EUROFUSION WP15ER-PR(15) 14214

C. Negulescu et al.

Numerical analysis of an asymptotic-preserving scheme for anisotropic elliptic equations

Preprint of Paper to be submitted for publication in
Numerische Mathematik

This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom research and training programme 2014-2018 under grant agreement No 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

Numerical analysis of an asymptotic-preserving scheme for anisotropic elliptic equations

Alexei Lozinski[§] Jacek Narski[†] Claudia Negulescu[†]

June 26, 2015

Abstract

The main purpose of the present paper is to study from a numerical analysis point of view some robust methods designed to cope with stiff (highly anisotropic) elliptic problems. The so-called asymptotic-preserving schemes studied in this paper are very efficient in dealing with a wide range of ε-values, where $0 < \varepsilon \ll 1$ is the stiffness parameter, responsible for the high anisotropy of the problem. In particular, these schemes are even able to capture the macroscopic properties of the system, as ε tends towards zero, while the discretization parameters remain fixed. The objective of this work shall be to prove some ε-independent convergence results for these numerical schemes and put hence some more rigor in the construction of such AP-methods.

Keywords: Anisotropic elliptic problem, Asymptotic-Preserving scheme, Numerical analysis, Saddle-point problem, Inf-sup condition, Stabilization, Convergence.

1 Introduction

In a series of previous works [4, 5, 6, 11, 12] some efficient numerical schemes were introduced in the aim to solve at a moderate computational cost some highly anisotropic elliptic and parabolic problems. The interest in solving such problems comes for example from their regular occurrence in the modeling of magnetically confined plasmas [3, 9] and ionospheric plasmas [13], where the strong magnetic field creates anisotropy. An accurate and not resource demanding description of tokamak plasma dynamics is crucial for succeeding in the construction of a thermonuclear fusion reactor, producing clean energy for the future.

The problems cited above involve a small parameter $0 < \varepsilon \ll 1$ measuring the anisotropy ratio in the diffusion matrix. This feature makes their numerical treatment

[†]Institut de Mathématiques de Toulouse, UMR 5219 , Université de Toulouse, CNRS, UPS/IMT, 118 route de Narbonne, F-31062 Toulouse, France

[§]Université de Franche-Comté, Laboratoire de Mathématiques, 16 route de Gray, 25030 Besançon, France

rather involved, since the problems degenerate in the limit $\varepsilon \to 0$ leading to a breakdown of traditional schemes for ε very small. This is caused both by the huge, ε-dependent condition number of the discretized problem and by locking phenomena (the strong diffusion along a magnetic field line makes the solution to be almost constant along these lines, which is incompatible with an approximation by piecewise polynomials unless the computational mesh is well aligned with the field). In the previous works some efficient so-called *asymptotic-preserving* schemes were proposed and were shown to be able to cope with the deficiencies of traditional schemes. Their basic idea is to mimic on the discrete level the asymptotic behaviour of the continuous solution u^ε in the limit $\varepsilon \to 0$, thus making the diagram in Fig. 1 commutative.

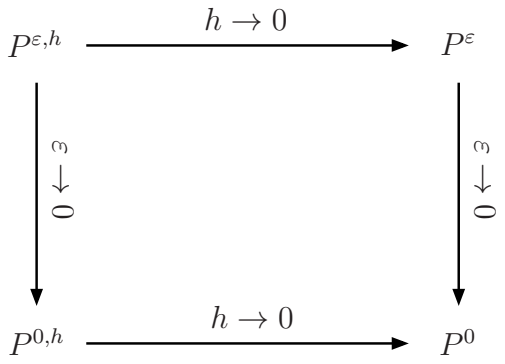


Figure 1: Properties of AP-schemes

In the present paper, we restrict out attention to the case of elliptic linear problems and are interested in two Asymptotic-Preserving schemes proposed in [6] and [12]. The efficiency and advantages of the different schemes was put into evidence numerically. However, the rigorous numerical analysis of these schemes is still lacking and is the subject of the present paper. The trick that makes these schemes work is the introduction of an auxiliary variable q^ε which serves as a Lagrange multiplier in the limit $\varepsilon \to 0$ corresponding to the constraint on u^ε, which results from the degeneracy of the governing equations. We are thus in the realm of mixed problems, their penalized variants and the discretizations thereof, as in [1, 8]. One cannot adapt though directly the techniques from these books to the present case as the inf-sup conditions are not satisfied on the discrete level, when one discretizes with standard finite elements as in the above cited papers. In fact, the choice of appropriate functional spaces even on the continuous level is not straightforward. We are going here to propose an adequate functional setting with the inf-sup condition being introduced in a non standard way. We shall develop then a complete analysis of the finite element schemes with ε-independent constants in the error estimates, relying on some discrete inf-sup conditions in h-dependent norms.

This paper is organized as follows. In Section 2, we introduce the anisotropic elliptic problem, which is the starting point of this work. A first Asymptotic-Preserving scheme for this problem, slightly modifying that proposed in [6] and well adapted for open field-line configurations, is then presented and analyzed in detail. Section 3 is concerned with the introduction of a different Asymptotic-Preserving reformulation

of the same anisotropic elliptic problem, being able to cope even with closed field-line configurations, which are often encountered in tokamak plasma modeling. This leads to the scheme proposed in [12]. A detailed numerical analysis of this scheme is then carried on. In Section 4 we validate numerically the error estimates obtained. Finally, some technical lemmas are postponed to the Appendices A and B.

2 An AP-scheme for open field-line configurations

Before presenting our model problem, let us first define some important quantities. Let b be a smooth field in a domain $\Omega \subset \mathbb{R}^d$, with $d = 2, 3$, and let us decompose the regular boundary $\Gamma = \partial\Omega$ into three components following the sign of the intersection with b:

$$\Gamma_D := \{x \in \Gamma \ / \ b(x) \cdot n(x) = 0\}\,, \quad \Gamma_N := \Gamma_{in} \cup \Gamma_{out} = \{x \in \Gamma \ / \ b(x) \cdot n(x) \lesseqgtr 0\}\,.$$

The vector n is here the unit outward normal to Γ.

The direction of the anisotropy of our problem is defined by this vector field $b \in (C^\infty(\Omega))^d$, which is supposed to satisfy $|b(x)| = 1$ for all $x \in \Omega$. Given this vector field b, one can decompose now vectors $v \in \mathbb{R}^d$, gradients $\nabla\phi$, with $\phi(x)$ a scalar function, and divergences $\nabla \cdot v$, with $v(x)$ a vector field, into a part parallel to the anisotropy direction and a part perpendicular to it. These parts are defined as follows:

$$\begin{array}{lll}
v_\parallel := (v \cdot b)\, b\,, & v_\perp := (Id - b \otimes b)v\,, & \text{such that} \quad v = v_\parallel + v_\perp\,, \\
\nabla_\parallel \phi := (b \cdot \nabla\phi)\, b\,, & \nabla_\perp \phi := (Id - b \otimes b)\, \nabla\phi\,, & \text{such that} \quad \nabla\phi = \nabla_\parallel\phi + \nabla_\perp\phi\,, \\
\nabla_\parallel \cdot\, v := \nabla \cdot v_\parallel\,, & \nabla_\perp \cdot\, v := \nabla \cdot v_\perp\,, & \text{such that} \quad \nabla \cdot v = \nabla_\parallel \cdot\, v + \nabla_\perp \cdot\, v\,.
\end{array} \tag{1}$$

Given these notations we can now introduce the highly anisotropic elliptic problem we are interested in, namely

$$\begin{cases}
-\frac{1}{\varepsilon}\nabla_\parallel \cdot \left(A_\parallel \nabla_\parallel u^\varepsilon\right) - \nabla_\perp \cdot (A_\perp \nabla_\perp u^\varepsilon) = f & \text{in } \Omega, \\
\frac{1}{\varepsilon} n_\parallel \cdot \left(A_\parallel \nabla_\parallel u^\varepsilon\right) + n_\perp \cdot (A_\perp \nabla_\perp u^\varepsilon) = 0 & \text{on } \Gamma_N, \\
u^\varepsilon = 0 & \text{on } \Gamma_D\,.
\end{cases} \tag{2}$$

The parameter $0 < \varepsilon \ll 1$ is very small, inducing rather sever numerical difficulties, when solving (2) via standard methods. Indeed, this elliptic system becomes degenerate in the limit $\varepsilon \to 0$, leading to the reduced problem

$$(R) \quad \begin{cases}
-\nabla_\parallel \cdot \left(A_\parallel \nabla_\parallel u\right) = 0 & \text{in } \Omega, \\
n_\parallel \cdot \left(A_\parallel \nabla_\parallel u\right) = 0 & \text{on } \Gamma_N, \\
u = 0 & \text{on } \Gamma_D\,,
\end{cases} \tag{3}$$

which has an infinite amount of solutions, all of them being constant along the field lines. Numerically this degeneracy translates in a very ill-conditioned linear system to be solved when $0 < \varepsilon \ll 1$.
The aim of the present section will be the mathematical study of the elliptic problem (2), in particular the investigation of its asymptotic behaviour as ε tends towards zero, the introduction of an *Asymptotic-Preserving* reformulation, better suited to pass to the limit $\varepsilon \to 0$, and the detailed numerical analysis of the designed AP-scheme. The reformulation of the singularly-perturbed problem (2) is based on asymptotic arguments and is a sort of "reorganization" of the problem into a form, which allows for an automatic numerical transition from (2) towards the limit-model (to be determined) as $\varepsilon \to 0$, while keeping the discretization parameters fixed.

2.1 Inflow Asymptotic-Preserving reformulation

In order to avoid all the above mentioned difficulties corresponding to the non-uniqueness of the reduced problem (R), one has to pick up within all its solutions the right limit solution, by fixing in an adequate manner its value on the field lines. This was done in the previous works [4, 5, 6], via the introduction of Lagrange multipliers, which are necessary to recover the uniqueness in the limit $\varepsilon \to 0$. The numerical resolution of the thus obtained Asymptotic-Preserving reformulations was shown to be stable and accurate independently on the parameter ε, which is a great advantage as compared to standard discretizations for (2). This essential property of the designed AP-scheme was proved numerically, its rigorous numerical analysis being the subject of the present paper.

For the mathematical study, let us assume that the diffusion coefficients and the source term satisfy the following hypothesis:

Hypothesis A *Let $f \in H^{-1}(\Omega)$, $0 < \varepsilon < 1$ be a fixed arbitrary parameter and $\overset{\circ}{\Gamma}_D \neq \varnothing$. The diffusion coefficients $A_\parallel \in W^{2,\infty}(\Omega)$ and $A_\perp \in \mathbb{M}_{d\times d}(W^{2,\infty}(\Omega))$ are supposed to verify the bounds*

$$0 < A_0 \leq A_\parallel(x) \leq A_1\,, \quad \text{for a.a. } x \in \Omega, \tag{4}$$

$$A_0\|v\|^2 \leq v^t A_\perp(x) v \leq A_1\|v\|^2\,, \quad \forall v \in \mathbb{R}^d \text{ with } v\cdot b(x) = 0 \text{ for a.a. } x \in \Omega, \tag{5}$$

with some constants $0 < A_0 \leq A_1$.

Before we shall pass to a brief presentation of an AP-reformulation of (2), we shall rewrite this problem in a slightly different form, masking the perpendicular derivatives, which turn out to be cumbersome for the numerical analysis. Indeed, the following reformulation, called in the following $(P)^\varepsilon$-problem

$$(P)^\varepsilon \quad \begin{cases} -\frac{1-\varepsilon}{\varepsilon}\nabla_\parallel \cdot \left(A_\parallel \nabla_\parallel u^\varepsilon\right) - \nabla\cdot\left(A\nabla u^\varepsilon\right) = f & \text{in } \Omega, \\ \frac{1-\varepsilon}{\varepsilon} n_\parallel \cdot \left(A_\parallel \nabla_\parallel u^\varepsilon\right) + n\cdot\left(A\nabla u^\varepsilon\right) = 0 & \text{on } \Gamma_N, \\ u^\varepsilon = 0 & \text{on } \Gamma_D\,. \end{cases} \tag{6}$$

is easily seen to be equivalent to problem (2) by setting

$$A := (b \otimes b)\, A_{\|}\, (b \otimes b) + (Id - b \otimes b)\, A_{\perp}\, (Id - b \otimes b).$$

Remark that by Hypothesis A, we have immediately $A_0\|v\|^2 \leq v^t A(x) v \leq A_1\|v\|^2$ for all $v \in \mathbb{R}^d$ and for a.a. $x \in \Omega$, which is a sort of coercivity and boundedness property for the diffusivity matrix A.

Let us now introduce the mathematical framework and define the Hilbert space $\mathcal{V}$ as follows

$$\mathcal{V} := \{v \in H^1(\Omega)\,,\ \text{such that } v|_{\Gamma_D} = 0\}, \tag{7}$$

equipped with the scalar product

$$(u, v)_{\mathcal{V}} = a(u, v) := \int_{\Omega} A\nabla u \cdot \nabla v \, dx\,. \tag{8}$$

In the following, the bracket $(\cdot, \cdot)$ will stand for the standard L^2-scalar product. We shall also frequently use the bilinear form $a_{\|} : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ and the corresponding semi-norm $|\cdot|_{\|} : \mathcal{V} \to \mathbb{R}$ defined by

$$a_{\|}(u, v) := \int_{\Omega} A_{\|}\nabla_{\|} u \cdot \nabla_{\|} v \, dx\,, \quad |u|_{\|} := \sqrt{a_{\|}(u, u)}\,. \tag{9}$$

The weak formulation of problem (6) can be now written as: Find $u^\varepsilon \in \mathcal{V}$ such that

$$(P)^\varepsilon \qquad \frac{1-\varepsilon}{\varepsilon} a_{\|}(u^\varepsilon, v) + a(u^\varepsilon, v) = (f, v)\,, \quad \forall v \in \mathcal{V}\,. \tag{10}$$

Thanks to Hypothesis A and to Lax-Milgram theorem, problem (10) admits a unique solution $u^\varepsilon \in \mathcal{V}$ for all $\varepsilon > 0$.

The design of efficient schemes, which are uniformly stable along the transition $\varepsilon \to 0$, is based on the fundamental fact, that the solutions $u^\varepsilon \in \mathcal{V}$ of (10) tend for $\varepsilon \to 0$ towards some function u^0, constant along the field lines of b, i.e. belonging to the following Hilbert-space

$$\mathcal{G} = \{v \in \mathcal{V}\,,\ \text{such that } \nabla_{\|} v = 0\}\,, \quad (u, v)_{\mathcal{G}} := (\nabla_{\perp} u, \nabla_{\perp} v)_{L^2(\Omega)}\,, \tag{11}$$

which consists of functions belonging to $\mathcal{V}$ with zero gradient along the field lines. Taking the test functions in (10) from $\mathcal{G}$, and passing formally to the limit $\varepsilon \to 0$, permits to identify the problem satisfied by $u^0 \in \mathcal{G}$, the so-called Limit model

$$(L) \quad a(u^0, v) = (f, v)\,, \quad \forall v \in \mathcal{G}\,. \tag{12}$$

Again, the Lax-Milgram theorem permits to show the existence and uniqueness of a solution $u^0 \in \mathcal{G}$ of this Limit problem (12). Remark that this Limit model is defined on a constrained space $\mathcal{G}$, and shall be equivalently reformulated in the sequel, on a constraint-less space.

The main idea behind the first AP-reformulation of problem (6) is to rescale the parallel derivative of u^ε by introducing the auxiliary variable q^ε such that

$$\nabla_\parallel q^\varepsilon = \frac{1}{\varepsilon}\nabla_\parallel u^\varepsilon. \tag{13}$$

To ensure the uniqueness of q^ε, we require in this section that $q^\varepsilon = 0$ on Γ_{in}. Remark that this is only possible if all the field lines are open and enter the domain (by Γ_{in}). For closed field lines, completely contained in Ω, fixing q^ε on Γ_{in} would be not enough for the uniqueness and other methods shall be developed in Section 3.
We thus introduce the Hilbert space

$$\mathcal{L}_{in} = \{q \in L^2(\Omega) \ / \ \nabla_\parallel q \in L^2(\Omega) \text{ and } q|_{\Gamma_{in}} = 0\}\,, \tag{14}$$

equipped with the scalar product $a_\parallel(\cdot,\cdot)$, inducing the norm $|\cdot|_\parallel$. Note that this is indeed a norm on $\mathcal{L}_{in}$ since $|q|_\parallel = 0$ for $q \in \mathcal{L}_{in}$ means $\nabla_\parallel q = 0$ on Ω, which in combination with the boundary condition on Γ_{in} implies $q = 0$.

Substituting the definition (13) of q^ε into (10) yields the following problem, called in the following Inflow Asymptotic-Preserving reformulation of $(P)^\varepsilon$: Find $(u^\varepsilon, q^\varepsilon) \in \mathcal{V} \times \mathcal{L}_{in}$ satisfying

$$(AP_{in})^\varepsilon \quad \begin{cases} a(u^\varepsilon, v) + (1-\varepsilon)a_\parallel(q^\varepsilon, v) = (f, v), & \forall v \in \mathcal{V} \\ a_\parallel(u^\varepsilon, w) - \varepsilon a_\parallel(q^\varepsilon, w) = 0, & \forall w \in \mathcal{L}_{in}\,. \end{cases} \tag{15}$$

The notation $(AP_{in})^\varepsilon$ emphasizes the fact that we are introducing an Asymptotic-Preserving reformulation of (10), based on the Lagrange multiplier q^ε, which is uniquely determined through the inflow boundary condition on Γ_{in}. The reformulation (15) is completely equivalent to the starting model (10). However, remark that putting formally $\varepsilon = 0$ in (15) and introducing for some technical reasons explained in the next subsection a larger space $\tilde{\mathcal{L}}_{in} \supset \mathcal{L}_{in}$ leads to the well-posed problem : Find $(u^0, q^0) \in \mathcal{V} \times \tilde{\mathcal{L}}_{in}$ such that

$$(L_{in}) \quad \begin{cases} a(u^0, v) + a_\parallel(q^0, v) = (f, v), & \forall v \in \mathcal{V} \\ a_\parallel(u^0, w) = 0, & \forall w \in \tilde{\mathcal{L}}_{in}\,, \end{cases} \tag{16}$$

which is an equivalent (saddle-point) reformulation of the Limit-problem (12). Indeed, instead of setting the problem on the constrained space $\mathcal{G}$, one introduces a Lagrange multiplier $q^0 \in \tilde{\mathcal{L}}_{in}$, enabling us to solve the problem on the constraint-free space $\mathcal{V} \times \tilde{\mathcal{L}}_{in}$.

2.2 The Inf-Sup condition

Let us now focus on the mathematical study of the continuous problem $(AP_{in})^\varepsilon$ and its asymptotic behaviour as ε tends towards zero. Due to the saddle-point structure of this problem, we shall make use of the traditional inf-sup theory [1, 7]. The goal is to prove an ε-independent inf-sup condition corresponding to (15), which ensures

the existence and uniqueness of a solution, as well as the convergence of the AP-solution $(u^\varepsilon, q^\varepsilon)$ towards the L-solution (u^0, q^0) as $\varepsilon \to 0$. For this, we shall need a more adequate norm on the space $\mathcal{L}_{in}$, in contrast to the one proposed in [6] (see (14)).

Indeed, we would like that the form $a_\parallel(\cdot,\cdot)$ satisfies an inf-sup estimate on the pair of spaces $\mathcal{V}$ plus the space of functions q. This would be trivially the case, if we would search for q in the space $\tilde{\mathcal{L}}_{in}$, defined as the closure of $\mathcal{L}_{in}$ in the following norm $|q|_*$

$$|q|_* := \sup_{v\in\mathcal{V}} \frac{a_\parallel(q,v)}{|v|_\mathcal{V}}\,. \tag{17}$$

Note that for all $q \in \mathcal{L}_{in}$, one has $|q|_* \le |q|_\parallel$, which means that the injection $\mathcal{L}_{in} \subset \tilde{\mathcal{L}}_{in}$ is continuous, however $\mathcal{L}_{in} \neq \tilde{\mathcal{L}}_{in}$ in general, as can be seen from the subsequent remarks.

Remark 1 *One can extend the continuous bilinear form to $a_\parallel : \tilde{\mathcal{L}}_{in} \times \mathcal{V} \to \mathbb{R}$ by defining for each $q \in \tilde{\mathcal{L}}_{in}\backslash\mathcal{L}_{in}$*

$$a_\parallel(q,v) := \lim_{n\to\infty} a_\parallel(q_n,v)\,, \quad \forall v \in \mathcal{V}\,,$$

for some $\{q_n\}_{n\in\mathbb{N}} \subset \mathcal{L}_{in}$ such that $q_n \to_{n\to\infty} q$ in $\tilde{\mathcal{L}}_{in}$.
Indeed, the sequence $\{q_n\}_{n\in\mathbb{N}}$ being a Cauchy-sequence in $\tilde{\mathcal{L}}_{in}$, one deduces immediately that $\{a_\parallel(q_n,v)\}_{n\in\mathbb{N}}$ is also a Cauchy-sequence for each fixed $v \in \mathcal{V}$, being hence convergent.

Remark 2 *For any fixed $q \in \tilde{\mathcal{L}}_{in}$, the maximum of $\frac{a_\parallel(q,v)}{|v|_\mathcal{V}}$ over $v \in \mathcal{V}$ is attained with the function $v^* \in \mathcal{V}$, which is solution to the problem*

$$(v^*, w)_\mathcal{V} = a_\parallel(q,w) \quad \forall w \in \mathcal{V}\,. \tag{18}$$

Indeed, let us fix $q \in \tilde{\mathcal{L}}_{in}$ and $v^* \in \mathcal{V}$ be the corresponding solution to (18). Then, any $v \in \mathcal{V}$ can be decomposed as $v = \alpha v^* + v'$ with $\alpha := \frac{(v,v^*)_\mathcal{V}}{|v^*|^2_\mathcal{V}}$ and $v' \in \mathcal{V}$ verifying $(v^*, v')_\mathcal{V} = 0$. We observe then that

$$\frac{a_\parallel(q,v)}{|v|_\mathcal{V}} = \frac{\alpha|v^*|^2_\mathcal{V}}{\sqrt{\alpha^2|v^*|^2_\mathcal{V} + |v'|^2_\mathcal{V}}} \le |v^*|_\mathcal{V} = \frac{a_\parallel(q,v^*)}{|v^*|_\mathcal{V}}\,.$$

Remark 3 *In order to prove that $\mathcal{L}_{in} \neq \tilde{\mathcal{L}}_{in}$ it suffices to verify that the norms in $\mathcal{L}_{in}$ and $\tilde{\mathcal{L}}_{in}$ are not equivalent, i.e. that there is no constant $c > 0$, such that $|q|_\parallel \le c|q|_*$ for all $q \in \mathcal{L}_{in}$. For this, it suffices to construct a sequence $\{q_k\}_{k\in\mathbb{N}} \subset \mathcal{L}_{in}$, such that*

$$\sup_{v\in\mathcal{V}} \frac{a_\parallel(q_k,v)}{|q_k|_\parallel |v|_\mathcal{V}} \le \frac{c}{k} \;\Rightarrow\; |q_k|_* \le \frac{c}{k}|q_k|_\parallel \;\Rightarrow\; |q_k|_\parallel \ge \frac{k}{c}|q_k|_*\,, \quad \forall k \in \mathbb{N}, \tag{19}$$

with $c > 0$ a constant. One can easily do it in the following simple setting: let $\Omega = (0,\pi)\times(0,\pi)$, $A_\parallel = 1$, $A_\perp = Id$, $b = e_2$. Taking $q_k = \sin kx(\cos y - \cos 2y)$,

which is a function in $\mathcal{L}_{in}$ for any integer $k > 0$, we can find the solution v_k^* to problem (18) corresponding to $q = q_k$ as

$$v_k^* = \sin kx \,(\frac{1}{k^2+1}\cos y - \frac{4}{k^2+4}\cos 2y).$$

Now, in view of Remark 2

$$\sup_{v\in\mathcal{V}} \frac{a_{\|}(q_k, v)}{|q_k|_{\|}|v|_{\mathcal{V}}} = \frac{a_{\|}(q_k, v_k^*)}{|q_k|_{\|}|v^*|_{\mathcal{V}}} = \frac{|v_k^*|_{\mathcal{V}}}{|q_k|_{\|}} = \frac{1}{\sqrt{5}}\sqrt{\frac{1}{k^2+1} + \frac{16}{k^2+4}}.$$

This gives an example of (19).

Searching now for a solution (u, q) belonging to $\mathcal{V} \times \tilde{\mathcal{L}}_{in}$ is the proper setting for our problem in the limit case $\varepsilon = 0$. Indeed, in this particular case, we have to cope with a standard saddle point problem (L_{in}) and the inf-sup condition is satisfied in the space $\mathcal{V} \times \tilde{\mathcal{L}}_{in}$. However, this choice does not work any more for $\varepsilon > 0$, as the term $a_{\|}(q, w)$ makes no more sense if we suppose only $(q, w) \in \tilde{\mathcal{L}}_{in} \times \tilde{\mathcal{L}}_{in}$. Hence, we propose to work for $\varepsilon > 0$ in the previous space $\mathcal{L}_{in}$ for the Lagrange multiplier q, associated however with the following slightly different norm

$$|q|_{\varepsilon} := (|q|_*^2 + \varepsilon|q|_{\|}^2)^{\frac{1}{2}}, \quad \forall q \in \mathcal{L}_{in}, \tag{20}$$

which is equivalent to the old norm $|\cdot|_{\|}$ of $\mathcal{L}_{in}$ with ε-dependent equivalence constants exploding as $\varepsilon \to 0$:

$$|q|_{\varepsilon} \le \sqrt{1+\varepsilon}|q|_{\|} \text{ and } |q|_{\|} \le \frac{1}{\sqrt{\varepsilon}}|q|_{\varepsilon}.$$

The space $(\mathcal{L}_{in}, |\cdot|_{\varepsilon})$ is a Hilbert one equipped with the scalar product

$$(q_1, q_2)_{\varepsilon} := (v_1^*, v_2^*)_{\mathcal{V}} + \varepsilon a_{\|}(q_1, q_2), \quad \forall q_1,\, q_2 \in \mathcal{L}_{in},$$

where v_1^*, v_2^* are the unique solutions of problem (18). In the limit $\varepsilon \to 0$, this space transforms into the Hilbert space $(\tilde{\mathcal{L}}_{in}, |\cdot|_*)$ with the scalar product $(q_1, q_2)_* := (v_1^*, v_2^*)_{\mathcal{V}}$.

We are finally able to introduce the right mathematical setting for a rigorous study of the AP-problem (15) and its convergence towards the Limit-problem (16). The Hilbert space adapted to our problem is

$$\mathcal{X}_{\varepsilon} := \begin{cases} \mathcal{V} \times \mathcal{L}_{in} & \text{for} \quad \varepsilon > 0 \\ \mathcal{V} \times \tilde{\mathcal{L}}_{in} & \text{for} \quad \varepsilon = 0, \end{cases} \qquad \|u, q\|_{\mathcal{X}_{\varepsilon}} := (|u|_{\mathcal{V}}^2 + |q|_*^2 + \varepsilon|q|_{\|}^2)^{1/2},$$

and the problem we are interested in, can now be simply written as: Find for each $\varepsilon \in [0, 1]$ the solution $(u^{\varepsilon}, q^{\varepsilon}) \in \mathcal{X}_{\varepsilon}$ to

$$(AP_{in})^{\varepsilon} \quad \begin{cases} a(u^{\varepsilon}, v) + (1-\varepsilon)a_{\|}(q^{\varepsilon}, v) = (f, v), \\ a_{\|}(u^{\varepsilon}, w) - \varepsilon a_{\|}(q^{\varepsilon}, w) = 0, \end{cases}, \quad \forall (v, w) \in \mathcal{X}_{\varepsilon}. \tag{21}$$

For the further developments, we shall also introduce the coupled bilinear form $C_\varepsilon : \mathcal{X}_\varepsilon \times \mathcal{X}_\varepsilon \to \mathbb{R}$ defined as

$$C_\varepsilon((u,q),(v,w)) := a(u,v) + (1-\varepsilon)a_\parallel(q,v) + a_\parallel(u,w) - \varepsilon a_\parallel(q,w)\,. \tag{22}$$

This bilinear form C_ε is uniformly continuous in $\varepsilon \in [0,1]$, *i.e.*

$$C_\varepsilon((u,q),(v,w)) \leq 2\|u,q\|_{\mathcal{X}_\varepsilon}\|v,w\|_{\mathcal{X}_\varepsilon}\,, \quad \forall (u,q),(v,w) \in \mathcal{X}_\varepsilon\,, \tag{23}$$

as, using Cauchy-Schwarz inequality, one has

$$\begin{aligned} C_\varepsilon((u,q),(v,w)) &\leq |u|_{\mathcal{V}}|v|_{\mathcal{V}} + |q|_*|v|_{\mathcal{V}} + |u|_{\mathcal{V}}|w|_* + \varepsilon|q|_\parallel|w|_\parallel \\ &\leq \left(2|u|_{\mathcal{V}}^2 + |q|_*^2 + \varepsilon|q|_\parallel^2\right)^{\frac{1}{2}} \left(2|v|_{\mathcal{V}}^2 + |w|_*^2 + \varepsilon|w|_\parallel^2\right)^{\frac{1}{2}}\,. \end{aligned}$$

The form C_ε enjoys furthermore the inf-sup property

$$\inf_{(u,q)\in\mathcal{X}_\varepsilon} \sup_{(v,w)\in\mathcal{X}_\varepsilon} \frac{C_\varepsilon((u,q),(v,w))}{\|u,q\|_{\mathcal{X}_\varepsilon}\|v,w\|_{\mathcal{X}_\varepsilon}} \geq \beta\,, \tag{24}$$

with a constant $\beta > 0$ that does not depend on ε. This is established in the following lemma which is recast to a slightly more general and abstract setting. Our particular result is recovered from this lemma setting the bilinear forms $a(\cdot,\cdot)$, $b(\cdot,\cdot)$ and $c(\cdot,\cdot)$ from the lemma to, respectively, $a(\cdot,\cdot)$, $a_\parallel(\cdot,\cdot)$ and $a_\parallel(\cdot,\cdot)$. Note that this result is very close to those from Section 4.3 of [8] but we do not require here an inf-sup condition for the form b in $V \times L$.

Lemma 4 ***(Inf-Sup condition)*** *Let V and L be Hilbert spaces with their respective scalar products $a(\cdot,\cdot)$ and $c(\cdot,\cdot)$ inducing the norms $\|\cdot\|_V$ and $\|\cdot\|_L$. Let moreover $\tilde{L} \supset L$ be another Hilbert space with the norm $\|\cdot\|_{\tilde{L}}$ such that $\|q\|_{\tilde{L}} \leq \|q\|_L$ for all $q \in L$. Let $b : \tilde{L} \times V \to \mathbb{R}$ be a bilinear form satisfying the continuity relation $\|b(q,v)\| \leq \|v\|_V\|q\|_{\tilde{L}}$ for all $v \in V$, $q \in L$ and*

$$\inf_{q\in\tilde{L}} \sup_{v\in V} \frac{b(q,v)}{\|q\|_{\tilde{L}}\|v\|_V} = \alpha > 0. \tag{25}$$

Set $L_\varepsilon := L$ for any $\varepsilon > 0$, $L_0 := \tilde{L}$ and let X_ε for any $\varepsilon \geq 0$ denote the Hilbert space $V \times L_\varepsilon$ equipped with the norm $\|u,q\|_{X_\varepsilon} := (\|u\|_V^2 + \|q\|_{\tilde{L}}^2 + \varepsilon\|q\|_L^2)^{1/2}$. Introduce for any $\varepsilon \in [0,1]$ the bilinear form $C_\varepsilon : X_\varepsilon \times X_\varepsilon \to \mathbb{R}$

$$C_\varepsilon((u,q),(v,w)) = a(u,v) + (1-\varepsilon)b(q,v) + b(w,u) - \varepsilon c(q,w)\,.$$

Then C_ε is continuous, with continuity constant $M = 2$, and satisfies moreover the inf-sup condition

$$\inf_{(u,q)\in X_\varepsilon} \sup_{(v,w)\in X_\varepsilon} \frac{C_\varepsilon((u,q),(v,w))}{\|u,q\|_{X_\varepsilon}\|v,w\|_{X_\varepsilon}} \geq \beta\,, \tag{26}$$

with a constant $\beta > 0$ that depends only on α.

Proof. To prove (26), let us fix an arbitrary $(u,q) \in X_\varepsilon$ and denote

$$Z := \sup_{(v,w)\in X_\varepsilon} \frac{C_\varepsilon((u,q),(v,w))}{\|v,w\|_{X_\varepsilon}}\,.$$

We want to prove that $Z \geq \beta\|u,q\|_{X_\varepsilon}$. First, we have

$$(1-\varepsilon)\alpha\|q\|_{\tilde{L}} \leq \sup_{v\in V} \frac{(1-\varepsilon)b(q,v)}{\|v\|_V} \leq \sup_{v\in V} \frac{C_\varepsilon((u,q),(v,0))}{\|v\|_V} + \sup_{v\in V} \frac{a(u,v)}{\|v\|_V} \leq Z + \|u\|_V\,.$$

Now, we take $v = u$, $w = -q$ and observe that

$$\begin{aligned} C_\varepsilon((u,q),(u,-q)) &= a(u,u) - \varepsilon b(q,u) + \varepsilon c(q,q) \\ &\geq (1-\frac{\varepsilon}{2})\|u\|_V^2 + \frac{\varepsilon}{2}\|q\|_L^2\,, \end{aligned}$$

implying altogether

$$\begin{aligned} \frac{1}{2}\|u\|_V^2 + \frac{\varepsilon}{2}\|q\|_L^2 + \frac{\alpha^2(1-\varepsilon)^2}{8}\|q\|_{\tilde{L}}^2 &\leq C_\varepsilon((u,q),(u,-q)) + \frac{1}{8}(Z+\|u\|_V)^2 \\ &\leq Z\|u,-q\|_{X_\varepsilon} + \frac{1}{4}Z^2 + \frac{1}{4}\|u\|_V^2\,. \end{aligned}$$

Thus,

$$\frac{1}{4}\|u\|_V^2 + \frac{\varepsilon}{2}\|q\|_L^2 + \frac{\alpha^2(1-\varepsilon)^2}{8}\|q\|_{\tilde{L}}^2 \leq Z\|u,q\|_{X_\varepsilon} + \frac{1}{2}Z^2 \leq \frac{1}{2\gamma}\|u,q\|_{X_\varepsilon}^2 + \frac{1+\gamma}{2}Z^2\,,$$

for any $\gamma > 0$ by Young inequality. Besides, we have for any $\varepsilon \in [0,1]$

$$\frac{\varepsilon}{2}\|q\|_L^2 + \frac{\alpha^2(1-\varepsilon)^2}{8}\|q\|_{\tilde{L}}^2 \geq c_0(\|q\|_{\tilde{L}}^2 + \varepsilon\|q\|_L^2)\,,$$

with a constant $c_0 > 0$ depending only on α. Indeed, for $\varepsilon \in [0,1/2]$ we can observe that $(1-\varepsilon)^2 \geq \frac{1}{4}$ and conclude. For $\varepsilon \in [1/2,1]$, we can neglect the term with $\|q\|_{\tilde{L}}^2$ on the left-hand side and use $\|q\|_{\tilde{L}} \leq \|q\|_L$.
Thus, assuming without loss of generality that $c_0 \leq \frac{1}{4}$ we have

$$c_0\|u,q\|_{X_\varepsilon}^2 \leq \frac{1}{2\gamma}\|u,q\|_{X_\varepsilon}^2 + \frac{1+\gamma}{2}Z^2\,.$$

Taking finally a sufficiently big γ gives immediately $\|u,q\|_{X_\varepsilon} \leq (1/\beta)Z$ with a constant $\beta > 0$, independent of ε. ■

The following theorem is the main theorem on the continuous level, which shows that the AP-reformulation (15) of problem (10) is well-posed and better adapted to capture the macro-scale behaviour of u^ε in the limit $\varepsilon \to 0$. This AP-model provides thus a link between the micro-scale ($\varepsilon \sim 1$) and the macro-scale ($\varepsilon \sim 0$) behaviour of the system.

Theorem 5 ***(Existence/Uniqueness/ε-Convergence)*** *Let hypothesis A be satisfied. The AP-problem (21) is well-posed for each $\varepsilon \in [0,1]$, i.e. for any $f \in \mathcal{V}'$ and any $\varepsilon \in [0,1]$ there exists a unique solution $(u^\varepsilon, q^\varepsilon) \in \mathcal{X}_\varepsilon$, which satisfies*

$$\|u^\varepsilon, q^\varepsilon\|_{\mathcal{X}_\varepsilon} \leq \frac{1}{\beta}\|f\|_{\mathcal{V}'}\,,$$

with $\beta > 0$ the constant given by the inf-sup condition (26). Moreover, we have the ε-convergences

$$\|u^\varepsilon - u^0, q^\varepsilon - q^0\|_{\mathcal{X}_\varepsilon} \to 0\,, \quad \textit{for} \quad \varepsilon \to 0\,.$$

If we suppose more regular data, as $f \in L^2(\Omega)$, then one has even $q^0 \in \mathcal{L}_{in}$ and the estimates

$$|u^\varepsilon - u^0|_{\mathcal{V}} \leq C\sqrt{\varepsilon}\,, \quad |q^\varepsilon - q^0|_* \leq C\sqrt{\varepsilon}\,, \tag{27}$$

with $C > 0$ some ε-independent constant.

Proof. The existence and uniqueness of a solution $(u^\varepsilon, q^\varepsilon) \in \mathcal{X}_\varepsilon$ for each $\varepsilon \geq 0$, is a simple consequence of the Banach-Nečas-Babuška (hereafter BNB) theorem [7]. To prove the convergence $(u^\varepsilon, q^\varepsilon) \to (u^0, q^0)$ we assume first that $f \in L^2(\Omega)$. We have proved in [6] that $q^0 \in \mathcal{L}_{in}$ in this case. Subtracting now (16) from (15) yields

$$C_\varepsilon((u^\varepsilon - u^0, q^\varepsilon - q^0), (v,w)) = \varepsilon a_\parallel(q^0, v+w)\,, \quad \forall (v,w) \in \mathcal{V} \times \mathcal{L}_{in}\,.$$

Thus, for any $\varepsilon > 0$, by the inf-sup property, there exist $(v,w) \in \mathcal{X}_\varepsilon = \mathcal{V} \times \mathcal{L}_{in}$ such that

$$\beta'\|u^\varepsilon - u^0, q^\varepsilon - q^0\|_{\mathcal{X}_\varepsilon}\,\|v,w\|_{\mathcal{X}_\varepsilon} \leq \varepsilon a_\parallel(q^0, v+w) \leq \varepsilon |q^0|_\parallel |v+w|_\parallel \leq \sqrt{2\varepsilon}|q^0|_\parallel \|v,w\|_{\mathcal{X}_\varepsilon}\,,$$

with some $0 < \beta' < \beta$, for ex. $\beta' := \beta/2$, which implies $\|u^\varepsilon - u^0, q^\varepsilon - q^0\|_{\mathcal{X}_\varepsilon} \leq \frac{\sqrt{2\varepsilon}}{\beta'}\,|q^0|_\parallel$, leading to the convergence estimates (27).
We are now going to generalize this result to any $f \in \mathcal{V}'$ by a density argument. Let us denote simply by $U^\varepsilon(f)$ the solution $(u^\varepsilon, q^\varepsilon) \in \mathcal{X}_\varepsilon$ of (21) associated to $f \in \mathcal{V}'$. Now fix some $f \in \mathcal{V}'$. Since $L^2(\Omega)$ is dense in $\mathcal{V}'$, for any $\delta > 0$ there exists $f_\delta \in L^2(\Omega)$ such that $f = f_\delta + R_\delta$ with $\|R_\delta\|_{\mathcal{V}'} < \frac{\delta\beta'}{4}$. Hence, there exists $\varepsilon_0 > 0$ such that for all $\varepsilon < \varepsilon_0$

$$\|U^\varepsilon(f) - U^0(f)\|_{\mathcal{X}_\varepsilon} \leq \|U^\varepsilon(f_\delta) - U^0(f_\delta)\|_{\mathcal{X}_\varepsilon} + \|U^\varepsilon(R_\delta) - U^0(R_\delta)\|_{\mathcal{X}_\varepsilon} < \frac{\delta}{2} + \frac{2}{\beta'}\|R_\delta\|_{\mathcal{V}'} < \delta\,.$$

Here we used the fact that $f_\delta \in L^2(\Omega)$ which implies $\|U^\varepsilon(f_\delta) - U^0(f_\delta)\|_{\mathcal{X}_\varepsilon} \leq C\frac{\sqrt{\varepsilon}}{\beta'} \to 0$ as $\varepsilon \to 0$. ■

2.3 The numerical analysis of the inflow AP-scheme

Having reformulated on the continuous level the singularly-perturbed problem $(P)^\varepsilon$ into a system $(AP_{in})^\varepsilon$ which is better suited to capture the macroscopic limit as $\varepsilon \to 0$, we shall now discretize via a standard approach this new system and analyse

the obtained AP-scheme in detail. In particular error estimates are deduced and the convergence of the scheme independently on the anisotropy parameter ε is shown.

Let us introduce a mesh $\mathcal{T}_h$ on Ω consisting of triangles (resp. rectangles) of maximal size h, let $V_h \subset \mathcal{V}$ be the finite dimensional space of $\mathbb{P}_k$ (resp. $\mathbb{Q}_k$) finite elements on $\mathcal{T}_h$, and let us define $L_h := V_h \cap \mathcal{L}_{in} = V_h \cap \tilde{\mathcal{L}}_{in}$ as well as $X_h := V_h \times L_h$. Note that we require $L_h \subset V_h$, which signifies that we enforce the boundary conditions on Γ_D for functions in L_h, cf. [6]. We are thus looking for a discrete solution $(u_h^\varepsilon, q_h^\varepsilon) \in V_h \times L_h$ of

$$(AP_{in})_h^\varepsilon \quad \begin{cases} a(u_h^\varepsilon, v_h) + (1-\varepsilon) a_\parallel(q_h^\varepsilon, v_h) = (f, v_h)\,, & \forall v_h \in V_h \\ a_\parallel(u_h^\varepsilon, w_h) - \varepsilon a_\parallel(q_h^\varepsilon, w_h) = 0, & \forall w_h \in L_h\,. \end{cases} \tag{28}$$

The analysis of this scheme would be straightforward if the discrete inf-sup condition

$$\inf_{q_h \in L_h} \sup_{v_h \in V_h} \frac{a_\parallel(q_h, v_h)}{|q_h|_* |v_h|_{\mathcal{V}}} \geq \alpha\,, \tag{29}$$

were satisfied with an ε- as well as mesh-independent constant $\alpha > 0$. However, this constant is unfortunately mesh dependent, as shown in Appendix C. In order to circumvent this difficulty we introduce the following mesh-dependent norm on L_h

$$|q_h|_{*h} := \sup_{v_h \in V_h} \frac{a_\parallel(q_h, v_h)}{|v_h|_{\mathcal{V}}}\,.$$

Note that this is indeed a norm, since $|q_h|_{*h} = 0$ implies $a_\parallel(q_h, q_h) = 0$ due to the inclusion $L_h \subset V_h$ and thus $\nabla_\parallel q_h = 0$, which in combination with the boundary conditions on Γ_{in} yields $q_h = 0$.

We now equip the space X_h with the norm

$$\|u_h, q_h\|_{X_{\varepsilon,h}} = (|u_h|_{\mathcal{V}}^2 + |q_h|_{*h}^2 + \varepsilon |q_h|_\parallel^2)^{1/2}\,.$$

By Lemma 4, the bilinear form C_ε is continuous on X_h with this norm, and enjoys the inf-sup property

$$\inf_{(u_h,q_h) \in X_h} \sup_{(v_h,w_h) \in X_h} \frac{C_\varepsilon((u_h,q_h),(v_h,w_h))}{\|u_h,q_h\|_{X_{\varepsilon,h}} \|v_h,w_h\|_{X_{\varepsilon,h}}} \geq \beta\,, \tag{30}$$

with a constant $\beta > 0$ that does not depend neither on the mesh nor on ε. This implies the discrete version of theorem 5.

Theorem 6 ***(Discrete Existence/Uniqueness/ε-Convergence)*** *The discrete AP-problem (28) admits for each fixed $h > 0$ and $\varepsilon \geq 0$ a unique solution $(u_h^\varepsilon, q_h^\varepsilon) \in V_h \times L_h$, satisfying*

$$\|u_h^\varepsilon, q_h^\varepsilon\|_{X_{\varepsilon,h}} \leq \frac{1}{\beta} \|f\|_{\mathcal{V}'}\,,$$

and one has the ε-convergence

$$\|u_h^\varepsilon - u_h^0, q_h^\varepsilon - q_h^0\|_{X_{\varepsilon,h}} \to 0 \quad \textit{for} \quad \varepsilon \to 0.$$

Moreover, the condition number of the matrix corresponding to problem (28) is bounded by a constant independent of ε (assuming that the same bases of V_h and L_h are chosen for all values of ε).

Proof. The existence and uniqueness of a solution $(u_h^\varepsilon, q_h^\varepsilon) \in X_h$ for each $\varepsilon \geq 0$, is a simple consequence of the BNB theorem [7]. The convergence $(u_h^\varepsilon, q_h^\varepsilon) \to_{\varepsilon\to 0} (u_h^0, q_h^0)$ can be established by the same arguments as in the proof of Theorem 5.

We turn now to the study of the condition number. Let $\{\phi_1^u, \ldots, \phi_{N^u}^u\}$ (resp. $\{\phi_1^q, \ldots, \phi_{N^q}^q\}$) be a basis of V_h (resp. L_h). We shall identify every function $u_h \in V_h$ (resp. $q_h \in L_h$) with a vector $\vec{u} \in \mathbb{R}^{N_u}$ (resp. $\vec{q} \in \mathbb{R}^{N_q}$) consisting of the expansion coefficients of u_h (resp. q_h) in these bases. Denoting the Euclidean norm of a vector by $||\cdot||_2$ and using the equivalence of norms on a finite dimensional space, we observe that for all $u_h \in V_h$ and $q_h \in L_h$ we have

$$\mu_u||\vec{u}||_2 \leq |u_h|_\nu \leq \nu_u||\vec{u}||_2\,, \quad \mu_q||\vec{q}||_2 \leq |q_h|_{||} \leq \nu_q||\vec{q}||_2, \quad \mu_*||\vec{q}||_2 \leq |q_h|_{*h} \leq \nu_*||\vec{q}||_2$$

with some positive constants μ's and ν's. We shall moreover identify any $\Phi_h = (u_h, q_h) \in X_h$ with a vector $\vec{\Phi} \in \mathbb{R}^N$, $N = N_u + N_q$, such that $\vec{\Phi} = (\vec{u}^T, \vec{q}^T)^T$. We observe for any such Φ_h that $||\vec{\Phi}||_2^2 = ||\vec{u}||_2^2 + ||\vec{q}||_2^2$, which in combination with the estimates above gives

$$\min\{\mu_u^2, \mu_*^2 + \varepsilon\,\mu_q^2\}||\vec{\Phi}||_2^2 \leq ||\Phi_h||_{X_{\varepsilon,h}}^2 \leq \max\{\nu_u^2, \nu_*^2 + \varepsilon\,\nu_q^2\}||\vec{\Phi}||_2^2\,.$$

Let now A denote the $N_u \times N_u$ matrix with entries $a_{ij} = a(\phi_i^u, \phi_j^u)$, B the $N_u \times N_q$ matrix with entries $b_{ij} = a_{||}(\phi_i^u, \phi_j^q)$, and C the $N_q \times N_q$ matrix with entries $c_{ij} = a_{||}(\phi_i^q, \phi_j^q)$. The matrix corresponding to problem (28) can be then written in the following block form

$$\mathbb{A}^\varepsilon = \begin{pmatrix} A & (1-\varepsilon)B \\ B^T & \varepsilon C \end{pmatrix}.$$

Its 2-norm denoted by $||\cdot||_2$ is bounded for all $\varepsilon \in [0,1]$ by

$$\begin{aligned} ||\mathbb{A}^\varepsilon||_2 &= \sup_{\vec{\Phi},\vec{\Psi}\in\mathbb{R}^N\setminus\{0\}} \frac{\vec{\Psi}\cdot\mathbb{A}^\varepsilon\vec{\Phi}}{||\vec{\Phi}||_2||\vec{\Psi}||_2} = \sup_{\Phi_h,\Psi_h\in X_h\setminus\{0\}} \frac{C_\varepsilon(\Phi_h,\Psi_h)}{||\vec{\Phi}||_2||\vec{\Psi}||_2} \\ &\leq M \sup_{\Phi_h,\Psi_h\in X_h\setminus\{0\}} \frac{||\Phi_h||_{X_{\varepsilon,h}}||\Psi_h||_{X_{\varepsilon,h}}}{||\vec{\Phi}||_2||\vec{\Psi}||_2} \leq M\max(\nu_u^2, \nu_*^2 + \nu_q^2) \end{aligned}$$

where M is the (ε-independent) continuity constant of C_ε. Similarly, using the inf-sup property of this bilinear form we know that for all $\Phi_h \in X_h$ there exists $\Psi_h \in X_h$ such that

$$\begin{aligned} \beta'\min\{\mu_u^2, \mu_*^2\}||\vec{\Phi}||_2||\vec{\Psi}||_2 &\leq \beta||\Phi_h||_{X_{\varepsilon,h}}||\Psi_h||_{X_{\varepsilon,h}} \leq C_\varepsilon(\Phi_h,\Psi_h) \\ &= \vec{\Psi}\cdot\mathbb{A}^\varepsilon\vec{\Phi} \leq ||\vec{\Psi}||_2||\mathbb{A}^\varepsilon\vec{\Phi}||_2\,, \end{aligned}$$

with an ε-independent constant $\beta > \beta' > 0$. This simplifies to

$$\beta'\min\{\mu_u^2, \mu_*^2\}||\vec{\Phi}||_2 \leq ||\mathbb{A}^\varepsilon\vec{\Phi}||_2\,,$$

or equivalently

$$\beta' \min\{\mu_u^2, \mu_*^2\} ||(\mathbb{A}^\varepsilon)^{-1}\vec{\Phi}||_2 \leq ||\vec{\Phi}||_2 \,, \quad \forall\, \vec{\Phi} \in \mathbb{R}^N \,.$$

Thus, the condition number can be estimated as

$$cond_2(\mathbb{A}^\varepsilon) = ||\mathbb{A}^\varepsilon||_2 ||(\mathbb{A}^\varepsilon)^{-1}||_2 \leq \frac{M \max(\nu_u^2, \nu_*^2 + \nu_q^2)}{\beta' \min\{\mu_u^2, \mu_*^2\}} \,, \tag{31}$$

which is an ε-independent bound. ■

Remark 7 *Let us try to be more quantitative in our estimate of $cond_2(\mathbb{A}^\varepsilon)$. In what follows, the symbols $\lesssim$ and $\sim$ will hide the constants of order 1, independent of the mesh. Consider the standard finite element setting: the bases of V_h and L_h are formed by the hat finite element functions on a quasi-uniform mesh. We know in this case that $||u_h||_{L^2}^2 \sim h^2 ||\vec{u}||_2^2$ and $|u_h|_{\mathcal{V}} \leq C_I h^{-1} ||u_h||_{L^2}$ by the inverse inequality with a constant $C_I > 0$ that depends only on the mesh regularity [7]. We also recall the Poincaré inequality $||u_h||_{L^2} \leq C_P |u_h|_{\mathcal{V}}$. The same holds for q_h and leads to*

$$\mu_u \sim \mu_q \sim h \text{ and } \nu_u \sim \nu_q \sim 1 \,.$$

We also have $|q_h|_{*h} \leq |q_h|_{||}$, hence $\nu_* \leq \nu_q$. Moreover, for any $q_h \in L_h$ we prove, using the inverse and Poincaré inequalities, that

$$\begin{aligned}
|q_h|_{*h} &\geq \frac{a_{\|}(q_h, q_h)}{|q_h|_{\mathcal{V}}} = \frac{|q_h|_{||}^2}{\left(|q_h|_{\perp}^2 + |q_h|_{||}^2\right)^{1/2}} \geq \frac{|q_h|_{||}^2}{\left(C_I^2 h^{-2} ||q_h||_{L^2}^2 + |q_h|_{||}^2\right)^{1/2}} \\
&= \frac{C_P^2 C_I^2 h^{-2} |q_h|_{||}^2 + |q_h|_{||}^2}{(C_P^2 C_I^2 h^{-2} + 1)\left(C_I^2 h^{-2} ||q_h||_{L^2}^2 + |q_h|_{||}^2\right)^{1/2}} \geq \frac{\left(C_I^2 h^{-2} ||q_h||_{L^2}^2 + |q_h|_{||}^2\right)^{1/2}}{C_P^2 C_I^2 h^{-2} + 1} \\
&\geq \frac{\left(C_I^2 h^{-2} + C_P^2\right)^{1/2}}{C_P^2 C_I^2 h^{-2} + 1} ||q_h||_{L^2} \sim h ||q_h||_{L^2} \sim h^2 ||\vec{q}||_2 \,.
\end{aligned}$$

This implies $\mu_* \gtrsim h^2$, so that (31) becomes finally

$$cond_2(\mathbb{A}^\varepsilon) \lesssim \frac{1}{h^4} \,.$$

Theorem 8 ***(h-Convergence)*** *Let $k \geq 1$ and $V_h \subset \mathcal{V}$ be the $\mathbb{P}_k$ (or $\mathbb{Q}_k$) finite element space on a regular mesh $\mathcal{T}_h$. Suppose moreover that problem (21) has a solution $(u^\varepsilon, q^\varepsilon) \in \mathcal{X}_\varepsilon$, having the regularity $u^\varepsilon \in H^{k+1}(\Omega)$, $q^\varepsilon \in H^{k+1}(\Omega)$. Then one has the estimate*

$$|u^\varepsilon - u_h^\varepsilon|_{\mathcal{V}} \leq c\, h^k (|u^\varepsilon|_{H^{k+1}} + |q^\varepsilon|_{H^{k+1}}) \,, \tag{32}$$

with a constant $c > 0$ that depends neither on the mesh, nor on ε.

Proof. Let $\hat{u}_h^\varepsilon \in V_h$ and $\hat{q}_h^\varepsilon \in L_h$ be the standard nodal interpolant of u^ε and q^ε satisfying [7]

$$|u^\varepsilon - \hat{u}_h^\varepsilon|_{H^1} \leq c\, h^k |u^\varepsilon|_{H^{k+1}} \text{ and } |q^\varepsilon - \hat{q}_h^\varepsilon|_{H^1} \leq c\, h^k |q^\varepsilon|_{H^{k+1}}\,.$$

We can now derive the error estimates in the H^1-norm for u^ε in the way similar to Cea's lemma: by the inf-sup property, there exists $(v_h, w_h) \in X_h$ with $\|v_h, w_h\|_{X_{\varepsilon,h}} = 1$ such that (with some $0 < \beta' < \beta$)

$$\begin{aligned}
|u_h^\varepsilon - \hat{u}_h^\varepsilon|_{H^1} &\leq \|u_h^\varepsilon - \hat{u}_h^\varepsilon, q_h^\varepsilon - \hat{q}_h^\varepsilon\|_{X_{\varepsilon,h}} \leq \frac{1}{\beta'} C_\varepsilon((u_h^\varepsilon - \hat{u}_h^\varepsilon, q_h^\varepsilon - \hat{q}_h^\varepsilon), (v_h, w_h)) \quad (33)\\
&= \frac{1}{\beta'} C_\varepsilon((u^\varepsilon - \hat{u}_h^\varepsilon, q^\varepsilon - \hat{q}_h^\varepsilon), (v_h, w_h))\\
&\leq c\,(|u^\varepsilon - \hat{u}_h^\varepsilon|_{\mathcal{V}}^2 + |q^\varepsilon - \hat{q}_h^\varepsilon|_{*h}^2 + \varepsilon |q^\varepsilon - \hat{q}_h^\varepsilon|_{\|}^2)^{1/2}\\
&\leq c\,(|u^\varepsilon - \hat{u}_h^\varepsilon|_{H^1} + |q^\varepsilon - \hat{q}_h^\varepsilon|_{H^1})\,,
\end{aligned}$$

since

$$|q^\varepsilon - \hat{q}_h^\varepsilon|_{*h} = \sup_{v_h \in V_h} \frac{a_\|(q^\varepsilon - \hat{q}_h^\varepsilon, v_h)}{|v_h|_{\mathcal{V}}} \leq |q^\varepsilon - \hat{q}_h^\varepsilon|_{\|} \leq |q^\varepsilon - \hat{q}_h^\varepsilon|_{H^1}.$$

We can now employ the interpolation error estimates to conclude. ■

Remark 9 *The error estimate (32) would be of course useless if the norms $|u^\varepsilon|_{H^{k+1}}$, $|q^\varepsilon|_{H^{k+1}}$ were $\varepsilon-$dependent and exploding in the limit $\varepsilon \to 0$. Fortunately, it is not the case. We expect indeed that $|u^\varepsilon|_{H^{k+1}}$ is bounded uniformly in ε by the norm of f in $H^{k-1}(\Omega)$ and $|q^\varepsilon|_{H^{k+1}}$ is bounded uniformly in ε by the norm of f in $H^{k+1}(\Omega)$. This can be easily proved in the case of a simple aligned geometry, see Appendix A. We conjecture that this remains true also in a general setting.*

Remark 10 *If we do not omit the norm of $q_h^\varepsilon - \hat{q}_h^\varepsilon$ in the left-hand side of the first inequality in (33), we also get an error estimate for q_h^ε*

$$|q^\varepsilon - q_h^\varepsilon|_{||} \leq c \frac{h^k}{\sqrt{\varepsilon}} (|u^\varepsilon|_{H^{k+1}} + |q^\varepsilon|_{H^{k+1}})\,,$$

which degenerates as ε goes to 0. We are not sure, if this estimate is sharp, but we recall that q^ε is an auxiliary variable, without any intrinsic meaning.

3 Second AP-reformulation for general field lines

The fundamental idea of the AP-reformulation introduced in Section 2 is the introduction of a Lagrange multiplier $q^\varepsilon \in \mathcal{L}_{in}$ in order to handle well with the constraint $\nabla_\| u^0 = 0$ in the limit $\varepsilon \to 0$. This Lagrange multiplier was uniquely determined up to a constant on the field lines, which was fixed by imposing $q^\varepsilon_{|\Gamma_{in}} = 0$. The disadvantage of this scheme is that it requires to identify the inflow part of the boundary, which can be cumbersome in practice or even not possible if some of the field lines

are closed and lie completely inside the domain Ω. It is thus tempting to abandon the zero inflow boundary condition and to search for the auxiliary q^ε-variable in the Hilbert space

$$\mathcal{L} = \{\xi \in L^2(\Omega) \ / \ \nabla_{||}\xi \in L^2(\Omega)\}\,, \quad (u,v)_{\mathcal{L}} := (u,w) + (\nabla_{||}u, \nabla_{||}w)\,. \tag{34}$$

The problem with this idea is that we loose now uniqueness of the solution if we attempt to implement the AP-reformulation (15) just changing $\mathcal{L}_{in}$ to $\mathcal{L}$. To circumvent this difficulty, it was proposed in [12] to introduce a stabilization term into the AP reformulation so that it becomes: Find $(u^{\varepsilon,\sigma}, \xi^{\varepsilon,\sigma}) \in \mathcal{V} \times \mathcal{L}$ such that

$$(AP_{\mathcal{S}})^{\varepsilon,\sigma} \quad \begin{cases} a(u^{\varepsilon,\sigma}, v) + (1-\varepsilon)a_{\|}(\xi^{\varepsilon,\sigma}, v) = (f,v), & \forall v \in \mathcal{V} \\ a_{\|}(u^{\varepsilon,\sigma}, w) - \varepsilon a_{\|}(\xi^{\varepsilon,\sigma}, w) - \sigma(\xi^{\varepsilon,\sigma}, w) = 0, & \forall w \in \mathcal{L}\,, \end{cases} \tag{35}$$

where $\sigma > 0$ is a small stabilization parameter, chosen consistently with the overall discretization error. It is this term which permits to have the uniqueness, as will be shown in Lemma 13.

In the limit $\varepsilon \to 0$ this system yields: Given $\sigma > 0$, find $(u^{0,\sigma}, \xi^{0,\sigma}) \in \mathcal{V} \times \tilde{\mathcal{L}}^2$, solution to

$$(L_{\mathcal{S}})^{\sigma} \quad \begin{cases} a(u^{0,\sigma}, v) + a_{\|}(\xi^{0,\sigma}, v) = (f,v), & \forall v \in \mathcal{V} \\ a_{\|}(u^{0,\sigma}, w) - \sigma(\xi^{0,\sigma}, w) = 0, & \forall w \in \tilde{\mathcal{L}}^2\,, \end{cases} \tag{36}$$

where $\tilde{\mathcal{L}}^2$ is, loosely speaking, the closure of $\mathcal{L}$ in the $|\cdot|_*$ semi-norm (17) intersected with $L^2(\Omega)$, i.e.

$$\tilde{\mathcal{L}}^2 = \left\{ \xi \in L^2(\Omega) \,/\, \sup_{v \in \mathcal{V}} \frac{a_{\|}(\xi, v)}{|v|_{\mathcal{V}}} < \infty \right\}.$$

This space is a Hilbert-space associated with the scalar product

$$(u,w)_{\tilde{\mathcal{L}}^2} := (u,w) + (u^*, w^*)_{L^2}\,, \quad \forall (u,w) \in \tilde{\mathcal{L}}^2\,,$$

where u^* resp. w^* are the unique solutions of (18) corresponding to u resp. w. We need this special space, first of all, to be able to treat the limit-problem $(L_{\mathcal{S}})^\sigma$ with the inf-sup theory, similar to the inflow-case, and also in order to be able to define the stabilization term $\sigma(\xi^{0,\sigma}, w)$.

Remark 11 *Remark also that we have $\tilde{\mathcal{L}}^2 \neq \mathcal{L}$. Let us prove it in the following simple setting: let $\Omega = (0,\pi) \times (0,\pi)$, $A_{\|} = 1$, $A_{\perp} = Id$, $b = e_2$. For any $q = \sum_{k,l=1}^{\infty} q_{kl} \sin kx \cos ly$, the calculation as in Remark 3 gives*

$$|q|_*^2 = \sum_{k,l=1}^{\infty} \frac{l^4}{k^2 + l^2} |q_{kl}|^2\,,$$

so that taking q_{kl} such that $q_{kl} = \frac{1}{l}$ if $k = l^2$ and $q_{kl} = 0$ for any $k \neq l^2$ we have

$$|q|_*^2 = \sum_{l=1}^{\infty} \frac{l^2}{l^4 + l^2} < \infty\,,$$

so that $q \in \tilde{\mathcal{L}}$. *Moreover, clearly* $q \in L^2(\Omega)$. *However,*

$$|q|_\parallel^2 = \sum_{k,l=1}^{\infty} l^2 |q_{kl}|^2 = \infty \,.$$

3.1 Mathematical analysis on the continuous level

To analyze the well-posedness of problem (35) ans its asymptotic limit behaviour for $\varepsilon \to 0$, we shall rewrite it in an equivalent manner, better suited for mathematical studies. For this, we observe first that the second equation in (35) gives for all $\varepsilon, \sigma > 0$

$$(\xi^{\varepsilon,\sigma}, w) = 0, \quad \forall w \in \mathcal{G}_{\mathcal{L}} \,, \quad \text{with} \quad \mathcal{G}_{\mathcal{L}} := \{v \in \mathcal{L} \,|\, \nabla_\parallel v = 0\} \,,$$

which means that $\xi^{\varepsilon,\sigma}$ belongs to the following space

$$\mathcal{A} = \{\xi \in \mathcal{L} \;/\; (\xi, w) = 0 \quad \forall w \in \mathcal{G}_{\mathcal{L}}\} \,, \tag{37}$$

which consists thus of functions from $\mathcal{L}$ with zero (weighted) average along the field lines. Remark that one has the decomposition $\mathcal{L} = \mathcal{G}_{\mathcal{L}} \oplus^\perp \mathcal{A}$. Problem (35) can be hence rewritten as: Find $(u^{\varepsilon,\sigma}, \xi^{\varepsilon,\sigma}) \in \mathcal{V} \times \mathcal{A}$ such that

$$(AP'_{\mathcal{S}})^{\varepsilon,\sigma} \quad \begin{cases} a(u^{\varepsilon,\sigma}, v) + (1-\varepsilon) a_\parallel(\xi^{\varepsilon,\sigma}, v) = (f, v), & \forall v \in \mathcal{V} \\ a_\parallel(u^{\varepsilon,\sigma}, w) - \varepsilon a_\parallel(\xi^{\varepsilon,\sigma}, w) - \sigma(\xi^{\varepsilon,\sigma}, w) = 0, & \forall w \in \mathcal{A} \,. \end{cases} \tag{38}$$

We emphasize that this reformulation is completely equivalent to (35) for all $\varepsilon > 0$ and $\sigma > 0$ and is done solely for the purposes of mathematical analysis. The formulation used for the numerical discretization will be (35).

Note that $\mathcal{A}$ becomes a Hilbert space when equipped with the scalar product $a_\parallel(\cdot,\cdot)$ and corresponding norm $|\cdot|_\parallel$. Indeed, if $\xi \in \mathcal{A}$ and $|\xi|_\parallel = 0$ then $\xi \in \mathcal{G}_{\mathcal{L}}$ which implies $\xi = 0$ since $\mathcal{A}$ is orthogonal to $\mathcal{G}_{\mathcal{L}}$. For the same reasons, the semi-norm $|\cdot|_*$ (17) is actually a norm when applied to $\mathcal{A}$. We can thus introduce the closure $\tilde{\mathcal{A}}$ of $\mathcal{A}$ with respect to $|\cdot|_*$, needed as usual, for the $\varepsilon \to 0$ limit model. The $(L_{\mathcal{S}})^\sigma$-problem will be shown to be equivalent to: Find $(u^{0,\sigma}, \xi^{0,\sigma}) \in \mathcal{V} \times (\tilde{\mathcal{A}} \cap L^2(\Omega))$, solution to

$$(L'_{\mathcal{S}})^\sigma \quad \begin{cases} a(u^{0,\sigma}, v) + a_\parallel(\xi^{0,\sigma}, v) = (f, v), & \forall v \in \mathcal{V} \\ a_\parallel(u^{0,\sigma}, w) - \sigma(\xi^{0,\sigma}, w) = 0, & \forall w \in \tilde{\mathcal{A}} \cap L^2(\Omega) \,. \end{cases} \tag{39}$$

As mentioned earlier, formulations (38) and (39) are better adapted for the mathematical study, then the completely equivalent ones (35) and (36). In particular in the limit $\sigma \to 0$, they permit to get the following problems: Find $(u^\varepsilon, \xi^\varepsilon) \in \mathcal{V} \times \mathcal{A}$ solution of

$$(AP_{\mathcal{A}})^\varepsilon \quad \begin{cases} a(u^\varepsilon, v) + (1-\varepsilon) a_\parallel(\xi^\varepsilon, v) = (f, v), & \forall v \in \mathcal{V} \\ a_\parallel(u^\varepsilon, w) - \varepsilon a_\parallel(\xi^\varepsilon, w) = 0, & \forall w \in \mathcal{A} \,. \end{cases} \tag{40}$$

which is equivalent to the original problem (10) and hence also to the inflow AP-reformulation (15). In the present case, we fix the Lagrangian variable ξ^ε by requiring zero average along the field lines, *i.e.* $\xi^\varepsilon \in \mathcal{A}$, in the former case we fixed the

corresponding Lagrangian variable q^ε by setting q^ε zero on the inflow boundary Γ_{in}, *i.e.* $q^\varepsilon \in \mathcal{L}_{in}$. Note that we do not want here to discretize the space $\mathcal{A}$ directly. This space arises only in the limit $\sigma \to 0$, which is never performed in practice when one implements the scheme of this paper. On the contrary, the scheme from [4] relies on a direct discretization of $\mathcal{A}$ which results in a rather complicated numerical method. Remark also that we abandoned in (40) the requirement that the ξ-variable has to belong to $L^2(\Omega)$, as there is no more need, for $\sigma = 0$.

Letting now formally $\varepsilon \to 0$ in (40), we obtain the problem: Find $(u^0, \xi^0) \in \mathcal{V} \times \tilde{\mathcal{A}}$ such that

$$(L_{\mathcal{A}}) \quad \begin{cases} a(u^0, v) + a_{\|}(\xi^0, v) = (f, v), & \forall v \in \mathcal{V} \\ a_{\|}(u^0, w) = 0, & \forall w \in \tilde{\mathcal{A}}\,, \end{cases} \tag{41}$$

which is an equivalent (saddle-point) reformulation of the original limit problem (12).

For the reader convenience, we draw in Figure 2 a scheme, with all the problems we introduced so far, and their relations. In the following Lemmata and Theorems, we shall prove some of these relations and convergences, adapting the results from the previous section 2.2 to the present case containing two parameters, ε and σ.

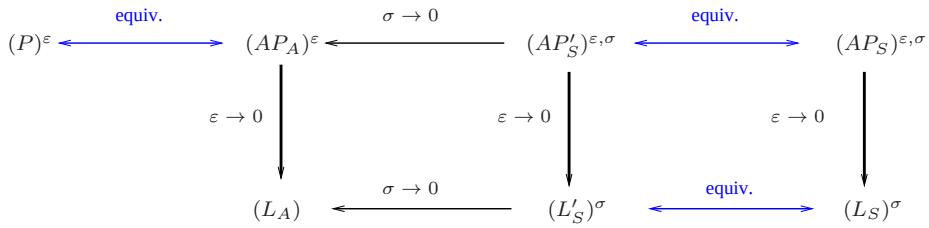


Figure 2: Stabilized reformulations of the original problem $(P)^\varepsilon$.

Lemma 12 ***(Inf-Sup condition)*** *Let V, L, $\tilde{L}$, $\hat{L}$ be Hilbert spaces such that $L \subset \tilde{L}$ and $L \subset \hat{L}$ with continuous inclusions and $\|\xi\|_{\tilde{L}} \leq \|\xi\|_L$ for all $\xi \in L$. Let $a(\cdot,\cdot)$, $c(\cdot,\cdot)$, $d(\cdot,\cdot)$ denote the scalar products on respectively V, $\tilde{L}$, $\hat{L}$ and $b(\cdot,\cdot) : \tilde{L} \times V \to \mathbb{R}$ be a bilinear form satisfying $\|b(\xi, v)\| \leq \|v\|_V \|\xi\|_{\tilde{L}}$ for all $v \in V$, $\xi \in L$ as well as the inf-sup condition*

$$\inf_{\xi \in \tilde{L}} \sup_{v \in V} \frac{b(\xi, v)}{\|\xi\|_{\tilde{L}} \|v\|_V} = \alpha > 0. \tag{42}$$

Define furthermore the Hilbert space $X_{\varepsilon,\sigma}$ for $\varepsilon \geq 0$, $\sigma \geq 0$ by

$$X_{\varepsilon,\sigma} := \begin{cases} V \times L, \text{ if } \varepsilon > 0, \sigma \geq 0 \\ V \times (\tilde{L} \cap \hat{L}), \text{ if } \varepsilon = 0, \sigma > 0 \\ V \times \tilde{L}, \text{ if } \varepsilon = 0, \sigma = 0 \end{cases},$$

and equip it with the norm $\|u,\xi\|_{X_{\varepsilon,\sigma}} := (\|u\|_V^2 + \|\xi\|_{\tilde{L}}^2 + \varepsilon\|\xi\|_L^2 + \sigma\|\xi\|_{\hat{L}}^2)^{1/2}$.

For any $\varepsilon \geq 0$ *and* $\sigma \geq 0$ *let* $C_{\varepsilon,\sigma} : X_{\varepsilon,\sigma} \times X_{\varepsilon,\sigma} \to \mathbb{R}$ *be the bilinear form defined by*

$$C_{\varepsilon,\sigma}((u,\xi),(v,w)) = a(u,v) + (1-\varepsilon)b(\xi,v) + b(w,u) - \varepsilon c(\xi,w) - \sigma d(\xi,w)\,.$$

Then $C_{\varepsilon,\sigma}$ *is continuous and satisfies the inf-sup condition*

$$\inf_{(u,\xi)\in X_{\varepsilon,\sigma}} \sup_{(v,w)\in X_{\varepsilon,\sigma}} \frac{C_{\varepsilon,\sigma}((u,\xi),(v,w))}{\|u,\xi\|_{X_{\varepsilon,\sigma}}\|v,w\|_{X_{\varepsilon,\sigma}}} \geq \beta\,, \tag{43}$$

with a constant $\beta > 0$ *that depends only on* α.

Proof. The proof of this lemma follows the same lines as that of Lemma 4 and we give here only a short version of it. For any $(u,\xi) \in X_{\varepsilon,\sigma}$, denoting

$$Z := \sup_{(v,w)\in X_{\varepsilon,\sigma}} \frac{C_{\varepsilon,\sigma}((u,\xi),(v,w))}{\|v,w\|_{X_{\varepsilon,\sigma}}}\,,$$

we can prove that $(1-\varepsilon)\alpha\|\xi\|_{\tilde{L}} \leq Z + \|u\|_V$. Now, taking $v = u$, $w = -\xi$ we observe that

$$\begin{aligned} C_{\varepsilon,\sigma}((u,\xi),(u,-\xi)) &= a(u,u) - \varepsilon b(\xi,u) + \varepsilon c(\xi,\xi) + \sigma d(\xi,\xi) \\ &\geq (1-\frac{\varepsilon}{2})\|u\|_V^2 + \frac{\varepsilon}{2}\|\xi\|_L^2 + \sigma\|\xi\|_{\hat{L}}^2\,, \end{aligned}$$

implying altogether

$$\frac{1}{2}\|u\|_V^2 + \frac{\varepsilon}{2}\|\xi\|_L^2 + \frac{\alpha^2(1-\varepsilon)^2}{8}\|\xi\|_{\tilde{L}}^2 + \sigma\|\xi\|_{\hat{L}}^2 \leq Z\|u,-\xi\|_{X_{\varepsilon,\sigma}} + \frac{1}{4}Z^2 + \frac{1}{4}\|u\|_V^2\,.$$

Following again the inequalities from the proof of Lemma 4, we see that there exists a constant $c_0 \in (0,\frac{1}{4}]$ depending only on α such that for any $\gamma > 0$ and $\varepsilon \in [0,1]$

$$c_0\|u,\xi\|_{X_{\varepsilon,\sigma}}^2 \leq \frac{1}{2\gamma}\|u,\xi\|_{X_{\varepsilon,\sigma}}^2 + \frac{1+\gamma}{2}Z^2\,.$$

Taking finally a sufficiently big γ yields $\|u,\xi\|_{X_\varepsilon} \leq (1/\beta)Z$ with a constant $\beta > 0$ depending only on α. ■

Lemma 13 ***(Existence/Uniqueness for*** $\varepsilon \geq 0$ ***and*** $\sigma > 0$***)*** *Let hypothesis A be satisfied. The stabilized AP-problem* $(AP_{\mathcal{S}})^{\varepsilon,\sigma}$ *(resp.* $(L_{\mathcal{S}})^{\sigma}$*) is well-posed for each* $\varepsilon \in (0,1]$ *and* $\sigma > 0$ *(resp.* $\varepsilon = 0, \sigma > 0$*), i.e. for any* $f \in \mathcal{V}'$ *there exists a unique solution* $(u^{\varepsilon,\sigma},\xi^{\varepsilon,\sigma}) \in \mathcal{V}\times\mathcal{L}$ *(resp.* $(u^{0,\sigma},\xi^{0,\sigma}) \in \mathcal{V}\times\tilde{\mathcal{L}}^2$*), which satisfies*

$$\|u^{\varepsilon,\sigma},\xi^{\varepsilon,\sigma}\|_{\mathcal{X}_{\varepsilon,\sigma}} \leq C\|f\|_{\mathcal{V}'}\,, \tag{44}$$

with $\|u,\xi\|_{\mathcal{X}_{\varepsilon,\sigma}} := (|u|_{\mathcal{V}}^2 + |\xi|_*^2 + \varepsilon|\xi|_\|^2 + \sigma\|\xi\|_{L^2}^2)^{1/2}$ *and some* $C > 0$ *independent on* ε *and* σ*. Moreover we have the* ε*-convergence*

$$||u^{\varepsilon,\sigma} - u^{0,\sigma}, \xi^{\varepsilon,\sigma} - \xi^{0,\sigma}||_{\mathcal{X}_{\varepsilon,\sigma}} \to 0 \quad \textit{for} \quad \varepsilon \to 0\,.$$

Proof. The existence and uniqueness of the solution to the reformulated problems $(AP'_{\mathcal{S}})^{\varepsilon,\sigma}$ and $(L'_{\mathcal{S}})^{\sigma}$ follows directly from Lemma 12 by setting $V = \mathcal{V}$, $L = \mathcal{A}$, $\tilde{L} = \tilde{\mathcal{A}}$, $\hat{L} = L^2(\Omega)$. Now, the equivalence of $(AP_{\mathcal{S}})^{\varepsilon,\sigma}$ and $(AP'_{\mathcal{S}})^{\varepsilon,\sigma}$ is easily seen from the decomposition $\mathcal{L} = \mathcal{G}_{\mathcal{L}} \oplus^{\perp} \mathcal{A}$. Similarly, the equivalence of $(L_{\mathcal{S}})^{\sigma}$ and $(L'_{\mathcal{S}})^{\sigma}$ can be derived from the decomposition $\tilde{\mathcal{L}} = \mathcal{G}_{\mathcal{L}} \oplus^{\perp} \widetilde{\mathcal{A}}$. ■

Theorem 14 ***(Existence/Uniqueness for*** $\varepsilon \geq 0$ ***and*** $\sigma = 0$***)*** *Let hypothesis A be satisfied. The* $(AP_{\mathcal{A}})^{\varepsilon}$*-problem (40) (resp.* $(L_{\mathcal{A}})$*-problem (41)) is well-posed for each* $\varepsilon \in (0, 1]$ *(resp.* $\varepsilon = 0$*), i.e. for any* $f \in \mathcal{V}'$ *and any* $\varepsilon \in [0, 1]$ *there exists a unique solution* $(u^{\varepsilon}, \xi^{\varepsilon}) \in \mathcal{X}_{\varepsilon,0}$*, which satisfies*

$$\|u^{\varepsilon}, \xi^{\varepsilon}\|_{\mathcal{X}_{\varepsilon,0}} \leq C \|f\|_{\mathcal{V}'} \,,$$

with some $C > 0$ *independent on* ε*. Furthermore, one has the* ε*-convergence*

$$\|u^{\varepsilon} - u^0, \xi^{\varepsilon} - \xi^0\|_{\mathcal{X}_{\varepsilon,0}} \to 0 \,, \quad \textit{for} \quad \varepsilon \to 0 \,.$$

If we suppose more regular data, as $f \in L^2(\Omega)$*, then one has even* $\xi^0 \in \mathcal{A}$ *and the estimates*

$$|u^{\varepsilon} - u^0|_{\mathcal{V}} \leq C\sqrt{\varepsilon} \,, \quad |\xi^{\varepsilon} - \xi^0|_* \leq C\sqrt{\varepsilon} \,,$$

with $C > 0$ *some* ε*-independent constant.*

Proof. The existence and uniqueness of a solution $(u^{\varepsilon}, \xi^{\varepsilon}) \in \mathcal{V} \times \mathcal{A}$ to $(AP_{\mathcal{A}})^{\varepsilon}$ resp. $(u^0, \xi^0) \in \mathcal{V} \times \tilde{\mathcal{A}}$ to $(L_{\mathcal{A}})$ is easily established using Lemma 4. The statements about the convergence as $\varepsilon \to 0$ follow in the same way as in the proof of Theorem 5. ■

Theorem 15 ***(*** σ***-Convergence)*** *Let hypothesis A be satisfied and moreover,* $(u^{\varepsilon,\sigma}, \xi^{\varepsilon,\sigma}) \in \mathcal{V} \times \mathcal{A}$ *be solution to* $(AP_{\mathcal{S}})^{\varepsilon,\sigma}$ *and* $(u^{\varepsilon}, \xi^{\varepsilon}) \in \mathcal{V} \times \mathcal{A}$ *solution of* $(AP_{\mathcal{A}})^{\varepsilon}$*, with* $\varepsilon > 0$*. Suppose that* $\xi^{\varepsilon} \in H^1(\Omega)$*. Then*

$$\|u^{\varepsilon,\sigma} - u^{\varepsilon}, \xi^{\varepsilon,\sigma} - \xi^{\varepsilon}\|_{\mathcal{X}_{\varepsilon,\sigma}} \leq c\sigma |\xi^{\varepsilon}|_{H^1} \,, \tag{45}$$

with a constant $c > 0$ *independent of* σ *and* ε*.*

Proof. We turn now to the convergence as $\sigma \to 0$. Using the combined bilinear form $C_{\varepsilon,\sigma}$ and recalling the problems $(AP'_{\mathcal{S}})^{\varepsilon,\sigma}$ resp. $(AP_{\mathcal{A}})^{\varepsilon}$, we can write

$$C_{\varepsilon,\sigma}((u^{\varepsilon,\sigma}, \xi^{\varepsilon,\sigma}), (v, w)) = (f, v) \,, \quad \forall\, (v, w) \in \mathcal{V} \times \mathcal{A} \,,$$

$$C_{\varepsilon,0}((u^{\varepsilon}, \xi^{\varepsilon}), (v, w)) = (f, v) \,, \quad \forall\, (v, w) \in \mathcal{V} \times \mathcal{A} \,,$$

where

$$C_{\varepsilon,\sigma}((u, \xi), (v, w)) := a(u, v) + (1 - \varepsilon) a_{\|}(\xi, v) + a_{\|}(u, w) - \varepsilon a_{\|}(\xi, w) - \sigma(\xi, w) \,.$$

Note that $C_{\varepsilon,0}$ coincides with C_{ε} as defined by (22). Taking the difference gives

$$C_{\varepsilon,\sigma}((u^{\varepsilon,\sigma} - u^{\varepsilon}, \xi^{\varepsilon,\sigma} - \xi^{\varepsilon}), (v, w)) = \sigma(\xi^{\varepsilon}, w) \leq \sigma |\xi^{\varepsilon}|_{\mathcal{V}} |w|_{\mathcal{V}'} \leq c\sigma |\xi^{\varepsilon}|_{\mathcal{V}} |w|_* \,.$$

We have used here the bound $|w|_{\mathcal{V}'} \leq c|w|_*$ valid for $w \in \mathcal{A}$ as proved below (Corollary 19).

Now, remind that the form $C_{\varepsilon,\sigma}$ enjoys the inf-sup property

$$\inf_{(u,\xi)\in\mathcal{X}_{\varepsilon,\sigma}} \sup_{(v,w)\in\mathcal{X}_{\varepsilon,\sigma}} \frac{C_{\varepsilon,\sigma}((u,\xi),(v,w))}{\|u,\xi\|_{\mathcal{X}_{\varepsilon,\sigma}}\|v,w\|_{\mathcal{X}_{\varepsilon,\sigma}}} \geq \beta\,,$$

where $\|u,\xi\|_{\mathcal{X}_{\varepsilon,\sigma}} = (|u|^2_{\mathcal{V}} + |\xi|^2_* + \varepsilon|\xi|^2_{\|} + \sigma\|\xi\|^2_{L^2})^{1/2}$, so that $|w|_* \leq \|v,w\|_{\mathcal{X}_{\varepsilon,\sigma}}$. We can thus conclude that there exists $(v,w) \in \mathcal{X}_{\varepsilon,\sigma}$ such that $\|v,w\|_{\mathcal{X}_{\varepsilon,\sigma}} = 1$ and

$$\beta'\|u^{\varepsilon,\sigma}-u^\varepsilon, \xi^{\varepsilon,\sigma}-\xi^\varepsilon\|_{\mathcal{X}_{\varepsilon,\sigma}} \leq C_{\varepsilon,\sigma}((u^{\varepsilon,\sigma}-u^\varepsilon, \xi^{\varepsilon,\sigma}-\xi^\varepsilon),(v,w)) \leq c\,\sigma|\xi^\varepsilon|_{\mathcal{V}}\|v,w\|_{\mathcal{X}_{\varepsilon,\sigma}} = c\,\sigma|\xi^\varepsilon|_{H^1}\,,$$

with some $0 < \beta' < \beta$, for example $\beta' = \beta/2$. This concludes the proof. ■

Remark 16 *Without the additional hypothesis $\xi^\varepsilon \in H^1(\Omega)$, we can easily prove a sub-optimal estimate*

$$\|u^{\varepsilon,\sigma} - u^\varepsilon, \xi^{\varepsilon,\sigma} - \xi^\varepsilon\|_{\mathcal{X}_{\varepsilon,\sigma}} \leq c\sqrt{\sigma}\|\xi^\varepsilon\|_{L^2}\,.$$

Indeed,

$$C_{\varepsilon,\sigma}((u^{\varepsilon,\sigma}-u^\varepsilon, \xi^{\varepsilon,\sigma}-\xi^\varepsilon),(v,w)) = \sigma(\xi^\varepsilon, w) \leq \sigma\|\xi^\varepsilon\|_{L^2(\Omega)}\|w\|_{L^2(\Omega)} \leq c\sqrt{\sigma}\|\xi^\varepsilon\|_{L^2(\Omega)}\|v,w\|_{\mathcal{X}_{\varepsilon,\sigma}}\,.$$

Remark 17 *The conclusions of Theorem 14 remain true (after an obvious rephrasing) in the limit case $\varepsilon = 0$ since the proof relies on the estimates in the norm of $\mathcal{X}_{\varepsilon,\sigma}$ which remains a valid norm in the limit $\varepsilon \to 0$.*

It remains to prove the bound $|w|_{\mathcal{V}'} \leq c\,|w|_*$ valid for $w \in \mathcal{A}$. This will be done using the following result:

Lemma 18 *Let $u \in \mathcal{V}$ and consider $v \in \mathcal{A}$ being the unique solution to*

$$a_{\|}(v,w) = (u,w), \quad \forall w \in \mathcal{A}. \tag{46}$$

Then $v \in H^1(\Omega)$ and there exists a constant $c > 0$ such that $\|v\|_{H^1} \leq c\,|u|_{\mathcal{V}}$.

Proof. To simplify the notations, let us restrict ourselves to the 2D case in this proof (the extension to $d > 2$ is rather straightforward). There is an evident bound $||\nabla_{\|}v||_{L^2} \leq c||u||_{L^2}$ which implies $||v||_{L^2} \leq C||u||_{L^2}$ by a Poincaré type inequality [4]. To continue, let us change the coordinates on Ω and suppose that there exist new coordinates (ξ_1,ξ_2) so that Ω becomes the unit square $\Omega_\xi = (0,1)^2$ and $\nabla_{||}$ becomes $\alpha(\xi_1,\xi_2)\frac{\partial}{\partial\xi_2}$ with some positive function α. Problem (46) is written in these new coordinates as

$$\int_{\Omega_\xi} N\,\frac{\partial v}{\partial\xi_2}\frac{\partial w}{\partial\xi_2}\,d\xi_1 d\xi_2 = \int_{\Omega_\xi} J\,uw\,d\xi_1 d\xi_2\,,$$

where $J = J(\xi_1,\xi_2)$ is the Jacobian and $N = N(\xi_1,\xi_2) = A_{||}\,J\,\alpha^2$, which are positive functions given by the geometry.

Let us now replace here w by $\frac{\partial \omega}{\partial \xi_1}$ with arbitrary and sufficiently smooth function ω such that $\omega = 0$ at $\xi_1 = 0$ and at $\xi_1 = 1$. Integration by parts with respect to ξ_1 yields then

$$\int_{\Omega_\xi} \frac{\partial N}{\partial \xi_1} \frac{\partial v}{\partial \xi_2} \frac{\partial \omega}{\partial \xi_2} \, d\xi_1 d\xi_2 + \int_{\Omega_\xi} N \frac{\partial^2 v}{\partial \xi_1 \partial \xi_2} \frac{\partial \omega}{\partial \xi_2} \, d\xi_1 d\xi_2 = \int_{\Omega_\xi} \frac{\partial (Ju)}{\partial \xi_1} \, \omega \, d\xi_1 d\xi_2.$$

Noting that ω is not differentiated in the last formula wrt ξ_1 we can use density arguments and extend this relation to a broader class of test functions ω, not necessarily vanishing at $\xi_1 = 0, 1$. In particular, we can now set $\omega = \frac{\partial v}{\partial \xi_1}$ and get

$$\int_{\Omega_\xi} N \left(\frac{\partial^2 v}{\partial \xi_1 \partial \xi_2} \right)^2 d\xi_1 d\xi_2 = \int_{\Omega_\xi} \frac{\partial (Ju)}{\partial \xi_1} \frac{\partial v}{\partial \xi_1} \, d\xi_1 d\xi_2 - \int_{\Omega_\xi} \frac{\partial N}{\partial \xi_1} \frac{\partial v}{\partial \xi_2} \frac{\partial^2 v}{\partial \xi_1 \partial \xi_2} \, d\xi_1 d\xi_2 \, .$$

This, reminding $\left\| \frac{\partial v}{\partial \xi_2} \right\|_{L^2} \le c ||\nabla_\parallel v||_{L^2} \le c ||u||_{L^2}$, entails by Young inequality

$$\left\| \frac{\partial^2 v}{\partial \xi_1 \partial \xi_2} \right\|_{L^2}^2 \le c\gamma ||u||_{H^1}^2 + \frac{c}{\gamma} \left\| \frac{\partial v}{\partial \xi_1} \right\|_{L^2}^2 + c ||u||_{L^2}^2 \, , \tag{47}$$

with a fixed constant $c > 0$ and arbitrary $\gamma > 0$.

Applying a Poincaré type inequality to $J \frac{\partial v}{\partial \xi_1}$, we can write $\forall \xi_1 \in (0,1)$

$$\int_0^1 \left(\frac{\partial v}{\partial \xi_1} \right)^2 d\xi_2 \le C \int_0^1 \left(\frac{\partial^2 v}{\partial \xi_1 \partial \xi_2} \right)^2 d\xi_2 + C \left(\int_0^1 J \frac{\partial v}{\partial \xi_1} \, d\xi_2 \right)^2 . \tag{48}$$

Remind that $v \in \mathcal{A}$, which means

$$\int_0^1 J(\xi_1, \xi_2) \, v(\xi_1, \xi_2) \, d\xi_2 = 0 \quad \forall \xi_1 \in (0,1) \, ,$$

or, after differentiation wrt ξ_1,

$$\int_0^1 J \frac{\partial v}{\partial \xi_1} \, d\xi_2 + \int_0^1 \frac{\partial J}{\partial \xi_1} \, v \, d\xi_2 = 0 \quad \forall \xi_1 \in (0,1) \, .$$

Relation (48) can be now rewritten as

$$\int_0^1 \left(\frac{\partial v}{\partial \xi_1} \right)^2 d\xi_2 \le C \int_0^1 \left(\frac{\partial^2 v}{\partial \xi_1 \partial \xi_2} \right)^2 d\xi_2 + C \left(\int_0^1 \frac{\partial J}{\partial \xi_1} \, v \, d\xi_2 \right)^2$$

which, after integrating over $\xi_1 \in (0,1)$, with the aid of (47) and the the bound $||v||_{L^2} \le C ||u||_{L^2}$, gives

$$\left\| \frac{\partial v}{\partial \xi_1} \right\|_{L^2}^2 \le C \left\| \frac{\partial^2 v}{\partial \xi_1 \partial \xi_2} \right\|_{L^2}^2 + C \|v\|_{L^2}^2 \le Cc\gamma ||u||_{H^1}^2 + \frac{Cc}{\gamma} \left\| \frac{\partial v}{\partial \xi_1} \right\|_{L^2}^2 + \tilde{C} ||u||_{H^1}^2 .$$

This implies, taking γ sufficiently big.

$$\left\| \frac{\partial v}{\partial \xi_1} \right\|_{L^2} \le C ||u||_{H^1} \, ,$$

which gives the desired result since, as already noted, $\left\| \frac{\partial v}{\partial \xi_2} \right\|_{L^2} \le c ||u||_{L^2}$. ■

Corollary 19 *Let $\xi \in \mathcal{A}$. Then one has $|\xi|_{\mathcal{V}'} \leq c\,|\xi|_*$ with some constant $c > 0$.*

Proof. One can immediately see that $|\xi|_{\mathcal{V}'} = |u|_{\mathcal{V}}$ where $u \in \mathcal{V}$ solves

$$(\nabla u, \nabla w) = (\xi, w), \quad \forall w \in \mathcal{V}. \tag{49}$$

This means in particular that $\xi = -\Delta u$. Let now $v \in \mathcal{A}$ be the solution to (46), corresponding to u solution to (49). Lemma 18 implies thus that $v \in H^1(\Omega)$ and one has

$$|\xi|_{\mathcal{V}'} = |u|_{\mathcal{V}} = \frac{(-\Delta u, u)}{|u|_{\mathcal{V}}} = \frac{(\xi, u)}{|u|_{H^1}} = \frac{a_\parallel(v, \xi)}{|u|_{H^1}} \leq c\,\frac{a_\parallel(\xi, v)}{\|v\|_{H^1}} \leq c\,|\xi|_*.$$

■

3.2 Numerical analysis for the stabilized AP-scheme

Let us introduce a mesh $\mathcal{T}_h$ on Ω consisting of triangles (resp. rectangles) of maximal size h and let $V_h \subset \mathcal{V}$ be the space of $\mathbb{P}_k$ (resp. $\mathbb{Q}_k$) finite elements on $\mathcal{T}_h$. We want now to discretize the stabilized problem (35) and remark that we can use V_h for both variables u and ξ. We are thus looking for a discrete solution $(u_h^{\varepsilon,\sigma}, \xi_h^{\varepsilon,\sigma}) \in V_h \times V_h$ of

$$(AP_{\mathcal{S}})_h^{\varepsilon,\sigma} \quad \begin{cases} a(u_h^{\varepsilon,\sigma}, v_h) + (1-\varepsilon) a_\parallel(\xi_h^{\varepsilon,\sigma}, v_h) = (f, v_h)\,, & \forall v_h \in V_h \\ a_\parallel(u_h^{\varepsilon,\sigma}, w_h) - \varepsilon a_\parallel(\xi_h^{\varepsilon,\sigma}, w_h) - \sigma(\xi_h^{\varepsilon,\sigma}, w_h) = 0, & \forall w_h \in V_h\,. \end{cases} \tag{50}$$

Let us decompose now $V_h = G_h \oplus A_h$ with $G_h = V_h \cap \mathcal{G} = V_h \cap \mathcal{G}_{\mathcal{L}}$ and A_h being the L^2-orthogonal complement of G_h. Taking test functions from G_h in the second equation of (50), we see that $\xi_h^{\varepsilon,\sigma} \in A_h$ so that this problem can be in fact equivalently rewritten as: Find $(u_h^{\varepsilon,\sigma}, \xi_h^{\varepsilon,\sigma}) \in V_h \times A_h$ such that

$$(AP_{\mathcal{S}})_h^{\varepsilon,\sigma} \quad \begin{cases} a(u_h^{\varepsilon,\sigma}, v_h) + (1-\varepsilon) a_\parallel(\xi_h^{\varepsilon,\sigma}, v_h) = (f, v_h)\,, & \forall v_h \in V_h \\ a_\parallel(u_h^{\varepsilon,\sigma}, w_h) - \varepsilon a_\parallel(\xi_h^{\varepsilon,\sigma}, w_h) - \sigma(\xi_h^{\varepsilon,\sigma}, w_h) = 0, & \forall w_h \in A_h\,. \end{cases} \tag{51}$$

The advantage of the last reformulation is purely analytical, as we can now reintroduce the mesh dependent norm on A_h

$$|\xi_h|_{*h} = \sup_{v_h \in V_h} \frac{a_\parallel(\xi_h, v_h)}{|v_h|_{\mathcal{V}}}\,.$$

We now equip the space $X_h := V_h \times A_h$ with the norm $\|u_h, \xi_h\|_{X_{\varepsilon,\sigma,h}} := (|u_h|_V^2 + |\xi_h|_{*h}^2 + \varepsilon|\xi_h|_\parallel^2 + \sigma|\xi_h|_{L^2}^2)^{1/2}$. By Lemma 12, the bilinear form $C_{\varepsilon,\sigma}$ is continuous on $(X_h, ||\cdot,\cdot||_{X_{\varepsilon,\sigma,h}})$ and enjoys the inf-sup property

$$\inf_{(u_h,\xi_h)\in X_h} \sup_{(v_h,w_h)\in X_h} \frac{C_{\varepsilon,\sigma}((u_h,\xi_h),(v_h,w_h))}{|u_h,\xi_h|_{X_{\varepsilon,\sigma,h}}|v_h,w_h|_{X_{\varepsilon,\sigma,h}}} \geq \beta\,, \tag{52}$$

with a constant $\beta > 0$ that does not depend neither on the mesh nor on ε and σ. This implies

Theorem 20 ***(Discrete Existence/Uniqueness/σ,ε-Convergences)*** *The discrete AP-problem (50) admits a unique solution $(u_h^{\varepsilon,\sigma},\xi_h^{\varepsilon,\sigma})\in V_h\times V_h$, satisfying*

$$\|u_h^{\varepsilon,\sigma},\xi_h^{\varepsilon,\sigma}\|_{X_{\varepsilon,\sigma,h}}\leq\frac{1}{\beta}\|f\|_{\mathcal{V}'}\,.$$

Moreover, for any $\varepsilon\geq 0$ fixed, one has the convergences

$$u_h^{\varepsilon,\sigma}\to_{\sigma\to 0}u_h^{\varepsilon,0}\,;\qquad \xi_h^{\varepsilon,\sigma}\to_{\sigma\to 0}\xi_h^{\varepsilon,0}\quad,$$

where $(u_h^{\varepsilon,0},\xi_h^{\varepsilon,0})\in V_h\times A_h$ is the unique solution to (51) with $\sigma=0$. We also have

$$u_h^{\varepsilon,0}\to_{\varepsilon\to 0}u_h^{0,0}\,;\qquad \xi_h^{\varepsilon,0}\to_{\varepsilon\to 0}\xi_h^{0,0}\quad,$$

where $(u_h^{0,0},\xi_h^{0,0})\in V_h\times A_h$ is the unique solution to (51) with $\varepsilon=\sigma=0$.

The condition number of the matrix corresponding to problem (50) is bounded by a constant that depends on σ but not on ε (assuming that the same bases of V_h and L_h are chosen for all values of ε,σ).

Proof. The existence and uniqueness of a solution $(u_h^{\varepsilon,\sigma},\xi_h^{\varepsilon,\sigma})\in V_h\times V_h$ for each $\varepsilon\geq 0$, $\sigma>0$ is a simple consequence of the BNB theorem [7]. As mentioned already this solution lies in fact in $V_h\times A_h$ and it is thus also the solution to (51). By the same arguments, the latter problem admits a unique solution $(u_h^{\varepsilon,\sigma},\xi_h^{\varepsilon,\sigma})\in V_h\times A_h$ also in the case $\sigma=0$. To prove the convergence $(u_h^{\varepsilon,\sigma},\xi_h^{\varepsilon,\sigma})\to(u_h^{\varepsilon,0},\xi_h^{\varepsilon,0})$ for $\varepsilon\geq 0$ fixed, we observe

$$\begin{aligned}C_{\varepsilon,\sigma}((u_h^{\varepsilon,\sigma}-u_h^{\varepsilon,0},\xi_h^{\varepsilon,\sigma}-\xi_h^{\varepsilon,0}),(v_h,w_h))&=\sigma(\xi_h^{\varepsilon,0},w_h)\leq\sigma\|\xi_h^{\varepsilon,0}\|_{L^2}\|w_h\|_{L^2}\\&\leq\sqrt{\sigma}\,\|\xi_h^{\varepsilon,0}\|_{L^2}\|v_h,w_h\|_{X_{\varepsilon,\sigma,h}},\qquad\forall(v_h,w_h)\in V_h\times A_h\,,\end{aligned}$$

and we conclude using the discrete inf-sup property for $C_{\varepsilon,\sigma}$. The proof of the other convergence $\varepsilon\to 0$ while $\sigma=0$ is done exactly in the same way as in Theorem 6.

We turn now to the study of the condition number. We recall the notations from the proof of Theorem 6 with the only change that there is no longer the space L_h, which has been replaced by V_h. In particular, the constants μ_q, ν_q, μ_*, ν_* are now evaluated on V_h instead of L_h and one can have $\mu_q=\mu_*=0$. Denoting by $\hat{\mu}$ and $\hat{\nu}$ the minimal and maximal eigenvalues of the mass matrix $(\phi_i^u,\phi_j^u)_{L^2(\Omega)}$ we conclude for any $\varepsilon,\sigma\geq 0$ and any $\Phi_h=(u_h,\xi_h)\in V_h\times V_h$

$$\min(\mu_u^2,\sigma\hat{\mu}^2)||\vec{\Phi}||_2^2\leq||\Phi_h||_{X_{\varepsilon,\sigma,h}}^2\leq\max\{\nu_u^2,\nu_*^2+\varepsilon\,\nu_q^2+\sigma\hat{\nu}^2\}||\vec{\Phi}||_2\,.$$

Introducing the matrix of problem (50), denoted by $\mathbb{A}^{\varepsilon,\sigma}$, and repeating the calculations of Theorem 6 we obtain

$$||\mathbb{A}^{\varepsilon,\sigma}||_2\leq M\max\{\nu_u^2,\nu_*^2+\varepsilon\nu_q^2+\sigma\hat{\nu}^2\}\,,\qquad\beta'\min(\mu_u^2,\sigma\hat{\mu}^2)||\vec{\Phi}||_2\leq||\mathbb{A}^{\varepsilon,\sigma}\vec{\Phi}||_2\,,$$

for any $\vec{\Phi}\in\mathbb{R}^N$, so that one has finally

$$cond_2(\mathbb{A}^{\varepsilon,\sigma})=||\mathbb{A}^{\varepsilon,\sigma}||_2||(\mathbb{A}^{\varepsilon,\sigma})^{-1}||_2\leq\frac{M\max\{\nu_u^2,\nu_*^2+\varepsilon\nu_q^2+\sigma\hat{\nu}^2\}}{\beta\min(\mu_u^2,\sigma\hat{\mu}^2)}\,,$$

which is an ε-independent bound. ■

Remark 21 *As already observed in Remark 7 we have*

$$\hat{\mu} \sim \hat{\nu} \sim h, \quad \mu_u \sim h \text{ and } \nu_u \sim \nu_q \sim \nu_* \sim 1\,.$$

Hence, assuming $\varepsilon, \sigma \in [0,1]$*, one obtains*

$$cond_2(\mathbb{A}^{\varepsilon,\sigma}) \lesssim \frac{1}{\sigma h^2}\,.$$

Theorem 22 ***(h-Convergence)*** *Let* $k \geq 1$ *and* V_h *be the* P_k *or* Q_k *finite element space on a regular mesh* $\mathcal{T}_h$*. Suppose moreover that problem (35) has the solution* $u^{\varepsilon,\sigma} \in H^{k+1}(\Omega)$*,* $\xi^{\varepsilon,\sigma} \in H^{k+1}(\Omega)$ *and problem (40) has a solution* $u^{\varepsilon} \in H^1(\Omega)$*,* $\xi^{\varepsilon} \in H^1(\Omega)$*. Then*

$$|u^{\varepsilon} - u_h^{\varepsilon,\sigma}|_{H^1} \leq Ch^k(|u^{\varepsilon,\sigma}|_{H^{k+1}} + |\xi^{\varepsilon,\sigma}|_{H^{k+1}}) + C\sigma|\xi^{\varepsilon}|_{H^1}\,, \tag{53}$$

with a constant $C > 0$ *that depends neither on the mesh, nor on* ε *or* σ*.*

Proof. In the same way as in the inflow case we prove that

$$|u^{\varepsilon,\sigma} - u_h^{\varepsilon,\sigma}|_{H^1} \leq Ch^k(|u^{\varepsilon}|_{H^{k+1}} + |\xi^{\varepsilon}|_{H^{k+1}}).$$

It remains to invoke Lemma 14 and the triangle inequality to conclude. ■

Remark 23 *The error estimate (53) would be of course useless if the norms* $|u^{\varepsilon,\sigma}|_{H^{k+1}}$*,* $|\xi^{\varepsilon,\sigma}|_{H^{k+1}}$ *were dependent on* ε *and* σ*. Fortunately, it is not the case. We expect indeed that* $|u^{\varepsilon,\sigma}|_{H^{k+1}}$ *is bounded uniformly in* ε *by the norm of* f *in* $H^{k-1}(\Omega)$ *and* $|\xi^{\varepsilon,\sigma}|_{H^{k+1}}$ *is bounded uniformly in* ε *by the norm of* f *in* $H^{k+1}(\Omega)$*. This can be easily proved in the case of simple aligned geometry, see Appendix A.*

Remark 24 *One can also easily obtain*

$$|\xi^{\varepsilon,\sigma} - \xi_h^{\varepsilon,\sigma}|_{H^1} \leq \frac{C}{\sqrt{\varepsilon}}\left[(|u^{\varepsilon,\sigma}|_{H^{k+1}} + |\xi^{\varepsilon,\sigma}|_{H^{k+1}}) + \sigma|\xi^{\varepsilon}|_{H^1}\right]$$

which degenerates as in the inflow case, as ε *goes to 0. Again, we are not sure, if this estimate is sharp, but we recall that* $\xi^{\varepsilon,\sigma}$ *is an auxiliary variable, without any intrinsic meaning.*

4 Numerical tests

Let us now study numerically both AP-reformulations, the inflow as well as the stabilized one. We consider in the following a square computational domain $\Omega = [0,1] \times [0,1]$ and the non-uniform and not coordinate-aligned b field:

$$b = \frac{B}{|B|}\,, \quad B = \begin{pmatrix} \alpha(2y-1)\cos(\pi x) + \pi \\ \pi\alpha(y^2 - y)\sin(\pi x) \end{pmatrix}\,, \tag{54}$$

as well as a sample function u^0 which is constant in the direction of the b-field:

$$u^0 = \sin\left(\pi y + \alpha(y^2 - y)\cos(\pi x)\right). \tag{55}$$

Here $\alpha \geq 0$ is a parameter to be fixed in the following different test cases and describes the variations of b. We choose u^0 to be the $\varepsilon \to 0$ limit solution of the anisotropic problem $(P)^\varepsilon$, hence solution of (12), and construct an exact solution of (10) by adding a perturbation proportional to ε, *i.e.*

$$u^\varepsilon = \sin\left(\pi y + \alpha(y^2 - y)\cos(\pi x)\right) + \varepsilon\cos\left(2\pi x\right)\sin\left(\pi y + \alpha(y^2 - y)\cos(\pi x)\right). \tag{56}$$

Note that the auxiliary variable q^ε, solution of (15), is in this case equal to

$$q^\varepsilon = \cos\left(2\pi x\right)\sin\left(\pi y + \alpha(y^2 - y)\cos(\pi x)\right) - \sin\left(\pi y + \alpha(y^2 - y)\cos(\pi x)\right). \tag{57}$$

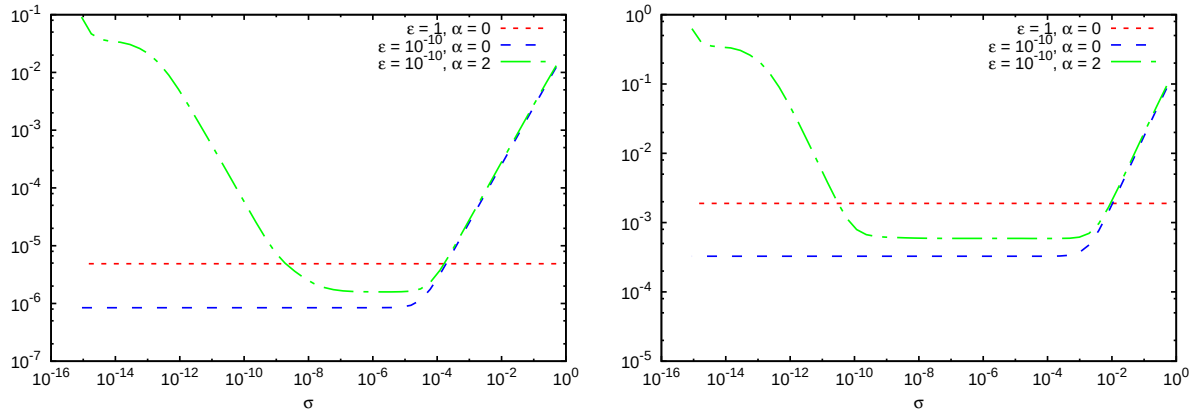
Finally, we compute the right hand side accordingly and have thus constructed an exact solution of problem (10). All simulations (unless stated otherwise) were performed using a $\mathbb{Q}_2$ finite element method.

Aim of this section is to study and validate from a numerical point of view the error estimates established in the last two sections. In particular, we investigate firstly the error introduced by the stabilization procedure in the $(AP_{\mathcal{S}})^{\varepsilon,\sigma}$ formulation, meaning the σ-convergence estimate of (45) in Theorem 14 is verified numerically. Then the h-convergence of both methods is studied and the estimates (32) and (53) are confirmed in both anisotropic ($\varepsilon \ll 1$) and isotropic ($\varepsilon \sim 1$) regimes. Next, we show that both methods are Asymptotic-Preserving in the parameter ε. The conditioning of the corresponding linear systems appear effectively to scale in agreement with Remarks 7 and 21. Finally, the case of a less regular force term f, belonging merely to $L^2(\Omega)$ (and not to $H^1(\Omega)$) is studied — the convergence of the schemes is tested beyond the validity of Theorems 8 and 22.

4.1 Stabilization error ($\varepsilon \geq 0$, $h > 0$ fixed, $\sigma \to 0$)

Let us start by studying the error introduced by a stabilization term proportional to σ in the $(AP_S)^{\varepsilon,\sigma}$ reformulation, in particular we shall estimate numerically for fixed $\varepsilon \geq 0$ and $h > 0$ the L^2- resp. H^1-errors between the exact solution u^ε constructed in (56) and the numerical stabilized solution $u_h^{\varepsilon,\sigma}$, solution of (35) or (50), *i.e.* $||u^\varepsilon - u_h^{\varepsilon,\sigma}||$. The mesh size h is fixed to 0.01. Numerical simulations are performed for the stabilization constant σ varying from 1 to 10^{-15}, considering three different regimes : no anisotropy ($\varepsilon = 1$), strong anisotropy with direction aligned with the coordinate system ($\varepsilon = 10^{-10}$, $\alpha = 0$) and strong anisotropy with variable direction ($\varepsilon = 10^{-10}$, $\alpha = 2$). The L^2- and H^1-errors are presented as a function of σ in Figure 3.

In the first regime, with no anisotropy present in the system, the stabilization constant does not influence the precision at all. Indeed, it is exactly what is expected as the terms involving $\xi^{\varepsilon,\sigma}$ do not appear for $\varepsilon = 1$ in the first equation of $(AP_{\mathcal{S}})^{\varepsilon,\sigma}$. In the second regime, with strong and aligned anisotropy, the precision of the scheme is



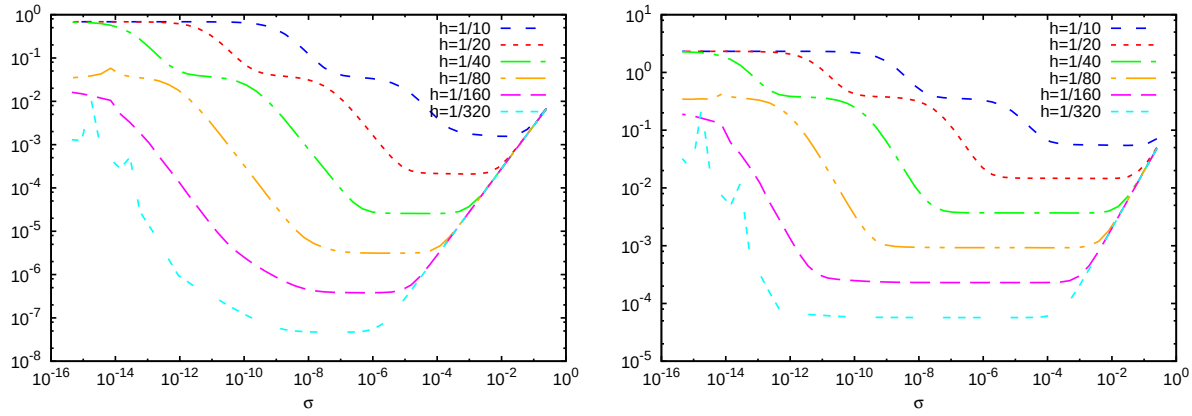
(a) L^2 error (b) H^1 error

Figure 3: Absolute error $||u^\varepsilon - u_h^{\varepsilon,\sigma}||_{L^2}$ (on the left) and $||u^\varepsilon - u_h^{\varepsilon,\sigma}||_{H^1}$ (on the right) with respect to the exact solution u^ε, as a function of σ, for $h = 0.01$ and three regimes : no anisotropy, strong and aligned anisotropy as well as strong anisotropy with variable direction.

influenced by the stabilization procedure only for σ-values greater than 10^{-5} in the L^2-norm and greater than 10^{-3} in the H^1-norm. The error dependence in σ is here linear, according to the Theorem 14, and is explained simply by the fact that the stabilization influences the results for large $\sigma > 0$.

For σ-values smaller than these critical values the accuracy of the scheme in both norms remains unchanged and is given only by the mesh size. This holds true even if the value of the stabilization constant is close to the machine precision (10^{-15}) and can be explained by the fact that we are in an aligned test-case. Indeed, normally for $\sigma \to 0$ the error should increase, due to the non-uniqueness of the ξ-solution. Here we are however plotting the error corresponding to the u^ε-function, which is uniquely determined. The non-uniqueness of $\xi^{\varepsilon,\sigma}$ steps in only in the not-aligned case, which is our third regime of strong anisotropy with variable direction. In this case, the curves show an expected σ-behaviour, the optimal value of σ being between 10^{-8} and 10^{-5} for the L^2-error and between 10^{-10} and 10^{-3} for the H^1-norm. To explain this, observe that in the limit $\sigma \to 0$ the auxiliary function $\xi^{\varepsilon,\sigma}$ is uniquely determined up to a constant on the field lines. This constant will normally not interfere in the computation of $u^{\varepsilon,\sigma}$, as only the parallel derivatives of $\xi^{\varepsilon,\sigma}$ are present in the $u^{\varepsilon,\sigma}$-equation. However, if the mesh is not aligned with the field lines, this parallel derivative mixes the directions, introducing errors which lead to the observed behaviour of the error as $\sigma \to 0$.

Having tested the σ-dependence of the error $||u^\varepsilon - u_h^{\varepsilon,\sigma}||$ for fixed $h > 0$, we are now interested in how these curves remodel for different h-meshes. The σ-convergence is hence compared for different mesh sizes in the most difficult setting, that is to say when a strong anisotropy with variable direction is present in the system ($\varepsilon = 10^{-10}$, $\alpha = 2$). Numerical simulations were performed for the mesh size ranging from 0.1 to



(a) L^2 error (b) H^1 error

Figure 4: Absolute error $||u^\varepsilon - u_h^{\varepsilon,\sigma}||_{L^2}$ (on the left) and $||u^\varepsilon - u_h^{\varepsilon,\sigma}||_{H^1}$ (on the right) as a function of σ, for different values of h and for $\varepsilon = 10^{-10}$ and $\alpha = 2$.

0.003125. Cumulative results are presented on the Figure 4. The plateau for which the accuracy of the scheme does not depend on the stabilization parameter is clearly dependent on the mesh size. As a consequence, the value of σ should be clearly made mesh dependent. We observe that in the case of $\mathbb{Q}_2$ finite elements the upper and lower bounds for the optimal value scale like h^3 and h^4 for the L^2-error, while for the H^1-error the respective scaling is approximately h^2 and h^6. It is therefore reasonable to put $\sigma = h^3$ (or $\sigma = h^2$ if one is interested in the H^1-precision only). Note that this scaling depends on the finite element method used. In general, if a $\mathbb{P}_k$ (or $\mathbb{Q}_k$) method is used, the optimal choice of σ is h^{k+1}, which ensures optimal h-convergence of the method in the L^2-norm.

4.2 h-convergence ($\varepsilon \geq 0$ fixed, $\sigma = h^3$, $h \to 0$)

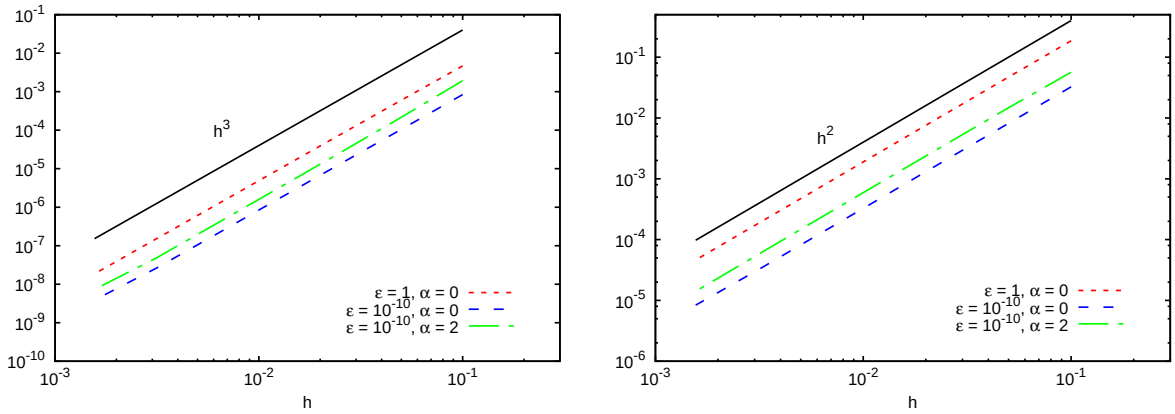
Let us now turn our attention to the h-convergence of both Asymptotic Preserving reformulations (21) and (35). Since the AP-scheme with inflow boundary conditions was studied in a previous work [6] we are mainly interested in the behaviour of the scheme with stabilization. As in the previous subsection, numerical tests are performed in three regimes : an isotropic one ($\varepsilon = 1$ and $\alpha = 0$) and two anisotropic regimes ($\varepsilon = 10^{-10}$ and $\alpha = 0$ or $\alpha = 2$). The stabilization coefficient is set to $\sigma = h^3$ as a consequence of the last subsection. The convergence rate in the L^2- and H^1-norms is presented on Figure 5. As expected the optimal convergence rate (of a $\mathbb{Q}_2$-FEM) is found in both norms. Next we compare the results with the convergence of the AP-scheme with inflow boundary conditions in Tables 1 and 2. Note that in the case of no anisotropy or anisotropy aligned with the coordinate system ($\alpha = 0$), both schemes give quasi exactly the same precision for both L^2 and H^1-norms. In the last regime the stabilized scheme is slightly less accurate compared to the $(AP_{in})^\varepsilon$ scheme. A small loss of the convergence rate of the stabilized scheme is observed for the smallest mesh size in both norms.

h	$\varepsilon = 1$, $\alpha = 0$		$\varepsilon = 10^{-10}$, $\alpha = 0$		$\varepsilon = 10^{-10}$, $\alpha = 2$	
	$(AP_{in})^{\varepsilon}$	$(AP_{\mathcal{S}})^{\varepsilon,\sigma}$	$(AP_{in})^{\varepsilon}$	$(AP_{\mathcal{S}})^{\varepsilon,\sigma}$	$(AP_{in})^{\varepsilon}$	$(AP_{\mathcal{S}})^{\varepsilon,\sigma}$
0.1	5.39×10^{-3}	5.39×10^{-3}	1.19×10^{-3}	1.19×10^{-3}	2.81×10^{-3}	2.18×10^{-3}
0.05	6.97×10^{-4}	6.97×10^{-4}	1.49×10^{-4}	1.49×10^{-4}	3.16×10^{-4}	2.87×10^{-4}
0.025	8.79×10^{-5}	8.79×10^{-5}	1.86×10^{-5}	1.86×10^{-5}	3.77×10^{-5}	3.53×10^{-5}
0.0125	1.10×10^{-5}	1.10×10^{-5}	2.33×10^{-6}	2.33×10^{-6}	4.57×10^{-6}	4.31×10^{-6}
0.00625	1.38×10^{-6}	1.38×10^{-6}	2.91×10^{-7}	2.91×10^{-7}	5.60×10^{-7}	5.29×10^{-7}
0.003125	1.72×10^{-7}	1.72×10^{-7}	3.64×10^{-8}	3.64×10^{-8}	6.89×10^{-8}	6.52×10^{-8}
0.0015625	2.15×10^{-8}	2.15×10^{-8}	5.51×10^{-9}	4.78×10^{-9}	1.07×10^{-9}	8.05×10^{-9}

Table 1: Comparison of the L^2 relative precision $||u^{\varepsilon} - u_h^{\varepsilon,\sigma}||_{L^2}/||u_h^{\varepsilon,\sigma}||_{L^2}$ of both reformulations in both isotropic and anisotropic regimes for different mesh sizes and stabilization constant set to $\sigma = h^3$.

h	$\varepsilon = 1$, $\alpha = 0$		$\varepsilon = 10^{-10}$, $\alpha = 0$		$\varepsilon = 10^{-10}$, $\alpha = 2$	
	$(AP_{in})^{\varepsilon}$	$(AP_{\mathcal{S}})^{\varepsilon,\sigma}$	$(AP_{in})^{\varepsilon}$	$(AP_{\mathcal{S}})^{\varepsilon,\sigma}$	$(AP_{in})^{\varepsilon}$	$(AP_{\mathcal{S}})^{\varepsilon,\sigma}$
0.1	4.48×10^{-2}	4.48×10^{-2}	1.46×10^{-2}	1.46×10^{-2}	2.44×10^{-2}	2.33×10^{-2}
0.05	1.13×10^{-2}	1.13×10^{-2}	3.67×10^{-3}	3.67×10^{-3}	6.34×10^{-3}	6.12×10^{-3}
0.025	2.84×10^{-3}	2.84×10^{-3}	9.19×10^{-4}	9.19×10^{-4}	1.60×10^{-3}	1.54×10^{-3}
0.0125	7.11×10^{-4}	7.11×10^{-4}	2.30×10^{-4}	2.30×10^{-4}	3.99×10^{-4}	3.83×10^{-4}
0.00625	1.78×10^{-4}	1.78×10^{-4}	5.75×10^{-5}	5.75×10^{-5}	9.93×10^{-5}	9.53×10^{-5}
0.003125	4.45×10^{-5}	4.45×10^{-5}	1.44×10^{-5}	1.44×10^{-5}	2.46×10^{-5}	2.37×10^{-5}
0.0015625	1.11×10^{-5}	1.11×10^{-5}	3.76×10^{-6}	3.76×10^{-6}	6.08×10^{-6}	5.87×10^{-6}

Table 2: Comparison of the H^1 relative precision $||u^{\varepsilon} - u_h^{\varepsilon,\sigma}||_{H^1}/||u_h^{\varepsilon,\sigma}||_{H^1}$ of both reformulations in both isotropic and anisotropic regimes for different mesh sizes and stabilization constant set to $\sigma = h^3$.



(a) L^2 error (b) H^1 error

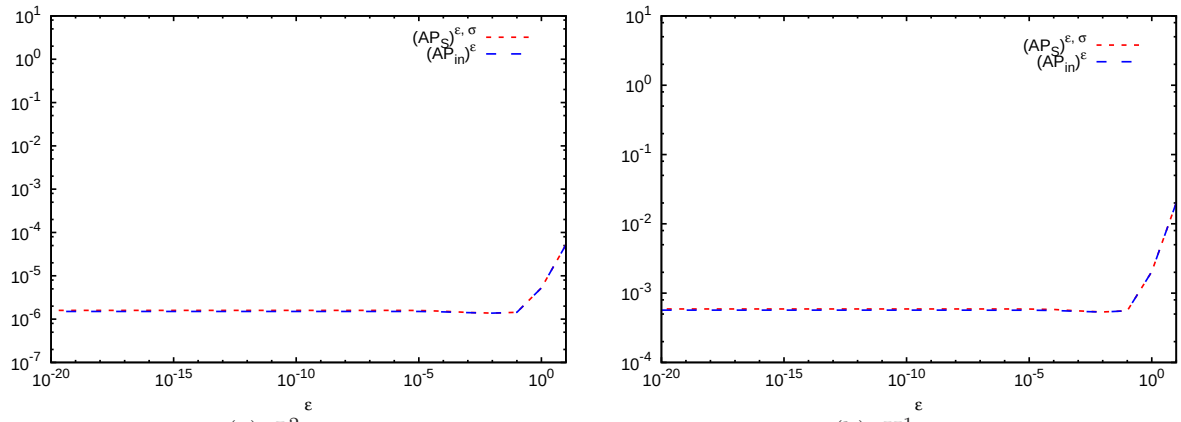
Figure 5: Absolute error $||u^\varepsilon - u_h^{\varepsilon,\sigma}||_{L^2}$ (on the left) and $||u^\varepsilon - u_h^{\varepsilon,\sigma}||_{H^1}$ (on the right) as a function of h and fixed $\sigma = h^3$. One isotropic regime ($\varepsilon = 1$, $\alpha = 0$) and two anisotropic ones: $\varepsilon = 10^{-10}$ and $\alpha = 0$ or $\alpha = 2$ are investigated. The optimal convergence rate is found.

4.3 AP-property ($h > 0$ fixed, $\sigma = h^3$, $\varepsilon \to 0$)

Next, we test if both schemes are indeed Asymptotic Preserving as $\varepsilon \to 0$. The mesh size is fixed to $h = 0.01$, σ is set to $\sigma = h^3$ and numerical simulations are performed for a variable anisotropy direction ($\alpha = 2$) with an anisotropy strength ε varying from 10^{-20} to 10. Both schemes exhibit the desired property, as shown in Figure 6. In particular, the absolute error for both reformulations and in both norms is independent of ε (for $\varepsilon < 0.1$). The error curves are practically indistinguishable. For large ε-values, the errors are increasing due to the fact that the here presented schemes are designed to cope with $\varepsilon \ll 1$ singularities.

4.4 Matrix conditioning

Finally, let us now turn our attention to the conditioning of the matrices associated with the numerical resolution of both schemes $(AP_{in})^\varepsilon$ resp. $(AP_{\mathcal{S}})^{\varepsilon,\sigma}$. The strong anisotropy case with variable direction ($\alpha = 2$) is considered for different mesh sizes $h > 0$. The stabilization constant σ is set to h^3 in the $(AP_S)^{\varepsilon,\sigma}$ reformulation and the anisotropy strength ε is set to 10^{-10}. Sparse matrices were assembled in every case and the condition number was estimated using the matlab function `condest()` returning the estimate of $cond_1$. The results are displayed on Figure 7. As expected, the conditioning scales as $1/h^4$ for the inflow reformulation and as $1/\sigma h^2 = 1/h^5$ for the stabilized method. The first method results in better conditioned matrices in this setting. However, if one is interested mainly in the H^1 precision the stabilization constant σ could be set to h^2 resulting in a conditioning proportional to $1/h^4$ for the stabilized method, discretized with the $\mathbb{Q}_2$ finite elements.



(a) L^2 error (b) H^1 error

Figure 6: Absolute error $||u^{\varepsilon}_{ex} - u^{\varepsilon,\sigma}_{num}||_{L^2}$ (on the left) and $||u^{\varepsilon}_{ex} - u^{\varepsilon,\sigma}_{num}||_{H^1}$ (on the right) as a function of ε for an anisotropy not aligned with the coordinate system ($\alpha = 0$) and the mesh size $h = 0.01$, $\sigma = h^3$. The error curves are superposed, both schemes show similar accuracy independently of ε.

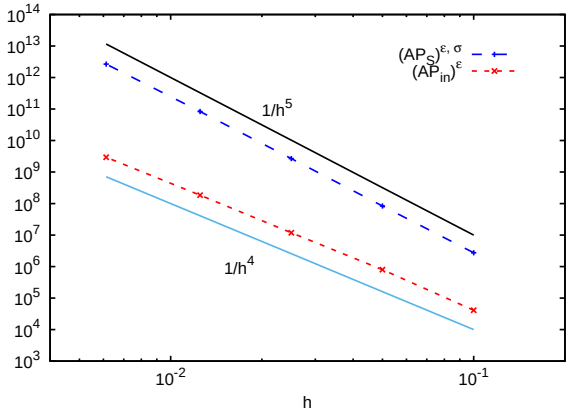


Figure 7: Conditioning ($cond_1$) of the matrices associated with both AP schemes as a function of the mesh size for strong and nonaligned anisotropy ($\varepsilon = 10^{-10}$, $\alpha = 2$). The predicted scaling is found.

4.5 The case of $f \notin H^1(\Omega)$

Aim of this subsection is to investigate the error estimates in a case where the right hand side f is less regular than supposed in the theoretical part of the last two sections. All simulations in this section are preformed with a $\mathbb{Q}_1$ finite element method and the stabilization parameter in the $(AP_{\mathcal{S}})^{\varepsilon,\sigma}$-formulation is set to h^2. In this case we have the h-estimates (32) resp. (53) with $k = 1$ and we recall Remark 9 resp. 23. Let us now choose u^0 to be defined by

$$u^0 = \Big(\big(y + \alpha(y^2 - y)\cos(\pi x)/\pi\big)^2 \ln\big(y + \alpha(y^2 - y)\cos(\pi x)/\pi\big) - 1.5\Big) + 7.5\big(y + \alpha(y^2 - y)\cos(\pi x)/\pi\big)\ , \tag{58}$$

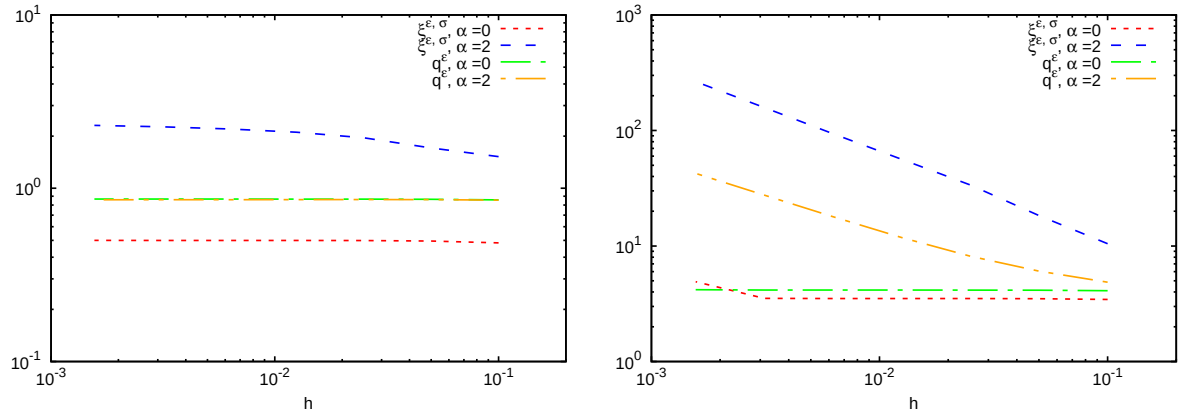
so that the right hand side for the limit problem is a function that belongs to L^2 and not to H^1. If $\alpha = 0$ (the field is aligned), then the right hand side of the limit problem equals to $\ln y$.

We remind that in view of our theoretical result, the H^1-norms of q^ε and $\xi^{\varepsilon,\sigma}$ are not guaranteed to be bounded if the force term is not $H^1(\Omega)$. Nothing can be said on the convergence of the numerical methods in this test case since the right hand side (58) is not in $H^2(\Omega)$. We consider two anisotropic regimes ($\varepsilon = 10^{-10}$): with anisotropy direction aligned with the coordinate system ($\alpha = 0$) resp. with variable direction ($\alpha = 2$).

Numerical simulations show that the H^1-norms of q_h^ε and $\xi_h^{\varepsilon,\sigma}$ grow as h approaches 0 for the variable anisotropy direction. This seems to confirm our expectations. On the other hand, the H^1-norm remains constant when the anisotropy is aligned with the coordinate system. The L^2-norm seems to be bounded regardless of the method for both studied regimes. The results are displayed on Figure 8. To our surprise, optimal convergence rate of u_h^ε and $u_h^{\sigma,\varepsilon}$ is conserved in the tested h range — see Figure 9.

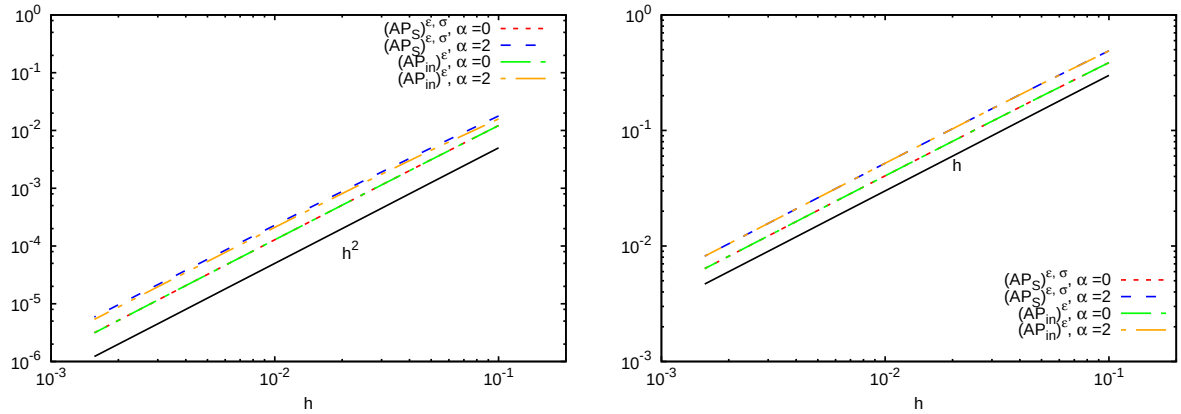
5 Conclusions

A detailed numerical analysis of some asymptotic-preserving numerical schemes, designed to cope with highly anisotropic elliptic problems, was carried out in the present work. In particular, we have shown rigorously that in the limit regimes where traditional schemes become inadequate, AP-schemes are perfectly able to capture the macroscopic behavior of the solution. Convergence results for the schemes were proven, with an accuracy and stability which are shown to be ε-independent, ε being the perturbation parameter responsible for the stiffness of the problem. The development of AP-schemes is based on asymptotic arguments and permit hence to create a link between the various scales in the considered problem, while the numerical parameters remain independent on the stiffness parameter.



(a) L^2 norm (b) H^1 norm

Figure 8: L^2 (on the left) and H^1 (on the right) norms of $\xi^{\varepsilon,\sigma}$ and q^{ε} as a function of the mesh size $h > 0$, for $\varepsilon = 10^{-10}$, $\sigma = h^2$ and $\alpha = 0$ or $\alpha = 2$. The H^1-norms of both auxiliary variables increase with decreasing mesh size for variable direction of anisotropy. The L^2-norms seem to be bounded.



(a) L^2 norm (b) H^1 norm

Figure 9: Absolute error $||u^{\varepsilon} - u_h^{\varepsilon,\sigma}||_{L^2}$ (on the left) and $||u^{\varepsilon} - u_h^{\varepsilon,\sigma}||_{H^1}$ (on the right) as a function of the mesh size $h > 0$, for $\varepsilon = 10^{-10}$, $\sigma = h^2$ and $\alpha = 0$ or $\alpha = 2$. Optimal convergence rate is observed for both methods and both anisotropy configurations.

A The regularity of the solution in the case of simple geometry

Consider the case of $\Omega = (0,\pi)^2$ and the field b looking upwards, *i.e.* $b = e_y$. Moreover let $A_{||} = 1$ and $A_\perp = Id$. We want to explore in this Appendix the regularity of the solution $(u^{\varepsilon,\sigma}, \xi^{\varepsilon,\sigma})$ to (35) when f belongs to $H^s(\Omega)$ and considering an aligned geometry case.
To start, let us first remark that the functions $\{\sqrt{2/\pi}\,\sin kx\}_{k\geq 1}$ as well as $\{\sqrt{2/\pi}\,\cos lx\}_{l\geq 0}$ form an orthogonal basis in $L^2(0,\pi)$ [2], such that each $f \in L^2(\Omega)$ can now be written under the form

$$f(x,y) = \sum_{k=1}^{\infty}\sum_{l=0}^{\infty} f_{kl}\sin kx\cos ly\,, \quad \{f_{kl}\}_{k,l\in\mathbb{N}} \subset l^2\,,$$

which implies immediately that

$$u^{\varepsilon,\sigma}(x,y) = \sum_{k=1}^{\infty}\sum_{l=0}^{\infty} \frac{1}{k^2+l^2+\frac{(1-\varepsilon)l^4}{\varepsilon l^2+\sigma}} f_{kl}\sin kx\cos ly\,,$$

$$\xi^{\varepsilon,\sigma}(x,y) = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \frac{l^2}{(\varepsilon l^2+\sigma)(k^2+l^2)+(1-\varepsilon)l^4} f_{kl}\sin kx\cos ly.$$

Now, if $f \in H^s(\Omega)$, Parseval's equality permits to show that

$$|f|^2_{H^s} \sim \sum_{k=1}^{\infty}\sum_{l=0}^{\infty}(k^2+l^2)^s f_{kl}^2\,,$$

so that

$$|u^{\varepsilon,\sigma}|^2_{H^{s+2}} \sim \sum_{k=1}^{\infty}\sum_{l=0}^{\infty} \frac{(k^2+l^2)^{s+2}}{\left(k^2+l^2+\frac{(1-\varepsilon)l^4}{\varepsilon l^2+\sigma}\right)^2} f_{kl}^2 \leq \sum_{k=1}^{\infty}\sum_{l=0}^{\infty}(k^2+l^2)^s f_{kl}^2 \sim |f|^2_{H^s}\,,$$

$$|\xi^{\varepsilon,\sigma}|^2_{H^s} \sim \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \frac{(k^2+l^2)^s l^4}{((\varepsilon l^2+\sigma)(k^2+l^2)+(1-\varepsilon)l^4)^2} f_{kl}^2 \leq \sum_{k=1}^{\infty}\sum_{l=1}^{\infty}(k^2+l^2)^s f_{kl}^2 \sim |f|^2_{H^s}\,.$$

Moreover, in the case $\sigma = 0$ one has

$$|\partial_{yy}u^{\varepsilon}|^2_{H^s} \sim \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \frac{(k^2+l^2)^s l^4}{\left(k^2+\frac{l^2}{\varepsilon}\right)^2} f_{kl}^2 \leq \varepsilon^2 \sum_{k=1}^{\infty}\sum_{l=0}^{\infty}(k^2+l^2)^s f_{kl}^2 \sim \varepsilon^2 |f|^2_{H^s}\,.$$

In conclusion, if $f \in H^s(\Omega)$ then $u^{\varepsilon,\sigma} \in H^{s+2}(\Omega)$, $\xi^{\varepsilon,\sigma} \in H^s(\Omega)$ and $\partial_{yy}u^\varepsilon \in H^s(\Omega)$ and there is a constant $C > 0$ independent of ε and σ such that

$$|u^{\varepsilon,\sigma}|_{H^{s+2}} \leq C|f|_{H^s}, \quad |\xi^{\varepsilon,\sigma}|_{H^s} \leq C|f|_{H^s} \quad \text{and} \quad |\partial_{yy}u^\varepsilon|_{H^s} \leq C\,\varepsilon\,|f|_{H^s}\,.$$

The same estimates hold true in the inflow case, problem (15), *i.e.* when q^ε is associated with zero boundary conditions on the inflow part. Indeed, the link between q^ε and $\xi^{\varepsilon,0}$ can be explicited as

$$q^\varepsilon(x,y) = \xi^{\varepsilon,0}(x,y) - \xi^{\varepsilon,0}(x,0).$$

Hence, it suffices to study the regularity of the trace function $\chi^\varepsilon(x) := \xi^{\varepsilon,0}(x,0)$. We have

$$\chi^\varepsilon(x) = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \frac{l^2}{\varepsilon l^2(k^2+l^2)+(1-\varepsilon)l^4} f_{kl} \sin kx\,,$$

implying

$$\begin{aligned} |\chi^\varepsilon|^2_{H^s(\Gamma_{in})} &= \sum_{k=1}^{\infty} k^{2s} \left(\sum_{l=1}^{\infty} \frac{l^2}{\varepsilon l^2(k^2+l^2)+(1-\varepsilon)l^4} f_{kl} \right)^2 \le \sum_{k=1}^{\infty} k^{2s} \left(\sum_{l=1}^{\infty} \frac{f_{kl}}{l^2} \right)^2 \\ &\le \sum_{k=1}^{\infty} k^{2s} \left(\sum_{l=1}^{\infty} \frac{1}{l^4} \right) \left(\sum_{l=1}^{\infty} f_{kl}^2 \right) \le C \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} (k^2+l^2)^s f_{kl}^2 \sim |f|^2_{H^s}\,. \end{aligned}$$

We conclude thus $\chi^\varepsilon \in H^s(\Gamma_{in})$ so that $q^\varepsilon \in H^s(\Omega)$ with the $\varepsilon-$independent estimate $|q^\varepsilon|_{H^s} \le C|f|_{H^s}$.

B On the discrete inf-sup condition

As mentioned earlier, the numerical analysis in this paper would be more convenient, if the discrete inf-sup condition (29) were true, i.e.

$$\inf_{q_h \in L_h} \sup_{v_h \in V_h} \frac{a_\parallel(q_h, v_h)}{|q_h|_* |v_h|_{\mathcal{V}}} \ge \alpha \tag{59}$$

with a mesh independent $\alpha > 0$. In more explicit form, this means

$$\forall q_h \in L_h : \sup_{v_h \in V_h} \frac{a_\parallel(q_h, v_h)}{|v_h|_{\mathcal{V}}} \ge \alpha \sup_{v \in \mathcal{V}} \frac{a_\parallel(q_h, v)}{|v|_{\mathcal{V}}}\,. \tag{60}$$

We show first that (59) holds true in a simple aligned geometry. Secondly, we provide a numerical study of a non-aligned case where (59) turns out to be false.

B.1 The case of the aligned geometry

Assume $\Omega = (0, L_x) \times (0, L_y)$, $b = e_2$. We choose moreover V_h as Q_k finite elements ($k \ge 1$) on a rectangular grid aligned with the coordinate axes.

We want to prove first that for all $q_h \in L_h$ there exists $v_h \in V_h$ such that

$$\left(\frac{1}{h} \sum_{E \in E_h} \int_E |[\partial_y q_h]]|^2 ds + \sum_{K \in T_h} \int_K |\partial_{yy} q_h|^2 dx \right)^{\frac{1}{2}} ||v_h||_{L^2} \le C a_{||}(q_h, v_h). \tag{61}$$

A convenient reformulation of this is: all $q_h \in L_h$ there exists $v_h \in V_h$ such that

$$||v_h||^2_{L^2} \leq C_1 \left(\frac{1}{h} \sum_{E \in E_h} \int_E |[\partial_y q_h]|^2 ds + \sum_{K \in T_h} \int_K |\partial_{yy} q_h|^2 dx \right), \tag{62}$$

$$a_{||}(q_h, v_h) \geq C_2 \left(\frac{1}{h} \sum_{E \in E_h} \int_E |[\partial_y q_h]|^2 ds + \sum_{K \in T_h} \int_K |\partial_{yy} q_h|^2 dx \right). \tag{63}$$

Denoting $y_0, \ldots, y_{N_y}$ the y-coordinates of the nodes in our mesh, we can rewrite the quantities above as

$$\begin{aligned}
\frac{1}{h} \sum_{E \in E_h} \int_E |[\partial_y q_h]|^2 ds \; + & \sum_{K \in T_h} \int_K |\partial_{yy} q_h|^2 dx \\
= & \frac{1}{h} \sum_{i=0}^{N_y} \int_0^{L_x} |[\partial_y q_h]|^2 (y_i) dx + \sum_{i=0}^{N_y} \int_0^{L_x} \int_{y_{i-1}}^{y_i} |\partial_{yy} q_h|^2 dy dx \\
||v_h||^2_{L^2} = & \int_0^{L_x} \int_0^{L_y} v_h^2 dy dx \\
a_{||}(q_h, v_h) = & \sum_{i=0}^{N_y} \int_0^{L_x} [\partial_y q_h](y_i) v_h(y_i) dx - \sum_{i=1}^{N_y} \int_0^{L_x} \int_{y_{i-1}}^{y_i} \partial_{yy} q_h v_h dy dx.
\end{aligned}$$

The construction of v_h is particularly easy in the case of piecewise linear finite elements ($k = 1$): we can take $v_h \in V_h$ such that for all $i = 0, \ldots, N_y$

$$v_h(x, y_i) = \frac{1}{h} [\partial_y q_h](x, y_i) \tag{64}$$

Then (63) becomes an equality with $C_2 = 1$. Moreover, (62) is easily verified by a scaling argument, which gives for all $x \in [0, L_x]$

$$\int_0^{L_y} v_h^2(x, y) dy \leq C_1 h \sum_{i=0}^{N_y} v_h^2(x, y_i). \tag{65}$$

We describe now a more complicated construction in the case $k \geq 2$. We observe first that the space V_h can be decomposed as

$$V_h = V_h^{(1)} \oplus V_h^{(2)},$$

with

$V_h^{(1)} = \{v_h \in V_h$ such that $v_h(x, y) = p_{i,x}(y)(y - y_{i-1})(y_i - y)$ for $y \in [y_{i-1}, y_i]$ with $\deg p_{i,x} \leq k-2\}$

and $V_h^{(2)}$ the L^2-orthogonal complement of $V_h^{(1)}$. Now, for a given $q_h \in V_h$ we construct $v_h \in V_h$ as $v_h = v_h^{(1)} + v_h^{(2)}$ with

$$\begin{aligned}
v_h^{(1)} \in & \; V_h^{(1)} : v_h^{(1)}(x, y) = -\partial_{yy} q_h(x, y) \frac{(y - y_{i-1})(y_i - y)}{h^2} \text{ for } y \in [y_{i-1}, y_i] \\
v_h^{(2)} \in & \; V_h^{(2)} : v_h^{(2)}(x, y_i) = \frac{1}{h} [\partial_y q_h](x, y_i).
\end{aligned}$$

In order to see that such $v_h^{(2)}$ exists, we can first take any $v_h \in V_h$ satisfying (64) and then take $v_h^{(2)}$ as the L^2-orthogonal projection of v_h on $V_h^{(2)}$. Since all the functions from $V_h^{(1)}$ vanish at all y_i we shall have $v_h = v_h^{(2)}$ at all y_i. We observe again that (63) becomes an equality with $C_2 = 1$. In order to prove (62) we employ again a scaling inequality of type (65):

$$\int_0^{L_y} v_h^2 dy \leq 2\int_0^{L_y} |v_h^{(1)}|^2 dy + 2\int_0^{L_y} |v_h^{(2)}|^2 dy \leq C\left(\sum_{i=1}^{N_y}\int_{y_{i-1}}^{y_i} |\partial_{yy} q_h|^2 dy + \frac{1}{h}\sum_{i=0}^{N_y} [\partial_y q_h]^2(y_i)\right).$$

Now, (61) being established, we can prove (59) by a Verfürth trick. Take any $q_h \in L_h$. By the definition of the norm $\tilde{\mathcal{L}}$, there exists $v \in V$ such that $|v|_V = |q_h|_{\tilde{\mathcal{L}}}$ and $a_{||}(q_h, v) = (|q_h|_{\tilde{\mathcal{L}}})^2$. Let $\tilde{v}_h \in V_h$ be the Clement interpolant if v such that

$$|v - \tilde{v}_h|_{L^2(\Omega)} \leq Ch|v|_V, \quad |\tilde{v}_h|_{H^1(\Omega)} \leq C|v|_V \quad \text{and} \quad |v - \tilde{v}_h|_{L^2(E_h)} \leq C\sqrt{h}|v|_V.$$

Here E_h denotes the set of all the edges of the mesh with the exception of those lying on Γ_D and the norm in $L^2(E_h)$ is defined by

$$|v|^2_{L^2(E_h)} = \sum_{E\in E_h}\int_E v^2 ds\,.$$

The estimate for $|v - \tilde{v}_h|_{L^2(E_h)}$ follows from the first two estimates and the inequality $|v|^2_{L^2(\partial K)} \leq C||v||_{L^2(K)}||v||_{H^1(K)}$ that we can apply on every $K \in T_h$ and sum up.

We thus observe that

$$\begin{aligned} |q_h|^2_{\tilde{\mathcal{L}}} &= a_{||}(q_h, v) = a_{||}(q_h, v - \tilde{v}_h) + a_{||}(q_h, \tilde{v}_h) \\ &\leq a_{||}(q_h, v - \tilde{v}_h) + |\tilde{v}_h|_{\mathcal{V}} \sup_{v_h\in V_h} \frac{a_{||}(q_h, v_h)}{|v_h|_{\mathcal{V}}} \leq a_{||}(q_h, v - \tilde{v}_h) + C|q_h|_{\tilde{\mathcal{L}}} \sup_{v_h\in V_h} \frac{a_{||}(q_h, v_h)}{|v_h|_{\mathcal{V}}}. \end{aligned}$$

We rewrite now the first term in the last line

$$\begin{aligned} a_{||}(q_h, v - \tilde{v}_h) &= \sum_{K\in T_h}\int_K \nabla_{||} q_h \cdot \nabla_{||}(v - \tilde{v}_h) dx \\ &= \sum_{E\in E_h}\int_E [n_{||} \cdot \nabla_{||} q_h](v - \tilde{v}_h) ds - \sum_{K\in T_h}\int_K (\nabla_{||} \cdot \nabla_{||} q_h)(v - \tilde{v}_h) dx \\ &\leq \left(\frac{1}{h}\sum_{E\in E_h}\int_E |[\partial_y q_h]|^2 ds + \sum_{K\in T_h}\int_K |\partial_{yy} q_h|^2 dx\right)^{\frac{1}{2}} h|v||_V \\ &\leq Ch|v|_V \sup_{v_h\in V_h} \frac{a_{||}(q_h, v_h)}{||v_h||_{L^2}} \leq C|v|_V \sup_{v_h\in V_h} \frac{a_{||}(q_h, v_h)}{|v_h|_V}. \end{aligned}$$

We have used here (61) and the standard inverse inequality. This enables us to conclude

$$|q_h|^2_{\tilde{\mathcal{L}}} \leq C|v|_{\mathcal{V}} \sup_{v_h\in V_h} \frac{a_{||}(q_h, v_h)}{|v_h|_{\mathcal{V}}} + C|q_h|^* \sup_{v_h\in V_h} \frac{a_{||}(q_h, v_h)}{|v_h|_{\mathcal{V}}} \leq C|q_h|_{\tilde{\mathcal{L}}} \sup_{v_h\in V_h} \frac{a_{||}(q_h, v_h)}{|v_h|_{\mathcal{V}}},$$

since $q_h \in V_h$ and $|v|_{\mathcal{V}} = |q_h|_{\tilde{\mathcal{L}}}$. The last inequality gives the desired result (59).

B.2 A numerical study in the case of a general geometry

The aim of this section is to investigate the validity of (59) or equivalently (60) in a more general context by a series of numerical experiments. As in Appendix B.1, we shall assume that $\Omega = (0, L_x) \times (0, L_y)$, $b = e_2$, however this time the grid is no more aligned with the field lines of b. Indeed, we are using here a regular grid made of triangles such that their hypotenuses are no longer aligned with b. Numerical simulations are performed with FreeFem++ [10].

Let V_h be the $\mathcal{P}_1$ finite element space on a mesh described above of size $h > 0$. Observe that the first supremum in (60) is attained on $v_h^* \in V_h$ that satisfies

$$a(v_h^*, w_h) = a_\parallel(q_h, w_h)\,, \quad \forall w_h \in V_h\,. \tag{66}$$

In order to explore the second supremum in (60), we use the finer finite element space $V_{h/2}^f$, constructed via $\mathcal{P}_2$ finite elements on mesh of size $h/2$, i.e. a two-times refinement of the mesh above. The goal in introducing this finer space $V_{h/2}^f$ is to approximate the infinite-dimensional space in (60).

Consider $v_{h/2}^{*f} \in V_{h/2}^f$ that satisfies

$$a(v_{h/2}^{*f}, w_{h/2}^f) = a_\parallel(q_{h/2}, w_{h/2}^f)\,, \quad \forall w_{h/2}^f \in V_{h/2}^f\,. \tag{67}$$

If (60) holds true, than we have

$$\forall q_h \in L_h : \frac{|v_h^*|_\mathcal{V}}{\left|v_{h/2}^{*f}\right|_\mathcal{V}} \geq \alpha\,.$$

Unfortunately, this is false as shown in the following numerical experiment. Let $L_x = L_y = 1$ and let us choose on each mesh of size $h = 1/n$ the function $q_h \in V_h$ defined by its values at the mess nodes as

$$q_h(x_i, y_j) = x_i \sin(\pi n\, y_j/2)\,, \tag{68}$$

where $x_i = ih$, $y_j = jh$, $i, j = 0, \ldots, n$. Note that this function satisfies all the boundary conditions provided n is even. In Fig. 10 we plot the quantity $\frac{|v_h^*|_\mathcal{V}}{|v_{h/2}^{*f}|_\mathcal{V}}$ computed for such a q_h on a series of meshes versus $h = \frac{1}{n}$. It shows clearly that the constant α in (59) is mesh dependent, i.e. it tends to 0 (in general) when the mesh size tends to 0.

Acknowledgments

This work has been supported by the ANR project MOONRISE (MOdels, Oscillations and NumeRIcal SchEmes, 2015-2019). This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom research and training programme 2014-2018 under grant agreement No 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

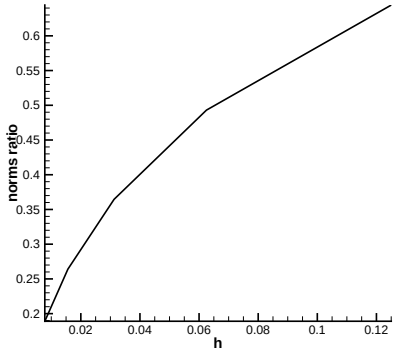


Figure 10: The ratio $\frac{|v_h^*|_{\mathcal{V}}}{|v_h^{*f}|_{\mathcal{V}}}$ computed for q_h given by (68).

References

[1] D. Boffi, F. Brezzi, and M. Fortin. *Mixed finite element methods and applications*. Springer, 2013.

[2] H. Brezis. *Analyse fonctionnelle*. Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1983. Théorie et applications. [Theory and applications].

[3] F. F. Chen. *Plasma Physics and controlled fusion. Plasma Physics*. Springer-Verlag, 2006.

[4] P. Degond, F. Deluzet, A. Lozinski, J. Narski, and C. Negulescu. Duality-based asymptotic-preserving method for highly anisotropic diffusion equations. *Commun. Math. Sci.*, 10(1):1–31, 2012.

[5] P. Degond, F. Deluzet, and C. Negulescu. An asymptotic preserving scheme for strongly anisotropic elliptic problems. *Multiscale Model. Simul.*, 8(2):645–666, 2009/10.

[6] P. Degond, A. Lozinski, J. Narski, and C. Negulescu. An asymptotic-preserving method for highly anisotropic elliptic equations based on a micro-macro decomposition. *Journal of Computational Physics*, 231(7):2724–2740, 2012.

[7] A. Ern and J.-L. Guermond. *Theory and practice of finite elements*, volume 159. Springer, 2004.

[8] V. Girault and P.-A. Raviart. *Finite element methods for Navier-Stokes equations. Theory and algorithms*, volume 5 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1986.

[9] R. D. Hazeltine and J. D. Meiss. *Plasma confinement*. Dover Publications, 2003.

[10] F. Hecht. New development in freefem++. *J. Numer. Math.*, 20(3-4):251–265, 2012.

[11] A. Lozinski, J. Narski, and C. Negulescu. Highly anisotropic nonlinear temperature balance equation and its numerical solution using asymptotic-preserving schemes of second order in time. *ESAIM: Mathematical Modelling and Numerical Analysis*, 48(06):1701–1724, 2014.

[12] J. Narski and M. Ottaviani. Asymptotic preserving scheme for strongly anisotropic parabolic equations for arbitrary anisotropy direction. *Computer Physics Communications*, 185(12):3189–3203, 2014.

[13] R. Schunk and A. Nagy. *Ionospheres: physics, plasma physics, and chemistry*. Cambridge University Press, 2009.