Ciencias Matemáticas, Vol. 30, No. 1, Pag. 1-84, 2016
Recibido 05-2016

Introducción al método de los elementos finitos Introduction to the finite element method

Franck Boyer[1], Sébastien Martin[2]*

Resumen Estas notas presentan los aspectos básicos del método de elemento finito. Se presentan las bases matemáticas del método como las del análisis funcional, la formulación variacional de problemas de contorno para ecuaciones en derivadas parciales, así como el buen planteamiento de los problemas resultantes. Se introducen las aproximaciones de Galerkin que proporcionan la idea general del método de elemento finito. Se muestran ejemplos de espacios de elementos finitos y sus propiedades incluyéndose el análisis de error. La inclusión de aspectos prácticos se realiza a través del análisis de problemas con enfoque en los aspectos matemáticos presentados. El software libre `FreeFem++` se emplea en 2D para ilustrar los principales teoremas mostrados, así como sus limitaciones, en particular tomando presupuestos que no están necesariamente presentes en los teoremas. Se incluyen programas `FreeFem++` y ejercicios.

Abstract This notes presents the basic aspects of the finite element method. The mathematical foundations such as the functional framework, the variational formulation of boundary value partial differential equations and the well-posedness of the problems are presented. Galerkin approximations are introduced, providing the general idea of the finite element method. Examples of finite element spaces and their related properties are presented, including the error analysis. Practical aspects are included through the analysis of problems which focus on the mathematical issues of the notes: The free finite element solver `FreeFem++` is used in 2D to illustrate the main theorems and their limitations, in particular by tackling assumptions that are not necessarily met in the theorems. `FreeFem++` programs and exercises are included.

Palabras Clave
finite element method — elliptic problems — saddle-point problems — `FreeFem++`

[1] *Université Toulouse 3 Paul Sabatier, Institut de Mathématiques de Toulouse CNRS UMR 5219, franck.boyer@math.univ-toulouse.fr*
[2] *Université Paris Descartes, MAP5 CNRS UMR 8145, sebastien.martin@parisdescartes.fr*
***Autor para Correspondencia**

Índice

Introduction

In engineering sciences, finite element solvers are likely to compute an approximate solution of boundary value problems. The goal of designing a numerical method is to ensure the convergence of the method, i.e. to guarantee that the approximate solution is close to the unique solution of the continuous problem (if hopefully it exists) defined by the mathematical model. But understanding the properties of the obtained solution may strongly rely on mathematical aspects such as

1. the mathematical formulation and the well-posedness of the continuous problem: existence, uniqueness, regularity of the solution and stability with respect to the data are required in order to provide a strong basis for the discretization process;

2. the choice of the discretization process: the subsequent approximate finite-dimensional problem has to be well-posed in order to be numerically solved;

3. the error analysis: the quantitative analysis of the approximation should guarantee that the solution to the approximate problem is “close” to the solution to the initial problem.

The aim of this course is to provide an introduction to the mathematical analysis of the finite element method. Practical issues are also addressed.

In section 1 we present the main results in functional analysis, which are the basis of the mathematical analysis at both continuous and discrete levels. In section 2 we present the mathematical formulation of elliptic problems with related results and examples. In section 3 we present the mathematical formulation of saddle-point problems with related results and examples. In section 4 we introduce the so-called Galerkin approximation which defines a class of finite-dimensional problems associated to the continuous problems. In section 5 we introduce the main approximation spaces and establish the main approximation properties of these spaces along with subsequent error estimates of the method. In section 6 we focus on the finite element approximation of saddle-point problems such as the Stokes system for which additional difficulties have to be targetted. In section 7 we present some mathematical problems that are solved with a free finite element solver `FreeFem++` whose formalism respects the variational formulation of problems. Section 8 is dedicated to the `FreeFem++` programs.

1. Sobolev spaces

1.1 Definitions

For any open subset Ω of $\mathbb{R}^d$, we define $\mathscr{D}(\Omega)$ (resp. $\mathscr{D}'(\Omega)$) to be the set of $\mathscr{C}^\infty$ functions compactly supported in Ω (resp. the set of distributions on Ω). For any function $u \in L^1_{loc}(\Omega)$, and any multi-index $\alpha = (\alpha_1, ..., \alpha_d) \in \mathbb{N}^d$, we set

$$\partial^\alpha u = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}} u,$$

in the sense of distributions.

Definition 1.1 (Sobolev space) *Let Ω be an open subset of $\mathbb{R}^d$. For all integer $k \geq 0$, we define the Sobolev space $H^k(\Omega)$ as*

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

The space is endowed with the norm

$$\|u\|_{H^k} = \left(\sum_{|\alpha| \leq k} \|\partial^\alpha u\|^2_{L^2} \right)^{\frac{1}{2}}.$$

.

Remark 1.2 *An equivalent definition consists in defining, by induction, $H^0(\Omega) = L^2(\Omega)$ and*

$$H^{k+1}(\Omega) = \{u \in H^k(\Omega),\ \nabla u \in (H^k(\Omega))^d\}.$$

Definition 1.3 (Domains of $\mathbb{R}^d$) *A non empty open subset Ω of $\mathbb{R}^d$ is said to be a C^k domain (resp. a Lipschitz domain) if its boundary $\partial\Omega$ is a submanifold of $\mathbb{R}^d$ of class C^k (resp. Lipschitz continuous) and if Ω is locally situated only on one side of its boundary.*

For such domains, by using local charts, we can define the outward unit normal field $n : \partial\Omega \to \mathbb{R}^d$ as well as surface integrals $\int_\partial f$ and corresponding Lebesgue spaces $L^p(\partial)$.

Proposition 1.4 (Density of smooth functions) *Let Ω be a Lipschitz bounded domain of $\mathbb{R}^d$, then the set $C^\infty(\overline{\Omega})$ made of smooth functions up to the boundary of Ω is dense in $H^k(\Omega)$ for any $k \geq 0$.*

The following subspace of $H^k(\Omega)$ will be very important in the study of boundary-value problems.

Definition 1.5 *For any open subset Ω in $\mathbb{R}^d$ we define $H_0^k(\Omega)$ as the closure of $\mathscr{D}(\Omega)$ in $H^k(\Omega)$.*

The topological dual of $H_0^k(\Omega)$ (i.e. the space made of all continuous linear forms defined on this space) may be characterized as follows.

Definition 1.6 *For all $k \geq 0$, we define the following space of distributions:*

$$H^{-k}(\Omega) = \left\{ f \in \mathscr{D}'(\Omega),\ f = \sum_{|\alpha| \leq k} \partial^\alpha f_\alpha,\ \text{with } f_\alpha \in L^2(\Omega) \right\}.$$

The space is endowed with the norm

$$\|f\|_{H^{-k}} = \inf \left(\sum_{|\alpha| \leq k} \|f_\alpha\|_{L^2}^2 \right)^{\frac{1}{2}}.$$

the infimum being taken among all the possible decompositions of f.

Proposition 1.7 (Dual of Sobolev spaces) *The space $H^{-k}(\Omega)$, $k \geq 0$, is a Hilbert space which is isomorphic to the dual space of $H_0^k(\Omega)$. To be more precise, the duality bracket can be expressed as*

$$\langle f, u \rangle_{H^{-k}, H_0^k} = \sum_{|\alpha| \leq k} (-1)^{|\alpha|} \int_\Omega f_\alpha \, \partial^\alpha u,$$

for all $f \in H^{-k}(\Omega)$ and for all $u \in H_0^k(\Omega)$. The formula does not depend on the choice of decomposition for f.

Note that the topological dual of $H^k(\Omega)$ is not a distribution space and cannot be characterized so easily.

Remark 1.8 *The main application of Proposition 1.7 is the following result: any element $f \in H^{-1}(\Omega)$ can be written as*

$$f = u + \operatorname{div}(G), \quad \text{with } u \in L^2(\Omega) \text{ and } (G \in L^2(\Omega))^d$$

in the sense of distributions.

1.2 Basic properties

Theorem 1.9 (Chain rule) *Let Ω be an open subset of $\mathbb{R}^d$. For any function $u \in H^1(\Omega)$ and any function $T \in C^1(\mathbb{R};\mathbb{R})$ with a bounded derivative, we have*

$$T(u) \in H^1(\Omega), \quad \nabla T(u) = T'(u) \nabla u.$$

Moreover,

$$\begin{array}{rccc} T & : & H^1(\Omega) & \to & H^1(\Omega) \\ & & u & \mapsto & T(u) \end{array}$$

is a (non-linear) continuous map.

Actually the chain rule is still valid if T is Lipschitz continuous and piecewise C^1, for instance. In this case, T' may be defined in a non univoque way in some points but the theorem still applies, which implies that the values of T' on these particular points are not important.

Example. If $T(x) = x^+ = \max(x,0)$, we define $T'(x) = 1$ if $x > 0$ and $T'(x) = 0$ if $x \leq 0$. Then we obtain

$$\nabla(u^+) = \mathbf{1}_{u>0} \nabla u.$$

But if we define $T'(x) = 1$ if $x \geq 0$ and $T'(x) = 0$ if $x < 0$, we obtain

$$\nabla(u^+) = \mathbf{1}_{u \geq 0} \nabla u.$$

The theorem states that $\nabla u = 0$ almost everywhere on the set $\{u = 0\}$. □

Let us recall the main results on Sobolev embeddings: in this part, we focus on $H^1(\Omega)$ even if similar results can be proved for spaces $H^k(\Omega)$, $k > 1$.

Theorem 1.10 (Sobolev embeddings) *Let Ω be a bounded Lipschitz domain of $\mathbb{R}^d$.*

- *If $d = 1$, the embedding $H^1(\Omega) \subset L^\infty(\Omega)$ is continuous.*
- *If $d = 2$, the embedding $H^1(\Omega) \subset L^p(\Omega)$, for all $p \in [2, +\infty[$, is continuous.*
- *If $d \geq 3$, the embedding $H^1(\Omega) \subset L^{p^*}(\Omega)$ is continuous, with $p^* = \frac{2d}{d-2}$.*

Theorem 1.11 (Rellich-Kondrachov) *Let Ω be a bounded Lipschitz domain of $\mathbb{R}^d$.*

- *If $d = 1$, the embedding $H^1(\Omega) \subset C^0(\bar{\Omega})$ is compact (completely continuous).*
- *If $d = 2$, the embedding $H^1(\Omega) \subset L^p(\Omega)$, for all $p \in [1, +\infty[$, is compact.*
- *If $d \geq 3$, the embedding $H^1(\Omega) \subset L^p(\Omega)$, for all $p \in [1, p^*[$ with $p^* = \frac{2d}{d-2}$, is compact.*

In particular, the embedding $H^{k+1}(\Omega) \subset H^k(\Omega)$, $k \geq 0$, is compact.

Theorem 1.12 (Morrey's inequality) *Let Ω be a bounded Lipschitz domain of $\mathbb{R}^d$. If*

$$k - \alpha = d/2$$

with $\alpha \in (0,1)$, then one has the following embedding:

$$H^k(\Omega) \subset C^{0,\alpha}(\bar{\Omega}).$$

As a consequence, $H^k(\Omega) \subset C^0(\bar{\Omega})$ if $k > d/2$.

1.3 Trace operator

A Lipschitz continuous function which is defined over an open subset Ω can be naturally extended up to the boundary of Ω. This allows us to define the notion of *trace* of such a function on the boundary $\partial\Omega$. In some way, it is possible to build a similar notion for functions which lack such a regularity.

Theorem 1.13 (Trace) *Let Ω be a* bounded *Lipschitz domain of $\mathbb{R}^d$. The mapping*

$$\begin{array}{cccc} \gamma_0 & : & C^\infty(\bar{\Omega}) & \to & L^2(\partial\Omega) \\ & & u & \mapsto & \gamma_0(u) = u_{|\partial\Omega} \end{array}$$

has a unique continuous extension to the whole Sobolev space $H^1(\Omega)$. This operator, still denoted γ_0, is called the trace operator. *Besides,*

- *the trace operator $\gamma_0 : H^1(\Omega) \to L^2(\partial\Omega)$ is not surjective onto $L^2(\partial\Omega)$;*
- *the image of $H^1(\Omega)$ by the trace operator is a fractional Sobolev space called $H^{\frac{1}{2}}(\partial\Omega)$ which is a Hilbert space if it is endowed with the norm*

$$\|v\|_{H^{\frac{1}{2}}(\partial\Omega)} = \inf_{\substack{u \in H^1(\Omega) \\ \gamma_0(u) = v}} \|u\|_{H^1(\Omega)}.$$

Moreover, there exists a (non unique) continuous linear operator $R_0 : H^{\frac{1}{2}}(\partial\Omega) \to H^1(\Omega)$, called lift operator, which satisfies

$$\gamma_0 \circ R_0 = \mathrm{Id}_{H^{\frac{1}{2}}(\partial\Omega)}.$$

Proposition 1.14 (Integration by parts) *The trace operator extends the notion of integration by parts: for all $u \in H^1(\Omega)$, for all $\Phi \in (H^1(\Omega))^d$,*

$$\int_\Omega u \,\mathrm{div}(\Phi) + \int_\Omega \nabla u \cdot \Phi = \int_{\partial\Omega} \gamma_0(u)\, \gamma_0(v).$$

Example. The case of the unit square is quite interesting. Let us consider $\Omega =]0,1[^2$ and $u \in C^\infty(\Omega)$. We have, for all $(x,y) \in]0,1[^2$,

$$u^2(x,0) = u^2(x,y) - 2\int_0^y u(x,z)\, \partial_y u(x,z)\, \mathrm{d}z.$$

Then we get

$$u^2(x,0) \le u^2(x,y) + 2\int_0^y |u(x,z)|\, \left|\partial_y u(x,z)\right| \mathrm{d}z.$$

Integrating with respect to x and using the Cauchy-Schwarz inequality, we get

$$\begin{array}{rcl} \int_0^1 u^2(x,0)\,\mathrm{d}x & \le & \int_\Omega u^2 + 2\|u\|_{L^2}\, \|\partial_y u\|_{L^2} \\ & \le & \|u\|_{L^2}^2 + 2\|u\|_{L^2} \|\nabla u\|_{L^2}, \\ & \le & 2\|u\|_{L^2} \|u\|_{H^1}, \end{array}$$

and finally

$$\|u\|_{L^2(]0,1[\times\{0\})} \le \sqrt{2}\|u\|_{H^1}.$$

The above inequality shows that the mapping $\gamma_0 : C^\infty(\bar{\Omega}) \to L^2(\partial\Omega)$ is continuous for the H^1-topology. Its natural continuous extension to $H^1(\Omega)$ is then straightforward by density (see Proposition 1.4).

Observe finally that we have in fact shown the more precise inequality

$$\|u\|_{L^2(]0,1[\times\{0\})} \le C\|u\|_{L^2}^{\frac{1}{2}} \|u\|_{H^1}^{\frac{1}{2}},$$

which implies that we can estimate the L^2 norm of the trace by using only *the square root of the norm of the gradient of u.* This is a formal justification of the notation $H^{\frac{1}{2}}(\partial)$ that we adopted for the range of the trace operator.

□

Exercise 1 1. *Let $Q = (0,1)^2$ be the unit cube in $\mathbb{R}^2$. Prove that, there exists a $C > 0$, such that for any $u \in H^1(Q)$, we have*

$$\left| \int_0^1 (\gamma_0 u)(x,0)\,\mathrm{d}x - \int_Q u \right|^2 \le C\int_Q |\nabla u|^2.$$

2. *Let $\alpha \in (0,1)$ and $Q_\alpha = (0,\alpha)^2$. Prove that for any $u \in H^1(Q)$ we have*

$$\left| \frac{1}{\alpha}\int_0^\alpha (\gamma_0 u)(x,0)\,\mathrm{d}x - \frac{1}{\alpha^2}\int_{Q_\alpha} u \right|^2 \le C\int_{Q_\alpha} |\nabla u|^2,$$

and

$$\left| \frac{1}{\alpha}\int_0^\alpha (\gamma_0 u)(0,y)\,\mathrm{d}y - \frac{1}{\alpha^2}\int_{Q_\alpha} u \right|^2 \le C\int_{Q_\alpha} |\nabla u|^2.$$

3. *Let $g \in L^2(\partial Q)$ be defined as follows*

$$g(x,0) = g(x,1) = 0, \quad \forall x \in (0,1),$$

$$g(0,y) = g(1,y) = 1, \quad \forall y \in (0,1).$$

Using the inequalities of question 2, prove that it does not exist a function $u \in H^1(Q)$ such that $\gamma_0 u = g$. We have thus proved that this function g does not belong to the trace space $H^{\frac{1}{2}}(\partial Q)$.

In the case of regular domains of $\mathbb{R}^d$, the space $H_0^1(\Omega)$ can be characterized by the following properties:

Proposition 1.15 *Let Ω be a bounded Lipschitz domain of $\mathbb{R}^d$ and $u \in H^1(\Omega)$. The following are equivalent:*

- *$u \in H_0^1(\Omega)$;*
- *$\bar{u} \in H^1(\mathbb{R}^d)$, where $\bar{u}$ is the extension of u over $\mathbb{R}^d$ satisfying $\bar{u} = 0$ on $\mathbb{R}^d \setminus \Omega$.*

It is also possible to consider the trace of functions in $H^1(\Omega)$ when combined with a nonlinear function:

Proposition 1.16 *Let Ω be a bounded Lipschitz domain of $\mathbb{R}^d$ and $T \in C^1(\mathbb{R};\mathbb{R})$ with a bounded derivative. Then*

$$\forall u \in H^1(\Omega), \quad \gamma_0(Tu) = T(\gamma_0(u)).$$

This property still holds if T is Lipschitz continuous and piecewise C^1 with a countable number of discontinuity points for T'.

1.4 Poincaré inequalities

The trace operator allows us to give a clear characterization of $H_0^1(\Omega)$.

Proposition 1.17 *Let Ω be a bounded Lipschitz domain of $\mathbb{R}^d$. We have*

$$H_0^1(\Omega) = \mathrm{Ker}(\gamma_0).$$

The space $H_0^1(\Omega)$ is thus the set of functions in $H^1(\Omega)$ which are equal to 0 on the boundary in the sense of traces.

Let us go back to the previous example, dealing with the unit square, and assume that u is equal to 0 on the boundary. A similar computation gives

$$\int_\Omega u^2 \leq 2\|u\|_{L^2}\|\nabla u\|_{L^2},$$

hence

$$\|u\|_{L^2} \leq C\|\nabla u\|_{L^2}.$$

This inequality states that the L^2-norm is controlled by the L^2-norm of its derivatives, in the case of functions that are equal to 0 on the boundary.

Theorem 1.18 (Poincaré-Friedrichs I) *Let Ω be a bounded Lipschitz domain of $\mathbb{R}^d$. There exists a constant $C > 0$ such that*

$$\forall u \in H_0^1(\Omega), \quad \|u\|_{L^2} \leq C\|\nabla u\|_{L^2}.$$

In particular the map $u \mapsto \|\nabla u\|_{L^2}$ is a norm on $H_0^1(\Omega)$, which is equivalent to the classical H^1-norm. Other similar inequalities are available:

Theorem 1.19 (Poincaré-Friedrichs II) *Let Ω be a* connected *bounded Lipschitz domain of $\mathbb{R}^d$. Consider Γ a closed part of $\partial\Omega$ with a non-empty relative interior and define*

$$H_\Gamma^1(\Omega) = \{u \in H^1(\Omega),\ \gamma_0(u) = 0 \text{ on } \Gamma\}.$$

There exists a constant $C > 0$ such that

$$\forall u \in H_\Gamma^1(\Omega), \quad \|u\|_{L^2} \leq C\|\nabla u\|_{L^2}.$$

Theorem 1.20 (Poincaré-Wirtinger) *Let Ω be a* connected *bounded Lipschitz domain of $\mathbb{R}^d$. We define*

$$\tilde{H}^1(\Omega) = \left\{u \in H^1(\Omega),\ \int_\Omega u = 0\right\}.$$

There exists a constant $C > 0$ such that

$$\forall u \in \tilde{H}^1(\Omega), \quad \|u\|_{L^2} \leq C\|\nabla u\|_{L^2}.$$

Actually these theorems follow from the *abstract* inequality:

Theorem 1.21 (Poincaré) *Let Ω be a bounded open subset of $\mathbb{R}^d$. Let $\mathscr{H}$ be a Hilbert space, $L : H^1(\Omega) \to \mathscr{H}$ be a continuous linear operator. We assume that L is nonzero over the set of nonzero locally constant functions. Then there exists a constant $C > 0$ such that*

$$\forall u \in H^1(\Omega), \quad \|u\|_{L^2} \leq C(\|\nabla u\|_{L^2} + \|L(u)\|_{\mathscr{H}}).$$

In particular the map $u \mapsto \|\nabla u\|_{L^2} + \|L(u)\|_{\mathscr{H}}$ is a norm on $H^1(\Omega)$, which is equivalent to the classical H^1-norm.

Proof of Theorem 1.21. We proceed by *reductio ad absurdum*. Assume that the property is false. This implies that there exists a sequence of functions u_n in $H^1(\Omega)$ satisfying

$$\|u_n\|_{L^2} \geq n(\|\nabla u_n\|_{L^2} + \|L(u_n)\|_{\mathscr{F}}).$$

By homogeneity we may assume that $\|u_n\|_{L^2} = 1$ so that

$$\|\nabla u_n\|_{L^2} + \|L(u_n)\|_{\mathscr{F}} \leq \frac{1}{n}. \tag{1}$$

Thus $\{u_n\}$ is bounded in $H^1(\Omega)$ and consequently there exists some $u \in H^1(\Omega)$ such that $\{u_n\}$ weakly converges to u in $H^1(\Omega)$, up to a subsequence. The embedding $H^1 \subset L^2$ is compact (here we use the fact that Ω is bounded) so that the sequence strongly converges in $L^2(\Omega)$, which implies that $\|u\|_{L^2} = 1$. Then the sequence $\{\nabla u_n\}$

- weakly converges to ∇u in $L^2(\Omega)$, as $\{u_n\}$ weakly converges to u in $H^1(\Omega)$,
- strongly converges to 0 in $L^2(\Omega)$, because of (1),

hence $\nabla u = 0$. This implies that u is locally constant. Besides, we have

- $L(u_n)$ strongly converges to 0 in $\mathscr{F}$, because of (1),
- $L(u_n)$ weakly converges to $L(u)$ in $\mathscr{F}$, see[1].

[1] The result follows from the proposition:

Proposition 1.22 *Let $\mathscr{H}$ and $\mathscr{F}$ be Hilbert spaces. Assume that $L : \mathscr{H} \to \mathscr{F}$ is a continuous linear operator. If $\{u_n\}$ weakly converges to u in $\mathscr{H}$, then $\{L(u_n)\}$ weakly converges to $L(u)$ in $\mathscr{F}$.*

Proof. Assume that $u_n \rightharpoonup u$ in $\mathscr{H}$. We fix $T \in \mathscr{F}'$ and we aim at proving that $T(Lu_n) \to T(Lu)$. We have $T \circ L \in \mathscr{H}'$ and as $\{u_n\}$ weakly converges to u in $\mathscr{H}$, we get $(T \circ L)(u_n) \to (T \circ L)u$, i.e. $T(Lu_n) \to T(Lu)$. ■

and thus $L(u) = 0$. Since u is locally constant, by assumption on L, we obtain that necessarily $u = 0$ which is in contradiction with the property $\|u\|_{L^2} = 1$. This concludes the proof. ■

1.5 Vector fields in L^2 with divergence in L^2. Definition of the normal trace

Definition 1.23 (Space H_{div}) *The space*

$$H_{\text{div}}(\Omega) = \{u \in (L^2(\Omega))^d,\ \text{div}(u) \in L^2(\Omega)\}$$

endowed with the norm

$$\|u\|_{H_{\text{div}}} = (\|u\|_{L^2}^2 + \|\text{div}(u)\|_{L^2}^2)^{\frac{1}{2}}$$

is a Hilbert space.

Theorem 1.24 *Let Ω be a regular bounded open subset of $\mathbb{R}^d$. The set of vector fields of class $(C^\infty(\bar{\Omega}))^d$ is dense in $H_{\text{div}}(\Omega)$.*

The proof of Proposition 1.24 is based upon the Hahn-Banach theorem that we recall:

Theorem 1.25 (Hahn-Banach) *Let E be a Banach space and F a subspace of E. Then F is dense in E if, and only if, any continuous linear form on E which is zero over F is the zero form.*

Proof of Proposition 1.24. In order to apply the Hahn-Banach theorem, we consider a continuous linear form $\mathscr{F}$ on $H_{\text{div}}(\Omega)$ which is zero on $(C^\infty(\bar{\Omega}))^d$. Let us prove that it is zero on the whole space $H_{\text{div}}(\Omega)$.

Since $H_{\text{div}}(\Omega)$ is a Hilbert space, the Riesz representation theorem guarantees that $\mathscr{F}$ is represented by an element $f \in H_{\text{div}}(\Omega)$ by

$$\langle \mathscr{F}, u\rangle_{H'_{\text{div}},H_{\text{div}}} = \int_\Omega f \cdot u + \int_\Omega \text{div}(f)\,\text{div}(u), \quad \forall u \in H_{\text{div}}(\Omega).$$

By assumption, we have

$$\int_\Omega f \cdot \phi + \int_\Omega \text{div}(f)\,\text{div}(\phi) = 0, \quad \forall \phi \in (C^\infty(\bar{\Omega}))^d.$$

We denote $\overline{f}$ (resp. $\overline{\text{div}(f)}$) the extension of f (resp. $\text{div}(f)$) by 0 over $\mathbb{R}^d$ we get

$$\int_{\mathbb{R}^d} \overline{f} \cdot \phi + \int_{\mathbb{R}^d} \overline{\text{div}(f)}\,\text{div}(\phi) = 0, \quad \forall \phi \in (\mathscr{D}(\mathbb{R}^d))^d.$$

This shows in particular that $\overline{\text{div}(f)} \in H^1(\mathbb{R}^d)$ and

$$\nabla(\overline{\text{div}(f)}) = \overline{f}.$$

Since $\text{div}(f) \in H^1(\Omega)$ and $\overline{\text{div}(f)} \in H^1(\mathbb{R}^d)$, then $g := \text{div}(f) \in H^1_0(\Omega)$ (see Proposition 1.15) and $\nabla g = f$. Then the linear form $\mathscr{F}$ writes

$$\langle \mathscr{F}, u\rangle_{H'_{\text{div}},H_{\text{div}}} = \int_\Omega u \cdot \nabla g + \int_\Omega \text{div}(u)\, g,\ \forall u \in H_{\text{div}}(\Omega).$$

As $\mathscr{D}(\Omega)$ is dense in $H^1_0(\Omega)$ there exists a sequence $\{g_n\} \in \mathscr{D}(\Omega)$ which converges to g in $H^1_0(\Omega)$. But for any n, we have

$$\int_\Omega u \cdot \nabla g_n + \int_\Omega \text{div}(u)\, g_n = 0,\ \forall u \in H_{\text{div}}(\Omega),$$

by definition of the divergence in the sense of distributions. Thus we can pass to the limit and get

$$\int_\Omega u \cdot f + \int_\Omega \text{div}(u)\,\text{div}(f) = 0,\ \forall u \in H_{\text{div}}(\Omega),$$

i.e. $\mathscr{F} = 0$ as it is expected. ■

Let us recall that $H^{\frac{1}{2}}(\partial\Omega)$ is the image of the trace operator of $\gamma_0 : H^1(\Omega) \to L^2(\partial\Omega)$ and that the norm of this space can be defined by

$$\|\phi\|_{H^{\frac{1}{2}}(\partial\Omega)} = \inf_{\substack{u \in H^1(\Omega)\\ \gamma_0(u)=\phi}} \|u\|_{H^1(\Omega)}.$$

We define $H^{-\frac{1}{2}}(\partial\Omega)$ as the dual space of $H^{\frac{1}{2}}(\partial\Omega)$ after identifying $L^2(\partial\Omega)$ to its own dual space.

Proposition 1.26 (Stokes formula) *Let Ω be a Lipschitz continuous open subset of $\mathbb{R}^d$. The mapping*

$$\begin{array}{cccc} \gamma_n & : & (C^\infty(\bar{\Omega}))^d & \to & H^{-\frac{1}{2}}(\partial\Omega) \\ & & u & \mapsto & \gamma_n(u) = (u \cdot n)_{|\partial\Omega} \end{array}$$

where n denotes the outward normal unit vector at $\partial\Omega$, has an unique continuous extension on the whole space $H_{\text{div}}(\Omega)$. Moreover we have, for all $u \in H_{\text{div}}(\Omega)$ and for all $w \in H^1(\Omega)$,

$$\int_\Omega u \cdot \nabla w + \int_\Omega w\,\text{div}(u) = \langle \gamma_n u, \gamma_0 w\rangle_{H^{-\frac{1}{2}},H^{\frac{1}{2}}}.$$

Remark 1.27 *If $u \in (H^1(\Omega))^d$, we have in particular $u \in H_{\text{div}}(\Omega)$ and the so-called* normal trace *γ_n has two definitions. In fact, they coincide and we have*

$$\gamma_0 u \cdot n = \gamma_n u.$$

In this case, $\gamma_n u$ belongs to $H^{\frac{1}{2}}(\Omega)$.

Proof of Proposition 1.26. Let us recall that, by Theorem 1.13, there exists a continuous linear lift operator $R_0 : H^{\frac{1}{2}}(\partial\Omega) \to H^1(\Omega)$ satisfying

$$\gamma_0 \circ R_0 = \text{Id}_{H^{\frac{1}{2}}(\partial\Omega)}.$$

Let us proceed in two steps:

Step 1. For all $u \in H_{\text{div}}(\Omega)$ and $\phi \in H^{\frac{1}{2}}(\partial\Omega)$, we define $X_u \in H^{-\frac{1}{2}}(\partial\Omega)$ as

$$X_u(\phi) = \int_\Omega (R_0(\phi)\,\text{div}(u) + u \cdot \nabla R_0(\phi)),\ \forall \phi \in H^{\frac{1}{2}}(\partial\Omega).$$

We have

$$|X_u(\phi)| \le \|R_0(\phi)\|_{H^1}\|u\|_{H_{\mathrm{div}}} \le C\|\phi\|_{H^{\frac{1}{2}}(\partial\Omega)}\|u\|_{H_{\mathrm{div}}},$$

using the continuity of R_0. This shows that, u being fixed, X_u is a continuous linear form on $H^{\frac{1}{2}}(\partial\Omega)$ which satisfies

$$\|X_u\|_{H^{-\frac{1}{2}}(\partial\Omega)} \le C\|u\|_{H_{\mathrm{div}}}.$$

This inequality shows that the mapping

$$u \in H_{\mathrm{div}} \mapsto X_u \in H^{-\frac{1}{2}}(\partial\Omega)$$

is linear and continuous. We denote $\gamma_n(u) := X_u$ and it remains to prove that the desired property is satisfied.

Step 2. Let $u \in (C^\infty(\bar\Omega))^d$. If w_1 and w_2 are functions in $H^1(\Omega)$ such that $\gamma_0 w_1 = \gamma_0 w_2$, then we obtain by integration by parts

$$\int_\Omega ((w_1 - w_2)\,\mathrm{div}(u) + u\cdot\nabla(w_1 - w_2)) = \underbrace{\int_{\partial\Omega}\gamma_0(w_1 - w_2)(u\cdot n)}_{=0}.$$

This shows that, for $u \in (C^\infty(\bar\Omega))^d$ for all $w \in H^1(\Omega)$, by taking $w_1 = w$ and $w_2 = R_0(\gamma_0(w))$, we get

$$\begin{aligned}\langle u\cdot n, \gamma_0 w\rangle &= \int_\Omega (w\,\mathrm{div}(u) + u\cdot\nabla w)\\ &= \int_\Omega (R_0((\gamma_0(w))\,\mathrm{div}(u) + u\cdot\nabla R_0((\gamma_0(w)))\\ &= X_u(\gamma_0(w)) = \langle\gamma_n u, \gamma_0 w\rangle_{H^{-\frac{1}{2}},H^{\frac{1}{2}}}.\end{aligned}$$

This holds for any regular function and the proof is concluded by density, using the continuity of γ_n. ■

Theorem 1.28 *Let Ω be a Lipschitz continuous bounded open subset of $\mathbb{R}^d$ and let us denote $H_{0,\mathrm{div}}(\Omega)$ the closure of $(\mathscr{D}(\Omega))^d$ in $H_{\mathrm{div}}(\Omega)$. Then*

$$H_{0,\mathrm{div}}(\Omega) = \mathrm{Ker}(\gamma_n).$$

Proof of Theorem 1.28. It is clear that $H_{0,\mathrm{div}}(\Omega) \subset \mathrm{Ker}(\gamma_n)$ as $(\mathscr{D}(\Omega))^d \subset \mathrm{Ker}(\gamma_n)$ by definition of the trace operator. Let us prove that $\mathrm{Ker}(\gamma_n) \subset H_{0,\mathrm{div}}(\Omega)$.

Let $u \in H_{\mathrm{div}}(\Omega)$ such that $\gamma_n(u) = 0$. Since Ω is bounded and Lipschitz continuous, we can prove that there exists a *finite open covering* of $\bar\Omega$ by open subsets ω_i, $i \in \{1,...,m\}$, such that $\omega_i \cap \Omega$ is a star-shaped subdomain of $\mathbb{R}^d$. We now introduce a C^∞ partition of unity $(\alpha_i)_i$ subordinate to this open covering: there exists a family $\{f_i\}_{i=1,...,m}$ such that

- $\forall i \in \{1,...,m\}$, $f_i \in C^\infty(\mathbb{R}^d)$,
- $\forall i \in \{1,...,m\}$, $\mathrm{supp}(f_i) \subset \omega_i$,
- $\chi_{\bar\Omega} \le \sum_{i=1}^m f_i \le 1$.

Let $\phi \in \mathscr{D}(\mathbb{R}^d)$ which we restrict to Ω and we apply the Stokes formula, see Proposition 1.26 :

$$\int_\Omega \phi\ \mathrm{div}(u) + \int_\Omega u\cdot\nabla\phi = 0.$$

We replace ϕ by $\alpha_i\phi$ so that we get

$$\int_{\Omega\cap\omega_i}\alpha_i\phi\ \mathrm{div}(u) + \int_{\Omega\cap\omega_i} u\cdot(\phi\nabla\alpha_i + \alpha_i\nabla\phi) = 0.$$

We may rewrite this equality as

$$\begin{aligned}&\int_{\mathbb{R}^d}(\mathbf{1}_{|\Omega\cap\omega_i}\alpha_i\ \mathrm{div}(u) + \mathbf{1}_{|\Omega\cap\omega_i}u\cdot\nabla\alpha_i)\phi\\ &\qquad + \int_{\mathbb{R}^d}(\mathbf{1}_{|\Omega\cap\omega_i}u\alpha_i)\cdot\nabla\phi \quad = \quad 0.\end{aligned}$$

The above equality holds for any test function ϕ, which means that

$$v_i := \mathbf{1}_{|\Omega\cap\omega_i}u\alpha_i,$$

which belongs to $(L^2(\mathbb{R}^d))^d$, has a divergence in the sense of distributions, which is identified as

$$\mathbf{1}_{|\Omega\cap\omega_i}\alpha_i\ \mathrm{div}(u) + \mathbf{1}_{|\Omega\cap\omega_i}u\cdot\nabla\alpha_i$$

which also belongs to $L^2(\mathbb{R}^d)$. As a consequence, $v_i \in H_{\mathrm{div}}(\mathbb{R}^d)$ and its support is in $\Omega\cap\omega_i$ which is an open star-shaped subset by assumption. In order to simplify the notations we assume that the open subset is a star-shaped domain with respect to 0. We now introduce the family of functions defined by the homothetic ratio $\theta \in (0,1)$

$$v_i^\theta(x) = v_i\left(\frac{x}{\theta}\right).$$

The compact support of each function v_i^θ is in $\Omega\cap\omega_i$. Moreover it is clear that v_i^θ converges to v_i in $H_{\mathrm{div}}(\mathbb{R}^d)$ as θ goes to 1. For a fixed $\varepsilon > 0$, we may find $\theta \in (0,1)$ such that

$$\|v_i - v_i^\theta\|_{H_{\mathrm{div}}(\Omega)} \le \varepsilon.$$

But since v_i^θ has a compact support in $\Omega\cap\omega_i$, we may use a convolution process that allows us to regularize the function and prevent the support of the function from spreading out of $\Omega\cap\omega_i$. Thus there exists $\eta > 0$ such that

$$\rho_\eta \star v_i^\theta \in (\mathscr{D}(\Omega\cap\omega_i))^d,$$
$$\|v_i^\theta - \rho_\eta * v_i^\theta\|_{H_{\mathrm{div}}} \le \varepsilon.$$

As a consequence, we obtain

$$\|v_i - \rho_\eta * v_i^\theta\|_{H_{\mathrm{div}}} \le 2\varepsilon.$$

This process allows us to approximate each v_i by functions of $(\mathscr{D}(\Omega))^d$ for the $H_{\mathrm{div}}-$norm. Using the properties of α_i, we get

$$u = \sum_{i=1}^m u\alpha_i = \sum_{i=1}^m v_i.$$

As the sum is finite, the homothetic mapping combined with the regularization may be used on each v_i, which allows us to conclude. ■

2. Weak formulation of elliptic problems

Solving a PDE leads to different questions:

- the solution or the coefficients of the PDE may be not regular enough to guarantee a classical sense to the equation;
- the solution may be regular but the functional space may not have suitable properties to guarantee existence and / or uniqueness of the solution in this space.

Thus we define a weaker notion of solution, even if we have in mind that it should match the classical notion when regularity is met. The general method consists in proceeding this way:

- to find a solution in a functional space which is weaker than the one that was primarily targetted;
- to determine a mathematical formulation by using test functions and *formal* integration by parts;
- to ensure that 1) classical solutions are weak solutions, 2) weak solutions with additional regularity properties are classical solutions.

The difficulties are the following ones:

- the choice of the functional space is crucial: it is non trivial and not necessarily unique;
- if the notion of solution is too weakened, the existence issue is easier but the uniqueness properties may not hold or be harder to prove;
- if the notion of solution is not weakened enough, the uniqueness issue is easier but the existence of the solution may be too hard to prove.

2.1 Lax-Milgram theorem

Many problems in applied mathematics arise out of a variational formulation. In the case of so-called elliptic partial differential equations, the basic existence and uniqueness result is based on an abstract theorem, the Lax-Milgram theorem, which provides a powerful guideline in the analysis of these problems.

Theorem 2.1 (Lax-Milgram) *Let H be a Hilbert space, $a(\cdot,\cdot)$ a bilinear form on H, L a linear form on H. We assume that*

1. *$a(\cdot,\cdot)$ is continuous:*
$$\forall u,v \in H, \quad |a(u,v)| \leq \|a\|\|u\|_H\|v\|_H.$$
2. *$a(\cdot,\cdot)$ is coercive:*
$$\exists \alpha > 0, \quad \forall u \in H, \quad a(u,u) \geq \alpha\|u\|_H^2.$$
3. *L is continuous:*
$$\forall u \in H, \quad |L(u)| \leq \|L\|\|u\|_H.$$

We consider the abstract problem

$$(\mathrm{P})\begin{cases} \text{Find } u \in H \text{ such that} \\ a(u,v) = L(v), \quad \forall v \in H. \end{cases}$$

Problem (P) *admits a unique solution. Moreover,*

$$\|u\|_H \leq \frac{\|L\|}{\alpha}.$$

If a is symmetric, then u is the unique minimizer on H of the functional

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

Proof of Theorem 2.1.
Symmetric case I. Assume that $a(\cdot,\cdot)$ is symmetric. Let us prove that the variational problem and the minimization problems are equivalent.

- Writing $v = u+v-u$, we have
$$\begin{aligned} J(v) &= \frac{1}{2}\left(a(u,u)+2a(u,v-u)+a(v-u,v-u)\right) \\ &\qquad -L(u)-L(v-u) \\ &= J(u)+\left(a(u,v-u)-L(v-u)\right) \\ &\qquad +\frac{1}{2}a(v-u,v-u). \end{aligned}$$
If u is a solution of $a(u,\cdot) = L$, we have
$$J(v)-J(u) = \frac{1}{2}a(v-u,v-u) \geq \alpha\|v-u\|^2,$$
hence $J(u) = \min_{v\in V} J(v)$ and it is the unique solution of the minimization problem.
- Conversely if $u \in V$ is a solution of
$$J(u) \leq J(v), \quad \forall v \in V,$$
then for all $h \in V$ and for all $t \in \mathbb{R}$, we have $J(u+th) \geq J(u)$, i. e.
$$t(a(u,h)-L(h)) + \frac{t^2}{2}a(h,h) \geq 0, \ \forall t \in \mathbb{R}.$$
Necessarily we get $a(u,h) = L(h)$, for all $h \in V$, i. e. $a(u,\cdot) = L$.

Symmetric case II. The well-posedness of the variational problem can be proved with two methods:

- By assumption, $a(\cdot,\cdot)$ is a scalar product on H which is equivalent to the usual scalar product on H. Then, $(H, a(\cdot,\cdot))$ is the same topological space as $(H,(\cdot,\cdot)_H)$ and the linear form L is continuous on this *new* Hilbert space. The result follows from the Riesz representation theorem.

- We directly study the minimization problem by considering a minimizing sequence and proving that it is a Cauchy sequence by using the parallelogram law.

General case. The well-posedness of the variational problem can be proved as follows. For all $u \in H$, the form $v \in H \mapsto a(u,v)$ is continuous so that there exists a unique $Au \in H$ such that

$$a(u,v) = (Au,v)_H, \quad \forall v \in H,$$

as a consequence of the Riesz representation theorem. It is clear that

- A is linear;

- A is continuous: if one takes $v = Au$ in the continuity assumption of a we find

  $$\|Au\|_H^2 = a(u,Au) \leq \|a\|\|u\|_H\|Au\|_H,$$

  so that

  $$\|Au\|_H \leq \|a\|\|u\|_H, \quad \forall u \in H,$$

- A is injective: indeed, by using the coercivity property of a, we find that

  $$(Au,u)_H = a(u,u) \geq \alpha\|u\|_H^2, \quad \forall u \in H,$$

  and thus

  $$\|Au\|_H \geq \alpha\|u\|_H, \quad \forall u \in H.$$

  In particular, we have $Au = 0 \Rightarrow u = 0$.

By the Riesz representation theorem, L can be represented by an element $\ell \in H$ and we finally have to prove that there exists $u \in H$ such that $Au = \ell$. Let $\rho > 0$ and

$$\begin{array}{cccc} T & : & H & \mapsto & H \\ & & u & \mapsto & Tu = u - \rho(Au - \ell). \end{array}$$

Our problem reduces to the proof of existence of a fixed-point for T. In a Hilbert space, this can be achieved by showing that T is a contraction.

$$\begin{array}{rcl} \|Tu - Tv\|_H^2 & = & \|u-v\|_H^2 + \rho^2\|Au - Av\|_H^2 \\ & & -2\rho(u-v, A(u-v))_H \\ & \leq & \|u-v\|_H^2 + \rho^2\|a\|^2\|u-v\|_H^2 \\ & & -2\rho\alpha\|u-v\|_H^2 \\ & \leq & (1 - 2\rho\alpha + \rho^2\|a\|^2)\|u-v\|_H^2. \end{array}$$

For ρ sufficiently small, T is a contraction, which concludes the proof. ■

Remark 2.2 *The Lax-Milgram theorem provides sufficient conditions to find a unique solution of an abstract problem. These conditions are not necessary! For instance in finite dimension, i.e. $H = \mathbb{R}^n$, the continuity assumptions are obvious and the coercivity assumption is expressed as a condition over the matrix that represents the operator. Indeed denoting $\{e_i\}_{i=1,\dots,n}$ the canonical basis of $\mathbb{R}^n$, $X = (x_i)_{i=1,\dots,n}$ the coordinates of $u \in \mathbb{R}^n$ (i.e. $u = \sum_{i=1}^n x_i e_i$), we denote $A = (a(e_j,e_i))_{ij}$ so that we have $a(u,u) = (AX,X)_{\mathbb{R}^n}$ where $(\cdot,\cdot)_{\mathbb{R}^n}$ is the usual scalar product on $\mathbb{R}^n$. Then, by assumption of coercivity*

$$\begin{array}{rcl} \frac{1}{2}\left((AX,X) + (X,A^{\mathrm{t}}X)\right)_{\mathbb{R}^n} & = & (AX,X)_{\mathbb{R}^n} \\ & = & a(u,u) \\ & \geq & \alpha\|u\|_{\mathbb{R}^n}^2 = \alpha\|X\|_{\mathbb{R}^n}^2. \end{array}$$

Thus coercivity of $a(\cdot,\cdot)$ means that $A + A^{\mathrm{t}}$ is s.p.d. The theorem then becomes: if $A + A^{\mathrm{t}}$ is s.p.d. then A is invertible. This is clearly a sufficient condition but not a necessary one.

In the problems that will be considered, H is the functional space (L^2, H^1, H_0^1, etc.), u denotes the (weak) solution of the problem and functions v serve as test functions. Besides,

- the Lax-Milgram theorem requires some constraints: in particular, the test functions and the solution should belong to the same functional space. This is the reason why the derivation of the weak formulation relies on an equilibrium between the derivatives of the solution and the derivatives of the test functions.

- The regularity of the functions in H cannot be completely independent from the bilinear form a. If the functions in H are too regular, the coercivity property will not be obtained; conversely if the functions in H are not sufficiently regular, the continuity of a will not be obtained.

  - Consider $a(u,v) = \int_\Omega u(-\Delta v)$. Then Δv has to be defined, which leads us to define $H = H^2(\Omega) \cap H_0^1(\Omega)$. The bilinear form is continuous, since

    $$\begin{array}{rcl} |a(u,v)| & \leq & \|u\|_{L^2}\|\Delta v\|_{L^2} \\ & \leq & \|u\|_{L^2}\|v\|_{H^2} \\ & \leq & \|u\|_H\|v\|_H. \end{array}$$

    Unfortunately the coercivity is missing: if $u \in H$ we have

    $$\begin{array}{rcl} a(u,u) & = & \int_\Omega u(-\Delta u) \\ & = & \int_\Omega u\,\mathrm{div}(-\nabla u) \\ & = & \int_{\partial\Omega} u(-\nabla u)\cdot n + \int_\Omega |\nabla u|^2 \\ & = & \int_\Omega |\nabla u|^2. \end{array}$$

Clearly there is no $\alpha > 0$ such that

$$\|\nabla u\|_{L^2}^2 \geq \alpha \|u\|_H^2$$

since it would imply that $H^2(\Omega) \cap H_0^1(\Omega) = H_0^1(\Omega)$.

– Consider $a(u,v) = \int_\Omega \nabla u \nabla v$ and $H = H^1(\Omega)$. The continuity of the bilinear form is guaranteed but the coercivity is missing because $a(u,u) = 0$ implies that u is a constant function. Then $u \equiv 1$ is a counter example for the existence of a coercivity constant.

2.2 Example 1: Poisson problem

Homogeneous Dirichlet boundary conditions. Assume that $f \in L^2(\Omega)$. We consider the following PDE problem:

$$\left(\mathrm{P}_s^{(1)}\right) \left\{ \begin{array}{rcll} -\Delta u &=& f & \text{in } \Omega, \\ u &=& 0 & \text{on } \partial\Omega. \end{array} \right.$$

Consider a regular test function and let us deal with formal computations. Multiplying by v, integrating over Ω and integrating by parts, we get

$$\int_\Omega \nabla u \cdot \nabla v - \int_{\partial\Omega} v\,(\nabla u \cdot n) = \int_\Omega f v.$$

Now let us discuss this equality:

- Integrals over Ω only deal with derivatives of order 1 at most. Thus looking for the continuity and coercivity of the bilinear form leads us to consider (a subspace of) $H^1(\Omega)$.

- No boundary condition on $\nabla u \cdot n$ is prescribed. But we target the definition of a functional space that contains the solution u and the test functions v. As the solution should be zero on the boundary, we may impose this condition on the test functions as well, hence killing the boundary term. The resulting functional space is therefore $H_0^1(\Omega)$.

The weak problem (or variational problem) associated to the strong problem $\left(\mathrm{P}_s^{(1)}\right)$ is

$$\left(\mathrm{P}_w^{(1)}\right) \left\{ \begin{array}{l} \text{Find } u \in H_0^1(\Omega) \text{ such that} \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v = \int_\Omega f v, \\ \qquad \text{for all } v \in H_0^1(\Omega). \end{array} \right.$$

Exercise 2 *Prove that the variational problem* $\left(\mathrm{P}_w^{(1)}\right)$ *admits a unique solution.*

Exercise 3 *Let* $f \in L^p(\Omega)$, $p \in [1, +\infty]$. *Which values of* p *can be considered in order to apply the same method as above?*

The variational problem admits a unique solution u. What are the regularity properties of the weak solution? What is the link with the initial problem?

Note that $\nabla u \in (L^2(\Omega))^d$ and $\mathscr{D}(\Omega) \subset H$ so that we can choose $v \in \mathscr{D}(\Omega)$ as a test function:

$$\int_\Omega \nabla u \cdot \nabla \phi = \int_\Omega f \phi, \quad \forall \phi \in \mathscr{D}(\Omega).$$

By definition, the divergence of ∇u is equal to $-f$ in the sense of distributions. Then u is a solution of $-\Delta u = f$ in the sense of distributions. Actually it is possible to prove the following regularity result:

Theorem 2.3 (Elliptic regularity) *Assume that* Ω *is a* C^2 *bounded domain of* $\mathbb{R}^d$. *Let* $f \in L^2(\Omega)$ *and* $u \in H_0^1(\Omega)$ *the unique variational solution of the Poisson problem with homogeneous Dirichlet conditions* $\left(\mathrm{P}_w^{(1)}\right)$. *We have*

- $u \in H^2(\Omega)$,
- $\|u\|_{H^2} \leq C \|f\|_{L^2}$, *where* C *is a constant which only depends on* Ω.

The proof of this theorem is technical and voluntarily omitted. Interestingly all the second order partial derivatives of u are functions in L^2. But the theorem contains an important regularity assumption on the domain. This condition can be weakened (for instance, if Ω is convex and polygonal the result is still valid) but some counter-examples do exist otherwise: typically if Ω has a *re-entrant corner*, one may find variational solutions in $H_0^1(\Omega) \setminus H^2(\Omega)$.

Example. Consider the domain defined in polar coordinates by $\Omega = \{(r,\theta), r \leq 1, 0 \leq \theta \leq \alpha\}$, as in Fig. 1.

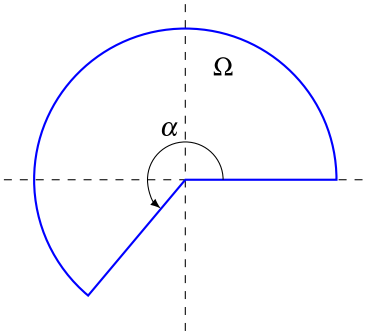


Figure 1. The domain Ω

The function $\tilde{u} : (r,\theta) \mapsto r^{\frac{\pi}{\alpha}} \sin(\frac{\pi\theta}{\alpha})$ is harmonic. Thus $\tilde{u}$ satisfies

$$\left\{ \begin{array}{rcll} -\Delta \tilde{u} &=& 0 & \text{in } \Omega, \\ \tilde{u} &=& g & \text{on } \partial\Omega. \end{array} \right.$$

where g is a continuous function on $\partial\Omega$ satisfying:

$$g(r,\alpha) = g(r,0) = 0, \quad 0 \leq r \leq 1.$$

Define $\chi_\varepsilon \in C^\infty(\mathbb{R}^2)$ such that

- $\chi_\varepsilon(r,\theta) = 1$ for $0 \le r \le \frac{1}{2}$,
- $\chi_\varepsilon(r,\theta) = 0$ for $\frac{1}{2} + \varepsilon \le r$.

and define $u := \tilde{u}\chi_\varepsilon$. Thus u satisfies

$$\left\{ \begin{array}{rcll} -\Delta u &=& f_\varepsilon & \text{in } \Omega, \\ u &=& 0 & \text{on } \partial\Omega. \end{array} \right.$$

with

- $f_\varepsilon := -\Delta\tilde{u} \in L^2(\Omega)$,
- $\mathrm{supp}(f_\varepsilon) \subset \Omega \cap \{(r,\theta), \frac{1}{2} \le r \le \frac{1}{2} + \varepsilon\}$.

Moreover, $u \equiv \tilde{u}$ on $\Omega \cap \{(r,\theta), 0 \le r \le \frac{1}{2}\}$. In particular, because of the behaviour of u (or $\tilde{u}$) near the corner, we can prove that $u \in H^1(\Omega) \setminus H^2(\Omega)$ for $\alpha > \pi$. □

Non-homogeneous Dirichlet boundary conditions. Assume that $f \in L^2(\Omega)$ and $g \in H^{\frac{1}{2}}(\partial\Omega)$. We consider the following PDE problem:

$$\left(\mathrm{P}_s^{(2)}\right) \left\{ \begin{array}{rcll} -\Delta u &=& f & \text{in } \Omega, \\ u &=& g & \text{on } \partial\Omega. \end{array} \right.$$

The choice of the functional space for g relies on the fact that the functional space for the variational formulation is likely to be $H^1(\Omega)$, according to the previous example. The trace of the solution should be in $H^{\frac{1}{2}}(\partial\Omega)$. Dealing with less regular boundary data would lead us outside the Lax-Milgram framework.

By Theorem 1.13, there exists $R_0 g \in H^1(\Omega)$ such that $\gamma_0(R_0 g) = g$. We define $\tilde{u} = u - R_0 g$ so that the problem $\left(\mathrm{P}_q^{(2)}\right)$ is equivalent to the following one

$$\left(\mathrm{P}_s^{(2')}\right) \left\{ \begin{array}{rcll} -\Delta\tilde{u} &=& f + \Delta(R_0 g) & \text{in } \Omega, \\ \tilde{u} &=& 0 & \text{on } \partial\Omega. \end{array} \right.$$

As $R_0 g \in H^1(\Omega)$, we have $\nabla(R_0 g) \in (L^2(\Omega))^d$ and $\Delta(R_0 g) = \mathrm{div}(\nabla(R_0 g)) \in H^{-1}(\Omega)$ by Proposition 1.7. Thus the variational formulation of $\left(\mathrm{P}_s^{(2')}\right)$ reduces to

$$\left(\mathrm{P}_w^{(2')}\right) \left\{ \begin{array}{l} \text{Find } \tilde{u} \in H_0^1(\Omega) \text{ such that} \\ \displaystyle\int_\Omega \nabla\tilde{u} \cdot \nabla\tilde{v} = \int_\Omega f\tilde{v} + \langle \Delta(R_0 g), \tilde{v} \rangle_{H^{-1},H_0^1}, \\ \qquad \text{for all } \tilde{v} \in H_0^1(\Omega). \end{array} \right.$$

This variational formulation falls into the scope of the Lax-Milgram theorem, hence $\tilde{u}$ is uniquely defined. Let us remark that the duality term can be expressed as:

$$\langle \Delta(R_0 g), \tilde{v} \rangle_{H^{-1},H_0^1} = -\int_\Omega \nabla(R_0 g) \cdot \nabla\tilde{v}.$$

Remark 2.4 *Considering $H_g^1(\Omega) := \{u \in H^1(\Omega), \gamma_0 u = g\}$, the function $u = \tilde{u} + R_0 g$ is the unique solution of*

$$\left(\mathrm{P}_w^{(2'')}\right) \left\{ \begin{array}{l} \textit{Find } u \in H_g^1(\Omega) \textit{ such that} \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v = \int_\Omega fv, \\ \qquad \textit{for all } v \in H_0^1(\Omega). \end{array} \right.$$

The Lax-Milgram cannot directly be applied to this formulation: $H_g^1(\Omega)$ is not a vector space, let alone a Hilbert space! Nevertheless this formulation does not require the introduction of the lift of the boundary term, whose computation would be difficult in practice; moreover this formulation is used in practical computations.

Exercise 4 *Prove that the solution u of the problem $\left(\mathrm{P}_w^{(2'')}\right)$ is also the unique solution of the minimisation problem*

$$J(u) = \inf_{v \in H_g^1(\Omega)} J(v),$$

with

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

Remark 2.5 *The elliptic regularity properties proved in Theorem 2.3 can be extended to the Poisson problem with non-homogeneous Dirichlet boundary conditions if we assume furthermore that $g \in H^{\frac{3}{2}}(\Omega)$ which is, by definition, the image of $H^2(\Omega)$ by the trace operator.*

Remark 2.6 *We consider the Poisson problem defined on a unit square $\Omega =]0,1[^2$ with a source term $f \equiv 0$. We consider Dirichlet boundary conditions: for this purpose, we denote*

$$\Gamma :=]0,1[\times \{0\}$$

and consider the problem

$$\left\{ \begin{array}{rcll} -\Delta u &=& f & \textit{in } \Omega, \\ u &=& 0 & \textit{on } \Gamma, \\ u &=& 1 & \textit{on } \partial\Omega \setminus \Gamma. \end{array} \right.$$

The boundary term

$$g(x) = \left\{ \begin{array}{ll} 0 & \textit{if } x \in \Gamma, \\ 1 & \textit{if } x \in \partial\Omega \setminus \Gamma. \end{array} \right.$$

does not belong to $H^{\frac{1}{2}}(\partial\Omega)$ (see Exercise 1). As a consequence, we cannot use the lift operator to define a variational formulation in $H^1(\Omega)$. In fact the solution of this problem can not *belong to $H^1(\Omega)$ (by definition of $H^{\frac{1}{2}}(\partial O)$). Interestingly, when using the finite element method, the discretization of this problem leads to a well-posed problem. But when considering the solution u_h, then $\|u_h\|_{H^1}$ explodes as h goes to 0, illustrating the ill-posedness of the (continuous) variational problem in $H^1(\Omega)$. See also the Problem given in Section 7.3.*

Non-homogeneous Neumann boundary conditions. Assume that Ω is connected, $f \in L^2(\Omega)$ and $g \in H^{-\frac{1}{2}}(\partial\Omega)$. We consider the following PDE problem:

$$\left(\mathrm{P}_s^{(3)}\right) \left\{ \begin{array}{rcll} -\Delta u &=& f & \text{in } \Omega, \\ \nabla u \cdot n &=& g & \text{on } \partial\Omega. \end{array} \right.$$

The choice of the functional space for g is quite natural: the functional space for the variational formulation is likely to be (a subspace of) $H^1(\Omega)$, according to the previous examples, so that $\nabla u \in (L^2(\Omega))^d$. As we ask for $\mathrm{div}(\nabla u) = \Delta u = -f \in L^2(\Omega)$, if a solution exists, we will necessarily get $\nabla u \in H_{\mathrm{div}}$, which allows us to give a weak sense to the normal trace at the boundary of $\nabla u \cdot n$ in the space $H^{-\frac{1}{2}}(\partial\Omega)$.

In this problem, compatibility conditions emerge:

- If the problem admits a solution, then integrating over Ω we obtain (the data are assumed to be regular):
$$\int_\Omega f = \int_\Omega (-\Delta u) = -\int_{\partial\Omega} \nabla u \cdot n = -\int_{\partial\Omega} g.$$
Existence of a solution requires a necessary condition. If $g \in H^{-\frac{1}{2}}(\partial\Omega)$, the compatibility condition writes
$$\int_\Omega f = -\langle g, 1\rangle_{H^{-\frac{1}{2}},H^{\frac{1}{2}}}.$$
- The solution *cannot* be unique: we may add any constant to a solution, thus defining another solution. An additional condition is required in order to select a unique solution. A natural way, since Ω is assumed to be connected, consists in fixing the constant by imposing
$$\int_\Omega u = 0.$$

In the general approach we choose a regular test function v and proceed as before. The boundary value of the solution is not prescribed so that no condition is imposed on the test functions v. By contrast we impose that u has zero mean value so that the same condition should be imposed to the test functions. We finally obtain the variational formulation

$$(\mathrm{P}_w^{(3)}) \left\{ \begin{array}{l} \text{Find } u \in \tilde{H}^1(\Omega) \text{ such that} \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v = \int_\Omega f v + \langle g, v\rangle_{H^{-\frac{1}{2}},H^{\frac{1}{2}}}, \\ \qquad \text{for all } v \in \tilde{H}^1(\Omega). \end{array} \right.$$

Exercise 5 *Prove that the variational problem admits a unique solution.*

The variational problem admits a unique solution. It is possible to get a formulation with test functions in $H^1(\Omega)$ instead of $\tilde{H}^1(\Omega)$. Indeed, for any $v \in H^1(\Omega)$, we consider $\tilde{v} = v - m(v)$ where $m(v) = |\Omega|^{-1} \int_\Omega v$. Then we get
$$\int_\Omega \nabla u \cdot \nabla \tilde{v} = \int_\Omega f \tilde{v} + \langle g, \tilde{v}\rangle_{H^{-\frac{1}{2}},H^{\frac{1}{2}}}.$$
Using $\nabla v = \nabla \tilde{v}$,
$$\begin{aligned} \int_\Omega \nabla u \cdot \nabla v \quad = \quad & \int_\Omega f v - m(v) \left(\int_\Omega f \right) \\ & + \langle g, v\rangle_{H^{-\frac{1}{2}},H^{\frac{1}{2}}} - m(v) \langle g, 1\rangle_{H^{-\frac{1}{2}},H^{\frac{1}{2}}}. \end{aligned}$$

Using the compatibilty condition we get
$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega f v + \langle g, v\rangle_{H^{-\frac{1}{2}},H^{\frac{1}{2}}}. \tag{2}$$

Taking test functions $\phi \in \mathscr{D}(\Omega)$, the influence of the boundary terms is dropped and we obtain, as in the case of Dirichlet boundary conditions, $-\Delta u = f$ in the sense of distributions. Since $\nabla u \in (L^2(\Omega))^d$ and $f \in L^2(\Omega)$, the equation states that $\nabla u \in H_{\mathrm{div}}$, hence the normal trace $\nabla u \cdot n$ is defined in $H^{-\frac{1}{2}}(\partial\Omega)$ and by the Stokes formula, see Proposition 1.26,
$$\int_\Omega \nabla u \cdot \nabla v + \int_\Omega \mathrm{div}(\nabla u)\, v = \langle \nabla u \cdot n, v\rangle_{H^{-\frac{1}{2}},H^{\frac{1}{2}}}. \tag{3}$$
By comparison between (2) and (3), as $f = -\Delta u$, we obtain
$$\nabla u \cdot n = g, \quad \text{on } \partial\Omega$$
in a weak sense.

Exercise 6 *We consider the Poisson problem with non-homogeneous Fourier (or Robin) boundary conditions:*
$$\left\{ \begin{array}{rcll} -\Delta u &=& f & \textit{in } \Omega, \\ \nabla u \cdot n + \alpha u &=& g & \textit{on } \partial\Omega, \end{array} \right.$$
with $\alpha > 0$.

- *Define the variational formulation.*
- *Is the problem well-posed?*

Let us present and briefly discuss more sophisticated models.

2.3 Example 2: General linear elliptic operators

Let $x \mapsto \mathscr{K}(x) \in \mathbb{R}^{d\times d}$ be a bounded measurable mapping. We assume that $\mathscr{K}$ is symmetric positive definite and uniformly coercive:
$$\exists \alpha > 0,\ (\mathscr{K}(x)\xi, \xi) \geq \alpha |\xi|^2,\ \forall \xi \in \mathbb{R}^d, \text{ for a.e. } x \in \Omega.$$
We consider the problem
$$(\mathrm{P}_s^{(4)}) \left\{ \begin{array}{rcll} -\mathrm{div}(\mathscr{K}\nabla u) &=& f & \text{in } \Omega, \\ u &=& g & \text{on } \partial\Omega, \end{array} \right.$$
with $g \in H^{\frac{1}{2}}(\partial\Omega)$. The associated variational problem is
$$(\mathrm{P}_w^{(4)}) \left\{ \begin{array}{l} \text{Find } u \in H_g^1(\Omega) \text{ such that} \\ \displaystyle\int_\Omega (\mathscr{K}\nabla u) \cdot \nabla v) = \int_\Omega f v, \\ \text{for all } v \in H_0^1(\Omega). \end{array} \right.$$

The adaptation to the non-homogeneous case is straightforward. In the Neumann case, the difficulty only relies on the understanding of the boundary condition which should be read as
$$(\mathscr{K}\nabla u) \cdot n = g.$$

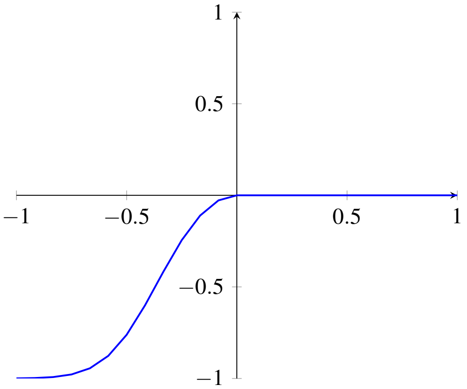


Figure 2. The nonlinear function T

Let us conclude this subsection with an important property of elliptic problems: the weak maximum principle.

Theorem 2.7 (Weak maximum principle) *Let $x \mapsto \mathscr{K}(x) \in \mathbb{R}^{d\times d}$ be a bounded measurable mapping. We assume that $\mathscr{K}$ is symmetric positive definite and uniformly coercive. Let $f \in L^2(\Omega)$ and $g \in H^{\frac{1}{2}}(\partial\Omega)$. Let $u \in H^1(\Omega)$ the unique solution of $\left(\mathrm{P}_w^{(4)}\right)$.*

If f and g are non-negative a.e., then u is non-negative a.e.

Proof of Theorem 2.7. Let $T \in C^1(\mathbb{R};\mathbb{R})$ be a nondecreasing function, with bounded derivative, such that $T(s) = 0$ if and only if $s \geq 0$, see Figure 2.

As $u \in H^1(\Omega)$, by Theorem 1.9 we deduce that $T(u) \in H^1(\Omega)$ and $\gamma_0 T(u) = T(\gamma_0 u) = T(g)$ almost every where. Since $g \geq 0$ and by construction of T, we have that $T(g) = 0$ and thus $T(u) \in H_0^1(\Omega)$. We can therefore take $T(u)$ as a test function in the variational formulation of the problem:

$$\int_\Omega \mathscr{K}\nabla u \cdot \nabla(T(u)) = \int_\Omega f\, T(u).$$

We have $\nabla T(u) = T'(u)\nabla u$ and, then

$$\int_\Omega \underbrace{T'(u)}_{\geq 0}\, \underbrace{(\mathscr{K}\nabla u, \nabla u)}_{\geq 0} = \int_\Omega \underbrace{f}_{\geq 0}\, \underbrace{T(u)}_{\leq 0}.$$

Then each integral is equal to 0 and the integrands vanish almost everywhere. By the coercivity assumption on $\mathscr{K}$, it follows that

$$T'(u)|\nabla u|^2 = 0, \quad \text{almost everywhere,}$$

and thus

$$T'(u)\nabla u = 0, \quad \text{almost everywhere.}$$

By Theorem 1.9, this proves that $\nabla(T(u)) = 0$ and since $T(u) \in H_0^1(\Omega)$, we eventually obtain that $T(u) = 0$. By construction of T, this proves that u is non-negative a.e. on Ω. ■

The weak maximum principle may admit equivalent versions in the discrete framework: in particular the weak maximum principle holds in the case of a $\mathbb{P}^1$ finite element approximation, see Proposition 5.12 whereas it is not valid in the $\mathbb{P}^2$ approximation.

2.4 Example 3: Reaction-diffusion.

We consider the problem

$$\left\{ \begin{array}{rcll} -\Delta u + \eta u & = & f & \text{in } \Omega, \\ \nabla u \cdot n & = & g & \text{on } \partial\Omega, \end{array} \right.$$

with $\eta > 0$. The condition $\eta > 0$ is important because there exists $\eta < 0$ and source terms f for which no solution exists (this is related to the existence of positive eigenvalues of the operator $-\Delta$). When dealing with Dirichlet boundary conditions, there is no difficulty and the variational formulation can be derived in straightforward way. In the case of Neumann boundary conditions, it is different: integrating over Ω,

$$\int_\Omega f = \eta \int_\Omega u - \int_{\partial\Omega} g$$

so that there is no requirement such as the compatibility condition on the data. Moreover, the mean value of u is prescribed and cannot be fixed. The variational formulation is

$$\left\{ \begin{array}{l} \text{Find } u \in H^1(\Omega) \text{ such that} \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v + \eta \int_\Omega u v = \int_\Omega f v, \\ \text{for all } v \in H^1(\Omega). \end{array} \right.$$

Exercise 7 *Prove that the above variational problem admits a unique solution.*

2.5 Example 4: Convection-diffusion.

Let $f \in L^2(\Omega)$ and $b \in (C^1(\bar{\Omega}))^d$. We consider the problem

$$\left\{ \begin{array}{rcll} -\Delta u + b \cdot \nabla u & = & f & \text{in } \Omega, \\ u & = & 0 & \text{on } \partial\Omega, \end{array} \right.$$

The order of derivative for the convection term $b \cdot \nabla u$ is lower than the one of the diffusion term. Thus the functional space is $H_0^1(\Omega)$ and the variational formulation writes

$$\left\{ \begin{array}{l} \text{Find } u \in H_0^1(\Omega) \text{ such that} \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v + \int_\Omega (b \cdot \nabla u)\, v = \int_\Omega f v, \\ \text{for all } v \in H_0^1(\Omega). \end{array} \right.$$

The bilinear form considered here is not symmetric and this is unavoidable due to the advection term. The continuity of $a(\cdot,\cdot)$ is obvious but its coercivity is questionable since

$$a(u,u) = \|\nabla u\|_{L^2}^2 + \int_\Omega (b \cdot \nabla u) u$$

has not necessarily the good sign. More precisely, the second term should be controlled by the first one to ensure the coercivity. For this there are two possible ways to proceed:

- **Method 1:**

  $$\begin{aligned}\left|\int_\Omega (b\cdot\nabla u)u\right| &\le \|b\|_\infty \|\nabla u\|_{L^2}\|u\|_{L^2}\\ &\le C_\Omega \|b\|_\infty \|\nabla u\|_{L^2}^2\end{aligned}$$

  where C_Ω denotes the Poincaré-Friedrichs constant. Thus if $C_\Omega\|b\|_\infty < 1$, the bilinear form is coercive.

- **Method 2**, based upon an integration by parts:

  $$\begin{aligned}-\int_\Omega (b\cdot\nabla u)u &= \int_\Omega \mathrm{div}(b)\frac{u^2}{2}\\ &\le \tfrac{1}{2}\|(\mathrm{div}(b))^+\|_{L^\infty} C_\Omega^2 \|\nabla u\|_{L^2}^2.\end{aligned}$$

  Thus if $\|(\mathrm{div}(b))^+\|_{L^\infty} C_\Omega^2 < 2$ the bilinear form is coercive. Most important cases deal with $\mathrm{div}(b) \le 0$ in which case the condition is obviously satisfied.

If one of the above smallness assumptions on $\|b\|_{L^\infty}$ or $\|\mathrm{div}(b)^+\|_{L^\infty}$ is satisfied, then the variational problem is well-posed.

2.6 Example 5: Linear elasticity.

We consider the PDE which describes the displacement of an elastic solid when submitted to a force field. The domain $\Omega \subset \mathbb{R}^d$ represents the solid at rest. The vector field $u : \Omega \to \mathbb{R}^d$ represents the displacement of a material point under the effect of the force field. Thus a material point at position x at rest is displaced at position $x + u(x)$. We now introduce the Jacobian matrix ∇u and the strain tensor $\mathbb{D}(u) = \frac{1}{2}(\nabla u + (\nabla u)^{\mathrm{t}})$. The Cauchy stress tensor σ in the solid is ruled by a law which takes the form

$$\sigma = \lambda\,\mathrm{Tr}(\mathbb{D}(u))\,\mathrm{Id} + 2\mu\,\mathbb{D}(u) = \lambda\,\mathrm{div}(u)\,\mathrm{Id} + 2\mu\,\mathbb{D}(u),$$

where λ and μ are the so-called Lamé coefficients which quantify the elastic behaviour of the solid. The equilibrium of the solid under the load imposed by an external force f reads

$$-\mathrm{div}(\sigma) = f,$$

i.e.

$$-2\mu\,\mathrm{div}(\mathbb{D}(u)) - \lambda\nabla(\mathrm{div}(u)) = f.$$

The physical boundary conditions that are classically used lead us to consider homogeneous Dirichlet conditions $u = 0$ on a part Γ_{D} of the domain models the clamping of the solid whereas an homogeneous Neuman condition $\sigma\cdot n = 0$ on $\Gamma_{\mathrm{N}} = \partial\Omega\setminus\Gamma_{\mathrm{D}}$ models a no-stress situation on the other part of the boundary. The natural framework is

$$H_{\mathrm{D}} = \{u \in (H^1(\Omega))^d,\ \gamma_0(u) = 0 \text{ on } \Gamma_{\mathrm{D}}\}$$

and the variational problem is defined as

$$\left\{\begin{array}{l}\text{Find } u \in H_{\mathrm{D}} \text{ such that}\\ \displaystyle\int_\Omega 2\mu\mathbb{D}(u) : \mathbb{D}(v) + \int_\Omega \lambda\,\mathrm{div}(u)\,\mathrm{div}(v) = \int_\Omega f\cdot v,\\ \qquad\text{for all } v \in H_{\mathrm{D}}.\end{array}\right.$$

This formulation is well-posed: existence and uniqueness of the solution can be proved by using the Korn inequality to prove the coercivity:

Proposition 2.8 (Korn inequality) *Let Ω be a Lipschitz connected bounded domain of $\mathbb{R}^d$ and assume that $|\Gamma_{\mathrm{D}}| > 0$. The general case, is more intricate.*

There exists a constant $C > 0$ such that

$$\forall u \in H_{\mathrm{D}},\quad \|\nabla u\|_{L^2} \le C\|\mathbb{D}(u)\|_{L^2}.$$

Proof of Proposition 2.8. Let us prove the inequality in the case $\Gamma_{\mathrm{D}} = \partial\Omega$, i.e. $H_{\mathrm{D}} = H_0^1(\Omega)$. We have

$$\int_\Omega |\mathbb{D}(u)|^2 = \int_\Omega \mathbb{D}(u) : \nabla u = \frac{1}{2}\int_\Omega |\nabla u|^2 + \frac{1}{2}\int_\Omega (\nabla u)^{\mathrm{t}} : \nabla u.$$

The result is proved if the last term is non-negative. We write

$$\int_\Omega (\nabla u)^{\mathrm{t}} : \nabla u = \int_\Omega \sum_{i,j} \partial_i u_j \partial_j u_i$$

then we integrate by parts using the boundary conditions

$$\begin{aligned}\textstyle\int_\Omega (\nabla u)^{\mathrm{t}} : \nabla u &= -\textstyle\int_\Omega \sum_{i,j} \partial_{ij}^2 u_j\, u_i\\ &= -\textstyle\int_\Omega \nabla(\mathrm{div}(u))\cdot u\\ &= \textstyle\int_\Omega (\mathrm{div}(u))^2 \ge 0.\end{aligned}$$

■

The Korn inequality provides the coercivity of the bilinear form in the variational formulation of the elasticity problem.

3. Weak formulation of saddle-point problems

We aim at targetting mathematical problems that do not fall into the scope of the Lax-Milgram theorem. The main result will provide necessary and sufficient conditions for the solvability of a class of variational problems.

Let us recall the Banach theorem:

Theorem 3.1 (Banach) *Let E and F be Banach spaces and $T : E \to F$ a continuous linear operator. If T is bijective, then T^{-1} is continuous.*

The proof of the theorem is a consequence of the theorem of the open mapping, which follows from Baire's lemma. It admits a corollary:

Corollary 3.2 (Banach closed range theorem) *Let E and F be Banach spaces and $T : E \to F$ a continuous linear operator. The following are equivalent:*

1. *There exists $\alpha > 0$ such that*
$$\forall x \in E, \quad \|x\|_E \le \alpha \|T(x)\|_F.$$
2. *T is injective and its image is closed.*

Proof of Corollary 3.2. *Assume that T is injective and its image is closed.* By assumption, $T(E)$ is closed, so that $T(E)$ is a Banach space and T is a bijective continuous linear operator from E onto $T(E)$: it is an isomorphism. Thus there exists $\alpha > 0$ such that $\|T^{-1}(y)\|_E \le \alpha \|y\|_F$, for all $y \in T(E)$, i. e. $\|x\|_E \le \alpha \|T(x)\|_F$, for all $x \in E$. *Conversely, assume that there exists $\alpha > 0$ such that*

$$\forall x \in E, \quad \|x\|_E \le \alpha \|T(x)\|_F. \tag{4}$$

From this inequality T is obviously injective. Let us prove that $T(E)$ is closed. Let $\big(T(x_n)\big)_n$ be a sequence in $T(E)$ which converges to some $y \in F$; $(T(x_n))_n$ is a Cauchy sequencein F . Inequality (4) and the linearity of T imply that $(x_n)_n$ is a Cauchy sequence in E which is complete. Hence $(x_n)_n$ converges to some $x \in E$. By continuity of T, we have $y = T(x)$ and in particular $y \in T(E)$. Thus $T(E)$ is closed. ■

3.1 Banach-Nečas-Babuška theorem

In some situations that we will illustrate later, it is necessary to deal with variational formulations for which the unknown u and the test function v do not belong to the same space. In this case, the Lax-Milgram theorem is inoperant. Moreover, even if u and v belong to the same space, the Lax-Milgram theorem only deals with coercive bilinear forms and do not say anything on the non-coercive case.

The following result is a generalisation of the Lax-Milgram theorem that gives necessary and suffisant well-posedness conditions in such cases.

Theorem 3.3 (Banach-Nečas-Babuška) *Let V and W be Hilbert spaces and $\tilde{a}(\cdot,\cdot)$ a bilinear continuous form on $V \times W$. Then the following properties are equivalent:*

- *For any continuous linear form L on W, there exists a unique $u \in V$ such that*
$$\tilde{a}(u,w) = L(w), \quad \forall w \in W.$$
- *The following conditions are satisfied:*
$$\exists \alpha > 0, \quad \inf_{v \in V} \left(\sup_{w \in W} \frac{\tilde{a}(v,w)}{\|v\|_V \|w\|_W} \right) \ge \alpha, \tag{5}$$
$$\Big(\forall v \in V, \ \tilde{a}(v,w) = 0 \Big) \ \Rightarrow \ w = 0. \tag{6}$$

If one of the properties is satisfied, the unique solution u satisfies

$$\|u\|_V \le \frac{\|L\|_{W'}}{\alpha}.$$

Proof of Theorem 3.3. Since $\tilde{a}(\cdot,\cdot)$ is linear and continuous with respect to the variable w, the Riesz representation theorem gives the existence of an operator $\tilde{A} : V \to W'$ such that

$$\big\langle \tilde{A}v, w \big\rangle_{W',W} = \tilde{a}(v,w), \quad \forall w \in W.$$

As $\tilde{a}(\cdot,\cdot)$ is a continuous bilinear form, $\tilde{A}$ is itself linear and continuous and its norm is controlled by $\|\tilde{a}\|$. Let us prove the equivalence.

- Assume that for any continuous linear form L on W, there exists a unique $u \in V$ such that
$$\tilde{a}(u,w) = L(w), \quad \forall w \in W.$$
i.e. with the notation above
$$\forall L \in W', \quad \exists!\, u \in V, \quad \tilde{A}u = L.$$
Then $\tilde{A}$ is bijective. By Banach's theorem (see Theorem 3.1), $\tilde{A}^{-1}$ is a continuous operator. Thus there exists $\alpha^{-1} > 0$ such that for all $L \in W'$ the unique $u \in V$ such that $\tilde{a}(u,\cdot) = L$ satisfies
$$\|u\|_V \le \alpha^{-1} \|L\|_{W'}.$$
Let us consider $v \in V$. We note $L(w) = \tilde{a}(v,w)$ so that v is the unique solution of $\tilde{a}(v,\cdot) = L$. Hence
$$\|v\|_V \le \alpha^{-1} \sup_{w \in W} \frac{\tilde{a}(v,w)}{\|w\|_W}.$$
This inequality holds for any $v \in V$ and therefore Eq. (5) holds.

  Assume now that, for some $w \in W$, we have $\tilde{a}(v,w) = 0$ for all $v \in V$. We introduce the continuous linear form L on W defined as $L(w') = (w,w')_W$. Applying the assumption to $v = \tilde{A}^{-1}L$, we get
$$0 = \tilde{a}(\tilde{A}^{-1}L, w) = \big\langle \tilde{A}\tilde{A}^{-1}L, w \big\rangle_{W',W} = L(w) = \|w\|_W^2,$$
hence $w = 0$. This proves Eq. (6).

- Conversely, assume that Eqs. (5) and (6) hold.

  Let us show that Eq. (6) implies that $\mathrm{Im}(\tilde{A})$ is dense in W'. This is a consequence of the Hahn-Banach theorem, see Theorem 1.25, which provides a characterization of the density of a subspace of a Banach space: let us consider a continuous linear form $\phi \in (W')'$ which is zero on $\mathrm{Im}(\tilde{A})$. As W is a Hilbert space, it is reflexive and ϕ is represented by an element $w \in W$ as $\phi(L) = \langle L, w \rangle_{W',W}$. The assumption implies that, for all $v \in V$, we have

$$0 = \phi(\tilde{A}v) = \langle \tilde{A}v, w \rangle_{W',W} = \tilde{a}(v,w) = 0.$$

By Eq. (6), we obtain $w = 0$ and thus $\phi = 0$ which shows that $\mathrm{Im}(\tilde{A})$ is dense in W'. Now let us deal with Eq. (5), which can be written as

$$\|\tilde{A}v\|_{W'} \geq \alpha \|v\|_V, \quad \forall v \in V,$$

which implies in particular, see Corollary 3.2, that $\tilde{A}$ is injective and that $\mathrm{Im}(\tilde{A})$ is closed. So far the following properties have been proved: 1) $\mathrm{Im}(\tilde{A})$ is dense in W' and 2) $\mathrm{Im}(\tilde{A})$ is closed. Thus $\mathrm{Im}(\tilde{A}) = W'$: $\tilde{A}$ is surjective. As a consequence, since $\tilde{A}$ is also injective, the variational problem is well-posed.

■

Remark 3.4 *Assume that $V = W$ and that $\tilde{a}(\cdot,\cdot)$ is α-coercive. We first note that, for any $v \in V$,*

$$\sup_{w \in V} \frac{\tilde{a}(v,w)}{\|w\|_V} \geq \frac{c(v,v)}{\|v\|_V} \geq \alpha \|v\|_V,$$

so that (5) *holds. Similarly we have*

$$\sup_{v \in V} \frac{\tilde{a}(v,w)}{\|v\|_V} \geq \frac{\tilde{a}(w,w)}{\|w\|_V} \geq \alpha \|w\|_V,$$

so that (6) *holds.*

Therefore, the previous theorem is indeed a generalization of the Lax-Milgram theorem.

Remark 3.5 *In the case of the finite dimensional framework, i.e. $V = \mathbb{R}^n$, $W = \mathbb{R}^p$ and $\tilde{a}(v,w) = (\tilde{A}v, w)$ with $\tilde{A} \in \mathscr{M}_{p \times n}(\mathbb{R})$, Theorem 3.3 can be read as follows. Eq.* (5) *should be read as*

$$\forall v \in \mathbb{R}^n, \quad \|\tilde{A}v\| \geq \alpha \|v\|,$$

which means that $\tilde{A}$ is injective. Eq. (6) *states that $\tilde{A}^{\mathrm{t}}$ is injective, i.e. $\tilde{A}$ is surjective. Clearly the two conditions are equivalent to the solvability of the linear system.*

3.2 Saddle-point problems

Let X and M be Hilbert spaces, $a(\cdot,\cdot)$ a continuous bilinear form on $X \times X$, b a continuous linear form on $X \times M$. For all $L \in X'$, for all $G \in M'$, we aim at solving the variational problem

$$(\mathrm{Q}) \left\{ \begin{array}{l} \text{Find } (u,p) \in X \times M \text{ such that} \\ \quad a(u,v) + b(v,p) \;=\; L(v), \\ \qquad\qquad\;\; b(u,q) \;=\; G(q), \\ \text{for all } (v,q) \in X \times M. \end{array} \right.$$

The second equation is often referred as the *constraint* and the unknown scalar field p is the *Lagrange multiplier* associated to the constraint. This comes from the fact that in the symmetric case the variational problem is equivalent to the Euler-Lagrange problem which consists in minimizing the functional

$$\frac{1}{2} a(v,v) - L(v)$$

on the constrained space

$$Z = \{v \in X,\ b(v,q) = G(q),\ \forall q \in M\}.$$

We can also see the solution of this problem as a saddle-point (u,p) of the functional L defined by

$$L(v,q) = \frac{1}{2} a(v,v) + b(v,p) - L(v) - G(q),$$

which is called *the Lagrangian of the problem.*

Remark 3.6 *The variational formulation can be written in terms of operators. Defining $A : X \to X'$ as*

$$\langle Au, v \rangle_{X',X} = a(u,v), \quad \forall v \in X$$

and $B : X \to M'$ as

$$\langle Bv, p \rangle_{M',M} = b(v,p), \quad \forall p \in M$$

the formulation is equivalent to

$$\left\{ \begin{array}{rcl} \multicolumn{3}{l}{\textit{Find } (u,p) \in X \times M \textit{ such that}} \\ Au + B'p &=& L, \\ Bu &=& G, \end{array} \right.$$

where $B' : M \to X'$ is the adjoint operator of B, the bidual of M being identified to M itself.

Remark 3.7 *The variational problem can take the general form:*

$$\left\{ \begin{array}{l} \textit{Find } (u,p) \in X \times M \textit{ such that} \\ a(u,v) + b(v,p) - b(u,q) = L(v) - G(q), \\ \textit{for all } (v,q) \in X \times M. \end{array} \right.$$

Defining $V = X \times M$ and $\tilde{a}((u,p),(v,q)) = a(u,v) + b(v,p) - b(u,q)$, the well-posedness of the variational problem depends on conditions satisfied by $\tilde{a}$, see Banach-Nečas-Babuška theorem (i.e. Theorem 3.3). Note also that $\tilde{a}((u,p),(u,p)) = a(u,u)$ so that $\tilde{a}$ has no chance to be coercive on $X \times M$ as this term does not include any control on the Lagrange multiplier p.

Theorem 3.8 (Saddle-point problem) *If $a(\cdot,\cdot)$ is coercive on X, the problem* (Q) *is well-posed if and only if the so-called* inf-sup condition *is satisfied:*

$$\exists \beta > 0, \quad \inf_{p \in M} \left(\sup_{v \in X} \frac{b(v,p)}{\|v\|_X \|p\|_M} \right) \geq \beta. \tag{7}$$

Proof of Theorem 3.8. *Assume that the problem is well-posed.* By the Banach-Nečas-Babuška theorem, see Theorem 3.3, this implies that the solution (u,p) continuously depends on the data L and G. Thus there exists a constant $\alpha^{-1} > 0$ such that

$$\|u\|_X + \|p\|_M \le \alpha^{-1}(\|L\|_{X'} + \|G\|_{M'}).$$

Let $\tilde{p} \in M$ and let us consider $L = 0$ and G defined by $G(q) = (\tilde{p}, q)_M$ (note that $\|G\|_{M'} = \|\tilde{p}\|_M$). In particular we obtain that there exists $u \in X$ such that

$$b(u,q) = (\tilde{p}, q)_M, \quad \forall q \in M,$$

and

$$\|u\|_X \le C\|\tilde{p}\|_M.$$

Then

$$\|\tilde{p}\|_M^2 = (\tilde{p},\tilde{p})_M = b(u,\tilde{p}) \le \alpha^{-1}\frac{b(u,\tilde{p})}{\|u\|_X}\|\tilde{p}\|_M,$$

which yields

$$\alpha\|\tilde{p}\|_M \le \sup_{v\in X}\frac{b(v,\tilde{p})}{\|v\|_X}.$$

The last inequality is exactly the inf-sup condition given by Eq. (7).

Conversely, assume that the inf-sup condition is satisfied and let us prove that the bilinear form $\tilde{a}$ satisfies the conditions given by Eqs. (5) and (6) in order to apply Banach-Nečas-Babuška Theorem (see Theorem 3.3).

- Let us prove that Eq. (6) is satisfied. Assume that $(v,q) \in X \times M$ is such that

  $$\tilde{a}((u,p),(v,q)) = 0, \quad \forall (u,p) \in X \times M.$$

  We aim at proving that $(v,q) = (0,0)$. We take $(u,p) = (v,q)$ in the above equality which gives: $a(v,v) = 0$ hence $v = 0$ because a is coercive on X. Thus the previous equality is reduced to $b(u,q) = 0$ for all $u \in X$. The inf-sup condition given by Eq. (7) implies that $q = 0$.

- Let us prove that Eq. (5) is satisfied. Let $(u,p) \in X \times M$. By the inf-sup condition , we have

  $$\beta\|p\|_M \le \sup_{v\in X}\frac{b(v,p)}{\|v\|_X}.$$

  There exists $\tilde{u} \in X$ such that

  $$\frac{b(\tilde{u},p)}{\|\tilde{u}\|_X} \ge \beta\|p\|_M.$$

  We can choose the norm of $\tilde{u}$ and take for instance

  $$\|\tilde{u}\|_X = \|p\|_M.$$

  We now define $(v,q) = (u + \gamma\tilde{u}, p)$ and compute the term $\tilde{a}((u,p),(v,q))$

  $$\begin{aligned}&\tilde{a}((u,p),(v,q))\\ &\quad= a(u,u) + \gamma a(u,\tilde{u}) + \gamma b(\tilde{u},p)\\ &\quad\ge \alpha\|u\|_X^2 - \gamma\|a\|\|u\|_X\|\tilde{u}\|_X + \gamma\beta\|\tilde{u}\|_X^2\\ &\quad\ge \tfrac{\alpha}{2}\|u\|_X^2 - \tfrac{\gamma^2\|a\|^2}{2\alpha}\|\tilde{u}\|_X^2 + \gamma\beta\|\tilde{u}\|_X^2,\end{aligned}$$

  where we have used the Young's inequality for the last step[2]. Then choosing $\gamma < \frac{\alpha\beta}{\|a\|^2}$ we obtain

  $$\begin{aligned}\tilde{a}((u,p),(v,q)) &\ge \tfrac{\alpha}{2}\|u\|_X^2 + \tfrac{\gamma\beta}{2}\|\tilde{u}\|_X^2\\ &\ge \tfrac{\alpha}{2}\|u\|_X^2 + \tfrac{\gamma\beta}{2}\|p\|_M^2\\ &\ge \delta\|(u,p)\|_{X\times M}^2\end{aligned}$$

  where δ only depends on γ, β, α and $\|a\|$. Moreover we have

  $$\begin{aligned}\|(v,q)\|_{X\times M} &= \|v\|_X + \|q\|_M\\ &\le \|u\|_X + \gamma\|\tilde{u}\|_X + \|p\|_M\\ &\le \|u\|_X + (1+\gamma)\|p\|_M\end{aligned}$$

  and thus

  $$\|(v,q)\|_{X\times M} \le (1+\gamma)\|(u,p)\|_{X\times M}.$$

  We finally obtain

  $$\frac{\tilde{a}((u,p),(v,q))}{\|(v,q)\|_{X\times M}} \ge \frac{\delta}{1+\gamma}\|(u,p)\|_{X\times M}$$

  which proves that Eq. (5) is satisfied.

■

Remark 3.10 *An alternate proof of Theorem 3.3 is based on a so-called artificial compressibility method. Assume that $a(\cdot,\cdot)$ is coercive and that the inf-sup condition is satisfied. We introduce the approximate problem:*

$$\begin{cases}\textit{Find } (u_\varepsilon, p_\varepsilon) \in X \times M \textit{ such that}\\ a(u_\varepsilon, v)_X + b(v, p_\varepsilon) - b(u_\varepsilon, q) + \varepsilon(p_\varepsilon, q)_M = L(v) - G(q),\\ \textit{for all } (v,q) \in X \times M.\end{cases}$$

The bilinear form $\tilde{a}_\varepsilon$ which defines this problem is obviously continuous and satisfies:

$$\tilde{a}_\varepsilon((v,p),(v,p)) = a(v,v) + \varepsilon(p,p)_M, \ \forall v \in X, \forall p \in M,$$

[2] The Young's inequality reads:

Proposition 3.9 (Young's inequality) *For all $a > 0$, $b > 0$,*

$$ab \le \frac{a^2}{2\varepsilon} + \frac{\varepsilon b^2}{2}$$

for any $\varepsilon > 0$.

As a consequence:

$$\gamma\|a\|\|u\|_X\|\tilde{u}\|_X \le \frac{\gamma^2\|a\|^2}{2\alpha}\|\tilde{u}\|_X^2 + \frac{\alpha}{2}\|u\|_X^2,$$

by identifying $a := \gamma\|a\|\|\tilde{u}\|_X$, $b := \|u\|_X$ and $\varepsilon = \alpha$.

so that it is coercive. By the Lax-Milgram theorem the approximate problem admits a unique solution $(u_\varepsilon, p_\varepsilon)$. *Note that the problem can be defined by means of operators as*

$$\left\{\begin{array}{lcl} \text{Find } u_\varepsilon \in X \text{ and } p_\varepsilon \in M \text{ such that} \\ Au_\varepsilon + B' p_\varepsilon & = & L, \\ Bu_\varepsilon - \varepsilon p_\varepsilon & = & G, \end{array}\right.$$

We use first the inf-sup condition (7)*, then the first equation of the system, and finally the continuity of* a *and* L*, to get*

$$\begin{array}{rcl} \|p_\varepsilon\|_M & \leq & \frac{1}{\beta} \sup_{v\in X} \frac{b(v,p_\varepsilon)}{\|v\|_X} \\ & = & \frac{1}{\beta} \sup_{v\in X} \frac{-a(u_\varepsilon,v)+L(v)}{\|v\|_X} \\ & \leq & \frac{\|a\|}{\beta}\|u_\varepsilon\|_X + \frac{\|L\|_{X'}}{\beta}. \end{array}$$

This gives us an estimate of p_ε *in terms of* u_ε*. We now take* $v = u_\varepsilon$ *and* $q = p_\varepsilon$ *in the approximate problem:*

$$a(u_\varepsilon, u_\varepsilon) + \varepsilon \|p_\varepsilon\|_M^2 = L(u_\varepsilon) - G(p_\varepsilon).$$

Using the coercivity of $a(\cdot,\cdot)$*, we have*

$$\begin{array}{rcl} \alpha\|u_\varepsilon\|_X^2 & \leq & \|L\|_{X'}\|u_\varepsilon\|_X + \|G\|_{M'}\|p_\varepsilon\|_M \\ & \leq & \left(\|L\|_{X'} + \frac{\|a\|\|G\|_{M'}}{\beta}\right)\|u_\varepsilon\|_X + \frac{\|L\|_{X'}^2}{\beta}. \end{array}$$

With the Young inequality, we conclude that

$$\frac{\alpha}{2}\|u_\varepsilon\|_X^2 \leq \frac{1}{2\alpha}\left(\|L\|_{X'} + \frac{\|a\|\|G\|_{M'}}{\beta}\right)^2 + \frac{\|L\|_{X'}^2}{\beta}.$$

Thus $\{u_\varepsilon\}$ *and* $\{p_\varepsilon\}$ *are bounded in* X *and* M *respectively. Up to a subsequence, they weakly converge to a some* $u \in X$ *and* $p \in M$ *respectively. Obviously we may pass to the limit in the approximate problem so that* (u,p) *is a solution of* (Q)*. Uniqueness of the solution follows by taking* (u,p) *as a test function in the* homogeneous *abstract problem (i.e. the one with* $L = 0$ *and* $G = 0$*). This shows that* $u = 0$*. Then the inf-sup condition implies that* $p = 0$*.*

In fact we can prove that the whole sequence $\{(u_\varepsilon, p_\varepsilon)\}$ *strongly converges to* (u,p)*. More precisely, we can prove that there exists* $C > 0$ *which only depends on the data such that*

$$\|u - u_\varepsilon\|_X + \|p_\varepsilon - p\|_M \leq C\varepsilon, \quad \forall \varepsilon > 0.$$

The difference $(u_\varepsilon - u, p_\varepsilon - p) \in X \times M$ *satisfies*

$$a(u_\varepsilon - u, v) + b(v, p_\varepsilon - p) - b(u_\varepsilon - u, q) + \varepsilon(p_\varepsilon, q)_M = 0,$$

for all $v \in X$ *and for all* $q \in M$*. Taking* $q = 0$ *and using the inf-sup condition,*

$$\begin{array}{rcl} \beta\|p_\varepsilon - p\|_M & \leq & \sup_{v\in X} \frac{b(v,p_\varepsilon - p)}{\|v\|_X} \\ & \leq & \sup_{v\in X} \frac{a(u_\varepsilon - u,v)}{\|v\|_X} \\ & \leq & \|a\|\|u_\varepsilon - u\|_X. \end{array}$$

Then we take $v = u_\varepsilon - u$ *and* $q = p_\varepsilon - p$ *in the above formulation and we use the coercivity of* $a(\cdot,\cdot)$ *to get*

$$\alpha\|u_\varepsilon - u\|_X^2 + \varepsilon(p_\varepsilon - p, p_\varepsilon - p)_M = -\varepsilon(p, p_\varepsilon - p)_M,$$

hence

$$\|u - u_\varepsilon\|_X \leq \varepsilon\frac{\|a\|}{\alpha\beta}\|p\|_M, \quad \|p_\varepsilon - p\|_M \leq \varepsilon\frac{\|a\|^2}{\alpha\beta^2}\|p\|_M,$$

for all $\varepsilon > 0$*.*

Corollary 3.11 (Right inversibility of B**)** *Assume that* b *satisfies the inf-sup condition* (7)*. Then, there exists a continuous linear operator* $\Phi : M' \to X$ *with* $\|\Phi\| \leq \beta^{-1}$ *such that*

$$B \circ \Phi = \mathrm{Id}_{M'}.$$

In particular for any $G \in M'$ *there exists* $u_G = \Phi(G) \in X$ *such that*

$$b(u_G, q) = G(q), \quad \forall q \in M,$$

$$\|u_G\|_X \leq \frac{\|G\|_{M'}}{\beta}.$$

Such an element u_G *is not necessarily unique.*

Remark 3.12 *Actually the existence of such an operator is equivalent to the inf-sup condition.*

Proof of Corollary 3.11. We apply Theorem 3.8 to the abstract problem

$$\left\{\begin{array}{lcl} \text{Find } (u,p) \in X \times M \text{ such that} \\ (u,v)_X + b(v,p) & = & 0, \\ b(u,q) & = & G(q), \\ \text{for all } (v,q) \in X \times M. \end{array}\right.$$

As $(\cdot,\cdot)_X$ is obviously coercive on X, the theorem applies and we have a unique solution $(u_G, p_G) \in X \times M$ satisfying the above problem. Then if we take $v = u_G$ in the first equation, we obtain

$$\|u_G\|_X^2 = -G(p_G).$$

Besides the inf-sup condition gives

$$\beta\|p_G\|_M \leq \sup_{v\in X}\left(\frac{b(v,p_G)}{\|v\|_X}\right) = \sup_{v\in X}\left(\frac{(u_G,v)_X}{\|v\|_X}\right) = \|u_G\|_X.$$

Thus we get

$$\|u_G\|_X^2 \leq \|G\|_{M'}\|p_G\|_M \leq \frac{\|G\|_{M'}}{\beta}\|u_G\|_X,$$

which provides the estimate on $\|u_G\|_X$. Moreover by construction, u_G depends linearly on G which gives the existence of the operator Φ. ■

It is now possible to determine *a priori* bounds for the solution of the abstract problem:

Theorem 3.13 (A priori estimates) *Assume that $a(\cdot,\cdot)$ is coercive on X (with α the constant of coercivity). Assume that b satisfies the inf-sup condition* (7)*. Then the unique solution (u,p) of the abstract problem* (Q) *satisfies:*

$$\|u\|_X \le \frac{\|L\|_{X'}}{\alpha} + \frac{1}{\beta}\left(1+\frac{\|a\|}{\alpha}\right)\|G\|_{M'}, \qquad (8)$$

$$\|p\|_M \le \left(1+\frac{\|a\|}{\alpha}\right)\left(\frac{1}{\beta}\|L\|_{X'} + \frac{\|a\|}{\beta^2}\|G\|_{M'}\right) \qquad (9)$$

Proof of Theorem 3.13. Let us use the properties of the right inverse of B that was considered in Corollary 3.11. Let $(u,p)\in X\times M$ be the solution of the saddle-point problem (Q). We take $v = u-\Phi(G)$ as a test function in the first equation:

$$a(u,u-\Phi(G)) + b(u-\Phi(G),p) = L(u-\Phi(G)).$$

Note that, thanks to the second equation of the system $b(u-\Phi(G),p) = b(u,p) - b(\Phi(G),p) = G(p) - G(p) = 0$. Thus we have

$$a(u-\Phi(G),u-\Phi(G)) = L(u-\Phi(G)) - a(\Phi(G),u-\Phi(G)).$$

Using the coercivity of $a(\cdot,\cdot)$ and the estimates on Φ,

$$\alpha\|u-\Phi(G)\|_X \le \|L\|_{X'} + \frac{\|a\|}{\beta}\|G\|_{M'}.$$

This yields

$$\begin{array}{rcl} \|u\|_X & \le & \|u-\Phi(G)\|_X + \|\Phi(G)\|_X \\ & \le & \frac{\|L\|_{X'}}{\alpha} + \frac{1}{\beta}\left(1+\frac{\|a\|}{\alpha}\right)\|G\|_{M'}. \end{array}$$

The estimate on p is obtained with the inf-sup condition:

$$\begin{array}{rcl} \beta\|p\|_M & \le & \sup_{v\in X}\frac{b(v,p)}{\|v\|_X} \\ & \le & \sup_{v\in X}\frac{a(u,v)-L(v)}{\|v\|_X} \\ & \le & \|a\|\|u\|_X + \|L\|_{X'}. \end{array}$$

Inserting the estimate on u into the above inequality allows us to conclude the proof. ■

Let us discuss the consequences of the previous results. In particular, we present some extension of Theorem 3.8 by weakening some assumption and we consider the finite dimensional problem.

A stronger result

Definition 3.14 (Kernel) *The kernel of a bilinear form b is the closed subspace $Z\subset X$ defined as*

$$Z := \mathrm{Ker}(B) = \{v\in X,\ b(v,q) = 0,\ \forall q\in M\}.$$

Corollary 3.15 *Assume that b satisfies the inf-sup condition* (7)*. Let L be a continuous linear form on X such that $L(v) = 0$ for all $v\in Z$. Then there exists a unique $p\in M$ such that*

$$L(v) = b(v,p), \quad \forall v\in X,$$

and

$$\|p\|_M \le \frac{\|L\|_{X'}}{\beta}.$$

Proof of Corollary 3.15. By Theorem 3.8 and assumptions on L, there is a unique $(u,p)\in X\times M$ such that

$$\left\{\begin{array}{rcll} (u,v)_X + b(v,p) & = & L(v), & \forall v\in X, \\ b(u,q) & = & 0, & \forall q\in M. \end{array}\right.$$

The second equation states that $u\in Z$. Thus if we take $v = u$ in the first equation we obtain by using the assumption $L_{|Z} = 0$ that $\|u\|_X^2 = 0$ hence $u = 0$. Thus we have

$$b(v,p) = L(v), \quad \forall v\in X.$$

Moreover by the inf-sup inequality, we have:

$$\beta\|p\|_M \le \sup_{v\in X}\frac{b(v,p)}{\|v\|_X} = \sup_{v\in X}\frac{L(v)}{\|v\|_X} = \|L\|_{X'}.$$

■

Now it is possible to prove a result which is stronger than Theorem 3.8: the existence and uniqueness result still holds if $a(\cdot,\cdot)$ is coercive on the kernel of b.

Theorem 3.16 *Theorem 3.8 still holds if we only assume that $a(\cdot,\cdot)$ is coercive on Z.*

Proof of Theorem 3.16. We use the operator Φ defined in Corollary 3.11 so as to find (u,p) the solution of the saddle-point problem as $u = \tilde{u} + \Phi(G)$ where $\tilde{u}$ belongs to Z, by definition of $\Phi(G)$. Thus we aim at finding $\tilde{u}\in Z$ which satisfies

$$a(\tilde{u},\tilde{v}) = L(\tilde{v}) - a(\Phi(G),\tilde{v}), \quad \forall v\in Z.$$

Since Z is a closed subspace of X, it is a Hilbert space and the coercivity assumption on Z leads to the existence and uniqueness of $\tilde{u}$ by the Lax-Milgram theorem. Now we define

$$\tilde{L}(v) = -a(\Phi(G)+\tilde{u},v) + L(v), \quad \forall v\in X.$$

By definition of $\tilde{u}$, $\tilde{L}$ is zero on Z and, by Corollary 3.15 there exists a unique $p\in M$ such that

$$\tilde{L}(v) = b(v,p), \quad \forall v\in X.$$

This exactly expresses the fact that $(u = \tilde{u}+\Phi(G),p)$ is the solution of our problem. The estimate on u and p readily adapts from the proof of Theorem 3.8. ■

The finite dimensional case

Assume that $X=\mathbb{R}^n$ and $M=\mathbb{R}^p$. Consider a basis $(\phi_i)_{i=1,\dots,n}$ (resp. $(\psi_i)_{i=1,\dots,p}$) of X (resp. M). Any element $u\in X$ and $p\in M$ can be decomposed on these bases as follows

$$u=\sum_{i=1}^{n}U_i\phi_i, \qquad p=\sum_{j=1}^{p}P_i\psi_i.$$

Denoting by U the vector $(U_1,\dots,U_n)^{\mathrm{t}}\in\mathbb{R}^n$ and by P the vector $(P_1,\dots,P_p)^{\mathrm{t}}\in\mathbb{R}^p$, we will use this type of notation when it is necessary. Although the norms are equivalent in a finite dimensional space, it is useful to introduce the following specific ones. For any $V\in\mathbb{R}^n$,

$$\|V\|_{\mathbb{R}^n}:=\left(\sum_{i=1}^{n}V_i^2\right)^{\frac{1}{2}}, \qquad \|V\|_X:=\|\sum_{i=1}^{n}V_i\phi_i\|_X.$$

A similar notation can be introduced for M. Note that the basis $(\phi_i)_{i=1,\dots,n}$ may differ from the canonical one (in particular it is not necessarily orthonormal) so that the norms $\|\cdot\|_X$ and $\|\cdot\|_{\mathbb{R}^n}$ may differ (although they are equivalent).

Define $F=(\langle L,\phi_1\rangle_{X',X},\dots,\langle L,\phi_n\rangle_{X',X})^{\mathrm{t}}\in\mathbb{R}^n$ and the following matrices

$$A=[a(\phi_j,\phi_i)]_{i,j=1,\dots,n}, \qquad B=[b(\phi_j,\psi_i)]_{i=1,\dots,p,\ j=1,\dots,n}$$

and assuming, for the sake of simplicity, that $G\equiv 0$, the problem consists in studying the linear system

$$\begin{pmatrix} A & B^{\mathrm{t}}\\ B & 0\end{pmatrix}\begin{pmatrix} U\\ P\end{pmatrix}=\begin{pmatrix} F\\ 0\end{pmatrix}.$$

Proposition 3.17 (Inf-sup condition in finite dimension) *The inf-sup condition* (7) *is equivalent to the injectivity of* B^{t} *(hence to the surjectivity of* B*).*

Proof of Proposition 3.17. The inf-sup condition states that there exists $\beta>0$ such that

$$\forall q\in M, \quad \sup_{v\in X}\frac{b(v,q)}{\|v\|_X}\geq\beta\|q\|_M,$$

which after introducing the vectors $V\in\mathbb{R}^n$ and $Q\in\mathbb{R}^p$ representing v and q gives

$$\forall Q\in\mathbb{R}^p, \quad \sup_{V\in\mathbb{R}^n}\frac{(BV,Q)_{\mathbb{R}^p}}{\|V\|_X}\geq\beta\|Q\|_M,$$

Using the equivalence of the norms, we have

$$\forall Q\in\mathbb{R}^p, \quad \sup_{V\in\mathbb{R}^n}\frac{(BV,Q)_{\mathbb{R}^p}}{\|V\|_{\mathbb{R}^n}}\geq\tilde\beta\|Q\|_{\mathbb{R}^p},$$

with $\tilde\beta>0$. Then we obtain

$$\forall Q\in\mathbb{R}^p, \quad \sup_{V\in\mathbb{R}^n}\frac{(B^{\mathrm{t}}Q,V)_{\mathbb{R}^n}}{\|V\|_{\mathbb{R}^n}}\geq\tilde\beta\|Q\|_{\mathbb{R}^p},$$

i.e.

$$\forall Q\in\mathbb{R}^p, \quad \|B^{\mathrm{t}}Q\|_{\mathbb{R}^n}\geq\tilde\beta\|Q\|_{\mathbb{R}^p}.$$

As a consequence, since all the previous steps are equivalent, we have proved that the following are equivalent:

1. b satisfies the inf-sup condition;
2. $\exists\tilde\beta>0, \quad \forall Q\in\mathbb{R}^p, \quad \|B^{\mathrm{t}}Q\|_{\mathbb{R}^n}\geq\tilde\beta\|Q\|_{\mathbb{R}^p}$.

Let us conclude the proof in two steps:

- The inf-sup condition leads to the inequality that obviously implies that B^{t} is injective.
- Conversely, assume that B^{t} is injective. This implies that $Q\in\mathbb{R}^p\mapsto\|B^{\mathrm{t}}Q\|_{\mathbb{R}^n}$ is a norm on $\mathbb{R}^p$. Since all the norms on a finite dimension space are equivalent, we deduce that there exists $\alpha>0$ such that

  $$\forall Q\in\mathbb{R}^p, \quad \|Q\|_{\mathbb{R}^p}\leq\alpha\|B^{\mathrm{t}}Q\|_{\mathbb{R}^n}.$$

  Thus we have proved that item 2. is satisfied with $\tilde\beta=1/\alpha$, hence item 1. is satisfied. This concludes the proof.

■

Moreover we can prove the following result:

Theorem 3.18 *The following are equivalent:*

1. *A is coercive on $Z=\mathrm{Ker}(B)$,*
2. *$\exists\varepsilon>0$, $A+A^{\mathrm{t}}+\frac{1}{\varepsilon}B^{\mathrm{t}}B$ is s.p.d.*

Proof of Theorem 3.18. Assume that property 1. holds. We proceed by *reductio ad absurdum.* Assume that property 2. is false. Thus for any $\varepsilon>0$ there exists u_ε such that $\|u_\varepsilon\|=1$ and

$$2(Au_\varepsilon,u_\varepsilon)+\frac{1}{\varepsilon}\|Bu_\varepsilon\|^2\leq 0. \tag{10}$$

The sequence $\{u_\varepsilon\}$ converges, up to an extraction, to an element u with $\|u\|=1$. Moreover,

$$\|Bu_\varepsilon\|^2\leq 2\varepsilon\|A\|\|u_\varepsilon\|=2\varepsilon\|A\|\xrightarrow[\varepsilon\to 0]{}0.$$

Thus $Bu_\varepsilon\to 0$, hence $Bu=0$ and $u\in\mathrm{Ker}(B)$. As $a(\cdot,\cdot)$ is coercive on $\mathrm{Ker}(B)$ and $\|u\|=1$, we have $(Au,u)\geq\alpha$. Thus for ε sufficiently small, we have $(Au_\varepsilon,u_\varepsilon)>0$ which is in contradiction with Eq. (10).

Conversely, assume that property 2. holds so that there exists $\alpha>0$ satisfying

$$\forall u, \quad 2(Au,u)+\frac{1}{\varepsilon}\|Bu\|^2\geq\alpha\|u\|^2.$$

As a consequence, for $u\in\mathrm{Ker}(B)$,

$$(Au,u)\geq\alpha\|u\|^2,$$

and property 2. is proved. ■

3.3 Example 1: A very simple linear constraint

Consider a domain Ω and a non empty subdomain $B \subset \Omega$. Let $\alpha > 0$ and $f \in L^2(\Omega)$. Define the functional

$$J(v) := \frac{1}{2}\int_\Omega |\nabla v|^2 + \frac{\alpha}{2}\int_\Omega v^2 - \int_\Omega fv,$$

and the space

$$V = \left\{ v \in H^1(\Omega),\ \int_B v = 0 \right\}.$$

We consider the minimization problem which consists in finding a (unique) minimizer of J over V, i.e.

$$\left\{ \begin{array}{l} \text{Find } u \in V \text{ such that} \\ J(u) = \min\limits_{v\in V} J(v). \end{array} \right.$$

As V is a closed subspace of $H^1(\Omega)$ and J is a strictly convex and coercive functional, then the minimization problem admits a unique solution. But what kind of variational formulation is associated to this minimization problem?

Remark 3.19 *If the space was $H^1(\Omega)$ only, i.e. without the constraint on the mean value over B, then the variational formulation would write:*

$$\left\{ \begin{array}{l} \textit{Find } u \in H^1(\Omega) \textit{ such that} \\ \displaystyle\int_\Omega \nabla u\cdot\nabla v + \alpha\int_\Omega uv = \int_\Omega fv, \\ \textit{for all } v \in H^1(\Omega). \end{array} \right.$$

This is a classical elliptic problem but it does not take into account the constraint $\int_B u = 0$! And the solution of the above (unconstrained) problem does not satisfy the constraint $\int_B u = 0$ in general.

Remark 3.20 *By the Lax-Milgram theorem, a possible variational formulation consists in dealing with the constraint in the functional space:*

$$\left\{ \begin{array}{l} \textit{Find } u \in V \textit{ such that} \\ \displaystyle\int_\Omega \nabla u\cdot\nabla v + \alpha\int_\Omega uv = \int_\Omega fv, \\ \textit{for all } v \in V. \end{array} \right.$$

This is also a classical elliptic problem for which the constraint on $\int_B u$ is taken into account through the definition of the functional space. But the functional space is not so classical and, if we aim at solving this problem with the finite element method, we need to define a finite dimensional subspace of V, denoted V_h, thus building a basis of V_h; it means in particular that each element should satisfy the constraint! In order to avoid such a difficulty, we aim at preserving the natural *space $H^1(\Omega)$ (for which finite dimensional subspaces are well known and* easy to build*) but there is some price to pay: relaxing the constraint in the functional space requires the introduction of a new unknown: a so-called Lagrange multiplier (associated to the constraint).*

The variational framework is

$$(\mathrm{Q}^{(1)}) \left\{ \begin{array}{rcl} \multicolumn{3}{l}{\text{Find } (u,\lambda) \in H^1(\Omega)\times\mathbb{R} \text{ such that}} \\ \displaystyle\int_\Omega \nabla u\cdot\nabla v + \alpha\int_\Omega uv + \lambda\int_B v &=& \displaystyle\int_\Omega fv, \\ \displaystyle\mu\int_B u &=& 0, \\ \multicolumn{3}{l}{\text{for all } (v,\mu) \in H^1(\Omega)\times\mathbb{R}.} \end{array} \right.$$

The above system clearly falls into the scope of the saddle-point problems. Besides, the second equation ensures that the solution (if it exists) satisfies the constraint whereas the Lagrange multiplier λ in the first equation quantifies an external force (with support in B) that is necessary to impose this constraint: indeed we have

$$\lambda\int_B v = \int_\Omega \lambda \mathbf{1}_B v$$

so that the first equation can be read as

$$\int_\Omega \nabla u\cdot\nabla v + \alpha\int_\Omega uv = \int_\Omega (f - \lambda\mathbf{1}_B)\,v.$$

From the mathematical point of view, we may prove that this saddle-point problem is well-posed. The main required properties are obviously satisfied and we only focus to the inf-sup condition: we aim at identifying some $\beta > 0$ such that

$$\sup_{v\in H^1(\Omega)} \frac{\lambda\int_B v}{\|v\|_{H^1}} \geq \beta|\lambda|, \quad \forall\lambda\in\mathbb{R}$$

or, after an obvious simplification (this is not possible in general!),

$$\sup_{v\in H^1(\Omega)} \frac{\int_B v}{\|v\|_{H^1}} \geq \beta.$$

Consider $\tilde{u} \equiv 1$. As an obvious fact, $\tilde{u} \in H^1(\Omega)$ and

$$\frac{\int_B \tilde{u}}{\|\tilde{u}\|_{H^1}} = \frac{|B|}{|\Omega|^{\frac{1}{2}}}.$$

Thus we obtain

$$\sup_{v\in H^1(\Omega)} \frac{\int_B v}{\|v\|_{H^1}} \geq \frac{\int_B \tilde{u}}{\|\tilde{u}\|_{H^1}} = \frac{|B|}{|\Omega|^{\frac{1}{2}}} > 0.$$

The inf-sup condition is satisfied. As a conclusion, the saddle-point problem $(\mathrm{Q}^{(1)})$ is well-posed.

Exercise 8 *Consider a domain Ω and a nonempty subdomain $B \subset \Omega$. Let $\alpha > 0$ and $f \in L^2(\Omega)$. Define*

$$J(v) := \frac{1}{2}\int_\Omega |\nabla v|^2 + \frac{\alpha}{2}\int_\Omega v^2 - \int_\Omega fv,$$

and V is a subspace of $H^1(\Omega)$ to be precised. We consider the minimization problem which consists in finding the minimizer of J over V, i.e. find u such that

$$u = \mathrm{argmin}_{v\in V} J(v).$$

Define saddle-point formulations associated to the following constraints:

1. $V = \left\{ v \in H^1(\Omega),\ v = \fint_B v \text{ on } B \right\}$*;*

2. $V = \left\{ v \in H^1(\Omega),\ v = \fint_{\partial B} v \text{ on } B \right\}$*;*

3. $V = \left\{ v \in H^1(\Omega),\ \fint_{\Omega \setminus B} v = \fint_B v \text{ on } B \right\}$*.*

3.4 Example 2: Porous medium

We consider the same example as in section 2.3, namely the general elliptic equation $-\operatorname{div}(\mathscr{K}\nabla u) = f$, also refered to as the Darcy equation, and we propose an alternative formulation of this problem. This formulation is called *mixed formulation* and reads

$$\left\{ \begin{array}{rcl} -\operatorname{div}(\sigma) & = & f, \\ \nabla u - \mathscr{K}^{-1}\sigma & = & 0. \end{array} \right.$$

One possible weak formulation is the following

$$(\mathrm{Q}^{(2)}_{[a]}) \left\{ \begin{array}{l} \text{Find } (\sigma, u) \in (L^2(\Omega))^d \times H^1_0(\Omega) \text{ such that} \\ \displaystyle \int_\Omega \mathscr{K}^{-1}\sigma \cdot \tau - \int_\Omega \nabla u \cdot \tau \quad = \quad 0, \\ \displaystyle \qquad\qquad\qquad \int_\Omega \sigma \cdot \nabla v \quad = \quad \int_\Omega f v, \\ \text{for all } (\tau, v) \in (L^2(\Omega))^d \times H^1_0(\Omega). \end{array} \right.$$

In this formulation, $X = (L^2(\Omega))^d$ and $M = H^1_0(\Omega)$. Thus σ plays the role of the main unknown whereas u is the Lagrange multiplier associated with the constraint $-\operatorname{div}\sigma = f$. Consequently, we introduce

$$a(\sigma, \tau) = \int_\Omega \mathscr{K}^{-1}\sigma \cdot \tau, \quad b(\sigma, u) = -\int_\Omega \sigma \cdot \nabla u.$$

As $\mathscr{K}$ is uniformly bounded and coercive, the bilinear form a is continuous and coercive on X and the bilinear form b is continuous on $X \times M$. the well-posedness depends on the inf-sup condition which writes

$$\sup_{\sigma \in (L^2(\Omega))^d} \frac{\int_\Omega \sigma \cdot \nabla u}{\|\sigma\|_{L^2}} \geq \beta \|u\|_{H^1_0}, \quad \forall u \in H^1_0(\Omega).$$

This inequality holds with $\beta = 1$ by taking $\sigma = \nabla u$: indeed this leads to

$$\frac{\|\nabla u\|^2_{L^2}}{\|u\|_{H^1_0}} = \|u\|_{H^1_0}.$$

Thus the problem admits a unique solution.

Exercise 9 *Build explicitely a right inverse* $\Phi : M' \to X$ *for the operator*

$$\begin{array}{rcccc} B & : & X = (L^2(\Omega))^d & \to & M' = H^{-1}(\Omega) \\ & & \sigma & \mapsto & \operatorname{div}(\sigma). \end{array}$$

If we formally integrate by parts the term $b(\sigma, u)$ in the previous equations, we can write another weak formulation as follows

$$(\mathrm{Q}^{(2)}_{[b]}) \left\{ \begin{array}{l} \text{Find } (\sigma, u) \in H_{\operatorname{div}}(\Omega) \times L^2(\Omega) \text{ such that} \\ \displaystyle \int_\Omega \mathscr{K}^{-1}\sigma \cdot \tau + \int_\Omega u \operatorname{div}(\tau) \quad = \quad 0, \\ \displaystyle \qquad\qquad\qquad -\int_\Omega v \operatorname{div}(\sigma) \quad = \quad \int_\Omega f v, \\ \text{for all } (\tau, v) \in H_{\operatorname{div}}(\Omega) \times L^2(\Omega). \end{array} \right.$$

In this formulation, $X = H_{\operatorname{div}}(\Omega)$ and $M = L^2(\Omega)$. Here also σ plays the role of the main unknown and u is the Lagrange multiplier but the regularity assumed on those unknowns are not the same (σ is more regular and u is less regular than in the previous formulation). Consequently, we introduce:

$$a(\sigma, \tau) = \int_\Omega \mathscr{K}^{-1}\sigma \cdot \tau, \quad b(\sigma, u) = -\int_\Omega u \operatorname{div}(\sigma).$$

The bilinear form a is not coercive on X because the L^2-norm of $\operatorname{div}(\sigma)$ is not controlled. But it is coercive on the kernel of b, denoted Z. Indeed Z is $H_{0,\operatorname{div}}(\Omega) = \{\sigma \in H_{\operatorname{div}}(\Omega),\ \operatorname{div}(\sigma) = 0\}$ and the L^2-norm is equivalent to the $H_{\operatorname{div}}-$norm on Z. Let us prove that the inf-sup condition is satisfied:

$$\sup_{\sigma \in H_{\operatorname{div}}(\Omega)} \frac{\int_\Omega u \operatorname{div}(\sigma)}{\|\sigma\|_{H_{\operatorname{div}}}} \geq \beta \|u\|_{L^2}, \quad \forall u \in L^2(\Omega).$$

For this, for a given $u \in L^2(\Omega)$, let us consider the solution $\phi \in H^1_0(\Omega)$ of $-\Delta\phi = u$ and we define $\sigma = -\nabla\phi \in H_{\operatorname{div}}(\Omega)$. We have

$$\|\sigma\|^2_{H_{\operatorname{div}}} = \|\nabla\phi\|^2_{L^2} + \|u\|^2_{L^2}.$$

We test the equation for ϕ against u and, by the Poincaré inequality, we find

$$\int_\Omega |\nabla\phi|^2 = \int_\Omega u\phi \leq \|u\|_{L^2}\|\phi\|_{L^2} \leq C\|u\|_{L^2}\|\nabla\phi\|_{L^2}$$

so that $\|\sigma\|_{H_{\operatorname{div}}} \leq \sqrt{1 + C^2}\|u\|_{L^2}$ and we get

$$\frac{\int_\Omega u \operatorname{div}(\sigma)}{\|\sigma\|_{H_{\operatorname{div}}}} \geq \frac{1}{\sqrt{1 + C^2}}\|u\|_{L^2}$$

which proves the inf-sup condition.

3.5 Example 3: Stokes problem

The strong formulation of the Stokes system in a bounded connected domain of $\mathbb{R}^d$ writes

$$\left\{ \begin{array}{rcl} -\Delta u + \nabla p & = & f, \\ \operatorname{div}(u) & = & 0, \end{array} \right.$$

with homogeneous boundary conditions for the velocity field $u : \to \mathbb{R}^d$ and an additional conditon: the pressure $p : \to \mathbb{R}$ has zero mean value. Actually, this strong formulation comes from the following minimization problem: define the functional

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

and the space

$$V = \{v \in (H_0^1(\Omega))^d,\ \mathrm{div}(v) = 0\},$$

and find the unique minimizer of J over V, i.e. find u such that

$$u = \mathrm{argmin}_{v \in V} J(v).$$

As V is a closed subspace of $(H_0^1(\Omega))^d$ and J is a strictly convex and coercive functional, then the minimization problem admits a unique solution. But what kind of variational formulation is associated to this minimization problem?

Remark 3.21 *If the space were $(H_0^1(\Omega))^d$ only, i.e. without the constraint on the divergence, then the variational formulation would write:*

$$\left\{\begin{array}{l} \textit{Find } u \in (H_0^1(\Omega))^d \textit{ such that} \\ \displaystyle\int_\Omega \nabla u : \nabla v \quad = \quad \int_\Omega f \cdot v, \\ \textit{for all } v \in (H_0^1(\Omega))^d. \end{array}\right.$$

Here we used the notation $A : B = \sum_{i,j} a_{i,j} b_{i,j}$ for the inner product in the set of matrices.

The above equation is a classical elliptic problem but it does not take into account the constraint on the divergence !

Remark 3.22 *By the Lax-Milgram theorem, a possible variational formulation consists in dealing with the constraint in the functional space:*

$$\left\{\begin{array}{l} \textit{Find } u \in V \textit{ such that} \\ \displaystyle\int_\Omega \nabla u : \nabla v = \int_\Omega f \cdot v, \\ \textit{for all } v \in V. \end{array}\right.$$

This is also a classical elliptic problem which does take into account the constraint on the divergence. But the functional space is not so classical and, if we aim at solving this problem with the finite element method, we need to define a finite dimensional subspace of V, denoted V_h, thus building a basis of V_h; it means in particular that each element should satisfy the constraint! In order to avoid such a difficulty, we aim at preserving the natural *space $(H_0^1(\Omega))^d$ (for which finite dimensional subspaces are well known) but there is some price to pay: relaxing the constraint in the functional space requires the introduction of a new unknown: a so-called Lagrange multiplier (associated to the constraint).*

The variational framework is

$$(\mathrm{Q}^{(3)}) \left\{\begin{array}{l} \text{Find } (u,p) \in (H_0^1(\Omega))^d \times L_0^2(\Omega) \text{ such that} \\ \displaystyle\int_\Omega \nabla u : \nabla v - \int_\Omega p\,\mathrm{div}(v) \quad = \quad \int_\Omega f \cdot v, \\ \displaystyle\qquad\qquad\qquad \int_\Omega q\,\mathrm{div}(u) \quad = \quad 0, \\ \text{for all } (v,q) \in (H_0^1(\Omega))^d \times L_0^2(\Omega). \end{array}\right.$$

The bilinear forms are continuous and a is coercive on $X = (H_0^1(\Omega))^d$. The inf-sup condition writes:

$$\sup_{v \in (H_0^1(\Omega))^d} \frac{\int_\Omega p\,\mathrm{div}(v)}{\|v\|_{H^1}} \geq \beta \|p\|_{L^2}, \quad \forall p \in L_0^2(\Omega).$$

In this problem, the pressure p is the Lagrange multiplier associated to the incompressibility equation $\mathrm{div}(u) = 0$. Assuming that the solution is regular and proceeding with integrations by parts, we can see easily that the strong formulation is recovered.

This inf-sup inequality is a consequence of the following result (which is rather difficult to prove in the general case) which is an adaptation of Corollary 3.11:

Lemma 3.23 *For any function $p \in L_0^2(\Omega)$ there exists $v \in (H_0^1(\Omega))^d$ such that $\mathrm{div}(v) = p$. Besides we can choose v such that*

$$\|v\|_{H_0^1} \leq C \|p\|_{L^2},$$

where $C > 0$ only depends on Ω.

A proof of this lemma can found in [10, 3, 14].

4. Basic principles of the Galerkin approximation

4.1 Elliptic problems: Galerkin approximation

Let V be a Hilbert space, $a(\cdot,\cdot)$ a coercive continuous bilinear form and L a continuous linear form. Thus the problem

$$(\mathrm{P}) \left\{\begin{array}{l} \text{Find } u \in V \text{ such that} \\ a(u,v) = L(v), \\ \text{forall } v \in V. \end{array}\right.$$

admits a unique solution thanks to the Lax-Milgram theorem (Theorem 2.1). We aim at studying strategies that allow us to describe the solution by approximation, by means of computations. This pragmatic constraint leads us to target approximation procedures in a finite-dimesional framework. Different methods can be described: we will focus on the Galerkin approximation and the Petrov-Galerkin approximation.

We introduce a finite-dimensional subspace $V_h \subset V$ and looking for an approximate solution in this subspace. Thus we define the approximate problem

$$(\mathrm{P}_h) \left\{\begin{array}{l} \text{Find } u_h \in V_h \text{ such that} \\ a(u_h,v_h) = L(v_h), \\ \text{forall } v_h \in V_h. \end{array}\right.$$

Exercise 10

1. *Prove that the approximate problem (P_h) is equivalent to a finite-dimensional linear system to be determined.*
2. *Prove that (P_h) is well-posed.*

We may notice that:

- Subscript h refers to a *mesh size* or more generally to the quality of the approximation of V.
- Other approximations may be defined: it is possible to replace $a(\cdot,\cdot)$ by a bilinear form $a_h(\cdot,\cdot)$ (for instance using interpolation formula for the approximation of the integrals formula); it is also possible to consider spaces V_h which are not included in V (the approximation is said to be non-conforming) in which case an extension of the continuous bilinear form $a(\cdot,\cdot)$ is required in order to be defined on $V_h \times V_h$. The well-posedness becomes questionable in these cases.
- It is possible to consider two finite-dimensional spaces V_h and W_h and look for a solution $u_h \in V_h$ with test functions in W_h. This is called a *Petrov-Galerkin* approximation (which will be studied in the next subsection). The well-posedness of the approximate problem becomes questionable as well.

The interest of such an approximation only makes sense if we may guarantee that the solution of the approximate problem u_h is indeed an approximation of the solution u. Thus the error has to be estimated in order to ensure that the process is valid. Note that by linearity we have

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

The error $e_h := u - u_h$ is a−orthogonal to V_h. This is the basis which will alow us to perform the analysis of the convergence of the method.

Lemma 4.1 (Error estimate) *Under the approximability property of V_h, i.e.*

$$\forall v \in V, \quad \lim_{h\to 0} d(v, V_h) = 0.$$

then $\{u_h\}$ converges to u in V. Moreover,

$$\|u-u_h\|_V \le \frac{\|a\|}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_V = \frac{\|a\|}{\alpha} d(u, V_h).$$

Proof of Lemma 4.1. Let us proceed in three steps.

- **Step 1. Weak convergence of u_h to u.** Applying the Lax-Milgram theorem for (P_h) not only provides the existence and uniqueness result for its solution but also provides an estimate on u_h, namely

  $$\|u_h\|_V \le \frac{\|a\|}{\alpha} \|L\|_{V'}.$$

  Thus the sequence $\{u_h\}$ is bounded in V and, as a consequence[3], there exists $\bar{u} \in V$ such that, up to a subsequence still denoted h, $\{u_h\}$ weakly converges to $\bar{u}$. Let us prove that $\bar{u} = u$. For this, let us fix some $v \in V$. By assumption, there exists a sequence $\{v_h\}$ of elements in V such that

  - $v_h \in V_h$, for all h,
  - $\lim_{h\to 0} \|v - v_h\|_V = 0$.

  Let us select v_h as a test function in the approximate problem:

  $$a(u_h, v_{h)} = L(v_h).$$

  Using the “strong-weak convergence” argument[4], we pass to the limit in the left-hand side. Passing to the limit in the right-hand side as well, we obtain:

  $$a(\bar{u}, v) = L(v).$$

  This result holds for any $v \in V$ and, as the solution of the initial problem is unique, $\bar{u} = u$. It shows also that the sequence $\{u_h\}$ admits a unique adherent value (or closure point) in V for the weak topology. Therefore the whole sequence $\{u_h\}$ weakly converges to u (by Theorem 4.2).

- **Step 2. Strong convergence of u_h to u.** Let us prove that $\{u_h\}$ strongly converges to u. Let us choose u_h as a test function in the approximate problem. We have

  $$a(u_h, u_h) = L(u_h) \xrightarrow[h\to 0]{} L(u) = a(u,u).$$

[3] We have:

Theorem 4.2 *Let H be a Hilbert space and $\{u_n\}$ a bounded sequence in H. Then*

- *there exists a least one subsequence u_{n_k} which weakly converges;*
- *if, furthermore, the set of weak limits reduces to a single element, then the whole sequence $\{u_n\}_n$ weakly converges to this element .*

[4] The “strong-weak convergence” argument relies on the following proposition:

Proposition 4.3 *Let E be a normed vector space. We consider a sequence $\{x_n\}_n$ of elements in E, and $x \in E$. Assume that $\{x_n\}_n$ weakly converges to x. Then,*

(i) (x_n) is bounded;

(ii) if $\|f_n - f\|_{E'} \to 0$, then $f_n(x_n) \to f(x)$;

In particular, the proof of item (ii) (combined with item (i)) allows us to prove the “strong-weak convergence” argument in our specific case:

$$\begin{array}{rcl} |a(u_h,v_h) - a(\tilde{u},v)| & = & |a(u_h,v_h) - a(u_h,v) + a(u_h,v) - a(\tilde{u},v)| \\ & \le & |a(u_h,v_h) - a(u_h,v)| + |a(u_h,v) - a(\tilde{u},v)| \\ & \le & |(a(u_h, v_h - v)| + |a(u_h - \tilde{u}, v)| \\ & \le & \|a\| \underbrace{\|u_h\|_V}_{(*)} \underbrace{\|v_h - v\|_V}_{(**)} + \underbrace{|a(u_h - \tilde{u}, v)|}_{(***)}. \end{array}$$

By item (i), $(*)$ is bounded. Besides $(**)$ goes to 0 as $\{v_h\}$ (strongly) converges to v and $(***)$ goes to 0 as $\{u_h\}$ weakly converges to $\tilde{u}$ in V. Thus $a(u_h, v_h)$ converges to $a(\tilde{u}, v)$.

Then we get

$$\begin{aligned} a(u_h-u,u_h-u) &= a(u_h,u_h)-a(u_h,u)\\ &\quad -a(u,u_h)+a(u,u), \end{aligned}$$

which goes to 0 as $h\to 0$. By coercivity of a, the last convergence results implies that $\|u_h-u\|_V \xrightarrow[h\to 0]{} 0$.

- **Step 3. Estimate (Céa's lemma).** Let us prove the estimate. Taking v_h-u_h in the orthogonality equation of $e_h := u-u_h$, we have

  $$0 = a(e_h, v_h-u_h) = a(e_h, v_h-u)+a(e_h,e_h).$$

  Then we get

  $$\alpha\|e_h\|_V^2 \leq \|a\|\|e_h\|_V\|v_h-u\|_V.$$

  This inequality holds for any $v_h \in V_h$ so that the proof is concluded.

■

This lemma shows that the approximation error expressed in the $V-$norm e_h is directly related to the distance between V and V_h. Thus it is necessary to build suitable approximation spaces V_h that allow us to estimate the distance between the solution u and V_h.

4.2 Elliptic problems: Petrov-Galerkin approximation

Let us consider a generalized version of the Galerkin approximation: we consider two finite-dimensional subspaces of V, namely V_h and W_h. We assume that those spaces have the same dimension, which is a mandatory condition to expect the underlying linear problem to be well-posed. The approximate problem reads

$$(\tilde{\mathrm{P}}_h)\left\{\begin{array}{l}\text{Find } u_h\in V_h \text{ such that}\\ a(u_h,w_h)=L(w_h),\\ \text{for all } w_h\in W_h.\end{array}\right.$$

The coercivity of $a(\cdot,\cdot)$ on $V\times V$ is not sufficient to ensure existence and uniqueness of a solution for this approximate problem. The well-posedness of the problem is ensured by the two conditions of the Banach-Nečas-Babuška theorem (see Theorem 3.3, page 15). Actually, in finite dimension, the two conditions are equivalent[5] and the approximate problem is well-posed if, and only if, there exists $\alpha_h>0$ such that

$$\inf_{v_h\in V_h}\left(\sup_{w_h\in W_h}\frac{a(v_h,w_h)}{\|v_h\|_V\|w_h\|_V}\right)\geq \alpha_h.$$

In this case, the approximate problem admits a unique solution u_h which satisfies

$$\|u_h\|_V \leq \frac{\|L\|_{V'}}{\alpha_h}.$$

Note that u_h admits an *a priori* bound only if α_h does not tend to 0 when h tends to 0.

Lemma 4.4 (Error estimate) *Assume that*

- $\mathrm{dim}(V_h)=\mathrm{dim}(W_h)$,
- $\forall h>0,\ \exists\alpha_h>0,\ \inf_{v_h\in V_h}\left(\sup_{w_h\in W_h}\frac{a(v_h,w_h)}{\|v_h\|_V\|w_h\|_V}\right)\geq \alpha_h.$

Then we have

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

Assume furthermore that

- *a uniform inf-sup condition holds:*

  $$\exists\underline{\alpha}>0,\quad \forall h>0,\quad \alpha_h\geq\underline{\alpha},$$

- *the approximability property of V_h holds:*

  $$\forall v\in V,\quad \lim_{h\to 0} d(v,V_h)=0.$$

Then $\{u_h\}$ converges to u in V:

$$\lim_{h\to 0}\|u-u_h\|_V=0,$$

Proof of Lemma 4.4. The well-posedness of the approximate problem is ensured by the two conditions of the Banach-Nečas-Babuška theorem (see Theorem 3.3, page 15) and, in finite dimension with $\mathrm{dim}(V_h)=\mathrm{dim}(W_h)$, the two conditions are equivalent. Thus u_h is uniquely determined.

Besides the error $e_h:=u-u_h$ satisfies the orthogonality equation

$$a(e_h,w_h)=0,\quad \forall w_h\in W_h. \tag{11}$$

Let $v_h\in V_h$. We have

$$e_h=u-u_h=(u-v_h)+(v_h-u_h).$$

[5] In the case of the finite dimensional framework, consider a basis $(\phi_i)_{i=1,\dots,n}$ (resp. $(\psi_i)_{i=1,\dots,n}$) of V_h (resp. W_h). Any element $u_h\in V_h$ and $w_h\in W_h$ can be decomposed on these bases as follows: $u_h=\sum_{i=1}^n U_i\phi_i$, $w_h=\sum_{j=1}^n W_i\psi_i$. We denote U the vector $(U_1,\dots,U_n)^{\mathrm{t}}\in\mathbb{R}^n$ and W the vector $(W_1,\dots,W_n)^{\mathrm{t}}\in\mathbb{R}^n$ and we define the vector $F=(\langle L,\psi_1\rangle,\dots,\langle L,\psi_n\rangle)^{\mathrm{t}}$ and the following matrix $A=[a(\phi_j,\psi_i)]_{i,j=1,\dots,n}$. The problem consists in studying the linear system $AU=F$.

- Eq. (5) should be read as (see also Proposition 3.17, page 20, for a similar proof)

  $$\exists\tilde{\alpha}>0,\quad \forall V\in\mathbb{R}^n,\quad \|AV\|_{\mathbb{R}^n}\geq\tilde{\alpha}\|V\|_{\mathbb{R}^n},$$

  which means that A is injective;
- Eq. (6) states that A^{t} is injective, i.e. A is surjective.

Clearly for a $n\times n$ *linear system* the two conditions are equivalent and any of them provides the well-posedness.

Using first the inf-sup condition, then (11) and finally the continuity of a, we get

$$\begin{aligned}\|v_h - u_h\|_V &\leq \frac{1}{\alpha_h} \sup_{w_h \in W_h} \frac{a(v_h - u_h, w_h)}{\|w_h\|_V} \\ &= \frac{1}{\alpha_h} \sup_{w_h \in W_h} \frac{a(v_h - u, w_h)}{\|w_h\|_V} \\ &\leq \frac{\|a\|}{\alpha_h} \|u - v_h\|_V.\end{aligned}$$

Thus we obtain

$$\|e_h\|_V \leq \left(1 + \frac{\|a\|}{\alpha_h}\right) \|u - v_h\|_V.$$

This inequality holds for any $v_h \in V_h$. Finally the uniform control on α_h and the approximability property of V_h allow us to conclude the proof. ■

In the Petrov-Galerkin approximation, different properties are required:

- control of the approximation error associated to V_h,
- control of the constant in the inf-sup condition: subspaces V_h and W_h should not be *too* $a-$orthogonal as h goes to 0.

Example. Let us consider the standard example of the problem $-u'' = f$ with Dirichlet boundary conditions in 1D: $V = H_0^1(]0,1[)$, $a(u,v) = \int_0^1 u'v'$.

- We take a one-dimensional subspace $V_h \subset V$ generated by $p(x) = 1 - |2x - 1|$ (its gradient is equal to 2 on $[0,1/2[$ and -2 on $[1/2,1]$).
- We take a one-dimensional subspace $W_h \subset V$ generated by $q(x) = \sin(4\pi x)$ (its gradient is equal to $4\pi\cos(4\pi x)$).

Observe that the inf-sup condition is *not* satisfied for this problem since $a(v_h, w_h) = 0$ for any $v_h \in V_h, w_h \in W_h$. Indeed, we have

$$\begin{aligned}a(p,q) &= \int_0^1 p'q' \\ &= 8\pi \left(\int_0^{\frac{1}{2}} \cos(4\pi x)\,\mathrm{d}x - \int_{\frac{1}{2}}^1 \cos(4\pi x)\,\mathrm{d}x\right) \\ &= 0.\end{aligned}$$

□

Remark 4.5 *Even if $a(\cdot,\cdot)$ is symmetric the resulting Petrov-Galerkin approximate problem is not necessarily symmetric.*

4.3 Saddle-point problems and Galerkin approximation

Let X and M be two Hilbert spaces. We consider, as in section 3.2, the following problem:

$$\left\{\begin{array}{l}\text{Find } (u,p) \in X \times M \text{ such that} \\ \quad a(u,v) + b(v,p) \;=\; L(v), \\ \qquad\qquad\;\; b(u,q) \;=\; G(q), \\ \text{for any } (v,q) \in X \times M.\end{array}\right.$$

Here L and G are continuous linear forms over X and M respectively. We have seen that this problem is well posed if $a(\cdot,\cdot)$ is coercive on X (or on the kernel of b) and b satisfies the inf-sup condition

$$\inf_{p \in M} \left(\sup_{v \in X} \frac{b(v,p)}{\|v\|_X \|p\|_M}\right) \geq \beta > 0.$$

If we now consider X_h and M_h two finite-dimensional subspaces of X and M respectively, we may define an approximate problem:

$$\left\{\begin{array}{l}\text{Find } (u_h,p_h) \in X_h \times M_h \text{ such that} \\ \quad a(u_h,v_h) + b(v_h,p_h) \;=\; L(v_h), \\ \qquad\qquad\quad\; b(u_h,q_h) \;=\; G(q_h), \\ \text{for any } (v_h,q_h) \in X_h \times M_h.\end{array}\right.$$

Remark 4.6 *Note that a Petrov-Galerkin approach is possible by using test functions v_h and q_h in other spaces than X_h and M_h. The subsequent analysis is more intricate so that we do not want to enter the details here.*

Let us discuss the existence and convergence issue. What can we say about this approximate problem?

- If $a(\cdot,\cdot)$ is coercive on X, then a is coercive on X_h with the same constant of coercivity. In order to guarantee that the approximate problem is well-posed, it suffices for the inf-sup condition to be satisfied:

$$\inf_{p_h \in M_h} \left(\sup_{v_h \in X_h} \frac{b(v_h,p_h)}{\|v_h\|_X \|p_h\|_M}\right) \geq \beta_h > 0. \qquad (12)$$

Because of the finite dimensional framework, a compactness argument allows us to show that this condition is satisfied if, and only if,

$$\left.\begin{array}{ll}1) & p_h \in M_h \\ 2) & \forall v_h \in X_h,\ b(v_h,p_h) = 0\end{array}\right\} \Rightarrow p_h = 0.$$

If we consider the restriction B_h' of operator $B : M \to X'$ as an operator from M_h onto X_h', then the above condition states that B_h' should be injective (see also Proposition 3.17, page 20). In particular, the dimension of M_h should be lower than the dimension of X_h. Thus, *the inf-sup condition* *is not satisfied if M_h is too big with respect to X_h.* If the discrete inf-sup condition is satisfied, there exists a unique solution $(u_h,p_h)\in X_h\times M_h$ of the approximate problem and we have the following bounds:

$$\|u_h\|_X \leq \frac{\|L\|_{X'}}{\alpha}+\frac{1}{\beta_h}\left(1+\frac{\|a\|}{\alpha}\right)\|G\|_{M'},$$

$$\|p_h\|_M \leq \left(1+\frac{\|a\|}{\alpha}\right)\left(\frac{1}{\beta_h}\|L\|_{X'}+\frac{\|a\|}{\beta_h^2}\|G\|_{M'}\right).$$

- If $a(\cdot,\cdot)$ is coercive only on the kernel of b, then nothing guarantees that it is coercive on Z_h the kernel of the restriction $b:X_h\times M_h\to\mathbb{R}$. Indeed, we do not have in general the inclusion $Z_h\subset Z$ beacuse an element $v_h\in Z_h$ is such that $b(v_h,p_h)=0$ for all $p_h\in M_h$ but there is no reason why $b(v_h,p)=0$ for any $p\in M\setminus M_h$. Thus, it is necessary to impose to $a(\cdot,\cdot)$ a coercivity constraint on Z_h as a supplementary condition in order the discrete problem to be well-posed.

Then we have the convergence result (the proof readily adapts from the proof of Lemma 4.1):

Lemma 4.7 (Convergence) *If there exists $\underline{\beta}>0$ such that $\beta_h\geq\underline{\beta}$ for all $h>0$ and if $d(v,X_h)\to 0$ and $d(q,M_h)\to 0$ for all $v\in X$ and for all $q\in M$, then $(u_h,p_h)\to(u,p)$ in $X\times M$ when h goes to 0.*

Let us focus on error estimates. We establish the following result:

Proposition 4.8 *Assume that b satisfies the discrete inf-sup condition* (12) *for the spaces X_h and M_h. For all $h>0$, there exists a continuous linear operator $\Psi_h:=X\to X_h$ such that $\|\Psi_h\|\leq\frac{\|b\|}{\beta_h}$ and*

$$\forall v\in X,\quad \forall q_h\in M_h,\quad b(\Psi_h v,q_h)=b(v,q_h).$$

Proof of Proposition 4.8. Using the inf-sup condition (12) and Corollary 3.11, there exists a continuous right inverse $\Phi_h:M_h'\to X_h$ to the operator

$$\begin{array}{rcccc} B_h & : & X_h & \to & M_h' \\ & & v_h & \mapsto & b(v_h,\cdot) \end{array}$$

with a norm which is bounded by $1/\beta_h$. We define, for all $v\in X$,

$$\Psi_h v=\Phi_h((Bv)_{|M_h}).$$

By construction,

$$\|\Psi_h v\|_X\leq\frac{1}{\beta_h}\|(Bv)_{|M_h}\|_{M_h'}\leq\frac{\|B\|}{\beta_h}\|v\|_X,$$

with $\|B\|=\|b\|$. Moreover, we have

$$\begin{array}{rcl} b(\Psi_h v,q_h) & = & b(\Phi_h((Bv)_{|M_h}),q_h) \\ & = & \left\langle B_h\Phi_h(Bv)_{|M_h},q_h\right\rangle_{M_h',M_h} \\ & = & \left\langle (Bv)_{|M_h},q_h\right\rangle_{M_h',M_h} \\ & = & \left\langle Bv,q_h\right\rangle_{M',M} \\ & = & b(v,q_h). \end{array}$$

The claim is proved. ■

Remark 4.9 *We will see that the existence of such an operator with a norm which does not depend on h is a necessary and sufficient condition for the spaces X_h and M_h to satisfy a uniform inf-sup condition for b (see Lemma 6.1 (Fortin's lemma), page 45).*

Let us now focus on the error estimate.

Lemma 4.10 (Error estimate) *Assume that a is coercive on V and that the discrete inf-sup condition* (12) *is satisfied, then we have*

$$\begin{array}{rcl} \|u-u_h\|_X & \leq & \left(1+\frac{\|a\|}{\alpha}\right)\left(1+\frac{\|b\|}{\beta_h}\right)d(u,X_h) \\ & & \qquad +\frac{\|b\|}{\alpha}d(p,M_h), \end{array}$$

$$\begin{array}{rcl} \|p-p_h\|_M & \leq & \frac{\|a\|}{\beta_h}\left(1+\frac{\|a\|}{\alpha}\right)\left(1+\frac{\|b\|}{\beta_h}\right)d(u,X_h) \\ & & +\left(1+\frac{\|b\|}{\beta_h}+\frac{\|a\|}{\beta_h}\frac{\|b\|}{\alpha}\right)d(p,M_h). \end{array}$$

Proof of Proposition 4.10. We introduce

$$e_h:=u-u_h,\qquad \pi_h=p-p_h.$$

We may observe that:

$$\left\{\begin{array}{rcll} a(e_h,v_h)+b(v_h,\pi_h) & = & 0, & \forall v_h\in X_h, \\ b(e_h,q_h) & = & 0, & \forall q_h\in M_h. \end{array}\right.$$

Let us proceed in two steps:

1. The first equation is rewritten, for any $q_h\in M_h$, as

   $$a(e_h,v_h)+b(v_h,p-q_h)=b(v_h,p_h-q_h),\quad \forall v_h\in X_h.$$

   By using forst the discrete inf-sup condition, then the equation above, and finally the continuity of a and b, we have

   $$\begin{array}{rcl} \|q_h-p_h\|_M & \leq & \frac{1}{\beta_h}\sup_{v_h\in X_h}\frac{b(v_h,p_h-q_h)}{\|v_h\|_X} \\ & = & \frac{1}{\beta_h}\sup_{v_h\in X_h}\frac{a(e_h,v_h)+b(v_h,p-q_h)}{\|v_h\|_X} \\ & \leq & \frac{1}{\beta_h}(\|a\|\|e_h\|_X+\|b\|\|p-q_h\|_M). \end{array}$$

By the triangle inequality, we have, for all $q_h \in M_h$,

$$\|\pi_h\|_M \leq \|p - q_h\|_M + \|q_h - p_h\|_M,$$

hence

$$\|\pi_h\|_M \leq \left(1 + \frac{\|b\|}{\beta_h}\right) d(p, M_h) + \frac{\|a\|}{\beta_h}\|e_h\|_X.$$

We have thus obtained an estimate on the error π_h in function of the error e_h.

2. Let $v_h \in X_h$. We define $r_h = \Psi_h(u - v_h)$, where Ψ_h is the operator defined in (the proof of) Proposition 4.8. By definition, we have that

$$b(u - v_h - r_h, q_h) = 0, \quad \forall q_h \in M_h.$$

As $b(e_h, q_h) = 0$ for all $q_h \in M_h$, we get

$$b(u_h - (v_h + r_h), q_h) = 0, \quad \forall q_h \in M_h. \tag{13}$$

Thus we choose $(v_h + r_h) - u_h$ as a test function in the first equation and we get

$$a(e_h, (v_h + r_h) - u_h) + b((v_h + r_h) - u_h, \pi_h) = 0.$$

Since, by (13), $(v_h + r_h) - u_h$ is $b-$orthogonal to M_h, we may replace π_h in the second term by $p - q_h$, with an arbitrary $q_h \in M_h$:

$$a(e_h, (v_h + r_h) - u_h) + b((v_h + r_h) - u_h, p - q_h) = 0.$$

Then we deal with the quantity $\mathscr{A} := a((v_h + r_h) - u_h, (v_h + r_h) - u_h)$. On the one hand, by coercivity,

$$\alpha\|(v_h + r_h) - u_h\|_X^2 \leq |\mathscr{A}|. \tag{14}$$

On the other hand, as $u_h = u - e_h$,

$$\begin{aligned} \mathscr{A} &= a((v_h + r_h) - u + e_h, (v_h + r_h) - u_h) \\ &= a((v_h + r_h) - u, (v_h + r_h) - u_h) \\ &\quad + a(e_h, (v_h + r_h) - u_h) \\ &= a((v_h + r_h) - u, (v_h + r_h) - u_h) \\ &\quad - b((v_h + r_h) - u_h, p - q_h), \end{aligned}$$

We obtain

$$\begin{aligned} |\mathscr{A}| &\leq \|a\|\|u - (v_h + r_h)\|_X\|(v_h + r_h) - u_h\|_X \\ &\quad + \|b\|\|v_h + r_h - u_h\|_X\|p - q_h\|_M. \end{aligned} \tag{15}$$

Combining Eqs. (14) and (15), we get

$$\begin{aligned} \|(v_h + r_h) - u_h\|_X &\leq \frac{\|a\|}{\alpha}\|u - (v_h + r_h)\|_X \\ &\quad + \frac{\|b\|}{\alpha}\|p - q_h\|_M, \end{aligned}$$

where $q_h \in M_h$ is arbitrary. Hence,

$$\begin{aligned} \|(v_h + r_h) - u_h\|_X &\leq \frac{\|a\|}{\alpha}\|u - (v_h + r_h)\|_X \\ &\quad + \frac{\|b\|}{\alpha} d(p, M_h), \end{aligned}$$

and then

$$\begin{aligned} \|u - u_h\|_X &\leq \|u - (v_h + r_h)\|_X + \|(v_h + r_h) - u_h\|_X \\ &\leq \left(1 + \frac{\|a\|}{\alpha}\right)\|u - (v_h + r_h)\|_X \\ &\quad + \frac{\|b\|}{\alpha} d(p, M_h). \end{aligned}$$

Besides, the definition of r_h and the properties of operator Ψ_h imply

$$\begin{aligned} \|u - (v_h + r_h)\|_X &\leq \|u - v_h\|_X + \|r_h\|_X \\ &\leq \left(1 + \frac{\|b\|}{\beta_h}\right)\|u - v_h\|_X. \end{aligned}$$

This holds for all $v_h \in X_h$ and we thus obtain

$$\begin{aligned} \|u - u_h\|_X &\leq \left(1 + \frac{\|a\|}{\alpha}\right)\left(1 + \frac{\|b\|}{\beta_h}\right) d(u, X_h) \\ &\quad + \frac{\|b\|}{\alpha} d(p, M_h). \end{aligned}$$

The pressure estimate is obtained by using the above estimate.

■

The estimate depends not only on the approximation error for every space X_h and M_h but also on the dependency of the constant β_h in the inf-sup condition with respect to the discretization parameter. Let us now study how approximation spaces (V_h on the one hand, X_h and M_h on the other hand) can be built.

4.4 Approximation spaces

Some principles rule the choice of approximation spaces V_h and a basis $(\phi_i)_i$ of such spaces. Let us note that finding a solution $u_h = \sum_j u_j \phi_j$ of a variational problem consists in solving a linear system with matrix A defined by $a_{ij} = a(\phi_j, \phi_i)$. In practive, it is thus necessary to:

- compute the coefficients,
- build the most simple matrix, i.e. the related system should be "easily" solved.

The definition of the space and the choice of a basis is crucial. Let us discuss some examples:

- If we define V_h as the set of polynomial functions of degree less than N, with its canonical basis defined by the monomials, then the related matrix is likely to be dense. Of course, if we choose a basis of polynomials which are orthogonal for the scalar product defined by a (when a is symmetric, positive definite), then the matrix is diagonal and solving the related linear system is easy. This can possibly done by using a Gram orthonormalisation process.

- An efficient method consists in defining the approximation space as the subspace generated by the eigenfunctions of the operator that defines the PDE. If the solution is regular, we can prove under suitable assumptions that the method is quite efficient: this is the very basis of the so-called *spectral methods.*

  Example. We denote V_N the space generated by the first N eigenfunctions of the Dirichlet-Laplace operator in 1D (i.e. $x \mapsto \sin(k\pi x)$ on $]0,1[$, then we have (Parseval's identity):

  $$\forall u \in H^m(]0,1[), \quad d(u, V_N) \leq \frac{C_k}{N^m} \|u\|_{H^m}.$$

  □

  This method has a major drawback, as we only know the eigenfunctions of the operator in a limited number of cases. In the other cases, we use the eigenfunctions of another operator with the secret hope that it will work fine. Besides, the matrix is dense, which may lead to heavy computational costs.

- In the next section, we will focus on the *finite element method.* The main idea consists in cutting the domain into small pieces or *elements* and defining the approximation space as a set of piecewise regular functions with a particular structure on each element. Piecewise polynomials are classically used in this prospect.

5. Examples of finite element spaces

5.1 Finite elements in 1D: $\mathbb{P}^1$ finite element

Approximation space $\mathbb{P}^1$ in 1D. We consider $\Omega =]0,1[$ et we target the definition of a suitable approximation space for $V = H^1(\Omega)$. For this we *mesh* the domain Ω. In the 1D framework: the domain is divided in segments $[x_i, x_{i+1}]$, $i \in \{0, ..., N\}$ so that $x_i < x_{i+1}$ and $x_0 = 0$ and $x_{N+1} = 1$. We define the mesh size $h := \sup_i |x_{i+1} - x_i|$ and we denote by $K_i := [x_i, x_{i+1}]$ the cells of the mesh.

Let us consider the set of piecewise affine functions:

$$V_h = \{u \in V,\ u_{|K_i} \in \mathbb{P}^1\},$$

where $\mathbb{P}^1$ is the set of polynomials of degree 1. Notice that we consider a *conforming* approximation space since, by construction, $V_h \subset V$.

Lemma 5.1 *The space V_h writes*

$$V_h = \{u \in C^0(\bar{\Omega}),\ u_{|K_i} \in \mathbb{P}^1\},$$

and, moreover, the mapping

$$\Phi : u \mapsto (u(x_0), ..., u(x_{N+1}))$$

is an isomorphism from V_h onto $\mathbb{R}^{N+2}$. In particular $\dim(V_h) = N+2$.

Proof of Lemma 5.1. In 1D, $H^1(\Omega) \subset C^0(\bar{\Omega})$. Conversely, continuous functions that are affine on each element belong to $H^1(\Omega)$. The properties of the mapping, i.e. Φ is linear and bijective, follow from the fact that a piecewise affine function is uniquely defined by its values at the nodes of the mesh. ■

Remark 5.2 *The choice of a* conforming *approximation space may have important consequences: assume that we are interested in $V = H^2(\Omega)$. Then we can show that $\{u \in V,\ u_{|K_i} \in \mathbb{P}^1\}$ is an approximation space with dimension 2 (it is the set of affine functions over the whole domain)! Thus it is clear that this approximation has a very limited interest for the numerical approximation of a variational formulation dealing with $H^2(\Omega)$.*

As Φ is bijective, any element $u \in V_h$ can be identified to $\Phi(u) \in \mathbb{R}^{N+2}$. The elements of $\Phi(u)$ are refered to as *degrees of freedom* in the approximation space V_h.

When the degrees of freedom take the form $u \mapsto u(a)$ with $a \in \bar{\Omega}$, we refer the method to *Lagrange finite elements.* The points a_i are the *nodes* associated to the approximation space. In the case of $\mathbb{P}^1$ finite elements, the nodes a_i coincide with the points x_i which define the mesh. However, this is not always the case as we will see below for instance for the $\mathbb{P}^2$ elements.

Definition 5.3 *We denote e_i for $i = 0, ..., N+1$ the vectors of the canonical basis of $\mathbb{R}^{N+2}$. We denote ϕ_i the shape function associated to the node x_i such that $\Phi(\phi_i) = e_i$. Thus the functions $\phi_i \in V_h$ are defined by the property*

$$\phi_i(x_j) = \delta_{ij}, \quad \forall i, j \in \{0, ..., N+1\},$$

see Figure 3.

Let us estimate the approximation error associated to V_h. For this let us introduce the notion of *interpolation operator.* We denote by $v \mapsto |v|_{H^k}$ the seminorm on $H^k(\Omega)$, defined by

$$|v|_{H^k} = \left(\sum_{|\alpha| = k} \|\partial^\alpha v\|_{L^2}^2 \right)^{\frac{1}{2}},$$

with the consistent convention:

$$H^0(\Omega) = L^2(\Omega), \qquad |v|_{H^0} = \|v\|_{L^2}.$$

Definition 5.4 (Interpolation operator $\mathscr{I}_h^1$) *The interpolation operator $\mathscr{I}_h^1$ from V onto V_h is defined as*

$$\forall u \in V, \quad \mathscr{I}_h^1 u = \Phi^{-1}\big(u(x_0), ..., u(x_{N+1})\big)$$

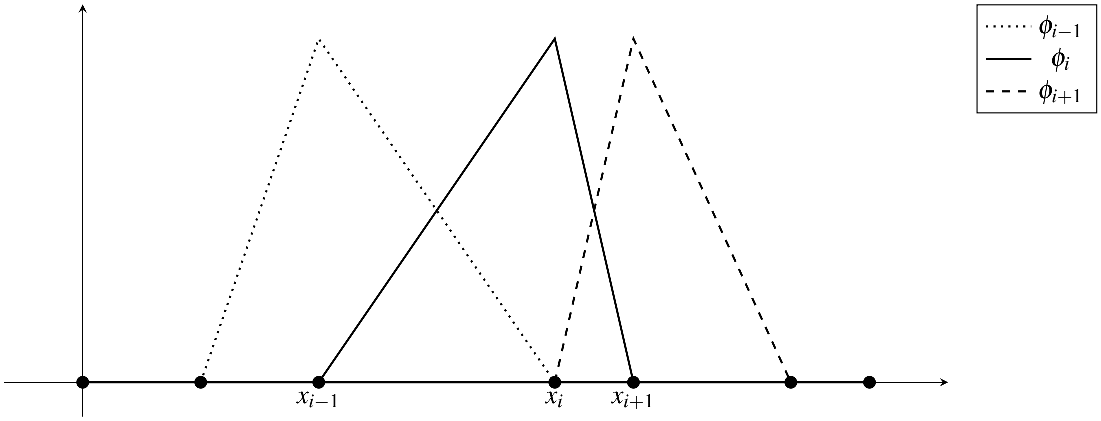


Figure 3. $\mathbb{P}^1$ finite element in 1D: global shape functions ϕ_{i-1}, ϕ_i and ϕ_{i+1}.

i.e.

$$\forall u \in V, \quad \forall x \in \bar{\Omega}, \quad \mathscr{I}_h^1 u(x) = \sum_{i=0}^{N+1} u(x_i)\, \phi_i(x).$$

Let us notice that if $u \in H_0^1(\Omega)$, then $\mathscr{I}_h^1 u \in H_0^1(\Omega)$ and thus

$$V_{h,0} := V_h \cap H_0^1(\Omega)$$

is a conforming approximation space for $H_0^1(\Omega)$. Let us now derive the estimates on $\mathscr{I}_h^1$.

We illustrate the shape of the $\mathbb{P}^1$ interpolation for a given function in Figure 5 (solid black line).

Theorem 5.5 (Properties of $\mathscr{I}_h^1$) *The following properties hold:*

1. *There exists $C > 0$ such that*

$$\forall u \in V, \quad \|\mathscr{I}_h^1 u - u\|_{L^2} \leq Ch\,|u|_{H^1}, \tag{16}$$

$$\forall u \in V, \quad \left|\mathscr{I}_h^1 u - u\right|_{H^1} \leq C\,|u|_{H^1}. \tag{17}$$

2. *There exists $C > 0$ such that*

$$\forall u \in V \cap H^2(\Omega), \quad \|\mathscr{I}_h^1 u - u\|_{L^2} \leq Ch^2\,|u|_{H^2}, \tag{18}$$

$$\forall u \in V \cap H^2(\Omega), \quad \left|\mathscr{I}_h^1 u - u\right|_{H^1} \leq Ch\,|u|_{H^2}. \tag{19}$$

3. *We have*

$$\forall u \in V, \quad \lim_{h \to 0} \|\mathscr{I}_h^1 u - u\|_{H^1} = 0.$$

In particular, for all $u \in V$, $\lim_{h\to 0} d(u, V_h) = 0$.

Remark 5.6 *The estimate* (19) *cannot be improved even if u is more regular. Besides Eq.* (16)*–*(19) *can be gathered in the following estimate: for all $m \in \{0,1\}$, for all $u \in H^{m+1}(\Omega)$,*

$$\|\mathscr{I}_h^1 u - u\|_{L^2} + h\left|\mathscr{I}_h^1 u - u\right|_{H^1} \leq Ch^{m+1}\,|u|_{H^{m+1}}. \tag{20}$$

Proof of Proposition 5.5. The basic idea relies on the analysis of $\mathscr{I}_h^1 u - u$ on each element of the mesh. Let $i \in \{0,...,N\}$ and let $x \in K_i$.

Step 1. Derivation of estimates (16) **and** (17)**.** For $x \in [x_i, x_{i+1}]$ we have

$$\mathscr{I}_h^1(x) = \frac{u_i(x_{i+1} - x) + u_{i+1}(x - x_i)}{h_i} \tag{21}$$

with $u_i = u(x_i)$ and $h_i = x_{i+1} - x_i$. Thus

$$\mathscr{I}_h^1(x) - u(x) = \frac{(u_i - u(x))(x_{i+1} - x) + (u_{i+1} - u(x))(x - x_i)}{h_i},$$

so that we obtain by the Cauchy-Schwarz inequality,

$$\begin{array}{rcl} \left|\mathscr{I}_h^1(x) - u(x)\right| & \leq & |u_i - u(x)| + |u_{i+1} - u(x)| \\ & \leq & \left|\displaystyle\int_{x_i}^{x} u'(t)\,\mathrm{d}t\right| + \left|\displaystyle\int_{x}^{x_{i+1}} u'(t)\,\mathrm{d}t\right| \\ & \leq & 2h_i^{\frac{1}{2}} \left(\left|\displaystyle\int_{x_i}^{x_{i+1}} \left|u'(t)\right|^2 \mathrm{d}t\right|\right)^{\frac{1}{2}} \end{array}$$

and then

$$\int_{x_i}^{x_{i+1}} \left|\mathscr{I}_h^1(x) - u(x)\right|^2 \mathrm{d}x \leq 4h_i^2 \left(\left|\int_{x_i}^{x_{i+1}} \left|u'(t)\right|^2 \mathrm{d}t\right|\right)$$

and, by summation,

$$\|\mathscr{I}_h^1 u - u\|_{L^2} \leq 2h\|u'\|_{L^2} = 2h\,|u|_{H^1}, \tag{22}$$

which is the estimate described in Eq. (16).

Besides, a consequence of the above estimate is that

$$\|\mathscr{I}_h^1 u\|_{L^2} \leq C\|u\|_{H^1}.$$

Back to Eq. (21), we have

$$\begin{array}{rcl} \left|(\mathscr{I}_h^1 u)'(x)\right| & = & \left|\dfrac{u_{i+1} - u_i}{h_i}\right| \\ & = & \dfrac{1}{h_i}\left|\displaystyle\int_{x_i}^{x_{i+1}} u'(t)\,\mathrm{d}t\right| \\ & \leq & \dfrac{1}{h_i^{\frac{1}{2}}}\left(\displaystyle\int_{x_i}^{x_{i+1}} \left|u'(t)\right|^2 \mathrm{d}t\right)^{\frac{1}{2}}, \end{array}$$

by the Cauchy-Schwarz inequality, hence

$$\int_{x_i}^{x_{i+1}} \left|(\mathscr{I}_h^1 u)'(x)\right|^2 \mathrm{d}x \le \int_{x_i}^{x_{i+1}} \left|u'(t)\right|^2 \mathrm{d}t.$$

Thus we have $\|(\mathscr{I}_h^1 u)'\|_{L^2} \le \|u'\|_{L^2}$, i.e. $\left|\mathscr{I}_h^1 u\right|_{H^1} \le |u|_{H^1}$. Thus we obtain

$$\left|\mathscr{I}_h^1 u - u\right|_{H^1} \le \left|\mathscr{I}_h^1 u\right|_{H^1} + |u|_{H^1} \le 2\,|u|_{H^1}\,,$$

which is the estimate described in Eq. (17).

Remark 5.7 *The estimate proves that if $u \in V$, then $\mathscr{I}_h^1 u$ weakly converges to u in $H^1(\Omega)$ when $h \to 0$. As $\{\mathscr{I}_h^1 u\}$ is bounded in $H^1(\Omega)$, there is a subsequence $\{h_n\}$ such that $h_n \to 0$ and $\{\mathscr{I}_{h_n}^1 u\}$ weakly converges to some u^* in $H^1(\Omega)$. By Eq. (22) the strong limit of $\mathscr{I}_{h_n}^1 u$ in L^2 is u. By uniqueness of the limit in the sense of distributions, we conclude that the weak limit in H^1 of this sequence is also u, hence $u^* = u$. By uniqueness of the adherent value (or closure point), the whole sequence $\mathscr{I}_h^1 u$ weakly converges to u in H^1, see Theorem 4.2. In fact we will prove below that $\mathscr{I}_h^1 u$* strongly *converges to u in H^1.*

Step 2. Derivation of estimates (18) **and** (19)**.** By Eq. (21),

$$(\mathscr{I}_h^1 u)'(x) - u'(x) = \frac{u_{i+1} - u_i}{h_i} - u'(x).$$

Using Taylor's formula, we have

$$\begin{array}{rcl} u_i &=& u(x) + (x_i - x)u'(x) + R_i(x), \\ u_{i+1} &=& u(x) + (x_{i+1} - x)u'(x) + R_{i+1}(x), \end{array}$$

with

$$R_i(x) := \frac{1}{2}\int_0^1 (1-t)u''(x + t(x_i - x))(x_i - x)^2\,\mathrm{d}t.$$

We obtain the following estimates:

- We write $(*) = \mathscr{I}_h^1 u(x) - u(x)$ as:

$$\begin{aligned} (*) &= \frac{(u_i - u(x))(x_{i+1} - x) + (u_{i+1} - u(x))(x - x_i)}{h_i} \\ &= \underbrace{\frac{(x_i - x)u'(x)(x_{i+1} - x) + (x_{i+1} - x)u'(x)(x - x_i)}{h_i}}_{=0} \\ &\quad + \frac{R_i(x)(x_{i+1} - x) + R_{i+1}(x)(x - x_i)}{h_i} \end{aligned}$$

  hence, by Jensen's inequality[6],

$$\begin{aligned} &\left|\mathscr{I}_h^1(x) - u(x)\right|^2 \\ &\quad \le \frac{R_i^2(x)(x_{i+1} - x) + R_{i+1}^2(x)(x - x_i)}{h_i} \\ &\quad \le R_i^2(x) + R_{i+1}^2(x). \end{aligned}$$

  We deduce

$$\int_{x_i}^{x_{i+1}} \left|\mathscr{I}_h^1 - u\right|^2 \le \int_{x_i}^{x_{i+1}} R_i^2 + \int_{x_i}^{x_{i+1}} R_{i+1}^2 \qquad (23)$$

  Let us now estimate $\int_{x_i}^{x_{i+1}} R_i^2$. First we have

$$\begin{aligned} \left|R_i^2(x)\right| &\le \frac{1}{2}\int_0^1 \left|u''(x + t(x_i - x))\right|^2 (x_i - x)^4\,\mathrm{d}t \\ &\le \frac{h_i^3}{2}\int_{x_i}^{x} \left|u''\right|^2 \qquad (24) \\ &\le \frac{h_i^3}{2}\int_{x_i}^{x_{i+1}} \left|u''\right|^2, \end{aligned}$$

  which yields

$$\int_{x_i}^{x_{i+1}} R_i^2 \le \frac{h_i^4}{2}\int_{x_i}^{x_{i+1}} \left|u''\right|^2,$$

  ans the same estimate for R_{i+1}. Putting this inequality into Eq. (23), we get

$$\int_{x_i}^{x_{i+1}} \left|\mathscr{I}_h^1 u - u\right|^2 \le h_i^4 \int_{x_i}^{x_{i+1}} \left|u''\right|^2.$$

  which provides Eq. (18) by summation over i.

- We write $(**) := (\mathscr{I}_h^1 u)'(x) - u'(x)$ as:

$$(**) = \frac{u_{i+1} - u_i}{h_i} - u'(x) = \frac{R_{i+1}(x) - R_i(x)}{h_i},$$

  and then, by (24),

$$\begin{aligned} &\left|(\mathscr{I}_h^1 u)'(x) - u'(x)\right|^2 \\ &\quad \le \frac{2}{h_i^2}(|R_i(x)|^2 + |R_{i+1}(x)|^2) \\ &\quad \le h_i\int_{x_i}^{x} \left|u''(t)\right|^2 \mathrm{d}t + h_i\int_{x}^{x_{i+1}} \left|u''(t)\right|^2 \mathrm{d}t. \end{aligned}$$

  Thus,

$$\int_{x_i}^{x_{i+1}} \left|(\mathscr{I}_h^1 u)' - u'\right|^2 \le h_i^2 \int_{x_i}^{x_{i+1}} \left|u''(t)\right|^2 \mathrm{d}t,$$

  which provides Eq. (19) by summation over i.

Step3. Convergence in H^1. By density of $H^2(\Omega)$ in $H^1(\Omega)$, the proof may be concluded. For every $u \in H^1(\Omega)$, and any $n \ge 1$, there exists $u_{(n)} \in C^\infty(\Omega)$ such that

$$\|u' - u'_{(n)}\|_{L^2} < \frac{1}{n}, \quad \forall n \ge 1.$$

Since $\mathscr{I}_h^1$ is a linear mapping, and using the uniform estimate (17), we have for any $n \ge 1$ and any $h > 0$

$$\|(\mathscr{I}_h^1 u)' - (\mathscr{I}_h^1 u_{(n)})'\|_{L^2} \le C\|u' - u'_{(n)}\|_{L^2} \le \frac{C}{n},$$

[6] By convexity of $x \mapsto x^2$, we have, for $\tau \in [0,1]$,

$$(a(1-\tau) + b\tau)^2 \le (1-\tau)a^2 + \tau b^2.$$

where C does not depend on h. By the triangle inequality, and (19), it comes

$$\begin{aligned}\|u' - (\mathscr{I}_h^1 u)'\|_{L^2} &\leq \|u' - u'_{(n)}\|_{L^2} + \|u'_{(n)} - (\mathscr{I}_h^1 u_{(n)})'\|_{L^2} \\ &\quad + \|(\mathscr{I}_h^1 u)' - (\mathscr{I}_h^1 u_{(n)})'\|_{L^2} \\ &\leq \tfrac{C}{n} + Ch\left|u_{(n)}\right|_{H^2}.\end{aligned}$$

It follows that, for any $n \geq 1$,

$$\limsup_{h\to 0} \|u' - (\mathscr{I}_h^1 u)'\|_{L^2} \leq \frac{C}{n}.$$

Letting $n \to \infty$, we conclude that

$$\limsup_{h\to 0} \|u' - (\mathscr{I}_h^1 u)'\|_{L^2} \leq 0,$$

which proves the claim. ■

From the reference element $\hat{K}$ to element K_i.

The previous analysis is done on each cell K_i separately and then global estimates are obtained by summation over all the cells. In view of the generalisation of the analysis to higher dimensions and more general approximation spaces, it is interesting to rewrite the previous analysis by means of changes of variable that transform each element K_i into a single *reference element* $\hat{K}$. In this strategy, we only need to analyse the interpolation properties on the reference element, and to analyse the properties of the change of variables.

The advantages are twofold: 1) computations are easier and 2) the method applies for higher order finite elements in higher dimension!

Let us give the main idea of the process in the case of the $\mathbb{P}^1$ finite element in 1D that we just analysed before. We consider the unit interval $\hat{K} = [0,1]$ as a reference element and we denote by $\mathscr{I}_0^1$ the Lagrange interpolation operator of degree 1 on this interval with the nodes $\{0,1\}$. If we assume that there exists $C > 0$ such that

$$\forall v \in H^1(\hat{K}), \quad \|\mathscr{I}_0^1 v - v\|_{L^2(\hat{K})} \leq C\,|v|_{H^1(\hat{K})}, \tag{25}$$

$$\forall v \in H^1(\hat{K}), \quad \left|\mathscr{I}_0^1 v - v\right|_{H^1(\hat{K})} \leq C\,|v|_{H^1(\hat{K})}, \tag{26}$$

$$\forall v \in H^2(\hat{K}), \quad \|\mathscr{I}_0^1 v - v\|_{L^2(\hat{K})} \leq C\,|v|_{H^2(\hat{K})}, \tag{27}$$

$$\forall v \in H^2(\hat{K}), \quad \left|\mathscr{I}_0^1 v - v\right|_{H^1(\hat{K})} \leq C\,|v|_{H^2(\hat{K})}, \tag{28}$$

then we can deduce immediately the results stated in Theorem 5.5. Indeed, for $i \in \{0,...,N\}$ we recall the notation $K_i = [x_i, x_{i+1}]$ and we introduce the affine mapping that transforms any cell K_i to the unit cell $\hat{K} = [0,1]$:

Definition 5.8 (Affine mapping T_i) *We introduce the affine change of variables*

$$\begin{array}{cccc} T_i & : & \hat{K} & \to & K_i \\ & & t & \mapsto & (1-t)x_i + t x_{i+1}. \end{array}$$

Straightforward computations lead to the following properties:

Proposition 5.9 (Properties of T_i) *For any function $\phi \in H^1(K_i)$,*

$$\|\phi \circ T_i\|_{L^2(\hat{K})} = \frac{1}{\sqrt{h_i}} \|\phi\|_{L^2(K_i)}, \quad |\phi \circ T_i|_{H^1(\hat{K})} = \sqrt{h_i}\,|\phi|_{H^1(K_i)}.$$

Moreover we have

$$(\mathscr{I}_h^1 u) \circ T_i = \mathscr{I}_0^1 (u \circ T_i). \tag{29}$$

Now let us derive the estimates:

- We apply Eq. (25) to $\phi = u \circ T_i$ and use (29):
  $$\|(\mathscr{I}_h^1 u) \circ T_i - u \circ T_i\|_{L^2(\hat{K})} \leq C\,|u \circ T_i|_{H^1(\hat{K})},$$
  then with the change of variables
  $$\|\mathscr{I}_h^1 u - u\|_{L^2(K_i)} \leq C h_i\,|u|_{H^1(K_i)},$$
  By summation over i, we get the estimate (16).
- The other proofs can be adapted in a similar way to derive estimates (17), (19) and (18).

As a conclusion, the estimate of the interpolation error over a fixed interval $\hat{K}$ allows us to derive the local estimate over the elements of the mesh.

Example of $\mathbb{P}^1$ approximation of an elliptic problem in 1D. Consider $\Omega =]0,1[$ and

$$\left\{\begin{array}{rcll} -(\mathscr{K} u')' + \alpha u &=& f, & \text{in } \Omega, \\ u &=& 0, & \text{on } \{0,1\}, \end{array}\right.$$

with $\alpha \geq 0$, $f \in L^2(\Omega)$, $\mathscr{K} \in L^\infty(\Omega)$ and $\inf(\mathscr{K}) > 0$. The corresponding variational formulation writes

$$(\mathrm{P}) \left\{\begin{array}{l} \text{Find } u \in V \text{ such that} \\ a(u,v) = L(v), \\ \text{for all } v \in V, \end{array}\right.$$

with $V = H_0^1(\Omega)$ and

$$a(u,v) = \int_0^1 \mathscr{K} u'v' + \alpha \int_0^1 uv,$$

$$L(v) = \int_0^1 fv.$$

Problem (P) admits a unique solution by the Lax-Milgram theorem. The Galerkin approximation of the elliptic abstract problem consists in solving the following problem

$$(\mathrm{P}_h) \left\{\begin{array}{l} \text{Find } u_h \in V_{h,0} \text{ such that} \\ a(u_h, v_h) = L(v_h), \\ \text{for all } v_h \in V_{h,0}. \end{array}\right.$$

Problem (P_h) admits a unique solution u_h by the Lax-Milgram theorem. By Theorem 5.5 we have $d(u, V_{h,0}) \leq \|u - \mathscr{I}_h^1 u\|_V \to 0$ as $h \to 0$. Combined with Céa's lemma, see Lemma 4.1, we deduce the convergence of u_h towards u without any additional assumption. Moreover, we can estimate the error as follows.

Theorem 5.10 ($\mathbb{P}^1$ error estimate) *Assume that the solution of* (P) *is regular, i.e.* $u \in H^2(\Omega)$, *then*

$$\|u-u_h\|_V \leq C\frac{\|a\|}{\alpha} h\, |u|_{H^2}.$$

In some cases the elliptic problem has some regularity properties which imply that, if the source term is L^2 then the solution is H^2. In the current case this property holds if A is sufficiently regular (Lipschitz continuity is enough) and we get the elliptic regularity property: there exists $C > 0$ such that the unique solution of the variational problem belongs to $H^2(\Omega)$ and satisfies

$$\|u\|_{H^2(\Omega)} \leq C\|f\|_{L^2(\Omega)}.$$

The error estimate becomes:

$$\|u-u_h\|_V \leq Ch\|f\|_{L^2(\Omega)}.$$

Let us discuss the derivation of estimates in a weaker norm (e.g. in $L^2(\Omega)$). In order to derive a new estimate, we need to assume furthermore that the *adjoint problem*[7] satisfies a so called *elliptic regularity property*: there exists $C > 0$ such that for all $v \in L^2(\Omega)$ the unique function $\phi_v \in V$ such that

$$a(w, \phi_v) = (v, w)_{L^2}, \quad \forall w \in V$$

satisfies

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

In our example this property is clearly satisfied because the adjoint problem is identical to the initial one by symmetry of the bilinear form a.

Theorem 5.11 (Aubin-Nitsche) *Assume that the adjoint problem of* (P) *satisfies the elliptic regularity property. Then the solution u of* (P) *and its $\mathbb{P}^1$ finite element approximation u_h satisfy*

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

In particular, if $u \in H^2(\Omega)$, Theorem 5.10 leads to

$$\|u-u_h\|_{L^2} \leq Ch^2\, |u|_{H^2}.$$

Proof of Theorem 5.11. Denote $e_h = u - u_h \in V$. Taking $v = w = e_h$ in the adjoint problem, we get

$$\|e_h\|^2_{L^2(\Omega)} = a(e_h, \phi_{e_h}).$$

As the error $e_h := u - u_h$ is $a-$orthogonal to V_h, we have

$$\begin{array}{rcl} \|e_h\|^2_{L^2(\Omega)} & = & a(e_h, \phi_{e_h} - \mathscr{I}^1_h \phi_{e_h}) \\ & \leq & \|a\|\|e_h\|_V \|\phi_{e_h} - \mathscr{I}^1_h \phi_{e_h}\|_{H^1} \\ & \leq & Ch\|e_h\|_V \|\phi_{e_h}\|_{H^2} \\ & \leq & Ch\|e_h\|_V \|e_h\|_{L^2(\Omega)}, \end{array}$$

hence the result. ■

Actually the above result is somehow general and not particular to $\mathbb{P}^1$ approximation.

Practical aspects related to the computation of the approximate solution.

Exercise 11 *Let $\{\phi_i\}_{i=1,\dots,N}$ be a basis for $V_{h,0}$. Prove that* (P_h) *is equivalent to a linear system (to be determined): find $U \in \mathbb{R}^N$ such that $A \cdot U = b$, where $A \in \mathscr{M}_{N\times N}(\mathbb{R})$ and $b \in \mathbb{R}^N$.*

A function $u_h \in V_{h,0}$ takes the form $u_h = \sum_{i=1}^N u_i \phi_i$, since the degrees of freedom corresponding to boundary nodes are 0. Hence, this solution is completely determined by $U := (u_i)_{i=1,\dots,N}$. The variational formulation then takes the form $A \cdot U = b$ with

$$A = (a(\phi_j, \phi_i))_{1\leq i,j\leq N}, \qquad b = (L(\phi_i))_{1\leq i\leq N}.$$

In our example, matrix A splits into $A^{(r)} + \alpha A^{(m)}$

$$A_{ij} = \underbrace{\int_0^1 \mathscr{K}\phi'_i\phi'_j}_{A^{(r)}_{i,j}} + \alpha \underbrace{\int_0^1 \phi_i\phi_j}_{A^{(m)}_{i,j}}$$

where $A^{(r)}$ is the so-called rigidity matrix whereas $A^{(m)}$ is the so-called mass matrix. In the case of the $\mathbb{P}^1$ finite element, integrals are zero as soon as $|i-j| > 1$ because the supports of ϕ_i and ϕ_j are disjoint in that case. Thus the remaining computations are the following ones:

$$\begin{array}{rcl} A^{(r)}_{i,i-1} & = & -\dfrac{1}{h^2_{i-1}}\displaystyle\int_{x_{i-1}}^{x_i} \mathscr{K}, \\ A^{(r)}_{i,i} & = & \dfrac{1}{h^2_{i-1}}\displaystyle\int_{x_{i-1}}^{x_i} \mathscr{K} + \dfrac{1}{h^2_i}\displaystyle\int_{x_i}^{x_{i+1}} \mathscr{K}, \\ A^{(r)}_{i,i+1} & = & -\dfrac{1}{h^2_i}\displaystyle\int_{x_i}^{x_{i+1}} \mathscr{K}, \\ A^{(m)}_{i,i-1} & = & \displaystyle\int_{x_{i-1}}^{x_i} \phi_{i-1}\,\phi_i = \dfrac{h_{i-1}}{6} \\ A^{(m)}_{i,i} & = & \displaystyle\int_{x_{i-1}}^{x_{i+1}} \phi_i^2 = \dfrac{h_{i-1}+h_i}{3}, \\ A^{(m)}_{i,i+1} & = & \displaystyle\int_{x_i}^{x_{i+1}} \phi_i\,\phi_{i+1} = \dfrac{h_i}{6} \end{array}$$

with suitable adaptations for $i = 1$ and $i = N$.

Proposition 5.12 *The matrix A is tridiagonal, symmetric and positive-definite. Moreover it satisfies the discrete maximum principle*

$$A^{-1} \geq 0.$$

Proof of Proposition 5.12. The first part follows from the fact that A is a Gram matrix[8] of $(\phi_i)_{i=1,\dots,N}$ in V for

[7] This means that the unknown is now the second variable of the bilinear form a and the test function is the first variable of a.

[8] A Gram matrix G of a set of vectors $(v_1,\dots,v_n)$ in an inner product space is the self-adjoint matrix of inner products, whose entries are given by $G_{ij} = (v_i, v_j)$. A Gram matrix is positive semi-definite.

some scalar product. The discrete maximum principle consists in showing that A is a M-matrix[9]. ■

In practical computations, integrals are numerically computed with quadrature formula. This may induce additional work in order to guarantee that this process does not induce a loss of precision.

5.2 Finite elements in 1D: $\mathbb{P}^2$ finite element

Approximation space $\mathbb{P}^2$ in 1D. Assume that the solution of a problem is much more regular (e.g of class C^∞) than it is expected from the variational formulation of the problem. The $\mathbb{P}^1$ finite element method does not allow us to get a better precision: indeed the precision of the interpolation operator which is associated to the approximation space does not increase when increasing the regularity of the interpolated function. In order to take advantage from the regularity of the solution (without modifying the mesh) it is necessary to adapt the approximation space. Instead of considering piecewise affine functions, we now consider *piecewise quadratic functions.*

Using the same notations as before, we introduce the center of the elements $x_{i+\frac{1}{2}} = \frac{x_i + x_{i+1}}{2}$, for $i = 0, ..., N$. Let us now consider the set of piecewise quadratic functions:

$$V_h = \{u \in V,\ u_{|K_i} \in \mathbb{P}^2\},$$

where $\mathbb{P}^2$ is the set of polynomials of degree 2.

Lemma 5.15 *The space V_h writes*

$$V_h = \{u \in C^0(\bar{\Omega}),\ u_{|K_i} \in \mathbb{P}^2\},$$

and, moreover, the mapping

$$\Phi : u \mapsto (u(x_0), u(x_{\frac{1}{2}}), u(x_1), u(x_{\frac{3}{2}}), ..., u(x_{N+\frac{1}{2}}), u(x_{N+1}))$$

is an isomorphism from V_h onto $\mathbb{R}^{2N+3}$. In particular $\dim(V_h) = 2N+3$.

Proof of Lemma 5.15. The injectivity of Φ comes from the fact that a polynomial of degree 2 with three distinct roots is necessary zero. Surjectivity of Φ emerges from the existence of a Lagrange interpolation polynomial on each K_i. ■

Although the degree of the polynomials has been increased, we do not build a *conforming* approximation space for H^2 since the *continuity of the derivatives* at the interfaces would be required.

Each of the coordinates functions of Φ is a linear form on V_h which is called degree of freedom. They all consist in evaluating the function at some point, which leads to the Lagrange terminology. The definition of the shape functions for the $\mathbb{P}^2$ finite element follows the same rule as for the $\mathbb{P}^1$ finite element, up to the dimension modifications.

The shape functions $(\phi_i)_i$ and $(\phi_{i+\frac{1}{2}})_i$ are defined by

$$\begin{array}{lcll} \phi_i(x_j) & = & \delta_{ij}, & \forall j \in \{0, ..., N+1\}, \\ \phi_i(x_{j+\frac{1}{2}}) & = & 0, & \forall j \in \{0, ..., N\}, \\ & & & \\ \phi_{i+\frac{1}{2}}(x_j) & = & 0, & \forall j \in \{0, ..., N+1\}, \\ \phi_{i+\frac{1}{2}}(x_{j+\frac{1}{2}}) & = & \delta_{ij}, & \forall j \in \{0, ..., N\}, \end{array}$$

see Figure 4. Thus

- the support of ϕ_i is $K_{i-1} \cup K_i$;
- the support of $\phi_{i+\frac{1}{2}}$ is K_i.

Definition 5.16 (Interpolation operator $\mathscr{I}_h^2$) *The interpolation operator $\mathscr{I}_h^2$ from V onto V_h is defined as*

$$\mathscr{I}_h^2 u(x) = \sum_{i=0}^{N+1} u(x_i)\, \phi_i(x) + \sum_{i=0}^{N} u(x_{i+\frac{1}{2}})\, \phi_{i+\frac{1}{2}}(x),$$

for all $u \in V$, for all $x \in \Omega$.

Let us notice that if $u \in H_0^1(\Omega)$, then $\mathscr{I}_h^2 u \in H_0^1(\Omega)$ and thus $V_h \cap H_0^1(\Omega)$ is a conforming approximation space for $H_0^1(\Omega)$. As before, $\mathscr{I}_h^2$ is a continuous projection of $H^1(\Omega)$ into itself and we can derive the following interpolation properties:

Theorem 5.17 (Properties of $\mathscr{I}_h^2$) *There exists $C > 0$ such that, for all $m \in \{0,1,2\}$, for all $u \in H^{m+1}(\Omega)$,*

$$\|\mathscr{I}_h^2 u - u\|_{L^2} + h\left|\mathscr{I}_h^2 u - u\right|_{H^1} \leq C h^{m+1} |u|_{H^{m+1}}.$$

Remark 5.18 *Notice that if $u \in H^2(\Omega)$ the interpolation result does not provide a better estimate than $\mathbb{P}^1$ finite elements. The use of such element should be motivated by the construction of an approximate solution for a solution which belongs to $H^3(\Omega)$.*

Proof of Theorem 5.17. The proof is similar to the proof of Theorem 5.5. We can also use the technique of

[9] We recall the definition:

Definition 5.13 *A so-called M-matrix which satisfies the following conditions:*

- *it is a Z-matrix, i.e. off-diagonal entries are less than or equal to zero;*
- *the real part of the eigenvalues are positive.*

Besides we recall the following property

Proposition 5.14 *The following are equivalent:*

- *A is a non-singular M-matrix;*
- *A is inverse-positive. That is, A^{-1} exists and $A^{-1} \geq 0$.*

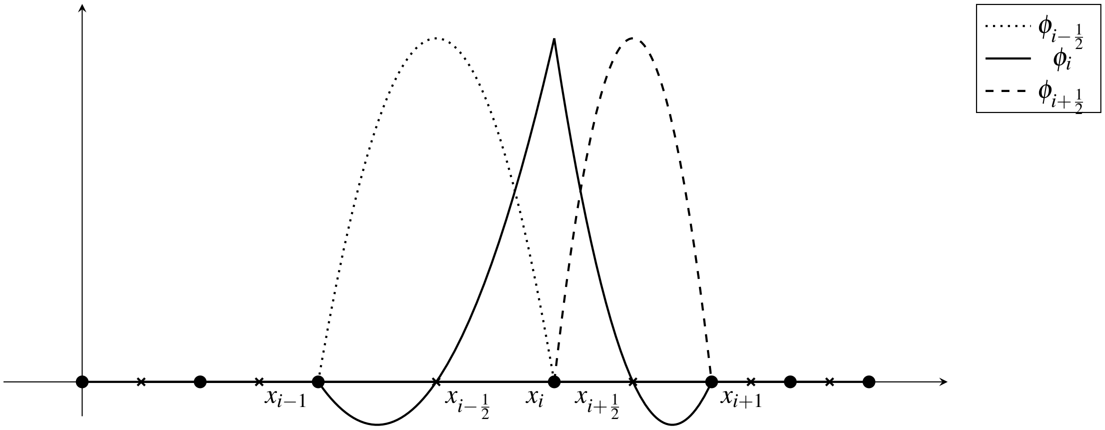


Figure 4. $\mathbb{P}^2$ finite element in 1D: shape functions $\phi_{i-\frac{1}{2}}$, ϕ_i and $\phi_{i+\frac{1}{2}}$.

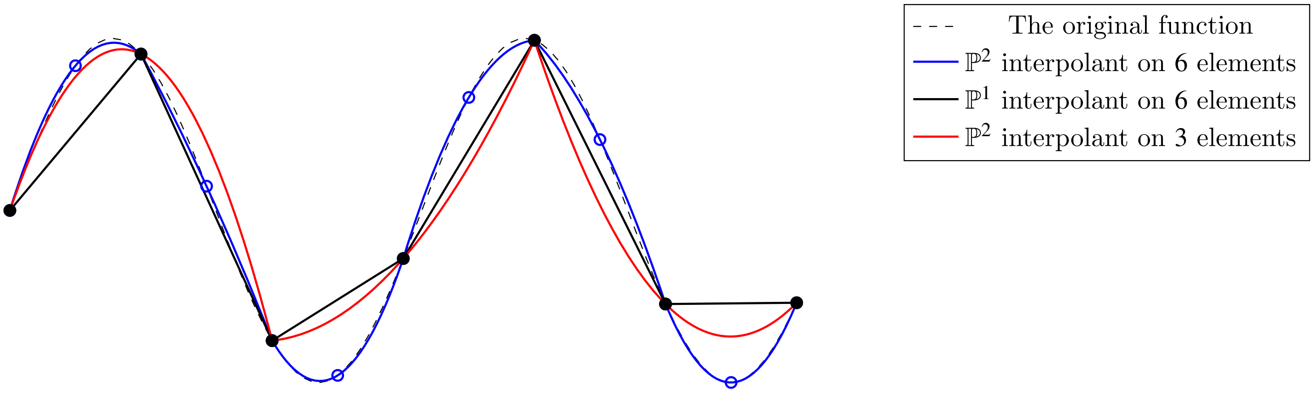


Figure 5. A smooth function (dashed line) and different kind of interpolations.

the change of variables by reducing the study to the interpolation properties of the operator on the unit domain $[0,1]$. ■

Example of $\mathbb{P}^2$ approximation of an elliptic problem in 1D. Considering the $\mathbb{P}^2$ approximation space as V_h and $V_{h,0} = V_h \cap H_0^1(\Omega)$. We obtain the convergence in the general case because $d(u, V_{h,0}) \to 0$ as $h \to 0$ for all $u \in V$.

Theorem 5.19 ($\mathbb{P}^2$ error estimate) *Assume that the solution $u \in V$ of the problem belongs to $H^3(\Omega)$. Then we have*

$$\begin{aligned} \|u-u_h\|_{H^1(\Omega)} &\leq Ch^2 |u|_{H^3(\Omega)}, \\ \|u-u_h\|_{L^2(\Omega)} &\leq Ch^3 |u|_{H^3(\Omega)}. \end{aligned}$$

Proof of Theorem 5.19. The first property follows from the above analysis whereas the second property follows from the Aubin-Nitsche trick, see the proof of Theorem 5.11. ■

We illustrate the differences between various interpolations of a same function in Figure 5.

Practical aspects related to the computation of the approximate solution. In practice the numbering of the unknowns is done continuously between 1 and $2N+1$ (we recall that x_0 and x_{N+1} are not taken into account because of the boundary conditions). More precisely, any $k \in \{1, ..., 2N+1\}$ can be written as $k = 2i + p$, with $p \in \{0, ..., N\}$ and $p \in \{0, 1\}$. We will denote $k = [i, p]$. The numbering of the basic functions writes:

$$\psi_k = \psi_{[i,p]} = \begin{cases} \phi_i, & \text{if } p = 0, \\ \phi_{i+\frac{1}{2}}, & \text{if } p = 1. \end{cases}$$

In the same way the numbering of the unknowns U_k related to the coordinates of the solution u_h in the basis $\{\psi_k\}$ follows the same rule.

Let us consider the matrix $A = (a_{k,l})_{kl}$ with

$$a_{k,l} = a(\psi_l, \psi_k) = \int_\Omega \psi_k' \psi_l'.$$

- If $k = [i, 0]$ the support of ψ_k is $K_{i-1} \cup K_i$. As a consequence the only coefficients l for which $a_{k,l}$ may be non-zero are:
  - for $l = k = [i, 0]$,
  $$a_{k,k} = \int_{x_{i-1}}^{x_{i+1}} |\phi_i'|^2,$$
  - for $l = k+1 = [i, 1]$,
  $$a_{k,k+1} = \int_{x_i}^{x_{i+1}} \phi_i' \phi_{i+\frac{1}{2}}',$$
  - for $l = k+2 = [i+1, 0]$,
  $$a_{k,k+2} = \int_{x_i}^{x_{i+1}} \phi_i' \phi_{i+1}',$$
  - for $l = k-1 = [i-1, 1]$,
  $$a_{k,k-1} = \int_{x_{i-1}}^{x_i} \phi_i' \phi_{i-\frac{1}{2}}',$$
  - for $l = k-2 = [i-1, 0]$,
  $$a_{k,k-2} \int_{x_{i-1}}^{x_i} \phi_i' \phi_{i-1}'.$$
- If $k = [i, 1]$ the support of ψ_k is K_i. As a consequence the only coefficients l for which a_{kl} may be non-zero are:
  - for $l = k = [i, 1]$,
  $$a_{k,k} = \int_{x_i}^{x_{i+1}} \left|\phi_{i+\frac{1}{2}}'\right|^2,$$
  - for $l = k+1 = [i+1, 0]$,
  $$a_{k,k+1} = \int_{x_i}^{x_{i+1}} \phi_{i+\frac{1}{2}}' \phi_{i+1}',$$
  - for $l = k-1 = [i, 0]$,
  $$a_{k,k-1} = \int_{x_i}^{x_{i+1}} \phi_{i+\frac{1}{2}}' \phi_i'.$$

Thus the matrix A is pentadiagonal. The linear system to solve is more complicated than with the $\mathbb{P}^1$ approximation. The matrix is symmetric positive-definite but it is not an M-matrix and the discrete maximum principle is not satisfied anymore.

5.3 $\mathbb{P}^k$ finite element in dimension $d > 1$

Meshes. We assume that Ω is a bounded, connected, polygonal in 2D or polyhedral in 3D.

Proposition 5.20 *Let $\mathscr{T}$ be a polygonal mesh, i.e. a set of polygonal/polyhedral cells $(K)_{K \in \mathscr{T}}$ such that*

$$\bar{\Omega} = \cup_{K \in \mathscr{T}} \bar{K}, \qquad \mathring{K} \cap \mathring{L} = \emptyset, \text{ if } K \neq L.$$

Let $m \in \mathbb{N} \setminus \{0\}$ and u a function defined on Ω. The following properties are equivalent:

- *$u \in H^m(\Omega)$;*
- *For all $K \in \mathscr{T}$, $u_{|K} \in H^m(K)$ and, for all $K \neq L$ such that the codimension of $\bar{K} \cap \bar{L}$ is 1, then the trace of $\partial^\alpha u_{|K}$ and the trace of $\partial^\alpha u_{|L}$ coincide on $\partial K \cap \partial L$, for any $|\alpha| \leq m-1$.*

If the properties hold, then for all $|\alpha| \leq m$,

$$\|\partial^\alpha u\|_{L^2(\Omega)} = \sum_{K \in \mathscr{T}} \|\partial^\alpha u_{|K}\|_{L^2(K)}.$$

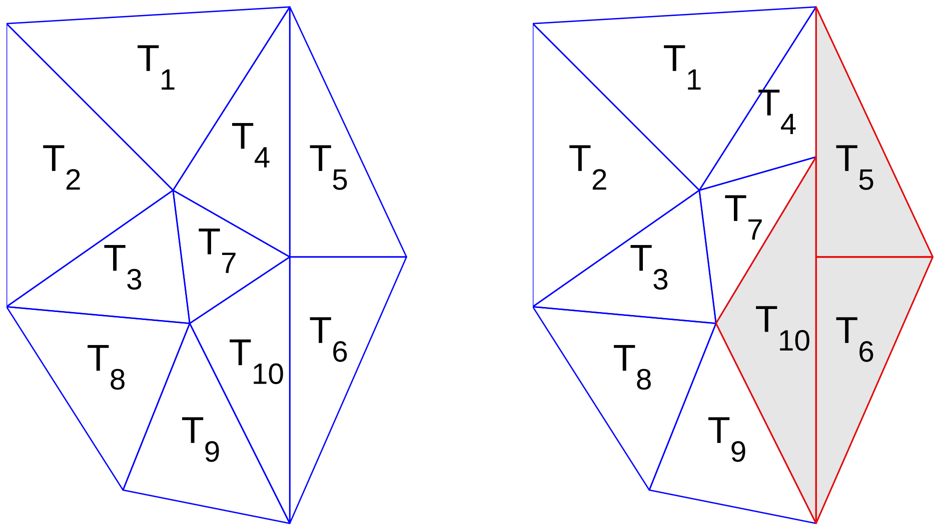


Figure 6. Admissible mesh and non-admissible mesh.

This result explains how it is possible to design finite element spaces that are H^m−conforming when the dimension is greater than 1. We take piecewise regular functions with additional constraint: the jumps of u and its derivatives at the interfaces of the mesh have to be zero. Then the geometrical conformity of the mesh is essential: indeed in Figure 6 the atypical node in the non-admissible mesh cannot be associated to a degree of freedom of the approximation space: if we consider a $\mathbb{P}^1$ approximation, then the value of the approximate solution in this particular node is related to the values at the vertices of triangle T_{10}.

We will now consider a mesh of Ω made of simplices, i.e. the convex hulls of $d+1$ points that do not belong to an hyperplane. Roughly speaking they are triangles in 2D, tetrahedra in 3D etc.

Simplicial finite element. The $\mathbb{P}^k$ simplicial Lagrange element. As it has been previously outlined, the finite element may be defined on a reference element. All the geometrical quantities of the mesh and the approximation space are based upon this object.

Let us denote $\hat{K}$ the unit simplex:

$$\hat{K} = \left\{ x \in \mathbb{R}^d,\ (\forall i,\ x_i \geq 0),\ \sum_i x_i \leq 1 \right\}.$$

Proposition 5.21 (Affine mapping T_K) *Any simplex $K = \operatorname{conv}(a_0, ..., a_d)$ of the mesh is the image of $\hat{K}$ by an affine mapping $T_K : \hat{K} \to K$ of the form*

$$T_K(\hat{x}) = a_0 + B_K \hat{x}$$

where the ith column of B_K is defined as the coordinates of $a_i - a_0$ in the canonical basis of $\mathbb{R}^d$. We have the following properties:

- $|\det(B_K)| = d!\ |K|$,
- $\|B_K\|_2 \leq \dfrac{h_K}{\rho_{\hat{K}}}$,
- $\|B_K^{-1}\|_2 \leq \dfrac{h_{\hat{K}}}{\rho_K}$.

Here h_K and $h_{\hat{K}}$ denote the diameters of K and $\hat{K}$ respectively whereas ρ_K and $\rho_{\hat{K}}$ denote the diameters of the incircles of K and $\hat{K}$ respectively.

Proof of Proposition 5.21. Let us prove the estimates on B_K. Using a change of variables,

$$|K| = \int_K 1\,\mathrm{d}x = \int_{\hat{K}} 1\ |\det(B_K)|\ \mathrm{d}\hat{x} = |\det(B_K)|\left|\hat{K}\right|,$$

where the volume of the reference simplex is given by $\left|\hat{K}\right| = 1/d!$. By definition of $\|\cdot\|_2$ we have

$$\|B_K\|_2 = \sup_{\|\hat{x}\| = \rho_{\hat{K}}} \frac{\|B_K \cdot \hat{x}\|}{\rho_{\hat{K}}}.$$

This supremum is attained (by a finite dimension argument): thus there exists $\hat{a}, \hat{b} \in \hat{K}$ such that $\|\hat{a} - \hat{b}\| = \rho_{\hat{K}}$ and

$$\|B_K(\hat{a} - \hat{b})\| = \|T_K(\hat{a}) - T_K(\hat{b})\| \leq h_K,$$

hence the result. The last inequality is obtained by exchanging the roles of K and $\hat{K}$. ■

Theorem 5.22 *Let $v : K \to \mathbb{R}$. Then $v \in H^m(K)$ if, and only if, $\hat{v} = v \circ T_K \in H^m(\hat{K})$. Moreover for any $0 \leq k \leq m$ we have*

$$|v|_{H^k(K)} \leq C \frac{|K|^{\frac{1}{2}}}{\rho_K^k} |\hat{v}|_{H^k(\hat{K})},$$

$$|\hat{v}|_{H^k(\hat{K})} \leq C \frac{h_K^k}{|K|^{\frac{1}{2}}} |v|_{H^k(K)}.$$

Remark 5.23 *Let us recall that for $k = 0$, $H^0 = L^2$ and $|\cdot|_{H^0} = \|\cdot\|_{L^2}$.*

Proof of Theorem 5.22. Let us prove the theorem for $m = 1$ (the general case readily adapts).

- *Case $k = 0$.* By using the change of variables

$$\begin{aligned}\|v\|^2_{L^2(K)} &= \int_K |v(x)|^2\,\mathrm{d}x\\ &= \int_{\hat{K}} |\hat{v}(\hat{x})|^2\ |\det(B_K)|\ \mathrm{d}\hat{x}\\ &= |K|\ \|\hat{v}\|^2_{L^2(\hat{K})}.\end{aligned}$$

- *Case $k = 1$.* We have

$$\hat{\nabla}\hat{v} = B_K^{\mathrm{t}}(\nabla v) \circ T_K = B_K^{\mathrm{t}}\widehat{\nabla v},$$

hence

$$\|\hat{\nabla} v\|^2_{L^2(\hat{K})} \leq \frac{h_K^2}{\rho_{\hat{K}}^2} \|\widehat{\nabla v}\|^2_{L^2(\hat{K})} = \frac{h_K^2}{\rho_{\hat{K}}^2 |K|} \|\nabla v\|^2_{L^2(K)}.$$

The other inequality is obtained by exchanging the roles of K and $\hat{K}$. ■

Proposition 5.24 (Local simplicial Lagrange f.e.) *We consider $\Sigma \subset \hat{K}$ the set of points defined as*

$$\left(\frac{I_1}{k}, ..., \frac{I_d}{k}\right), \quad \forall (I_1, ..., I_d) \in \mathbb{N}^d, \quad I_1 + ... + I_d \leq k.$$

Then $|\Sigma| = C^k_{k+d} = \frac{(k+d)!}{d!\ k!}$.

We denote $(\hat{a}_j)_{1 \leq j \leq |\Sigma|}$ the elements of this set. Then the mapping $p \in \mathbb{P}^k \mapsto (p(\hat{a}_j))_{1 \leq j \leq |\Sigma|} \in \mathbb{R}^{|\Sigma|}$ is an isomorphism (in particular $\mathbb{P}^k$ and Σ have the same cardinality).

- *the triplet $(\hat{K}, \mathbb{P}^k, \Sigma)$ is the so-called* $\mathbb{P}^k$ simplicial Lagrange finite element.

- *the set $(\hat{a}_j)_j$ is the set of nodes of this finite element and the linear form $v \in C^0(\hat{K}) \mapsto v(\hat{a}_j)$ is denoted the* degree of freedom *associated to $\hat{a}_j$.*

- *For all $1 \leq i \leq |\Sigma|$ there exists a unique function $\theta_i \in \mathbb{P}^k$ such that*

  $$\theta_i(\hat{a}_j) = \delta_{ij}, \quad \forall 1 \leq j \leq |\Sigma|.$$

  These functions are the so-called local shape functions.

Example. In 1D, the nodes of the $\mathbb{P}^k$ finite elements are described in Figure 8:

- $\mathbb{P}^0$ finite element: the node is the center of mass of $\hat{K}$ (this element will be discussed later).

- $\mathbb{P}^1$ finite element: the nodes are the vertices of $\hat{K}$. Figure 7 a) presents the degrees of freedom along with the local shape functions in this case.

- $\mathbb{P}^2$ finite element: the nodes are the vertices and the center of $\hat{K}$. Figure 7 b) presents the degrees of freedom along with the local shape functions in this case.

□

Example. In 2D, the nodes of the $\mathbb{P}^k$ finite elements are described in Figure 8:

- $\mathbb{P}^0$ finite element: the node is the center of mass of $\hat{K}$ (this element will be discussed later).

- $\mathbb{P}^1$ finite element: the nodes are the vertices of $\hat{K}$.

- $\mathbb{P}^2$ finite element: the nodes are the vertices and the center of the edges of $\hat{K}$.

- $\mathbb{P}^3$ finite element: the nodes are the vertices, the center of mass and the points located at $1/3$ and $2/3$ in each edge of $\hat{K}$.

□

Example. In 3D, the nodes of the $\mathbb{P}^k$ finite elements are described in Figure 9:

- $\mathbb{P}^1$ finite element: the nodes are the vertices of $\hat{K}$.

- $\mathbb{P}^2$ finite element: the nodes are the vertices and the center of the edges of $\hat{K}$.

□

Definition 5.25 (Simplicial conforming mesh) *Let $\mathscr{T}$ be a mesh made of simplicial elements K. We say that $\mathscr{T}$ is* geometrically conforming *if, and only if, for any distinct elements K and L, $\mathscr{E} = \bar{K} \cap \bar{L}$ satisifes one of the following properties:*

- *either* $\dim(F) \leq d-2$ *(note that F may be empty);*

- *or F is face for K and a face for L.*

Proposition 5.26 (Global simplicial Lagrange f.e.) *Let $\mathscr{T}$ be a simplicial conforming mesh of Ω.*

- *The set of discretization nodes is denoted $\Sigma_h = \cup_K \Sigma_K = \cup_K T_K(\Sigma)$.*

- *We call $\mathbb{P}^k$ approximation space on the mesh $\mathscr{T}$ the space*

  $$V_h := \{v \in C^0(\bar{\Omega}),\ \forall K \in \mathscr{T},\ v_{|K} \in \mathbb{P}^k\}.$$

- *For any $a \in \Sigma_h$ there exists a unique function in V_h, denoted ϕ_a, such that*

  $$\phi_a(a) = 1, \qquad \phi_a(b) = 0, \quad \forall b \in \Sigma_h \setminus \{a\}.$$

  These functions are the so-called global shape functions *of the approximation space. They form a basis for V_h.*

- *For any $a \in \Sigma_h$ the mapping $v \in C^0(\bar{\Omega}) \mapsto v(a)$ is a continuous linear form which is called a* degree of freedom *associated to node a.*

The space V_h satisfies

$$V_h = \{v \in H^1(\Omega),\ \forall K \in \mathscr{T},\ v_{|K} \in \mathbb{P}^k\}.$$

Moreover the link between the local and global shape functions is given by:

Proposition-Definition 5.27 (Shape functions) *Let $a \in$ Let $a \in \Sigma_h$ and let $K \in \mathscr{T}$ such that $a \in K$.*

- *There exists a unique $J(a,K) \in \{1,...,|\Sigma|\}$ such that $a = T_K(\hat{a}_{J(a,K)})$ (we say that $J(a,K)$ is the local index of node a in the element K).*

- *We have*

  $$(\phi_a)_{|K} \circ T_K = \theta_{J(a,K)},$$

Remark 5.28 *The function $\theta_{J(a,K)}$ is the local shape function associated to the node $J(a,K)$ in $\hat{K}$.*

Local interpolation operator. Global interpolation operator.

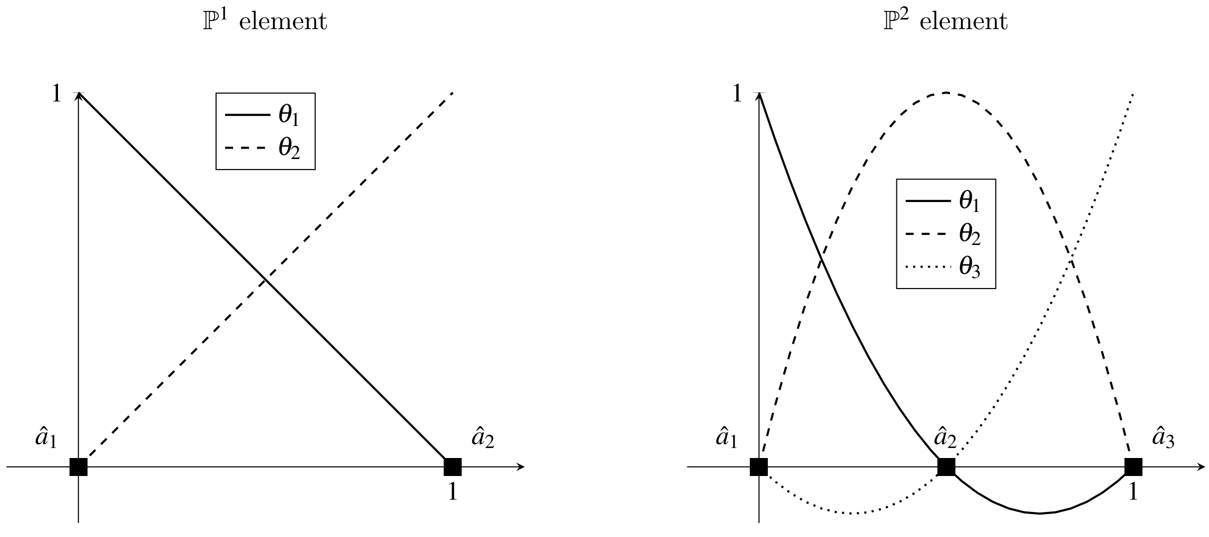


Figure 7. Local shape functions for the $\mathbb{P}^1$ and $\mathbb{P}^2$ elements in 1D

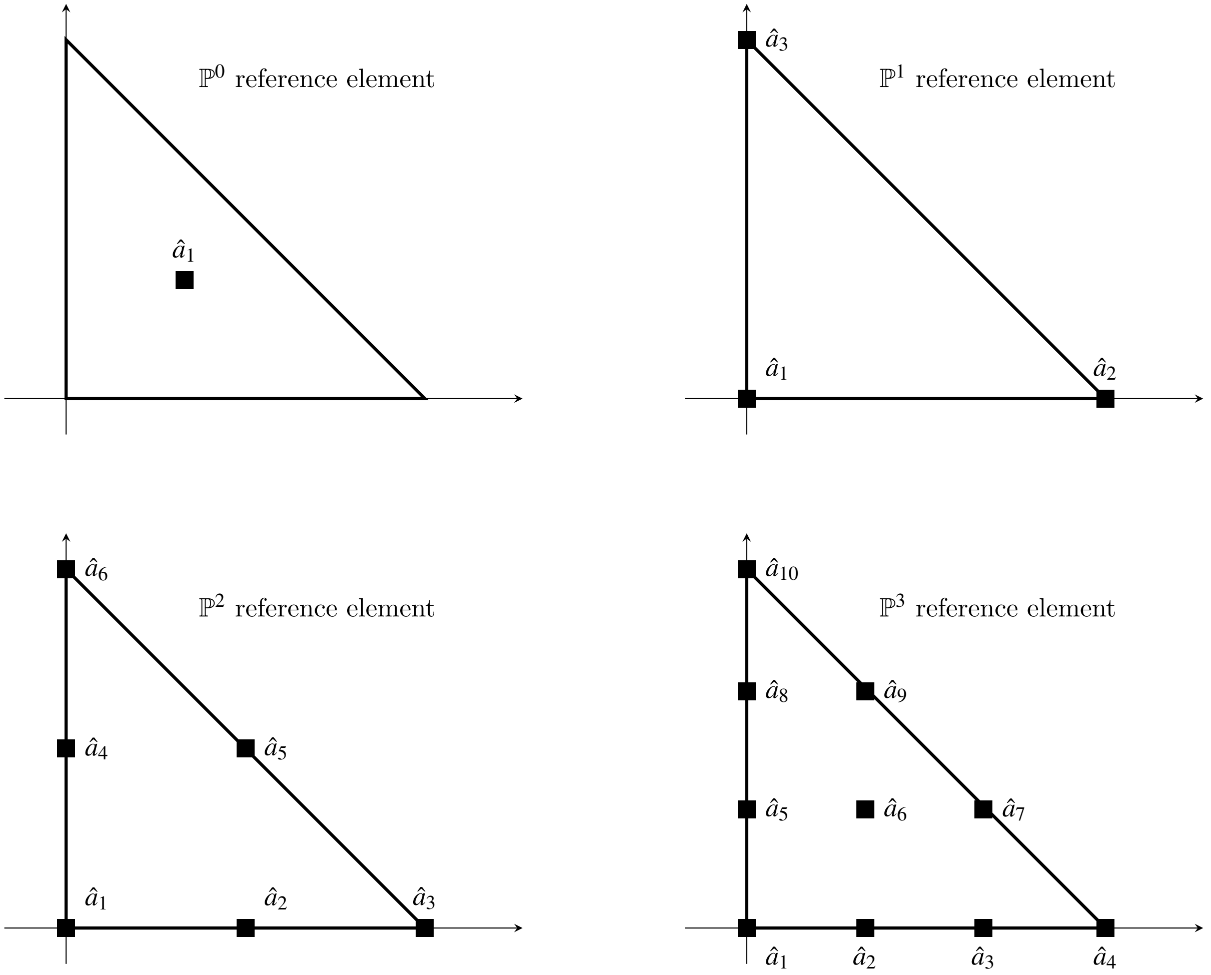


Figure 8. Nodes of the $\mathbb{P}^k$ finite elements in 2D

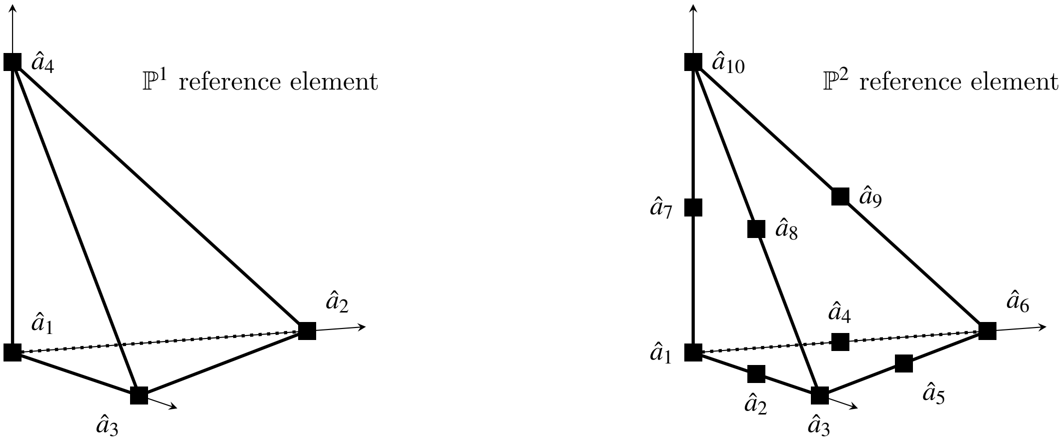


Figure 9. Nodes of the $\mathbb{P}^k$ finite elements in 3D

Definition 5.29 (Local interpolation operator) *We define the local interpolation operator*

$$\begin{array}{cccc} \mathscr{I}_0^k & : & C^0(\hat{K}) & \to & \mathbb{P}^k \\ & & \hat{v} & \mapsto & \mathscr{I}_0^k \hat{v} \end{array}$$

with

$$\mathscr{I}_0^k \hat{v}(\hat{x}) = \sum_{i=1}^{|\Sigma|} \hat{v}(\hat{a}_i)\theta_i(\hat{x}).$$

$\mathscr{I}_0^k \hat{v}$ is the unique polynomial in $\mathbb{P}^k$ which coincides with $\hat{v}$ on the nodes of the reference element.

Definition 5.30 (Global interpolation operator) *We define the global interpolation operator*

$$\begin{array}{cccc} \mathscr{I}_h^k & : & C^0(\bar{\Omega}) & \to & V_h \\ & & v & \mapsto & \mathscr{I}_h^k v \end{array}$$

with

$$\mathscr{I}_h^k v(x) = \sum_{a \in \Sigma_h} v(a)\phi_a(x).$$

$\mathscr{I}_h^k v$ is the unique function in V_h which coincides with v on the nodes of the discretization.

The operators are linked by the following formula:

Lemma 5.31

- *For any $K \in \mathscr{T}$,*

  $$(\mathscr{I}_h^k v)_{|K} \circ T_K = \mathscr{I}_0^k (v \circ T_K).$$

- *If $v \in C^0(\bar{\Omega})$, $v = 0$ on $\partial\Omega$, then $\mathscr{I}_h^1 v = 0$ on $\partial\Omega$. In particular $\mathscr{I}_h^1$ is an interpolation operator that maps $H_0^1() \cap C^0(\bar{\Omega})$ into $V_{h,0} = V_h \cap H_0^1(\Omega)$.*

Proof of Lemma 5.31.

- Notice that $(\mathscr{I}_h^k v)_{|K} \circ T_K$ and $\mathscr{I}_0^k (v \circ T_K)$ are two elements of $\mathbb{P}^k$ which coincide on the nodes of the reference element. By uniqueness of the Lagrange interpolation polynomial, the property is proved.
- The restriction of $\mathscr{I}_h^k v$ to a face F of the mesh only depends on the degrees of freedom associated to the nodes of face F. If the values of these degrees of freedom are zero, then $(\mathscr{I}_h^k v)_{|F} \equiv 0$. Thus the proof is concluded by taking $F \subset \partial\Omega$.

■

The interpolation operators have a major drawback: they apply to *continuous* functions only. But functions in $H^1(\Omega)$ are not necessarily continuous in dimension $d \geq 2$. Nevertheless for $d = 2$ or $d = 3$, $H^2(\Omega)$ is embedded in the space of continuous functions so that these operators may be used.

Analysis of the interpolation error. Let us start with two general tools.

Lemma 5.32 (Deny-Lions) *Let U a bounded Lipschitz domain of $\mathbb{R}^d$ and $k \in \mathbb{N}$. We denote $\mathbb{P}^k$ the set of polynomials of degree k on U. There exists a constant $C > 0$ such that*

$$\forall u \in H^{k+1}(\Omega), \quad \inf_{\pi \in \mathbb{P}^k} \|u - \pi\|_{H^{k+1}(U)} \leq C\,|u|_{H^{k+1}(U)}\,.$$

Proof of Lemma 5.32. For any multi-index $\alpha \in \mathbb{N}^d$ such that $|\alpha| \leq k$, we denote f_α the linear form defined on $H^{k+1}(U)$ as

$$f_\alpha(v) = \int_U \partial^\alpha v.$$

The dimension of the set of such multi-indices is exactly the dimension of $\mathbb{P}^k$.

- We first prove that the linear mapping

  $$F : \pi \in \mathbb{P}^k \mapsto (f_\alpha(\pi))_{|\alpha| \leq k} \in \mathbb{R}^{\dim(\mathbb{P}^k)} \tag{30}$$

  is bijective. Using the dimension argument, we just have to prove that it is injective. Assume that a non-zero polynomial $\pi \in \mathbb{P}^k$ is such that

$(f_\alpha(\pi))_{|\alpha|\leq k}=0$. We can find a non-zero monomial with maximal index in π. But then $\partial^\alpha \pi$ is a non-zero constant and $\int_U \partial^\alpha \pi$ cannot be zero...

- Let us prove that there exists $C>0$ such that, for all $u\in H^{k+1}(\Omega)$,

$$\|u\|_{H^{k+1}(U)} \leq C\left(|u|_{H^{k+1}(U)}+\sum_{|\alpha|\leq k}|f_\alpha(u)|\right). \quad (31)$$

Assume that the property is false. Then there exists a sequence $\{u_n\}$ of elements in $H^{k+1}(U)$ such that

$$\|u_n\|_{H^{k+1}(U)}=1, \quad (32)$$

$$|u_n|_{H^{k+1}(U)}+\sum_{|\alpha|\leq k}|f_\alpha(u_n)|\leq \frac{1}{n}. \quad (33)$$

By Eq. (32) we can extract a sequence $(u_{\phi(n)})_n$ which weakly converges in $H^{k+1}(U)$ to a function u. Moreover by compactness, $(u_{\phi(n)})_n$ strongly converges to u in $H^k(U)$ (see the Rellich-Kondrachov theorem). By Eq. (33) all the derivatives of order $k+1$ of $(u_{\phi(n)})_n$ tend to 0. This proves that the partial derivatives of u of order $k+1$ are zero, so that $u\in\mathbb{P}^k$. Moreover passing to the limit in $f_\alpha(u_{\phi(n)})$, we get: $f_\alpha(u)=0$ for all $|\alpha|\leq k$. From the first item, we deduce that $u=0$.

Since Eq. (32) leads to $\|u_{\phi(n)}\|_{H^k(U)}\to 1$, by strong convergence we have $\|u\|_{H^k(U)}=1$ which contradicts the property $u=0$.

- Let $u\in H^{k+1}(U)$. As the map F defined in (30) is surjective there exists $\tilde{\pi}\in\mathbb{P}^k$ such that

$$\forall |\alpha|\leq k, \quad f_\alpha(u-\tilde{\pi})=0.$$

Then we obtain

$$\begin{aligned}
&\inf_{\pi\in\mathbb{P}^k}\|u-\pi\|_{H^{k+1}(U)}\\
&\quad\leq \|u-\tilde{\pi}\|_{H^{k+1}(U)}\\
&\quad\leq C\left(|u-\tilde{\pi}|_{H^{k+1}(U)}+\sum_{|\alpha|\leq k}\underbrace{|f_\alpha(u-\tilde{\pi})|}_{=0}\right)\\
&\quad= C|u-\tilde{\pi}|_{H^{k+1}(U)}=C|u|_{H^{k+1}(U)}.
\end{aligned}$$

■

Lemma 5.33 (Bramble-Hilbert) *Let U be a bounded Lipschitz domain of $\mathbb{R}^d$, $k\in\mathbb{N}$ and Φ a continuous linear operator from $H^{k+1}(U)$ onto some normed vector space E. If $\Phi\equiv 0$ on $\mathbb{P}^k$ then there exists $C>0$ such that*

$$\forall u\in H^{k+1}(U), \quad \|\Phi u\|_E\leq C|u|_{H^{k+1}(U)}.$$

Proof of Lemma 5.33. For all $u\in H^{k+1}(U)$ and for all $\pi\in\mathbb{P}^k$, we have

$$\|\Phi u\|_E=\|\Phi(u-\pi)\|_E\leq\|\Phi\|\|u-\pi\|_{H^{k+1}(U)}.$$

Taking the infimum over π and using the Deny-Lions Lemma (see Lemma 5.32) we get

$$\|\Phi u\|_E\leq C|u|_{H^{k+1}(U)}.$$

■

Definition 5.34 (Regular mesh)

- *Let $\mathscr{T}$ be a mesh of Ω. We define*

$$\sigma_K=\frac{h_K}{\rho_K}, \qquad \sigma_{\mathscr{T}}=\sup_{K\in\mathscr{T}}\frac{h_K}{\rho_K}.$$

- *A family of meshes $(\mathscr{T}_h)_h$ of Ω is* regular *if there exists a constant $C>0$ such that*

$$\forall h>0, \quad \sigma_{\mathscr{T}_h}\leq C.$$

Remark 5.35 *The regularity property is equivalent to the following constraints:*

- *The volume of each cell is of order h_K^d:*

$$\exists C>0, \quad \forall h>0, \quad \forall K\in\mathscr{T}_h, \quad |K|\geq Ch_K^d.$$

- *The diameter of the incircle is uniformly bounded from below by h_K:*

$$\exists C>0, \quad \forall h>0, \quad \forall K\in\mathscr{T}_h, \quad \rho_K\geq Ch_K.$$

We deduce from the last two lemmas the following interpolation theorem:

Theorem 5.36 (Interpolation operator) *Let Ω be a bounded Lipschitz domain of $\mathbb{R}^d$, $k\in\mathbb{N}$. Let $0\leq m\leq k$ and we assume that $m+1>d/2$ so that $H^{m+1}(\Omega)\subset C^0(\bar{\Omega})$.*

- Local interpolation estimate. *For $l\leq m+1$, there exists $C>0$ such that, for all $K\in\mathscr{T}$, $\forall v\in H^{m+1}(K)$,*

$$\left|v-\mathscr{I}_h^k v\right|_{H^l(K)}\leq C\sigma_K^l h_K^{m+1-l}|v|_{H^{m+1}(K)}. \quad (34)$$

- Global interpolation estimate. *For $l\in\{0,1\}$, there exists $C>0$ such that, for all $v\in H^{m+1}(\Omega)$,*

$$\left|v-\mathscr{I}_h^k v\right|_{H^l(\Omega)}\leq C\sigma_T^l h^{m+1-l}|v|_{H^{m+1}(\Omega)}. \quad (35)$$

Let us recall that for $k=0$, $H^0=L^2$ and $|\cdot|_{H^0}=\|\cdot\|_{L^2}$.
Proof of Theorem 5.36.

- The global estimate, see Eq. (35), is deduced from local estimates (34) combined with Proposition 5.20, see page 36. The restriction $l \in \{0,1\}$ is essential here because functions in V_h are not in $H^l(\Omega)$ for $l \geq 2$. But the local estimate holds for $l \leq m+1$.
- Let us prove the local estimate for $l \leq m+1$. By the Bramble-Hilbert lemma, as the local interpolation operator $\mathscr{I}_0^k$ is the identity on $\mathbb{P}^k$, then the mapping
$$\Phi : \hat{v} \in H^{m+1}(\hat{K}) \mapsto \hat{v} - \mathscr{I}_0^k \hat{v} \in H^l(\hat{K})$$
is zero on $\mathbb{P}^k$. Moreover it is a *continuous* linear operator (because the embedding $H^{m+1}(\hat{K}) \subset C^0(\hat{K})$ is continuous). By the Bramble-Hilbert lemma, there exists $C > 0$ such that
$$\forall \hat{v} \in H^{m+1}(\hat{K}), \quad \|\hat{v} - \mathscr{I}_0^k \hat{v}\|_{H^l(\hat{K})} \leq C \, |\hat{v}|_{H^{m+1}(\hat{K})} .$$
Now we use Theorem 5.22, see page 38. Let $v \in H^{m+1}(K)$. We define $\hat{v} = v \circ T_K$ and we get
$$\begin{aligned} \left|v - \mathscr{I}_h^k v\right|_{H^l(K)} &\leq C \frac{|K|^{\frac{1}{2}}}{\rho_K^l} \left|v \circ T_K - \mathscr{I}_h^k v \circ T_K\right|_{H^l(\hat{K})} \\ &\leq C \frac{|K|^{\frac{1}{2}}}{\rho_K^l} \left|\hat{v} - \mathscr{I}_0^k v\right|_{H^l(\hat{K})} \\ &\leq C' \frac{|K|^{\frac{1}{2}}}{\rho_K^l} \, |\hat{v}|_{H^{m+1}(\hat{K})} \\ &\leq C'' \frac{h_K^{m+1}}{\rho_K^l} \, |v|_{H^{m+1}(K)} \\ &\leq C'' \sigma_{\mathscr{K}}^l h_K^{m+1-l} \, |v|_{H^{m+1}(K)} . \end{aligned}$$

■

The definition of a regular mesh ensures that the global interpolation estimate leads to an *optimal* approximation error estimate, see Eq. (35), i.e. as it is expected for the approximation space V_h built upon a regular triangulation. Thus we get the following error estimate.

Theorem 5.37 ($\mathbb{P}^k$ error estimate) *Let $a(\cdot,\cdot)$ be a bilinear form on $H_0^1(\Omega)$ which is continuous and coercive; L a continuous linear form on $H_0^1(\Omega)$. Let $(\mathscr{T}_h)_h$ be a regular family of simplicial meshes of Ω. Let V_h be the $\mathbb{P}^k$ Lagrange approximation space built upon these meshes and $V_{h,0} = V_h \cap H_0^1(\Omega)$. We assume that the unique $u \in H_0^1(\Omega)$ satisfying*
$$a(u,v) = L(v), \quad \forall v \in H_0^1(\Omega)$$
belongs to $H^{m+1}(\Omega)$ for some $m \leq k$. Denote u_h the solution of the approximate problem built upon $V_{h,0}$. There exists $C > 0$ which only depends on Ω, $a(\cdot,\cdot)$, $\sup_h(\sigma_{\mathscr{T}_h})$ such that
$$\|u - u_h\|_{H^1} \leq C h^m \, |u|_{H^{m+1}(\Omega)} .$$
If furthermore the adjoint problem admits an elliptic regularity property, then
$$\|u - u_h\|_{L^2} \leq C h^{m+1} \, |u|_{H^{m+1}(\Omega)} .$$

Proof of Theorem 5.37. In the case $m+1 \geq d/2$, the theorem is a consequence of the properties of the Lagrange interpolation operator, see Theorem 5.36 with $l = 1$. Note that the norms $|\cdot|_{H^1(\Omega)}$ and $\|\cdot\|_{H^1(\Omega)}$ are equivalent since $u - u_h \in H_0^1(\Omega)$.

In the case $m+1 < d/2$ it is necessary to build other interpolation operators which overcome the difficulties related to the continuity argument. Among operators that have stability and interpolation properties, let us mention the Clément operator [7] (which fails at preserving the Dirichlet boundary conditions) and the Scott-Zhang interpolation operator [16] (which preserves homogeneous boundary conditions). The construction is more intricate and is not given here and we refer the reader to [9] for an overview on this issue. ■

5.4 Other classical finite elements

The $\mathbb{Q}^k$ Lagrange finite element. Considering a mesh of Ω which is made of quadrilaterals in 2D, parallelepipeds in 3D. What kind of finite elements is it possible to build?

We cannot hope for a degree of freedom on each vertex of the mesh with piecewise $\mathbb{P}^1$ functions. Thus it is necessary to enlarge the space of polynomials that should be considered:
$$\begin{aligned} \mathbb{Q}^k(\mathbb{R}^d) &= \left\{ u = \sum_{\alpha \in \mathbb{N}^d,\ \sup \alpha_i \leq k} a_\alpha x_1^{\alpha_1} \ldots x_d^{\alpha_d} \right\} \\ &= \mathbb{P}^k(\mathbb{R}) \otimes \ldots \otimes \mathbb{P}^k(\mathbb{R}), \end{aligned}$$
which are the polynomials with all partial degrees less than k. Note that, in 1D, $\mathbb{Q}^k = \mathbb{P}^k$ for any k.

The $\mathbb{Q}^1$ element is defined by 4 coefficients in 2D, 8 coefficients in 3D. This corresponds to the number of vertices for a quadrilateral in 2D, parallelepiped in 3D. More precisely,
$$\mathbb{Q}^1(\mathbb{R}^2) = \mathrm{span}(1, x_1, x_2, x_1x_2),$$
$$\mathbb{Q}^1(\mathbb{R}^3) = \mathrm{span}(1, x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1x_2x_3).$$
We consider the reference element $\hat{K} = [0,1]^d$ as the unit cube in $\mathbb{R}^d$.

Remark 5.38 *Notice that the restriction of a function in $\mathbb{Q}^k$ to an hyperplane parallel to the axes is a function in $\mathbb{P}^k$.*

Proposition 5.39 ($\mathbb{Q}^k$ Lagrange finite element) *We consider $\Sigma \subset \hat{K}$ the set of points defined as*
$$\left(\frac{I_1}{k}, \ldots, \frac{I_d}{k}\right), \quad \forall (I_1, \ldots, I_d) \in \mathbb{N}^d, \quad \forall j \in \{1, \ldots, d\},\ I_j \leq k.$$
Then $|\Sigma| = (k+1)^d$.

We denote $(\hat{a}_j)_{1 \leq j \leq |\Sigma|}$ the elements of this set. Then the mapping $p \in \mathbb{Q}^k \mapsto (p(\hat{a}_j))_{1 \leq j \leq |\Sigma|} \in \mathbb{R}^{|\Sigma|}$ is an isomorphism (in particular $\mathbb{Q}^k$ and Σ have the same cardinality).

- *the triplet* $(\hat{K},\mathbb{Q}^k,\Sigma)$ *is the so-called* $\mathbb{Q}^k$ Lagrange finite element.
- *the set* $(\hat{a}_j)_j$ *is the set of nodes of this finite element and the linear form* $v \in C^0(\hat{K}) \mapsto v(\hat{a}_j)$ *is denoted the* degree of freedom *associated to* $\hat{a}_j$.
- *For all* $1 \le i \le |\Sigma|$ *there exists a unique function* $\theta_i \in \mathbb{Q}^k$ *such that*

  $$\theta_i(\hat{a}_j) = \delta_{ij}, \quad \forall 1 \le j \le |\Sigma|.$$

  These functions are the so-called local shape functions.

Example. In 2D or 3D, the nodes of the $\mathbb{Q}^1$ finite elements are described in Figure 10.

□

The notion of regularity of the mesh is slightly modified: in particular for a quadrilateral, the diameter of the incircle for a triangle is replaced by

$$\rho_K = \min(\rho_{T_1},\rho_{T_2},\rho_{T_3},\rho_{T_4})$$

where T_i denote the four triangles obtained from the four vertices of the quadrilateral.

Let us consider a mapping $T_K : \hat{K} \to K$ such that

$$T_K(\hat{a}_j) = a_j, \quad \forall j \in \{1,\dots|\Sigma|\}.$$

The mapping T_K *cannot* be *affine*, except if K is a parallelogram. In the general case, the mapping is quadratic: $T_K \in (\mathbb{Q}^1)^d$. In particular the Jacobian matrix is not constant anymore.

Thus the approximation space is defined as

$$V_h = \{v \in C^0(\bar{\Omega}),\ \forall K \in \mathscr{T},\ v \circ T_K \in \mathbb{Q}^k\},$$

which is *not* equivalent to the condition "$v_{|K} \in \mathbb{Q}^k$, for all K", except if the cells are parallelograms (in which case T_K is affine). The interpolation operator is defined as previously. Moreover under regularity assumptions on the mesh we obtain the following interpolation result:

Proposition 5.40 *For a family of regular quadrilateral meshes, there exists a constant* $C > 0$ *such that, for all* $0 \le m \le k$ *and for all* $0 \le p \le m$,

$$\forall v \in H^{m+1}(\Omega), \quad \|v - \mathscr{I}_h^k v\|_{H^p(\Omega)} \le C_\sigma h^{m-p+1}\,|v|_{H^{m+1}(\Omega)}.$$

A conforming L^2 element: the discontinuous element $\mathbb{P}^0 = \mathbb{Q}^0$. We note $\mathbb{P}^0 = \mathbb{Q}^0$ the set of constant functions. Then, from a quadrilateral or simplicial mesh, we can easily build the set of piecewise constant functions:

$$V_h = \{u \in L^2(\Omega),\ u_{|K_i} \in \mathbb{P}^0\}.$$

This space is conforming in $L^2(\Omega)$ (but not in $H^1(\Omega)$).

In the case of regular functions, the degree of freedom in a cell is the value of the function at the center of mass of the cell. It is possible to define the Lagrange interpolation operator $\mathscr{I}_h^0$ as done previously[10] and then get suitable estimates. However, for this particular approximation space, there exists an interpolation operator which is much simpler and which applies to any function in $L^2(\Omega)$.

Definition 5.41 *For any function* $v \in V$, *we define a piecewise constant function* $\mathscr{I}_h^0 v$ *whose value on each cell* K *is* $|K|^{-1}\int_K v$:

$$\begin{array}{rcccl} \mathscr{I}_h^0 & : & V & \to & V_h \\ & & v & \mapsto & \mathscr{I}_h^0 v := \displaystyle\sum_{K\in\Sigma} \frac{\int_K v}{|K|}\mathbf{1}_K. \end{array}$$

Proposition 5.42 *The operator* $\mathscr{I}_h^0$ *is the* L^2*-orthogonal projection on* V_h. *Moreover (without regularity assumption on the mesh, provided that the cells are convex), for all* $0 \le m \le 1$, *there exists* $C > 1$ *such that*

$$\forall v \in H^m(\Omega), \quad \|v - \mathscr{I}_h^0 v\|_{L^2(\Omega)} \le C h^m\,|v|_{H^m(\Omega)}.$$

This finite element is often used in the discretization of the pressure in the Stokes problem.

6. Finite elements for saddle-point problems

In the previous section, finite element spaces were considered for coercive problems. However when considering non-coercive problems such as saddle-point problems, it is necessary to ensure the compatibility of the approximation spaces X_h and M_h, in the sense that a uniform inf-sup inequality has to be satisfied.

As already pointed out, in the finite dimensional framework, positivity of the inf-sup constant is equivalent to the injectivity of the operator $B_h' : M_h \to X_h'$ (which, in the case of the Stokes problem, is nothing but the *discrete pressure gradient*), and thus to the well-posedness of the discrete problem. In that case, we can define

$$\beta_h = \inf_{q_h\in M_h}\left(\sup_{v_h\in X_h} \frac{b(v_h,q_h)}{\|v_h\|_X\|q_h\|_M}\right) > 0.$$

[10] For instance, in 1D, the space V_h writes

$$V_h = \{u \in L^2(\Omega),\ u_{|K_i} \in \mathbb{P}^0\},$$

and the mapping

$$\Phi : u \mapsto (u(x_{\frac{1}{2}}), u(x_{\frac{3}{2}}), \dots, u(x_{N+\frac{1}{2}})))$$

is an isomorphism from V_h onto $\mathbb{R}^{N+1}$. In particular $\dim(V_h) = N+1$. Then the interpolation operator $\mathscr{I}_h^0$ from V onto V_h is defined as

$$\forall u \in V, \quad \forall x \in \Omega, \quad \mathscr{I}_h^0 u(x) = \sum_{i=0}^{N} u(x_{i+\frac{1}{2}})\,\phi_{i+\frac{1}{2}}(x).$$

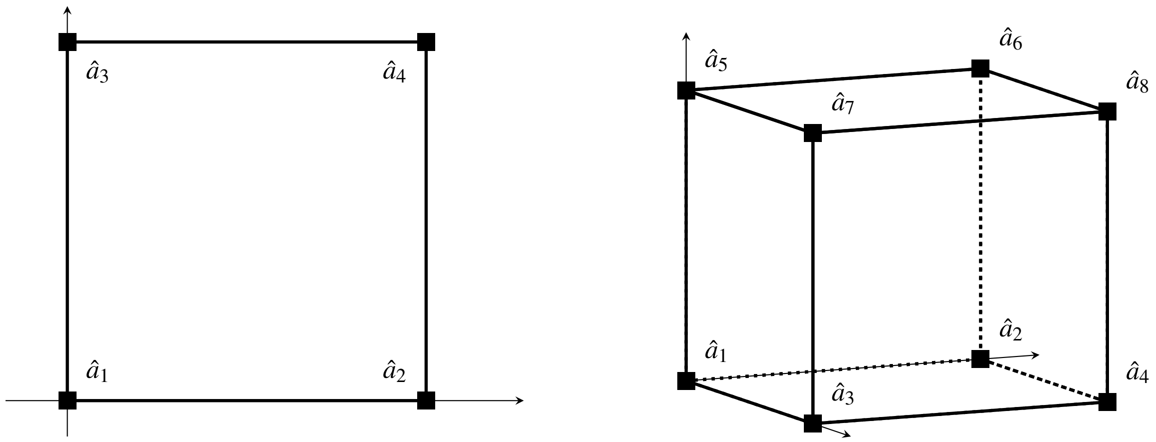


Figure 10. Nodes of the $\mathbb{Q}^1$ finite elements: in 2D and 3D.

Thus we need to be able to prove that B_h' is injective on the one hand and that β_h does not go to 0 as h goes to 0 on the other hand. This last property is called the *inf-sup stability* of the numerical method.

6.1 Fortin's lemma

We often consider *conforming* approximation spaces for X and M which satisfy the inf-sup condition. In that case the validity of a uniform discrete inf-sup condition is given by the following result:

Lemma 6.1 (Fortin) *Let $b : X \times M \to \mathbb{R}$ be a continuous bilinear form and $X_h \subset X$, $M_h \subset M$ two finite dimensional subspaces. We assume that b satisfies the inf-sup condition on $X \times M$:*

$$\exists \beta > 0, \quad \inf_{q \in M} \left(\sup_{v \in X} \frac{b(v,q)}{\|v\|_X \|q\|_M} \right) \geq \beta.$$

Then b satisfies a uniform inf-sup condition on $X_h \times M_h$ if, and only if, there exists a continuous linear operator $\Pi_h : X \to X_h$ and a $C > 0$, independent of h, such that

$$\forall v \in X, \quad \|\Pi_h v\|_X \leq C \|v\|_X$$

and

$$\forall q_h \in M_h, \quad b(v, q_h) = b(\Pi_h v, q_h).$$

In the literature, the operator Π_h is usually called the Fortin operator*.*

Proof of Lemma 6.1. Let us proceed in two steps.

- *Assume that there exists such a continuous linear operator $\Pi_h : X \to X_h$.* Then, for all $q_h \in M_h$, we have

$$\begin{aligned} \beta \|q_h\|_M &\leq \sup_{v \in X} \frac{b(v,q_h)}{\|v\|_X} = \sup_{v \in X} \frac{b(\Pi_h v, q_h)}{\|v\|_X} \\ &\leq C \sup_{v \in X} \frac{b(\Pi_h v, q_h)}{\|\Pi_h v\|_X} \\ &\leq C \sup_{v_h \in X_h} \frac{b(v_h, q_h)}{\|v_h\|_X}, \end{aligned}$$

which states the uniform inf-sup condition for b on $X_h \times M_h$ with a constant $\frac{\beta}{C}$.

- *Assume that the uniform inf-sup condition is satisfied with a constant $\bar{\beta}$.* Then there exists a unique $(u_h, q_h) \in X_h \times M_h$ such that

$$\left\{ \begin{aligned} (u_h, v_h) + b(v_h, p_h) &= 0, && \forall v_h \in X_h, \\ b(u_h, q_h) &= G(q_h), && \forall q_h \in M_h, \end{aligned} \right.$$

and (u_h, q_h) continuously depends on G with the estimate (see Theorem 3.13)

$$\|u_h\|_X \leq \frac{2}{\bar{\beta}} \|G\|_{M'}.$$

Let $v \in X$. We choose the linear form G as

$$\forall q \in M, \quad G(q) = b(v,q),$$

whose norm is bounded by $\|b\| \|v\|_X$. Then we find an element $u_h \in X_h$ which linearly depends on v, such that

$$\|u_h\|_X \leq \frac{2\|b\|}{\bar{\beta}} \|v\|_X,$$

and

$$b(u_h, q_h) = b(v, q_h), \quad \forall q_h \in M_h.$$

Thus the operator $\Pi_h : v \mapsto u_h$ satisfies the desired properties.

■

6.2 Stokes problem

For the sake of simplicity, we consider only here the 2D case. We review the most classical conforming finite element for the Stokes problem.

We shall consider the unit square $\Omega =]0,1[^2$ equipped with either a uniform rectangular mesh or a uniform triangular (simplicial) mesh obtained by dividing all the rectangles into two isometric parts.

The geometry of all the finite elements that we propose to analyse here is summarized in Figure 11.

6.2.1 The $\mathbb{P}^1 - \mathbb{P}^0$ element

Consider now the uniform triangle mesh of Ω. We consider the following approximation spaces

$$X_h = \{v \in (H_0^1(\Omega))^2,\ \forall K \in \mathscr{T},\ v \circ T_K \in (\mathbb{P}_1)^2\},$$

$$M_h = \{v \in L_0^2(\Omega),\ \forall K \in \mathscr{T},\ v \circ T_K \in \mathbb{P}_0\}.$$

Observe that the degrees of freedom for the velocity field are exactly the same as for the $\mathbb{Q}^1$ element. However, the approximation space X_h is not the same as before, since the fields in X_h are linear on each triangle whereas the in the previous section the fields in X_h were bilinear on each rectangle.

We also observe that, for a same number of vertices in the mesh, the number of degrees of freedom for the pressure is twice the one of the previous section.

Proposition 6.2 ($\mathbb{P}^1 - \mathbb{P}^0$ is unstable) *The $\mathbb{P}^1 - \mathbb{P}^0$ finite elements are not inf-sup stable for the Stokes problem.*

Proof of Proposition 6.2. Let us remark that

- there are $2NM$ elements and thus $\mathrm{d(p)} = 2NM - 1$;
- there are $(N+1)(M+1)$ nodes, including $(N-1)(M-1)$ interior nodes, hence $\mathrm{d(v)} = 2(N-1)(M-1)$.

Then

$$\begin{array}{rcl} \mathrm{d(p)} - \mathrm{d(v)} & = & (2NM-1) - (2(N-1)(M-1)) \\ & = & 2(N+M) - 3 \\ & > & 0 \end{array}$$

As $\mathrm{d(p)} > \mathrm{d(v)}$ the kernel of B'_h is not reduced to 0. In fact, it means that the kernel of B'_h is very large! This choice of approximation spaces is thus far from being inf-sup stable.

We may prove that the kernel of B_h (which contains the functions with discrete free divergence) is $\{0\}$, i.e. the only velocity field $v_h \in X_h$ which is likely to be a solution of the system is the null function.

By definition, $v_h \in \mathrm{Ker}(B_h)$ if

$$\int_\Omega \mathrm{div}(v_h)\, q_h = 0, \quad \forall q_h \in M_h. \tag{36}$$

Since v_h is 0 at the boundary, we have by the Stokes formula (see Proposition 1.26, page 6),

$$\int_\Omega \mathrm{div}(v_h)\, \bar{q} = 0, \quad \forall \bar{q} \in \mathbb{R}.$$

Thus Eq. (36) also holds for piecewise constant test functions with non-zero mean values. In practice we thus have a condition which writes

$$\int_K \mathrm{div}(v_h) = 0, \quad \forall K \in \mathscr{T}.$$

Since $v_h \in (\mathbb{P}^1)^2$ on each cell K, v_h takes the form $v_h(x,y) = \left(a_1^K + b_1^K x + c_1^K y, a_2^K + b_2^K x + c_2^K y\right)$ on K, hence $\mathrm{div}(v_h) = b_1^K + c_2^K$ is a constant. As a consequence $v_h \in \mathrm{Ker}(B_h)$ if, and only if, $\mathrm{div}(v_h) = 0$ in a distribution sense (this property does not hold for any discretization of the pressure!).

Let us now investigate the consequences on v_h. Let K be a triangle with nodes that are numbered 1, 2 and 3, and let $v \in (\mathbb{P}^1)^2$ on K, see Figure 12. By the Stokes formula, the divergence free condition on K writes

$$\sum_\sigma \int_\sigma (v \cdot n) = 0.$$

As v is affine, the integral along the edge is equal to the value of the function at the middle of the edge (weighted by the length of the edge), which is equal to the half-sum of the degrees of freedom of the corresponding nodes. Denoting m_i and n_i the measure and the outward normal unit vector of the edge at the opposite of node i, it yields

$$m_3 \frac{v_1 + v_2}{2} \cdot n_3 + m_2 \frac{v_1 + v_3}{2} \cdot n_2 + m_1 \frac{v_2 + v_3}{2} \cdot n_1 = 0$$

that is to say

$$\begin{aligned} v_1 \cdot (m_2 n_2 + m_3 n_3) + v_2 \cdot (m_1 n_1 + m_3 n_3) \\ + v_3 \cdot (m_1 n_1 + m_2 n_2) = 0. \end{aligned}$$

Applying the Stokes formula to constant vector fields in the triangle, we find that the following equality is satisfied

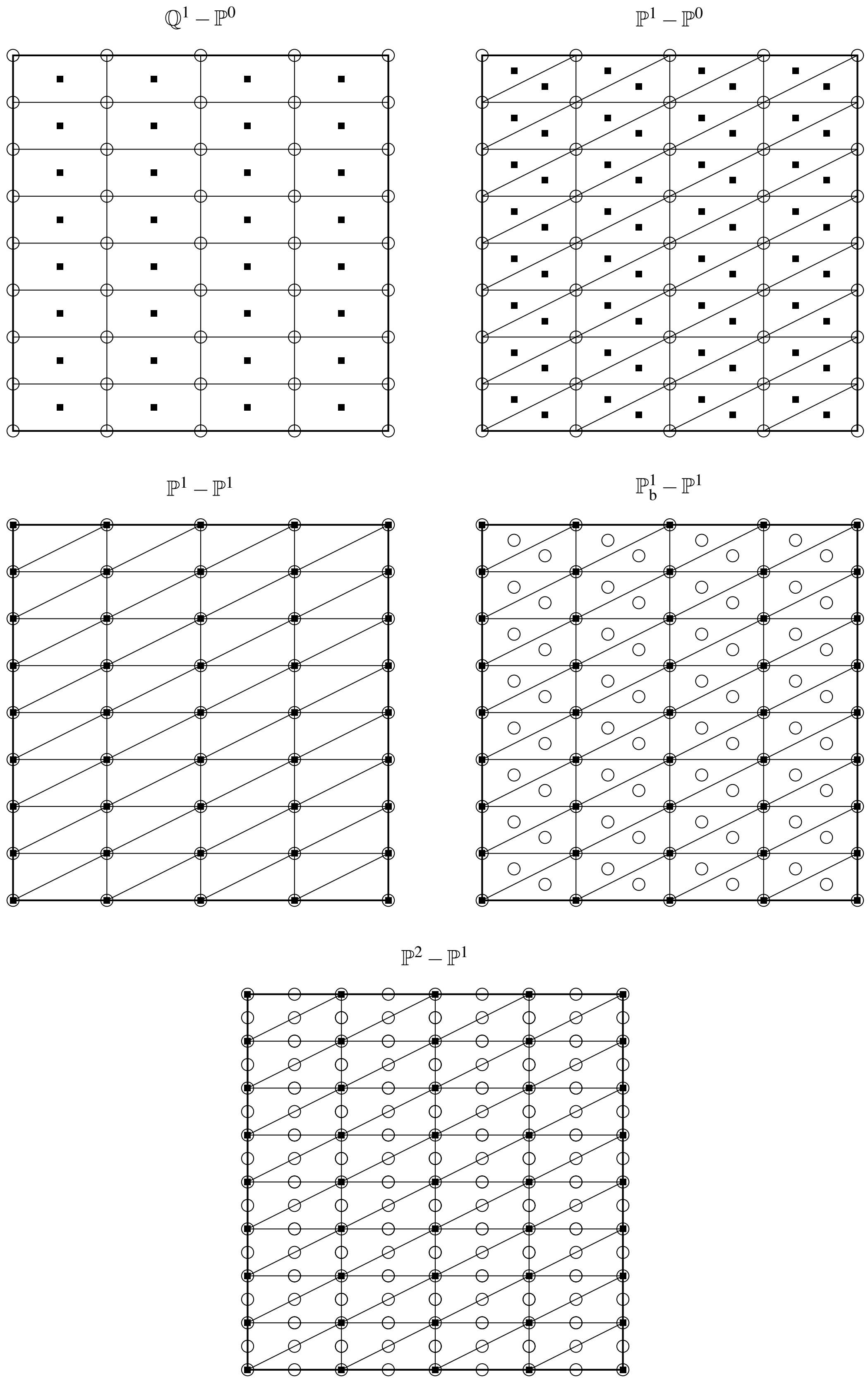
$$m_1 n_1 + m_2 n_2 + m_3 n_3 = 0,$$

which leads us back to the free divergence equation which becomes

$$m_1 v_1 \cdot n_1 + m_2 v_2 \cdot n_2 + m_3 v_3 \cdot n_3 = 0.$$

It means in particular that if $v_1 = v_2 = 0$ then v_3 and n_3 are orthogonal. Thus if K is a triangle with an edge at the boundary, then only one degree of freedom is at the *interior* of the domain: the orientation of the velocity field at this node is parallel to the boundary; if the same degree of freedom is the interior node of another triangle with an edge at the boundary, we obtain another orthogonality condition which proves that the velocity is zero.

This reasoning applies to structured triangular meshes and proves the so-called *locking effect*. ■



$\mathbb{Q}^1 - \mathbb{P}^0$

$\mathbb{P}^1 - \mathbb{P}^0$

$\mathbb{P}^1 - \mathbb{P}^1$

$\mathbb{P}^1_b - \mathbb{P}^1$

$\mathbb{P}^2 - \mathbb{P}^1$

Figure 11. Comparison of various elements for the Stokes problem on a square domain. Degrees of freedom for velocity (○) and pressure (▪)

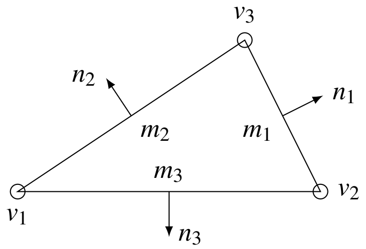


Figure 12. Notations in a triangle element. Three degrees of freedom for the velocity field.

6.2.2 The $\mathbb{Q}^1 - \mathbb{P}^0$ element

Here we consider the uniform cartesian mesh. The cells are thus denoted by K_{ij} with $i \in \{1,...,N\}$ and $j \in \{1,...,M\}$.

To simplify some computations, we assume that N and M are even and we set $N = 2n$, $M = 2m$. The mesh size in the x direction is denoted by $h = 1/N$ and the one in the y direction is denoted by $k = 1/M$. We introduce the functional spaces

$$X_h = \{v \in (H_0^1(\Omega))^2,\ \forall K \in \mathscr{T},\ v \circ T_K \in (\mathbb{Q}^1)^2\},$$

$$M_h = \{v \in L_0^2(\Omega),\ \forall K \in \mathscr{T},\ v_{|K} \in \mathbb{P}^0\},$$

The set of degrees of freedom for such discretization spaces is represented in Figure 11.

This element is the most simple that one can think of on a Cartesian grid. However, we will show that the discrete inf-sup condition is not satisfied by this finite element.

Proposition 6.3 ($\mathbb{Q}^1 - \mathbb{P}^0$ is unstable) *The $\mathbb{Q}^1 - \mathbb{P}^0$ finite elements are not inf-sup stable for the Stokes problem.*

Remark 6.4 *As it was already pointed out, we consider the restriction B_h' of operator $B : M \to X'$ as an operator from M_h onto X_h'. Then the inf-sup condition states that this operator should be injective. In particular,* the dimension of M_h should be lower than the dimension of X_h. *Thus,* the inf-sup condition is not satisfied if M_h is too big with respect to X_h.

Let us denote d(p) *(resp.* d(v)*) the number of degrees of freedom for the pressure (resp. for the velocity). Let us remark that*

- *there are NM elements and thus* $\mathrm{d(p)} = NM - 1$ *(as the mean value of the pressure is zero);*
- *there are $(N+1)(M+1)$ nodes, including $(N-1)(M-1)$ interior nodes (that are the only ones to be considered as velocity degrees of freedom because of the homogeneous Dirichlet conditions). Hence, since the velocity is a two-dimensional vector field, we have* $\mathrm{d(v)} = 2(N-1)(M-1)$ *in 2D.*

Then by the rank theorem, the dimension of the kernel of B_h' is at least

$$\begin{aligned} \mathrm{d(p)} - \mathrm{d(v)} &= (NM-1) - (2(N-1)(M-1)) \\ &= -NM + 2(N+M) \end{aligned}$$

which is negative (except for coarse meshes). Thus $\mathrm{d(p)} < \mathrm{d(v)}$ *and considerations on the size of the approximation spaces are not sufficient to conclude: the kernel of B_h' has to be investigated in details in order to show that the inf-sup condition is not satisfied.*

Proof of Proposition 6.3. We recall that the operator B_h' is defined by

$$\begin{array}{cccc} B_h' & : & M_h & \to & X_h' \\ & & p_h & \mapsto & B_h' p_h = b(\cdot, p_h) \end{array}$$

i.e.[11]

$$\forall v_h \in X_h, \quad \left\langle B_h' p_h, v_h \right\rangle_{X_h', X_h} = \int_\Omega \mathrm{div}(v_h)\, p_h$$

and we aim at proving that B_h' is *not* injective. More precisely, we shall prove that

$$\dim(\mathrm{Ker}(B_h')) = 1.$$

The velocity field v is described by its two components $v := (v^{(1)}, v^{(2)})$. We denote p_{ij} the discrete pressure in each element and $v^{(j)}_{i-\frac{1}{2}\,j-\frac{1}{2}}$, for $j \in \{1,2\}$, the components of the discrete velocity at node $(i-\frac{1}{2}, j-\frac{1}{2})$. We can compute

$$\int_\Omega \mathrm{div}(v_h)\, p_h = \sum_{i=1}^{N} \sum_{j=1}^{M} p_{ij} \left(\int_{K_{ij}} \mathrm{div}(v_h) \right).$$

Each cell has four vertices that we locally number counterclockwise starting from the lower left corner (from 1 to 4). The edges are also numbered:

- $\partial K_{ij}^{(1)}$ denotes the edge 1 → 2,
- $\partial K_{ij}^{(2)}$ denotes the edge 2 → 3,
- $\partial K_{ij}^{(3)}$ denotes the edge 3 → 4,
- $\partial K_{ij}^{(4)}$ denotes the edge 4 → 1.

[11] By definition, we have, for all $v_h \in X_h$,

$$\begin{aligned} \langle B_h' p_h, v_h \rangle_{X_h', X_h} &= \langle B' p_h, v_h \rangle_{X', X} \\ &= \langle B v_h, p_h \rangle_{M', M} \\ &= b(v_h, p_h) \\ &= \textstyle\int_\Omega \mathrm{div}(v_h)\, p_h. \end{aligned}$$

We have

$$\int_{K_{ij}} \mathrm{div}(v_h) = \int_{\partial K_{ij}} v_h \cdot n = \sum_{\ell=1}^{4} \int_{\partial K_{ij}^{(\ell)}} v_h \cdot n.$$

Let us compute the first term:

$$\begin{aligned}\int_{\partial K_{ij}^{(1)}} v_h \cdot n &= \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \left(\begin{array}{c} v_h^{(1)}\left(x, y_{j-\frac{1}{2}}\right) \\ v_h^{(2)}\left(x, y_{j-\frac{1}{2}}\right) \end{array} \right) \cdot \left(\begin{array}{c} 0 \\ -1 \end{array} \right) \mathrm{d}x \\ &= -\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} v_h^{(2)}(x, y_{j-\frac{1}{2}})\,\mathrm{d}x.\end{aligned}$$

The function $v_h(\cdot, y_{j-\frac{1}{2}})$ is linear on $\partial K_{ij}^{(1)}$ (see Remark 5.38) so that the (exact) trapezoidal rule yields

$$\int_{\partial K_{ij}^{(1)}} v_h \cdot n = -\frac{h}{2}\left(v^{(2)}_{i-\frac{1}{2}\, j-\frac{1}{2}} + v^{(2)}_{i+\frac{1}{2}\, j-\frac{1}{2}} \right).$$

In the same way the other terms can be computed as well:

$$\int_{\partial K_{ij}^{(2)}} v_h \cdot n = +\frac{k}{2}\left(v^{(1)}_{i+\frac{1}{2}\, j-\frac{1}{2}} + v^{(1)}_{i+\frac{1}{2}\, j+\frac{1}{2}} \right),$$

$$\int_{\partial K_{ij}^{(3)}} v_h \cdot n = +\frac{h}{2}\left(v^{(2)}_{i-\frac{1}{2}\, j+\frac{1}{2}} + v^{(2)}_{i+\frac{1}{2}\, j+\frac{1}{2}} \right),$$

$$\int_{\partial K_{ij}^{(4)}} v_h \cdot n = -\frac{k}{2}\left(v^{(1)}_{i-\frac{1}{2}\, j-\frac{1}{2}} + v^{(1)}_{i-\frac{1}{2}\, j+\frac{1}{2}} \right),$$

We finally obtain:

$$\begin{aligned}&\int_{K_{ij}} \mathrm{div}(v_h) \\ &= \frac{hk}{2}\left\{ \frac{v^{(2)}_{i-\frac{1}{2}\, j+\frac{1}{2}} - v^{(2)}_{i-\frac{1}{2}\, j-\frac{1}{2}}}{k} + \frac{v^{(2)}_{i+\frac{1}{2}\, j+\frac{1}{2}} - v^{(2)}_{i+\frac{1}{2}\, j-\frac{1}{2}}}{k} \right. \\ &\qquad \left. + \frac{v^{(1)}_{i+\frac{1}{2}\, j+\frac{1}{2}} - v^{(1)}_{i-\frac{1}{2}\, j+\frac{1}{2}}}{h} + \frac{v^{(1)}_{i+\frac{1}{2}\, j-\frac{1}{2}} - v^{(1)}_{i-\frac{1}{2}\, j-\frac{1}{2}}}{h} \right\}\end{aligned}$$

As $v_h = 0$ on the boundary, by a discrete integration by parts, we have

$$\begin{aligned} b(v_h, p_h) &= -\sum_{i,j} hk v_{i-\frac{1}{2}\, j+\frac{1}{2}} (\delta_y p)_{i-\frac{1}{2}\, j-\frac{1}{2}} \\ &\quad -\sum_{i,j} hk u_{i-\frac{1}{2}\, j-\frac{1}{2}} (\delta_x p)_{i-\frac{1}{2}\, j-\frac{1}{2}} \end{aligned}$$

with

$$(\delta_x p)_{i-\frac{1}{2}\, j-\frac{1}{2}} = \frac{1}{2}\left(\frac{p_{i\,j} - p_{i-1\,j}}{h} + \frac{p_{i\,j-1} - p_{i-1\,j-1}}{h} \right),$$

$$(\delta_y p)_{i-\frac{1}{2}\, j-\frac{1}{2}} = \frac{1}{2}\left(\frac{p_{i\,j} - p_{i\,j-1}}{k} + \frac{p_{i-1\,j} - p_{i-1\,j-1}}{k} \right).$$

Thus, $b(v_h, p_h) = 0$ for all $v_h \in X_h$ if, and only if, $\delta_x p = \delta_y p = 0$, i.e.

$$\begin{aligned} p_{i\,j} &= p_{i-1\,j-1}, \quad \forall i, j \\ p_{i-1\,j} &= p_{i\,j-1}, \quad \forall i, j. \end{aligned}$$

As $p_h \in M_h$, the mean value of p_h is zero, hence

$$p_{i\,j} = -p_{i-1\,j}, \forall i, j.$$

We thus obtain a *spurious mode* which is highly oscillating. This so-called *checkerboard mode* exactly generates the kernel of $B'_h = \mathrm{span}(\psi_h)$ with

$$\psi_h = \sum_i \sum_j \mathbf{1}_{K_{ik}} (-1)^{i+j}.$$

■

One may hope to recover the stability by replacing M_h by the space $\tilde{M}_h$ obtained from M_h by "removing" span(ψ_h), but we will see that it is not the case. More precisely, if we define $\tilde{M}_h$ as the orthogonal complement of span(ψ_h) in M_h, then we can prove that X_h and $\tilde{M}_h$ satisfy the discrete inf-sup condition but unfortunately not in a uniform way with respect to h.

Proposition 6.5 *Define*

$$\tilde{M}_h := \psi_h^{\perp} = \{p_h \in M_h,\ (p_h, \psi_h)_{L^2} = 0\},$$

where ψ_h denotes the spurious mode of the $\mathbb{Q}^1 - \mathbb{P}^0$ finite element. For all $h > 0$, we have

$$\beta_h := \inf_{p_h \in \tilde{M}_h} \sup_{v_h \in X_h} \frac{b(v_h, p_h)}{\|v_h\|_{H^1} \|p_h\|_{L^2}} > 0.$$

Moreover there exists $C_1 > 0$ and $C_2 > 0$ such that

$$C_1 h \leq \beta_h \leq C_2 h.$$

Proof of Proposition 6.5. The positivity of β_h follows from the construction of $\tilde{M}_h$: the operator

$$\begin{array}{cccc} B'_h & : & \tilde{M}_h & \rightarrow & X'_h \\ & & p_h & \mapsto & B'_h p_h = b(\cdot, p_h) \end{array}$$

is injective by means of construction which, in the finite dimensional framework, guarantees that $\beta_h > 0$.

Let us prove the most interesting inequality. For this we construct a function q_h which satisfies suitable properties:

$$q_h = \sum_{i=1}^{2n} \sum_{j=1}^{2m} \mathbf{1}_{K_{ij}} (-1)^{i+j} \left(\mathscr{E}\left(\frac{i-1}{2}\right) - \frac{n-1}{2} \right),$$

where $\mathscr{E}$ denotes the floor function. By means of construction the mean value of q_h is zero. Moreover,

$$\begin{aligned}\int_\Omega q_h \psi_h &= hk\sum_{ij}\left(\mathscr{E}\left(\frac{i-1}{2}\right)-\frac{n-1}{2}\right)\\ &= hkM\sum_{i=1}^{2n}\left(\mathscr{E}\left(\frac{i-1}{2}\right)-\frac{n-1}{2}\right)\\ &= 2hkM\sum_{p=0}^{n-1}\left(p-\frac{n-1}{2}\right)\\ &= 0.\end{aligned}$$

As a consequence, $q_h \in \tilde{M}_h$.

- *Computation of* $\|q_h\|_{L^2(\Omega)}$.

  $$\begin{aligned}\|q_h\|^2_{L^2(\Omega)} &= hk\textstyle\sum_{i,j}\left(\mathscr{E}\left(\frac{i-1}{2}\right)-\frac{n-1}{2}\right)^2\\ &= 2hkM\textstyle\sum_{p=0}^{n-1}\left(p-\frac{n-1}{2}\right)^2\\ &\sim Chn^3\\ &\sim Ch^{-2}.\end{aligned}$$

  Thus $\|q_h\|_{L^2(\Omega)}$ is of order h^{-1} as $h \to 0$.

- *Computation of* $b(v_h, q_h)$. Let us consider an arbitrary $v_h \in X_h$. Using the previous notations we compute $\delta_x q_h$ and $\delta_y q_h$. On the one hand, $\delta_x q_h = 0$. On the other hand,

  $$(\delta_y q_h)_{i-\frac{1}{2}\,j-\frac{1}{2}} = \begin{cases} 0, & \text{if } i = 2p,\\ 2\frac{(-1)^{j+1}}{k}, & \text{if } i = 2p+1.\end{cases}$$

  Then we get

  $$\begin{aligned}b(v_h,q_h) &= 2\sum_{p=0}^{n-1}\sum_{j=1}^{2m} h(-1)^{j+1} v_{2p+\frac{1}{2}\,j-\frac{1}{2}}\\ &= 2\sum_{p=0}^{n-1}\sum_{r=1}^{m} h\left(v_{2p+\frac{1}{2}\,2r-\frac{3}{2}} - v_{2p+\frac{1}{2}\,2r-\frac{1}{2}}\right).\end{aligned}$$

  Hence, by the Cauchy-Schwarz inequality, we get

  $$\begin{aligned}&|b(v_h,q_h)|\\ &\quad\le 2h\sum_{p=0}^{n-1}\int_0^1 \left|\partial_y v_h((2p+1)h,y)\right| \mathrm{d}y\\ &\quad\le 2h\left(\int_0^1\left(\sum_{p=0}^{n-1}\left|\partial_y v_h((2p+1)h,y)\right|\right)^2 \mathrm{d}y\right)^{\frac{1}{2}}\\ &\quad\le 2h\sqrt{n}\left(\int_0^1\sum_{p=0}^{n-1}\left|\partial_y v_h((2p+1)h,y)\right|^2 \mathrm{d}y\right)^{\frac{1}{2}}\end{aligned}$$

  As $v_h \in \mathbb{Q}^1$, $x \mapsto \partial_y v_h(x,y)$ is piecewise affine and moreover it is continuous for a.e. $y \in \Omega$ (precisely for all the y's that do not belong to the boundary of the cells). We can then use the following basic inequality

  $$a^2+b^2 \le 6\int_0^1 |ax+b(1-x)|^2\,\mathrm{d}x,$$

  which readily adapts on an interval of width h as

  $$a^2+b^2 \le \frac{6}{h}\int_0^h \left|a\frac{x}{h}+b\left(1-\frac{x}{h}\right)\right)\right|^2 \mathrm{d}x.$$

  Thus we get

  $$\sum_{p=0}^{n-1}\left|\partial_y v_h((2p+1)h,y)\right|^2 \le \frac{6}{h}\int_0^1\left|\partial_y v_h(x,y)\right|^2 \mathrm{d}x$$

  and finally integrating with respect to y,

  $$\begin{aligned}|b(v_h,q_h)| &\le 2h\sqrt{n}\left(\tfrac{1}{h}\textstyle\int_0^1\int_0^1 |\partial_y v_h|^2 \mathrm{d}x\mathrm{d}y\right)^{\frac{1}{2}}\\ &\le 2\sqrt{h}\sqrt{n}\|\nabla v_h\|_{L^2}.\end{aligned}$$

  Thus we obtain, for any $v_h \in X_h$,

  $$\frac{|b(v_h,q_h)|}{\|v_h\|_{H^1}} \le C.$$

  The inf-sup condition implies

  $$\beta_h\|q_h\|_{L^2} \le \sup_{v_h\in X_h}\frac{|b(v_h,q_h)|}{\|v_h\|_{H^1}} \le C.$$

  As $\|q_h\|_{L^2(\Omega)}$ is of order h^{-1}, we have $\beta_h \le Ch$.

■

Previous examples shed a light on possible fails when considering finite element approximations for the Stokes problem. Let us give an outline of classical stable finite elements.

6.2.3 The $\mathbb{P}^1_b - \mathbb{P}^1$ element

As we have seen before, the main reason for the instability of the $\mathbb{Q}^1 - \mathbb{Q}^0$ or $\mathbb{P}^1 - \mathbb{P}^0$ approximation is the fact that the pressure approximation space is too large in some sense compared to the velocity approximation space.

That is the reason why the construction of uniformly stable approximation spaces for the Stokes problem are often built starting from an unstable discretisation by using one the two following strategies:

- Either one can add functions (that is degrees of freedom) in the velocity approximation space. This induces a higher computational cost.
- Or one can remove functions (that is degrees of freedom) in the pressure approximation space. This induces a lower accuracy of the approximation.

The choice of a suitable pair of approximation spaces is thus a sort of tradeof between stability/accuracy/computational effort.

A very popular element that is based on the first strategy is the so-called *mini-element*, also denoted by

$\mathbb{P}^1-$ bubble / $\mathbb{P}^1$ or $\mathbb{P}^1_\mathrm{b}-\mathbb{P}^1$, see [1]. It consists in adding to the $\mathbb{P}^1-\mathbb{P}^1$ element one degree of freedom for each component of the velocity on the barycenters of the cells. Let $\hat{b}\in H^1(\hat{K})$ denote a function which takes the value 1 at the barycenter of the reference cell $\hat{K}$, vanishes on its boundary $\partial\hat{K}$ and satisfies $0\leq\hat{b}\leq 1$. Such a function is known as a *bubble function.* Define the space

$$\mathbb{P}^1_\mathrm{b}=\{v\in(C^0(\bar{\Omega}))^2,\ v\circ T_K\in(\mathbb{P}^1\oplus\mathrm{span}\{\hat{b}\})^2,\ \forall K\in\mathscr{T}\}.$$

Taking $X_h=\mathbb{P}^1_\mathrm{b}$ and $M_h=\mathbb{P}^1$, the inf-sup constant does not depend on h which ensures an optimal convergence rate:

Theorem 6.6 ($\mathbb{P}^1_\mathrm{b}-\mathbb{P}^1$ estimates) *Let $(\mathscr{T}_h)_h$ be a regular family of meshes which satisfy the geometrical assumption: each element $K\in\mathscr{T}_h$ has at most one edge on the boundary of Ω. Let $X_h\times M_h$ be the approximation spaces related to the $\mathbb{P}^1_\mathrm{b}-\mathbb{P}^1$ approximation. Assume that the solution (u,p) of the Stokes problem belongs to $(H^2(\Omega))^2\times H^1(\Omega)$. Then we have the following estimate:*

$$\|u-u_h\|_{H^1(\Omega)}+\|p-p_h\|_{L^2(\Omega)}\leq Ch\left(\|u\|_{H^2(\Omega)}+\|p\|_{H^1(\Omega)}\right).$$

If furthermore the adjoint problem (which is still the Stokes problem) has the elliptic regularity property in Ω then

$$\|u-u_h\|_{L^2(\Omega)}\leq Ch^2\left(\|u\|_{H^2(\Omega)}+\|p\|_{H^1(\Omega)}\right).$$

The $\mathbb{P}^1-$bubble / $\mathbb{P}^1$ were defined by Arnold, Brezzi and Fortin (1984), see [1], and its analysis relies on the Clement interpolation operator which satisfies Fortin's lemma, hence the finite element satisfies a uniform inf-sup condition. The mini-element is the simplest stable element for the Stokes system.

6.2.4 The $\mathbb{P}^2-\mathbb{P}^1$ (Taylor-Hood) element

We consider a simplicial mesh and the $\mathbb{P}^2-\mathbb{P}^1$ elements:

$$X_h=\{v\in(C^0(\bar{\Omega}))^2,\ \forall K\in\mathscr{T},\ v\circ T_K\in(\mathbb{P}_2)^2\}\cap(H^1_0(\Omega))^2,$$

$$M_h=\{v\in C^0(\bar{\Omega}),\ \forall K\in\mathscr{T},\ v\circ T_K\in\mathbb{P}_1\}\cap L^2_0(\Omega).$$

We can prove for a quite general simplicial mesh of that the inf-sup constant β_h does not depend on h which ensures an optimal convergence rate.

Proposition 6.7 ($\mathbb{P}^2-\mathbb{P}^1$ finite elements) *Let us assume that each element $K\in\mathscr{T}$ has at most one edge on the boundary of Ω. The spaces X_h and M_h satisfy a uniform inf-sup condition: the constant β_h only depends on the regularity of the mesh.*

The geometrical assumption on the mesh is not restrictive.

Proof of Proposition 6.7. Let $q_h\in M_h\setminus\{0\}$. We aim at proving that there exists $C>0$ (which does not depend on h) and some $v_h\in X_h$ (which may depend on q_h) such that

$$\frac{b(v_h,q_h)}{\|v_h\|_{H^1(\Omega)}}\geq C\|q_h\|_{L^2(\Omega)}.$$

Step 1. Definition of v_h. For each edge σ of the mesh, we denote τ_σ a unit vector of this edge and $|\sigma|$ its length. The element $v_h\in X_h$ is uniquely determined by defining its values at the nodes as follows:

- $v_h(a)=0$ if a is a vertex of the mesh;
- $v_h(a)=-|\sigma|^2\,\tau_\sigma(\nabla q_h\cdot\tau_\sigma)$ if a is the middle of an *interior* edge σ;
- $v_h(a)=0$ if a is the middle of a boundary edge.

Let us note that

- v_h is zero on the boundary;
- the definition of $v_h(a)$ is consistent if a is the middle of an *interior* edge. Indeed the gradient of q_h is a priori constant on each element and thus it has no trace defined on the edges. But since q_h is continuous through the edge, the tangential gradient of q_h is uniqueley defined on the edge.

We now use the following quadrature formula, which is valid for the elements in $\mathbb{P}^2$: for all $\pi\in\mathbb{P}^2$, for all $K\in\mathscr{T}$,

$$\frac{1}{|K|}\int_K\pi(x)\,\mathrm{d}x=\sum_{a\in\mathscr{M}(K)}\frac{\pi(a)}{5}-\sum_{a\in\mathscr{S}(K)}\frac{\pi(a)}{20},$$

where $\mathscr{M}(K)$ denotes the set of the all the middles of the edges of K and $\mathscr{S}(K)$ denotes the set of the vertices of K (this formula can be proved on the reference element and then extended to any element with a change of variables). We thus obtain

$$\begin{aligned}b(v_h,q_h)&=-\int_\Omega v_h\nabla q_h\\&=-\sum_K\left(\int_K v_h\nabla q_h\right)\\&=\frac{1}{5}\sum_K|K|\left(\sum_{\sigma\in\partial K\setminus\partial\Omega}|\sigma|^2\,(\nabla q_h\cdot\tau_\sigma)^2\right)\quad(37)\end{aligned}$$

Step 2. Estimate of $|\nabla q_h|$. Let K be a triangle. We consider an affine function on K and denote u_i the value of the function at the vertices M_i of K. We aim at controlling the gradient of u with respect to the terms u_i-u_j. For this we can see that the norm of vector $\rho_K\frac{\nabla u}{|\nabla u|}$ is ρ_K (see the definition of ρ_K in Proposition 5.21). Thus there exists $X,Y\in K$ such that

$$X-Y=\rho_K\frac{\nabla u}{|\nabla u|}.$$

We denote $\{x_i\}_{i=1,2,3}$ and $\{y_i\}_{i=1,2,3}$ the (nonnegative!) barycentric coordinates[12] of X and Y in the triangle and we observe that, since u is affine, we have on the one hand

$$u(X)-u(Y)=\nabla u\cdot(X-Y)=\rho_K\,|\nabla u|$$

and, on the other hand,

$$u(X)-u(Y)=\sum_{i,j}x_iy_j(u(M_i)-u(M_j))=\sum_{i,j}x_iy_j(u_i-u_j).$$

By the Jensen inequality,

$$\begin{array}{rcl}\rho_K^2\,|\nabla u|^2 & \leq & \sum_{i,j}x_iy_j\left|u_i-u_j\right|^2\\ & \leq & |u_1-u_2|^2+|u_2-u_3|^2+|u_3-u_1|^2.\end{array}$$

Besides, since we have

$$\begin{array}{rcl}|u_1-u_2|^2 & \leq & (|u_1-u_3|+|u_3-u_2|)^2\\ & \leq & 2\,|u_3-u_1|^2+2\,|u_2-u_3|^2,\end{array}$$

the estimate still holds with two terms, up to a constant:

$$\rho_K^2\,|\nabla u|^2\leq 3\,|u_1-u_2|^2+3\,|u_1-u_3|^2.$$

Step 3. Bound from below for $b(v_h,q_h)$. Let us go back to the estimate of $b(v_h,q_h)$, see Eq. (37). In the last term,

$$|\sigma|^2\,(\nabla q_h\cdot\tau_\sigma)^2=|q_h(a_\sigma)-q_h(b_\sigma)|^2$$

where a_σ,b_σ denote the nodes associated to σ. Thus each term $|\sigma|^2\,(\nabla q_h\cdot\tau_\sigma)^2$ can be expressed as some

$$|q_h(a_\sigma)-q_h(b_\sigma)|^2.$$

Consequently, a_K,b_K,c_K denoting the vertices of K and assuming that the triangles could have at most one edge in $\partial\Omega$,

$$(*):=\sum_{\sigma\in\partial K\setminus\partial\Omega}|\sigma|^2\,(\nabla q_h\cdot\tau_\sigma)^2$$

can be calculated:

- if $\overline{K}\cap\partial\Omega=\emptyset$

$$\begin{array}{rl}(*)=|q_h(a_K)-q_h(b_K)|^2+|q_h(b_K)-q_h(c_K)|^2\\ +\,|q_h(c_K)-q_h(a_K)|^2\end{array}$$

- if $[b_K,c_K]\subset\partial\Omega$,

$$(*)=|q_h(a_K)-q_h(b_K)|^2+|q_h(a_K)-q_h(c_K)|^2.$$

By Step 2, each term $\sum_{\sigma\in\partial K\setminus\partial\Omega}|\sigma|^2\,(\nabla q_h\cdot\tau_\sigma)^2$ can be bounded from below by $C\rho_K^2\,|\nabla q_h|^2$ so that we get the estimate:

$$\begin{array}{rcll}b(v_h,q_h) & \geq & C\sum_K|K|\,\rho_K^2\,|\nabla q_h|^2 & \\ & \geq & C\sum_K\rho_K^2\,|q_h|^2_{H^1(K)} & \\ & \geq & C\sum_K h_K^2\,|q_h|^2_{H^1(K)}\,, & (38)\end{array}$$

as the mesh is regular.

Step 4. Estimate of $\|v_h\|_{H^1}$. We choose $K\in\mathscr{T}$ and ϕ_i a shape function such that

$$\mathrm{supp}(\phi_i)\cap K\neq\emptyset.$$

By Theorem 5.22, page 38,

$$|\phi_i|_{H^1(K)}\leq C\frac{|K|^{\frac{1}{2}}}{\rho_K}\qquad(39)$$

Using the definition of v_h, we have

$$v_h=-\sum_{a\in\mathscr{M}(K),\ a\in\sigma}\left(|\sigma|^2\,\tau_\sigma(\nabla q_h\cdot\tau_\sigma))\right)\phi_{I(a)},$$

where $I(a)$ denotes the global numbering of the shape function associated to a. Combined with Eq. (39) we obtain

$$\begin{array}{rcl}|v_h|^2_{H^1(K)} & \leq & \displaystyle\sum_{a\in\mathscr{M}(K),\ a\in\sigma}C\frac{|K|}{\rho_K^2}\,|\sigma|^4\,|\nabla q_h|^2\\ & \leq & C\,|K|\,\frac{h_K^2}{\rho_K^2}h_K^2\,|\nabla q_h|^2.\end{array}$$

Thus with the regularity property of the mesh,

$$|v_h|^2_{H^1(K)}\leq Ch_K^2\,|q_h|^2_{H^1(K)}$$

and then

$$|v_h|^2_{H^1(\Omega)}\leq C\sum_K h_K^2\,|q_h|^2_{H^1(K)}.\qquad(40)$$

Conclusion. Combining Eqs. (38) and (40), we have

$$\frac{b(v_h,q_h)}{\|v_h\|_{H^1(\Omega)}}\geq C\left(\sum_K h_K^2\,|q_h|^2_{H^1(K)}\right)^{\frac{1}{2}}.$$

Now since $q_h\in M_h\subset H^1(\Omega)$, its mean value is zero, so that $q_h\in\tilde{H}^1(\Omega)$ and we have

$$\left(\sum_K h_K^2\,|q_h|^2_{H^1(K)}\right)^{\frac{1}{2}}=|q_h|_{H^1(\Omega)}=\|\nabla q_h\|_{L^2(\Omega)}\geq C\|q_h\|_{L^2(\Omega)}$$

by the Poincaré-Wirthinger inequality (see Theorem 1.20). We finally obtain

$$\frac{b(v_h,q_h)}{\|v_h\|_{H^1(\Omega)}}\geq C\|q_h\|_{L^2(\Omega)}.$$

Thus the uniform inf-sup condition is proved. ■

[12] i.e. (x_1,x_2,x_3) is the unique triplet satisfying

- $x_i\geq 0$, for all $i=1,2,3$,
- $x_1+x_2+x_3=1$,
- $X=x_1M_1+x_2M_2+x_3M_3$.

Theorem 6.8 **($\mathbb{P}^2-\mathbb{P}^1$ estimates)** *Let $(\mathscr{T}_h)_h$ a regular family of meshes which satisfy the geometrical assumption: each element $K\in\mathscr{T}_h$ has at most one edge on the boundary of Ω. Let $X_h\times M_h$ the approximation spaces related to the $\mathbb{P}^2-\mathbb{P}^1$ approximation. Assume that the solution (u,p) of the Stokes problem belongs to $(H^3(\Omega))^2\times H^2(\Omega)$. Then we have the following estimate:*

$$\|u-u_h\|_{H^1(\Omega)}+\|p-p_h\|_{L^2(\Omega)}\leq Ch^2\left(\|u\|_{H^3(\Omega)}+\|p\|_{H^2(\Omega)}\right).$$

If furthermore the adjoint problem (which is still the Stokes problem) has the elliptic regularity property in Ω then

$$\|u-u_h\|_{L^2(\Omega)}\leq Ch^3\left(\|u\|_{H^3(\Omega)}+\|p\|_{H^2(\Omega)}\right).$$

Proof of Theorem 6.8. The interpolation results for X_h and M_h are known: we apply Theorem 5.36, page 42, with $k=2$ and $m=2$ and $l=1$ for the velocity, $k=1$ and $m=1$ and $l=0$ for the pressure):

$$\forall v\in H^3(\Omega),\quad \left|v-\mathscr{I}_h^k v\right|_{H^1(\Omega)}\leq Ch^2\,|v|_{H^3(\Omega)},$$

$$\forall p\in H^2(\Omega),\quad \|p-\mathscr{I}_h^k p\|_{L^2(\Omega)}\leq Ch^2\,|p|_{H^2(\Omega)}.$$

From Lemma 4.7 combined with Lemma 6.7 (uniform inf-sup condition), we deduce the convergence of the method.

From Lemma 4.10 (abstract error estimate) we get

$$\begin{aligned}\|u-u_h\|_X &\leq \left(1+\frac{\|a\|}{\alpha}\right)\left(1+\frac{\|b\|}{\beta_h}\right)d(u,X_h)\\ &\qquad +\frac{\|b\|}{\alpha}d(p,M_h),\\ \|p-p_h\|_M &\leq \frac{\|a\|}{\beta_h}\left(1+\frac{\|a\|}{\alpha}\right)\left(1+\frac{\|b\|}{\beta_h}\right)d(u,X_h)\\ &\qquad +\left(1+\frac{\|b\|}{\beta_h}+\frac{\|a\|}{\beta_h}\frac{\|b\|}{\alpha}\right)d(p,M_h).\end{aligned}$$

From these estimates combined with the above global interpolation estimates, we deduce the error estimate.

The estimate in L^2-norm readily adapts from the Aubin-Nitsche trick. ■

7. Problems

We now investigate the main results of the previous sections by means of numerical simulations. For this purpose we use a finite element solver, `FreeFem++`, which is developed and maintained in Université Pierre & Marie Curie and Laboratoire Jacques-Louis Lions [12]. Before focusing on mathematical problems, let us briefly introduce `FreeFem++`.

`FreeFem++` is a free software designed to compute the solution of initial- and boundary-value problems for partial differential equations in 2D or 3D with the finite element method. Its principle is based upon the *discretization* of a variational formulation and the *computation of the solution* of the resulting linear system. Multi-physics nonlinear problems can be addressed through iterative schemes that rely on a linear problem to solve at the basic level.

The requirements for the user are the following ones:

1. define a domain: boundaries can be defined with a simple parametrization;
2. define a mesh: the number of nodes on each labelled boundary is sufficient, as the software owns a mesh generator that is able to produce triangulations;
3. define approximation spaces ($\mathbb{P}^0$, $\mathbb{P}^1$, $\mathbb{P}^1_b$, $\mathbb{P}^2$...);

   Remark 7.1 *At this point,* `FreeFem++` *defines a finite element basis which "lives" on the mesh of the domain.*

4. define a variational formulation: this is the most specific part of the software, as a suitable syntax is required (note that it is very close to the mathematical formulation).

   Remark 7.2 *At this point,* `FreeFem++` *owns all the tools leading to the corresponding discretized problem: the software functionality consists in building the matrix and the right-hand side vector associated to 1) the bilinear / linear form, 2) the finite element space.*

5. choose a linear solver (optional): `FreeFem++` owns different solvers, such as `LU`, `Cholesky`, `Crout`, `CG`, `GMRES`, `UMFPack`, `sparsesolver`. Sparse systems can be solved with `sparsesolver`, `UMFPACK`, `GMRES`, whereas full systems can be solved with `LU`, `Crout`, `Cholesky`. The default choice is `sparsesolver` (equivalent to `UMFPACK` if no sparse solver is defined) or `LU` if no sparse solver is available.
   - `LU` is a *direct* method corresponding to the LU decomposition method.
   - `Crout` is a *direct* method for *symmetric* systems, based upon the LU decomposition method, thus using specific properties due to the symmetric property of the matrix.
   - `Cholesky` is a *direct* method for *symmetric positive-definite* systems, based upon the LU decomposition method, thus using specific properties the matrix leading to a simple decomposition.
   - `CG` (conjugate gradient method) is an *iterative* method for *symmetric positive-definite* systems.

- GMRES is an *iterative* method for sparse systems (no additional assumption on the structural properties of the matrix is required). It is a generalization of the conjugate gradient method.
- UMFPACK is a *direct* method for sparse systems (no additional assumption).

Details (including download, documentation, examples etc.) can be found on http://www.freefem.org/.

7.1 Analysis of the convergence

Problem. Write a FreeFem++ program to solve a Poisson problem with $\mathbb{P}^1$ or $\mathbb{P}^2$ finite elements. Quantify the error in H^1 and L^2 norms with respect to the mesh size. □

Solution. Define $u := xy(1-x)(1-y)$ and $f := 2y(1-y)+2x(1-x)$. It can be checked that the function u is the unique variational solution in $H_0^1(\Omega)$ of $-\Delta u = f$ in $\Omega =]0,1[^2$. For a fixed mesh, we may compute corresponding finite element solution u_h and we aim at estimating $\|u-u_h\|_{H^1}$ and $\|u-u_h\|_{L^2}$ (note that, for any h, the error $\|u-u_h\|$ should be computed on a fixed *very fine* mesh, not on the coarse one). Assume that the error behaves as

$$\|u-u_h\| = \mathcal{O}(h^\alpha),$$

where α is the order of the method in a chosen norm. Then we have

$$\log(\|u-u_h\|) = \alpha\log(h)+\log(C).$$

As a consequence, α is numerically determined by identifying the derivative of the linear function

$$\log(h) \mapsto \log(\|u-u_h\|) = \alpha\log(h)+\log(C).$$

Using different values of $\{h_i\}_i$ with corresponding values of $\{\|u-u_{h_i}\|\}_i$ (to be determined by solving the PDE problem), a linear regression, or a visual inspection, allows us to identify α.

- Let us discuss the $\mathbb{P}^1$ approximation, see the left-hand side of Figure 13.
  - Numerical results provide the estimate $\|u-u_h\|_{H^1} = \mathcal{O}(h)$. This illustrates Theorem 5.10, as $u \in H_0^1(\Omega)\cap H^2(\Omega)$.
  - Numerical results provide the estimate $\|u-u_h\|_{L^2} = \mathcal{O}(h^2)$: as (P) satisfies the elliptic regularity property, the Aubin-Nitsche lemma applies, see Lemma 5.11.
- Let us discuss now the $\mathbb{P}^2$ approximation, see the right-hand side of Figure 13.
  - Numerical results provide the estimate $\|u-u_h\|_{H^1} = \mathcal{O}(h^2)$. This illustrates Theorem 5.19, as $u \in H_0^1(\Omega)\cap H^3(\Omega)$.
  - Numerical results provide the estimate $\|u-u_h\|_{L^2} = \mathcal{O}(h^3)$: as (P) satisfies the elliptic regularity property, the Aubin-Nitsche lemma applies, see Lemma 5.11 and Theorem 5.19.

□

7.2 Numerical treatment of the boundary conditions

Problem. Let $f \equiv 1$, $\alpha > 0$ and $\varepsilon > 0$. Let us consider the Laplace-Robin problem:

$$(\mathrm{P}_\varepsilon)\left\{\begin{array}{rcll} -\Delta u+\alpha u &=& f & \text{in } \Omega =]0,1[^2, \\ \nabla u\cdot n+\dfrac{1}{\varepsilon}u &=& 0 & \text{on } \partial\Omega. \end{array}\right.$$

1. Write the variational formulation of the problem. Write a FreeFem++ program to solve the problem with $\mathbb{P}^1$ or $\mathbb{P}^2$ finite elements.
2. Discuss the behaviour of the solution of the Robin problem as ε goes to 0.
3. Discuss the behaviour of the solution of the Robin problem as ε goes to ∞.

□

Solution. Let us introduce the Laplace-Dirichlet and Laplace-Neumann problems:

$$(\mathrm{P}_0)\left\{\begin{array}{rcll} -\Delta u+\alpha u &=& f & \text{in } \Omega =]0,1[^2, \\ u &=& 0 & \text{on } \partial\Omega, \end{array}\right.$$

$$(\mathrm{P}_\infty)\left\{\begin{array}{rcll} -\Delta u+\alpha u &=& f & \text{in } \Omega =]0,1[^2, \\ \nabla u\cdot n &=& 0 & \text{on } \partial\Omega. \end{array}\right.$$

At least formally, we may expect that the solution u_ε of (P_ε)

- converges to the solution of (P_0) when ε goes to 0,
- converges to the solution of (P_∞) when ε goes to $+\infty$.

Let us use FreeFem++ as a validation tool for the above formal asymptotics. In that prospect we need to define the variational formulation of each problem: it will allow us to define the suitable functional framework of our problems and it will be the key for the computations with FreeFem++.

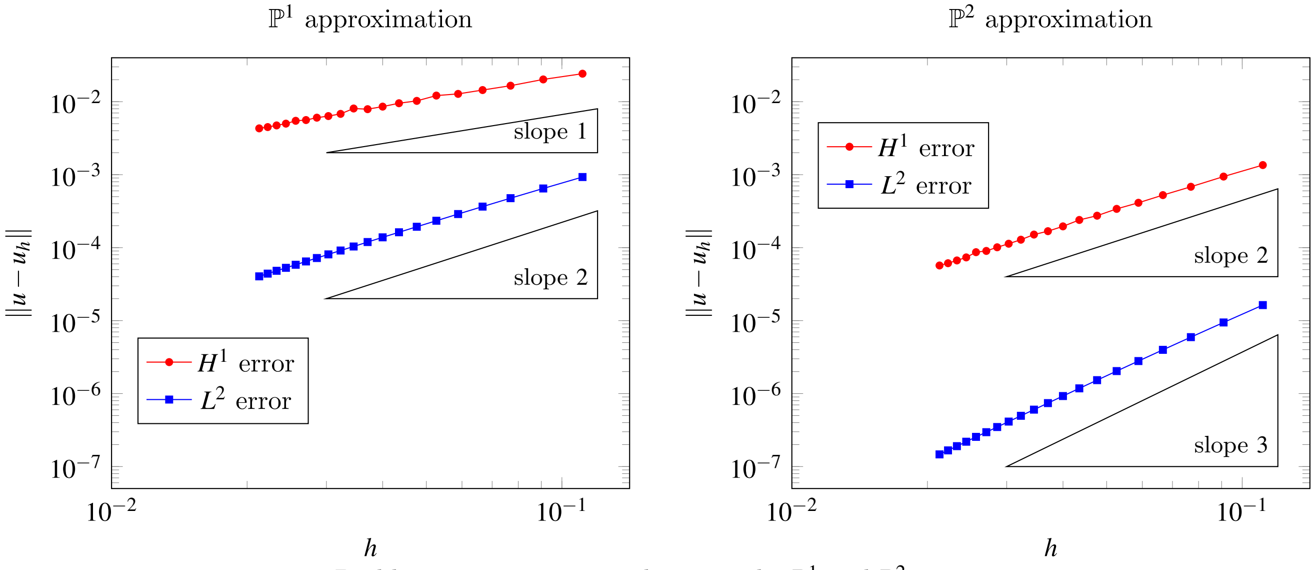


Figure 13. Problem 7.1 : comparison between the $\mathbb{P}^1$ and $\mathbb{P}^2$ approximations

1. The variational formulations of the problems write:

$$(\mathrm{P}_\varepsilon)\left\{\begin{array}{l}\text{Find } u\in H^1(\Omega) \text{ such that}\\ \displaystyle\int_\Omega \nabla u\cdot\nabla v+\alpha\int_\Omega uv+\frac{1}{\varepsilon}\int_{\partial\Omega} uv=\int_\Omega fv,\\ \text{for all } v\in H^1(\Omega),\end{array}\right.$$

$$(\mathrm{P}_0)\left\{\begin{array}{l}\text{Find } u\in H^1_0(\Omega) \text{ such that}\\ \displaystyle\int_\Omega \nabla u\cdot\nabla v+\alpha\int_\Omega uv=\int_\Omega fv,\\ \text{for all } v\in H^1_0(\Omega),\end{array}\right.$$

$$(\mathrm{P}_\infty)\left\{\begin{array}{l}\text{Find } u\in H^1(\Omega) \text{ such that}\\ \displaystyle\int_\Omega \nabla u\cdot\nabla v+\alpha\int_\Omega uv=\int_\Omega fv,\\ \text{for all } v\in H^1(\Omega).\end{array}\right.$$

Each problem is well-posed (for problem (P_ε), use the continuity of the trace operator for the mathematical treatment of the boundary term in the bilinear form) and we will denote u_ε, u_0 and u_∞ the respective solutions of the problems.

2. We fix $\alpha=1$ and $f=1$ for the computations. We use a mesh 200×200 and, in fact, we compute the error $\|u_{\varepsilon,h}-u_{0,h}\|_{H^1}$. In the left-hand side of Figure 14 we observe that, as ε goes to 0, $u_{\varepsilon,h}$ converges to $u_{0,h}$ in $H^1(\Omega)$ at order 1:

$$\|u_{\varepsilon,h}-u_{0,h}\|_{H^1}\simeq C\varepsilon.$$

Actually this is how `FreeFem++` imposes Dirichlet boundary conditions! The software considers *all* the nodes as unprescribed (including the boundary nodes). In order to prescribe $u=g$ at the boundary, the software considers a Robin condition

$$\nabla u\cdot n+\frac{1}{\varepsilon}u=\frac{1}{\varepsilon}g,$$

with a very small value for ε: this is a so-called *penalty* method, as the Dirichlet condition is mimicked by a penalized Robin condition. This also explains how `FreeFem++` is able to compute the solution of a problem with a non-homogeneous condition $u=g$, even if $g\notin H^{\frac{1}{2}}(\partial\Omega)$.

The numerical results are compatible with the following result:

Proposition 7.3 (From Robin to Dirichlet) *Let u_ε be the variational solution of the Laplace-Robin problem (P_ε) and let u_0 be the solution of the Laplace-Dirichlet problem (P_0). Then u_ε converges to u_0 in $H^1(\Omega)$ as ε goes to 0.*

Assume furthermore that $u_0\in H^2(\Omega)$. Moreover $\|u_\varepsilon-u_0\|_{H^1}$ converges to 0 at least at order $1/2$.

Proof of Proposition 7.3.

We first take $v=u_\varepsilon$ as a test function in (P_ε) and we use the Cauchy-Schwarz inequality to deduce that

$$\|\nabla u_\varepsilon\|^2_{L^2}+\frac{\alpha}{2}\|u_\varepsilon\|^2_{L^2}+\frac{1}{\varepsilon}\int_\partial |u_\varepsilon|^2\le\frac{1}{2\alpha}\|f\|^2_{L^2},\ \ \forall\varepsilon>0.$$

It follows that the family u_ε is bounded in $H^1()$ and that the traces $\gamma_0(u_\varepsilon)$ tends to 0 in $L^2(\partial\Omega)$ as $\varepsilon\to 0$. As a consequence, there exists a subsequence u_{ε_k} that weakly converges towards some u in $H^1()$ and moreover, this limit satisfies $\gamma_0(u)=0$ that is $u\in H^1_0(\Omega)$.

Taking now a test function $v\in H^1_0(\Omega)$ in (P_ε), the boundary term disappears and we can easily pass to the limit as k goes to infinity. It follows that

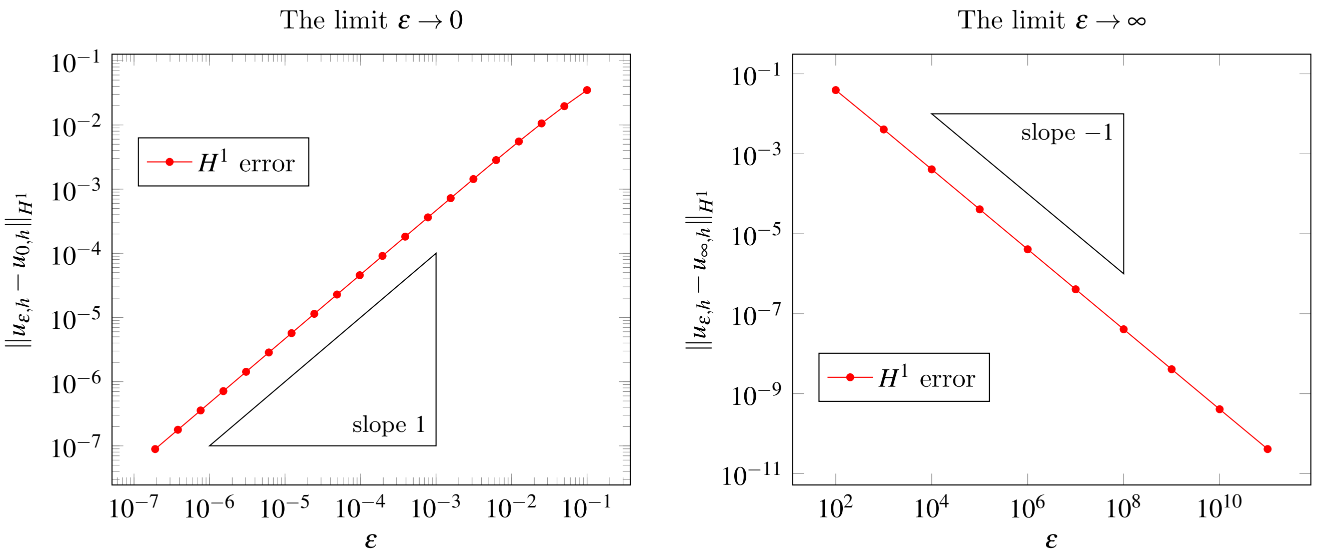


Figure 14. Problem 7.2. Illustrations of the convergence in the regimes $\varepsilon \to 0$ and $\varepsilon \to \infty$

$u = u_0$, the unique solution of (P_0). By Theorem 4.2, we obtain that the whole family u_ε weakly converges to u_0 in $H^1(\Omega)$ as $\varepsilon \to 0$. It remains to show the strong convergence.

We set $e_\varepsilon = u_\varepsilon - u_0$. We test the equation satisfied by u_0 by a test function $v \in H^1(\Omega)$ and we subtract the weak formulation of (P_ε) to get

$$\int_\Omega \nabla e_\varepsilon \cdot \nabla v + \alpha \int_\Omega e_\varepsilon v + \frac{1}{\varepsilon} \int_{\partial\Omega} e_\varepsilon v \\ = -\langle \nabla u_0 \cdot n, v \rangle_{H^{-1/2},H^{1/2}}, \quad \forall v \in H^1(\Omega).$$

Taking $v = e_\varepsilon$ as a test function in this last equation, we obtain

$$\|\nabla e_\varepsilon\|_{L^2}^2 + \alpha \|e_\varepsilon\|_{L^2}^2 + \frac{1}{\varepsilon} \|e_\varepsilon\|_{L^2(\partial)}^2 \\ = -\langle \nabla u_0 \cdot n, e_\varepsilon \rangle_{H^{-1/2},H^{1/2}}. \quad (41)$$

This last quantity converges to 0 since e_ε weakly converges to 0 in $H^1(\Omega)$ and the first claim is proved.

Assume now that $u_0 \in H^2()$, which implies in particular that $\nabla u_0 \cdot n \in L^2(\partial\Omega)$, it follows that the boundary term can be written as an integral instead of duality bracket and can be estimated as follows

$$\left| \int_{\partial\Omega} \nabla u_0 \cdot n \, e_\varepsilon \right| \\ \leq \left(\int_{\partial\Omega} |\nabla u_0 \cdot n|^2 \right)^{1/2} \left(\int_{\partial\Omega} |e_\varepsilon|^2 \right)^{1/2} \\ \leq \frac{\varepsilon}{2} \int_{\partial\Omega} |\nabla u_0 \cdot n|^2 + \frac{1}{2\varepsilon} \int_{\partial\Omega} |e_\varepsilon|^2,$$

by the Young's inequality. Using this estimate in the inequality (41) gives

$$\|\nabla e_\varepsilon\|_{L^2}^2 + \alpha \|e_\varepsilon\|_{L^2}^2 + \frac{1}{2\varepsilon} \|e_\varepsilon\|_{L^2(\partial)}^2 \leq \frac{\varepsilon}{2} \|u_0\|_{H^2}^2,$$

and the proof is complete. ■

3. We fix $\alpha = 1$ and $f = 1$ for the computations. We use a mesh 50×50 and we compute the error $\|u_{\varepsilon,h} - u_{0,h}\|_{H^1}$. In the right-hand side of Figure 14 we observe that, as ε goes to $+\infty$, $u_{\varepsilon,h}$ converges to $u_{\infty,h}$ in $H^1(\Omega)$ at order 1:

$$\|u_{\varepsilon,h} - u_{\infty,h}\|_{H^1} \simeq \frac{C}{\varepsilon}.$$

The numerical results are compatible with the following result:

Proposition 7.4 (From Robin to Neumann) *Let u_ε be the variational solution of the Laplace-Robin problem (P_ε) and let u_∞ be the solution of the Laplace-Neumann problem (P_∞). Then u_ε converges to the solution u_∞ as ε goes to $+\infty$. Moreover $\|u_\varepsilon - u_\infty\|_{H^1}$ converges to 0 at least at order 1.*

Proof of Proposition 7.4. We have

$$\int_\Omega \nabla u_\varepsilon \cdot \nabla v + \alpha \int_\Omega u_\varepsilon v + \frac{1}{\varepsilon} \int_{\partial\Omega} u_\varepsilon v = \int_\Omega f v,$$

for all $v \in H^1(\Omega)$. Taking $v = u_\varepsilon$ as a test function, we get

$$\int_\Omega |\nabla u_\varepsilon|^2 + \alpha \int_\Omega u^2 + \frac{1}{\varepsilon} \int_{\partial\Omega} |u_\varepsilon|^2 = \int_\Omega f u_\varepsilon.$$

On the one hand,

$$\int_\Omega |\nabla u_\varepsilon|^2 + \alpha \int_\Omega u^2 + \frac{1}{\varepsilon} \int_{\partial\Omega} |u_\varepsilon|^2 \geq \min(1, \alpha) \|u_\varepsilon\|_{H^1}^2$$

and, on the other hand,

$$\int_\Omega f u_\varepsilon \leq \|f\|_{L^2}\|u_\varepsilon\|_{H^1}.$$

Thus we obtain the estimate

$$\|u_\varepsilon\|_{H^1} \leq \frac{\|f\|_{L^2}}{\min(1,\alpha)}. \tag{42}$$

Remark 7.5 *At this stage, we prove that u_ε weakly converges to u_∞ in $H^1(\Omega)$. Indeed, $\{u_\varepsilon\}$ is bounded in $H^1(\Omega)$, $\{u_\varepsilon\}$ weakly converges, up to a subsequence, to some $\bar{u} \in H^1(\Omega)$. Passing to the limit in the variational formulation (note that $\gamma_0(u_\varepsilon)$ weakly converges to $\gamma_0(\bar{u})$ in $L^2(\partial\Omega)$, by continuity of the trace operator), we get :*

$$\int_\Omega \nabla\bar{u}\cdot\nabla v + \alpha\int_\Omega \bar{u}v = \int_\Omega fv,$$

for all $v \in H^1(\Omega)$. By uniqueness of the solution of the Laplace-Neumann problem, $\bar{u} = u_\infty$.

Now defining $e_\varepsilon = u_\varepsilon - u_\infty$, we use the variational formulations of the two problems and, by subtraction,

$$\int_\Omega \nabla e_\varepsilon\cdot\nabla v + \alpha\int_\Omega e_\varepsilon v + \frac{1}{\varepsilon}\int_{\partial\Omega} u_\varepsilon v = 0, \quad \forall v \in H^1(\Omega).$$

Then taking $v = e_\varepsilon$ as a test function, we obtain

$$\int_\Omega |\nabla e_\varepsilon|^2 + \alpha\int_\Omega e_\varepsilon^2 = -\frac{1}{\varepsilon}\int_{\partial\Omega} u_\varepsilon e_\varepsilon.$$

On the one hand,

$$\int_\Omega |\nabla e_\varepsilon|^2 + \alpha\int_\Omega e_\varepsilon^2 \geq \min(1,\alpha)\|e_\varepsilon\|_{H^1}^2$$

and, on the other hand,

$$-\frac{1}{\varepsilon}\int_{\partial\Omega} u_\varepsilon e_\varepsilon \leq \frac{C}{\varepsilon}\|u_\varepsilon\|_{H^1}\|e_\varepsilon\|_{H^1},$$

where we have used the Cauchy-Schwarz inequality and the continuity of the trace operator. Combining the previous inequalities and the estimate on $\|u_\varepsilon\|_{H^1}$, see Eq. (42), we get

$$\|e_\varepsilon\|_{H^1} \leq \frac{\|f\|_{L^2}}{\min(1,\alpha)^2}\frac{1}{\varepsilon},$$

which concludes the proof. ■

□

7.3 On a Dirichlet boundary term $g \notin H^{1/2}(\partial\Omega)$

Problem. Consider the Poisson problem on a unit square $\Omega =]0,1[^2$ with a source term $f \equiv 1$ and Dirichlet boundary conditions: for this purpose, we denote $\Gamma :=]0,1[\times\{0\}$ and consider the problem

$$\left\{\begin{array}{rcll} -\Delta u &=& f & \text{in } \Omega, \\ u &=& g & \text{on } \partial\Omega, \end{array}\right.$$

where the boundary data is defined as

$$g(x) = \left\{\begin{array}{ll} 1 & \text{if } x \in \Gamma, \\ 0 & \text{if } x \in \partial\Omega\setminus\Gamma. \end{array}\right.$$

We recall that g does not belong to $H^{\frac{1}{2}}(\partial\Omega)$, see Exercise 1.

Compute the $\mathbb{P}^1$ finite element solution u_h and discuss the behaviour of $\|u_h\|_{H^1}$ as h goes to 0. □

Solution. Let us define two problems $(\mathrm{P}^{(i)})$, for $i \in \{0,1\}$:

$$(\mathrm{P}^{(i)})\left\{\begin{array}{rcll} -\Delta u &=& f & \text{in } \Omega, \\ u &=& g^{(i)} & \text{on } \partial\Omega, \end{array}\right.$$

with the corresponding boundary terms:

$$g^{(0)}(x) = \left\{\begin{array}{ll} 0 & \text{if } x \in \Gamma, \\ 0 & \text{if } x \in \partial\Omega\setminus\Gamma, \end{array}\right.$$

$$g^{(1)}(x) = \left\{\begin{array}{ll} 0 & \text{if } x \in \Gamma, \\ 1 & \text{if } x \in \partial\Omega\setminus\Gamma. \end{array}\right.$$

We denote $u_h^{(0)}$ and $u_h^{(1)}$ the corresponding finite element solutions, see Figures 15 and 16.

Analysis of Problem $(\mathrm{P}^{(0)})$. The finite element solution of $(\mathrm{P}^{(0)})$ converges to the unique variational solution $u^{(0)} \in H_0^1(\Omega)$, since $f \in L^2(\Omega)$ (actually, $f \in H^{-1}(\Omega)$ would be sufficient). In our case, the domain being polygonal and convex, we may even prove that $u^{(0)} \in H^2(\Omega)$. As a consequence, if we use $\mathbb{P}^1$ finite elements, $u_h^{(0)}$ strongly converges to $u^{(0)}$ in $H^1(\Omega)$ at order 1. Figure 17 illustrates the convergence of $\|u_h^{(0)}\|_{H^1}$ to $\|u^{(0)}\|_{H^1}$.

Analysis of Problem $(\mathrm{P}^{(1)})$. The behaviour of the finite element solution of $(\mathrm{P}^{(1)})$ is quite different: indeed $g^{(1)} \notin H^{\frac{1}{2}}(\partial\Omega)$ and thus the lift operator cannot be applied. Actually the discrete linear problem is well-posed but the finite element solution $u_h^{(1)}$ does not converge to an element in $H^1(\Omega)$: problem $(\mathrm{P}^{(1)})$ admits no solution in $H^1(\Omega)$. In particular singularities concentrate at the corners $(0,0)$ and $(0,1)$, which means that the gradient locally explodes, see Figure 16. Figure 17 illustrates the divergence of $\|u_h^{(1)}\|_{H^1}$, as the limit of $u_h^{(1)}$ does not belong to $H^1(\Omega)$.

These computations provide an example of the critical assumptions on the non-homogeneous boundary term:

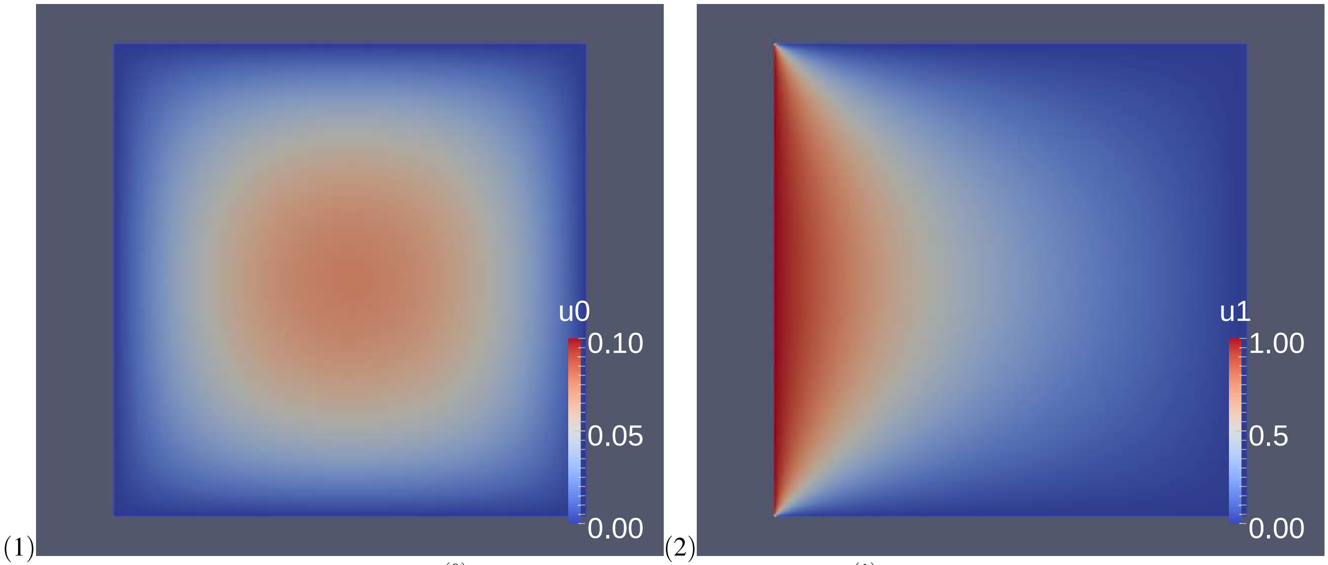


Figure 15. (1) finite element solution $u_h^{(0)}$ and (2) finite element solution $u_h^{(1)}$. In both cases the mesh size is $h = 2 \cdot 10^{-3}$ and the solution is computed with $\mathbb{P}^1$ finite elements.

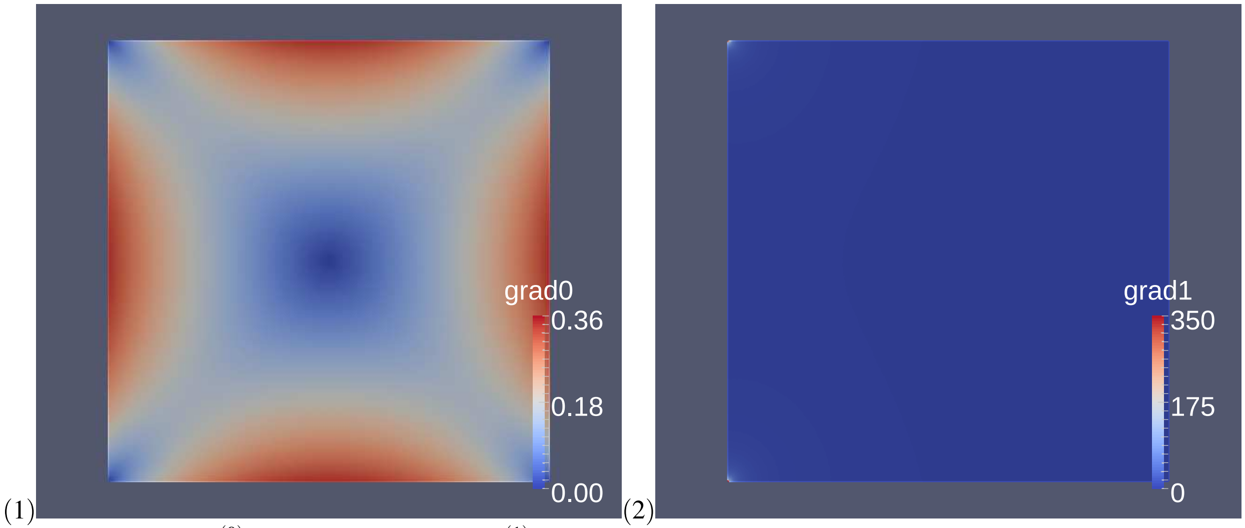


Figure 16. (1) $\mathbf{x} \mapsto \|\nabla u_h^{(0)}(\mathbf{x})\|_2$ and (2) $\mathbf{x} \mapsto \|\nabla u_h^{(1)}(\mathbf{x})\|_2$. In both cases the mesh size is $h = 2 \cdot 10^{-3}$ and the solution is computed with $\mathbb{P}^1$ finite elements.

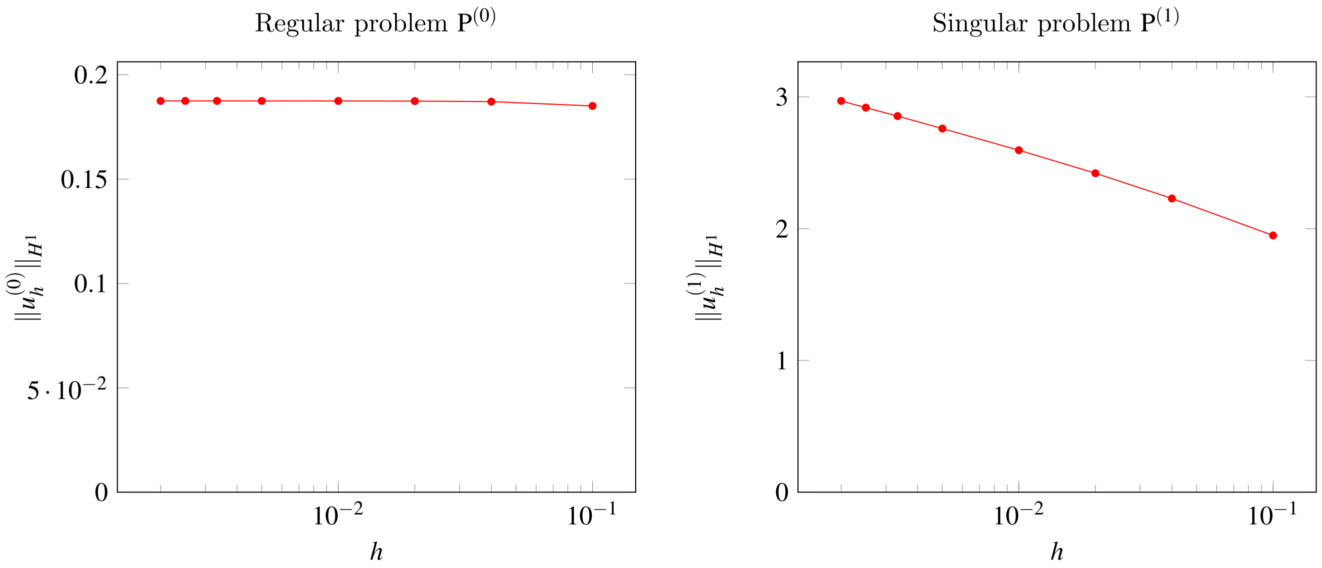


Figure 17. Problem 7.3. Influence of the regularity of the boundary data.

if g does not belong to $H^{\frac{1}{2}}(\partial\Omega)$ then the lift operator cannot be used in order to settle a classical variational formulation in $H^1(\Omega)$, see Section 2.2.

□

7.4 Compatibility conditions in PDE

Problem. Solve with `FreeFem++` the Poisson-Neumann problem:

$$\left\{ \begin{array}{rcll} -\Delta u & = & f & \text{in } \Omega =]0,1[^2, \\ \nabla u \cdot n & = & g & \text{on } \partial\Omega. \end{array} \right.$$

with $f \in L^2(\Omega)$ and $g \in H^{-\frac{1}{2}}(\partial\Omega)$. □

Solution.

Mathematical framework. The mathematical framework and the analysis of the problem are required before thinking about writing a `FreeFem++` solver for this problem. The mathematical analysis was done in section 2. In this problem, we recall the two main issues:

- The well-posedness requires the (necessary) compatibility condition:
$$\int_\Omega f + \langle g, 1 \rangle_{H^{-\frac{1}{2}}, H^{\frac{1}{2}}} = 0. \tag{43}$$
- The solution *cannot* be unique in $H^1(\Omega)$: we may add any constant to a solution, thus defining another solution. In order to select a unique solution (thus fixing the constant) we impose
$$\int_\Omega u = 0, \tag{44}$$
hence dealing with the functional space $\tilde{H}^1(\Omega)$.

Thus the variational formulation writes

$$(\mathrm{P}) \left\{ \begin{array}{l} \text{Find } u \in \tilde{H}^1(\Omega) \text{ such that} \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v = \int_\Omega f v + \langle g, v \rangle_{H^{-\frac{1}{2}}, H^{\frac{1}{2}}}, \\ \text{for all } v \in \tilde{H}^1(\Omega) \end{array} \right.$$

which is a well-posed problem under the compatibility condition of Eq. (43).

How to deal with the constraint in the functional space? Basically the main idea consists in fixing the constant which is equivalent to providing some coercivity to the problem. Instead of solving problem (P), we may solve a penalized version of the problem:

$$(\mathrm{P}_\varepsilon) \left\{ \begin{array}{l} \text{Find } u_\varepsilon \in H^1(\Omega) \text{ such that} \\ \displaystyle\int_\Omega \nabla u_\varepsilon \cdot \nabla v + \varepsilon \int_\Omega u_\varepsilon v = \int_\Omega f v + \langle g, v \rangle_{H^{-\frac{1}{2}}, H^{\frac{1}{2}}}, \\ \text{for all } v \in H^1(\Omega), \end{array} \right.$$

which is the variational formulation of problem

$$\left\{ \begin{array}{rcll} -\Delta u_\varepsilon + \varepsilon u_\varepsilon & = & f & \text{in } \Omega =]0,1[^2, \\ \nabla u_\varepsilon \cdot n & = & g & \text{on } \partial\Omega. \end{array} \right.$$

In problem (P_ε), the functional framework relies on $H^1(\Omega)$ (and not $\tilde{H}^1(\Omega)$) as the mean value of the solution is not necessarily zero). Note that, under the compatibility assumption, by taking $v \equiv 1$ as a test function we have

$$\varepsilon \int_\Omega u_\varepsilon = \int_\Omega f v + \langle g, v \rangle_{H^{-\frac{1}{2}}, H^{\frac{1}{2}}} = 0.$$

Thus the approximate solution u_ε has also zero mean value (this property emerges from the compatibility condition, it is not imposed by the functional framework). We now may prove the following:

Proposition 7.6 (Convergence) *Let $f \in L^2(\Omega)$ and $g \in H^{-\frac{1}{2}}(\partial\Omega)$. Assume that the compatibility condition* (43) *is satisfied. Let u_ε be the variational solution of the Laplace-Neumann problem* (P_ε) *and let u be the solution of the Poisson-Neumann problem* (P)*. Then u_ε converges to u as ε goes to 0. Moreover $\|u_\varepsilon - u\|_{H^1}$ converges to 0 at least at order* 1*.*

Proof of Proposition 7.6.
Bound for $|u_\varepsilon|_{H^1(\Omega)}$. We observe that $u_\varepsilon \in \tilde{H}^1(\Omega)$ for each ε and we recall that, by the Poincaré-Wirtinger inequality, see Theorem 1.20, page 5, $|\cdot|_{H^1(\Omega)}$ is a norm on $\tilde{H}^1(\Omega)$. Taking $v = u_\varepsilon$ as a test function in (P_ε) yields

$$\begin{aligned} |u_\varepsilon|^2_{H^1(\Omega)} + \varepsilon\|u_\varepsilon\|^2_{L^2(\Omega)} \quad \leq \quad & \|f\|_{L^2(\Omega)}\|u_\varepsilon\|_{L^2(\Omega)} \\ & + C_{\gamma_0}\|g\|_{H^{-\frac{1}{2}}(\partial\Omega)}\|u_\varepsilon\|_{H^1(\Omega)}, \end{aligned}$$

where we have used the Cauchy-Schwarz inequality and the continuity of the trace operator. Thus, we have

$$\begin{aligned} |u_\varepsilon|^2_{H^1(\Omega)} & \leq |u_\varepsilon|^2_{H^1(\Omega)} + \varepsilon\|u_\varepsilon\|^2_{L^2(\Omega)} \\ & \leq (\|f\|_{L^2(\Omega)} + C_{\gamma_0}\|g\|_{H^{-\frac{1}{2}}(\partial\Omega)})\|u_\varepsilon\|_{H^1(\Omega)} \\ & \leq C|u_\varepsilon|_{H^1(\Omega)}, \end{aligned}$$

where C only depends on f, g, γ_0 and Ω. Thus $\{u_\varepsilon\}$ is bounded in $H^1(\Omega)$.

Estimate for $e_\varepsilon := u_\varepsilon - u$. In the variational formulation of problem (P), thanks to the compatibility condition, test functions can be taken in $H^1(\Omega)$ and not only $\tilde{H}^1(\Omega)$, see Section 2.2. Thus we have, for all $v \in H^1(\Omega)$:

$$\begin{aligned} \int_\Omega \nabla u_\varepsilon \cdot \nabla v \quad + \quad \varepsilon\int_\Omega u_\varepsilon v \quad & = \quad \int_\Omega fv + \langle g, v\rangle_{H^{-\frac{1}{2}},H^{\frac{1}{2}}}, \\ \int_\Omega \nabla u \cdot \nabla v \quad & = \quad \int_\Omega fv + \langle g, v\rangle_{H^{-\frac{1}{2}},H^{\frac{1}{2}}}, \end{aligned}$$

and, taking the difference,

$$\int_\Omega \nabla e_\varepsilon \cdot \nabla v + \varepsilon\int_\Omega u_\varepsilon v = 0, \ \forall v \in H^1(\Omega).$$

Note that $e_\varepsilon \in \tilde{H}^1(\Omega)$ for each ε. Then taking $v = e_\varepsilon$ in the above equation, we get

$$\begin{aligned} |e_\varepsilon|^2_{H^1(\Omega)} & = -\varepsilon\textstyle\int_\Omega u_\varepsilon e_\varepsilon \\ & \leq \varepsilon\|u_\varepsilon\|_{L^2(\Omega)}\|e_\varepsilon\|_{L^2(\Omega)} \\ & \leq \varepsilon\|u_\varepsilon\|_{H^1(\Omega)}\|e_\varepsilon\|_{H^1(\Omega)}, \end{aligned}$$

by the Cauchy-Schwarz inequality. Then using the boundedness of $\{u_\varepsilon\}$ and the equivalence of $|\cdot|_{H^1(\Omega)}$ and $\|\cdot\|_{H^1(\Omega)}$ on $\tilde{H}^1(\Omega)$, we obtain

$$|e_\varepsilon|_{H^1(\Omega)} \leq C\varepsilon,$$

where C only depends on f, g, γ_0 and Ω. ■

Thus, in practical computations with `FreeFem++`, we may solve (P_ε) with e.g. $\varepsilon = 10^{-6}$: it is sufficient to ensure the stability of the computations and Proposition 7.6 guarantees that the penalized solution is close to the exact solution. *But we should be very careful with the compatibility condition*:

- *if the compatibility condition is not satisfied*, taking too small values of ε in the penalized problem does not work: as ε goes to 0, the limit problem is ill-posed (recall that the compatibility condition is a necessary condition for the well-posedness of (P));

- the compatibility condition has to be satisfied *at the discrete level* as well: thus, the projections of the data f and g over the finite element spaces have to be done carefully.

Figure 18 was obtained with `FreeFem++` computations by solving the Poisson-Neumann problem (in fact, the penalized version with $\varepsilon = 10^{-8}$) on the unit square in two situations:

1. By choosing
$$f(x,y) = 0,$$
and
$$g(x,y) = \begin{cases} x(1-x), & \text{on }]0,1[\times\{0\}, \\ 0, & \text{on } \{0\}\times]0,1[, \\ -x(1-x), & \text{on }]0,1[\times\{1\}, \\ 0, & \text{on } \{1\}\times]0,1[, \end{cases}$$
the compatibility condition is satisfied and we obtained the (penalized) solution of our problem.

2. Replacing $f \equiv 0$ by $f \equiv 1$, the compatibility condition is *not* satisfied anymore and `FreeFem++` provides a nonsense solution with an amplitude of 10^{+8}: when ε is small, the L^∞−norm of the (penalized) solution behaves as ε^{-1} illustrating the fact that the limit problem is ill-posed.

□

7.5 Lack of regularity and error estimates

Problem. Let $\Omega =]0,1[^2$, $(x_0, y_0) \in \Omega$. The source term and the boundary conditions are chosen so as

$$u(x,y) = \left((x-x_0)^2 + (y-y_0)^2\right)^{\frac{\alpha}{2}}$$

is the solution of the Poisson problem.

1. Prove that the regularity of u critically depends on α:
$$u \in \begin{cases} H^1(\Omega), & \text{if } 0 < \alpha \leq 1, \\ H^2(\Omega), & \text{if } 1 < \alpha \leq 2, \\ H^3(\Omega), & \text{if } 2 < \alpha. \end{cases}$$

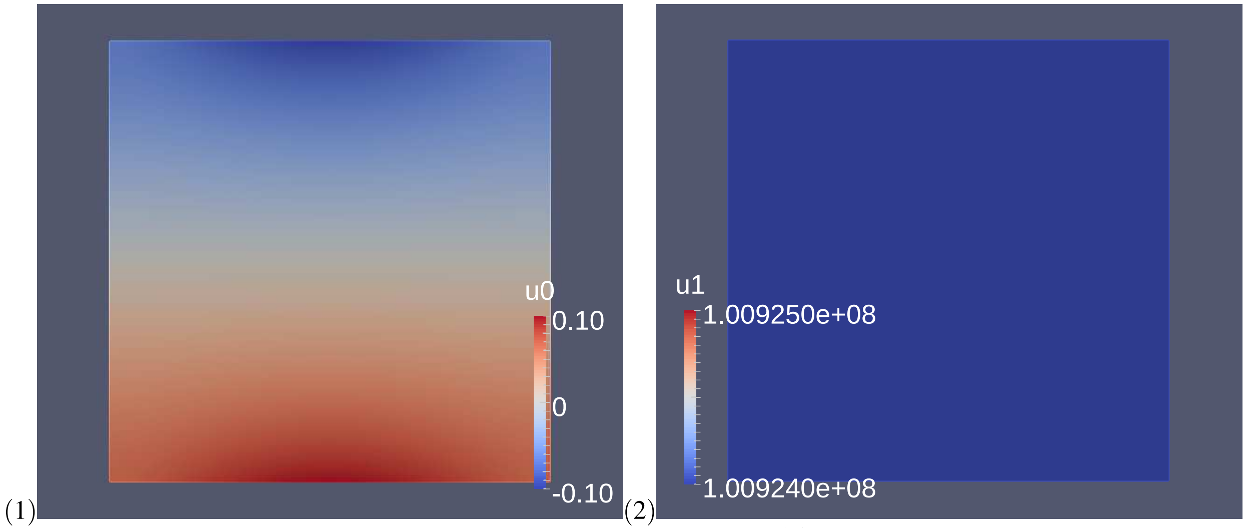


Figure 18. Solution of the Poisson-Neumann problem by penalization: (1) the compatibility condition is satisfied; (2) the compatibility condition is *not* satisfied.

2. Using `FreeFem++`, discuss the convergence rate of the approximations obtained by the $\mathbb{P}^1$ or $\mathbb{P}^2$ finite element method, depending on the value of α.

□

Solution.

1. For the sake of simplicity (and without loss of generality), we consider $\Omega = \mathscr{B}(0,1)$ (the unit ball) and $(x_0, y_0) = (0,0)$. Let us compute the derivatives of the function and their related L^2 integrability.

   - **At order 0**, we have
   $$u(x,y) = (x^2+y^2)^{\frac{\alpha}{2}}$$
   and, as a consequence,
   $$\int_{\mathscr{B}(0,1)} u^2 = \int_0^{2\pi}\int_0^1 r^{2\alpha}\, r\,\mathrm{d}r\,\mathrm{d}\theta = 2\pi\int_0^1 r^{2\alpha+1}\,\mathrm{d}r.$$
   Note that $r \mapsto r^{2\alpha+1} \in L^1(0,1)$ if, and only if, $2\alpha+1 > -1$.

   As a conclusion, $u \in L^2(\mathscr{B}(0,1))$ if, and only if, $\alpha > -1$.
   - **At order 1**, we have
   $$\begin{array}{rcl} \partial_x u(x,y) & = & \alpha x(x^2+y^2)^{\frac{\alpha}{2}-1}, \\ \partial_y u(x,y) & = & \alpha y(x^2+y^2)^{\frac{\alpha}{2}-1} \end{array}$$
   and, as a consequence,
   $$\begin{array}{rcl} \displaystyle\int_{\mathscr{B}(0,1)} |\nabla_x u|^2 & = & \displaystyle\alpha^2\int_0^{2\pi}\int_0^1 r^{2\alpha-2}\, r\,\mathrm{d}r\,\mathrm{d}\theta \\ & = & \displaystyle 2\pi\alpha^2\int_0^1 r^{2\alpha-1}\,\mathrm{d}r. \end{array}$$
   Note that $r \mapsto r^{2\alpha-1} \in L^1(0,1)$ if, and only if, $2\alpha-1 > -1$.

   As a conclusion, $u \in H^1(\mathscr{B}(0,1))$ if, and only if, $\alpha > 0$.
   - **At order 2**, we have
   $$\begin{array}{rcl} \partial_{xx} u(x,y) & = & \alpha(x^2+y^2)^{\frac{\alpha}{2}-1} \\ & & +\alpha(\alpha-2)x^2(x^2+y^2)^{\frac{\alpha}{2}-2}, \\ \partial_{xy} u(x,y) & = & \alpha(\alpha-2)xy(x^2+y^2)^{\frac{\alpha}{2}-2}, \\ \partial_{yy} u(x,y) & = & \alpha(x^2+y^2)^{\frac{\alpha}{2}-1} \\ & & +\alpha(\alpha-2)y^2(x^2+y^2)^{\frac{\alpha}{2}-2} \end{array}$$

and
$$\partial_{yx} u = \partial_{xy} u.$$

Then, computing $(\partial_{xx} u)^2$, $(\partial_{xy} u)^2$, $(\partial_{yy} u)^2$, integrating over $\mathscr{B}(0,1)$ and using the change of coordinates $(x,y) = (r\cos\theta, r\sin\theta)$, we get
$$\begin{array}{rcl} \displaystyle\int_{\mathscr{B}(0,1)} |\partial_{xx} u|^2 & = & \displaystyle C_1\int_0^1 r^{2\alpha-3}\,\mathrm{d}r, \\ \displaystyle\int_{\mathscr{B}(0,1)} |\partial_{xy} u|^2 & = & \displaystyle C_2\int_0^1 r^{2\alpha-3}\,\mathrm{d}r, \\ \displaystyle\int_{\mathscr{B}(0,1)} |\partial_{yy} u|^2 & = & \displaystyle C_3\int_0^1 r^{2\alpha-3}\,\mathrm{d}r \end{array}$$
with
$$\begin{array}{rcl} C_1 & = & \displaystyle 2\pi\alpha^2 + \alpha^2(\alpha-2)^2\int_0^{2\pi}\cos^4\theta\,\mathrm{d}\theta \\ & & \displaystyle +\alpha^2(\alpha-2)\int_0^{2\pi}\cos^2\theta\,\mathrm{d}\theta, \\ C_2 & = & \displaystyle \alpha^2(\alpha-2)^2\int_0^{2\pi}\cos^2\theta\,\sin^2\theta\,\mathrm{d}\theta, \\ C_3 & = & \displaystyle 2\pi\alpha^2 + \alpha^2(\alpha-2)^2\int_0^{2\pi}\sin^4\theta\,\mathrm{d}\theta \\ & & \displaystyle +\alpha^2(\alpha-2)\int_0^{2\pi}\sin^2\theta\,\mathrm{d}\theta. \end{array}$$

Note that $r \mapsto r^{2\alpha-3} \in L^1(0,1)$ if, and only if, $2\alpha - 3 > -1$.

As a conclusion, $u \in H^2(\mathscr{B}(0,1))$ if, and only if, $\alpha > 1$.

- **At order 3**, we have

$$\begin{aligned}
&\partial_{xxx}u(x,y)\\
&= C_0\left(3x(x^2+y^2)^{\frac{\alpha}{2}-2} + x^3(\alpha-4)(x^2+y^2)^{\frac{\alpha}{2}-3}\right),\\
&\partial_{xxy}u(x,y)\\
&= C_0\left(y(x^2+y^2)^{\frac{\alpha}{2}-2} + (\alpha-4)x^2y(x^2+y^2)^{\frac{\alpha}{2}-3}\right),\\
&\partial_{yyx}u(x,y)\\
&= C_0\left(x(x^2+y^2)^{\frac{\alpha}{2}-2} + (\alpha-4)y^2x(x^2+y^2)^{\frac{\alpha}{2}-3}\right),\\
&\partial_{yyy}u(x,y)\\
&= C_0\left(3y(x^2+y^2)^{\frac{\alpha}{2}-2} + y^3(\alpha-4)(x^2+y^2)^{\frac{\alpha}{2}-3}\right)
\end{aligned}$$

with $C_0 = \alpha(\alpha-2)$, and

$$\partial_{yxx}u = \partial_{xyx}u = \partial_{xxy}u, \quad \partial_{xyy}u = \partial_{yxy}u = \partial_{yyx}u.$$

Then, computing $(\partial_{xxx}u)^2$, $(\partial_{xxy}u)^2$, $(\partial_{yyx}u)^2$, $(\partial_{yyy}u)^2$, integrating over $\mathscr{B}(0,1)$ and using the change of coordinates $(x,y) = (r\cos\theta, r\sin\theta)$, we get

$$\begin{aligned}
\int_{\mathscr{B}(0,1)} |\partial_{xxx}u|^2 &= C_1' \int_0^1 r^{2\alpha-5}\,\mathrm{d}r,\\
\int_{\mathscr{B}(0,1)} |\partial_{xxy}u|^2 &= C_2' \int_0^1 r^{2\alpha-5}\,\mathrm{d}r,\\
\int_{\mathscr{B}(0,1)} |\partial_{yyx}u|^2 &= C_3' \int_0^1 r^{2\alpha-5}\,\mathrm{d}r,\\
\int_{\mathscr{B}(0,1)} |\partial_{yyy}u|^2 &= C_4' \int_0^1 r^{2\alpha-5}\,\mathrm{d}r,
\end{aligned}$$

where C_i' denotes a constant[13]. Note that $r \mapsto r^{2\alpha-5} \in L^1(0,1)$ if, and only if, $2\alpha - 5 > -1$.

As a conclusion, $u \in H^3(\mathscr{B}(0,1))$ if, and only if, $\alpha > 2$.

2. Figure 19 represents the solution for different values of α. In particular the behaviour of the solution near the point $(0.5, 0.5)$ illustrates the regularity issue that has been discussed above.

Let us focus on the numerical computation of the solution. By Lemma 4.1 and the approximability property of $\mathbb{P}^1$ and $\mathbb{P}^2$ with respect to $H^1(\Omega)$, the numerical solution u_h converges to the exact solution u in $H^1(\Omega)$. But what is the rate of convergence?

Theorem 5.10 states that the $\mathbb{P}^1$ method converges at first order in H^1, provided the solution is at least H^2.

Theorem 5.19 states that the $\mathbb{P}^2$ method converges at second order in H^1, provided the solution is at least H^3.

The convergence in the L^2 norm illustrates Theorem 5.11 when $u \in H^2(\Omega)$. Note that the adjoint problem is identical to the initial problem by symmetry of the bilinear form so that it satisfies the *elliptic regularity property* since the domain Ω is convex.

Focusing on the error in the H^1 norm, Figure 20 illustrates the above theorems by observing the numerical orders of convergence for $\alpha = 0.5$, 1.5 and 2.5. Optimality of the finite element method is achieved with the $\mathbb{P}^1$ elements for $\alpha = 1.5$ and $\alpha = 2.5$ (with a numerical order that is close to 1) but not for $\alpha = 0.5$: in this case the convergence is suboptimal because the solution does not belong to H^2. Optimality of the finite element method is achieved with the $\mathbb{P}^2$ elements for $\alpha = 2.5$ (with a numerical order that is close to 2) but not for $\alpha = 0.5$ and $\alpha = 1.5$: in these cases the convergence is suboptimal because the solution does not belong to H^3. Similar observations can be led with the L^2 analysis.

□

7.6 Finite elements for the Stokes system

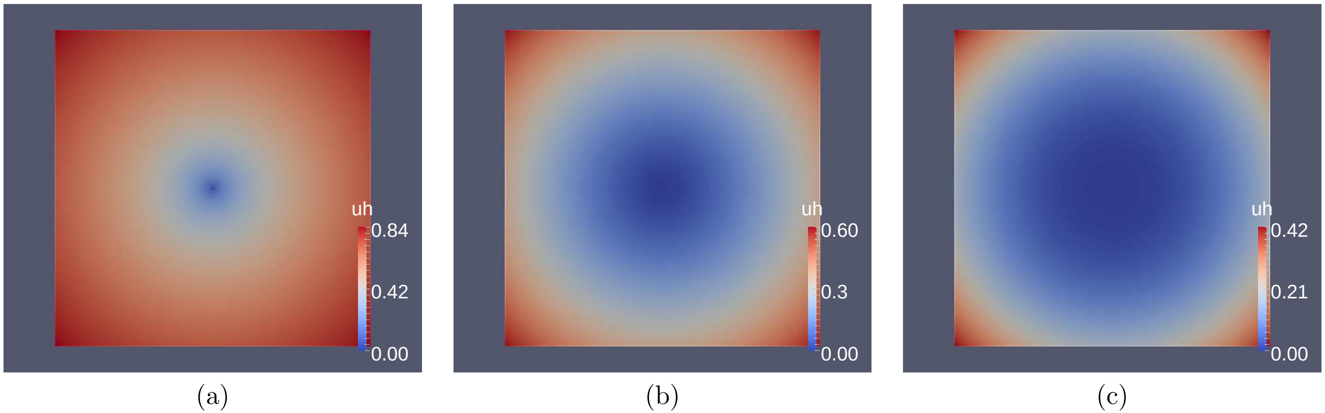
Problem. Write a `FreeFem++` program to investigate the finite element method applied to the Stokes problem. □

Solution. The major issue discussed in Section 6, is whether or not the couple of approximation spaces chosen for a variationnal problem in mixed form satisfies the (uniform) inf-sup condition at the discrete level. In order to get well-posedness of the discretized Stokes problem, the functional spaces X_h and M_h (for the velocity and pressure, respectively) should indeed satisfy the inf-sup condition. Figures 21 to 23 exhibit the numerical solution of a Stokes problem with different approximation spaces.

Consider the domain $\Omega =]0,1[^2$ and $\mathscr{B}$ denotes the ball of center $(0.5, 0.5)$ and radius $r = 0.25$ and define the components of the source term f as $f_1 = f_2 = 50 \times \mathbf{1}_{\mathscr{B}}$.

[13] with

$$\begin{aligned}
C_1' &= C_0^2\Big(9\textstyle\int_0^{2\pi}\cos^2\theta\,\mathrm{d}\theta + (\alpha-4)^2\int_0^{2\pi}\cos^6\theta\,\mathrm{d}\theta\\
&\qquad +6(\alpha-4)\textstyle\int_0^{2\pi}\cos^4\theta\,\mathrm{d}\theta\Big),\\
C_2' &= C_0^2\Big(\textstyle\int_0^{2\pi}\sin^2\theta\,\mathrm{d}\theta + (\alpha-4)^2\int_0^{2\pi}\cos^4\theta\,\sin^2\theta\,\mathrm{d}\theta\\
&\qquad +(\alpha-4)\textstyle\int_0^{2\pi}\cos^2\theta\,\sin^2\theta\,\mathrm{d}\theta\Big),\\
C_3' &= C_0^2\Big(\textstyle\int_0^{2\pi}\cos^2\theta\,\mathrm{d}\theta + (\alpha-4)^2\int_0^{2\pi}\sin^4\theta\,\cos^2\theta\,\mathrm{d}\theta\\
&\qquad +(\alpha-4)\textstyle\int_0^{2\pi}\cos^2\theta\,\sin^2\theta\,\mathrm{d}\theta\Big),\\
C_4' &= C_0^2\Big(9\textstyle\int_0^{2\pi}\sin^2\theta\,\mathrm{d}\theta + (\alpha-4)^2\int_0^{2\pi}\sin^6\theta\,\mathrm{d}\theta\\
&\qquad +6(\alpha-4)\textstyle\int_0^{2\pi}\sin^4\theta\,\mathrm{d}\theta\Big).
\end{aligned}$$



(a) (b) (c)

Figure 19. Function $u : (x,y) \mapsto ((x-0.5)^2+(y-0.5)^2)^{\frac{\alpha}{2}}$ for different values of α: (a) $\alpha = 0.5$, (c) $\alpha = 1.5$, (c) $\alpha = 2.5$.

We consider the (variational) Stokes problem:

$$\left\{\begin{array}{l} \text{Find } (u,p) \in (H_0^1(\Omega))^2 \times L_0^2(\Omega) \text{ such that} \\ \begin{array}{rcl} \displaystyle\int_\Omega \nabla u : \nabla v - \int_\Omega p \operatorname{div}(v) & = & \displaystyle\int_\Omega f \cdot v, \\ \displaystyle\int_\Omega q \operatorname{div}(u) & = & 0, \end{array} \\ \text{for all } (v,q) \in (H_0^1(\Omega))^2 \times L_0^2(\Omega). \end{array}\right.$$

In order to illustrate the influence of the choice of the approximation spaces, we use a *structured* mesh[14].

- Figure 21 deals with $\mathbb{P}^1 - \mathbb{P}^0$ finite elements: the *locking effect* is illustrated, see Proposition 6.2, as the numerical velocity field is 0, which evidences the failure of the approximation.
- Figure 22 deals with $\mathbb{P}^1 - \mathbb{P}^1$ finite elements: the *checkerboard effect* on the pressure field is illustrated. In particular, spurious oscillations strongly depend on the mesh size, producing an unrealistic pressure field. The $\mathbb{P}^1 - \mathbb{P}^1$ finite elements have the same drawback as the $\mathbb{Q}^1 - \mathbb{P}^0$ finite elements for which the checkerboard effect has been proved, see Proposition 6.3.
- Figure 23 deals with $\mathbb{P}_b^1 - \mathbb{P}^1$ finite elements, which are inf-sup stable.
- Figure 24 deals with $\mathbb{P}^2 - \mathbb{P}^1$ finite elements, which are also inf-sup stable. Besides, these finite elements are more precise than the $\mathbb{P}_b^1 - \mathbb{P}^1$ finite elements. As a consequence, as the exact solution is regular, the numerical simulation provides a better approximation at fixed size mesh.

□

7.7 Uzawa algorithm for saddle-point problems

Problem. Consider a domain Ω and a subdomain $\mathscr{B} \subset \Omega$. Let $f \in L^2(\Omega)$. Define

$$J(v) := \frac{1}{2}\int_\Omega |\nabla u|^2 + \frac{1}{2}\int_\Omega u^2 - \int_\Omega fv,$$

$$V = \left\{ v \in H^1(\Omega),\ \int_{\mathscr{B}} v = 0 \right\}.$$

Solve with `FreeFem++` the minimization problem:

$$\left\{\begin{array}{l} \text{Find } u \in V \text{ such that} \\ J(u) = \min\limits_{v \in V} J(v). \end{array}\right.$$

□

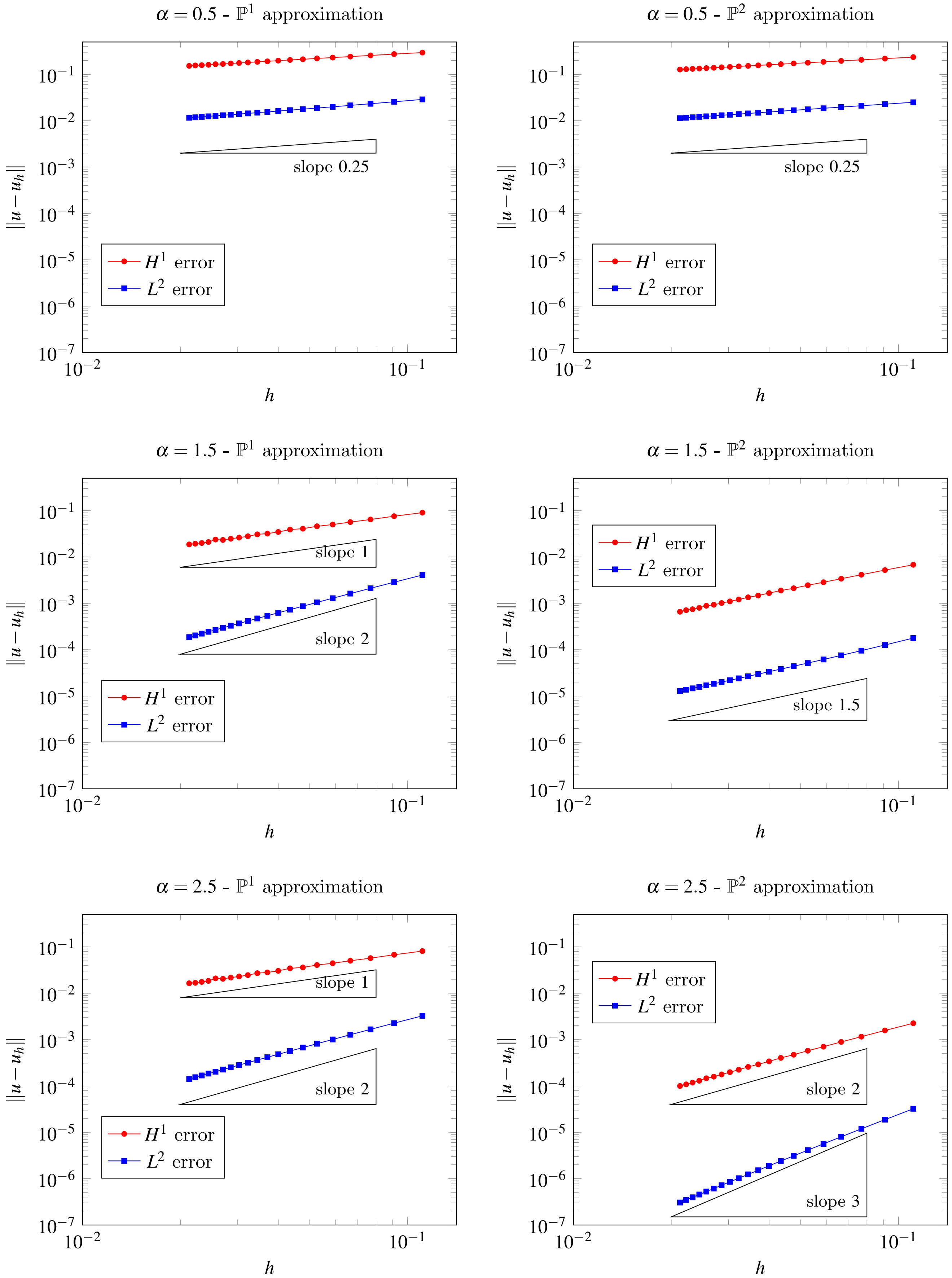
Solution. This problem is the application of Example 1 developped in Section 3. By the Lax-Milgram theorem, the minimization problem is equivalent to a variational formulation associated to the functional space V. From the numerical point of view, dealing with V is difficult because we cannot build finite elements in a finite dimensional subspace of V, because of the constraint. Nevertheless, the solution u of the minimization problem is the first component of the solution of the saddle-point problem:

$$(Q)\left\{\begin{array}{l} \text{Find } (u,\lambda) \in H^1(\Omega) \times \mathbb{R} \text{ such that} \\ \begin{array}{rcl} \displaystyle\int_\Omega \nabla u \cdot \nabla v + \int_\Omega uv + \lambda \int_{\mathscr{B}} v & = & \displaystyle\int_\Omega fv, \\ \displaystyle\mu \int_{\mathscr{B}} u & = & 0, \end{array} \\ \text{for all } (v,\mu) \in H^1(\Omega) \times \mathbb{R}. \end{array}\right.$$

We can prove that problem (Q) is well-posed (see Subsection 3.3). The main advantage of this formulation is that (Q) is now a problem without constraint.

Uzawa algorithm. From the numerical point of view, the solution of the (discretized version of the) saddle-point problem can be defined by solving directly the

[14] The *locking* effect associated to the $\mathbb{P}^1 - \mathbb{P}^0$ finite elements is proved under the assumption of a *structured triangular mesh*, see Proposition 6.2, page 46.



$\alpha = 0.5$ - $\mathbb{P}^1$ approximation

$\alpha = 0.5$ - $\mathbb{P}^2$ approximation

$\alpha = 1.5$ - $\mathbb{P}^1$ approximation

$\alpha = 1.5$ - $\mathbb{P}^2$ approximation

$\alpha = 2.5$ - $\mathbb{P}^1$ approximation

$\alpha = 2.5$ - $\mathbb{P}^2$ approximation

Figure 20. Error analysis of the finite element method in H^1 and L^2 for various regularity of the exact solution. From top to bottom $\alpha = 0.5$, $\alpha = 1.5$, and $\alpha = 2.5$.

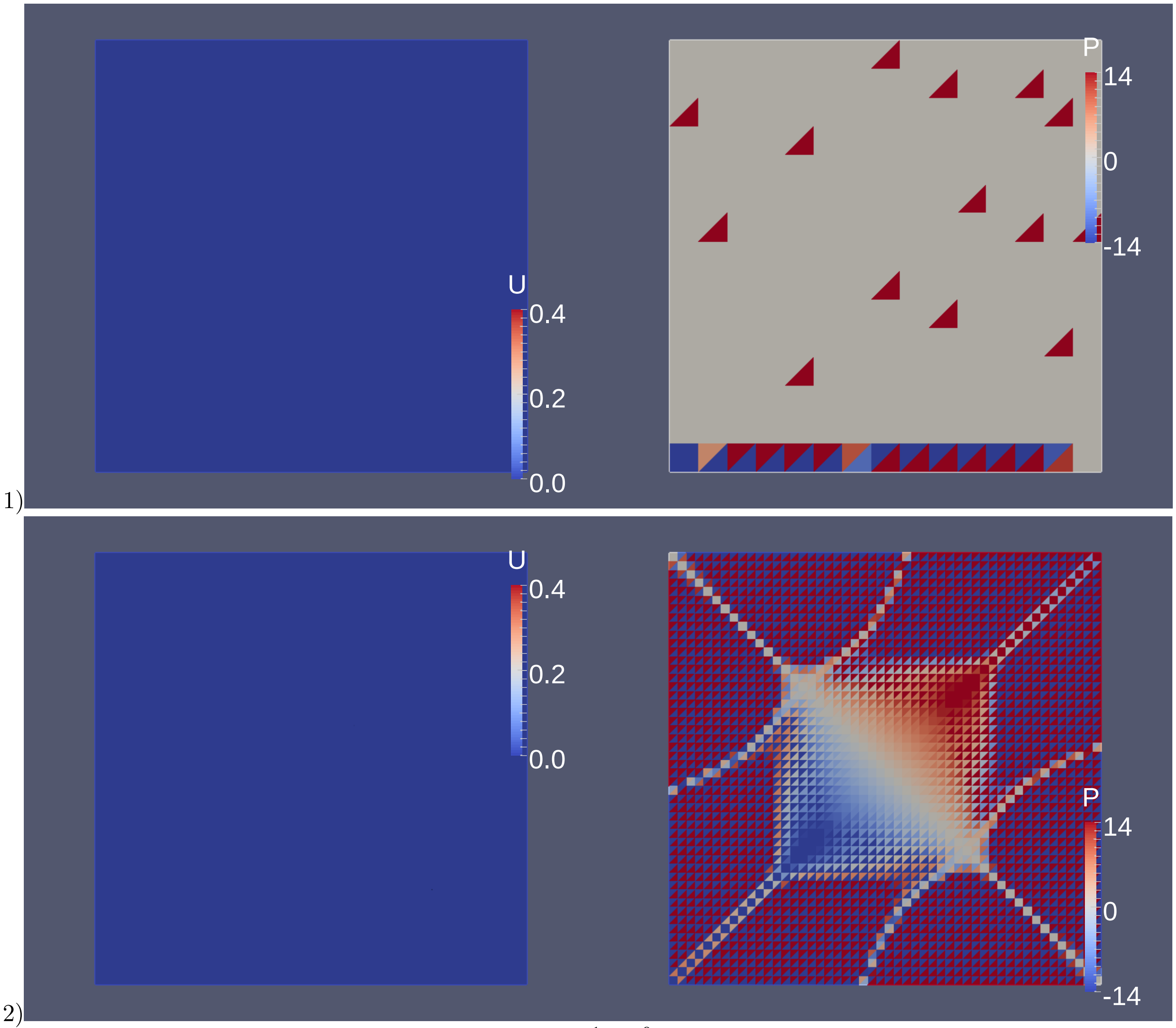


Figure 21. Numerical solution of the Stokes system with $\mathbb{P}^1 - \mathbb{P}^0$ finite elements. Velocity field and pressure field for 1) $h = 0.067$ and 2) $h = 0.020$.

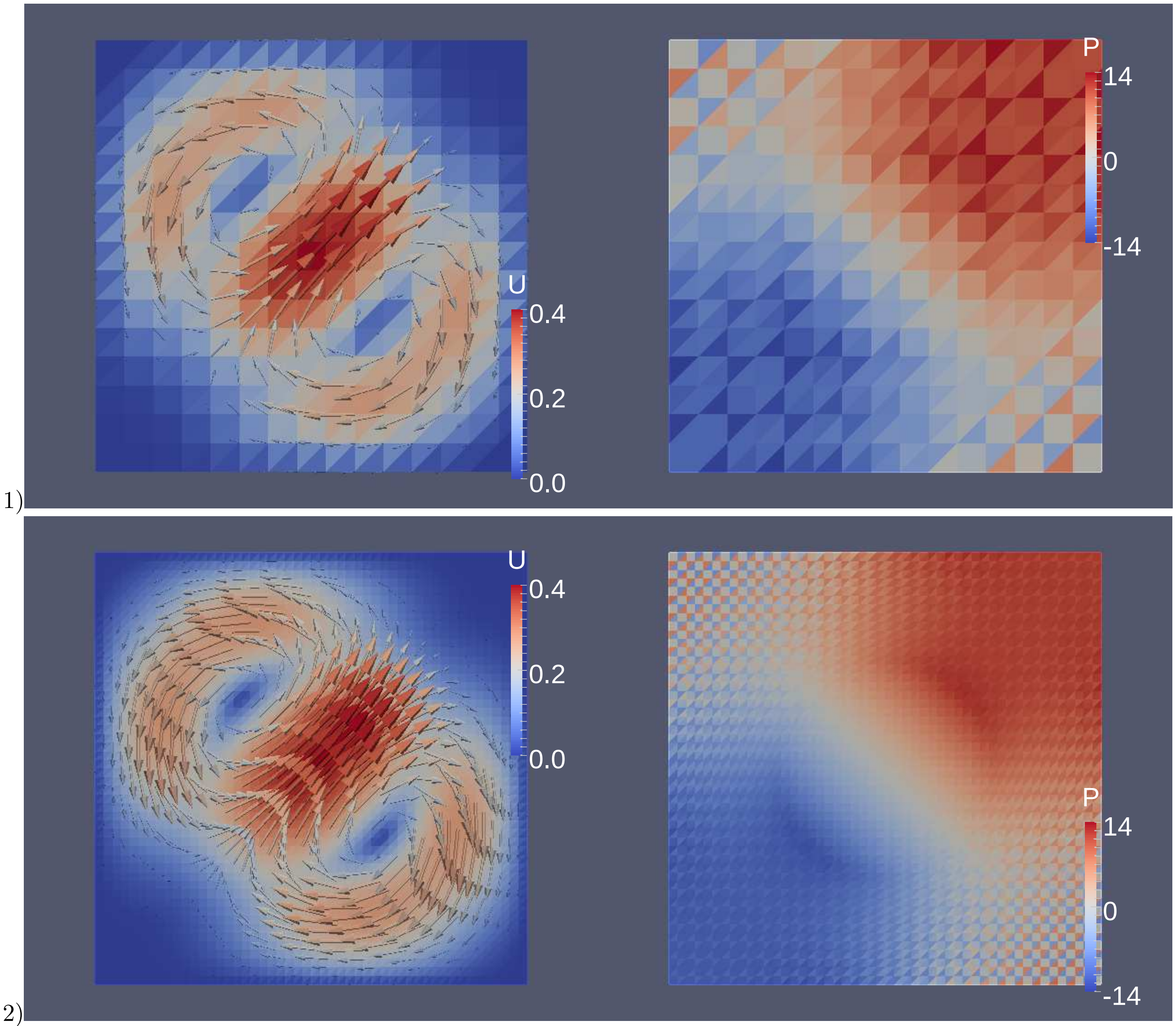


Figure 22. Numerical solution of the Stokes system with $\mathbb{P}^1 - \mathbb{P}^1$ finite elements. Velocity field and pressure field for 1) $h = 0.067$ and 2) $h = 0.020$.

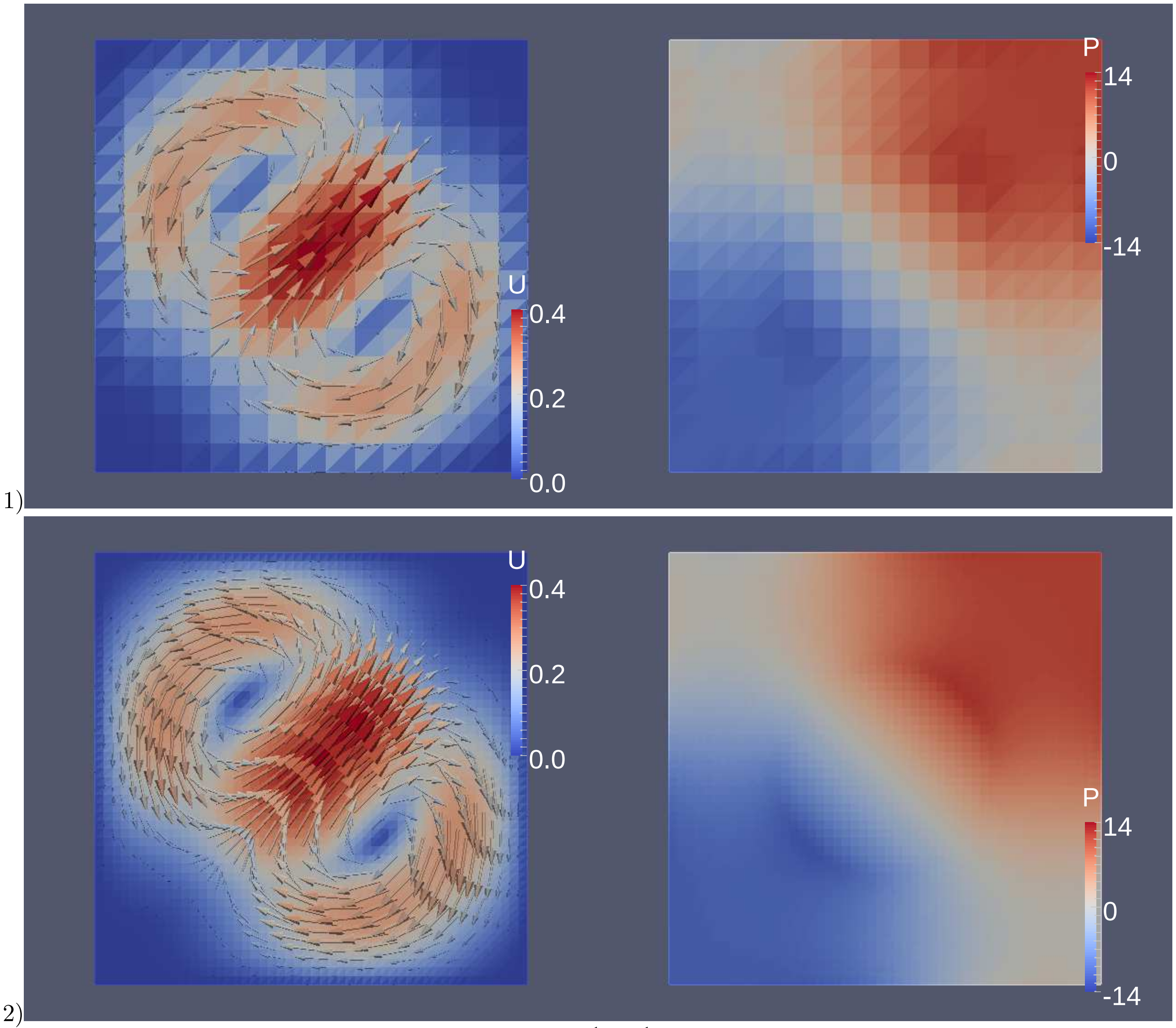


Figure 23. Numerical solution of the Stokes system with $\mathbb{P}_b^1 - \mathbb{P}^1$ finite elements. Velocity field and pressure field for 1) $h = 0.067$ and 2) $h = 0.020$.

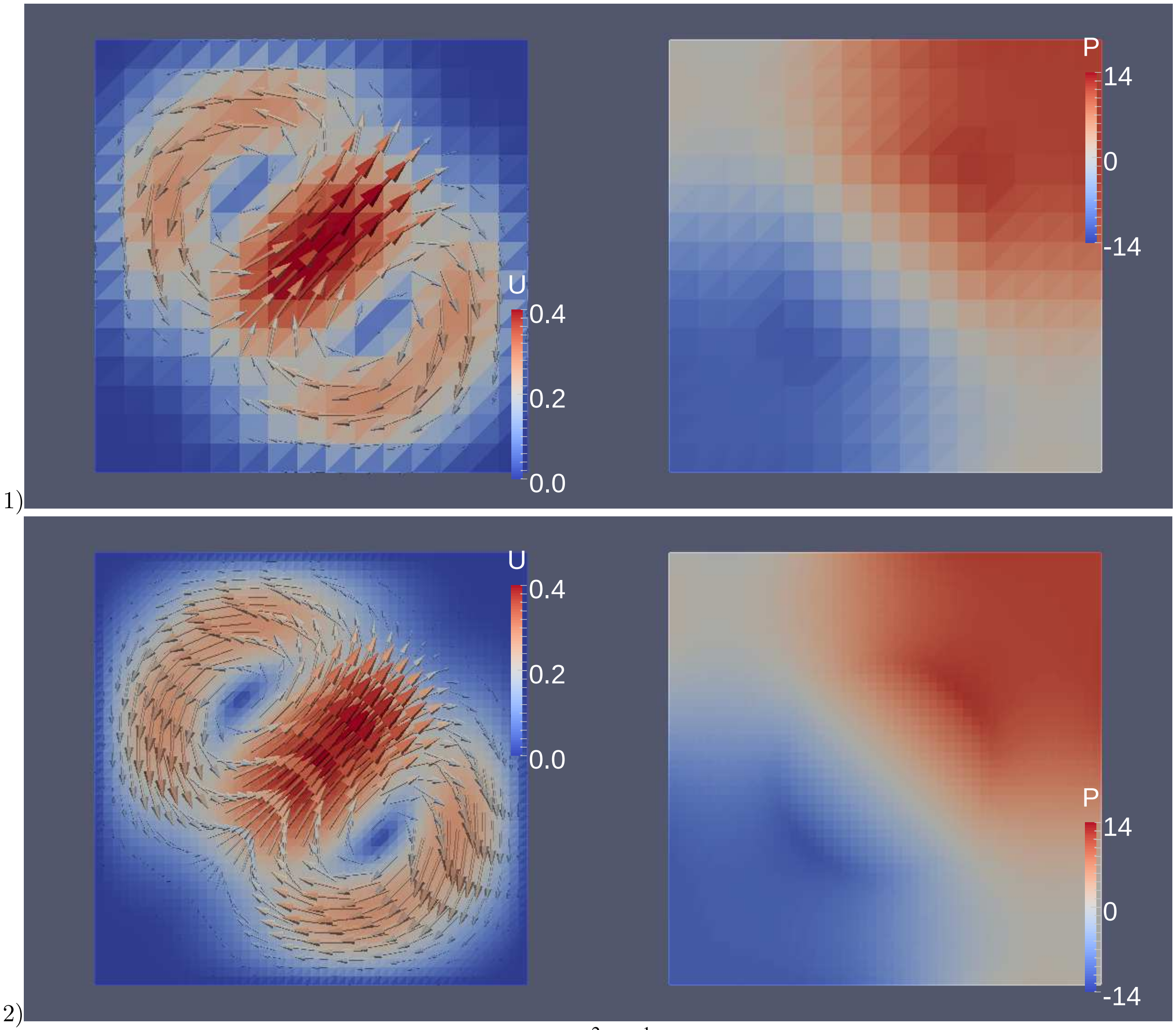


Figure 24. Numerical solution of the Stokes system with $\mathbb{P}^2 - \mathbb{P}^1$ finite elements. Velocity field and pressure field for 1) $h = 0.067$ and 2) $h = 0.020$.

corresponding linear system. However, by nature, this linear system does not have a nice symmetric definite positive structure for which efficient solver can be used. That is the reason why many specific iterative algorithms have been developed in the literature to solve such linear systems.

Here we present one of those alternative methods, which is iterative: for each iteration, a first step only requires to solve the *elliptic* contribution of the problem and then a second step consists in updating the Lagrange multiplier. The Uzawa algorithm is often used to solve saddle-point problems. In the literature it is often presented in a finite dimensional framework but it makes sense also in the infinite dimensional framework: if a solution exists, then the algorithm converges! Let us present the algorithm in different forms.

- **Abstract problem**. Let X and M be Hilbert spaces, $a(\cdot,\cdot)$ a continuous bilinear form on $X\times X$, b a continuous bilinear form on $X\times M$. For all $L\in X'$, for all $G\in M'$, we aim at solving the variational problem

$$\left\{\begin{array}{l} \text{Find } (u,p)\in X\times M \text{ such that} \\ \quad a(u,v)+b(v,p) \;=\; L(v), \\ \qquad\qquad\;\; b(u,q) \;=\; G(q), \\ \text{for all } (v,q)\in X\times M. \end{array}\right.$$

Then, for a given parameter $\rho>0$, the Uzawa algorithm writes:

```
Choose $p^0 \in M$.
for $n = 0,...,+\infty$
    1. Solve the elliptic problem
       Find $u^{n+1} \in X$ such that
       $a(u^{n+1},v) = L(v) - b(v,p^n)$,
       for all $v \in X$.
    2. Update the Lagrange multiplier
       Find $p^{n+1} \in M$ such that
       $(p^{n+1},q) = (p^n,q) + \rho(b(u^{n+1},q) - G(q))$,
       for all $q \in M$.
enddo
```

- **Operators**. Defining $A:X\to X'$ as $\langle Au,\cdot\rangle_{X',X} = a(u,\cdot)$ and $B:X\to M'$ as $\langle Bv,\cdot\rangle_{M',M} = b(v,\cdot)$, the formulation is equivalent to

$$\left\{\begin{array}{l} \text{Find } (u,p)\in X\times M \text{ such that} \\ \quad Au+B'p \;=\; L, \\ \qquad\quad Bu \;=\; G, \end{array}\right.$$

where $B':M\to X'$ is the adjoint operator of B, the bidual of M being identified to M itself. Then the Uzawa algorithm writes:

```
Choose $p^0 \in M$.
for $n = 0,...,+\infty$
    1. Solve for the principal unknown:
       Find $u^{n+1} \in X$ such that
       $Au^{n+1} = L - B'p^n$,
    2. Update the Lagrange multiplier:
       Define $p^{n+1} \in M$ by
       $p^{n+1} = p^n + \rho(Bu^{n+1} - G)$.
enddo
```

Proposition 7.7 (Uzawa algorithm) *Assume that the saddle-point problem admits a solution. Then the sequence $\{u^n\}$ defined by the Uzawa algorithm converges to the solution of the corresponding minimization problem if*

$$0<\rho<\frac{2\alpha}{\|B\|^2},$$

where α is the coercivity constant of $a(\cdot,\cdot)$.

Proof of Proposition 7.7. We have

$$\begin{array}{lcccl} p^{n+1} & = & p^n & + & \rho(Bu^{n+1}-G), \\ p & = & p & + & \rho(Bu-G), \end{array}$$

hence

$$p^{n+1}-p = p^n-p+\rho B(u^{n+1}-u).$$

Then

$$\begin{array}{rl} \|p^{n+1}-p\|_M^2 & \\ = & \|p^n-p\|_M^2+2\rho\left(p^n-p,Bu^{n+1}-u\right)_M \\ & +\rho^2\|B(u^{n+1}-u)\|_M^2 \\ = & \|p^n-p\|_M^2+2\rho\left(B'(p^n-p),u^{n+1}-u\right)_X \\ & +\rho^2\|B(u^{n+1}-u)\|_M^2 \\ = & \|p^n-p\|_M^2-2\rho\left(A(u^{n+1}-u),u^{n+1}-u\right)_X \\ & +\rho^2\|B(u^{n+1}-u)\|_M^2 \\ \leq & \|p^n-p\|_M^2-\rho(2\alpha-\rho\|B\|^2)\|u^{n+1}-u\|_X^2. \end{array}$$

Thus if the condition on ρ is satisfied, the series

$$\sum_n \|u^{n+1}-u\|_X^2,$$

is convergent. As a consequence, $\{u^n\}$ converges to u.

The equation satisfied by p^n-p is thus

$$B'(p^n-p) = -A(u^{n+1}-u),$$

which gives, for any $v\in X$

$$b(p^n-p,v) = -a(u^{n+1}-u,v)\leq \|a\|\|u^{n+1}-u\|\|v\|.$$

By the inf-sup condition we obtain

$$\beta\|p^n-p\|\leq \sup_{v\in X}\frac{b(p^n-p,v)}{\|v\|}\leq \|a\|\|u^{n+1}-u\|,$$

and the convergence of p^n to p is proved. ■

In practice, the algorithm stops running when an error criterion on the Lagrange multiplier is met.

Application. Let us apply the Uzawa algorithm to our problem. It reads

```
Define $\lambda^0 = 0$.
for $n = 0, ..., +\infty$
    1. Solve the problem
       $\left\{ \begin{array}{l} \text{Find } u^{n+1} \in H^1(\Omega) \text{ such that} \\ \int_\Omega \nabla u^{n+1} \cdot \nabla v + \int_\Omega u^{n+1} v = \int_\Omega f v - \lambda^n \int_{\mathscr{B}} v, \\ \text{for all } v \in H^1(\Omega). \end{array} \right.$
    2. Solve the problem
       $\left\{ \begin{array}{l} \text{Find } \lambda^{n+1} \in \mathbb{R} \text{ such that} \\ \lambda^{n+1} = \lambda^n + \rho \int_{\mathscr{B}} u^{n+1}. \end{array} \right.$
enddo
```

Let us compute the range of admissible values for the Uzawa parameter ρ in our constrained problem. *For the sake of simplicity, we assume that* $\Omega =]0,1[^2$. The coercivity constant is $\alpha = 1$. Then, for all $u \in H^1(\Omega)$,

$$|Bu| = \left| \int_{\mathscr{B}} u \right| \leq \int_{\mathscr{B}} |u| \leq |\mathscr{B}|^{\frac{1}{2}} \, \|u\|_{L^2(\Omega)} \leq |\mathscr{B}|^{\frac{1}{2}} \, \|u\|_{H^1(\Omega)},$$

so that

$$\|B\| := \sup_{u \in H^1(\Omega)} \frac{|Bu|}{\|u\|_{H^1(\Omega)}} \leq |\mathscr{B}|^{\frac{1}{2}}.$$

Moreover, taking $\tilde{u} \equiv 1$ yields

$$\frac{|B\tilde{u}|}{\|\tilde{u}\|_{H^1(\Omega)}} = \frac{|\mathscr{B}|}{|\Omega|^{\frac{1}{2}}} = |\mathscr{B}|.$$

Thus,

- if $|\mathscr{B}| = 1$ (i.e. if $\mathscr{B} = \Omega$), then the bound is obviously attained and
$$\|B\| = 1.$$
- if $|\mathscr{B}| < 1$, we have
$$\|B\| = |\mathscr{B}|,$$
which is consistent with the case $|\mathscr{B}| = 1$.

Then the Uzawa parameter should be chosen as

$$0 < \rho < \rho_{\max} := \frac{2}{|\mathscr{B}|^2}.$$

Numerical results. Let us present some numerical simulations: $\mathscr{B}$ is the disk of center $(0.5, 0.5)$ and radius $r = 0.2$. We define the source term as

$$f(x,y) = \left\{ \begin{array}{rl} 1, & \text{if } (x,y) \in \mathscr{B}, \\ -1, & \text{if } (x,y) \notin \mathscr{B}. \end{array} \right.$$

In practice the Uzawa algorithm stops running when the user estimates that convergence of $\{\lambda^n\}$ has been numerically reached. Thus we define for instance $\texttt{tol} = 10^{-8}$ and the computations will stop as soon as

$$\left| \lambda^{n+1} - \lambda^n \right| < \texttt{tol},$$

or, alternatively,

$$\frac{\left| \lambda^{n+1} - \lambda^n \right|}{\rho} < \texttt{tol}$$

in which case the stopping test addresses the numerical constraint $\|Bu^n - G\|$.

- Choosing
$$\rho = \frac{\rho_{\max}}{2},$$
the algorithm converges and produces the numerical solution (u_h, λ_h) where u_h is represented on Figure 25 and $\lambda_h = -5.25807$. Actually, with our fixed tolerance and our choice for ρ, the algorithm converges in 11 iterations and, in the end, satisfies $\int_B u_h = 5.86 \cdot 10^{-11}$ which is quite satisfactory.
- As it was outlined the choice of ρ has a critical impact on the behaviour of the algorithm. Figure 26 exhibits the behaviour of the sequence $\{\lambda^n\}$ for various values of the Uzawa parameter: for $\rho > \rho_{\max}$, the method does not converge. For $\rho < \rho_{max}$ the method converges but the choice of ρ critically rules the the rate of convergence: it is optimal for $\rho = 0.5 \cdot \rho_{max}$ whereas it is deteriorated when too small or too close to ρ_{max}.

Remark 7.8 *The Stokes problem can be solved iteratively with the Uzawa algorithm: the elliptic part reduces to a vector Laplace problem (for which all the component of the velocity can be uncoupled) whereas the update of the Lagrange multiplier (the pressure field) reduces to a simple computation. Let* Ω *be a bounded domain,* $f \in L^2(\Omega)$ *and consider the Stokes problem:*

$$\left\{ \begin{array}{l} \textit{Find } (u,p) \in (H_0^1(\Omega))^2 \times L_0^2(\Omega) \textit{ such that} \\ \begin{array}{rcl} \int_\Omega \nabla u : \nabla v - \int_\Omega p \operatorname{div}(v) & = & \int_\Omega f \cdot v, \\ \int_\Omega q \operatorname{div}(u) & = & 0, \end{array} \\ \textit{for all } (v,q) \in (H_0^1(\Omega))^2 \times L_0^2(\Omega). \end{array} \right.$$

Then the Uzawa algorithm writes:

```
Choose $p^0 \equiv 0$.
for $n = 0, ..., +\infty$
    1. Solve the elliptic problem
       $\left\{ \begin{array}{l} \textit{Find } u^{n+1} \in (H_0^1(\Omega))^2 \textit{ such that} \\ \int_\Omega \nabla u^{n+1} : \nabla v = \int_\Omega p^n \operatorname{div}(v) + \int_\Omega f \cdot v, \\ \textit{for all } v \in (H_0^1(\Omega))^2. \end{array} \right.$
    2. Update the Lagrange multiplier
       $\left\{ \begin{array}{l} \textit{Find } p^{n+1} \in L^2(\Omega) \textit{ such that} \\ \int_\Omega p^{n+1} q = \int_\Omega p^n q - \rho \int_\Omega q \operatorname{div}(u^{n+1}), \\ \textit{for all } q \in L^2(\Omega). \end{array} \right.$
enddo
```

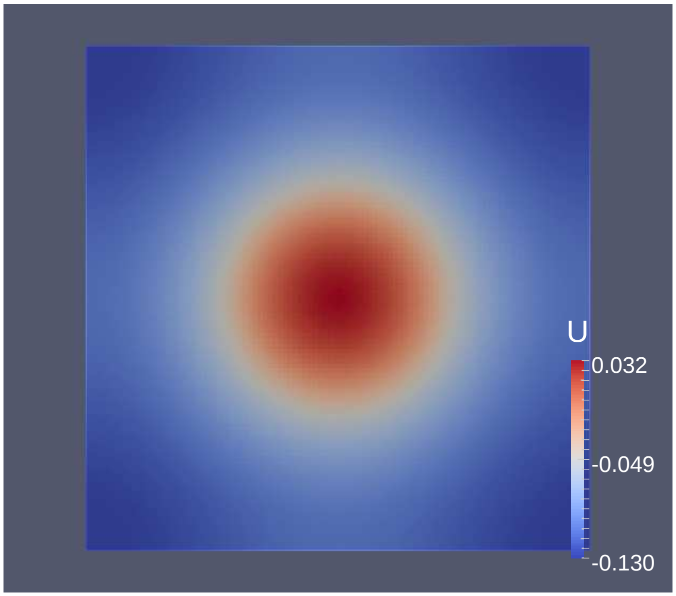


Figure 25. Solution of the saddle-point problem with the Uzawa algorithm. The solution u_h numerically satisfies $\int_{\mathscr{B}} u_h = 0$.

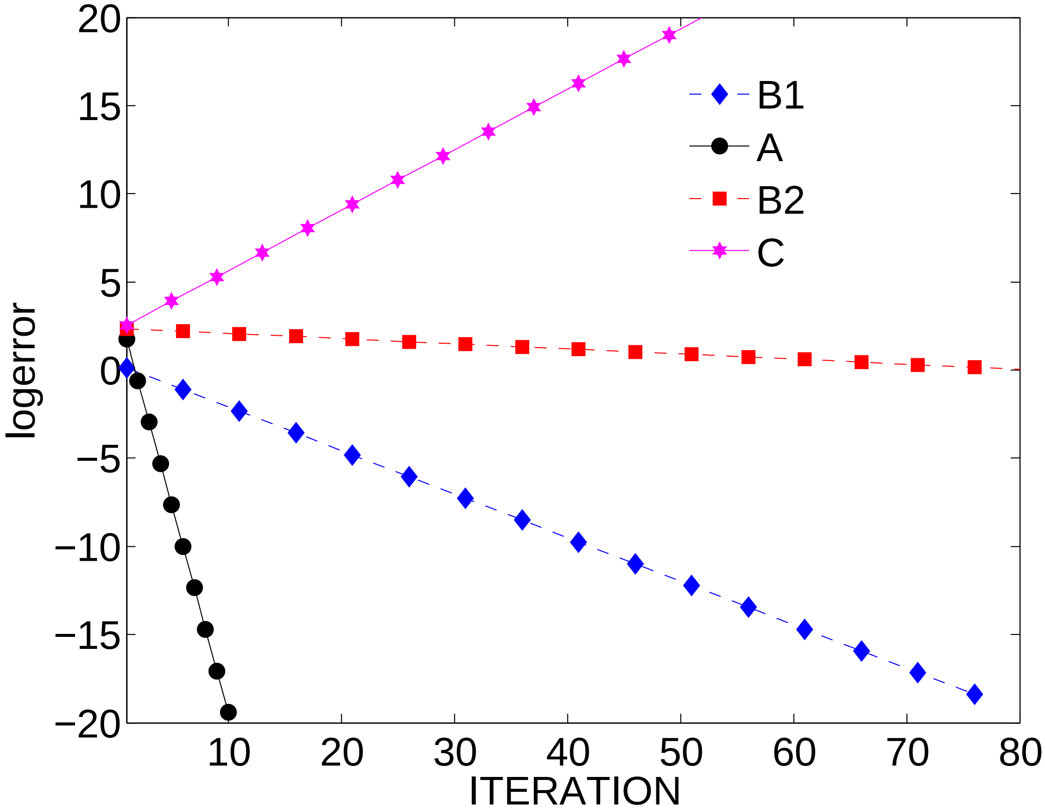


Figure 26. Convergence of the Lagrange multiplier with the Uzawa algorithm for different values of ρ.

Notice that the initialization $p^0 \equiv 0$ guarantees that the Lagrange multiplier has zero mean value at each time step: choosing $q \equiv 1$ as a test function, we get by the Lagrange multiplier equation

$$\begin{aligned}\int_\Omega p^{n+1} &= \int_\Omega p^n - \rho \int_\Omega \operatorname{div}(u^{n+1}) \\ &= \int_\Omega p^n - \rho \int_{\partial\Omega} u^{n+1} \cdot n \\ &= \int_\Omega p^n,\end{aligned}$$

hence the property inherited from the initialization step.

□

7.8 A fluid-structure interaction problem

Problem. We aim at describing the interaction between a rigid particle and an incompressible Newtonian fluid. Let Ω be the unit square in $\mathbb{R}^2$. Let $B(t) \subset \Omega$ a rigid particle with a center of mass $x_{B(t)}$, angular position $\theta_{B(t)}$. Inertial effects are neglected, which leads us to consider the instantaneous equilibrium of forces within the fluid and for the particle. Thus, at each time t, we consider the Stokes equations in the fluid domain:

$$(\mathrm{FSI}_1)\left\{\begin{array}{rcll} -\operatorname{div}(2\mu\mathbb{D}(u) - p\mathbb{I}) &=& f_{\mathrm{f}} & \text{in } \Omega \setminus B(t), \\ \operatorname{div}(u) &=& 0 & \text{in } \Omega \setminus B(t), \\ u &=& U^r & \text{on } \partial B(t). \end{array}\right.$$

Here μ denotes the fluid viscosity and $f_{\mathrm{f}} := -\rho_{\mathrm{f}}\, g\, \mathbf{e}_y$ denotes the gravity force exerted on the fluid (with density ρ_{f}). $\mathbb{D}(u) := \frac{1}{2}(\nabla u + (\nabla u)^{\mathrm{t}})$ is the strain tensor. The boundary condition at the fluid-particle interface $\partial B(t)$ is a no-slip condition: the particle has a rigid movement which is decomposed into a translational movement (with translational velocity U) and a rotational movement (with rotational velocity ω),

$$U^r(t,x) = U(t) + \omega(t)(x - x_{B(t)})^\perp, \quad t > 0, \quad x \in \partial B(t).$$

The translational and rotational velocities are *a priori* unknown. We use the notation $x^\perp = (-x_2, x_1)$. The fluid-structure interaction emerges from the coupling with the Newton equation which expresses the instantaneous equilibrium of the forces applied to the particle:

$$(\mathrm{FSI}_2)\left\{\begin{array}{rcl} \displaystyle\int_B f_B - \int_{\partial B} \sigma.n &=& 0, \\ \displaystyle\int_B (x - x_B)^\perp \cdot f_B - \int_{\partial B} (x - x_B)^\perp \cdot (\sigma.n) &=& 0. \end{array}\right.$$

where $\sigma := 2\mu\mathbb{D}(u) - p\mathbb{I}$ is the total stress tensor for a Newtonian fluid. Here f_B denotes the external *non-hydrodynamical* forces exerted on the particle. In our case we restrict our study to the gravity forces: $f_B := -\rho_B\, g\,(0,1)$, where ρ_B is the density of the particle. Let us recall that buoyancy denotes the power to float or rise in a fluid; therefore a particle is said *buoyant* if $\rho_B \neq \rho_{\mathrm{f}}$ and *neutrally buoyant* if $\rho_B = \rho_{\mathrm{f}}$.

1. Compute the instantaneous velocity field generated by the inclusion of a rigid sphere in the fluid.
2. Compute the dynamics of a buoyant sphere in a fluid at rest.
3. Compute the dynamics of a neutrally buoyant ellipsoid in a linear shear flow.

□

Solution. We present a fictitious domain approach that allows us to address the fluid-structure interaction problem.

"Direct" formulation of the problem. The computational method that we propose is based upon a fictitious domain approach: the velocity field u and the pressure field p, defined on $\Omega \setminus \overline{B}$ are extended over Ω by

$$\begin{array}{rcll} u(t,x) &=& U(t) + \omega(t)(x - x_{B(t)})^\perp, & \text{on } B(t), \\ p(t,x) &=& 0, & \text{on } B(t). \end{array}$$

The extension of p follows from the fact that the pressure field is the Lagrange multiplier associated to the incompressibility condition. Note that if u describes a rigid movement, then it is divergence free, hence the extention by 0 for the associated pressure field is natural.

We introduce the functional spaces:

$$\begin{array}{rcl} X_B &=& \{u \in (H^1_0(\Omega))^2,\ \exists (U,\omega) \in \mathbb{R}^2 \times \mathbb{R}, \\ && \quad u(x) = U + \omega(x - x_B)^\perp \text{ a.e. in } B\}, \\ M_B &=& \{p \in L^2_0(\Omega),\ p = 0 \text{ a.e. in } B\}. \end{array}$$

We introduce the source term

$$f := f_{\mathrm{f}}\mathbf{1}_{\Omega\setminus\overline{B}} + f_B\mathbf{1}_B.$$

By the forthcoming Proposition 7.9, the fluid flow is determined as the solution of the variational problem:

$$(\mathrm{F})\left\{\begin{array}{l} \text{Find } (u,p) \in K_B \times M_B \text{ such that} \\ \begin{array}{rcl} 2\mu \displaystyle\int_\Omega \mathbb{D}(u) : \mathbb{D}(v) - \int_\Omega p \operatorname{div}(v) &=& \displaystyle\int_\Omega f \cdot v, \\ \displaystyle\int_\Omega q \operatorname{div}(u) &=& 0, \end{array} \\ \text{for all } (v,q) \in K_B \times M_B. \end{array}\right.$$

Proposition 7.9 *Let $(u,p) \in (H^1_0(\Omega))^2 \times L^2_0(\Omega)$. Assume that the restriction to $\Omega \setminus \overline{B}$ of (u,p) belongs to $(H^2(\Omega \setminus \overline{B}))^2 \times H^1(\Omega \setminus \overline{B})$. Then (u,p) is a solution of Equations $(\mathrm{FSI}_1) - (\mathrm{FSI}_2)$ if, and only if, (u,p) is a solution of problem (F).*

The dynamics of the rigid particle is described by:

$$(\mathrm{P})\begin{cases} \dfrac{d}{dt}x_{B(t)} = U(t) := \dfrac{\displaystyle\int_{B(t)} \rho_B u(t,x)\,\mathrm{d}x}{\displaystyle\int_{B(t)} \rho_B\,\mathrm{d}x}, \\[2ex] \dfrac{d}{dt}\theta_{B(t)} = \omega(t) := \dfrac{\displaystyle\int_{B(t)} \rho_B\, u(t,x)\cdot(x-x_{B(t)})^{\perp}\,\mathrm{d}x}{\displaystyle\int_{B(t)} \rho_B\, \|x-x_{B(t)}\|^2\,\mathrm{d}x}. \end{cases}$$

The computational method relies on the computation of the solution, at a given time, of the variational problem (F) which determines the flow generated by the inclusion of the particle and, then update the position of the particle by solving (P) (with an explicit Euler scheme, for instance). Thus let us focus on the computation of the solution of problem (F).

Remark 7.10 *For the sake of simplicity (and without any consequences on the velocity field) we may replace the source term modelling the gravity*

$$f := f_{\mathrm{f}}\mathbf{1}_{\Omega\setminus\overline{B}} + f_B\mathbf{1}_B$$

by a source term modelling the buoyancy only:

$$\tilde{f} := f - f_{\mathrm{f}} = (f_B - f_{\mathrm{f}})\,\mathbf{1}_B.$$

Penalized formulation of problem (F)**.** Solving problem (F) with a finite element solver is not easy: the elements should belong to the *constrained* functional spaces. Moreover, as the rigid domain may evolve in time, so do the constraints. As a consequence, a finite element basis should be built at each time step, which is prohibitive! In order to avoid these difficulties, we may use an approximation method which consists in relaxing the constraints in the functional spaces, thus leading to the possibility of using standard finite element solvers. In the variational formulation, the relaxation of the constraints should be associated with the introduction of integrals which tend to mimick / impose the constraint on the solution: this additional term is called a penalty term.

Let us introduce another characterization of the rigid movement:

Proposition 7.11 *We have*

$$X_B = \{u \in (H^1_0(\Omega))^2,\ \mathbb{D}(u) = 0 \ \text{ a.e. in } B\}.$$

Roughly speaking, the above proposition states that rigid movements do not deform the domain (as $\mathbb{D}(u)$ is the deformation tensor).

The penalty method applied to the variational formulation leads to the following problem:

$$(\mathrm{F}_\varepsilon)\begin{cases} \text{Find } (u_\varepsilon, p_\varepsilon) \in (H^1_0(\Omega))^2 \times L^2_0(\Omega) \text{ such that} \\ \quad 2\mu \displaystyle\int_\Omega \mathbb{D}(u_\varepsilon) : \mathbb{D}(v) + \frac{2}{\varepsilon}\int_B \mathbb{D}(u_\varepsilon) : \mathbb{D}(v) \\ \qquad\qquad - \displaystyle\int_\Omega p_\varepsilon \operatorname{div}(v) \;=\; \int_\Omega f\cdot v, \\ \qquad\qquad \displaystyle\int_\Omega q \operatorname{div}(u_\varepsilon) \;=\; 0, \\ \text{for all } (v,q) \in (H^1_0(\Omega))^2 \times L^2_0(\Omega). \end{cases}$$

Roughly speaking, the penalty term $\frac{2}{\varepsilon}\int_B \mathbb{D}(u_\varepsilon) : \mathbb{D}(v)$ ensures that when ε tends to 0, then $\mathbb{D}(u_\varepsilon)$ tends to 0 on B, thus satisfying the constraint in the asymptotic regime. It means also that rigid domains are modelled as highly viscous domains (with viscosity $1/\varepsilon$).

It is possible to prove the following result:

Proposition 7.12 *Let* (u,p) *be the solution of* (F) *and let* $(u_\varepsilon, p_\varepsilon)$ *be the solution of* (F_ε)*. Then*

$$\|u - u_\varepsilon\|_{H^1(\Omega)} = \mathscr{O}(\varepsilon).$$

From the computational point of view, it is possible to use a standard finite element solver to compute the solution of (F_ε). This approach does not require mesh adaptation techniques: a fixed (structured or unstructured) mesh can be used.

Algorithm for the dynamics of a rigid particle in a fluid. The dynamics of a particle which evolves in a fluid has been modelled by a strongly coupled fluid-structure interaction problem: we aim at solving problems (F) and (P). Let us describe how to handle this coupled problem with the computational aspects. Problem (P) is solved using an explicit Euler scheme but requires the knowledge of the instantaneous velocity field generated by the inclusion of the rigid particle, hence the solution of problem (F). As it was outlined, problem (F) can be solved with the penalty formulation or the saddle-point formulation. We present the algorithm with the penalty method (the adaptation for the saddle-point formulation is straightforward).

Assume that the position of a particle is known at time t_n. We aim at computing the velocity field in the fluid at time t_n and update the position of the particle at time $t_{n+1} = t_n + \Delta t$. The computational process writes:

Step 0 (initialization). The position of the center of mass $x^n_B = x_B(t_n)$ and angle $\theta^n_B = \theta_B(t_n)$ are known. The rigid domain

$$B^n := B(t_n)$$

is completely characterized by x^n_B, θ^n_B and the geometrical properties of the particle.

Step 1. Define the generalized viscosity and the source term:

$$\mu^n := \mu\mathbf{1}_{\Omega\setminus B^n} + \frac{1}{\varepsilon}\mathbf{1}_{B^n}, \qquad f^n := (f_{B^n} - f_{\mathrm{f}})\,\mathbf{1}_{B^n}.$$

Step 2. Solve (the penalized version of) problem (F):

$$\left\{\begin{array}{l} \text{Find } (u^n,p^n)\in (H^1_0(\Omega))^2\times L^2_0(\Omega) \text{ such that} \\ \displaystyle\int_\Omega 2\mu^n\,\mathbb{D}(u^n):\mathbb{D}(v)-\int_\Omega p^n\operatorname{div}(v) \;=\; \int_\Omega f^n\cdot v, \\ \displaystyle\qquad\qquad\qquad\qquad\int_\Omega q\operatorname{div}(u^n) \;=\; 0, \\ \text{for all } (v,q)\in (H^1_0(\Omega))^2\times L^2_0(\Omega). \end{array}\right.$$

Note that the definition of μ^n leads to write the penalized problem as

$$\int_\Omega 2\mu^n\,\mathbb{D}(u):\mathbb{D}(v) = 2\mu\int_\Omega \mathbb{D}(u):\mathbb{D}(v)+\frac{2}{\varepsilon}\int_{B^n}\mathbb{D}(u):\mathbb{D}(v).$$

Note also that the above problem can be used with a standard finite element solver.

Step 3. Solve problem (P) with an explicit Euler scheme:

- Compute the translational and rotational velocities of the particle at time t_n:

$$U^n := \frac{\displaystyle\int_{B^n}\rho_B u^n(x)\,\mathrm{d}x}{\displaystyle\int_{B^n}\rho_B\,\mathrm{d}x},$$

$$\omega^n := \frac{\displaystyle\int_{B^n}\rho_B u^n(x)\cdot(x-x_{B^n})^\perp\,\mathrm{d}x}{\displaystyle\int_{B^n}\rho_B\|x-x_{B^n}\|^2\,\mathrm{d}x}.$$

- Update the position of the particle:

$$x_B^{n+1}=x_B^n+\Delta t\,U^n, \qquad \theta_B^{n+1}=\theta_B^n+\Delta t\,\omega^n.$$

1. *Inclusion of a rigid sphere in a fluid.* We consider the domain $\Omega=]0,1[^2$ which divides into two (moving) subdomains: a part is occupied by an incompressible Newtonian fluid (viscosity $\mu=1$) and the other part is a rigid sphere (radius 0.1) at $(x^0,y^0)=(0.5,0.8)$. The sphere is not neutrally buoyant ($\rho_B=1.0$, $\rho_\mathrm{f}=0.1$). The instantaneous velocity field generated by the inclusion of the sphere may be computed by using the penalized formulation of the problem. In the penalized formulation, a generalized Stokes problem is solved at each time step. This requires the use of finite elements that are inf-sup stable. In that prospect we use $\mathbb{P}^1_\mathrm{b}-\mathbb{P}^1$ finite elements. Note that it is no worth using $\mathbb{P}^2-\mathbb{P}^1$ elements in order to increase the accuracy of the solution: the extension of the velocity field from $\Omega\setminus B$ to Ω is not in $H^3(\Omega)$. Actually the velocity field is not even in $H^2(\Omega)$ (at the fluid-particle interface, the velocity field is continuous but its gradient admits a jump discontinuity through the interface). As a consequence, higher order finite elements would increase the computational costs *without additional accuracy.* Nevertheless the analysis guarantees that, by using the mini element, the numerical solution converges to the exact solution with a suboptimal order of convergence (in order to guarantee a convergence rate of order 1, the solution should be in H^2, see Theorem 6.6). Figure 27 presents the velocity field generated by a rigid sphere of radius $r=0.1$ and coordinates $(0.5,0.6)$. The sphere is not neutrally buoyant ($\rho_B>\rho_\mathrm{f}$).

2. *Sedimentation of a rigid sphere in a fluid at rest*, see Section 8, Program 8.2. We consider the domain $\Omega=]0,1[^2$ made of an incompressible Newtonian fluid and a rigid sphere of radius 0.1 with a center of mass located at $(0.5,0.8)$. The sphere is not neutrally buoyant ($\rho_B>\rho_\mathrm{f}$) so that it is expected that the sphere falls down to the ground. By a symmetry argument, if the initial position of the sphere is symmetric with respect to the axis $x=0.5$ then the sphere falls along this axis with a zero angular velocity. As a consequence we only compute the translational velocity in order to update the position of the sphere. We compute the solution of the penalized version of the fluid-structure interaction problem with $\mathbb{P}^1_\mathrm{b}-\mathbb{P}^1$ finite elements on a 100×100 structured mesh. The fall of a rigid sphere in a fluid at rest is reproduced in Figure 28.

3. *Rotating ellipsoid in a linear shear flow*, see Section 8, Section 8.8, Program 8.3. We consider the domain $\Omega=]0,1[^2$. A linear shear flow may be described by the velocity profile

$$u(x,y)=(1-2y)\begin{pmatrix}1\\0\end{pmatrix},$$

which is obtained by solving the Stokes equations in Ω with the boundary conditions

$$\begin{array}{ll} u(\cdot,0)=+1, & \text{on } y=0,\\ u(\cdot,1)=-1, & \text{on } y=1, \end{array}$$

and periodic conditions with respect to x. Consider an ellipsoid (with half semi-axes 0.1 and 0.25) which is neutrally buoyant (i.e. $\rho_B=\rho_\mathrm{f}$). When inserting the ellipsoid in the shear flow, the velocity profile in the fluid is modified and, as a result, the backflow produces a movement of the ellipsoid. We expect that, if the center of mass is initially located at $(0.5,0.5)$, it does not evolve in time (no gravity effect), i.e. the ellipsoid does not translate. But

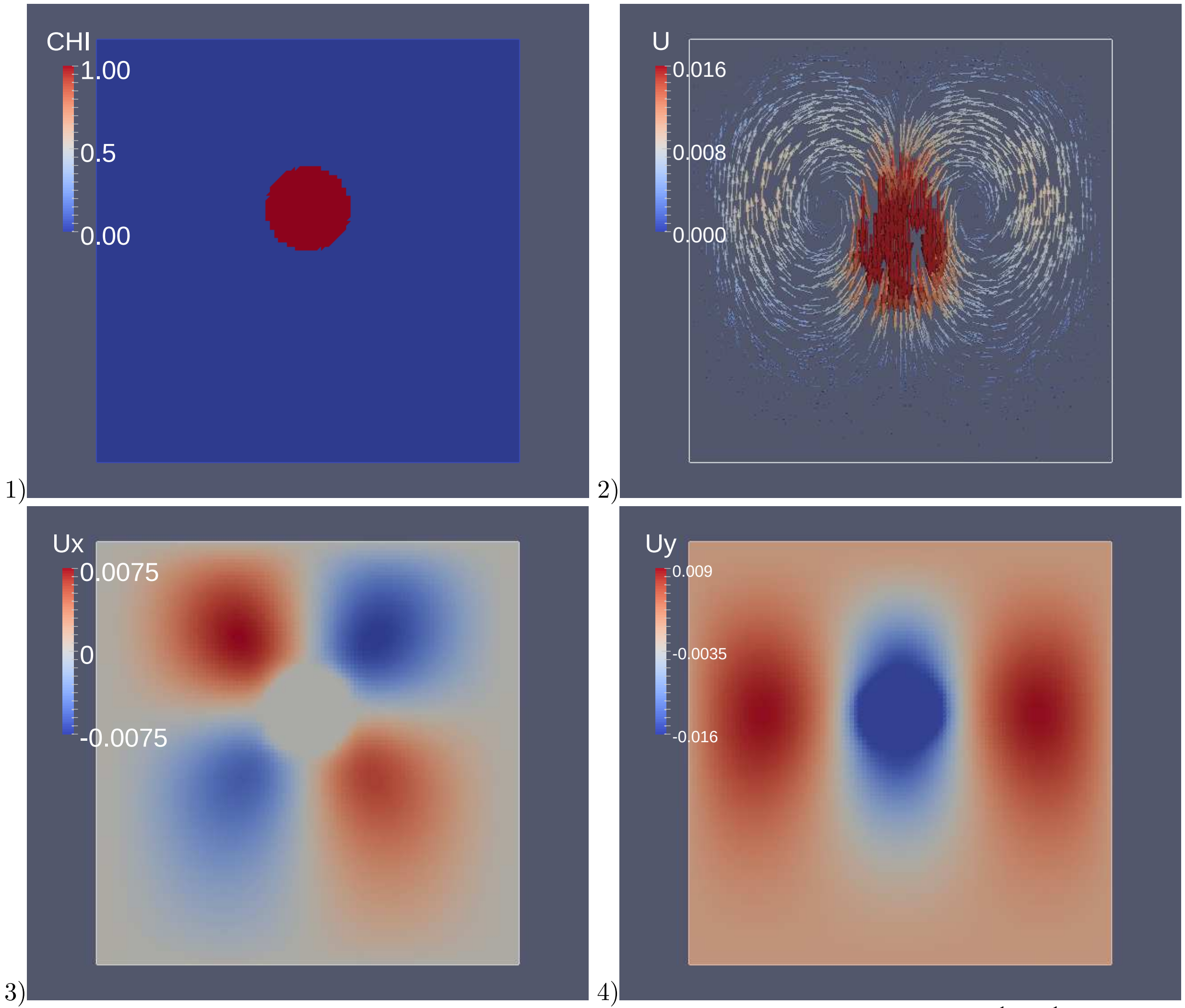


Figure 27. Solution of the fluid-structure interaction problem with the penalty method with $\mathbb{P}^1_b - \mathbb{P}^1$ finite elements on a 100×100 structured mesh: 1) characteristic function of $\boldsymbol{B}$. 2) velocity field. 3) first component of the velocity field. 4) second component of the velocity field.

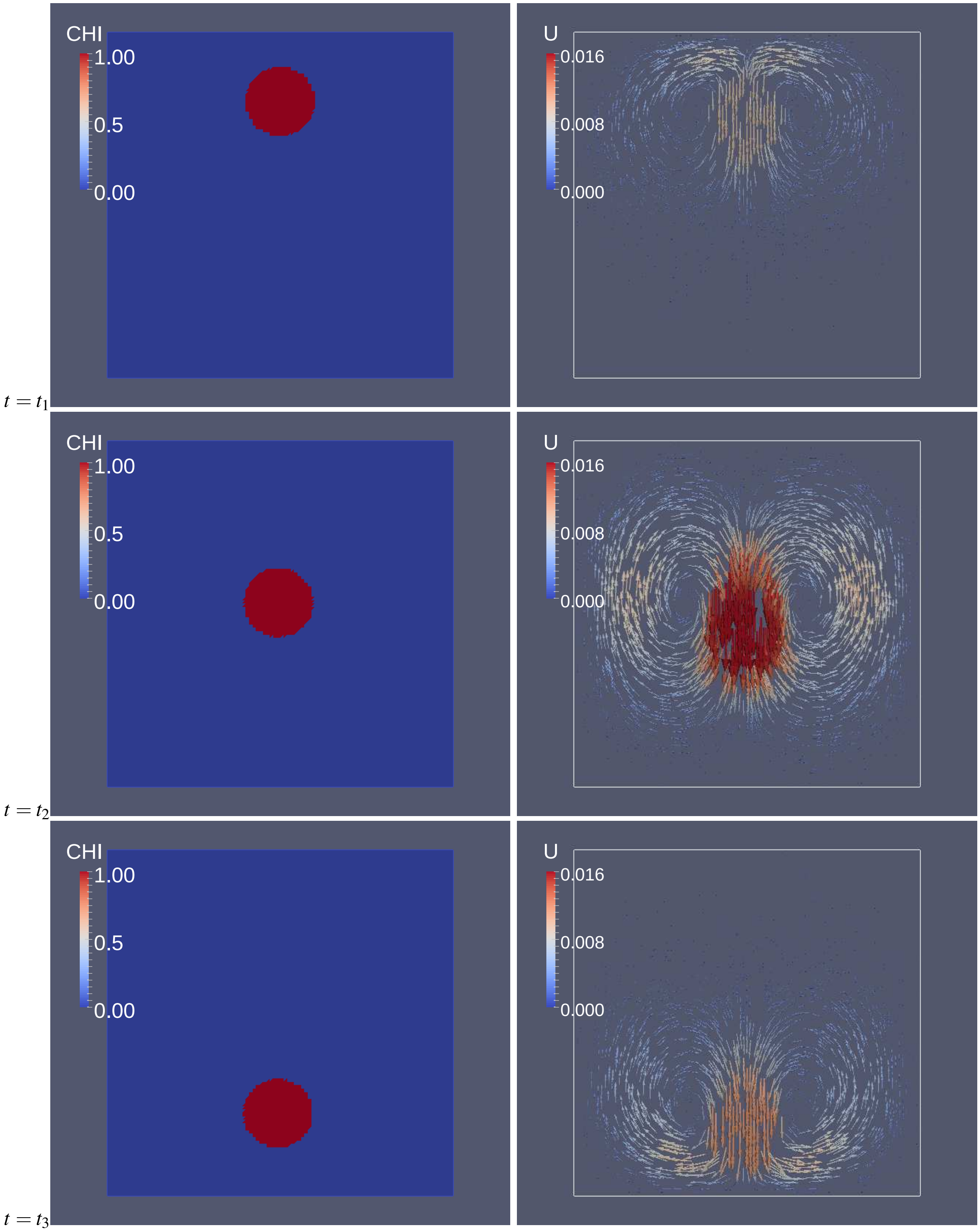


Figure 28. Position of the sphere and velocity field at different times t_i.

the angular position of the ellipsoid does evolve in time: the ellipsoid rotates around its center of mass. We compute the solution of the penalized version of the fluid-structure interaction problem with $\mathbb{P}^1_b - \mathbb{P}^1$ finite elements on a 100×100 structured mesh. Figure 29 represents the ellipsoid and the velocity field at different time steps.

□

8. FreeFem++ programs

8.1 Problem 1

```
/********************************************/
/* PROBLEM 1                                */
/********************************************/

/* == Fine mesh *****************************/
mesh Th0=square(300,300);

/* == Solution of the Poisson problem *******/
func u=x*y*(1.-x)*(1.-y);

/* == Source term ***************************/
func f=2*y*(1.-y)+2*x*(1.-x);

/* == Boundary conditions *******************/
func g=0;

/* == Number of computed errors *************/
int Niter=20;

/* == Storage of the error in L2 or H1 ******/
real[int] eL2(Niter);
real[int] eH1(Niter);
real[int] hL2(Niter);
real[int] hH1(Niter);

/* == Finite element space on the fine mesh */
fespace Vh0(Th0,P2);

/* == Projection on the fine mesh ***********/
Vh0 u0=u;

/* == Initialization of the number of nodes */
int np=10;

/* == Initialization of the mesh ************/
mesh Th=square(np,np);

/* == Finite element on the coarse mesh *****/
fespace Vh(Th,P2);
Vh uh,vh,errh;

/* == Variational problem *******************/
problem Poisson(uh,vh,solver=LU) =
    int2d(Th)(dx(uh)*dx(vh)+dy(uh)*dy(vh))
   -int2d(Th)(f*vh)
   +on(1,2,3,4,uh=g) ;

/* == Compute uh and uh-u for different h ***/
for (int i=0;i<Niter;i++){
   Th=square(np,np);
   Poisson;
   errh=u-uh;
   Vh0 uh0=uh;
   Vh0 errh0=u0-uh0;

   real errL2=sqrt(int2d(Th0)(errh0^2));
   real errH1=sqrt(int2d(Th0)(dx(errh0)^2
                             +dy(errh0)^2));

   hL2(i)=1./(np-1);
   eL2(i)=errL2;
   hH1(i)=1./(np-1);
   eH1(i)=errH1;

   plot(uh,wait=0);
   np=np+2;
}

ofstream out1("P2_errL2.txt");
for (int i=0;i<Niter;i++){
   out1<<log(hL2(i))<<" "<<log(eL2(i))<<endl;
}

ofstream out2("P2_errH1.txt");
for (int i=0;i<Niter;i++){
   out2<<log(hH1(i))<<" "<<log(eH1(i))<<endl;
}
/********************************************/
```

8.2 Problem 2

From the Robin condition to the Dirichlet condition

```
/********************************************/
/* PROBLEM 2-1: FROM THE ROBIN CONDITION    */
/*              TO THE DIRICHLET CONDITION  */
/********************************************/

/* == Mesh **********************************/
int np=200;
mesh Th=square(np,np);
plot(Th,wait=0);

/* == P1 finite element space ***************/
fespace Vh(Th,P1);
Vh uh,vh,ueh,e;

func f=1.;
```

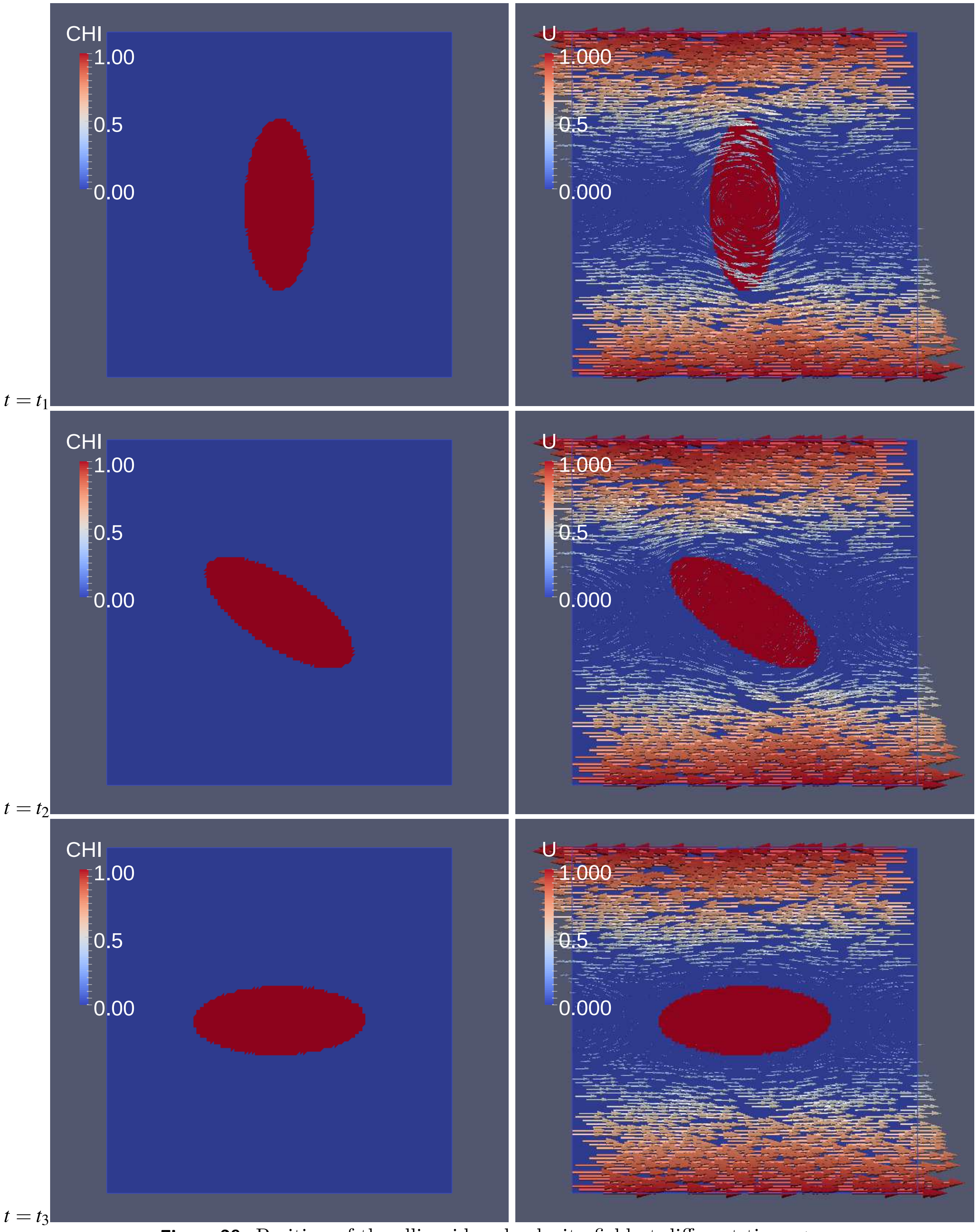


Figure 29. Position of the ellipsoid and velocity field at different times t_i.

```
func g=0.;

int Niter=20;
real[int] errH1(Niter);
real[int] ee(Niter);

real eps=0.1;
real alpha=1;

/* == Variational problem *******************/
/* == with Dirichlet conditions *************/
problem laplDir(uh,vh,solver=LU)=
    int2d(Th)(dx(uh)*dx(vh)+dy(uh)*dy(vh))
   +int2d(Th)(alpha*uh*vh)
   -int2d(Th)(f*vh)
   +on(1,2,3,4,uh=g) ;

laplDir;

plot(uh,wait=0);

/* == Variational problem *******************/
/* == with Robin condition ******************/
problem laplRobin(ueh,vh,solver=LU)=
    int2d(Th)(dx(ueh)*dx(vh)+dy(ueh)*dy(vh))
   +int2d(Th)(alpha*ueh*vh)
   +int1d(Th,1,2,3,4)((1./eps)*ueh*vh)
   -int2d(Th)(f*vh);

/* == Compute the error ue-u0 ***************/
for (int i=0;i<Niter;i++){
   laplRobin;
   e=uh-ueh;
   real eH1=sqrt(int2d(Th)(e^2)
                +int2d(Th)(dx(e)^2+dy(e)^2));
   errH1(i)=eH1;
   ee(i)=eps;
   plot(ueh,wait=0);
   eps=eps/2.;
}

ofstream out1("q2_errorH1.txt");
for (int i=0;i<Niter;i++){
   out1<<log(ee(i))<<" "<<log(errH1(i))<<endl;
}
/********************************************/
```

From the Robin condition to the Neuman condition

```
/********************************************/
/* PROBLEM 2-2: FROM THE ROBIN CONDITION    */
/*              TO THE NEUMANN CONDITION    */
/********************************************/

/* == Mesh **********************************/
int np=200;
mesh Th=square(np,np);
plot(Th,wait=0);

/* == P1 finite element space ***************/
fespace Vh(Th,P1);
Vh uh,vh,fh,ueh,e;

func f=1;
real alpha=1;

int Niter=10;
real[int] errH1(Niter);
real[int] ee(Niter);

real eps=100;

/* == Variational problem *******************/
/* == with Neumann conditions ***************/
problem laplNeum(uh,vh,solver=LU) =
    int2d(Th)(dx(uh)*dx(vh)+dy(uh)*dy(vh))
   +int2d(Th)(alpha*uh*vh)
   -int2d(Th)(f*vh);

laplNeum;

plot(uh,wait=0,value=true);

/* == Variational problem *******************/
/* == with Robin condition ******************/
problem laplRobin(ueh,vh,solver=LU) =
    int2d(Th)(dx(ueh)*dx(vh)+dy(ueh)*dy(vh))
   +int2d(Th)(alpha*ueh*vh)
   +int1d(Th,1,2,3,4)((1./eps)*ueh*vh)
   -int2d(Th)(f*vh) ;

for (int i=0;i<Niter;i++){
   laplRobin;
   e=uh-ueh;
   real eL2=sqrt(int2d(Th)(e^2));
   real eH1=sqrt(int2d(Th)(e^2)
                +int2d(Th)(dx(e)^2+dy(e)^2));
   errH1(i)=eH1;
   ee(i)=eps;
   plot(ueh,wait=0,value=true);
   cout<<i<<endl;
   eps=eps*10.;
}


ofstream out1("q3_errorH1.txt");
for (int i=0;i<Niter;i++){
   out1<<log(ee(i))<<" "<<log(errH1(i))<<endl;
}

/********************************************/
```

8.3 Problem 3

```
/********************************************/
/* PROBLEM 3                                */
/********************************************/

/* == Define the boundaries of the domain ***/
border A(t=0.,1) {x=t  ;y=0;  label=1;};
border B(t=0.,1) {x=1  ;y=t;  label=2;};
border C(t=0.,1) {x=1-t;y=1;  label=3;};
border D(t=0.,1) {x=0  ;y=1-t;label=4;};

/* == Number of elements on each boundary ***/
int n=500;

/* == Mesh generation ***********************/
mesh Th = buildmesh(A(n)+B(n)+C(n)+D(n));
plot(Th,wait=1);

/* == Dirichlet boundary function ***********/
/* == - g0 for problem (P^0) ****************/
/* == - g1 for problem (P^1) ****************/
func g0=0;
func g1=1;

/* == Finite element space (here: P1) *******/
fespace Vh(Th,P1);
Vh u,v,w;

/* == Define the source term ****************/
func f=1;

/* == Define the variational formulation ****/
/* == - select g0 on boundary 4 for (P^0) ***/
/* == - select g1 on boundary 4 for (P^1) ***/
problem Poisson(u,v,solver=LU)=
int2d(Th)(dx(u)*dx(v)+dy(u)*dy(v))
   -int2d(Th)(f*v)
   +on(1,2,3,u=0)
   +on(4,u=g0);

/* == Solve the problem *********************/
Poisson;

/* == Compute the H1-norm of the solution ***/
real normH1=sqrt(int2d(Th)(dx(u)^2+dy(u)^2));
cout<< normH1 <<endl;
plot(u,fill=0,wait=0,value=1);
w=sqrt(dx(u)^2+dy(u)^2);
plot(w,fill=0,wait=0,value=1);

/********************************************/
```

8.4 Problem 4

```
/********************************************/
/* PROBLEM 4                                */
/********************************************/

/* == Mesh **********************************/
int np=200;
mesh Th=square(np,np);

/* == P1 finite element space ***************/
fespace Vh(Th,P1);
Vh uh,vh;

/* == Source term f *************************/
func f=1.;
//Vh fh=f;

/* == Source term g on each boundary ********/
func g1= x*(1.-x);
func g2= 0.;
func g3=-x*(1.-x);
func g4= 0.;

/* == Penalization parameter ****************/
real eps=1.0E-8;

/* == Variational problem *******************/
problem laplace(uh,vh,solver=LU)=
    int2d(Th)(dx(uh)*dx(vh)+dy(uh)*dy(vh))
   +int2d(Th)(eps*uh*vh)
   -int2d(Th)(f*vh)
   -int1d(Th,1)(g1*vh)
   -int1d(Th,2)(g2*vh)
   -int1d(Th,3)(g3*vh)
   -int1d(Th,4)(g4*vh);

real CC=int2d(Th)(f)
        +int1d(Th,1)(g1)+int1d(Th,2)(g2)
        +int1d(Th,3)(g3)+int1d(Th,4)(g4);
cout<<"Value (should be 0): "<<CC<<endl;

laplace;

plot(uh,wait=0,value=true);
/********************************************/
```

8.5 Problem 5

```
/********************************************/
/* PROBLEM 5                                */
/********************************************/

/* == Fine mesh *****************************/
mesh Th0=square(300,300);

/* == Data of the problem *******************/
real a =0.5;
real x0=0.5;
real y0=0.5;

/* == Solution of the Poisson problem *******/
```

```
func u=((x-x0)^2+(y-y0)^2)^(a/2.);
/* == Source term ***************************/
func f=a^2*((x-x0)^2+(y-y0)^2)^(a/2.-1.);
/* == Boundary conditions ******************/
func g=u;

/* == Number of computed errors ************/
int Niter=20;
/* == Storage of the error in L2 or H1 ******/
real[int] eL2(Niter);
real[int] eH1(Niter);
real[int] hL2(Niter);
real[int] hH1(Niter);

/* == Finite element space on the fine mesh */
fespace Vh0(Th0,P2);

/* == Projection on the fine mesh ***********/
Vh0 u0=u;
plot(u0,value=1,fill=0,wait=0);

/* == Initialization of the number of nodes */
int np=10;
/* == Initialization of the mesh ************/
mesh Th=square(np,np);

/* == Finite element on the coarse mesh *****/
fespace Vh(Th,P2);
Vh uh,vh,errh;

/* == Variational problem ******************/
problem Poisson(uh,vh) =
    int2d(Th)(dx(uh)*dx(vh)+dy(uh)*dy(vh))
   -int2d(Th)(f*vh)
   +on(1,2,3,4,uh=g) ;

Poisson;

/* == Compute uh and uh-u for different h ***/
for (int i=0;i<Niter;i++){
   Th=square(np,np);
   Poisson;
   errh=u-uh;
   Vh0 uh0=uh;
   Vh0 errh0=u0-uh0;
   real errL2=sqrt(int2d(Th0)(errh0^2));
   real errH1=sqrt(int2d(Th0)(dx(errh0)^2
                          +dy(errh0)^2));
   hL2(i)=1./(np-1);
   eL2(i)=errL2;
   hH1(i)=1./(np-1);
   eH1(i)=errH1;

   plot(uh,value=1,fill=1,wait=0);
   np=np+2;
}

ofstream out1("errL2.txt");
for (int i=0;i<Niter;i++){
   out1<<log(hL2(i))<<" "<<log(eL2(i))<<endl;
}

ofstream out2("errH1.txt");
for (int i=0;i<Niter;i++){
   out2<<log(hH1(i))<<" "<<log(eH1(i))<<endl;
}
/*******************************************/
```

8.6 Problem 6

```
/*******************************************/
/* PROBLEM 6                               */
/*******************************************/

int n=50;

mesh Th=square(n,n);

//fespace Xh(Th,[P1,P1,P0]);
//fespace Xh(Th,[P1,P1,P1]);
//fespace Xh(Th,[P1b,P1b,P1]);
fespace Xh(Th,[P2,P2,P1]);

/* === Velocity and pressure field **********/
Xh [u1,u2,p],[v1,v2,q];

/* === Source term *************************/
real r=0.25;
V0h f1h=50*((x-0.5)^2+(y-0.5)^2<r^2);
V0h f2h=50*((x-0.5)^2+(y-0.5)^2<r^2);

/* == Stabilization parameter ***************/
real eps=1E-6;

/*******************************************/
/* == Stokes problem ***********************/
/********************************************/

problem Stokes([u1,u2,p],[v1,v2,q])=
int2d(Th)(2*dx(u1)*dx(v1)+dy(u1)*dy(v1)
          +dx(u2)*dx(v2)+2*dy(u2)*dy(v2)
          +dy(u1)*dx(v2)+dx(u2)*dy(v1))
   +int2d(Th)(eps*p*q-p*dx(v1)-p*dy(v2)
                  -q*dx(u1)-q*dy(u2))
   -int2d(Th)(f1h*v1+f2h*v2)
   +on(1,2,3,4,u1=0,u2=0);

Stokes;

plot([u1,u2],value=1);
plot(p,fill=1,value=1);
```

```
/********************************************/
```

8.7 Problem 7

```
/********************************************/
/* PROBLEM 7                                */
/********************************************/

/* == Mesh **********************************/
int np=100;
mesh Th=square(np,np);

/* == P1 finite element space ***************/
fespace Vh(Th,P1);
Vh uh,vh;

/* == P0 finite element space ***************/
fespace Ph(Th,P0);

/* == Characteristic function ***************/
Ph chih=((x-0.5)^2+(y-0.5)^2<0.2^2);

/* == Source term f *************************/
Ph fh=((x-0.5)^2+(y-0.5)^2<0.2^2)
     -((x-0.5)^2+(y-0.5)^2>0.2^2);

/* == Lagrange multiplier *******************/
/* == lm0: Lagrange multiplier at time n ****/
/* == lm : Lagrange multiplier at time n+1 **/
real lm,lm0=0.;

/* == Uzawa parameters **********************/
real rhomax=2./(int2d(Th)(chih))^2;
real rho=0.5*rhomax;
real tol=1.E-8;
int  Nmax=100;

/* == Variational problem *******************/
problem laplace(uh,vh,solver=LU)=
    int2d(Th)(dx(uh)*dx(vh)+dy(uh)*dy(vh))
   +int2d(Th)(uh*vh)
   -int2d(Th)(fh*vh)
   +int2d(Th)(lm0*chih*vh);

/* == Uzawa algorithm ***********************/
real err=2*tol;
int  i=1;

while ((i<=Nmax)&&(err>tol)){

   laplace;
   lm=lm0+rho*int2d(Th)(chih*uh);

   err=sqrt((lm-lm0)^2);

   lm0=lm;

   i=i+1;

   cout<<"i="<<i<<" ; error="<<err<<endl;
}

plot(uh,wait=0,value=true);
cout<< "*** Constraint:"<<endl;
cout<< "  Bu="<< int2d(Th)(chih*uh)<<endl;
cout<< "  lambda="<<lm<<endl;
/********************************************/
```

8.8 Problem 8

Rigid sphere in a fluid: computing the velocity field with a penalized formulation

```
/********************************************/
/* PROBLEM 8-1                              */
/* FSI PROBLEM A PENALIZED FORMULATION      */
/********************************************/

mesh Th =square(100,100);

fespace Vh(Th,P1b);
fespace Xh(Th,P1);
Vh u1,u2,v1,v2;
Xh p,q;

/* == Coordinates / radius of the sphere ****/
real xB=0.5, yB=0.6, rB=0.1;
/* == Gravity and buoyancy ******************/
real g=9.81, rhoB=1.0, rhoF=0.1;

/* == Source term, characteristic function **/
fespace Ph(Th,P0);
Ph chiB = ((x-xB)^2+(y-yB)^2 < rB^2);
Ph f1 =  0.00;
Ph f2 = -(rhoB-rhoF)*g*chiB;

real eps=1.0E-5, delta=1.0E-8;

/* == Penalized formulation *****************/
problem FSI([u1,u2,p],[v1,v2,q])=
int2d(Th)(2*dx(u1)*dx(v1)+dy(u1)*dy(v1)
           +dx(u2)*dx(v2)+2*dy(u2)*dy(v2)
           +dy(u1)*dx(v2)+dx(u2)*dy(v1))
   +int2d(Th)(delta*p*q-p*dx(v1)-p*dy(v2)
                       +q*dx(u1)+q*dy(u2))
   -int2d(Th)(f1*v1+f2*v2)
   +int2d(Th)((2*dx(u1)*dx(v1)+dy(u1)*dy(v1)
      +dx(u2)*dx(v2)+2*dy(u2)*dy(v2)
      +dy(u1)*dx(v2)+dx(u2)*dy(v1))*chiB/eps)
   +on(1,2,3,4,u1=0,u2=0);

FSI;

/* == Computing the translational velocity **/
real Mh=int2d(Th)(rhoB*chiB);
real uB=int2d(Th)(rhoB*chiB*u1)/Mh;
real vB=int2d(Th)(rhoB*chiB*u2)/Mh;

/* == Computing the angular velocity ********/
real Jh,oB;
Jh=int2d(Th)(rhoB*chiB*((x-xB)^2+(y-yB)^2));
oB=int2d(Th)(rhoB*chiB*(-(y-yB)*u1+(x-xB)*u2))/Jh;

cout<<"(u,v)=("<<uB<< ","<<vB<<")"<<endl;
cout<<"omega="<<oB<<endl;
plot(chiB,fill=true,wait=0);
plot([u1,u2],value=true,wait=0,coef=5);
```

Sedimentation of a rigid sphere in a fluid at rest

```
/********************************************/
/* PROBLEM 8-2                              */
/* SEDIMENTATION OF A RIGID SPHERE          */
/********************************************/

mesh Th =square(100,100);

fespace Xh(Th,P1b);
fespace Mh(Th,P1);
Xh u1,u2,v1,v2;
Mh p,q;

/* == Coordinates / radius of the sphere ****/
real xB=0.5, yB=0.8, rB=0.1;
/* == Gravity and buoyancy ******************/
real g=9.81, rhoB=1.0, rhoF=0.1;

/* == Source term, characteristic function **/
fespace Ph(Th,P0);
Ph chiB = ((x-xB)^2+(y-yB)^2 < rB^2);
Ph f1=0.00;
Ph f2=-(rhoB-rhoF)*g*chiB;

/* == Parameters ****************************/
real eps=1.0E-2, delta=1.0E-8;

/* == Variational formulation ***************/
problem FSI([u1,u2,p],[v1,v2,q])=
int2d(Th)(2*dx(u1)*dx(v1)+dy(u1)*dy(v1)
           +dx(u2)*dx(v2)+2*dy(u2)*dy(v2)
           +dy(u1)*dx(v2)+dx(u2)*dy(v1))
   +int2d(Th)(delta*p*q-p*dx(v1)-p*dy(v2)
                       +q*dx(u1)+q*dy(u2))
   -int2d(Th)(f1*v1+f2*v2)
   +int2d(Th)((2*dx(u1)*dx(v1)+dy(u1)*dy(v1)
      +dx(u2)*dx(v2)+2*dy(u2)*dy(v2)
      +dy(u1)*dx(v2)+dx(u2)*dy(v1))*chiB/eps)
   +on(1,2,3,4,u1=0,u2=0);

/* == Time evolution ************************/
int imax=50;
real t=0.0;
real vpart1=0.0;
real vpart2=0.0;
real dt=1.0;
for (int i=0 ; i<=imax; i++){
   FSI;
   plot([u1,u2],ps="u"+i+".eps",coef=1);
   plot(chiB,ps="chi"+i+".eps",fill=1);
   t=t+dt;
   vpart1=u1(xB,yB);
   vpart2=u2(xB,yB);
   xB=xB+vpart1*dt;
   yB=yB+vpart2*dt;
   chiB= ((x-xB)^2+(y-yB)^2 < rB^2);
   f2=-(rhoB-rhoF)*g*chiB;
   plot(chiB,fill=1);
}
```

Rigid ellipsoid in a linear shear flow

```
/********************************************/
/* PROBLEM 8-3                              */
/* ROTATING ELLIPSOID IN A SHEAR FLOW       */
/********************************************/

mesh Th =square(100,100);

fespace Vh(Th,P1b);
fespace Wh(Th,P1);
Vh u1,u2,v1,v2;
Wh p,q;

/* == Coordinates / radii of the ellipsoid **/
real xB=0.5,yB=0.5,thetaB=0,rB=0.1,dB=0.25;
/* == Gravity and buoyancy ******************/
real g=9.81, rhoB=1., rhoF=1.;

real Mh,Jh,uB,vB,omegaB;
real cosB=cos(thetaB);
real sinB=sin(thetaB);

/* == Source term, characteristic function **/
fespace Ph(Th,P0);
Ph chiB;

chiB=(((x-xB)*cosB+(y-yB)*sinB)^2/rB^2
    +(-(x-xB)*sinB+(y-yB)*cosB)^2/dB^2<=1.0);

Ph f1=0.00;
Ph f2=-(rhoB-rhoF)*g*chiB;

/* == Boundary condition (Couette flow) *****/
func g1=1.0*(1.-2*y);
func g2=0;

/* == Parameters ****************************/
real eps=1.0E-2, delta=1.0E-8;
```

```
/* == Variational formulation ***************/
problem FSI([u1,u2,p],[v1,v2,q],solver=Crout)=
int2d(Th)(2*dx(u1)*dx(v1)+dy(u1)*dy(v1)
         +dx(u2)*dx(v2)+2*dy(u2)*dy(v2)
         +dy(u1)*dx(v2)+dx(u2)*dy(v1))
   +int2d(Th)(delta*p*q-p*dx(v1)-p*dy(v2)
                       +q*dx(u1)+q*dy(u2))
   -int2d(Th)(f1*v1+f2*v2)
   +int2d(Th)((2*dx(u1)*dx(v1)+dy(u1)*dy(v1)
      +dx(u2)*dx(v2)+2*dy(u2)*dy(v2)
      +dy(u1)*dx(v2)+dx(u2)*dy(v1))*chiB/eps)
   +on(1,2,3,4,u1=g1,u2=g2);

/* == Time evolution ************************/
int imax=100;
real t = 0.0;
real dt= 0.1;
for (int i=0 ; i<=imax; i++){
   FSI;
   plot([u1,u2],ps="u"+i+".eps",wait=0,coef=1);
   plot(chiB,ps="chi"+i+".eps",wait=0,fill=1);
   t=t+dt;

   /* == Translational velocity *************/
   Mh=int2d(Th)(rhoB*chiB);
   uB=int2d(Th)(rhoB*chiB*u1)/Mh;
   vB=int2d(Th)(rhoB*chiB*u2)/Mh;

   /* == Angular velocity *******************/
   Jh=int2d(Th)(rhoB*chiB*((x-xB)^2+(y-yB)^2));
   omegaB=int2d(Th)(rhoB*chiB*
               (-(y-yB)*u1+(x-xB)*u2))/Jh;

   /* == Update of the position *************/
   xB=xB+uB*dt;
   yB=yB+vB*dt;
   thetaB=thetaB+omegaB*dt;
   cosB=cos(thetaB);
   sinB=sin(thetaB);

   chiB=(((x-xB)*cosB+(y-yB)*sinB)^2/rB^2
    +(-(x-xB)*sinB+(y-yB)*cosB)^2/dB^2<=1.0);
   f2=-(rhoB-rhoF)*g*chiB;
   cout<<"thetaB="<<<<thetaB<<endl;
}
```

References

[1] D.N. Arnold, F. Brezzi and M. Fortin, A stable finite element for the Stokes equations. *Calcolo*, 23(4):337–344, 1985.

[2] I. Babuska, The finite element method with Lagrangian multipliers. *Num. Math.*, 20:179–192, 1973.

[3] F. Boyer and P. Fabrie, Mathematical tools for the study of the incompressible Navier-Stokes equations and related models. *Springer*, 2013.

[4] H. Brezis, Analyse fonctionnelle : théorie et applications. *Masson*, 1983.

[5] F. Brezzi, On the existence uniqueness and approximation of saddle-point problems arising from Lagrange multipliers. *RAIRO Mathematical Modelling and Numerical Analysis*, 8(R2):129–151, 1974.

[6] F. Brezzi and M. Fortin, Mixed and hybrid finite element method. *Springer*, 1991.

[7] P. Clément, Approximation by finite element functions using local regularization. *RAIRO Numerical Analysis*, 9(R-2), 77–84, 1975.

[8] J. Douglas and J. Wang, An absolutely stabilized finite element method for the Stokes problem. *Math. of Comp.*, 52(186):495–508, 1989.

[9] A. Ern and J.-L. Guermond, Theory and practice of finite elements. *Springer*, 2004.

[10] G.P. Galdi, An introduction to the mathematical theory of the Navier-Stokes equations: steady-state problems. *Springer*, 1994.

[11] V. Girault and P.-A. Raviart, Finite element methods for Navier-Stokes equations. *Springer*, 1986.

[12] F. Hecht, New development in FreeFem++. *J. Numer. Math.*, 20(3-4):251–265, 2012.

[13] T.J.R. Hughes, L.P. Franca, and M. Balestra, A new finite element formulation for computational fluid dynamics: V. Circumventing the Babuska-Brezzi condition: a stable Petrov-Galerkin formulation of the Stokes problem accommodating equal-order interpolations. *Comp. Meth. App. Mech. Eng.*, 59(1):85–99,1986.

[14] A. Novotny and I. Straskraba, Introduction to the mathematical theory of compressible flow. *Oxford University Press*, 2004.

[15] A. Quarteroni and A. Valli, Approximation of partial differential equations: theory and numerical analysis. *North-Holland*, 1979.

[16] L.R. Scott and S. Zhang, Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Math. Comp.*, 54(190):483–493, 1990.

[17] R. Temam, Navier-Stokes Equations: theory and numerical analysis. *North-Holland*, 1979.