Unification of distance inequalities for linear variational problems

José Alberto Cuminato, Vitoriano Ruas

▶ To cite this version:

José Alberto Cuminato, Vitoriano Ruas. Unification of distance inequalities for linear variational problems. Computational & Applied Mathematics, 2015, 34 (3), pp.1009-1033. 10.1007/s40314-014-0163-6 . hal-01218082

HAL Id: hal-01218082

https://hal.sorbonne-universite.fr/hal-01218082

Submitted on 20 Oct 2015

Unification of Distance Inequalities for Linear Variational Problems

Published in Computational & Applied Mathematics, DOI 10.1007/s40314-014-0163-6

José Alberto CUMINATO [a] Vitoriano RUAS [a,b] [1]

[a] *Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação, São Carlos, BRAZIL*

[b] *Sorbonne Universités, Université Pierre et Marie Curie, UMR 7190, Institut Jean Le Rond d'Alembert/CNRS, Paris, FRANCE*

Abstract

In this work a unifying approach is presented that leads to bounds for the distance in natural norms between solutions belonging to different spaces, of well-posed linear variational problems with the same input data. This is done in a general hilbertian framework, and in this sense, well-known inequalities such as Céa's or Babuška's for coercive and non coercive problems are extended and/or refined, as mere by-products of this unified setting. More particularly such an approach gives rise to both an improvement and a generalization to the weakly coercive case, of second Strang's inequality for abstract coercive problems. Additionally several aspects specific to linear variational problems are the subject of a thorough analysis beforehand, which also allows for clarifications and further refinements about the concept of weak coercivity.

Key words: Babuška, Brezzi, Céa, Dupire, error bounds, linear, Strang, variational problems, weak coercivity.

[1] . Visiting professor in ^{a}. Corresponding author
Email addresses: jacuminato@gmail.com, vitoriano.ruas@upmc.fr (Vitoriano RUAS).

Preprint submitted to *October 20, 2015*

Notation

— 0_E : the null element of a vector space E;

— $\| e \|_E$: the norm of an element e of a normed vector space E;

— S_E : the unit sphere of E (i.e. the set of elements $e \in E$ such that $\| e \|_E = 1$);

— $\mathcal{L}_c(X, Y)$: the space of continuous linear operators from a normed space X into another Y;

— X' : the topological dual space of X (i.e. $\mathcal{L}_c(X, \Re)$);

— $R(A)$: the range of an operator $A \in \mathcal{L}_c(X, Y)$;

— $Ker(A)$: the null space of A;

— $\mathcal{L}_{2c}(X \times Y)$: the space of continuous bilinear forms from $X \times Y$ into $\Re$;

— $Isom_c(X, Y)$: the space of bijective operators in $\mathcal{L}_c(X, Y)$;

— I : the identity operator;

— Π_V : the orthogonal projection operator onto a closed subspace V of a Hilbert space X;

— $V \oplus W$: the direct sum of two subspaces V and W;

— $B \circ A$: the operator in $\mathcal{L}_c(X, Z)$ given by $B \circ A(x) = B(y)$ for $y = A(x)$, $A \in \mathcal{L}_c(X, Y)$, $B \in \mathcal{L}_c(Y, Z)$;

— $A_{|V}$: the restriction to subspace V of X of operator A;

— $A_{W,V}$: the operator from subspace W of X into subspace V of Y equals $\Pi_V \circ A_{|W}$ for $A \in \mathcal{L}_c(X, Y)$;

— $d_X(u, V)$: the distance of an element $u \in X$ to a subset V $(= \inf\limits_{v \in V} \| u - v \|_X)$;

— $\overline{S}$: the closure of a set S in a normed vector space;

— A^* : the adjoint of an operator A;

— $E \setminus S$: the set of elements in E that do not belong to its subset S.

— $L^2(\Omega)$: the space of square (Lebesgue) integrable functions in a bounded open set Ω of $\Re^N$.

— $H^m(\Omega)$: the subspace of $L^2(\Omega)$ of functions having all derivatives up to the m-th order in $L^2(\Omega)$.

— $C^0(\bar{\Omega})$: the space of continuous functions in the closure of a bounded open set $\Omega \subset \Re^N$.

— $C^1(\bar{\Omega})$: the subspace of $C^0(\bar{\Omega})$ of functions having all first order derivatives in $C^0(\bar{\Omega})$.

1. Introduction

This work presents optimized bounds for the distance between solutions to *linear variational problems* in different spaces. For the sake of conciseness throughout the paper we refer to such problems as *LVP*'s.

The main approach here consists of working in a hilbertian framework, under minimum hypotheses on the existence and uniqueness of solutions. In such a way the classical bound that applies to the particular case of coercive problems approximated using closed sub-spaces - known as Céa's inequality -, is recovered. In this respect we observe that both the First and the Second Strang's inequality [7] do not have this property. For this reason we prove corresponding versions that do reduce to Céa's inequality in the absence of so-called variational crimes, not only for coercive bilinear forms, but also in the case of weakly coercive ones in the sense of Nečas, Babuška and Brezzi. All this is achieved by starting from an inequality proved by Bruno Dupire in his doctoral dissertation [10] defended at PUC-Rio - the Catholic University of Rio de Janeiro -, under the second author's supervision. This result is the exact counterpart of Céa's inequality, in the sense that the coercivity constant is simply replaced by the weak coercivity constant, called the co-norm of the bilinear form. The proof of Dupire's inequality is recalled in Proposition 5.1, on the basis of two other results of his, namely, Lemma 4.1 and Proposition 4.3. Extensions of Proposition 5.1 to more general cases usually called *variational crimes* are considered in Sections 5 and 6, in the form of Theorems 5.1, 6.1 and 6.2, which can thus be viewed as the main novelty of this work.

The authors would like to point out that one of their main concerns was rendering this paper as self-contained as possible. This is first because they feel that this subject is treated neither so clearly, nor in a sufficiently accurate or optimal way in specialized text books. In this sense their paper can also be regarded as a review on the subject. Moreover the adopted presentation seems to the authors more likely to be attractive to non mathematicians such as engineers, who are nevertheless assiduous users of variational techniques. For both reasons some results of Functional Analysis necessary for the purpose of this study are integrated to the text, including a few well-known ones with corresponding short proofs. In contrast it should be noted that some propositions given in Sections 3 and 4 are included in the text because, to some extent, they are not so well-known. Actually most of them are aimed at refining or unifying different presentations of weak coercivity usually encountered in the literature.

The paper is organized as follows. Section 2 presents a brief survey of classical distance inequalities in the context of LVP's, together with the paper's motivation and main contributions in this context. In Section 3 the theory of well-posedness of LVP's is recalled, in connection with the concepts of *weak coercivity*, *minored operators* and their *co-norm*. These results are enriched with aspects inherent to finite dimensional problems, more particularly Corollary 3.1, often neglected in text books on the subject. In Section 4 some results of Functional Analysis with applications to LVP's are supplied, and we bring up relevant properties for the subsequent analysis. More especially those related to co-norm are studied in depth, which marginally leads to an alternative definition of weak coercivity. This clarifies the relationship between different definitions commonly encountered. Most of Section 5 and the whole of Section 6 are dedicated to the paper's main contribution. This consists of applications of Dupire's inequality to the case of variational crimes in the broader sense of Strang and Fix [29]. While the so-called *non conforming case* with exact linear and bilinear forms is addressed in Section 5 the case of perturbed forms is considered in Section 6. We conclude in Section 7 with a few important remarks. Additionally we supply an Appendix devoted to an illustration of the theory presented in Sections 3 through 6. We chose an example in which a novel type of convergence analysis for the finite volume method is carried out, by exploiting the concept of weak coercivity. Although the problem under study is rather simple, the authors believe to be providing a convenient framework for the analysis of this popular method, when applied to more complex problems.

2. Motivation

We consider LVP's set in the following framework. Let X and Y be two real Hilbert spaces equipped with inner products $(\cdot|\cdot)_X$ and $(\cdot|\cdot)_Y$ respectively, with corresponding norms denoted by $\| \cdot \|_X$ and $\| \cdot \|_Y$. Given a continuous linear real functional L on Y (i.e. $L \in Y^{'}$) and a a continuous real bilinear form defined on $X \times Y$ (i.e. $a \in \mathcal{L}_{2c}(X \times Y)$), we wish to

$$(P_{X,Y}) \begin{cases} \text{Find } u \in X \quad \text{ such that} \\ a(u,v) = L(v) \;\; \forall v \in Y. \end{cases}$$

Problems of the type $(P_{X,Y})$ are frequent in Applied Sciences. For instance, linear boundary value differential equations can be recast in this form, known as *weak formulation.* The concept of weak coercivity in connection with such a weak formulation seems to have first been developed by Nečas [22]. However it was in the framework of research on variational formulations with Lagrange multipliers - i.e. *mixed formulations* -, mostly conducted by Babuška [2] and Brezzi [5], that it became better known. The underlying theory was first exploited more thoroughly in numerical applications by Babuška himself (cf [2], [3]), and a little later by Raviart and Thomas (cf. [30], [24]) in their pioneering work on mixed finite element methods.
Since weak coercivity is the main property the results given in this work rely upon, we recall it right away:

A bilinear form $a \in \mathcal{L}_{2c}(X \times Y)$ is said to be *weakly coercive* if the following two conditions are satisfied:

— (I) $\exists \alpha_{X,Y} > 0$ such that $\displaystyle \inf_{u \in X \setminus \{0_X\}} \sup_{v \in Y \setminus \{0_Y\}} \frac{a(u,v)}{\| u \|_X \| v \|_Y} \geq \alpha_{X,Y};$

— (II) $\forall v \in Y \setminus \{0_Y\}$, $\exists u \in X$ such that $a(u,v) \neq 0$.

$\alpha_{X,Y}$ is commonly called the *weak coercivity constant* for spaces X and Y. Here we shall rather refer to it as the *co-norm* of a over the pair (X,Y) (more precisely the co-norm is the sup of all $\alpha_{X,Y}$ satisfying (I)).

We emphasize that weak coercivity is a necessary and sufficient condition for problem $(P_{X,Y})$ to have a unique solution, for all functional $L \in Y^{'}$ (cf. Theorem 3.1 hereafter). However when the spaces X and Y are the same, we may consider coercive bilinear forms a on $X \times X$, or more simply on X. This means that there exists a strictly positive constant α such that

$$a(v,v) \geq \alpha \| v \|_X^2 \;\; \forall v \in X. \tag{1}$$

Clearly, a coercive bilinear form on X is weakly coercive not only on $X \times X$ but also on $W \times W$ for every subspace W of X with the same constant. Actually both co-norms $\alpha_{X,X}$ and $\alpha_{W,W}$ can be taken to be equal to the *coercivity constant* α in this case. However, it is possible to derive lower bounds for the co-norm greater than this for coercive bilinear forms.
Notice that, in contrast to coercivity, weak coercivity is a property that cannot be transfered to subspaces in general. Indeed, if W is a subspace of X, (I) clearly holds if in the infimum X is replaced with W. However it is not certain that the same condition holds for the constant $\alpha_{X,Y}$, and not even for some $\alpha_{W,V} > 0$ smaller than $\alpha_{X,Y}$, if Y is also replaced by a subspace $V \subsetneq Y$.

Consider now the linear variational problem $(P_{X,X})$, which under the coercivity assumption has a unique solution u according to the *Lax-Milgram Theorem* [19] together with its *approximate problem*,

$$\begin{cases} \text{Find } u_W \in W \quad \text{such that} \\ a(u_W, v) = L(v) \ \forall v \in W. \end{cases} \tag{2}$$

Then a celebrated inequality due to Céa [6], known as *Céa's Lemma* (cf. [7]), gives an upper bound for the distance between u and u_W measured in the norm of X, namely,

$$\| u - u_W \|_X \leq \frac{\| a \|}{\alpha} \inf_{v \in W} \| u - v \|_X, \tag{3}$$

where $\| a \|$ is the infimum of all constants M such that $a(u,v) \leq M \| u \|_X \| v \|_X$, or equivalently,

$$\| a \| := \sup_{u \in X \setminus \{0_X\},\ v \in X \setminus \{0_X\}} \frac{|a(u,v)|}{\| u \|_X \| v \|_X}. \tag{4}$$

The quantity $\inf_{v \in W} \| u - v \|_X$ is the one to be bounded in turn, if one wishes to derive an *error estimate* for the *approximate solution* u_W of problem $(P_{X,X})$ in case a is coercive over X. Henceforth we will denote by $d_X(v, W)$ the minimum distance measured in the norm of X of an element $v \in X$ to a closed subspace W. In view of this definition we can rewrite (3) as follows:

$$\| u - u_W \|_X \leq \frac{\| a \|}{\alpha} d_X(u, W). \tag{5}$$

One of the ideas exploited in this work is the fact that a bound of the same nature as (5) applies to the case where a is only weakly coercive. More specifically, W being a closed subspace of X and V being a closed subspace of Y, assume that problem $(P_{X,Y})$ has a solution u, and that the continuous bilinear form a on $X \times Y$ is weakly coercive on $W \times V$ with co-norm $\alpha_{W,V}$. Then if u_W is the (unique) solution of a problem derived from $(P_{X,Y})$ when the pair (X, Y) is replaced with (W, V), Dupire [10] proved that:

$$\| u - u_W \|_X \leq \frac{\| a \|}{\alpha_{W,V}} d_X(u, W). \tag{6}$$

Although (6) was demonstrated almost thirty years ago, since Babuška developed his approximation theory for non coercive problems (cf. [2] and [3]) the inequality widely in use is,

$$\| u - u_W \|_X \leq \left[1 + \frac{\| a \|}{\alpha_{W,V}} \right] d_X(u, W), \tag{7}$$

This is probably because up to now the proof of Dupire's result has appeared only in his doctoral dissertation [10]. Actually, except for some citations and particular cases studied in the literature, his work had not yet been properly disseminated, neither in scientific journals nor in a language other than Portuguese. This fact is even more surprising, because much more recently (6) was derived in [31] using a different technique of analysis. However in contrast to [10], the authors in [31] restricted themselves to the above functional setting, and gave an emphasis to the particular case of mixed problems with Lagrange multipliers. In this respect we would like to point out that the latter case had also been addressed by Dupire in [10], who obtained the same qualitative results in terms of distance inequalities as in [31]. However the framework considered by Dupire is more general than the one in [31], as reported at the end of this section. Incidentally since the seventies thorough studies of mixed formulations directly using Babuška and Brezzi theory or not were carried out by many authors. For this reason we refrain from elaborating on such applications here (cf. Section 7 and the Appendix).

In any case inequality (6) allows recovering the one given in Céa's Lemma, by calling to the fact that coercivity implies weak coercivity with the same constant. However a refinement is possible since the constant α may be replaced in (5) by a larger one. Another point to be stressed is that in our approach the case $W \subset X$ and $V \subset Y$, to which [31] is restricted, can be viewed as only a particular case of our general setting. More specifically one of our main results can be viewed as an extension of *Second Strang's inequality* [7], which applies when $X = Y$ and $W(= V)$ is not a subspace of X, known as the *non conforming* case. This lies among the so-called *variational crimes* in the sense of Strang (cf. [28]). Before stating such a result let us recall this celebrated inequality.

Assume that a is a continuous bilinear form on the smallest Hilbert space Z containing both W and X, with norm $\| \cdot \|_Z$ such that $\| x \|_Z = \| x \|_X$ for every $x \in X$ and $\| w \|_Z = \| w \|_W$ for every $w \in W$. This means that both X and W are closed subspaces of Z and moreover,

$$a(t, z) \leq \| a \| \| t \|_Z \| z \|_Z \ \ \forall t \in Z, \ \forall z \in Z.$$

Assuming further that a is also coercive on $W \times W$, that is,

$$a(v, v) \geq \alpha_W \| v \|_W^2 \ \ \forall v \in W,$$

let $u \in X$ be a solution to problem $(P_{X,X})$ with a defined on $X \times X$, and the linear form L in Z'. Setting the *approximate problem*

$$\begin{cases} \text{Find } u_W \in W & \text{such that} \\ a(u_W, v) = L(v) \ \forall v \in W, \end{cases} \tag{8}$$

the second Strang's inequality states that (cf. [7]):

$$\| u - u_W \|_Z \leq \left[1 + \frac{\| a \|}{\alpha_W} \right] d_Z(u, W) + \frac{1}{\alpha_W} \sup_{v \in W \setminus \{0_W\}} \frac{a(u, v) - L(v)}{\| v \|_W}. \tag{9}$$

First of all, under the same boundedness assumptions on a and L, we will prove that a result analogous to (9) but stronger than it, holds for problems $(P_{X,X})$ and (8) in the case where a is only weakly coercive on $W \times W$ with co-norm α_W. More specifically we shall show that,

$$\| u - u_W \|_Z \leq \frac{1}{\alpha_W} \left[\| a \| \, d_Z(u, W) + \sup_{v \in W \setminus \{0_W\}} \frac{a(u, v) - L(v)}{\| v \|_W} \right]. \tag{10}$$

Actually inequality (10) was proved in the Appendix of a paper co-authored by the second author [26], in such a manner that it provides by itself an improvement of (9) and hence of (5). However we point out that a similar result can be proved in an even more general framework, namely:

Assume that Z and T are Hilbert spaces and that X and W are closed subspaces of Z, and Y and V are closed subspaces of T. If $a : Z \times T \longrightarrow \Re$ belongs to $\mathcal{L}_{2c}(Z \times T)$, $L \in T'$, $(P_{X,Y})$ has a solution u and a is weakly coercive on $W \times V$ with constant $\alpha_{W,V}$, then:

$$\| u - u_W \|_Z \leq \frac{1}{\alpha_{W,V}} \left[\| a \| \, d_Z(u, W) + \sup_{v \in V \setminus \{0_T\}} \frac{a(u, v) - L(v)}{\| v \|_T} \right]. \tag{11}$$

Notice that (11) holds irrespective of the fact that W or X (resp. V or Y) is included in one another. In this sense techniques of analysis employed by other authors fail to apply to this case, in particular because they rely upon projections (see. e.g. [31]). We would like to stress that (11) will be proved here by exploiting the concept of co-norm.

3. Well-posed LVP's

We return to the functional framework introduced in the beginning of Section 2 in connection with problem $(P_{X,Y})$, to address in this section the issue of its well-posedness for any datum $L \in Y^{'}$. We also present and/or recall related definitions and properties to be used in the sequel.

3.1. *Preparatory material*

Before going into the study of LVP's, we make a few remarks on the concept of weak coercivity of a bilinear form briefly mentioned in the previous section.
First we note that whenever $X = Y$ and $a \in \mathcal{L}_{2c}(X \times Y)$ is symmetric, that is, $a(u,v) = a(v,u)\ \forall u, v \in X$, the definition of weak coercivity reduces to condition (I), which in this case becomes,

$$\exists \alpha_X > 0 \text{ such that } \inf_{u \in X \backslash \{0_X\}} \sup_{v \in X \backslash \{0_X\}} \frac{a(u,v)}{\parallel u \parallel_X \parallel v \parallel_X} \geq \alpha_X.$$

Next we observe that even when a is not symmetric, property (II) can be recast in a form similar to (I), by interchanging the arguments of a, that is,

$$\exists \alpha^{'}_{X,Y} > 0 \text{ such that } \inf_{v \in Y \backslash \{0_Y\}} \sup_{u \in X \backslash \{0_X\}} \frac{a(u,v)}{\parallel u \parallel_X \parallel v \parallel_Y} \geq \alpha^{'}_{X,Y}. \quad (12)$$

The fact that (12) implies (II) is obvious. The converse statement is also true, as long as (I) holds, and in this case we can even take $\alpha^{'}_{X,Y} = \alpha_{X,Y}$. All this is a consequence of a series of properties to be recalled or proven in the sequel, among which lies the fact that weak coercivity is a necessary and sufficient condition for well-posedness of problem $(P_{X,Y})$ for all possible data $L \in Y^{'}$. Let us go into details.

First of all let A be the linear operator defined by $(Au|v)_Y = a(u,v)$ for every $u \in X$ and for every $v \in Y$. As a trivial consequence of the fact that $a \in \mathcal{L}_{2c}(X \times Y)$, we have $A \in \mathcal{L}_c(X,Y)$. Let A^* be the adjoint operator of A in $\mathcal{L}_c(Y,X)$, given by $(u|A^*v)_X = (Au|v)_Y = a(u,v)$ for every $u \in X$ and for every $v \in Y$. According to well-known results (cf. [33]) the standard norm of A in $\mathcal{L}_c(Y,X)$, simply denoted by $\parallel A \parallel$, equals the norm of A^* in $\mathcal{L}_c(Y,X)$, i.e. $\parallel A^* \parallel = \parallel A \parallel$. Actually we have $\parallel A \parallel = \parallel a \parallel$. Indeed denoting by S_E the unit sphere of a normed space E (i.e. $S_E := \{e \in E \text{ such that } \parallel e \parallel_E = 1\}$), we have,

$$\parallel a \parallel = \sup_{u \in S_X,\, v \in S_Y} a(u,v) = \sup_{u \in S_X} \sup_{v \in S_Y} (Au|v)_Y = \sup_{u \in S_X} \parallel Au \parallel_Y = \parallel A \parallel$$

Although the celebrated *inf-sup* condition (I), together with (II), are the usual way of defining weak coercivity, it is often more convenient to equivalently express it by

- (I$^{'}$) $\exists \alpha_{X,Y} > 0$ such that $\forall u \in S_X$, $\exists v \in S_Y$ satisfying $(Au|v)_Y \geq \alpha_{X,Y}$.

Proposition 3.1 *Conditions (I) and (I$^{'}$) are equivalent.*

Proof. Let (I) hold. Recalling the definition of A, we first note that condition (I) is equivalent to the existence of $\alpha_{X,Y} > 0$ such that $\forall u \in X \setminus \{0_X\}$, $\sup_{v \in Y \setminus \{0_Y\}} \frac{(Au|v)_Y}{\| u \|_X \| v \|_Y} \geq \alpha_{X,Y}$.

Then by a scaling argument, (I) is also seen to be equivalent to

$$\exists \alpha_{X,Y} > 0 \text{ such that } \forall u \in S_X \sup_{v \in Y \setminus \{0_Y\}} \frac{(Au|v)_Y}{\| v \|_Y} = \sup_{v \in S_Y} (Au|v)_Y = \| Au \|_Y \geq \alpha_{X,Y}. \tag{13}$$

Then (I) is readily seen to imply that A is one-to-one. Therefore $Au \neq 0_Y$, $\forall u \in S_X$, and we may define $v_0 = Au / \| Au \|_Y \in S_Y$. It follows that $(Au|v_0)_Y = \| Au \|_Y$, and thus taking into account (13), (I) $\Longrightarrow$ (I$'$) with $v = v_0$.

The converse statement is easily seen to be true. ■

Notice that condition (II) in turn can be equivalently rewritten as

- (II*) $Ker(A^*) = \{0_Y\}$.

In view of this, by extension we shall qualify operator $A \in \mathcal{L}_c(X, Y)$ as weakly coercive if it satisfies both (I$'$) and (II*), or yet if the underlying bilinear form $a \in \mathcal{L}_{2c}(X \times Y)$ satisfies (I) and (II).

3.2. *Minored operators and the Generalized Lax-Milgram Theorem*

Given an operator $B \in \mathcal{L}_c(X, Y)$ we define the co-norm of B, denoted $co(B)$, by

$$co(B) := \inf_{x \in X \setminus \{0_X\}} \frac{\| Bx \|_Y}{\| x \|_X} = \inf_{x \in S_X} \| Bx \|_Y .$$

If $co(B)$ is strictly positive, B is said to be minored, and in this case we can prove,

Proposition 3.2 *[5] An operator $B \in \mathcal{L}_c(X, Y)$ is minored iff $Ker(B) = \{0_X\}$ and $R(B)$ is closed.*

Proof. Let us first assume that $co(B) = \gamma > 0$. If $Bx = 0_Y$ and $x \neq 0_X$ then $0 = \frac{\| Bx \|_Y}{\| x \|_X} \geq \gamma$, which is impossible. Hence $Ker(B) = \{0_X\}$. On the other hand if $\{y_n\}_n$ is a sequence of $R(B)$ that converges to $y \in Y$, let $\{x_n\}$ be the sequence of X defined by $Bx_n = y_n$. We have $\gamma \| x_n - x_m \|_X \leq \| y_n - y_m \|_Y \; \forall m, n$. Since necessarily $\{y_n\}_n$ is a Cauchy sequence of Y, this implies that $\{x_n\}_n$ is also a Cauchy sequence of X. Therefore the latter sequence has a limit $x \in X$, and since B is continuous, $y_n = Bx_n$ converges to Bx. Therefore $y = Bx$ and $R(B)$ is thus closed.

Conversely, assume that $R(B)$ is closed and $Ker(B) = \{0_X\}$. In this case the operator B is a continuous isomorphism from X onto $R(B)$, which is complete since it is a closed subspace of a Hilbert space. Thus from the Banach Inverse Theorem (see e.g. [33]) an inverse operator B^{-1} is defined from $R(B)$ onto X, and moreover it is continuous. It follows that $\forall x \in X$, $\| x \|_X = \| B^{-1}Bx \|_X \leq \| B^{-1} \| \| Bx \|_Y$. Therefore $\| Bx \|_Y \geq \gamma \| x \|_X$ for every $x \in X$ with $\gamma = \| B^{-1} \|^{-1}$ and the Proposition is thus proved. ■

The next result is a crucial one in this work, strongly related to Theorem 3.1 given hereafter

Proposition 3.3 *An operator $A \in \mathcal{L}_c(X, Y)$ is weakly coercive if and only if it is invertible.*

Proof. Let (I) and (II) hold. We know that (I), and hence (I$'$), implies that A is minored. More specifically we have $co(A) = \alpha_{X,Y}$. Therefore from Proposition 3.2 A is one-to-one and furthermore $R(A) = \overline{R(A)}$. On the other hand (II) states that there is no $v \in Y \setminus \{0_Y\}$ orthogonal to Au provided

$u \neq 0_X$. It follows that $R(A)^{\perp} = \{0_Y\}$ or yet that $\overline{R(A)} = Y$. Since $R(A)$ is closed A must be also onto.

Conversely let A^{-1} be well-defined on Y; since A is bounded by assumption, from the Banach Inverse Theorem $\| A^{-1} \|$ is bounded. Hence $\| u \|_X = \| A^{-1}Au \|_X \leq \| A^{-1} \| \| Au \|_Y$. It follows that $\| Au \|_Y \geq \| A^{-1} \|^{-1} \| u \|_X$, and therefore A is minored with $co(A) = \| \underline{A^{-1}} \|^{-1}$, that is (I) holds. Finally (II) also holds for if $Ker(A^*)$ is not reduced to $\{0_Y\}$ we have $R(A) = \overline{R(A)} = Ker(A^*)^{\perp} \subsetneq Y$. But this contradicts the fact that A is onto and the result follows. ■

Next we give the *Generalized Lax-Milgram Theorem* as stated by Brezzi [5].

Theorem 3.1 *[5] Let X and Y be two Hilbert spaces and $a \in \mathcal{L}_{2c}(X \times Y)$. The following two properties are equivalent:*

— 1) a is weakly coercive;

— 2) $\forall L \in Y^{'}$, problem $(P_{X,Y})$ has a unique solution $u \in X$ which satisfies the stability condition

$$\| u \|_X \leq \frac{\| L \|_{Y'}}{\alpha_{X,Y}}, \text{ where } \alpha_{X,Y} \text{ is the co-norm of } a. \tag{14}$$

PROOF. Let a be weakly coercive. According to the Riesz Representation Theorem for every $L \in Y^{'}$ there exists a unique $f_L \in Y$ such that $(f_L|v)_Y = L(v)$ $\forall v \in Y$, and moreover $\| f_L \|_Y = \| L \|_{Y'}$. Hence problem $(P_{X,Y})$ can be equivalently rewritten as $a(u,v) = (f_L|v)_Y$ $\forall v \in Y$ or yet as $(Au|v)_Y = (f_L|v)_Y$ $\forall v \in Y$. This implies in turn that $(P_{X,Y})$ is equivalent to $Au = f_L$ in Y. Since from Proposition 3.3 A is invertible, this problem has a unique solution for all f_L, i.e. for all L. Moreover from (I) we have $\alpha_{X,Y} \| u \|_X \leq \| Au \|_Y = \| f_L \|_Y = \| L \|_{Y'}$ and thus (14) holds.

Conversely, if $(P_{X,Y})$ has a unique solution for all $L \in Y^{'}$, it is clear from the above argument that A must be invertible. It follows from Proposition 3.3 that A is weakly coercive and hence so is a. ■

Remark 1 The direct statement 1) $\Longrightarrow$ 2) of Theorem 3.1 is frequently quoted as the Babuška-Lax-Milgram Theorem [25] or the Banach-Nečas-Babuška Theorem [12], or yet as the Lions-Lax-Milgram Theorem [27]. One of the merits of Brezzi's work [5] dating back to 1974, was to prove that the converse statement 2) $\Longrightarrow$ 1) is also true. This means that in [5] the fact that weak coercivity corresponds to minimum conditions for well-posedness of LVP's with arbitrary data was formally established. Actually this implication has rather been taken for granted in the literature, but to the best of our knowledge it was not explicitly stated by any of the other authors quoted above. This is probably because the converse statement is not so relevant for practical purposes. ■

3.3. *Finite-dimensional problems*

In practice the subspaces W and V are often finite dimensional. The purpose of this subsection is to handle weak coercivity in connection with finite dimensional subspaces. In particular the fact that condition (II) can be replaced with $dim\ W = dim\ V$ for bilinear forms associated with finite dimensional spaces is highlighted in Corollary 3.1. Referring to any classical book on Matrix Analysis such as [15] for the basic tools employed here, let us show this in more detail:

Setting $n = dim\ W$, $m = dim\ V$ and letting $\{\varphi_j\}_{j=1}^{n}$, $\{\zeta_i\}_{i=1}^{m}$ be bases of W and V, assume that $(P_{W,V})$ has a solution u_W. For suitable real coefficients u_j, u_W can be uniquely expanded as,

$$u_W = \sum_{j=1}^{n} u_j \varphi_j$$

Then taking $v = \zeta_i$, $(P_{W,V})$ leads to the rectangular *linear system of algebraic equations* for the unknowns $u_1, u_2, \ldots, u_n$,

$$\sum_{j=1}^{n} u_j a(\varphi_j, v) = L(\zeta_i) \ \text{ for } i = 1, 2, \ldots, m, \ \text{ i.e.,}$$

$$A\overrightarrow{u} = \overrightarrow{b}, \tag{15}$$

where

— A is the $m \times n$ matrix with entries $a_{ij} = a(\varphi_j, \zeta_i)$;

— $\overrightarrow{b}$ is the vector in $\Re^m$ with i-th component $b_i = L(\zeta_i)$;

— $\overrightarrow{u} \in \Re^n$ is the vector of unknown coefficients of u_W.

Trivially enough if system (15) is fulfilled then $(P_{W,V})$ also holds, that is, both problems are equivalent. Now we prove,

Corollary 3.1 *If W and V are finite dimensional spaces, the bilinear form a is weakly coercive over $W \times V$ if and only if either*

— *Condition (I) holds and $m = n$;*

— *Matrix A associated with the bilinear form a is a square invertible matrix,*

both conditions being equivalent.

PROOF. First assume that a is weakly coercive on $W \times V$. This means not only that (I) holds, but also that Problem $(P_{W,V})$ has a unique solution according to Theorem 3.1. The equivalence of problems $(P_{W,V})$ and (15) then means that the weak coercivity of a on $W \times V$ implies that the linear system $A\overrightarrow{u} = \overrightarrow{b}$ has a unique solution for all $\overrightarrow{b}$. But this is only possible if $m = n$ and A is invertible.

Next let A be an invertible $m \times m$ matrix. The equivalence of $(P_{W,V})$ and (15), implies that a is weakly coercive according to Theorem 3.1, i.e., (I) and (II) hold.

Now assume that (I) holds and $m = n$. If A were not an invertible matrix, there would exist $\overrightarrow{u} \neq \overrightarrow{0}$ in $\Re^m$ such that $A\overrightarrow{u} = \overrightarrow{0}$ and hence $(A\overrightarrow{u}|\overrightarrow{v}) = 0 \ \forall \overrightarrow{v} \in \Re^m$. Noticing that if $u = \sum_{j=1}^{m} u_j \varphi_j \in W$ and $v = \sum_{i=1}^{m} v_i \zeta_i \in V$, $a(u, v) = (A\overrightarrow{u}|\overrightarrow{v})$ where $(\cdot|\cdot)$ is the euclidean inner product of $\Re^m$, clearly enough (I) is violated for such a u. Hence A must be an invertible $m \times m$ matrix.

Finally all that is left to prove is that (I) also implies (II) if $m = n$. Indeed if (II) does not hold, there exists $v \in V$, $v \neq 0_Y$, such that $a(u, v) = 0 \ \forall u \in W$. Then $(A\overrightarrow{u}|\overrightarrow{v}) = (\overrightarrow{u}|A^T\overrightarrow{v}) = 0 \ \forall \overrightarrow{u} \in \Re^m$, where $v = \sum_{i=1}^{m} v_i \zeta_i$ with $\overrightarrow{v} \neq \overrightarrow{0}$. This means that $A^T\overrightarrow{v} = \overrightarrow{0}$, but since A is a square invertible matrix so is A^T and therefore $v = 0_Y$. It follows that (II) must hold. ■

To complete this section it is interesting to point out that in the finite-dimensional case one may take

$$\alpha_{W,V} = C_{\varphi,\zeta}\lambda(A)/\mu(A), \text{ where}$$

— $\lambda(A) > 0$ is the smallest eigenvalue of the (positive definite) matrix $A^T A$;

— $\mu(A)$ is the square root of the largest eigenvalue of matrix $A^T A$;

— $C_{\varphi,\zeta}$ is a constant depending only on the bases $\{\varphi_j\}_{j=1}^m$ and $\{\zeta_i\}_{i=1}^m$.

Indeed, let $u \in W$, i.e. $u = \sum_{j=1}^m u_j \varphi_j$ with $\overrightarrow{u} = [u_1, u_2, \dots, u_m]^T \in \Re^m$. Denoting by $\parallel \overrightarrow{v} \parallel$ the euclidean norm of $\overrightarrow{v} \in \Re^n$, for any n, and by $\parallel B \parallel$ the spectral norm of a square matrix B, setting $\overrightarrow{\varphi} := [\parallel \varphi_1 \parallel_X, \parallel \varphi_2 \parallel_X, \dots, \parallel \varphi_m \parallel_X]^T$, we easily derive,

$$\parallel u \parallel_X \leq C_\varphi \parallel \overrightarrow{u} \parallel \quad \text{where } C_\varphi := \parallel \overrightarrow{\varphi} \parallel .$$

Now define $v = \sum_{i=1}^m v_i \zeta_i \in V$ such that $\overrightarrow{v} = [v_1, v_2, \dots, v_m]^T \ = \ A\overrightarrow{u}$.
From well-known results we have $a(u,v) = (A^T A \overrightarrow{u} | \overrightarrow{u}) \geq \lambda(A) \parallel \overrightarrow{u} \parallel^2$.
On the other hand $\parallel \overrightarrow{v} \parallel \leq \parallel A \parallel \parallel \overrightarrow{u} \parallel$ and $\parallel v \parallel_Y \leq C_\zeta \parallel \overrightarrow{v} \parallel$ where $C_\zeta := \parallel \overrightarrow{\zeta} \parallel$ with $\overrightarrow{\zeta} := [\parallel \zeta_1 \parallel_Y, \parallel \zeta_2 \parallel_Y, \dots, \parallel \zeta_m \parallel_Y]^T$. Thus $a(u,v)/ \parallel v \parallel_Y \geq \lambda(A)[C_\zeta \parallel A \parallel]^{-1} \parallel \overrightarrow{u} \parallel \geq \lambda(A)[C_\varphi C_\zeta \parallel A \parallel]^{-1} \parallel u \parallel_X$.
Since $\parallel A \parallel = \mu(A)$, condition (I) follows for $\alpha_{W,V} = C_{\varphi,\zeta}\lambda(A)/\mu(A)$, with $C_{\varphi,\zeta} = [C_\varphi C_\zeta]^{-1}$.

4. Properties of the co-norm

The purpose of this section is to provide preparatory material for the subsequent ones, in the form of a series of properties related to the co-norm. As a by-product we formally establish the equivalence between different definitions of weak coercivity in a finer way than the one usually encountered in the literature. First of all we need some definitions and notations. W being a closed subspace of X and V being a closed subspace of Y, we define:

— Π_V is the orthogonal projection operator from Y onto V;

— $A_{W,V} \in \mathcal{L}_{2c}(W \times V)$ is the operator given by $\Pi_V \circ A_{|W}$

— $A \in \mathcal{L}_c(X,Y)$ is said to be (W,V)-weakly coercive if $A_{W,V}$ is weakly coercive;

— $\alpha_{W,V} = co(A_{W,V})$.

Clearly, stating that A is (W,V)-weakly coercive is equivalent to stating that a is weakly coercive on $W \times V$.

Now we prove the following:

Proposition 4.1 *If $A \in Isom_c(X,Y)$ then $co(A) = \parallel A^{-1} \parallel^{-1}$.*

PROOF. $co(A) := \inf_{u \in X \setminus \{0_X\}} \frac{\| Au \|_Y}{\| u \|_X} \implies co(A) = \inf_{v \in Y \setminus \{0_Y\}} \frac{\| v \|_Y}{\| A^{-1}v \|_X} = \left[\sup_{v \in Y \setminus \{0_Y\}} \frac{\| A^{-1}v \|_X}{\| v \|_Y} \right]^{-1}$, and the result follows. ■

Proposition 4.2 *Let $A \in \mathcal{L}_c(X, Y)$.*

— *a) If $A \in Isom_c(X, Y)$ then $co(A^*) = \| A^{-1} \|^{-1} = co(A)$*

— *b) If $A \in \mathcal{L}_c(X, Y)$ is onto then $co(A) \leq co(A^*)$.*

PROOF. a) Since $A^* \in Isom_c(Y, X)$, according to Proposition 4.1 we have $co(A^*) = \| [A^*]^{-1} \|^{-1}$. On the other hand $[A^*]^{-1} = [A^{-1}]^*$ and therefore $co(A^*) = \| A^{-1} \|^{-1} = co(A)$.
b) If A is also one-to-one we know from a) that $co(A) = co(A^*)$. On the other hand if A is not one-to-one $co(A) = 0 \leq co(A^*)$ and the result follows. ■

The following Corollary of Proposition 4.2 allows for a more symmetric definition of weak coercivity.

Corollary 4.1 *Conditions (I) and (II) on $a \in \mathcal{L}_{2c}(X \times Y)$ are equivalent to conditions ($I^{'}$) and ($II^{'}$) on the related operator $A \in \mathcal{L}_c(X, Y)$, where*

($II^{'}$) $\forall v \in S_Y \ \exists u \in S_X$ such that $(Au|v)_Y \geq \alpha_{X,Y} > 0$.

PROOF. Let (I) and (II) hold. Then the operator A associated with a belongs to $Isom_c(X, Y)$ with $co(A) = \alpha_{X,Y}$. Moreover $\inf_{v \in S_Y} \sup_{u \in S_X} a(u, v) = \inf_{v \in S_Y} \sup_{u \in S_X} (u|A^*v)_X = \inf_{v \in S_Y} \| A^*v \|_X = co(A^*)$. On the other hand, a simple scaling argument like to the one in the proof of Proposition 3.1, implies that for every $v \in S_Y$, there exists $u \in S_X$ such that $\| A^*v \|_X = (A^*v|u)_X$. Then using Proposition 4.2 we derive ($II^{'}$).
Conversely, from Proposition 3.1 conditions (I) and ($I^{'}$) are equivalent and ($II^{'}$) trivially implies (II). ■

In conclusion we give the proofs of two properties of the co-norm that shall play a key role in this work.

Lemma 4.1 *[10] Let $B \in \mathcal{L}_c(X, Y)$ and V be a closed subspace of Y. Set $B_1 := \Pi_V \circ B$ and $B_2 := \Pi_{V^\perp} \circ B$. Then*

$$co(B_1)^2 + \| B_2 \|^2 \leq \| B \|^2 . \tag{16}$$

PROOF. For every $x \in S_X$ we have $\| B \|^2 \geq \| Bx \|_Y^2 = \| B_1 x \|_Y^2 + \| B_2 x \|_Y^2 \geq co(B_1)^2 + \| B_2 x \|_Y^2$. Hence $co(B_1)^2 + \| B_2 \|^2 \leq \| B \|^2$. ■

Proposition 4.3 *[10] Let $A \in \mathcal{L}_c(X, Y)$ and $W \subset X$, $V \subset Y$ be closed subspaces. Set $A_1 := A_{W,V}$ and $A_2 := A_{W^\perp,V}$. If A_1 is onto then*

$$co(A_1)^2 + \| A_2 \|^2 \leq \| \Pi_V \circ A \|^2 . \tag{17}$$

PROOF. Let us apply Lemma 4.1 with $B = A^*_{|V}$. We have $B_1 = \Pi_W \circ A^*_{|V}$ and $B_2 = \Pi_{W^\perp} \circ A^*_{|V}$. It is easy to check that $B = [\Pi_V \circ A]^*$, $B_1 = A_1^*$ and $B_2 = A_2^*$. Therefore $[co(A_1^*)]^2 + \| A_2^* \|^2 \leq \| [\Pi_V \circ A]^* \|^2$. Since A_1 is onto by assumption Proposition 4.2 implies that $co(A_1) \leq co(A_1^*)$. The result is then a simple consequence of the fact that $\| B^* \| = \| B \| \ \forall B \in \mathcal{L}_c(X, Y)$. ■

5. Weakly coercive sub-problems

In this section we apply the properties derived in Section 4 in order to prove the validity of bound (11). In particular this permits comparing the exact solution with alternative solutions to $(P_{X,Y})$ belonging to a Hilbert space W equipped with an inner product $(\cdot|\cdot)_W$, that is not necessarily a subspace of X. We can also use a test-function space V, that is a Hilbert space when equipped with an inner product $(\cdot|\cdot)_V$, but may not be a subspace of Y. This will lead to a more general form of (6), which reduces to this inequality itself in case $W \subset X$, $V \subset Y$, and the inner products on X and W and on Y and V coincide.

Let us first consider the case where W and V are closed subspaces of X and Y. The associated sub-problem of $(P_{X,Y})$ is,

$$(P_{W,V}) \begin{cases} \text{Find } u_W \in W \quad \text{such that} \\ a(u_W, v) = L(v) \ \forall v \in V. \end{cases}$$

We know that if a is weakly coercive on $W \times V$ there exists a unique u_W. Moreover u_W satisfies,

$$(A_{|W} u_W|v)_X = (A_{W,V} u_W|v)_X = L_{|V}(v) \ \forall v \in V \text{ for } L \in Y^{'}. \tag{18}$$

Remark 2 From the Hahn-Banach Theorem (cf. [33]), $\forall M \in V^{'}$, $\exists L \in Y^{'}$ such that $L_{|V} = M$ with $\| L \|_{Y'} = \| M \|_{V'}$. Hence, conversely, $A_{W,V}$ being (W,V)-weakly coercive is a necessary condition for problem $(P_{W,V})$ to have a unique solution $\forall M \in V^{'}$. ■

Now given two other closed subspaces $R \subset X$ and $S \subset Y$ suppose A is (R,S)-weakly coercive with constant $\alpha_{R,S}$. This implies that problem $(P_{R,S})$ has a unique solution $u_R \in R$.
Next we recall the optimal bound of the type (6) for the distance between u_R and u_W measured in the norm of X, proved by Dupire in [10].

Proposition 5.1 *[10] Under the above assumptions on R, W, S and V, if $S \subset V$ it holds,*

$$\| u_R - u_W \|_X \leq \frac{\| a \|}{\alpha_{R,S}} d_X(u_W, R). \tag{19}$$

Proof. Since $S \subset V$, $\forall v \in S$ we have $(A(u_W - u_R|v)_Y = 0$. Hence $\Pi_S \circ A(u_W - u_R) = \{0_Y\}$. Set $u_1 = \Pi_R(u_W - u_R)$ and $u_2 = \Pi_{R^\perp}(u_W - u_R)$. Then $u_2 = \Pi_{R^\perp} u_W$ and therefore $\| u_2 \|_X = d_X(u_W, R)$. Furthermore $\Pi_S \circ A u_1 = -\Pi_S \circ A u_2$ and thus

$$\alpha_{R,S} \| u_1 \|_X \leq \| \Pi_S \circ A_{|R^\perp} \| \| u_2 \|_X .$$

Therefore using the property of orthogonal direct sums it follows that,

$$\alpha^2_{R,S} \| u_R - u_W \|^2_X \leq [\| \Pi_S \circ A_{|R^\perp} \|^2 + \alpha^2_{R,S}] \| u_2 \|^2_X . \tag{20}$$

Now we observe that the operator $\Pi_S \circ A_{|R} \in \mathcal{L}_c(R,S)$ is onto since it is (R,S)-weakly coercive by assumption. Hence using Proposition 4.3 we have $\| \Pi_S \circ A_{|R^\perp} \|^2 + \alpha^2_{R,S} \leq \| \Pi_S \circ A \|^2$. The result follows taking into account that $\| \Pi_S \circ A \| \leq \| a \|$. ■

Now we are ready to prove the bound (11). The corresponding functional setting is the one described above, except for the fact that $S \subset Y$ is no longer assumed to be a subspace of V. This result was proved in [26] by the second author and collaborator in a more particular case. It had also been stated in [10], though without a rigorous proof. The one given in [26] was communicated to B. Dupire [11], and triggered some fruitful exchanges with him on the present work.

Theorem 5.1 *Let $a \in \mathcal{L}_{2c}(X \times Y)$ be weakly coercive on $R \times S$ with constant $\alpha_{R,S}$ and $u_R \in R$ be the unique solution of $(P_{R,S})$. If $(P_{W,V})$ has a solution $u_W \in W$ the following bound holds for the distance between u_R and u_W:*

$$\| u_W - u_R \|_X \leq \frac{1}{\alpha_{R,S}} \left[\| a \| \, d_X(u_W, R) + \sup_{s \in S \setminus \{0_Y\}} \frac{a(u_W, s) - L(s)}{\| s \|_Y} \right]. \qquad (21)$$

Proof. Let $L_u : Y \longrightarrow \Re$ be given by

$$L_u(v) = a(u_W, v) - L(v) \; \forall v \in Y. \qquad (22)$$

Clearly $L_u \in Y^{'}$ with $\| L_u \|_{Y'} \leq \| a \| \| u_W \|_X + \| L \|_{Y'}$. Let us consider the auxiliary problem,

$$\text{Find } r \in R \text{ such that } a(r, s) = L_u(s) \; \forall s \in S. \qquad (23)$$

According to Theorem 3.1 problem (23) has a unique solution. Moreover by construction $a(r, s) = a(u_W, s) - a(u_R, s)$, $\forall s \in S$, that is $w := r + u_R \in R$ satisfies $a(w, s) = a(u_W, s) \; \forall s \in S$. Hence we may apply Proposition 5.1 to derive the bound

$$\| w - u_W \|_X \leq \frac{\| a \|}{\alpha_{R,S}} d_X(u_W, R).$$

This in turn implies that

$$\| u_R - u_W \|_X \leq \frac{\| a \|}{\alpha_{R,S}} d_X(u_W, R) + \| r \|_X \; . \qquad (24)$$

On the other hand recalling the definition of the co-norm we have

$$\alpha_{R,S} \| r \|_X \leq \sup_{s \in S \setminus \{0_Y\}} \frac{a(r, s)}{\| s \|_Y}. \qquad (25)$$

Combining (24) and (25) and using (23) together with (22) the result follows. ■

6. The case of perturbed problems

In this section we consider extensions of the studies carried out in Section 5 to variational sub-problems, where the bilinear form and right hand side functional are perturbations of those in the original problem $(P_{X,Y})$ having the sought-after solution. This study is relevant for several reasons, one of them being that in practice it is sometimes impossible to compute with the exact expressions of a and L. The replacement of a and L by a bilinear form $\tilde{a}$ and a linear form $\tilde{L}$ gives rise to another type of variational crime (cf. [28]), which incidentally may occur together with the one considered in the previous section, that is, non conformity. Whatever the case, it is necessary to bound the distance between the theoretically exact solution and the solution of the perturbed sub-problem.

Let us first consider two closed subspaces W and V of Hilbert spaces X and Y equipped with the inner products $(\cdot|\cdot)_X$ and $(\cdot|\cdot)_Y$, and corresponding norms $\| \cdot \|_X$ and $\| \cdot \|_Y$. Given $\tilde{a} \in \mathcal{L}_{2c}(W \times V)$ and $\tilde{L} \in V^{'}$, we wish to solve

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

Notice that $\tilde{a}$ may not even be defined on $X \times Y$. Assuming that $\tilde{a}$ is weakly coercive on $W \times V$ with constant $\tilde{\alpha} > 0$, problem $(\tilde{P}_{W,V})$ has a unique solution according to Theorem 3.1.

Next we endeavor to give fine bounds for the distance between $\tilde{u}$ and u measured in the norm of X, where u is the solution of $(P_{X,Y})$.

$(P_{W,V})$ is the problem associated with $(\tilde{P}_{W,V})$ when the bilinear form $\tilde{a}$ and the functional $\tilde{L}$ are replaced with $a \in \mathcal{L}_{2c}(X \times Y)$ and $L \in Y^{'}$, for which problem $(P_{X,Y})$ is defined. However here it is convenient to rename such a problem and its solution $u^* \in W$, as follows:

$$(P^*_{W,V}) \begin{cases} \text{Find } u^* \in W \quad \text{ such that} \\ a(u^*,v) = L(v) \ \forall v \in V. \end{cases}$$

Assuming that a is weakly coercive on $W \times V$ with constant $\alpha^* > 0$, Theorem 3.1 implies that $(P^*_{W,V})$ has a unique solution. Moreover from Proposition 5.1 this solution satisfies

$$\| u - u^* \|_X \leq \frac{\| a \|}{\alpha^*} d_X(u,W). \tag{26}$$

Next we estimate the distance between $\tilde{u}$ and u^*.
First we have

$$\tilde{\alpha} \| u^* - \tilde{u} \|_X \leq \sup_{v \in V \setminus \{0_Y\}} \frac{\tilde{a}(u^* - \tilde{u}, v)}{\| v \|_Y}. \tag{27}$$

On the other hand taking into account $(P^*_{W,V})$ and $(\tilde{P}_{W,V})$ we can write

$$\tilde{a}(\tilde{u} - u^*, v) = \tilde{a}(\tilde{u}, v) - [\tilde{a} - a](u^*, v) - L(v) = [\tilde{L} - L](v) - [\tilde{a} - a](u^*, v). \tag{28}$$

Combining (27) and (28), it follows that

$$\tilde{\alpha} \| \tilde{u} - u^* \|_X \leq \sup_{v \in V \setminus \{0_Y\}} \frac{[\tilde{L} - L](v) - [\tilde{a} - a](u^*, v)}{\| v \|_Y}. \tag{29}$$

Finally (29) yields

$$\| \tilde{u} - u^* \|_X \leq \frac{1}{\tilde{\alpha}} \left\{ \sup_{v \in V \setminus \{0_Y\}} \frac{[a - \tilde{a}](u^*, v)}{\| v \|_Y} + \sup_{v \in V \setminus \{0_Y\}} \frac{[\tilde{L} - L](v)}{\| v \|_Y} \right\}. \tag{30}$$

Using the triangle inequality, together with (26) and (30), we have thus proved,

Theorem 6.1 *The distance between the unique solutions to $(\tilde{P}_{W,V})$ and $(P_{X,Y})$ can be bounded by*

$$\| u - \tilde{u} \|_X \leq \frac{\| a \|}{\alpha^*} d_X(u,W) + \frac{1}{\tilde{\alpha}} \left\{ \sup_{v \in V \setminus \{0_Y\}} \frac{[a - \tilde{a}](u^*, v)}{\| v \|_Y} + \sup_{v \in V \setminus \{0_Y\}} \frac{[\tilde{L} - L](v)}{\| v \|_Y} \right\} . \blacksquare \tag{31}$$

To conclude this section we expand the scope of the bounds for problem $(\tilde{P}_{W,V})$, by considering the case where W (resp. V) is not a subspace of X (resp. Y).

Similarly but in a way slightly different from Section 5, we consider two larger Hilbert spaces Z and T equipped with norms $\| \cdot \|_Z$ and $\| \cdot \|_T$ such that both X and W are closed subspaces of Z and both Y and V are closed subspaces of T. For instance we may consider direct sums $Z = X + W$ and $T = Y + V$ and norms such that $\| z \|_Z = \| z \|_X$, for $z \in X$ and $\| t \|_T = \| t \|_Y$, for $t \in Y$. We proceed analogously for W and V, after appropriately equipping these spaces with norms $\| \cdot \|_W$ and $\| \cdot \|_V$, respectively, assuming them to be Hilbert spaces.

Whatever the case, let $a \in \mathcal{L}_{2c}(Z \times T)$ and $L \in T^{'}$. As for $\tilde{a}$ and $\tilde{L}$ we make the same assumptions as above, that is, $\tilde{a} \in \mathcal{L}_{2c}(W \times V)$ and $\tilde{L} \in [V]^{'}$. As long as both a and $\tilde{a}$ are weakly coercive on $W \times V$, with constants α^* and $\tilde{\alpha}$, respectively, the existence and uniqueness of the solutions u^* and $\tilde{u}$ to $(P^*_{W,V})$ and $(\tilde{P}_{W,V})$ are ensured. We still denote by $\| a \|$ the norm of a on $Z \times T$.

Recalling Theorem 5.1 we may compare to u^* the solution u to $(P_{X,Y})$ assumed to exist, as follows

$$\| u - u^* \|_Z \leq \frac{1}{\alpha^*} \left[\| a \| \, d_Z(u, W) + \sup_{v \in V \setminus \{0_T\}} \frac{a(u,v) - L(v)}{\| v \|_T} \right]. \tag{32}$$

Then the remaining of the analysis, that is, bounding $\| u^* - \tilde{u} \|_Z$ is similar to the case where $W \subset X$ and $V \subset Y$. This immediately leads to a more general form of Theorem 6.1, namely,

Theorem 6.2 *The distance between the unique solutions to $(\tilde{P}_{W,V})$ and $(P_{X,Y})$ can be bounded by*

$$\begin{aligned} \| u - \tilde{u} \|_Z \leq & \frac{1}{\alpha^*} \left[d_Z(u, W) + \sup_{v \in V \setminus \{0_V\}} \frac{a(u,v) - L(v)}{\| v \|_V} \right] \\ & + \frac{1}{\tilde{\alpha}} \left\{ \sup_{v \in V \setminus \{0_V\}} \frac{[a - \tilde{a}](u^*, v)}{\| v \|_V} + \sup_{v \in V \setminus \{0_V\}} \frac{[\tilde{L} - L](v)}{\| v \|_V} \right\}. \blacksquare \end{aligned} \tag{33}$$

In fact the second term in brackets on the right hand side of (33) vanishes whenever $V \subset Y$, and if $W \subset X$ we have $Z = X$.

7. Miscellaneous remarks

To conclude we make a couple of comments on different aspects of this work.

1. It is noteworthy that the inequality (31) can be viewed as a generalization to weakly coercive problems of a celebrated bound for the same kind of distance, known as the *First Strang inequality* (cf. [28]). However our bound differs from another one given by Ciarlet in [8] Chapter 4, Section 25, also named this way. Indeed in such a bound u^* is replaced by a suitable *interpolate* $I_W(u) \in W$, such that $\| u - I_W(u) \|_X$ is very close to $d_X(u, W)$. However if we attempt to use the same trick, the multiplicative constant $\| a \| / \tilde{\alpha}$ on the right hand side of (31) will have to be adjusted to $1 + \| a \| / \tilde{\alpha}$. This is a price we didn't want to pay, since from the beginning our purpose was to work in such a manner that we recover exactly Céa's inequality in the simplest case where $X = Y$, $\tilde{a} = a$, $\tilde{L} = L$ and a is coercive on $X \times X$. Notice that in many practical situations the inequality as given in Strang and Fix [29] - and hence (31) -, is sufficient to derive optimal estimates of the distance between $\tilde{u}$ and u.

2. As far as problem $(P_{X,Y})$ is concerned, the existence of a solution u for a particular data set is the only ingredient required in order to use the distance inequalities studied in this work. Eventually neither existence nor uniqueness of such a solution will be assured in general, if at least one assumption of Theorem 3.1 is violated. This situation occurs for instance in the example given in the Appendix, in which the space X equipped with $\| \cdot \|_X$ is not a Hilbert space. In contrast all such assumptions must be satisfied by the *approximate* variational problem.

3. Condition (I) is commonly called the *inf-sup* condition in papers devoted to the study of stability and convergence of numerical approximations of partial differential equations. As already stressed in Section 2 this condition was mostly exploited in the context of mixed finite element approximations, in connection with underlying finite dimensional spaces. In this case (I) is sometimes called the *discrete inf-sup* condition. Particularly Petrov-Galerkin formulations of such problems involving mesh-dependent terms - also known as stabilized formulations -, heavily used this condition, such as [14]. Nevertheless in those works the problem was not really considered in the functional setting of Sections 5 and 6. The authors believe that, as long as suitable mesh-dependent norms are employed, Theorems 5.1, 6.1 and 6.2 provide a handy framework to study this type of formulations. They intend to show this in more detail in a forthcoming work.

To summarize the authors would like to stress that in this work optimal bounds for the distance measured in hilbertian norms, between solutions of linear variational problems of similar nature were derived, in the widest possible functional framework. Moreover in doing so the authors also attempted to complete and clarify a series of definitions and underlying equivalence aspects related to weak coercivity, not properly addressed in the literature. To the best of their knowledge such a comprehensive study on the topic had not been carried out before.

Acknowledgement : This work was partially supported by CNPq, the Brazilian National Research Council and FAPESP, the Research Funding Agence of São Paulo state. The authors would like to thank their colleague Gustavo Buscaglia for helpful comments.

References

[1] R.A. Adams, Sobolev Spaces, Academic Press, New York, 1975.

[2] I. Babuška, Error bound for the finite element method, Numerische Mathematik, **16** (1971), 322-333.

[3] I. Babuška, The finite element method with Lagrangian multipliers, Numerische Mathematik, **20** (1973), 179-192.

[4] D. Braess, **Finite Elements Theory, Fast Solvers, and Applications in Solid Mechanics**, Cambridge University Press, 1997.

[5] F. Brezzi, On the existence, uniqueness and approximation of saddle-point problems arising from Lagrange multipliers, RAIRO Analyse Numérique, **8-2** (1974), 129-151.

[6] J. Céa, Approximation variationnelle des problèmes aux limites, Annales de l'Institut Fourier, **14-2** (1964), 345-444.

[7] P.G. Ciarlet, **The Finite Element Methods for Elliptic Problems**, North-Holland, Amsterdam, 1978.

[8] P.G. Ciarlet, Basic Error Estimates for Elliptic Problems, Vol.II Finite Element Methods (Part 1), in: **Handbook of Numerical Analysis**, J.-L. Lions and P.G. Ciarlet eds., North Holland, Amsterdam, 1991.

[9] R. Courant, Variational methods for the solution of problems of equilibrium and vibrations, Bulletin of the American Mathematical Society **49** (1943), 1-23.

[10] B. Dupire, **Problemas Variacionais Lineares, sua Aproximação e Formulações Mistas**, doctoral disssertation, PUC-Rio the Catholic University of Rio de Janeiro, Brazil, 1985.

[11] B. Dupire, **Personal communication**, Quantitative Research, Bloomberg L.P., N.Y., 2010-2013.

[12] A. Ern and J.L. Guermond. **Theory and Practice of Finite Elements**, Appl. Math. Sci. 159, Springer, N.Y., 2004.

[13] R. Eymard, T.R. Gallouët and R. Herbin, The Finite Volume Method, In: **Handbook of Numerical Analysis, Vol. VII**, Ciarlet, P.G. and Lions, J.-L. eds, North Holland, Amsterdam, 2000.

[14] L. Franca, T.J.R. Hughes and R. Stenberg. Stabilized Finite Element Methods, in: **Incompressible Computational Fluid Dynamics**, M.D. Gunzburger and R.A. Nicolaides eds., Cambridge University Press, p. 87-107, 1994.

[15] F.R. Gantmacher, **Matrix Theory**, Vol. 1 (2nd ed.), American Mathematical Society, 1990.

[16] P. Grisvard, **Elliptic Problems in Non-smooth Domains**, Pittman, Boston, 1985.

[17] S. Idelsohn and E. Oñate, Finite element and finite volume, two good friends, Int. J. Num. Methods Eng., **37** (1994), 3323-3341.

[18] P. Knabner and L. Angermann, **Numerical Methods for Elliptic and Parabolic Partial Differential Equations**, Springer, New York, 2003.

[19] P. Lax and A.N. Milgram, **Parabolic Equations. Contributions to the theory of partial differential equations**, Annals of Mathematical Studies no. 33, Princeton University Press, pp. 167-190, 1954.

[20] R.J. Leveque, **Finite Volume Methods for Hyperbolic Problems**, Cambridge University Press, UK, 2002.

[21] J.-L. Lions and E. Magenes, Problèmes aux limites non homogènes et applications, Dunod, 1968.

[22] J. Nečas, Sur une méthode pour résoudre les équations aux dérivées partielles du type elliptique, voisine de la variationnelle, Annali della Scuola Normale Superiore di Pisa, Classe di Scienze, **3-16-4** (1962), 305-326.

[23] A. Quarteroni, R. Sacco, F. Saleri, **Numerical Mathematics**, Springer, 2010.

[24] P.-A. Raviart & J.-M. Thomas, Mixed Finite Element Methods for Second Order Elliptic Problems, Lecture Notes in Mathematics, Springer Verlag, **606**, 292-315, 1977.

[25] I. Roşca, The Babuška-Lax-Milgram theorem, **in: M. Hazewinke, Encyclopedia of Mathematics**, Springer, 2001.

[26] V. Ruas and J.H. Carneiro de Araujo, A family of methods of the DG-Morley type for polyharmonic equations, The Advances in Applied Mathematics and Mechanics, **3-2** (2010), 303-332.

[27] R.E. Showalter, Monotone operators in Banach space and nonlinear partial differential equations, **Mathematical Surveys and Monographs 49**, American Mathematical Society, Providence, RI, p. xiv+278, 1997.

[28] G. Strang, Variational crimes in the finite element method, in: **The Mathematical Foundations of the Finite Element Method**, A.K. Aziz ed., Academic Press, 1972.

[29] G. Strang and G. Fix, **An Analysis of the Finite Element Method**, Prentice Hall, Englewood Cliffs, 1973.

[30] J.-M. Thomas, Sur l'analyse numérique des méthodes d'éléments finis hybrides et mixtes, Thèse de Doctorat d'Etat, Université Paris VI, 1977.

[31] J. Xu and L. Zikatanov, Some observations on Babuška and Brezzi theories, Numerische Mathematik, **94** (2003), 195-202.

[32] X. Ye, On the relationship between finite volume and finite element methods applied to the Stokes equations, Numerical Methods for Partial Differential Equations, **17-5** (2001), 440-453.

[33] K. Yosida, **Functional Analysis**, Grundlehren der mathematischen Wissenschaften 123, Springer Verlag, Berlin, 1965.

Appendix

The purpose of this Appendix is to illustrate the use of the theory developed in this work.

Certainly the most common practical application of the study of non coercive variational problems arise in the framework of the numerical solution of differential equations. The particular case we consider here is far from new, and the estimates we derive for it are also well-known. The point deserving the reader's attention is that we set the problem in a special functional framework, using the tools developed in Sections 3 through 6. As a consequence we provide a different interpretation of the classical piecewise linear finite element method first introduced by Courant [9] to solve second order elliptic equations, which can be helpful to users.

Actually the application we consider below can be viewed as a non standard *finite volume method* (see e.g. [20] or [13]). This is because the unknown function is represented like in the (piecewise linear) finite element method, but the test functions are replaced by characteristic functions of control volumes. More specifically the problem to solve is either the homogeneous Poisson equation or a second order ordinary differential equation in a bounded domain Ω of $\Re^N$ for $N = 1, 2, 3$, with boundary Γ, namely:

$$\begin{cases} -\Delta u + \sigma u = f \text{ in } \Omega \\ u = 0 \qquad\qquad \text{on } \Gamma, \end{cases} \tag{34}$$

where f is a given function in $C^0(\bar{\Omega})$, and σ is a function satisfying $0 < \sigma_m \leq \sigma(x) \leq \sigma_M \; \forall x \in \bar{\Omega}$ for $N = 1$ and $\sigma = 0$ for $N = 2, 3$. For the sake of simplicity we assume that Ω is an interval when $N = 1$, a polygon when $N = 2$ and a polyhedron when $N = 3$.

The existence and uniqueness of u was established a long time ago, and we know that u lies in the Sobolev space $H^2(\Omega)$ (cf. [1]), if Ω is convex [16].

The (almost) standard equivalent variational form of problem (34) is

$$\begin{cases} \text{Find } u \in X \quad \text{ such that} \\ \bar{a}(u, v) = L(v) \; \forall v \in Y, \end{cases} \tag{35}$$

where $Y = H_0^1(\Omega)$, that is, the subspace of Sobolev space $H^1(\Omega)$ consisting of functions that vanish a.e. on Γ. X is the subspace of Y consisting of functions whose laplacian belongs to $L^2(\Omega)$, $\bar{a}(u, v) := \int_\Omega \mathbf{grad}\, u \cdot \mathbf{grad}\, v \, d\mathbf{x}$ and $L(v) := \int_\Omega f v \, d\mathbf{x}$.

Now referring to classical books on the finite element method such as [4], [7] or [12], among many others, let $\mathcal{P} = \{\mathcal{T}_h\}_h$ be a quasi-uniform family of partitions of Ω into disjoint intervals, triangles or tetrahedra, according to the value of N, satisfying certain compatibility conditions such as $\cup_{K \in \mathcal{T}_h} \bar{K} = \bar{\Omega}$, $\forall \mathcal{T}_h \in \mathcal{P}$. We refer to any of the above references for the other compatibility conditions. The usual interpretation of the subscript h is a reference length characterizing the partition $\mathcal{T}_h$, most commonly the maximum edge length of all its elements. We denote by G_K the centroid of an element $K \in \mathcal{T}_h$.

Now for a given h, let X_h be the sub-space of Y consisting of continuous functions, whose restriction to each element in $\mathcal{T}_h$ is a polynomial of degree less than or equal to one. Let P_i, $i = 1, 2, \ldots, I_h$, be a node - i.e. a vertex - of the partition $\mathcal{T}_h$ in the interior of Ω, and Π_i be a *control volume* associated with P_i: the union of all the elements in $\mathcal{T}_h$ whose closure contains P_i. We denote by Γ_i the boundary of Π_i and further introduce the space Y_h spanned by the characteristic functions χ_i of the Π_i's. We denote by v_i the coefficients in the expansion of a function $v \in Y_h$ with respect to the χ_i's, i.e. $v = \displaystyle\sum_{i=1}^{I_h} v_i \chi_i$.

We may equip X_h with the same norm as we equip both X and Y, namely,

$$\| v \|_X = \| v \|_Y := \left\{ \int_\Omega \left[v^2 + \sum_{j=1}^{N} (\partial v / \partial x_j)^2 \right] d\mathbf{x} \right\}^{1/2}$$

for which Y is a Hilbert space (cf. [21]), and of course so is X_h, but not X. On the other hand we must define a discrete analogue $\| v \|_{Y_h}$ of this norm for $v \in Y_h$, more specifically, $\| v \|_{Y_h} = \left[\int_\Omega v^2 \, \mathbf{x} + \int_\Omega |\mathbf{grad}_h v|^2 \, \mathbf{x} \right]^{1/2}$, where $\mathbf{grad}_h v$ is a discrete analogue of the gradient operator for discontinuous functions on the interfaces of the Π_i's. For practical purposes it is possible to avoid the exact definition of such an operator, as seen hereafter. Nevertheless this definition will be given for $N = 1$, the only case we address in detail.

As one can easily check, $Y_h \cap Y = \{0_Y\}$. Hence the norm of an element $t = v + v_h$ in the direct sum T of Y and Y_h, for $v \in Y$ and $v_h \in Y_h$, is given by $\| t \|_T = [\| v \|_Y^2 + \| v_h \|_{Y_h}^2]^{1/2}$. On the other hand, noting that $X \cap X_h = \{0_X\}$, the direct sum $Z = X + X_h$ is simply normed by $\| \cdot \|_X$.
Using the identity $\int_{\Pi_i} \Delta u \; d\mathbf{x} \equiv \oint_{\Gamma_i} \partial u_{|\Pi_i} / \partial n_i dS$ for $u \in X$, where $\partial \cdot / \partial n_i$ denotes the outer normal derivative over Γ_i, we consider the following problem to approximate (34):

$$\begin{cases} \text{Find } u_h \in X_h \quad \text{ such that} \\ a_h(u_h, v) = L_h(v) \;\; \forall v \in Y_h, \end{cases} \tag{36}$$

where $a_h \in \mathcal{L}_{2c}(X_h \times Y_h)$ and $L_h \in Y^{'}$ are defined by

— $a_h(u,v) := \displaystyle\sum_{i=1}^{I_h} v_i \left[-\frac{N+1}{N} \oint_{\Gamma_i} \frac{\partial u_{|\Pi_i}}{\partial n_i} \; dS + \int_{\Pi_i} \sigma u \; d\mathbf{x} \right]$;

— $L_h(v) = \int_\Omega f_h v \; d\mathbf{x}$.

f_h being defined by $f_h(\mathbf{x}) = f(G_K)$ for every $\mathbf{x} \in K$, $\forall K \in \mathcal{T}_h$.

The fact that (36) has a unique solution for $N = 2$ or $N = 3$, is a consequence of its equivalence with another well-posed problem, corresponding to the classical approximation of (34) by the piecewise linear finite element method, namely,

$$\begin{cases} \text{Find } \bar{u}_h \in X_h \quad \text{ such that} \\ \bar{a}(\bar{u}_h, v) = L_h(v) \;\; \forall v \in X_h, \end{cases} \tag{37}$$

Indeed it is possible to prove that the $I_h \times I_h$ matrix corresponding to problem (36) is exactly the matrix corresponding to problem (37) multiplied by $N+1$. This assertion can be verified by geometric arguments, which we refrain from developing here, since they require a series of definitions and calculations, that would divert attention from the essence of our purposes. On the other hand the right hand side vector of (36) equals the one corresponding to (37) multiplied by $N+1$, and thus this factor accounts for the only difference between both problems for $N > 1$.

There is a large amount of work devoted to the relationship between finite volume and finite element schemes for two- and three-dimensional boundary value problems. Far from being exhaustive, we refer to [17], [18] and [32] for details on this issue. Actually the one-dimensional case provides an ideal framework for illustrating Theorem 6.2, and hence Theorems 6.1 and 5.1 too. For this reason in the rest of this section we address this case in detail, assuming first of all that σ is constant. Afterwards we briefly extend the study to the case of a variable σ.
First we adjust the x-coordinate in such a way that $\Omega = (0, L)$, and label the elements in $\mathcal{T}_h$ as K_i, $i = 1, 2, \ldots, I_h + 1$, where $K_i = (x_{i-1}, x_i)$, with $0 = x_0 < x_1 < \ldots < x_{I_h} < x_{I_h+1} = L$. We further set $h_i := x_i - x_{i-1}$, $\forall i \in \{1, 2, \ldots, I_h + 1\}$.
Next setting $v_0 = v_{I_h+1} = 0$, we define a norm in Y_h as follows:

$$\| v \|_{Y_h}^2 := \int_0^L v^2 \; dx + \sum_{i=1}^{I_h+1} \left[\frac{v_i - v_{i-1}}{h_i} \right]^2 h_i.$$

Clearly T equipped with the underlying norm is a Hilbert space, and moreover we have

$$a_h(u,v) \leq M_h \| u \|_X \| v \|_{Y_h} \;\; \forall u \in X, \; \forall v \in Y_h, \tag{38}$$

for a suitable constant M_h that we will specify later on, after having extended a_h to X.
For the sake of simplicity we consider the particular case where $x_i - x_{i-1} = h$, $\forall i \in \{1, 2, \ldots, I_h + 1\}$, with $h = L/(I_h + 1)$. Since $\tilde{a}(u, v) \leq \parallel u \parallel_X \parallel v \parallel_Y$ $\forall u \in X$, $\forall v \in Y$, it is clear that the bilinear form $a \in \mathcal{L}_{2c}(Z \times T)$ defined by $a(z, t) = \bar{a}(z, v) + a_h(z, v_h)$, where $z = u + u_h$ with $u \in X$ and $u_h \in X_h$, and $t = v + v_h$ with $v \in Y$ and $v_h \in Y_h$, satisfies

$$a(z, t) \leq M \parallel z \parallel_X \parallel t \parallel_T \;\; \forall z \in Z, \; \forall t \in T, \tag{39}$$

for a constant M given as a function of σ and M_h. In order to determine M_h first we extend a_h to X as:

$$a_h(u, v) = -2 \sum_{i=1}^{I_h} h^{-1} \left[\int_{x_{i-1}}^{x_i} u^{'}(x) \, dx - \int_{x_i}^{x_{i+1}} u^{'}(x) dx \right] v_i + \sigma \int_0^L uv \, dx, \; \forall u \in X, \; \forall v \in Y_h. \tag{40}$$

Let us consider the case where $u \in X_h$. Noting that $u^{'}$ is constant in both K_i and K_{i+1}, setting $u_i = u(x_i)$ we have,

$$a_h(u, v) = 2 \sum_{i=1}^{I_h} \left[\frac{u_i - u_{i+1}}{h} + \frac{u_i - u_{i-1}}{h} \right] v_i + \sigma \int_0^L uv \, dx. \tag{41}$$

Applying a simple manipulation in the first term of the summation in (41), we easily derive

$$a_h(u, v) = 2 \sum_{i=1}^{I_h+1} \frac{(u_i - u_{i-1})(v_i - v_{i-1})}{h} + \sigma \int_0^L uv \, dx. \tag{42}$$

Now using the Cauchy-Schwarz inequality this gives,

$$a_h(u, v) \leq 2 \left[\sum_{i=1}^{I_h+1} \frac{(u_i - u_{i-1})^2}{h} \right]^{1/2} \left[\sum_{i=1}^{I_h+1} \frac{(v_i - v_{i-1})^2}{h} \right]^{1/2} + \sigma \left[\int_0^L u^2 \, dx \right]^{1/2} \left[\int_0^L v^2 \, dx \right]^{1/2}, \tag{43}$$

which finally yields $\forall u \in X_h$ and $\forall v \in Y_h$,

$$a_h(u, v) \leq \max[2, \sigma] \parallel u \parallel_X \parallel v \parallel_{Y_h} . \tag{44}$$

Next taking $u \in X$, similar manipulations produce:

$$a_h(u, v) = 2 \sum_{i=1}^{I_h+1} \int_{x_{i-1}}^{x_i} \frac{u^{'}(v_i - v_{i-1})}{h} dx + \sigma \int_0^L uv \, dx. \tag{45}$$

After application of the Cauchy-Schwarz inequality to (45) we obtain:

$$a_h(u, v) \leq 2 \left[\sum_{i=1}^{I_h+1} \int_{x_{i-1}}^{x_i} |u^{'}|^2 \right]^{1/2} \left[\sum_{i=1}^{I_h+1} \left(\frac{v_i - v_{i-1}}{h} \right)^2 h \right]^{1/2} + \sigma \left[\int_0^L u^2 \, dx \right]^{1/2} \left[\int_0^L v^2 \, dx \right]^{1/2}, \tag{46}$$

which gives (44) $\forall v \in Y_h$ and $\forall u \in X$. It follows that $M_h = \max[2, \sigma]$.

The next step is to prove that a is weakly coercive on $X_h \times Y_h$, and to exhibit the underlying constant $\alpha_h > 0$. For this purpose we proceed as follows:
Let u be given in X_h and $v \in Y_h$ be defined by $v_i = u_i$, $i = 1, 2, \ldots, I_h$. Recalling (42) and noticing that $v_{|K_i} = v_i + v_{i-1}$ for $i = 1, 2, \ldots, I_h + 1$ and $\int_{K_i} u \, dx = [u_i + u_{i-1}]h/2$, for $i = 1, 2, \ldots, I_h + 1$, we have

$$a_h(u,v) = 2\sum_{i=1}^{I_h+1}\left[\left(\frac{u_i - u_{i-1}}{h}\right)^2 + \sigma\frac{(u_i+u_{i-1})^2}{2}\right] h. \tag{47}$$

Straightforward calculations lead to

$$\int_0^L [u^{'}]^2\, dx = \sum_{i=1}^{I_h+1}\left(\frac{u_i-u_{i-1}}{h}\right)^2 h. \tag{48}$$

On the other hand by the Friedrichs-Poincaré inequality (cf. [21]) we have

$$\int_0^L u^2\, dx \leq 4L^2\int_0^L |u^{'}|^2\, dx. \tag{49}$$

Plugging (48) and (49) into (47) we easily obtain,

$$a_h(u,v) \geq 2(1+4L^2)^{-1} \parallel u \parallel_X^2 . \tag{50}$$

On the other hand we have, $\int_{K_i} v^2\, dx = (u_{i-1}+u_i)^2 h \leq [u_{i-1}^2 + (u_{i-1}+u_i)^2 + u_i^2]h = 6\int_{K_i} u^2\, dx$. Summing from $i=1$ through $i=I_h+1$ and recalling both (48) and the definition of $\parallel\cdot\parallel_{Y_h}$, we obtain,

$$\parallel v \parallel_{Y_h} \leq \sqrt{6} \parallel u \parallel_X . \tag{51}$$

Finally combining (50) and (51) we immediately conclude that a_h is weakly coercive on $X_h\times Y_h$ with constant $\alpha_h = 2(1+4L^2)^{-1}/\sqrt{6}$.

Theorem 6.2 (for the case $\tilde{a}=a$), can now be applied, leading to,

$$\parallel u-u_h \parallel_X \leq \frac{1}{\alpha_h}\left[M d_X(u,X_h) + \sup_{v\in Y_h\setminus\{0_Y\}}\frac{a(u,v)-L(v)}{\parallel v\parallel_{Y_h}}\right] + \frac{1}{\alpha}\sup_{v\in Y_h\setminus\{0_Y\}}\frac{L_h(v)-L(v)}{\parallel v\parallel_{Y_h}}, \tag{52}$$

where α is the constant of weak coercivity of a on $X_h\times Y_h$, that is $\alpha=\alpha_h$.

From standard approximation results (cf. [28]), there exists a constant C_I independent of h such that,

$$d_X(u,X_h) \leq C_I h \parallel u^{''} \parallel, \tag{53}$$

where $\parallel\cdot\parallel$ stands for the norm of $L^2(\Omega)$, that is, $\parallel [\cdot] \parallel := \left\{\int_\Omega [\cdot]^2\, dx\right\}^2$.

Using the Cauchy-Schwarz inequality, we have,

$$|L_h(v)-L(v)| = \left|\sum_{i=1}^{I_h+1}\int_{K_i}[f-f_h](x)[v_i+v_{i-1}]\, dx\right| \leq \left[\int_0^L [f-f_h]^2(x)\, dx\right]^{1/2}\left[\int_0^L v^2(x)\, dx\right]^{1/2}.$$

Now assuming that $f\in H^1(\Omega)$, from the same standard approximations results (see e.g. [7]), there exists a constant C_F independent of h such that $\left\{\int_0^L [f-f_h]^2(x)\, dx\right\}^{1/2} \leq C_F h \parallel f^{'} \parallel$. It follows that,

$$\sup_{v\in Y_h\setminus\{0_Y\}}\frac{L_h(v)-L(v)}{\parallel v\parallel_{Y_h}} \leq C_F h \parallel f^{'} \parallel . \tag{54}$$

We still have to estimate the *sup* term with numerator equal to $a(u,v)-L(v) = a_h(u,v) - \int_0^L fv\, dx$, for $v\in Y_h$. In this aim we first note that

$$\int_0^L [\sigma u - f]v \, dx = \int_0^L u^{''} v \, dx = \sum_{i=1}^{I_h} \int_{x_{i-1}}^{x_{i+1}} u^{''} v_i dx. \tag{55}$$

The assumption $f \in H^1(\Omega)$ allows us to legitimately assert that $u \in H^3(\Omega)$. Hence we may use the following identities, which can be proved by straightforward calculations:

$$\begin{cases} \displaystyle\int_{x_{i-1}}^{x_i} u^{'}(x)dx = \int_{x_{i-1}}^{x_i} [u^{'}(x_i) + (x - x_i)u^{''}(x_i) + \int_x^{x_i} u^{'''}(s)(s-x)ds]dx \\ \displaystyle\int_{x_i}^{x_{i+1}} u^{'}(x)dx = \int_{x_i}^{x_{i+1}} [u^{'}(x_i) + (x - x_i)u^{''}(x_i) - \int_{x_i}^{x} u^{'''}(s)(s-x)ds]dx. \end{cases} \tag{56}$$

Combining this with (55) we easily obtain,

$$\begin{aligned} a_h(u,v) - L(v) = \sum_{i=1}^{I_h} \Bigg\{ & \int_{x_{i-1}}^{x_{i+1}} u^{''}(x)dx - 2hu^{''}(x_i) \\ +2h^{-1} & \left[\int_{x_{i-1}}^{x_i} \int_x^{x_i} u^{'''}(s)(s-x)ds \, dx + \int_{x_i}^{x_{i+1}} \int_{x_i}^{x} u^{'''}(s)(s-x)ds \, dx \right] \Bigg\} v_i. \end{aligned} \tag{57}$$

But since $\int_{x_{i-1}}^{x_{i+1}} u^{''}(x)dx = 2hu^{''}(x_i) + \int_{x_{i-1}}^{x_{i+1}} \int_{x_i}^{x} u^{'''}(s)ds \, dx$, it follows from (57) that,

$$\begin{aligned} a_h(u,v) - L(v) \leq \sum_{i=1}^{I_h} \Bigg\{ & \int_{x_{i-1}}^{x_{i+1}} \int_{x_{i-1}}^{x_{i+1}} |u^{'''}(s)||v_i| ds \, dx \\ + & \left[\int_{x_{i-1}}^{x_i} \int_{x_{i-1}}^{x_i} |u^{'''}(s)||v_i| ds \, dx + 2 \int_{x_i}^{x_{i+1}} \int_{x_i}^{x_{i+1}} |u^{'''}(s)||v_i| ds \, dx \right] \Bigg\}. \end{aligned} \tag{58}$$

Now we adjust the range of i in the summation of the second term in brackets from $i = 2$ up to $i = I_h + 1$, taking into account that $v_0 = v_{I_h+1} = 0$. Combining this with the obvious relation $|v_i| + |v_{i-1}| \leq |v_i + v_{i-1}| + |v_i - v_{i-1}|L/h$, the Cauchy-Schwarz inequality leads to,

$$a_h(u,v) - L(v) \leq C_A h \parallel u^{'''} \parallel \parallel\!| \, v \parallel_{Y_h} . \tag{59}$$

for a suitable constant C_A independent of h.

For $f \in H^1(\Omega)$, collecting (52), (53), (54) and (59), we have thus proven the *error estimate*:

$$\parallel u - u_h \parallel_X \leq Ch[\parallel u^{''} \parallel + \parallel u^{'''} \parallel + \parallel f^{'} \parallel], \text{ with } C = \max[MC_I, C_F, C_A]/\alpha. \tag{60}$$

To complete this *example* let us briefly examine the case where σ is non constant. Here, in principle it is necessary to work with a polynomial approximation σ_h of σ. More precisely, assuming that $\sigma \in C^0(\bar{\Omega})$ we set $\sigma_h(x) = \sigma(G_{K_i}) \; \forall x \in K_i$, for $i = 1, 2, \ldots, I_h + 1$, and define the corresponding bilinear form $\tilde{a}_h \in \mathcal{L}_{2c}(X_h \times Y_h)$ by,

$$\tilde{a}_h(u,v) := 2 \sum_{i=1}^{I_h} [u^{'}(x_{i+1}^{-}) - u^{'}(x_{i-1}^{+})]v_i + \int_0^L \sigma_h uv \, dx. \tag{61}$$

An extension of $\tilde{a}_h$ to $X \times Y_h$ is defined in the same manner as the one of a_h. We also define $\bar{a}_h \in \mathcal{L}([X + X_h], Y)$ by replacing σ with σ_h in the expression of $\bar{a}$.

Accordingly we modify the approximate problem (36) into,

$$\begin{cases} \text{Find } \tilde{u}_h \in X_h \qquad \text{such that} \\ \tilde{a}_h(\tilde{u}_h, v) = L_h(v) \ \forall v \in Y_h. \end{cases} \tag{62}$$

We must also introduce a perturbed bilinear form $\tilde{a} \in \mathcal{L}_{2c}(Z \times T)$ given by $\tilde{a}(z,t) = \bar{a}_h(z,v)+\tilde{a}_h(z,v_h)$, where $z = u+u_h$ and $t = v+v_h$, with $u \in X$, $u_h \in X_h$, $v \in Y$ and $v_h \in Y_h$. It is not difficult to see that all the bounds that hold for a_h, similarly apply to $\tilde{a}_h$, if we replace the constant σ by σ_M in the upper bounds. In particular this yields constants $\tilde{M}_h$ and $\tilde{\alpha}_h$ that play the same role for $\tilde{a}_h$ as M_h and α_h do for a_h. Keeping the same definition of $a \in \mathcal{L}_{2c}(Z \times T)$ as before, we now denote the theoretical solution of (36) by u_h^*.

In fact the only issue really new here is that now we have to estimate another *sup* term in (33), namely,

$$\sup_{v \in Y_h \setminus \{0_T\}} \frac{[\tilde{a}-a](u_h^*,v)}{\parallel v \parallel_{Y_h}} = \sup_{v \in Y_h \setminus \{0_T\}} \frac{\int_0^L (\sigma_h - \sigma) u_h^* v \ dx}{\parallel v \parallel_{Y_h}} \leq \parallel \sigma_h - \sigma \parallel_\infty \parallel u_h^* \parallel, \tag{63}$$

where $\parallel g \parallel_\infty$ represents the standard maximum norm of a bounded function g in $\bar{\Omega}$, that is $\parallel g \parallel_\infty = \max_{x \in \bar{\Omega}} |g(x)|$. Assuming that $\sigma \in C^1(\bar{\Omega})$ again by standard approximation results (see e.g. [23]) we have $\parallel \sigma_h - \sigma \parallel_\infty \leq C_S h \parallel \sigma^{'} \parallel_\infty$, where C_S is a constant independent of h.

Adapting the analysis leading to (60) to the case of a variable σ, it can be shown that an estimate entirely analogous to (60) holds for u_h^*. More precisely, for a suitable constant C^* depending on L and σ_M but not on h we have,

$$\parallel u - u_h^* \parallel_X \leq C^* h[\parallel u^{''} \parallel + \parallel u^{'''} \parallel + \parallel f^{'} \parallel]. \tag{64}$$

From (64) it follows that $\parallel u_h^* \parallel \leq \parallel u \parallel + C^* h[\parallel u^{''} \parallel + \parallel u^{'''} \parallel + \parallel f^{'} \parallel]$. On the other hand $\forall v \in X$, $\parallel v^{'} \parallel^2 = -\int_0^L v v^{''} \ dx \leq \parallel v \parallel \parallel v^{''} \parallel$. Thus using again (49) we obtain $\parallel u \parallel \leq 4L^2 \parallel u^{''} \parallel$. This finally leads to estimate (65), where $\tilde{C}$ is a suitable constant depending on $L, \sigma_M, \parallel \sigma^{'} \parallel_\infty$ but not on h:

$$\parallel u - \tilde{u}_h \parallel_X \leq \tilde{C} h[\parallel u^{''} \parallel + \parallel u^{'''} \parallel + \parallel f^{'} \parallel]. \tag{65}$$

In short we can assert that, provided the data f and σ are sufficiently smooth, the approximate solution $\tilde{u}_h$ converges linearly to u in the natural norm $\parallel \cdot \parallel_X$, as h goes to zero.

Remark 3 The analysis carried out above was deliberately long. This is because we wanted to work in a very broad hilbertian framework, thereby showing how to handle methods that are not usually considered in a variational setting, such as the finite volume method. Indeed this leads to different exact and approximate bilinear forms and right hand side functionals, spaces not included in each other in the exact and the approximate problem, among other differences. Of course we could have simply compared the linear system corresponding to the approximate problem, to the one of the classical piecewise linear finite element method. Akin to the multi-dimensional case, we would have found that the latter is identical to the former, except for the numerical quadrature formulae employed to integrate non constant terms. However we avoided this line of argument, for we are persuaded that our strategy can serve as a guide to the analysis of other problems in future work, to which such a similarity does not apply. ■

Remark 4 Incidentally this example provides a noticeable physical interpretation of the classical piecewise linear finite element method. Indeed, as pointed out above, the underlying non standard finite volume scheme is roughly equivalent to such a finite element method. This means that the latter possesses the flux conservation property across control volumes, though overlapping ones, i.e. the Π_i's.■