functions of higher degree) such that (6.9) is satisfied. We refer to [4, Section 4.2] for details.

Proof. The idea of the proof is as follows. Consider $q_h \in M_h \subset M$. We know from Theorem 4.18 that there exists $v \in X$ such that div $v = -q_h$ with $\rho \|v\|_X \leq \|q_h\|_M$ for some $\rho > 0$.

Assume that we can find some $\Pi_h(v) \in X_h$ such that

$$\forall q_h \in M_h, \quad b(\Pi_h(v), q_h) = b(v, q_h), \tag{6.18}$$

$$\|\Pi_h(v)\|_X \leq c\|v\|_X. \tag{6.19}$$

We then write that

$$\begin{aligned} \sup_{v_h \in X_h,\, v_h \neq 0} \frac{b(v_h, q_h)}{\|v_h\|_X} &\geq \frac{b(\Pi_h(v), q_h)}{\|\Pi_h(v)\|_X} \\ &= \frac{b(v, q_h)}{\|\Pi_h(v)\|_X} \qquad \text{[using (6.18)]} \\ &\geq \frac{b(v, q_h)}{c\|v\|_X} \qquad \text{[using (6.19)]} \\ &= -\frac{1}{c\|v\|_X} \int_\Omega q_h \operatorname{div} v \\ &= \frac{\|q_h\|_M^2}{c\|v\|_X} \qquad \text{[by def. of } v \text{ in terms of } q_h] \\ &\geq \frac{\rho}{c}\|q_h\|_M, \end{aligned}$$

which is exactly the inf-sup condition (6.9) for some constant $\beta > 0$ independent of h.

We are thus left with showing (6.18), which reads

$$\forall q_h \in M_h, \quad \int_\Omega q_h \operatorname{div} \Pi_h(v) = \int_\Omega q_h \operatorname{div} v.$$

Since $q_h \in M_h$ belongs to $H^1(\Omega)$, this condition reads

$$\forall q_h \in M_h, \quad \int_\Omega \Pi_h(v) \cdot \nabla q_h = \int_\Omega v \cdot \nabla q_h.$$

Assume that $\Pi_h(v) \in X_h$ is such that

$$\forall K, \qquad \int_K \Pi_h(v) = \int_K v. \tag{6.20}$$

For any $q_h \in M_h$, we compute

$$\int_\Omega \Pi_h(v) \cdot \nabla q_h = \sum_K \nabla q_h|_K \cdot \int_K \Pi_h(v) = \sum_K \nabla q_h|_K \cdot \int_K v = \int_\Omega v \cdot \nabla q_h$$

and thus obatin (6.18). We are thus left with finding $\Pi_h(v) \in X_h$ satisfying (6.20) and (6.19).

We note that ensuring (6.20) for some $\Pi_h(v)$ which would be in $\mathbb{P}_1$ is challenging, since adjusting the average of $\Pi_h(v)$ on K requires to adjust the values of $\Pi_h(v)$ at the nodes of K, which modifies the average of $\Pi_h(v)$ on the neighboring elements.

Here, the space X_h is richer. In particular, we can adjust the average of $\Pi_h(v)$ on K by simply adjusting the coefficient in front of the bubble function λ_0^K, the support of which is included in K (this adjustment is hence local and does not modify $\Pi_h(v)$ on the other elements). We look for $\Pi_h(v)$ under the form

$$e_i \cdot \Pi_h(v) = e_i \cdot I_h v + \sum_K \gamma_i^K \lambda_0^K,$$

where $(e_i)_{1 \leq i \leq d}$ is the canonical basis of $\mathbb{R}^d$ and $I_h v$ is an accurate approximation of v in the $\mathbb{P}_1$ space (we can take the nodal interpolant of v if v is continuous and thus has well-defined values at the nodes), chosen such that $\|v - I_h v\|_{L^2(\Omega)} \leq Ch\|v\|_{H^1(\Omega)}$.

The parameters γ_i^K are chosen such that (6.20) holds. We hence request that, for any K and i, we have $\displaystyle\int_K e_i \cdot \Pi_h(v) = \int_K e_i \cdot v$, which implies

$$\int_K e_i \cdot v = \int_K e_i \cdot I_h v + \gamma_i^K \int_K \lambda_0^K$$

and hence

$$\gamma_i^K = \frac{\int_K e_i \cdot (v - I_h v)}{\int_K \lambda_0^K}.$$

We admit that, with such a definition of $\Pi_h(v)$, the estimate (6.19) is satisfied. This concludes the proof. □

Appendix A

Basic facts on the Lebesgue integral and L^p spaces

We briefly collect here some basic facts concerning the Lebesgue integral and L^p spaces. For more details, we refer to the first year ENPC course *Outils Mathématiques pour l'Ingénieur.*

In what follows, Ω is an open subset (which may or may not be bounded) of $\mathbb{R}^d$ (with $d \geq 1$) and p is a real number, $p \geq 1$.

We recall that two functions f and g defined on Ω are said to be equal almost everywhere on Ω if the measure of the set $\{x \in \Omega, \ f(x) \neq g(x)\}$ vanishes.

A.1 The space $L^1(\Omega)$

A.1.1 Definition

We first recall that the space $\mathcal{L}^1(\Omega)$ is defined by

$$\mathcal{L}^1(\Omega) = \left\{ f : \Omega \to \overline{\mathbb{R}}, \ f \text{ is measurable}, \ \int_\Omega |f(x)| \, dx < +\infty \right\}.$$

The vector space $L^1(\Omega)$ is obtained as the quotient of the space $\mathcal{L}^1(\Omega)$ with respect to the equivalence relation "$f \sim g$ if $f = g$ almost everywhere". When endowed with the norm

$$\|f\|_{L^1(\Omega)} = \int_\Omega |f(x)| \, dx,$$

the vector space $L^1(\Omega)$ is a Banach space.

A.1.2 Dominated convergence theorem

The following result is of paramount importance.

Theorem A.1. *Let f_n be a sequence of integrable functions in Ω (meaning $f_n \in L^1(\Omega)$) and let g be a non-negative function which is integrable on Ω. We assume that*

- *for almost every $x \in \Omega$, the sequence $(f_n(x))_n$ converges to some $f(x)$ when $n \to \infty$;*
- *for any $n \in \mathbb{N}$, and for almost every $x \in \Omega$, we have $|f_n(x)| \ \leq \ g(x)$.*

Then the function f is integrable on Ω and $\lim_{n\to\infty} \int_\Omega |f(x) - f_n(x)|\, dx = 0$, which implies that

$$\lim_{n\to\infty} \int_\Omega f_n(x)\, dx = \int_\Omega f(x)\, dx.$$

We admit the following result:

Theorem A.2. *Consider a sequence of functions $f_n \in L^1(\Omega)$ that converge to some f in $L^1(\Omega)$ when $n \to \infty$. Then there exists a subsequence of $\{f_n\}_{n\in\mathbb{N}}$ that converges to f almost everywhere when $n \to \infty$.*

A.2 The space $L^2(\Omega)$

Definition A.3. *The space $\mathcal{L}^2(\Omega)$ is defined by*

$$\mathcal{L}^2(\Omega) = \left\{ f : \Omega \to \overline{\mathbb{R}},\ f \text{ is measurable},\ \int_\Omega |f(x)|^2\, dx < +\infty \right\}.$$

The vector space $L^2(\Omega)$ is obtained as the quotient of the space $\mathcal{L}^2(\Omega)$ with respect to the equivalence relation "$f \sim g$ if $f = g$ almost everywhere".

Theorem A.4. *When endowed with the scalar product*

$$(f, g)_{L^2} = \int_\Omega f(x)\, g(x)\, dx,$$

the vector space $L^2(\Omega)$ is a Hilbert space.

A.3 The spaces $L^p(\Omega)$ and $L^p_{\text{loc}}(\Omega)$

Definition A.5. *The space $\mathcal{L}^p(\Omega)$ is defined as*

$$\mathcal{L}^p(\Omega) = \left\{ f : \Omega \to \overline{\mathbb{R}},\ f \text{ is measurable},\ \int_\Omega |f(x)|^p\, dx < +\infty \right\}.$$

The vector space $L^p(\Omega)$ is obtained as the quotient of the space $\mathcal{L}^p(\Omega)$ with respect to the equivalence relation "$f \sim g$ if $f = g$ almost everywhere".

Theorem A.6. *When endowed with the norm*

$$\|f\|_{L^p(\Omega)} = \left(\int_\Omega |f(x)|^p\, dx \right)^{1/p},$$

the vector space $L^p(\Omega)$ is a Banach space.

Theorem A.7 (Hölder inequality)**.** *Let $p \in \mathbb{R}$ with $1 < p < +\infty$, and let $q \in \mathbb{R}$ with $1/p + 1/q = 1$ (note that $1 < q < +\infty$). Let $f \in L^p(\Omega)$ and $g \in L^q(\Omega)$. Then $f\, g \in L^1(\Omega)$ and we have the following Hölder inequality:*

$$\|f\, g\|_{L^1(\Omega)} \le \|f\|_{L^p(\Omega)}\ \|g\|_{L^q(\Omega)}.$$

Lemma A.8. *Assume that the open set Ω is bounded. Let p and q two real numbers with $q > p \geq 1$. Then $L^q(\Omega) \subset L^p(\Omega)$.*

Definition A.9. *We say that f is locally integrable on Ω if, for any compact set K contained in Ω, the function f is integrable on K. We denote*

$$L^1_{\text{loc}}(\Omega) = \big\{f;\ f \in L^1(K) \textit{ for any compact subset } K \subset \Omega\big\}$$

the vector space of locally integrable functions. Likewise,

$$L^p_{\text{loc}}(\Omega) = \{f;\ f \in L^p(K) \textit{ for any compact subset } K \subset \Omega\}\,.$$

Lemma A.10. *We have that $L^p(\Omega) \subset L^p_{\text{loc}}(\Omega)$. For any real numbers p and q such that $q > p \geq 1$, we have $L^q_{\text{loc}}(\Omega) \subset L^p_{\text{loc}}(\Omega)$.*

Definition A.11. *Let f_n be a sequence of functions in $L^p_{\text{loc}}(\Omega)$ and $f \in L^p_{\text{loc}}(\Omega)$. We say that f_n converges to f in $L^p_{\text{loc}}(\Omega)$ when $n \to \infty$ if, for any compact subset $K \subset \Omega$, the sequence $f_n|_K$ converges to $f|_K$ in $L^p(K)$ when $n \to \infty$.*

A.4 The space $L^\infty(\Omega)$

Definition A.12. *A function f is said to be essentially bounded on Ω if there exists a non-negative real number M such that*

$$\mu\Big(\{x \in \Omega;\ |f(x)| \geq M\}\Big) = 0,$$

where μ is the Lebesgue measure. Hence, except on a null set (that is, a set the measure of which vanishes), we have $|f(x)| < M$.

The set of functions that are essentially bounded on Ω is denoted $\mathcal{L}^\infty(\Omega)$.

Definition A.13. *The vector space $L^\infty(\Omega)$ is obtained as the quotient of the space $\mathcal{L}^\infty(\Omega)$ with respect to the equivalence relation "$f \sim g$ if $f = g$ almost everywhere".*

Theorem A.14. *When endowed with the norm*

$$\|f\|_{L^\infty(\Omega)} = \inf\Big\{M \geq 0;\ \mu\Big(\{x \in \Omega;\ |f(x)| \geq M\}\Big) = 0\Big\}\,,$$

the vector space $L^\infty(\Omega)$ is a Banach space.

Lemma A.15. *If f is essentially bounded on Ω, then $|f(x)| \leq \|f\|_{L^\infty(\Omega)}$ almost everywhere on Ω.*

If in addition f is continuous on Ω, then $|f(x)| \leq \|f\|_{L^\infty(\Omega)}$ for any $x \in \Omega$.

Theorem A.16. *If $f \in L^1(\Omega)$ and $g \in L^\infty(\Omega)$, then the product $f\,g$ belongs to $L^1(\Omega)$ and*

$$\|f\,g\|_{L^1(\Omega)} \leq \|f\|_{L^1(\Omega)}\ \|g\|_{L^\infty(\Omega)}.$$

A.5 Other properties

Theorem A.17 (Interpolation)**.** *Let $f \in L^p(\Omega) \cap L^q(\Omega)$ with $1 \leq p \leq q \leq \infty$. Then, for any r such that $p \leq r \leq q$, we have that $f \in L^r(\Omega)$ and*

$$\|f\|_{L^r} \leq \|f\|_{L^p}^{\alpha} \; \|f\|_{L^q}^{1-\alpha}$$

with $\alpha \in (0,1)$ such that $\dfrac{1}{r} = \dfrac{\alpha}{p} + \dfrac{1-\alpha}{q}$.

Definition - Theorem A.18. *Let f and g in $L^1(\mathbb{R}^d)$. The function $f \star g$, defined by*

$$(f \star g)(x) = \int_{\mathbb{R}^d} f(x-y)\, g(y)\, dy$$

is called the convolution *of f and g. It belongs to $L^1(\mathbb{R}^d)$ and*

$$\|f \star g\|_{L^1(\mathbb{R}^d)} \leq \|f\|_{L^1(\mathbb{R}^d)} \; \|g\|_{L^1(\mathbb{R}^d)}.$$

Appendix B

Glossary of mathematical terms

English	French
arithmetic (operation)	(opération) arithmétique
axiom of choice	axiome du choix
almost everywhere (a.e.)	presque partout (p.p.)
ball	boule
basis	base
Banach space	espace de Banach
bilinear	bilinéaire
binomial coefficient	coefficient binomial
Borel set	ensemble borélien
Borel sigma-algebra	tribu borélienne
boundary value problem	problème aux limites
bounded set	ensemble borné
bounded function	fonction bornée
bounded above / bounded from above	majoré
bounded below / bounded from below	minoré
boundary	bord
to cancel	s'annuler
canonical	canonique
Cauchy sequence	suite de Cauchy
Cauchy-Schwarz inequality	inégalité de Cauchy-Schwarz
celestial motion	dynamique céleste
chain rule	dérivation des fonctions composées
characterisation	caractérisation
characteristic function	fonction charactéristique
closed set	ensemble fermé
closure	fermeture
compact set	ensemble compact
compactness	compacité
complement	complémentaire
conformal	conforme
consistency	consistance
concavity	concavité
completeness	complétude
continuity	continuité

English	French
contraction	contraction, fonction contractante
convolution (of functions)	(fonctions) convolées
convolution (as an operation)	convolution (comme opération)
to convolve	convoluer
coordinates	coordonnées
corollary	corollaire
countable	dénombrable
counter-example	contre-exemple
critical point	point critique
to deduce	déduire
determinant	déterminant
derivative	différentielle, dérivée
differentiable	différentiable
differentiation	dérivation
diffeomorphism	difféomorphisme
Dirac mass, Dirac delta	masse de Dirac
Dirichlet problem	problème de Dirichlet
discretisation	discrétisation
distribution	distribution
dominated convergence	convergence dominée
dynamics	dynamique
dynamical	dynamique
error analysis	analyse d'erreur
essentially bounded	essentiellement bornée
to establish	établir
Euclidean space	espace euclidien
explicit (scheme)	(schéma) explicite
exponent	exposant
exponentially	de manière exponentielle
extended	prolongé
factor	facteur
field (algebraic object)	corps
field (function)	champ
finite element method	méthode des éléments finis
finiteness	finitude
finite-dimensional	de dimension finie
fixed point	point fixe
form	forme
Fubini's theorem	théorème de Fubini
function	application, fonction
Galerkin method	méthode de Galerkin
Hardy inequality	inégalité de Hardy
Hamiltonian	Hamiltonienne
Hermitian space	espace hermitien
Heaviside function	fonction de Heaviside
Hessian (matrix)	(matrice) hessienne
Hilbert space	espace de Hilbert
Hilbertian	hilbertien
Hölder inequality	inégalité de Hölder
homeomorphism	homéomorphisme
hyperplane	hyperplan

English	French
hypothesis	hypothèse
implicit (scheme)	(schéma) implicite
to imply	impliquer
in the sense of	au sens de
induced	engendré, induit
inner product	produit scalaire
integral	intégrale
integer	entier
integrable	intégrable
integration by parts	intégration par parties
isomorphism	isomorphisme
isometry	isométrie
index, indices	indice, indices
Jacobian matrix	matrice jacobienne
Jacobian	jacobien
Lax-Milgram (theorem)	théorème de Lax-Milgram
lemma	lemme
Let ... be	Soit ...
linear	linéaire
Lipschitz	Lipschitzien
lower bound	minoration
Lyapunov function	fonction de Lyapunov
mapping	application, fonction
mean value theorem	théorème des accroissements finis
measurable set	ensemble mesurable
measure	mesure
measure space	espace mesuré
mesh	maillage
Minkowski inequality	inégalité de Minkowski
monotonicity	monotonie
neighbourhood	voisinage
norm	norme
normed	normé
null (set)	(ensemble) négligeable
open set	ensemble ouvert
ordinary differential equation	équation différentielle ordinaire
order (of approximation)	ordre (d'approximation)
orthogonal	orthogonal
orthonormal	orthonormé
outward normal vector	vecteur normal sortant
parallelogram law/identity	formule de la médiane
partial derivative	dérivée partielle
partial differential equation	équation aux dérivées partielles
partial ordering	order partiel
piecewise	par morceaux
Poincaré inequality	inégalité de Poincaré
Poisson problem	problème de Poisson
positivity	positivité
potential energy	énergie potentielle
power(s)	puissance(s)
preceding	précédant
probability measure	probabilité

English	French
probability space	espace probabilisé
principal value	valeur principale
quadratic	quadratique
to quotient	quotienter
real number, real	réel
regular	régulier
regularity	régularité
remainder	reste
Riesz representation theorem	Théorème de Riesz
robustness	robustesse
roots	racines
rounding error	erreur d'arrondi
scalar product	produit scalaire
scale	échelle
(numerical) scheme	schéma (numérique)
sequence	suite
(a) series	(une) série
sesquilinear	sesquilinéaire
set	ensemble
sigma-algebra	tribu
singularity	singularité
smooth	régulier, lisse
Sobolev space	espace de Sobolev
square (matrix)	(matrice) carrée
square-integrable	de carré intégrable
stability	stabilité
(time)step	pas (de temps)
step function	fonction etagée
stiffness matrix	matrice de rigidité
subset	sous-ensemble
such that (s.t.)	tel que (t.q.)
to suffice	suffire
sufficiently	suffisament
summable	sommable
symmetric	symétrique
symmetry	symétrie
to tend (to/towards)	tendre, converger (vers)
test function	fonction test
topology	topologie
trajectory	trajectoire
trapezium rule	méthode des trapèzes
triangle inequality	inégalité triangulaire
truncation	troncature
uniform	uniforme
uniqueness	unicité
units	unités
upper bound	majoration
variational problem	problème variationnel
vector space	espace vectoriel
with values in	à valeurs dans
with respect to	par rapport à
Young's inequality	inégalité de Young

Bibliography

[1] H. Brézis, **Analyse fonctionnelle: théorie et applications**, Dunod, 1999.

[2] E. Cancès and A. Ern, **Analyse**, ENPC lecture notes, Sept. 2011.

[3] E. Cancès and A. Ern, **Analyse: Distributions, transformée de Fourier et équations aux dérivées partielles**, in preparation.

[4] A. Ern and J.-L. Guermond, **Theory and practice of finite elements**, Texts in Applied Mathematics, Vol. 159, Springer, New York, 2004.

[5] A. Ern and G. Stoltz, **Calcul Scientifique**, ENPC lecture notes, Dec. 2014.

[6] G. Galdi, **An introduction to the mathematical theory of the Navier–Stokes equations**, volume I, Springer Tracts in Natural Philosophy, Vol. 38, Springer, New York, 1994.

[7] A. Quarteroni and A. Valli, **Numerical approximation of partial differential equations**, Springer Series in Computational Mathematics, vol. 23, Springer-Verlag, Berlin, 1994.