Galerkin approximation of linear problems in Banach and Hilbert spaces

Wolfgang Arendt, Isabelle Chalendar, Robert Eymard

GALERKIN APPROXIMATION OF LINEAR PROBLEMS IN BANACH AND HILBERT SPACES

W. ARENDT, I. CHALENDAR, AND R. EYMARD

ABSTRACT. In this paper we study the conforming Galerkin approximation of the problem: find $u \in \mathcal{U}$ such that $a(u,v) = \langle L, v\rangle$ for all $v \in \mathcal{V}$, where $\mathcal{U}$ and $\mathcal{V}$ are Hilbert or Banach spaces, a is a continuous bilinear or sesquilinear form and $L \in \mathcal{V}'$ a given data. The approximate solution is sought in a finite dimensional subspace of $\mathcal{U}$, and test functions are taken in a finite dimensional subspace of $\mathcal{V}$. We provide a necessary and sufficient condition on the form a for convergence of the Galerkin approximation, which is also equivalent to convergence of the Galerkin approximation for the adjoint problem. We also characterize the fact that $\mathcal{U}$ has a finite dimensional Schauder decomposition in terms of properties related to the Galerkin approximation. In the case of Hilbert spaces, we prove that the only bilinear or sesquilinear forms for which any Galerkin approximation converges (this property is called the *universal Galerkin property*) are the *essentially coercive* forms. In this case, a generalization of the Aubin-Nitsche Theorem leads to optimal a priori estimates in terms of regularity properties of the right-hand side L, as shown by several applications. Finally, a section entitled "Supplement" provides some consequences of our results for the approximation of saddle point problems.

1. INTRODUCTION

Due to its practical importance, the approximation of elliptic problems in Banach or Hilbert spaces has been the object of numerous works. In Hilbert spaces, a crucial result is the simultaneous use of the Lax-Milgram theorem and of Céa's Lemma to conclude the convergence of conforming Galerkin methods in the case that the elliptic problem is resulting from a coercive bilinear or sesquilinear form.

But the coercivity property is lost in many practical situations: for example, consider the Laplace operator perturbed by a convection term or a reaction term (see the example in Section 7.2), and the approximation of non-coercive forms must be studied as well. For particular bilinear or sesquilinear forms, the Fredholm alternative provides an existence result in the case where the problem is well-posed in the Hadamard sense. Such results have been extended by Banach, Nečas, Babuška and Brezzi in the case of bilinear forms on Banach spaces. The conforming approximation of such problems enters into the framework of the so-called Petrov–Galerkin methods, for which sufficient conditions

2010 *Mathematics Subject Classification.* 65N30,47A07,47A52,46B20.

Key words and phrases. Galerkin approximation, sesquilinear coercive forms, approximation properties in Banach spaces, essential coercivity, universal Galerkin convergence.

for the convergence are classical (see for example the references [2, 8, 12, 31] which also include the case of non-conforming approximations).
Nevertheless, these sufficient conditions do not guarantee that for a given problem, there exists a converging Galerkin approximation. Moreover, they do not answer the following question, which is important in practice: under which conditions does the Galerkin approximation exist and converge to the solution of the continuous problem for *any* sufficiently fine approximation (for example, letting the degree of an approximating polynomial or the number of modes in a Fourier approximation be high enough, or letting the size of the mesh for a finite element method be small enough, and, in the case of Hilbert spaces, using the Galerkin method and not the Petrov–Galerkin method)?
The aim of this paper is precisely to address such questions for not necessarily coercive bilinear or sesquilinear forms defined on some Banach or Hilbert spaces (we treat the real and complex cases simultaneously). We shall restrict this study to conforming approximations, in the sense that the approximation will be sought in subspaces of the underlying space, using the continuous bilinear or sesquilinear form.
In the first part we consider the Banach space framework. Given a continuous bilinear form $a : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$ where $\mathcal{U}$ and $\mathcal{V}$ are reflexive, separable Banach spaces, one is interested in the existence and the convergence of the Galerkin approximation to u, where u is the solution of the following problem:

$$\text{Find } u \in \mathcal{U} \text{ such that } a(u, v) = \langle L, v \rangle, \text{ for all } v \in \mathcal{V}, \tag{1.1}$$

where $L \in \mathcal{V}'$ is given (the existence and uniqueness of u are obtained under the Banach-Nečas-Babuška conditions, see for example [12, Theorem 2.6]). For approximating sequences $(\mathcal{U}_n)_{n\in\mathbb{N}^*}$, $(\mathcal{V}_n)_{n\in\mathbb{N}^*}$ (see Section 2 for the definition), the Galerkin approximation of (1.1) is given by the sequence $(u_n)_{n\in\mathbb{N}^*}$ such that, for any $n \in \mathbb{N}^*$, u_n is the solution of the following finite dimensional linear problem:

$$\text{Find } u_n \in \mathcal{U}_n \text{ such that } a(u_n, \chi) = \langle L, \chi \rangle, \text{ for all } \chi \in \mathcal{V}_n. \tag{1.2}$$

It is known that, if $\dim \mathcal{U}_n = \dim \mathcal{V}_n$, the uniform Banach-Nečas-Babuška condition (BNB) given in Section 2 is sufficient for these existence and convergence properties (see for example [12, Theorem 2.24]). We show here that this condition is also necessary and, surprisingly, that the convergence of the Galerkin approximation of (1.1) is equivalent to that of the Galerkin approximation of the dual problem.
These two results seem to be new and are presented in Section 2.
In Section 3, we ask the following: given a form a such that (1.1) is well-posed, do there always exist approximating sequences in $\mathcal{U}$ and $\mathcal{V}$ such that the Galerkin approximation converges? Surprisingly, the answer is negative (even though the spaces $\mathcal{U}$ and $\mathcal{V}$ are supposed to be reflexive and separable). In fact, such approximating sequences exist if and only if the Banach space $\mathcal{U}$ has a finite dimensional Schauder decomposition, a property which is strictly more general than having a Schauder basis.
In the remainder of the paper, merely Hilbert spaces are considered and moreover we assume that $\mathcal{U} = \mathcal{V}$ and $\mathcal{U}_n = \mathcal{V}_n$ for all $n \in \mathbb{N}^*$. Given is a continuous bilinear form

$a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, where $\mathcal{V}$ is a separable Hilbert space. Assuming that (1.1) is well-posed, we show that the convergence of the Galerkin approximation for *all* approximating sequences in $\mathcal{V}$ (which we call here the *universal Galerkin property*) is equivalent to a being *essentially coercive*, which means that a compact perturbation of a is coercive. This notion of essential coercivity can also be characterized by a certain *weak-strong inverse continuity* of a, which, in fact, we take as definition of essential coercivity (Definition 4.2).
We then derive improved a priori error estimates by generalizing the Aubin–Nitsche argument to non-symmetric forms and also allowing the given right hand side L of (1.1) to belong to arbitrary interpolation spaces in between $\mathcal{V}$ and $\mathcal{V}'$. These generalizations are applied to two cases: the approximation of selfadjoint positive operators with compact resolvent (in this case, it is seen that our a priori error estimate is optimal, with the fastest speed of convergence for L in $\mathcal{V}$, the slowest for $L \in \mathcal{V}'$) and the finite element approximation of a non-selfadjoint elliptic differential operator, including convection and reaction terms which is indeed essentially coercive.
We finally give some further historical remarks in Section 8, where we consider saddle point problems. As a consequence of our results, we show that Brezzi's conditions, implying the convergence of mixed approximations (which are the Galerkin ones in the case of saddle point problems), are also necessary for this convergence.
To avoid any ambiguity, in the sequel, we let $\mathbb{N} = \{0, 1, 2, \cdots\}$ and $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$.

The paper is organized as follows:

Contents

2. Petrov–Galerkin approximation

In this section we give a characterization of the convergence of Petrov–Galerkin methods, that, for short, we call Galerkin convergence. A basic definition is the following.

Definition 2.1 (Approximating sequences of Banach spaces)**.** *Let $\mathcal{V}$ be a separable Banach space. An* approximating sequence *of $\mathcal{V}$ is a sequence $(\mathcal{V}_n)_{n\in\mathbb{N}^*}$ of finite dimensional*

subspaces of $\mathcal{V}$ *such that*

$$\operatorname{dist}(v, \mathcal{V}_n) \to 0 \text{ as } n \to \infty$$

for all $v \in \mathcal{V}$, *where* $\operatorname{dist}(u, \mathcal{V}_n) := \inf\{\|u - \chi\| : \chi \in \mathcal{V}_n\}$.

Now let $\mathcal{U}$ and $\mathcal{V}$ be two separable, reflexive Banach spaces over $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ and $a : \mathcal{U} \times \mathcal{V} \to \mathbb{K}$ be a continuous sesquilinear form such that

$$|a(u, v)| \leq M\|u\|_{\mathcal{U}}\|v\|_{\mathcal{V}} \text{ for all } u \in \mathcal{U}, v \in \mathcal{V}$$

where $M > 0$ is a constant. We assume that $\mathcal{U}$ and $\mathcal{V}$ are infinite dimensional and that $(\mathcal{U}_n)_{n\in\mathbb{N}^*}$ and $(\mathcal{V}_n)_{n\in\mathbb{N}^*}$ are approximating sequences of $\mathcal{U}$ and $\mathcal{V}$ respectively. We also assume throughout that

$$0 \neq \dim \mathcal{U}_n = \dim \mathcal{V}_n \text{ for all } n \in \mathbb{N}^*.$$

Given $L \in \mathcal{V}'$ we search a solution u of the problem:

$$\text{find } u \in \mathcal{U} \text{ such that } a(u, v) = \langle L, v\rangle, \text{ for all } v \in \mathcal{V}. \tag{2.1}$$

Moreover we want to approximate such a solution by u_n, the solution of the problem:

$$\text{find } u_n \in \mathcal{U}_n \text{ such that } a(u_n, \chi) = \langle L, \chi\rangle, \text{ for all } \chi \in \mathcal{V}_n. \tag{2.2}$$

Note that, given $n \in \mathbb{N}^*$, there exists a unique $u_n \in \mathcal{U}_n$ satisfying (2.2) if and only if

$$\text{for all } u \in \mathcal{U}_n, \big(a(u, \chi) = 0 \text{ for all } \chi \in \mathcal{V}_n\big) \Rightarrow u = 0, \tag{2.3}$$

since, by assumption, $\mathcal{U}_n$ and $\mathcal{V}_n$ have the same finite dimension.
Let us briefly recall the origin of the Banach-Nečas-Babuška conditions for the well-posedness of (2.1) as stated for example in [12, 31, 2] (equivalent conditions are proposed in [8] in the case of Hilbert spaces). Let us consider the associated operator $\mathcal{A} : \mathcal{U} \to \mathcal{V}'$ defined by

$$\langle \mathcal{A}u, v\rangle = a(u, v) \quad (u \in \mathcal{U}, v \in \mathcal{V}).$$

Then $\mathcal{A}$ is linear, bounded with $\|\mathcal{A}\| \leq M$. By the Inverse Mapping Theorem, $\mathcal{A}$ has closed range and is injective if and only if there exists $\beta > 0$ such that

$$\|\mathcal{A}u\|_{\mathcal{V}'} \geq \beta\|u\|_{\mathcal{U}} \text{ for all } u \in \mathcal{U}. \tag{2.4}$$

By the definition of the norm of $\mathcal{V}'$, this can be reformulated by

$$\sup_{\|v\|_{\mathcal{V}}\leq 1} |a(u, v)| \geq \beta\|u\|_{\mathcal{U}} \text{ for all } u \in \mathcal{U}. \tag{2.5}$$

Recall that $\mathcal{A}$ is invertible if and only if $\mathcal{A}$ is injective and has a closed and dense range. By the Hahn-Banach theorem, $\mathcal{A}$ has dense range if and only if no non-zero continuous functional on $\mathcal{V}'$ annihilates the range of $\mathcal{A}$. By reflexivity, this is equivalent to the following uniqueness property:

$$\text{for all } v \in \mathcal{V}, \big(a(u, v) = 0 \text{ for all } u \in \mathcal{U}\big) \Rightarrow v = 0. \tag{2.6}$$

Thus (2.1) is *well-posed* (i.e. for all $L \in \mathcal{V}'$ there exists a unique $u \in \mathcal{U}$ satisfying (2.1)) if and only if (2.5) and (2.6) are satisfied. In fact, Hadamard's definition of well-posedness

also requires continuity of the inverse operator, which here automatically follows from bijectivity by the Inverse Mapping Theorem.
In order to obtain a result of convergence of the approximate solutions we consider the following *uniform Banach-Nečas-Babuška condition* (called *Ladyzenskaia-Babuška-Brezzi condition* in the framework of the mixed formulations, i.e. approximation of saddle point problems, see also Section 8), which is the estimate (2.5) for $a_{|\mathcal{U}_n\times\mathcal{V}_n}$ uniformly in $n \in \mathbb{N}^*$, namely

$$(BNB) \quad \exists\beta > 0; \forall n \in \mathbb{N}^*,\ \forall u \in \mathcal{U}_n, \sup_{v\in\mathcal{V}_n, \|v\|_{\mathcal{V}}=1} |a(u,v)| \geq \beta\|u\|_{\mathcal{U}}.$$

Remark 2.2. *Condition (BNB) is also called the* inf–sup condition *since by the Hahn-Banach Theorem it can be reformulated as*

$$\exists\beta > 0; \forall n \in \mathbb{N}^*, \inf_{u\in\mathcal{U}_n, \|u|_V=1} \sup_{v\in\mathcal{V}_n, \|v\|_{\mathcal{V}}=1} |a(u,v)| \geq \beta.$$

More precisely, this is the uniform *or* discrete *BNB-condition which is used for approximation whereas (2.5) is the* continuous *BNB-condition which expresses well-posedness of the problem and can also be expressed by an inf-sup-condition (see for example* [15, Lemma 6.95 and Lemma 6.110]*). The use of (LBB) relates this inequality to the work of Ladyzhenkaya* [18] *who, after a previous contribution due to Babuska* [3]*, used it to prove well-posedness. Brezzi* [4] *introduced the analogue of the uniform BNB-condition for the treatment of saddle point problems (see Section 8 for more details).*
Usually, in the numerical analysis community, one uses the name "inf-sup" condition (or LBB condition) only in the context of saddle point problems (see condition (8.1).(*iii*) *in Section 8). We keep the name "(BNB) condition", following the monograph* [12].

We recall that (BNB) implies that the approximate solutions converge to the solution if the problem is well-posed (see for example [12, 31, 2]). Here we will show that (BNB) is actually equivalent to Galerkin-convergence, and surprisingly also to Galerkin-convergence for the dual problem.

Definition 2.3 (Convergence of Galerkin approximation)**.** *We say that* the Galerkin-approximation converges *if (2.1) as well as (2.2) are well-posed for all $n \in \mathbb{N}^*$ and $L \in \mathcal{V}'$ and if, in addition, there exists a constant $\gamma > 0$ independent of n and L such that,*

$$\|u - u_n\|_{\mathcal{U}} \leq \gamma \operatorname{dist}(u, \mathcal{U}_n), \tag{2.7}$$

where u is the solution of (2.1) and u_n the solution of (2.2) for $n \in \mathbb{N}^*$ and $L \in \mathcal{V}'$. In particular, $\lim_{n\to\infty} u_n = u$ in $\mathcal{U}$.

We may also consider the dual problem of (2.1) where a is replaced by the *adjoint form* $a^* : \mathcal{V} \times \mathcal{U} \to \mathbb{K}$ given by

$$a^*(v,u) = \overline{a(u,v)} \quad (u \in \mathcal{U}, v \in \mathcal{V}).$$

If in Definition 2.3 the form a is replaced by a^*, then we say that *the dual Galerkin approximation converges*. Similarly we note the following dual *uniform Banach-Nečas-Babuška condition*

$$(BNB^*) \quad \exists \beta^* > 0; \forall n \in \mathbb{N}^*, \sup_{u \in \mathcal{U}_n, \|u\|_{\mathcal{U}}=1} |a^*(u,v)| \geq \beta^* \|v\|_{\mathcal{V}} \quad (v \in \mathcal{V}_n).$$

Then the following theorem holds.

Theorem 2.4. *The following assertions are equivalent:*

(i) *the Galerkin approximation converges;*
(ii) (BNB) *holds;*
(iii) (BNB^*) *holds;*
(iv) *the dual Galerkin approximation converges.*

It is surprising that (BNB) and (BNB^*) are equivalent even though the corresponding condition (2.5) is obviously not equivalent to its dual form. In fact, it can well happen that $\mathcal{A}$ is injective and has closed range (so that there exists $\beta > 0$ satisfying (2.5)) but the range of $\mathcal{A}$ is a proper subspace of $\mathcal{V}'$ so that there exists $v \in \mathcal{V}$ such that $v \neq 0$ and $a(u,v) = 0$ for all $u \in \mathcal{U}$; in particular the dual form of (2.5) does not hold for any $\beta^* > 0$. We will give the proof of Theorem 2.4 in several steps which give partly even stronger results. At first we show that (ii) implies (i), where γ can even be expressed in terms of β and M. Although the proof of this result is classical (see for example [31, 12]), we provide it for the convenience of the reader, but also to establish the well-posedness of (2.1) which we did not assume. This will be important for the proof of Theorem 2.4 and for the main result in Section 5.

Proposition 2.5. *Let $\beta > 0$. Assume that for all $n \in \mathbb{N}^*$,*

$$\sup_{v \in \mathcal{V}_n, \|v\|_{\mathcal{V}}=1} |a(u,v)| \geq \beta \|u\|_{\mathcal{U}} \quad (u \in \mathcal{U}_n). \tag{2.8}$$

Then the Galerkin-approximation converges and (2.7) holds with

$$\gamma = 1 + \frac{M}{\beta}.$$

Proof. Let $L \in \mathcal{V}'$. Note that (2.8) implies (2.3). Thus, for each $n \in \mathbb{N}^*$ there exists a unique solution u_n of (2.2). By (2.8),

$$\|u_n\|_{\mathcal{U}} \leq \frac{1}{\beta} \sup_{v \in \mathcal{V}_n, \|v\|_{\mathcal{V}} \leq 1} |\langle L, v \rangle| \leq \frac{1}{\beta} \|L\|_{\mathcal{V}'}. \tag{2.9}$$

Since $\mathcal{U}$ is reflexive, we find $u \in \mathcal{U}$ such that a subsequence of $(u_n)_n$, say, $(u_{n_k})_k$, converges weakly to u. Let $v \in \mathcal{V}$. By assumption we find $v_k \in \mathcal{V}_{n_k}$ such that $\lim_{k\to\infty} \|v - v_{n_k}\|_{\mathcal{V}} = 0$. It follows that

$$a(u,v) = \lim_{k\to\infty} a(u_{n_k}, v_k) = \lim_{k\to\infty} \langle L, v_k \rangle = \langle L, v \rangle.$$

Thus we find a solution u of (2.1). But so far we do not know its uniqueness. This will be a consequence of (2.7) which we prove now. Indeed, observe that

$$a(u,\chi) = \langle L, \chi\rangle = a(u_n,\chi) \text{ for all } \chi \in \mathcal{V}_n. \tag{2.10}$$

It follows that $a(u,\chi) = a(u_n,\chi)$ for all $\chi \in \mathcal{V}_n$ (*Galerkin orthogonality*). Using this, for all $w \in \mathcal{U}_n$,

$$\begin{aligned}
\|u-u_n\|_{\mathcal{U}} &\leq \|u-w\|_{\mathcal{U}} + \|w-u_n\|_{\mathcal{U}} \\
&\leq \|u-w\|_{\mathcal{U}} + \frac{1}{\beta}\sup_{v\in\mathcal{V}_n,\|v\|_{\mathcal{V}}=1} |a(w-u_n,v)| \\
&= \|u-w\|_{\mathcal{U}} + \frac{1}{\beta}\sup_{v\in\mathcal{V}_n,\|v\|_{\mathcal{V}}=1} |a(w-u,v)| \\
&\leq \left(1+\frac{M}{\beta}\right)\|w-u\|_{\mathcal{U}}.
\end{aligned}$$

Taking the infimum over all $w \in \mathcal{U}_n$ we obtain (2.7). In particular $\lim_{n\to\infty}\|u-u_n\|_{\mathcal{U}} = 0$ which shows uniqueness. □

The following result is due to Xu and Zikatanov [31, Theorem 2] (see also [2, Satz 9.41]). We nevertheless provide its proof for the sake of completeness.

Proposition 2.6. *Assume that $\mathcal{U}$ is a Hilbert space and that $\beta > 0$ is such that (2.8) holds. Then the Galerkin-approximation converges and (2.7) holds with $\gamma = \frac{M}{\beta}$.*

Proof. Note that (2.8) implies (2.3). Consequently for each $w \in \mathcal{U}$ there exists a unique $Q_n w \in \mathcal{U}_n$ such that

$$a(Q_n w, \chi) = a(w,\chi) \quad (\chi \in \mathcal{V}_n).$$

Then Q_n is a projection from $\mathcal{U}$ onto $\mathcal{U}_n$, which is calle the *Ritz projection.* Moreover,

$$\begin{aligned}
\beta\|Q_n w\|_{\mathcal{U}} &\leq \sup_{\chi\in\mathcal{V}_n,\|\chi\|_{\mathcal{V}}=1} |a(Q_n w,\chi)| \\
&= \sup_{\chi\in\mathcal{V}_n,\|\chi\|_{\mathcal{V}}=1} |a(w,\chi)| \\
&\leq M\|w\|_{\mathcal{U}}.
\end{aligned}$$

Thus $\|Q_n\| \leq \frac{M}{\beta}$.
Since $U_n \neq 0$ and $U \neq 0$, one has $Q_n \neq 0, \mathrm{Id}$. It follows from a result due to Kato [17, Lemma 4] that $\|Q_n\| = \|\mathrm{Id} - Q_n\|$.
Now let $L \in \mathcal{V}'$ and u the solution of (2.3), u_n the solution of (2.2). Then for any $\chi \in \mathcal{U}_n$,

$$u - u_n = (\mathrm{Id} - Q_n)u = (\mathrm{Id} - Q_n)(u-\chi).$$

Hence

$$\|u-u_n\|_{\mathcal{U}} \leq \|\mathrm{Id} - Q_n\|\|u-\chi\|_{\mathcal{U}} = \|Q_n\|\|u-\chi\|_{\mathcal{U}} \leq \frac{M}{\beta}\|u-\chi\|_{\mathcal{U}}.$$

This implies that

$$\|u - u_n\|_{\mathcal{U}} \leq \frac{M}{\beta} \operatorname{dist}(u, \mathcal{U}_n).$$

□

Remark 2.7. *Also in certain Banach spaces an improvement of the constant* $1 + \frac{M}{\beta}$ *is possible, see Stern* [29].

Next we show that even a weaker assumption than the convergence of the Galerkin-approximation implies (BNB^*).

Proposition 2.8. *Assume (2.3) for all* $n \in \mathbb{N}^*$ *and that*

$$\sup_{n \in \mathbb{N}^*} \|u_n\|_{\mathcal{U}} < \infty$$

whenever $L \in \mathcal{V}'$ *and* u_n *is the solution of (2.2). Then* (BNB^*) *holds.*

Proof. Since the spaces $\mathcal{V}_n$ and $\mathcal{U}_n$ have the same finite dimension, our assumption (2.3) implies also dual uniqueness, i.e. $a(\chi, v) = 0$ for all $\chi \in \mathcal{U}_n$ implies $v = 0$ whenever $v \in \mathcal{V}_n$, and this for all $n \in \mathbb{N}^*$. Thus

$$\|v\|_{\mathcal{V}_n} := \sup_{u \in \mathcal{U}_n, \|u\|_{\mathcal{U}} = 1} |a(u, v)|$$

defines a norm on $\mathcal{V}_n$. Moreover,

$$|a(u, v)| \leq \|u\|_{\mathcal{U}} \|v\|_{\mathcal{V}_n} \text{ for all } u \in \mathcal{U}_n, v \in \mathcal{V}_n.$$

We show that the set

$$\mathcal{B} := \left\{ \frac{v}{\|v\|_{\mathcal{V}_n}} : n \in \mathbb{N}^*, v \in \mathcal{V}_n, v \neq 0 \right\}$$

is bounded. For that purpose, let $L \in \mathcal{V}'$. By assumption there exist $c > 0$ and $u_n \in \mathcal{U}_n$ such that

$$a(u_n, v) = \langle L, v \rangle \text{ for all } v \in \mathcal{V}_n$$

and $\|u_n\|_{\mathcal{U}} \leq c$ for all $n \in \mathbb{N}^*$. Now, for $\frac{v}{\|v\|_{\mathcal{V}_n}} \in \mathcal{B}$,

$$\left| \langle L, \frac{v}{\|v\|_{\mathcal{V}_n}} \rangle \right| = |a(u_n, v)| \frac{1}{\|v\|_{\mathcal{V}_n}} \leq \|u_n\|_{\mathcal{U}} \leq c.$$

This shows that $\mathcal{B}$ is weakly bounded and thus, owing to the Banach–Steinhaus theorem, norm-bounded. Therefore there exists $\beta^* > 0$ such that $\|v\|_{\mathcal{V}} \leq \frac{1}{\beta^*} \|v\|_{\mathcal{V}_n}$, i.e.

$$\beta^* \|v\|_{\mathcal{V}} \leq \sup_{u \in \mathcal{U}_n, \|u\|_{\mathcal{U}} = 1} |a(u, v)| \text{ for all } v \in \mathcal{V}_n, n \in \mathbb{N}^*.$$

This is (BNB^*). □

Proof of Theorem 2.4. $(ii) \Rightarrow (i)$ and $(iii) \Rightarrow (iv)$ via Proposition 2.5, whereas $(i) \Rightarrow (iii)$ and $(iv) \Rightarrow (ii)$ follows from Proposition 2.8.

□

Remark: The hypothesis on $\mathcal{U}$ and $\mathcal{V}$ to be reflexive is not needed in Proposition 2.5.

Finally we mention that the best lower bounds β for (BNB) and β^* for (BNB^*) are the same if $\mathcal{U}$ and $\mathcal{V}$ are Hilbert spaces.

Proposition 2.9. *Assuming that $\mathcal{U}$ and $\mathcal{V}$ are Hilbert spaces, let $\beta > 0$. Then the two conditions* (2.11) *and* (2.12) *are equivalent:*

$$\sup_{\|v\|_{\mathcal{V}}\leq 1, v\in\mathcal{V}_n} |a(u,v)| \geq \beta\|u\|_{\mathcal{U}} \text{ for all } u \in \mathcal{U}_n \text{ and for all } n \in \mathbb{N}^*; \tag{2.11}$$

$$\sup_{\|u\|_{\mathcal{U}}\leq 1, u\in\mathcal{U}_n} |a(u,v)| \geq \beta\|u\|_{\mathcal{U}} \text{ for all } v \in \mathcal{V}_n \text{ and for all } n \in \mathbb{N}^*; \tag{2.12}$$

Proof. Let $n \in \mathbb{N}^*$ and $A_n : \mathcal{U}_n \to \mathcal{V}_n$ be given by

$$\langle A_n u, v\rangle_{\mathcal{V}} = a(u,v).$$

Then

$$\langle A_n^* v, u\rangle_{\mathcal{U}} = a^*(v,u) = \overline{a(u,v)},$$

where A_n^* is the adjoint of A. Moreover, since A_n is invertible,

$$\sup_{\|v\|_{\mathcal{V}}=1, v\in\mathcal{V}_n} |a(u,v)| \geq \beta\|u\|_{\mathcal{U}}$$

for all $u \in \mathcal{U}_n$ if and only if $\|A_n^{-1}\| \leq \frac{1}{\beta}$. Since $(A_n^*)^{-1} = (A_n^{-1})^*$, it follows that $\|(A_n^*)^{-1}\| = \|(A_n^{-1})^*\| = \|A_n^{-1}\| \leq \frac{1}{\beta}$ and hence

$$\sup_{\|u\|_{\mathcal{U}}\leq 1, u\in\mathcal{U}_n} |a^*(v,u)| \geq \beta\|v\|_{\mathcal{V}} \text{ for all } v \in \mathcal{V}_n.$$

□

W. V. Petryshyn, namely in Theorem 2 and 3 of [22], considers approximation of an operator equation by finite dimensional problems and characterizes strong convergence. However, besides in very special situations, it sems not possible to deduce from this convergence of a Galerkin approximation, formulated in terms of sesquilinear forms. Further results for operator equations and their approximation can be found in the monograph [25, p. 26 ff].

3. Existence of a converging Galerkin approximation

In this section, we again let $\mathcal{U}$ and $\mathcal{V}$ be separable reflexive real Banach spaces and let $a : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$ be a continuous sesquilinear form such that the problem (2.1) is well-posed; i.e. for all $L \in \mathcal{V}'$ there exists a unique $u \in \mathcal{U}$ satisfying (2.1). Since $\mathcal{U}$ and $\mathcal{V}$ are separable, there always exist approximating sequences $(\mathcal{U}_n)_{n\in\mathbb{N}^*}$ of $\mathcal{U}$ and $(\mathcal{V}_n)_{n\in\mathbb{N}^*}$ of $\mathcal{V}$. Our question is whether there is a choice of these sequences which is adapted to the problem (2.1); i.e. such that the associated Galerkin approximation converges. We will show that the answer is related to the approximation property. In fact, different versions of this property play a role; we recall them in the next definition.

Definition 3.1 (Approximation property and Schauder decomposition)**.** *Let $\mathcal{X}$ be a separable Banach space.*

a) *The space $\mathcal{X}$ has the* approximation property *(AP) if, for every compact subset K of $\mathcal{X}$ and every $\varepsilon > 0$, there exists a finite rank operator $R \in \mathcal{L}(\mathcal{X})$ such that*

$$\|Rx - x\| < \varepsilon \text{ for all } x \in K.$$

b) *The space $\mathcal{X}$ has the* bounded approximation property *(BAP) if there exists a sequence $(P_n)_{n\in\mathbb{N}^*}$ of finite rank operators in $\mathcal{X}$ such that*

$$\text{for all } x \in \mathcal{X},\ \lim_{n\to\infty} P_n x = x.$$

c) *The space $\mathcal{X}$ has the* bounded projection approximation property *(BPAP) if each P_n in b) can be chosen as a projection (i.e. such that $P_n^2 = P_n$).*

d) *The space $\mathcal{X}$ possesses a* finite dimensional decomposition *if one finds $(P_n)_{n\in\mathbb{N}^*}$ as in c) with the additional property*

$$P_m P_n = P_n P_m = P_m \text{ for all } n \geq m. \tag{3.1}$$

e) *The space $\mathcal{X}$ has a* Schauder basis *if d) holds with*

$$\dim(P_n - P_{n-1})\mathcal{X} = 1 \text{ for all } n \in \mathbb{N}^*.$$

It is known that (BAP) is equivalent to (AP) if $\mathcal{X}$ is reflexive. The first counterexample of a Banach space without (AP) has been given by Enflo [11]. He constructed a space which is even separable and reflexive.
Obviously the properties a)–e) have decreasing generality. It was Read [26] who showed that (BAP) does not imply (BPAP), even if reflexive and separable spaces are considered. Szarek [30] constructed a reflexive, separable Banach space having a finite dimensional Schauder decompositon but not a Schauder basis. Finally, it seems to be unknown whether (BPAP) implies the existence of a finite dimensional Schauder decomposition (see [24, Sec. 5.7.4.6] and [7, Problem 6.2]). However, if $\mathcal{X}$ is reflexive and separable, then these two properties are equivalent by [7, Theorem 6.4 (3)]).
Concerning the notion of finite dimensional Schauder decomposition, there is an equivalent formulation, namely the existence of finite dimensional subspaces $\mathcal{X}_n$ of $\mathcal{X}$ such that for each $x \in \mathcal{X}$ there exist unique $x_n \in \mathcal{X}_n$ such that $x = \sum_{n\in\mathbb{N}^*} x_n$ This explains the name. We refer to [20, Chapter I] , [7] for more information and to [24, Sec. 5.7.4] for the history of the approximation property. In the following theorem, by the hypothesis of well-posedness, the two Banach spaces $\mathcal{U}$ and $\mathcal{V}$ are isomorphic. For this reason they have the same Banach space properties.

Theorem 3.2. *Let $\mathcal{U}$ and $\mathcal{V}$ be separable reflexive Banach spaces and let $a : \mathcal{U}\times\mathcal{V} \to \mathbb{K}$ be a continuous sesquilinear form such that (2.1) is well-posed. Then the following assertions are equivalent.*

(i) *There exist approximating sequences $(\mathcal{U}_n)_{n\in\mathbb{N}^*}$ of $\mathcal{U}$ and $(\mathcal{V}_n)_{n\in\mathbb{N}^*}$ of $\mathcal{V}$ such that the associated Galerkin approximation converges.*

(ii) *The space $\mathcal{U}$ has the (BPAP).*

(iii) *The space $\mathcal{U}$ has a finite dimensional Schauder decomposition.*

Here convergence of the associated Galerkin approximation is understood in the sense of Definition 2.3.

Proof of Theorem 3.2. $(i) \Rightarrow (ii)$ Let $u \in \mathcal{V}$. Then $\langle L, v\rangle := a(u, v)$ defines an element $L \in \mathcal{V}'$. By Definition 2.3, for each $n \in \mathbb{N}^*$, there exists a unique $P_n u \in \mathcal{V}_n$ such that

$$a(P_n u, \chi) = a(u, \chi) \text{ for all } \chi \in \mathcal{V}_n.$$

Moreover, $\|P_n u - u\| \leq \gamma \operatorname{dist}(\mathcal{U}_n, u)$ for all $n \in \mathbb{N}^*$ and some $\gamma > 0$. In particular, $\lim_{n\to\infty} P_n u = u$. It follows from the definition that $P_n^2 = P_n$. Since $P_n\mathcal{U} \subset \mathcal{U}_n$, each P_n has finite rank. We have shown that the space $\mathcal{U}$ has the (BPAP).
$(ii) \Rightarrow (iii)$ See [7, Theorem 6.4 (3)].
$(iii) \Rightarrow (i)$ Let $\mathcal{A} : \mathcal{U} \to \mathcal{V}'$ be the operator defined by $\langle \mathcal{A}u, v\rangle = a(u, v)$. Then $\mathcal{A}$ is invertible. By hypothesis there exist finite rank projections $(P_n)_{n\in\mathbb{N}^*}$ such that $\lim_{n\to\infty} P_n u = u$ for all $u \in \mathcal{U}$. Let $L \in \mathcal{V}'$, $u := \mathcal{A}^{-1}L$ be the solution of (2.1). Then

$$u_n := P_n\mathcal{A}^{-1}L \to u \text{ in } \mathcal{U} \text{ as } n \to \infty. \tag{3.2}$$

We show that u_n is obtained as a Galerkin approximation. In fact, fix $n \in \mathbb{N}^*$. There exist $b_1, \cdots, b_m \in \mathcal{U}$, $\varphi_1, \cdots, \varphi_m \in \mathcal{U}'$ such that $\langle \varphi_i, b_j\rangle = \delta_{i,j}$ and

$$P_n x = \sum_{k=1}^{m} \langle \varphi_k, x\rangle b_k \tag{3.3}$$

for all $x \in \mathcal{U}$. Since $\mathcal{V}$ is reflexive there exist $v_k \in \mathcal{V}$ such that

$$\langle \varphi_k, \mathcal{A}^{-1}g\rangle = \langle g, v_k\rangle \tag{3.4}$$

for all $g \in \mathcal{V}'$ and $k = 1, \cdots, m$. Define $\mathcal{V}_n = \operatorname{Span}\{v_1, \cdots, v_m\}$ and $\mathcal{U}_n = \operatorname{Span}\{b_1, \cdots, b_m\}$. Now consider the given $L \in \mathcal{V}'$. Let $w = \sum_{k=1}^{m} \lambda_k b_k \in \mathcal{U}_n$. Then

$$a(w, \chi) = \langle L, \chi\rangle \text{ for all } \chi \in \mathcal{V}_n \tag{3.5}$$

if and only if

$$a(w, v_j) = \langle L, v_j\rangle \text{ for } j = 1, \cdots, m. \tag{3.6}$$

By (3.4),

$$a(w, v_j) = \sum_{k=1}^{m} \lambda_k a(b_k, v_j) = \sum_{k=1}^{m} \lambda_k \langle \mathcal{A}b_k, v_j\rangle = \sum_{k=1}^{m} \lambda_k \langle \varphi_j, b_k\rangle = \lambda_j.$$

Therefore $w = \sum_{k=1}^{m} \langle L, v_k\rangle b_k$ is the unique solution of (3.5). Again, by (3.4),

$$u_n = P_n\mathcal{A}^{-1}L = \sum_{k=1}^{m} \langle \varphi_k, \mathcal{A}^{-1}L\rangle b_k = \sum_{k=1}^{m} \langle L, v_k\rangle b_k = w,$$

and it follows from (3.2) that $\lim_{n\to\infty} u_n = u$. This also implies that $\operatorname{dist}(\mathcal{U}_n, u) \to 0$ as $n \to \infty$. Thus the sequence $(\mathcal{U}_n)_{n\in\mathbb{N}^*}$ is approximating.

It remains to show that the sequence $(\mathcal{V}_n)_{n\in\mathbb{N}^*}$ is approximating in $\mathcal{V}$. For this we need the the additional property (3.1). Consider the adjoint $P_n' \in \mathcal{L}(\mathcal{U}')$ of P_n. Then $P_n'\varphi$ weakly converges to φ as $n \to \infty$ for all $\varphi \in \mathcal{U}'$. Thus

$$\mathcal{W} := \cup_{n\in\mathbb{N}^*} P_n'\mathcal{U}'$$

is weakly dense in $\mathcal{U}'$. But, because of (3.1), $\mathcal{W}$ is a subspace of $\mathcal{U}'$. Thus, by Mazur's Theorem, $\mathcal{W}$ is dense in $\mathcal{U}'$. If $\psi \in \mathcal{W}$, then there exist $m \in \mathbb{N}^*, \varphi \in \mathcal{U}'$ such that $\psi = P_m'\varphi$. Thus

$$P_n'\psi = P_n'P_m'\varphi = P_m'\varphi = \psi,$$

for all $n \in \mathbb{N}^*$ by (3.1), and then $\lim_{n\to\infty} P_n'\psi = \psi$ for all $\psi \in \mathcal{W}$. Since $\sup_{n\in\mathbb{N}^*} \|P_n'\| < \infty$, it follows that $\lim_{n\to\infty} P_n'\varphi = \varphi$ for all $\varphi \in \mathcal{U}'$. This implies that the sequence $(P_n'\mathcal{U}')_{n\in\mathbb{N}^*}$ is approximating in $\mathcal{U}'$. It follows from (3.4) that $\mathcal{V}_n \supset (\mathcal{A}^{-1})'P_n'\mathcal{U}'$. In fact, fix n and consider P_n as in (3.3). Then (3.4) says that $v_k = (\mathcal{A}^{-1})'\varphi_k$. Since $(P_n'\mathcal{U}')_{n\in\mathbb{N}^*}$ is an approximating sequence in $\mathcal{U}'$ and $(\mathcal{A}^{-1})'$ is an isomorphism from $\mathcal{U}'$ to $\mathcal{V}$, it follows that $(\mathcal{V}_n)_{n\in\mathbb{N}^*}$ is an approximating sequence in $\mathcal{V}$. □

4. Essentially coercive forms

Let $\mathcal{V}$ be a separable Hilbert space over $\mathbb{K} = \mathbb{C}$ or $\mathbb{R}$ and $a : \mathcal{V} \times \mathcal{V} \to \mathbb{K}$ be a sesquilinear form satisfying

$$|a(u, v)| \le M\|u\|_{\mathcal{V}}\|v\|_{\mathcal{V}} \text{ for all } u, v \in \mathcal{V}$$

for some $M > 0$. Then we may associate with a the operator $\mathcal{A} \in \mathcal{L}(\mathcal{V}, \mathcal{V}')$ defined by

$$\langle \mathcal{A}u, v\rangle = a(u, v).$$

If a is *coercive*, i.e. if

$$|a(u, u)| \ge \alpha\|u\|_{\mathcal{V}}^2 \quad (u \in \mathcal{V})$$

for some $\alpha > 0$, then $\mathcal{A}$ is invertible. This consequence is the well-known Lax-Milgram lemma.

Remark 4.1. *The notion of* coercivity *is not uniform in the literature. Ours is the natural hypothesis of the Lax-Milgram Lemma and is conform with the Wikipedia entry "Babuska-Lax-MilgramTheorem". In non-linear analysis there is a wide agreement on this notion: In the real case, a possibly non-linear operator $\mathcal{A} \in \mathcal{L}(\mathcal{V}, \mathcal{V}')$ is called* coercive *if there exists a function $\eta : \mathbb{R} \to \mathbb{R}$ such that $\eta(t) \to \infty$ as $t \to \infty$ and $\langle \mathcal{A}v, v\rangle \ge \eta(\|u\|_{\mathcal{V}})\|v\|_{\mathcal{V}}$ for all $v \in \mathcal{V}$. If $\mathcal{A}$ is linear this is equivalent to the existence of $\alpha > 0$ such that*

$$\langle \mathcal{A}u, u\rangle \ge \alpha\|u\|_{\mathcal{V}}^2 \quad (u \in \mathcal{V}),$$

i.e. our condition without the absolute value. This is a "forcing condition" which justifies the name coercive*. Other authors prefer the word $\mathcal{V}-$*ellipticity*, see e.g.* [15], [21]*. We use* elliptic *for* shifted coercivity *in* [1]*, see also the remark at the end of this section.*

Our aim is to find weaker assumptions than coercivity which help to decide whether the operator $\mathcal{A}$ is invertible.
Note that a is coercive if and only if

$$\lim_{n\to\infty} a(u_n,u_n)=0 \text{ implies that } \lim_{n\to\infty}\|u_n\|_{\mathcal{V}}=0.$$

We weaken this property in the following way.

Definition 4.2 (Essential coercivity)**.** *The continuous sesquilinear form a (or the operator $\mathcal{A}$) is called essentially coercive if for each sequence $(u_n)_{n\in\mathbb{N}^*}$ in $\mathcal{V}$ weakly converging to 0 and such that $\lim_{n\to\infty} a(u_n,u_n)=0$, one has $\lim_{n\to\infty}\|u_n\|_{\mathcal{V}}=0$.*

The following is a characterization of this new property.

Theorem 4.3. *The following assertions are equivalent:*

(i) *the form a is essentially coercive;*
(ii) *there exist an orthogonal projection $P\in\mathcal{L}(\mathcal{V})$ of finite rank and $\alpha>0$ such that*
$$|a(u,u)|+\|Pu\|_{\mathcal{V}}^2\geq\alpha\|u\|_{\mathcal{V}}^2 \text{ for all } u\in\mathcal{V};$$
(iii) *there exist a Hilbert space $\mathcal{H}$, a compact operator $J:\mathcal{V}\to\mathcal{H}$ and $\alpha>0$ such that*
$$|a(u,u)|+\|Ju\|_{\mathcal{H}}^2\geq\alpha\|u\|_{\mathcal{V}}^2 \quad (u\in\mathcal{V});$$
(iv) *there exist a compact operator $\mathcal{K}\in\mathcal{L}(\mathcal{V},\mathcal{V}')$ and $\alpha>0$ such that*
$$|a(u,u)|+|\langle\mathcal{K}u,u\rangle|\geq\alpha\|u\|_{\mathcal{V}}^2 \quad (u\in\mathcal{V}).$$

Proof. $(i)\Rightarrow(ii)$: Let $(e_n)_{n\in\mathbb{N}^*}$ be an orthonormal basis of $\mathcal{V}$ and consider the orthogonal projections P_n given by

$$P_nv:=\sum_{k=1}^{n}\langle v,e_k\rangle_{\mathcal{V}}\,e_k.$$

Assume that (ii) is false for every P_n. Then there exists a sequence $(u_n)_{n\in\mathbb{N}^*}\subset\mathcal{V}$ such that $\|u_n\|_{\mathcal{V}}=1$ and

$$|a(u_n,u_n)|+\|P_nu_n\|_{\mathcal{V}}^2<\frac{1}{n}.$$

Note that, since $\mathrm{Id}-P_n$ is a self-adjoint operator,

$$|\langle(\mathrm{Id}-P_n)u_n,v\rangle_{\mathcal{V}}|=|\langle u_n,(\mathrm{Id}-P_n)v\rangle_{\mathcal{V}}|\leq\|(\mathrm{Id}-P_n)v\|_{\mathcal{V}},$$

with $\lim_{n\to\infty}\|(\mathrm{Id}-P_n)v\|_{\mathcal{V}}=0$ for all $v\in\mathcal{V}$. This implies that $(\mathrm{Id}-P_n)u_n$ converges weakly to 0. Since $\lim_{n\to\infty}\|P_nu_n\|_{\mathcal{V}}=0$, it follows that u_n converges weakly to 0. Moreover $\lim_{n\to\infty}|a(u_n,u_n)|\leq\lim_{n\to\infty}\frac{1}{n}=0$. Therefore a is not essentially coercive.
$(ii)\Rightarrow(iii)$: Choose $\mathcal{H}=\mathcal{V}$ and $J=P$.
$(iii)\Rightarrow(iv)$: There exists a unique operator $J^*:\mathcal{H}\to\mathcal{V}'$ such that

$$\langle J^*u,v\rangle=\langle u,Jv\rangle_{\mathcal{H}}$$

for all $v\in\mathcal{V}$. Choose $\mathcal{K}=J^*J$.
$(iv)\Rightarrow(i)$: Let $(u_n)_{n\in\mathbb{N}^*}\subset\mathcal{V}$ that tends weakly to 0 and such that $a(u_n,u_n)=\langle\mathcal{A}u_n,u_n\rangle$

tends to 0 as $n \to \infty$. Since $\mathcal{K}$ is compact, $\|\mathcal{K}u_n\|_{\mathcal{V}} \to 0$ as $n \to \infty$. Hence $|\langle \mathcal{K}u_n, u_n\rangle_{\mathcal{V}}| \to 0$ as $n \to \infty$. By assumption there exists $\beta > 0$ such that

$$|\langle \mathcal{A}u_n, u_n\rangle| + |\langle \mathcal{K}u_n, u_n\rangle| \geq \beta \|u_n\|_{\mathcal{V}}^2.$$

It follows that $\|u_n\|_{\mathcal{V}} \to 0$ as $n \to \infty$. □

Next we want to justify the notion "essentially coercive". We recall that by the Toeplitz–Hausdorff theorem [14], the *numerical range* of a,

$$W(a) := \{a(u,u) : u \in \mathcal{V}, \|u\|_{\mathcal{V}} = 1\},$$

is a convex set. Hence also $\overline{W(a)}$ is convex. For $\alpha > 0$,

$$|a(u,u)| \geq \alpha \|u\|_{\mathcal{V}}^2 \quad (u \in \mathcal{V})$$

if and only if

$$\overline{W(A)} \cap D_\alpha = \emptyset,$$

where $D_\alpha = (-\alpha, \alpha)$ in the real case and $D_\alpha = \{w \in \mathbb{C} : |w| < \alpha\}$ if $\mathbb{K} = \mathbb{C}$. This observation leads to the following more precise description of coercivity.

Lemma 4.4. *The form a is coercive if and only if there exist $\alpha > 0$ and $\lambda \in \mathbb{K}$ with $|\lambda| = 1$ such that*

$$Re(\lambda z) \geq \alpha \text{ for all } z \in W(a).$$

Proof. We give the proof for $\mathbb{K} = \mathbb{C}$. Assume that a is coercive. There exists a maximal $\alpha > 0$ such that $\overline{W(a)} \cap D_a = \emptyset$. Then there exists $z_0 \in \overline{W(a)}$ of modulus α; i.e. $z_0 = e^{i\theta}\alpha$ for some $\theta \in \mathbb{R}$. The set $C := e^{-i\theta}\overline{W(a)}$ is convex and closed. Moreover $\alpha \in C$ and $D_\alpha \cap C = \emptyset$. This implies that $Re(z) \geq \alpha$ for all $z \in C$. Indeed, let $z \in C$ such that $Re(z) < \alpha$. Then the segment $[\alpha, z]$ has a non-empty intersection with D_α. Since C is convex it follows that $z \notin C$.
Conversely, clearly, if there exists $\alpha > 0$ such that $Re(\lambda z) \geq \alpha$ for all $z \in W(a)$, then a is coercive. □

Theorem 4.5. *Let $\mathcal{A} \in \mathcal{L}(\mathcal{V}, \mathcal{V}')$. The following assertions are equivalent:*

(i) *the operator $\mathcal{A}$ is essentially coercive;*
(ii) *there exists a finite rank operator $\mathcal{K} : \mathcal{V} \to \mathcal{V}'$ such that $\mathcal{A} + \mathcal{K}$ is coercive;*
(iii) *there exists a compact operator $\mathcal{K} : \mathcal{V} \to \mathcal{V}'$ such that $\mathcal{A} + \mathcal{K}$ is coercive.*

Proof. $(i) \Rightarrow (ii)$: Choose the orthogonal finite rank projection P on $\mathcal{V}$ and $\alpha > 0$ as in Theorem 4.3 (ii). Let $\mathcal{V}_1 = \ker V$ and $\mathcal{V}_2 = range\, P$. Then $\dim \mathcal{V}_2 < \infty$ and $|a(u,u)| \geq \alpha \|u\|_{\mathcal{V}}^2$ for all $u \in \mathcal{V}_1$. Let $j : \mathcal{V} \to \mathcal{V}'$ be the Riesz isomorphism given by

$$\langle j(u), v\rangle = \langle u, v\rangle_{\mathcal{V}}.$$

Let $A = j^{-1} \circ \mathcal{A} \in \mathcal{L}(\mathcal{V})$. Then $a(u,v) = \langle Au, v\rangle_{\mathcal{V}}$ for all $u, v \in \mathcal{V}$. Moreover A has a matrix decomposition

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

according to the decomposition $\mathcal{V} = \mathcal{V}_1 \oplus \mathcal{V}_2$ of $\mathcal{V}$. Since P is orthogonal, A_{11} is coercive. Thus, by Lemma 4.4, there exists $z_0 \in \mathbb{C}$ such that $|z_0| = 1$ and

$$Re\, z_0 \langle A_{11} u, u\rangle \geq \alpha \|u\|_{\mathcal{V}}^2$$

for all $u \in \mathcal{V}_1$. Since $\dim \mathcal{V}_2 < \infty$, there exists a finite rank operator $K_1 \in \mathcal{L}(\mathcal{V})$ such that

$$A + K_1 = \begin{pmatrix} A_{11} & 0 \\ 0 & 0 \end{pmatrix}.$$

Choose a further finite rank perturbation K_2 such that

$$B := A + K_1 + K_2 = \begin{pmatrix} A_{11} & 0 \\ 0 & \alpha\overline{z_0}\,\mathrm{Id}_{\mathcal{V}_2} \end{pmatrix}.$$

Since P is orthogonal, for $Q = \mathrm{Id} - P$, we get

$$\langle Bu, u\rangle_{\mathcal{V}} = \langle A_{11} Qu, Qu\rangle_{\mathcal{V}} + \alpha\overline{z_0}\langle Pu, Pu\rangle_{\mathcal{V}}.$$

Hence

$$Re\langle z_0 Bu, u\rangle_{\mathcal{V}} \geq \alpha \|Qu\|_{\mathcal{V}}^2 + \alpha \|Pu\|_{\mathcal{V}}^2 = \alpha \|u\|_{\mathcal{V}}^2.$$

Now let $\mathcal{K} = j \circ (K_1 + K_2)$. Then $\mathcal{A} + \mathcal{K}$ is coercive.
$(ii) \Rightarrow (iii)$ is obvious.
$(iii) \Rightarrow (i)$: Condition (iii) implies clearly Condition (iv) of Theorem 4.3; thus the claim (i) follows from that theorem. □

Corollary 4.6. *Let a be a continuous essentially coercive sesquilinear form. The following assertions are equivalent:*

(i) *for all $L \in \mathcal{V}'$ there exists a unique $u \in \mathcal{V}$ such that*

$$a(u, v) = \langle L, v\rangle \text{ for all } v \in \mathcal{V};$$

(ii) $a(u, v) = 0$ *for all $v \in \mathcal{V}$ implies that $u = 0$ (*uniqueness*);*
(iii) *for all $L \in \mathcal{V}'$ there exists $u \in \mathcal{V}$ such that $a(u, v) = \langle L, v\rangle$ for all $v \in \mathcal{V}$ (*existence*).*

Proof. The assertion (i) means that $\mathcal{A}$ is invertible, the assertion (ii) means that $\mathcal{A}$ is injective and the assertion (iii) means that $\mathcal{A}$ is surjective. By Theorem 4.5, there exists a compact operator $\mathcal{K} \in \mathcal{L}(\mathcal{V}, \mathcal{V}')$ such that $\mathcal{A} + \mathcal{K} =: \mathcal{B}$ is invertible.
$(ii) \Rightarrow (i)$: Assume that $\mathcal{A}$ is injective. Write

$$\mathcal{A} = \mathcal{B} - \mathcal{K} = \mathcal{B}(\mathrm{Id} - \mathcal{B}^{-1}\mathcal{K}).$$

Then also $(\mathrm{Id} - \mathcal{B}^{-1}\mathcal{K})$ is injective. Since $\mathcal{B}^{-1}\mathcal{K}$ is compact, it follows from the classical Fredholm alternative that $(\mathrm{Id} - \mathcal{B}^{-1}\mathcal{K})$ is invertible. Consequently also $\mathcal{A}$ is invertible.
$(iii) \Rightarrow (i)$: If $\mathcal{A}$ is surjective, write $\mathcal{A} = (\mathrm{Id} - \mathcal{K}\mathcal{B}^{-1})\mathcal{B}$ to conclude that $(\mathrm{Id} - \mathcal{K}\mathcal{B}^{-1})$ is surjective. Again we deduce that $(\mathrm{Id} - \mathcal{K}\mathcal{B}^{-1})$ is invertible and so is $\mathcal{A}$. □

Remark 4.7. *In the previous corollary we deduced from Theorem 4.5 the Fredholm alternative. This conclusion is well-known, if a compact perturbation is given, see for example* [32, Theorem 22.D]*, or* [15, Lemma 6.108]*. Our point is that a priori it is not at all clear that the topological condition defining essential coercivity implies that the form is a compact perturbation of a coercive form. This is what Theorem 4.5 shows. Note that,*

in [23, p229]*, our notion of essential coercivity is attributed, under the name "condition (S)", to Felix Browder* [6] *if we identify the operator with a form.*

Moreover, we deduce from Theorem 4.5 the following properties of essential coercivity.

Corollary 4.8. (a) *The set of all essentially coercive operators on $\mathcal{V}$ is open in $\mathcal{L}(\mathcal{V}, \mathcal{V}')$.*
(b) *If $\mathcal{A} \in \mathcal{L}(\mathcal{V}, \mathcal{V}')$ is essentially coercive and $\mathcal{K} \in \mathcal{L}(\mathcal{V}, \mathcal{V}')$ is compact, then $\mathcal{A} + \mathcal{K}$ is essentially coercive.*
(c) *If $\mathcal{A} \in \mathcal{L}(\mathcal{V}, \mathcal{V}')$ is essentially coercive, then $\mathcal{A}$ is a Fredholm operator of index 0.*

The following example shows that the invertibility of $\mathcal{A}$ does not imply the essential coercivity of a.

Example 4.9. *Let $\mathcal{V} = \ell^2(\mathbb{N}^*)$, $\mathbb{K} = \mathbb{R}$ and*

$$a(u, v) = \sum_{n=0}^{\infty} (-1)^n u_n v_n.$$

Let j be the Riesz isomorphism introduced in the proof of Theorem 4.5. Then $A := j^{-1} \circ \mathcal{A}$ is a diagonal operator with merely 1 and -1 in the diagonal. Thus A and obviously $\mathcal{A}$ are clearly invertible. Let $f_n = (0, \cdots, 1, 1, 0, \cdots)$ where the 1 is a coordinate for $k = 2n$ and $k = 2n + 1$. Then $\|f_n\| = \sqrt{2}$ and $(f_n)_n$ tends weakly to 0 as $n \to \infty$. Moreover $a(f_n, f_n) = 0$ for all n, which shows that a is not essentially coercive.

Remark 4.10. *Let $\mathbb{K} = \mathbb{C}$. In* [1] *a continuous sesquilinear form a is called* compactly elliptic *if there exists a compact operator $J : \mathcal{V} \to \mathcal{H}$, where $\mathcal{H}$ is some Hilbert space and there exists $\alpha > 0$ such that*

$$Re\, a(u, u) + \|Ju\|_{\mathcal{H}}^2 \geq \alpha \|u\|_{\mathcal{V}}^2.$$

In view of Theorem 4.3, each compactly elliptic form is essentially coercive. In fact the following holds: the form a is essentially coercive if and only if there exists $\lambda \in \mathbb{C} \setminus \{0\}$ such that λa is compactly elliptic.

Proof. If λa is compactly elliptic, then λa is essentially coercive and hence also a is essentially coercive. Conversely, let a be essentially coercive. By Theorem 4.5, there exists a compact operator $\mathcal{K} : \mathcal{V} \to \mathcal{V}'$ such that the form b defined by

$$b(u, v) = a(u, v) + \langle \mathcal{K}u, v \rangle$$

is coercive. By Lemma 4.4 there exist $\lambda \in \mathbb{C}$ of modulus one and $\alpha > 0$ such that $Re(\lambda b(u, u)) \geq \alpha \|u\|_{\mathcal{V}}^2$ for all $u \in \mathcal{V}$. Now let $j : \mathcal{V} \to \mathcal{V}'$ be the Riesz isomorphism. Then $J := j^{-1} \circ \mathcal{K} : \mathcal{V} \to \mathcal{V}$ is compact. Choosing $\mathcal{H} = \mathcal{V}$ we see that λb is compactly elliptic. It follows from [1, Proposition 4.4 (b)] that λa is compactly elliptic. □

5. Characterization of the universal Galerkin property

In this section we want to characterize those forms on a Hilbert space for which every Galerkin approximation converges, whatever be the choice of the approximating sequence.

Let $\mathcal{V}$ be a separable, infinite dimensional separable Hilbert space over $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, and let $a : \mathcal{V} \times \mathcal{V} \to \mathbb{K}$ be a continuous sesquilinear form. Given $L \in \mathcal{V}'$ we again consider solutions of the problem:

$$\text{Find } u \in \mathcal{V},\ a(u, v) = \langle L, v\rangle \text{ for all } v \in \mathcal{V}. \tag{5.1}$$

We say that the form a satisfies *uniqueness* if for $u \in \mathcal{V}$,

$$a(u, v) = 0 \text{ for all } v \in \mathcal{V} \text{ implies } u = 0.$$

We say that (5.1) is *well-posed* if for all $L \in \mathcal{V}'$ there exists a unique solution $u \in \mathcal{V}$.

Definition 5.1 (Universal Galerkin property)**.** *The sesquilinear and continuous form a has the* universal Galerkin property *if (5.1) is well-posed and the following holds. Let $(\mathcal{V}_n)_{n\in\mathbb{N}^*}$ be an arbitrary approximating sequence of $\mathcal{V}$. Then there exist $n_0 \in \mathbb{N}^*$ and $\gamma > 0$ such that for each $L \in \mathcal{V}'$ and each $n \geq n_0$, there exists a unique $u_n \in \mathcal{V}_n$ solving*

$$a(u_n, \chi) = \langle L, \chi\rangle \text{ for all } \chi \in \mathcal{V}_n,$$

and

$$\|u - u_n\|_{\mathcal{V}} \leq \gamma \operatorname{dist}(u, \mathcal{V}_n) \text{ for all } n \geq n_0,$$

where u is the solution of (5.1).

As recalled in the introduction and in the preceding section, the Lax-Milgram Theorem and Céa's Lemma imply the universal Galerkin property if a is coercive. We now show that the weaker notion of essential coercivity also provides a sufficient condition for ensuring the universal Galerkin property, and moreover that it is necessary.

Theorem 5.2. *The following assertions are equivalent.*

(i) *The form a is essentially coercive and satisfies uniqueness.*
(ii) *The form a has the universal Galerkin property.*

Proof. $(i) \Rightarrow (ii)$: let $(\mathcal{V}_n)_{n\in\mathbb{N}^*}$ be an approximating sequence in $\mathcal{V}$. By Theorem 2.4 it suffices to show that there exist $\beta > 0$ and $n_0 \in \mathbb{N}^*$ such that

$$\sup_{v\in\mathcal{V}_n, \|v\|_{\mathcal{V}}=1} |a(u, v)| \geq \beta\|u\|_{\mathcal{V}} \text{ for all } u \in \mathcal{V}_n, n \geq n_0. \tag{5.2}$$

Assume that (5.2) is false. We then find a subsequence $(n_k)_{k\in\mathbb{N}^*}$ and $u_{n_k} \in \mathcal{V}_{n_k}$ such that $\|u_{n_k}\|_{\mathcal{V}} = 1$ and

$$\sup_{v\in\mathcal{V}_{n_k}, \|v\|_{\mathcal{V}}=1} |a(u_{n_k}, v)| < \frac{1}{k} \text{ for all } k \in \mathbb{N}^*.$$

We may assume that $(u_{n_k})_k$ converges weakly to u taking a further subsequence otherwise. Let $v \in \mathcal{V}$. Then there exist $v_k \in \mathcal{V}_{n_k}$ such that $\lim_{k\to\infty} \|v - v_k\|_{\mathcal{V}} = 0$. Thus

$$a(u, v) = \lim_{k\to\infty} a(u_{n_k}, v_k) = 0.$$

It follows from the uniqueness assumption that $u = 0$. Thus $(u_{n_k})_k$ converges weakly to 0, $\lim_{k\to\infty} a(u_{n_k}, u_{n_k}) = 0$, but $\|u_{n_k}\|_{\mathcal{V}} = 1$ for all k. Therefore the form a is not essentially coercive.

$(ii) \Rightarrow (i)$: the uniqueness condition is part of (ii). It remains to show that a is essentially coercive. Let $(e_n)_{n\in\mathbb{N}^*}$ be an orthonormal basis of $\mathcal{V}$ and $\mathcal{V}_n := \mathrm{Span}\{e_1, \cdots, e_n\}$. By our assumption, there exist $2 \leq n_0 \in \mathbb{N}^*$ and for all $n \geq n_0$ an operator $Q_n : \mathcal{V} \to \mathcal{V}_n$ such that

$$a(Q_n u, \chi) = a(u, \chi) \text{ for all } \chi \in \mathcal{V}_n \quad (n \geq n_0).$$

Denote by $P_n : \mathcal{V} \to \mathcal{V}_n$ the orthogonal projection. Define the operator

$$J_n : \mathcal{V} \to \mathcal{V} \times \mathcal{V}$$

by

$$J_n u = (P_n u, Q_n u), \quad n \geq n_0.$$

Now assume that a is not essentially coercive. Then it follows from Theorem 4.3 that for all $n \geq n_0$ we find $u_n \in \mathcal{V}$ such that $\|u_n\|_{\mathcal{V}} = 1$ and

$$|a(u_n, u_n)| + \|P_n u_n\|_{\mathcal{V}}^2 + \|Q_n u_n\|_{\mathcal{V}}^2 < \frac{1}{(n+2)^2}.$$

In particular $\|P_n u_n\|_{\mathcal{V}} < \frac{1}{(n+2)^2}$. This implies that $u_n \notin \mathcal{V}_n$. Let $\tilde{\mathcal{V}}_n = \mathrm{Span}\{\mathcal{V}_n \cup \{u_n\}\}$. Then $(\mathcal{V}_n)_{n\geq n_0}$ and $(\tilde{\mathcal{V}}_n)_{n\geq n_0}$ are both approximating sequences. Let $n \geq n_0$ and let $v \in \tilde{\mathcal{V}}_n$ be arbitrary with unit norm. There exist a unique $w_1 \in \mathcal{V}_n$ and $\lambda \in \mathbb{K}$ such that

$$v = w_1 + \lambda u_n = w + \lambda(u_n - P_n u_n),$$

where $w := w_1 + \lambda P_n u_n \in \mathcal{V}_n$. Thus

$$1 = \|v\|_{\mathcal{V}}^2 = \|w\|_{\mathcal{V}}^2 + |\lambda|^2 \|u_n - P_n u_n\|_{\mathcal{V}}^2.$$

Consequently $\|w\|_{\mathcal{V}}^2 \leq 1$ and, since $\|P_n u_n\|_{\mathcal{V}} < \frac{1}{2}$, it follows that

$$\|u_n - P_n u_n\|_{\mathcal{V}} \geq \frac{1}{2},$$

which implies that $|\lambda|^2 \leq 4$, i.e. $|\lambda| \leq 2$.
Observe that the definition of Q_n implies that $a(u_n - Q_n u_n, w) = 0$. Hence

$$\begin{aligned} |a(u_n, v)| &= |a(u_n, w) + \lambda a(u_n, u_n - P_n u_n)| \\ &= |a(u_n - Q_n u_n, w) + a(Q_n u_n, w) + \lambda a(u_n, u_n - P_n u_n)| \\ &\leq |a(Q_n u_n, w)| + 2|a(u_n, u_n)| + 2|a(u_n, P_n u_n)| \\ &\leq \frac{M}{n+2} + \frac{2}{(n+2)^2} + \frac{2M}{(n+2)^2}. \end{aligned}$$

Consequently

$$\lim_{n\to\infty} \sup_{v\in\tilde{V}_n, \|v\|_{\mathcal{V}}=1} |a(u_n, v)| = 0.$$

Thus (BNB) is violated for the approximating sequence $(\tilde{\mathcal{V}}_n)_{n\geq n_0}$. But then (ii) does not hold by Theorem 2.4, which shows that the assumption that a is not essentially coercive is false.

□

It is obvious that a form a is essentially coercive if and only if its adjoint a^* is essentially coercive. However, a surprising consequence of Theorem 5.2 is that, for an essentially coercive form, uniqueness for the form and uniqueness for its adjoint are equivalent, as the following corollary shows.

Corollary 5.3. *Let $\mathcal{V}$ be a separable Hilbert space on $\mathbb{K}$ and $a : \mathcal{V}\times\mathcal{V} \to \mathbb{K}$ be a continuous essentially coercive form. The following assertions are equivalent:*

(i) *for all $u \in \mathcal{V}$, $a(u, v) = 0$ for all $v \in \mathcal{V}$ implies $u = 0$;*
(ii) *for all $v \in \mathcal{V}$, $a(u, v) = 0$ for all $u \in \mathcal{V}$ implies $v = 0$;*
(iii) *for all $L \in \mathcal{V}'$ there exists u in $\mathcal{V}$ such that $a(u, v) = \overline{\langle L, v\rangle}$, for all $v \in \mathcal{V}$;*
(iv) *for all $L \in \mathcal{V}'$ there exists v in $\mathcal{V}$ such that $a(u, v) = \overline{\langle L, u\rangle}$, for all $u \in \mathcal{V}$.*

Proof. $(i) \iff (ii)$: this follows from Theorem 5.2 and Theorem 2.4. The other equivalences follow from Corollary 4.6. □

6. The Aubin-Nitsche trick revisited

In this section we want to prove that on suitable Hilbert spaces containing the space $\mathcal{V}$ continuously the approximation speed in the Galerkin approximation can be improved. We refer also to [28] for related, but different results in this direction.
Let $\mathcal{V}$ be a separable Hilbert space over $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, and $a : \mathcal{V} \times \mathcal{V} \to \mathbb{K}$ a sesquilinear form satisfying

$$|a(u, v)| \leq M\|u\|_{\mathcal{V}}\|v\|_{\mathcal{V}}.$$

Let $(\mathcal{V}_n)_{n\in\mathbb{N}^*}$ be an approximating sequence of $\mathcal{V}$. We assume that (BNB) holds; i.e. there exists $\beta > 0$ such that

$$\text{For all } n \in \mathbb{N}^*, \quad \sup_{v\in\mathcal{V}_n, \|v\|_{\mathcal{V}}=1} |a(u, v)| \geq \beta\|u\|_{\mathcal{V}} \text{ for all } u \in \mathcal{V}_n. \tag{6.1}$$

Given $L \in \mathcal{V}'$ and $n \in \mathbb{N}^*$, let $u_n \in \mathcal{V}_n$ be the solution of

$$a(u_n, \chi) = \langle L, \chi\rangle \text{ for all } \chi \in \mathcal{V}_n, \tag{6.2}$$

and $u \in \mathcal{V}$ the solution of

$$a(u, v) = \langle L, v\rangle \text{ for all } v \in \mathcal{V}. \tag{6.3}$$

Note that, by subtracting (6.3) and (6.2), we obtain the following *Galerkin orthogonality*:

$$a(u - u_n, v) = 0 \text{ for all } v \in \mathcal{V}_n. \tag{6.4}$$

We know from Proposition 2.5 and Proposition 2.6 that

$$\|u - u_n\|_{\mathcal{V}} \leq \frac{M}{\beta} \operatorname{dist}(u, \mathcal{V}_n) \tag{6.5}$$

for all $n \in \mathbb{N}^*$. We want to improve this estimate if the given data $L \in \mathcal{V}'$ is in a suitable subspace of $\mathcal{V}'$.
Let $\mathcal{X} \hookrightarrow \mathcal{V}'$; i.e. $\mathcal{X}$ is a Banach space such that $\mathcal{X} \subset \mathcal{V}'$ and

$$\|f\|_{\mathcal{X}} \leq c_{\mathcal{X}}\|f\|_{\mathcal{V}'}$$

for all $f \in \mathcal{X}$ and some $c_{\mathcal{X}} > 0$. We define for $n \in \mathbb{N}^*$

$$\gamma_n(\mathcal{X}) := \sup_{f\in\mathcal{X},\|f\|_{\mathcal{X}}=1} \operatorname{dist}(\mathcal{A}^{-1}f, \mathcal{V}_n), \tag{6.6}$$

where the distance is taken in $\mathcal{V}$. Thus

$$\operatorname{dist}(w, \mathcal{V}_n) \leq \gamma_n(\mathcal{X})\|\mathcal{A}w\|_{\mathcal{X}} \text{ for all } w \in \mathcal{A}^{-1}\mathcal{X}, \tag{6.7}$$

where $\mathcal{A} : \mathcal{V} \to \mathcal{V}'$ is the isomorphism given by

$$\langle \mathcal{A}u, v\rangle = a(u, v).$$

Thus, if u is the solution of (6.3) and u_n the approximate solution of (6.2), then, if $L \in \mathcal{X}$, we have the estimate

$$\|u - u_n\|_{\mathcal{V}} \leq \frac{M}{\beta}\gamma_n(\mathcal{X})\|L\|_{\mathcal{X}}, \tag{6.8}$$

which has the advantage of being uniform for L in the unit ball of $\mathcal{X}$.

Remark 6.1. *Let $Q_n : \mathcal{X} \to \mathcal{V}, L \mapsto u_n$ be the solution operator for (6.2). Then (6.8) says that*

$$\|\mathcal{A}^{-1} - Q_n\|_{\mathcal{L}(\mathcal{X},\mathcal{V})} \leq \frac{M}{\beta}\gamma_n(\mathcal{X}).$$

We can characterize when $\gamma_n(\mathcal{X}) \to 0$ as $n \to \infty$.

Proposition 6.2. *One has*

$$\lim_{n\to\infty} \gamma_n(\mathcal{X}) = 0 \text{ if and only if } \mathcal{X} \hookrightarrow \mathcal{V}' \text{ is compact.}$$

Proof. Denote by $P_n : \mathcal{V} \to \mathcal{V}_n$ the orthogonal projection onto $\mathcal{V}_n$. Then

$$\gamma_n(\mathcal{X}) = \sup_{f\in\mathcal{X},\|f\|_{\mathcal{X}}=1} \|\mathcal{A}^{-1}f - P_n\mathcal{A}^{-1}f\|_{\mathcal{V}} = \|\mathcal{A}^{-1} \circ j - P_n\mathcal{A}^{-1} \circ j\|_{\mathcal{L}(\mathcal{X},\mathcal{V})},$$

where $j : \mathcal{X} \to \mathcal{V}'$ is the canonical injection. If j is compact, then $K := \mathcal{A}^{-1} \circ j(B_{\mathcal{X}})$, where $B_{\mathcal{X}}$ is the unit ball of $\mathcal{X}$, is relatively compact in $\mathcal{V}$. Now, P_n converges strongly to the identity of $\mathcal{V}$. Since $\|P_n\| \leq 1$, this convergence is uniform on compact subsets of $\mathcal{X}$. This shows that $\gamma_n(\mathcal{X}) \to 0$ as $n \to \infty$.
Conversely, if $\gamma_n(\mathcal{X}) \to 0$, then $\mathcal{A}^{-1} \circ j$ is compact as limit of finite rank operators. Then also j is compact. □

Similarly, we define

$$\gamma_n^*(\mathcal{X}) := \sup_{f\in\mathcal{X},\|f\|_{\mathcal{X}}=1} \operatorname{dist}(\mathcal{A}^{*-1}f, \mathcal{V}_n),$$

where $\mathcal{A}^* : \mathcal{V} \to \mathcal{V}^*$ is given by

$$\langle \mathcal{A}^*u, v\rangle = a^*(u, v) := \overline{a(u, v)}.$$

As before we have $\gamma_n^*(\mathcal{X})$ defined as $\gamma_n(\mathcal{X})$ but with a replaced by the adjoint form a^* of a. Thus we have for all $w \in \mathcal{A}^{*-1}\mathcal{X}$,

$$\operatorname{dist}(w, \mathcal{V}_n) \leq \gamma_n^*(\mathcal{X})\|\mathcal{A}^*w\|. \tag{6.9}$$

Now we apply the Aubin–Nitsche trick in the following proof. In contrast to the literature [12] we allow non-selfadjoint forms and also let $L \in \mathcal{X}$ where $\mathcal{X} \hookrightarrow \mathcal{V}'$ is arbitrary. However, as usual, we fix a Hilbert space $\mathcal{H}$ such that $\mathcal{V} \hookrightarrow \mathcal{H}$ with dense range. Thus we have the Gelfand triple

$$\mathcal{V} \hookrightarrow \mathcal{H} \hookrightarrow \mathcal{V}'.$$

Now we let $\mathcal{X} \hookrightarrow \mathcal{V}'$ be another Banach space in which we choose the given data L, whereas our error estimate is done with respect to the norm of $\mathcal{H}$.

Theorem 6.3. *Let $L \in \mathcal{X}$ and let u be the solution of (6.3), u_n the solution of (6.2). Then*

$$\|u - u_n\|_{\mathcal{H}} \leq \frac{M^2}{\beta} \gamma_n(\mathcal{X}) \gamma_n^*(\mathcal{H}) \|L\|_{\mathcal{X}}, \tag{6.10}$$

for all $n \in \mathbb{N}^*$.

Proof. Let $n \in \mathbb{N}^*$. Then, on the footsteps of Aubin–Nitsche, we consider the solution $w \in \mathcal{V}$ of

$$a^*(w, v) = \langle u - u_n, v \rangle_{\mathcal{H}} \quad (v \in \mathcal{V}). \tag{6.11}$$

Then, by (6.11), for any $\chi \in \mathcal{V}_n$,

$$\begin{aligned} \|u - u_n\|_{\mathcal{H}}^2 &= \langle u - u_n, u - u_n \rangle_{\mathcal{H}} = a^*(w, u - u_n) = \overline{a(u - u_n, w)} \\ &= \overline{a(u - u_n, w - \chi)} \leq M \|u - u_n\|_{\mathcal{V}} \|w - \chi\|_{\mathcal{V}} \end{aligned}$$

where in the last identity we used the Galerkin orthogonality (6.4).
Since $\chi \in \mathcal{V}_n$ is arbitrary, this implies that

$$\|u - u_n\|_{\mathcal{H}}^2 \leq M \|u - u_n\|_{\mathcal{V}} \operatorname{dist}(w, \mathcal{V}_n).$$

Now we use (6.8) and (6.9) to deduce

$$\|u - u_n\|_{\mathcal{H}}^2 \leq M . \frac{M}{\beta} \gamma_n(\mathcal{X}) \|L\|_{\mathcal{X}} \gamma_n^*(\mathcal{H}) \|u - u_n\|_{\mathcal{H}}.$$

Consequently, we obtain

$$\|u - u_n\|_{\mathcal{H}} \leq \frac{M^2}{\beta} \gamma_n(\mathcal{X}) \gamma_n^*(\mathcal{H}) \|L\|_{\mathcal{X}}.$$

□

7. Applications

7.1. **Selfadjoint positive operators with compact resolvent.** As an illustration, we apply Theorem 6.3 to selfadjoint positive operators with compact resolvent. Let $\mathcal{V}, \mathcal{H}$ be infinite dimensional, separable Hilbert spaces over $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ such that $\mathcal{V}$ is compactly injected in $\mathcal{H}$ and dense in $\mathcal{H}$. Thus we have the Gelfand triple

$$\mathcal{V} \hookrightarrow \mathcal{H} \hookrightarrow \mathcal{V}'.$$

Let $a : \mathcal{V} \times \mathcal{V} \to \mathbb{K}$ be continuous, symmetric and coercive. Then the operator $\mathcal{A} : \mathcal{V} \to \mathcal{V}'$ given by

$$\langle \mathcal{A}u, v\rangle = a(u, v)$$

is invertible. Moreover, there exist an orthonormal basis $(e_n)_{n\geq 0}$ of $\mathcal{H}$ and $\lambda_n \in \mathbb{R}$ such that

$$0 < \lambda_0 \leq \lambda_1 \leq \cdots, \lim_{n\to\infty} \lambda_n = \infty$$

and

$$\mathcal{V} = \{u \in \mathcal{H} : \sum_{n=0}^{\infty} \lambda_n |\langle u, e_n\rangle_{\mathcal{H}}|^2 < \infty\}$$

(see e.g. [2, Satz 4.49]) and

$$a(u, v) = \sum_{n=0}^{\infty} \lambda_n \langle u, e_n\rangle_{\mathcal{H}} \langle e_n, v\rangle_{\mathcal{H}}.$$

Passing to an equivalent scalar product we may and will assume that

$$\langle u, v\rangle_{\mathcal{V}} = a(u, v) \quad (u, v \in V).$$

Thus $|a(u, v)| \leq \|u\|_{\mathcal{V}}\|v\|_{\mathcal{V}}$ and $\sup_{\|v\|_{\mathcal{V}}=1} |a(u, v)| = \|u\|_{\mathcal{V}}$; i.e. we have $M = \beta = 1$ in the above estimates.
Consider $\mathcal{V}_n = \text{Span}\{e_0, \cdots, e_{n-1}\}, n = 1, 2, \cdots$. Then $(\mathcal{V}_n)_{n\in\mathbb{N}^*}$ is an approximating sequence of $\mathcal{V}$. We define for $s \in [-1, 1]$

$$\mathcal{V}_s := \{f \in \mathcal{V}' : \sum_{n\geq 0} \lambda_n^s |\langle f, e_n\rangle_{\mathcal{H}}|^2 < \infty\},$$

which is a Hilbert space for the norm

$$\|f\|_{\mathcal{V}_s}^2 = \sum_{n\geq 0} \lambda_n^s |\langle f, e_n\rangle_{\mathcal{H}}|^2.$$

Then it is easy to see that $\mathcal{V}_{-1} = \mathcal{V}'$, $\mathcal{V}_0 = \mathcal{H}$, $\mathcal{V}_1 = \mathcal{V}$ with identity of the norms. Morever, for $s \in (0, 1)$,

$$\mathcal{V}_s = (\mathcal{V}_0, \mathcal{V}_1)_s$$

(the complex interpolation space) and for $s \in (-1, 0)$,

$$\mathcal{V}_s = (\mathcal{V}_0, \mathcal{V}_{-1})_{-s}.$$

Lemma 7.1. *One has for $s \in [-1, 1]$,*

$$\gamma_n(\mathcal{V}_s) = |\lambda_n|^{-(1+s)/2} \quad (n = 1, 2, \cdots).$$

In particular,

$$\gamma_n(\mathcal{H}) = |\lambda_n|^{-1/2}.$$

Proof. Let $\widehat{e_n} = \frac{1}{\sqrt{\lambda_n}} e_n$. Then $(\widehat{e_n})_{n\geq 0}$ is an orthonormal basis of $\mathcal{V}$. For $u \in \mathcal{V}$,

$$\langle u, \widehat{e_k}\rangle_{\mathcal{V}} = \sum_{n\geq 0} \lambda_n \langle u, e_n\rangle_{\mathcal{H}} \langle e_n, \widehat{e_k}\rangle_{\mathcal{H}} = \sqrt{\lambda_k}\langle u, e_k\rangle_{\mathcal{H}}$$

Thus

$$P_n u = \sum_{k=0}^{n-1} \langle u, \widehat{e_k}\rangle_{\mathcal{V}}\ \widehat{e_k} = \sum_{k=0}^{n-1} \langle u, e_k\rangle_{\mathcal{H}}\ e_k$$

defines the orthogonal projection of $\mathcal{V}$ onto $\mathcal{V}_n$. Moreover, in $\mathcal{V}$ one has

$$\operatorname{dist}(u, \mathcal{V}_n)^2 = \|u - P_n u\|_{\mathcal{V}}^2 = \sum_{k\geq n} \lambda_k |\langle u, e_k\rangle_{\mathcal{H}}|^2.$$

Let $f \in \mathcal{V}_s$, $u = \mathcal{A}^{-1} f$. Then

$$\langle f, e_k\rangle_{\mathcal{H}} = \langle \mathcal{A}u, e_k\rangle_{\mathcal{H}} = \lambda_k \langle u, e_k\rangle_{\mathcal{H}}.$$

Thus

$$\begin{aligned}
\gamma_n(\mathcal{V}_s)^2 &= \sup_{\mathcal{A}u=f} \frac{\|u - P_n u\|_{\mathcal{V}}^2}{\|f\|_{\mathcal{V}_s}^2} = \sup_{f\in\mathcal{V}_s, \mathcal{A}u=f} \frac{\sum_{k\geq n} \lambda_k |\langle u, e_k\rangle_{\mathcal{H}}|^2}{\sum_{k\geq 0} \lambda_k^s |\langle f, e_k\rangle_{\mathcal{H}}|^2} \\
&= \sup_{f\in\mathcal{V}_s} \frac{\sum_{k\geq n} \lambda_k^{-1} |\langle f, e_k\rangle_{\mathcal{H}}|^2}{\sum_{k\geq 0} \lambda_k^s |\langle f, e_k\rangle_{\mathcal{H}}|^2} = \sup_{f\in\mathcal{V}_s} \frac{\sum_{k\geq n} \lambda_k^{-1-s}\lambda_k^s |\langle f, e_k\rangle_{\mathcal{H}}|^2}{\sum_{k\geq 0} \lambda_k^s |\langle f, e_k\rangle_{\mathcal{H}}|^2} \\
&\leq \lambda_n^{-1-s}
\end{aligned}$$

since $(\lambda_k)_{k\geq 0}$ is increasing.
Taking $f = e_n$, one sees that $\gamma_n(\mathcal{V}_s)^2 \geq \frac{\lambda_n^{-1}}{\lambda_n^s} = \lambda_n^{-s-1}$. □

Now let $f \in \mathcal{V}_s$, where $-1 \leq s \leq 1$ and let $u = \mathcal{A}^{-1} f$. Let $u_n \in \mathcal{V}$ such that

$$a(u_n, \chi) = \langle f, \chi\rangle \quad (\chi \in \mathcal{V}_n),$$

i.e. u_n is the approximate solution. Then by Theorem 6.3

$$\|u - u_n\|_{\mathcal{H}} \leq \gamma_n(\mathcal{X})\gamma_n(\mathcal{H})\|f\|_{\mathcal{V}_s} = |\lambda_n|^{-(1+s)/2} |\lambda_n|^{-1/2} \|f\|_{\mathcal{V}_s}.$$

Thus we obtain the following error estimate

$$\|u - u_n\|_{\mathcal{H}} \leq |\lambda|^{-1-s/2} \|f\|_{\mathcal{V}_s}. \tag{7.1}$$

Remark 7.2. *In this special case one can compute the error directly. In fact* $u = \sum_{k=0}^{\infty} \frac{1}{\lambda_k}\langle f, e_k\rangle_{\mathcal{H}} e_k$ *and* $u_n = \sum_{k=0}^{n-1} \langle f, e_k\rangle_{\mathcal{H}} e_k$. *Thus*

$$\|u - u_n\|_{\mathcal{H}}^2 = \sum_{k=n}^{\infty} \frac{1}{\lambda_k^2} |\langle f, e_k\rangle_{\mathcal{H}}|^2 = \sum_{k=n}^{\infty} \lambda_k^{-2-s}\lambda_k^s |\langle f, e_k\rangle_{\mathcal{H}}|^2 \leq \lambda_n^{-2-s} \|f\|_{\mathcal{V}_s}^2,$$

which is exactly the estimate (7.1). This means that Theorem 6.3 gives the best possible estimate of the error.

Let us provide an example of application of (7.1). Let $\mathbb{K} = \mathbb{C}$, $\mathcal{H} = L^2(0, 2\pi)$ with norm $\|u\|_{\mathcal{H}}^2 = \frac{1}{2\pi}\int_0^{2\pi} |u(t)|^2 dt$. Let $\mathcal{V} = \{u \in H^1(0, 2\pi) : u(0) = u(2\pi)\}$ with norm

$$\|u\|_{\mathcal{V}}^2 = \frac{1}{2\pi}\int_0^{2\pi} |u'(t)|^2 dt + \frac{1}{2\pi}\int_0^{2\pi} |u(t)|^2 dt.$$

Then the injection $\mathcal{V} \hookrightarrow \mathcal{H}$ is compact. Let $a : \mathcal{V} \times \mathcal{V} \to \mathbb{C}$ be given by

$$a(u, v) = \frac{1}{2\pi}\int_0^{2\pi} u'(t)\overline{v'(t)}dt + \frac{1}{2\pi}\int_0^{2\pi} u(t)\overline{v(t)}dt.$$

Let $f \in L^2(0, 2\pi)$. Then there exists a unique $u \in H^2(0, 2\pi)$ such that

$$u - u'' = f, \quad u(0) = u(2\pi), \quad u'(0) = u'(2\pi).$$

In fact, u is the unique element of $\mathcal{V}$ such that $a(u, v) = \langle f, v\rangle$ for all $v \in \mathcal{V}$.
For $u \in \mathcal{H}$, let $\widehat{u}(k) = \frac{1}{2\pi}\int_0^{2\pi} u(t)e^{-ikt}dt$ be the k-th Fourier coefficient. Then

$$a(u, v) = \sum_{k\in\mathbb{Z}} (1 + k^2)\widehat{u}(k)\overline{\widehat{v}(k)}.$$

Let $e_k(t) = e^{ikt}, t \in (0, 2\pi)$. Then $(e_k)_{k\in\mathbb{Z}}$ is an orthonormal basis of $\mathcal{H}$ and $\widehat{u}(k) = \langle u, e_k\rangle_{\mathcal{H}}$. Let $\mathcal{V}_n = \text{Span}\{e_k : |k| < n\}$ and let u_n be the approximate solution i.e.

$$a(u_n, \chi) = \langle f, \chi\rangle_{\mathcal{H}} \quad (\chi \in \mathcal{V}_n).$$

Then our estimate shows that

$$\|u_n - u\|_{L^2} \le \frac{1}{(1 + n^2)^{1/2}}\|f\|_{L^2}.$$

Let $0 < s \le 1$ and $\mathcal{V}_s := \{u \in L^2(0, 2\pi) : \sum_{k\in\mathbb{Z}}(1 + k^2)^s|\widehat{u}(k)|^2 < \infty\}$. If $f \in \mathcal{V}_s$, then by (7.1),

$$\|u - u_n\|_{L^2} \le (1 + n^2)^{-1-s/2}\|f\|_{\mathcal{V}_s}. \tag{7.2}$$

7.2. **Finite elements for the Poisson problem.** In this section we want to apply our results to show the convergence of a numerical approximation via triangularization for the solution of a Poisson problem where coercivity is violated but essential coercivity holds. For simplicity we choose $\mathbb{K} = \mathbb{R}$ throughout this section. Let $\Omega \subset \mathbb{R}^d$ be an open, bounded, convex set and let $a_{ij} : \Omega \to \mathbb{R}$ $(1 \le i, j \le d)$ be Lipschitz continuous functions such that

$$a_{ij} = a_{ji} \text{ and } \sum_{i,j=1}^{d} a_{ij}(x)\xi_i\xi_j \ge \alpha|\xi|^2 \quad (\xi \in \mathbb{R}^d)$$

for all $x \in \Omega$, where $\alpha > 0$. Moreover, let $b_j, c_j \in W^{1,\infty}(\Omega)$ for $j = 1, \cdots, d$ and $b_0 \in L^\infty(\Omega)$. We consider the operator A given by

$$Au := -\sum_{i,j=1}^{d} D_i(a_{ij}D_j u) + \sum_{j=1}^{d} b_j D_j u - \sum_{j=1}^{d} D_j(c_j u) + b_0 u \quad (u \in H^2(\Omega)).$$

Note that $A : H^2(\Omega) \to L^2(\Omega)$ is linear and continuous.

Our aim is to study the Poisson equation

$$Au = f \tag{7.3}$$

where $f \in L^2(\Omega)$ is given and a solution $u \in H_0^1(\Omega) \cap H^2(\Omega)$ is to be determined and calculated by approximation. We will impose the uniqueness condition

$$\text{For all } u \in H_0^1(\Omega) \cap H^2(\Omega),\ Au = 0 \text{ implies } u = 0. \tag{7.4}$$

We use the continuous, coercive form

$$a_0 : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$$

given by

$$a_0(u, v) = \sum_{i,j=1}^{d} \int_\Omega a_{ij} D_j u D_i v$$

and also the perturbed form a given by

$$a(u, v) = a_0(u, v) + \sum_{j=1}^{d} \int_\Omega (b_j D_j uv + c_j u D_j v) + \int_\Omega b_0 uv.$$

Note that the adjoint form a^* defined by $a^*(u, v) = a(v, u)$ has the same form as a. This is the reason why we also consider the coefficients c_j.
Then the following well posedness result holds.

Theorem 7.3.

i) *The form a is essentially coercive.*
ii) *Assume (7.4). Then for each $f \in L^2(\Omega)$ there exists a unique solution $u \in H_0^1(\Omega) \cap H^2(\Omega)$ of (7.3).*

Proof. a) We first show H^2-regularity. Let $u \in H_0^1(\Omega)$, $f \in L^2(\Omega)$ such that $a(u, v) = \int_\Omega fv$ for all $v \in H_0^1(\Omega)$. Then $u \in H^2(\Omega)$ and $Au = f$. In fact, let

$$g := f - b_0 u - \sum_{j=1}^{d} (b_j D_j u - D_j(c_j u)).$$

Then $g \in L^2(\Omega)$ and $a_0(u, v) = \int_\Omega gv$ for all $v \in H_0^1(\Omega)$. Now it follows from the classical H^2-result of Kadlec [16] (see [13, Theorem 3.2.1.2]) that $u \in H^2(\Omega)$. It clearly follows that $Au = f$.
b) We show that a is essentially coercive. Let $u_n \rightharpoonup 0$ as $n \to \infty$ in $H_0^1(\Omega)$ and $a(u_n, u_n) \to 0$ as $n \to \infty$. Then $D_j u_n \rightharpoonup 0$ as $n \to \infty$ in $L^2(\Omega)$. Since the embedding of $H_0^1(\Omega)$ in $L^2(\Omega)$ is compact, it follows that $u_n \to 0$ in $L^2(\Omega)$. Consequently

$$\int_\Omega b_j D_j u_n . u_n \to 0, \int_\Omega c_j D_j u_n . u_n \to 0 \text{ and } \int_\Omega b_0 u_n . u_n \to 0 \text{ as } n \to \infty.$$

Thus also $a_0(u_n, u_n) \to 0$ as $n \to \infty$. Since a_0 is coercive this implies $\|u_n\|_{H^1} \to 0$ as $n \to \infty$.
c) The form a satisfies uniqueness. In fact, let $u \in H_0^1(\Omega)$ such that $a(u, v) = 0$ for all

$v \in H_0^1(\Omega)$. Then $u \in H^2(\Omega)$ by part a) of the proof. Hence $u = 0$ by our assumption (7.4).
d) Let $f \in L^2(\Omega)$. It follows from Corollary 4.6 that there exists a unique $u \in H_0^1(\Omega)$ such that $a(u, v) = \langle f, v\rangle_{L^2}$ for all $v \in H_0^1(\Omega)$. Now a) implies that $u \in H^2(\Omega)$ and $Au = f$. □

Concerning the uniqueness property, we make the following remark.

Remark 7.4 (Eigenvalues and uniqueness)**.** *Replace the operator A by $A_\lambda := A - \lambda \mathrm{Id}$ (i.e. b_0 by $b_0 - \lambda$) where $\lambda \in \mathbb{R}$. Then there exists a finite or countable infinite set such that*

$$\{\lambda : (7.4) \text{ is violated for } A_\lambda\} = \{\lambda_n : n \in \mathbb{N}^*, n < N\}$$

where $1 < N \leq \infty$ and $\lambda_n \in \mathbb{R}$, $\lim_{n\to\infty} \lambda_n = \infty$ if $N = \infty$.
If $b_1 = \cdots = b_d = c_1 = \cdots = c_d = 0$ and $b_0 \geq 0$, then $\lambda_n > 0$ for all $n \in \mathbb{N}^*$ and then we are in the coercive case. But in general there will be also negative eigenvalues. The uniqueness condition (7.4) for A is equivalent to saying that $\lambda_n \neq 0$ for all $n \in \mathbb{N}^*$.

Our final aim is to show that the finite element method yields an approximation of the solution of (7.3).
For that purpose we assume that $d = 2$ and that Ω is a convex polygon. Let $\{\tau_h\}_{h>0}$ be a quasi-uniform admissible triangularization of Ω (see [2, Definition 9.26]). In particular each τ_h consists of finitely many triangles covering Ω of outer radius $r_T \leq h$.
For $h > 0$, we consider the corresponding finite element space V_h (see [2, Equation (9.35)]). Thus V_h consists of those continuous functions on $\overline{\Omega}$ which vanish at $\partial\Omega$ and are affine on each triangle $T \in \tau_h$.
The following fundamental estimates are classical (see e.g. [2, Korollar 9.28]) .

Proposition 7.5. *There exists a constant $c > 0$ such that for all $h \in (0, 1)$ and for each $v \in H^2(\Omega)$,*

$$\inf_{\chi\in V_h} \|v - \chi\|_{H^1(\Omega)} \leq ch|v|_{H^2(\Omega)}, \tag{7.5}$$

where $|v|^2_{H^2(\Omega)} := \int_\Omega (|D_1^2 v|^2 + 2|D_1 D_2 v|^2 + |D_2 v|^2)$.

Note that Proposition 7.5 shows how we can approximate functions in $H^2(\Omega)$ by finite elements and so far there is no relation with the solutions of the Poisson equation.
We assume the uniqueness condition (7.4). Then by Theorem 5.2, since the form a is essentially coercive, there exists $h_0 \in (0, 1]$ such that for $0 < h \leq h_0$ and $u \in V_h$

$$a(u, \chi) = 0 \text{ for all } \chi \in V_h \text{ implies } u \in V_h. \tag{7.6}$$

Let $f \in L^2(\Omega)$. Since V_h is finite dimensional, it follows from (7.6) that for all $0 < h \leq h_0$, there exists a unique $u_h \in V_h$ such that

$$a(u_h, \chi) = \int_\Omega f\chi \text{ for all } \chi \in V_h. \tag{7.7}$$

The finite elements $(u_h)_{0<h\le h_0}$ are the approximation of the solution of (7.3) we are interested in. They converge in $H^1(\Omega)$ with convergence order 1 and in $L^2(\Omega)$ with convergence order 2. More precisely, the following is our main theorem of this section.

Theorem 7.6. *Let $f \in L^2(\Omega)$ and consider the approximate solutions u_h, $0 < h \le h_0$. Then there exist $0 < h_1 \le h_0$ and constants c_1, c_2 independent of f such that*

$$\|u - u_h\|_{H^1(\Omega)} \le c_1 h \|f\|_{L^2(\Omega)} \tag{7.8}$$

and

$$\|u - u_h\|_{L^2(\Omega)} \le c_2 h^2 \|f\|_{L^2(\Omega)} \tag{7.9}$$

where u is the solution of (7.3).

Proof. Applying the closed graph theorem in the situation of Theorem 7.3, we find a constant $c_3 > 0$ such that

$$\|u\|_{H^2(\Omega)} \le c_3 \|f\|_{L^2(\Omega)} \tag{7.10}$$

whenever $f \in L^2(\Omega)$ and u solves (7.3).
By Theorem 5.2, there exist $\gamma > 0$, $0 < h_1 \le h_0$, both independent of f, such that

$$\|u - u_h\|_{H^1(\Omega)} \le \gamma \inf_{\chi \in V_h} \|\chi - u\|_{H^1(\Omega)}$$

for all $0 < h \le h_1$. Thus (7.5) implies that for $0 < h \le h_1$,

$$\|u_h - u\|_{H^1(\Omega)} \le ch\gamma |u|_{H^2(\Omega)}.$$

Now (7.8) follows from (7.10).
Next we establish the L^2-estimate (7.9). For that we compute using (7.5),

$$\gamma_h(\mathcal{H}) = \sup_{w \in H_0^1(\Omega) \cap H^2(\Omega)} \frac{\operatorname{dist}(w, \mathcal{V}_h)}{\|Aw\|_{L^2(\Omega)}} \le \sup_{w \in H_0^1(\Omega) \cap H^2(\Omega)} \frac{ch|w|_{H^2(\Omega)}}{\|Aw\|_{L^2(\Omega)}}.$$

Since $|w|_{H^2(\Omega)} \le \|w\|_{H^2(\Omega)}$, it follows from (7.10) that $\gamma_h(\mathcal{H}) \le cc_3 h$ for all $h > 0$.
The same estimate is true for $\gamma_h^*(\mathcal{H})$. Now assume that (7.9) is false. Then there exists a sequence $h_n \downarrow 0$ as $n \to \infty$ such that (7.9) does not hold for all $h = h_n$ and any constant c_2. This contradicts Theorem 6.3. □

Remark 7.7. *There are other methods to approximate the solution of a non-coercive advection-diffusion equation as (7.3). In fact, Le Bris, Legoll and Madiot* [19] *use the Banach-Nečas-Babuska lemma (instead of essential coercivity as we do) and a special measure to construct an approximation.*
The advantage is that no initial mesh h_1 has to be considered; on the other hand there seems to be no such precise error estimate as our quadratic convergence obtained in Theorem 7.6 even though numerical examples are given in [19].
Still, another approach (based on Fredholm perturbation) is presented by Christensen [9], *which also involves the Babuska inf-sup condition.*

Finally, let us mention the works by Droniou, Gallouët and Herbin [10]*, based on finite volume methods, which also present the advantage to provide an approximate solution for this problem on any admissible mesh.*
One of the first results on the Galerkin method in a special non-coercive case are due to Schatz [27] *and Schatz–Wang* [28]*.*

8. Supplement: saddle point problems

Brezzi's contribution [4] is a version of (BNB) which implies the convergence of the Galerkin approximation in the case of saddle point problems. Let us consider the case where $\mathcal{W}$ and $\mathcal{Y}$ are real Hilbert spaces and $\widehat{a} : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ and $\widehat{b} : \mathcal{W} \times \mathcal{Y} \to \mathbb{R}$ are continuous bilinear forms in the sense that there exists $M > 0$ with

$$|\widehat{a}(w, v)| \leq M\|w\|_{\mathcal{W}}\|v\|_{\mathcal{W}} \text{ for all } w \in \mathcal{W}, v \in \mathcal{W}$$

and

$$|\widehat{b}(w, q)| \leq M\|w\|_{\mathcal{W}}\|q\|_{\mathcal{Y}} \text{ for all } w \in \mathcal{W}, q \in \mathcal{Y}.$$

Then, given $(f, g) \in \mathcal{W}' \times \mathcal{Y}'$, the *continuous saddle point problem* consists in finding $(w, p) \in \mathcal{W} \times \mathcal{Y}$ such that

$$\begin{aligned} &\forall z \in \mathcal{W}, \quad \widehat{a}(w, z) + \widehat{b}(z, p) = f(z), \\ &\forall q \in \mathcal{Y}, \quad \widehat{b}(w, q) = g(q). \end{aligned}$$

Example 8.1. *An important example is the Stokes problem (motivating some investigation by Ladyžhenskaya* [18]*), with* $\mathcal{W} = (H_0^1(\Omega))^d$*, where* d *is the space dimension,* $\mathcal{Y} = L_0^2(\Omega)$ *(the space of* L^2*-functions with null average),*

$$\widehat{a}(w, z) = \sum_{i=1}^{d} \int_\Omega \nabla w^{(i)}(x) \cdot \nabla z^{(i)}(x) \mathrm{d}x$$

and

$$\widehat{b}(z, p) = \int_\Omega (\nabla \cdot z)(x) p(x) \mathrm{d}x.$$

The approximation of the saddle point problem is then generally done by a mixed method [4, 5], letting, for $n = 1, 2, \ldots$, $(\mathcal{W})_{n\in\mathbb{N}^\star}$ and $(\mathcal{Y})_{n\in\mathbb{N}^\star}$ be approximating sequences in the spaces $\mathcal{W}$ and $\mathcal{Y}$, respectively, in the sense of Definition 2.1, and looking for $(w_n, p_n) \in \mathcal{W}_n \times \mathcal{Y}_n$ such that

$$\begin{aligned} &\forall z \in \mathcal{W}_n, \quad \widehat{a}(w_n, z) + \widehat{b}(z, p_n) = f(z), \\ &\forall q \in \mathcal{Y}_n, \quad \widehat{b}(w_n, q) = g(q). \end{aligned}$$

We call this the *approximate saddle point problem.*

The following result shows that conditions (8.1) (which are Brezzi's conditions [4, Hypotheses H1 and H2]) are sufficient for the convergence of the solutions of the approximate saddle point problems. This is proved by Brezzi [4, Theorem 2.1], where a solution

is assumed to exist. However, similar to the proof of our Proposition 2.5, one can show that Brezzi's conditions imply existence and uniqueness of the continuous saddle point problem. Indeed, following the proof of (8.2) given in the proof of [4, Theorem 2.1], letting $w = 0$ and $p = 0$, we get a bound on the approximate solution, and a solution of the continuous problem can be obtained by passing to the limit of a weakly converging subsequence. Uniqueness follows from the estimate (8.2) proved by Brezzi. For $n = 1, 2, \ldots$, define

$$\mathcal{W}_{0,n} = \{u \in \mathcal{W}_n;\ \forall q \in \mathcal{Y}_n,\ \widehat{b}(u,q) = 0\}$$

and assume that $\mathcal{W}_{0,n}^* := \mathcal{W}_{0,n} \setminus \{0\} \neq \emptyset$ and $\mathcal{Y}_n^* := \mathcal{Y}_n \setminus \{0\} \neq \emptyset$ for all $n \in \mathbb{N}^*$.

Theorem 8.2 (Brezzi)**.** *Assume that there exists $\beta > 0$ such that*

$$\left\{\begin{array}{ll} (i) & \forall n \in \mathbb{N}^\star,\ \inf_{w\in\mathcal{W}_{0,n}^\star} \sup_{z\in\mathcal{W}_{0,n}^\star} \frac{\widehat{a}(w,z)}{\|w\|_{\mathcal{W}}\|z\|_{\mathcal{W}}} \geq \beta \\ (ii) & \forall n \in \mathbb{N}^\star,\ \inf_{z\in\mathcal{W}_{0,n}^\star} \sup_{w\in\mathcal{W}_{0,n}^\star} \frac{\widehat{a}(w,z)}{\|w\|_{\mathcal{W}}\|z\|_{\mathcal{W}}} \geq \beta \\ (iii) & \forall n \in \mathbb{N}^\star,\ \inf_{p\in\mathcal{Y}_n^\star} \sup_{z\in\mathcal{W}_n^\star} \frac{\widehat{b}(z,p)}{\|z\|_{\mathcal{W}}\|p\|_{\mathcal{Y}}} \geq \beta. \end{array}\right. \tag{8.1}$$

Then, given $(f,g) \in \mathcal{W}' \times \mathcal{Y}'$, there exists a unique solution (w,p) of the continuous saddle point problem and for each $n \in \mathbb{N}^*$ a unique solution (w_n, p_n) of the approximate saddle point problem. Moreover,

$$\forall n \in \mathbb{N}^\star,\ \|w_n - w\|_W + \|p_n - p\|_Y \leq c\big(\operatorname{dist}(w, W_n) + \operatorname{dist}(p, Y_n)\big) \tag{8.2}$$

where the constant c depends only on β and M.

The saddle point problem can be cast in our framework by letting $\mathcal{V} = \mathcal{U} = \mathcal{W} \times \mathcal{Y}$, $u = (w,p)$, $v = (z,q)$ and

$$a(u,v) = \widehat{a}(w,z) + \widehat{b}(z,p) + \widehat{b}(w,q).$$

Given $(f,g) \in \mathcal{V}' = \mathcal{W}' \times \mathcal{Y}'$, define $L \in \mathcal{V}'$ by

$$\langle L, (z,q)\rangle = \langle f, z\rangle + \langle g, q\rangle.$$

Then $u = (w,p)$ is a solution of the continuous saddle point problem if and only if (1.1) is satisfied. Moreover, letting $\mathcal{V}_n = \mathcal{U}_n = \mathcal{W}_n \times \mathcal{Y}_n$, a vector $u_n = (w_n, p_n) \in \mathcal{V}_n$ satisfies (1.2) if and only if (w_n, p_n) is a solution of the approximate saddle point problem. Thus our Theorem 2.4 shows that the convergence property expressed in Brezzi's Theorem is equivalent to (BNB) for the form a and the approximating sequence $(\mathcal{V}_n)$. We can use this to show the following converse result of Brezzi's Theorem.

Theorem 8.3. *Assume that, given $(f,g) \in \mathcal{W}' \times \mathcal{Y}'$, for each $n \in \mathbb{N}^*$, there is a unique solution (w_n, p_n) of the discrete saddle point problem and that $\sup_{n\in\mathbb{N}^*}(\|w_n\|_W + \|p_n\|_Y) < \infty$. Then Brezzi's conditions (8.1) hold.*

Proof. We know from Theorem 2.4 and Proposition 2.5 that (BNB) is satisfied for some $\beta > 0$. We endow the space $\mathcal{V}$ with the norm $\|u\|_{\mathcal{V}} = \big(\|w\|_{\mathcal{W}}^2 + \|p\|_{\mathcal{Y}}^2\big)^{1/2}$ for $u = (w,p)$ (it

is then a Hilbert space as well). Let $n \in \mathbb{N}^\star$ be given. We then have,

$$\forall (w,p) \in \mathcal{W}_n \times \mathcal{Y}_n, \quad \sup_{(z,q)\in \mathcal{W}_n\times\mathcal{Y}_n\setminus\{(0,0)\}} \frac{|a((w,p),(z,q))|}{\|(z,q)\|_{\mathcal{V}}} \geq \beta \|(w,p)\|_{\mathcal{V}}. \tag{8.3}$$

Let us first choose, for any $p \in \mathcal{Y}_n^\star$, $u = (0,p)$, which means that $w = 0 \in \mathcal{W}_n$. Let $(z,q) \in \mathcal{W}_n \times \mathcal{Y}_n \setminus \{(0,0)\}$ attaining the supremum value in (8.3). We then have, from the definition of a in this framework of a saddle point problem,

$$\frac{|\widehat{b}(z,p)|}{\left(\|z\|_{\mathcal{W}}^2 + \|q\|_{\mathcal{Y}}^2\right)^{1/2}} \geq \beta \|p\|_{\mathcal{Y}},$$

which implies that $z \neq 0$ and

$$\frac{|\widehat{b}(z,p)|}{\|z\|_{\mathcal{W}}} \geq \beta \|p\|_{\mathcal{Y}}.$$

This proves (8.1).(iii) , and thus that the operator $\widehat{\mathcal{B}}_n \ : \ \mathcal{W}_n \to \mathcal{Y}_n$, defined for all $z \in \mathcal{W}_n$ by

$$\forall q \in \mathcal{Y}_n, \ \widehat{b}(z,q) = \langle \widehat{\mathcal{B}}_n z, q\rangle_{\mathcal{Y}},$$

is bijective from $\mathcal{W}_{0,n}^\perp$ to $\mathcal{Y}_n$.

Let $w \in \mathcal{W}_{0,n}^\star$ and let $p \in \mathcal{Y}_n$ be defined by

$$\forall q \in \mathcal{Y}_n, \ \langle q,p\rangle_{\mathcal{Y}} = \widehat{b}(\widehat{\mathcal{B}}_n^{(-1)} q, p) = -\widehat{a}(w, \widehat{\mathcal{B}}_n^{(-1)} q).$$

Choose an element $(z,q) \in \mathcal{W}_n \times \mathcal{Y}_n \setminus \{(0,0)\}$ attaining the supremum value in (8.3) for this choice of $u = (w,p)$. We then write $z = z_0 + z_1$, with $z_0 \in \mathcal{W}_{0,n}$ and $z_1 \in \mathcal{W}_{0,n}^\perp$, which can be written as $z_1 = \widehat{\mathcal{B}}_n^{(-1)} q_1$ for some $q_1 \in \mathcal{Y}_n$. We have

$$a((w,p),(z,q)) = \widehat{a}(w,z_0) + \widehat{a}(w,z_1) + \widehat{b}(z_0,p) + \widehat{b}(z_1,p) + \widehat{b}(w,q).$$

Moreover, $\widehat{b}(z_0,p) = \widehat{b}(w,q) = 0$ since $z_0 \in \mathcal{W}_{0,n}$ and $w \in \mathcal{W}_{0,n}$, and

$$\widehat{a}(w,z_1) + \widehat{b}(z_1,p) = 0,$$

by definition of p and of z_1. Hence

$$\frac{|\widehat{a}(w,z_0)|}{\left(\|z\|_{\mathcal{W}}^2 + \|q\|_{\mathcal{Y}}^2\right)^{1/2}} \geq \beta \left(\|w\|_{\mathcal{W}}^2 + \|p\|_{\mathcal{Y}}^2\right)^{1/2}.$$

This implies that $z_0 \neq 0$, and therefore $z_0 \in \mathcal{W}_{0,n}^\star$ is such that

$$\frac{|\widehat{a}(w,z_0)|}{\|z_0\|_{\mathcal{W}}} \geq \beta \|w\|_{\mathcal{W}},$$

where we take into account that $\|z\|_{\mathcal{W}} \geq \|z_0\|_{\mathcal{W}}$ by Pythagore's theorem. This concludes the proof of (8.1).(i).

The equivalence between (BNB) and $(BNB^\star)$ allows to obtain the proof of (8.1).(ii) (with the same β, see Proposition 2.9), following the same path.

□

In conclusion, Brezzi's conditions (8.1) are equivalent to the well posedness of the continuous saddle point problem together with the convergence of the approximate solutions to the solution, and they are also equivalent to (BNB) for the form a and the approximating sequence $(\mathcal{V}_n)$ of $\mathcal{V}$.
Note that [5, Chapter II, Remark 2.11] provides a comment on the fact that (8.1).(*iii*) is a necessary condition.

Acknowledgments: We are most grateful to Gilles Lancien about a discussion on the approximation property and pointing out the survey article of Casazza [7] to us. We also thank the anonymous referee for useful and inspiring comments. This research is partly supported by the Bézout Labex, funded by ANR, reference ANR-10-LABX-58.

References

[1] W. Arendt, A. F. M. ter Elst, J. B. Kennedy, and M. Sauter. The Dirichlet-to-Neumann operator via hidden compactness. *J. Funct. Anal.*, 266(3):1757–1786, 2014.
[2] W. Arendt and K. Urban. *Partielle Differenzialgleichungen. Eine Einführung in analytische und numerische Methoden.* Berlin: Springer Spektrum, 2nd edition edition, 2018.
[3] I. Babuška. Error-bounds for finite element method. *Numer. Math.*, 16:322–333, 1970/71.
[4] F. Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge*, 8(R-2):129–151, 1974.
[5] F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods*, volume 15 of *Springer Series in Computational Mathematics*. Springer-Verlag, New York, 1991.
[6] F. E. Browder. Nonlinear operators and nonlinear equations of evolution in Banach spaces. In *Nonlinear functional analysis (Proc. Sympos. Pure Math., Vol. XVIII, Part 2, Chicago, Ill., 1968)*, pages 1–308, 1976.
[7] P. G. Casazza. Chapter 7 - approximation properties. In W. Johnson and J. Lindenstrauss, editors, *Handbook of the Geometry of Banach Spaces*, volume 1 of *Handbook of the Geometry of Banach Spaces*, pages 271 – 316. Elsevier Science B.V., 2001.
[8] L. Chesnel and P. jun. Ciarlet. T-coercivity and continuous Galerkin methods: application to transmission problems with sign changing coefficients. *Numer. Math.*, 124(1):1–29, 2013.
[9] S. H. Christiansen. Discrete Fredholm properties and convergence estimates for the electric field integral equation. *Math. Comput.*, 73(245):143–167, 2004.
[10] J. Droniou, T. Gallouët, and R. Herbin. A finite volume scheme for a noncoercive elliptic equation with measure data. *SIAM J. Numer. Anal.*, 41(6):1997–2031, 2003.
[11] P. Enflo. A counterexample to the approximation problem in Banach spaces. *Acta Math.*, 130:309–317, 1973.
[12] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.
[13] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*, volume 24 of *Monographs and Studies in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1985.
[14] K. Gustafson. The Toeplitz–Hausdorff theorem for linear operators. *Proc. Amer. Math. Soc.*, 25:203–204, 1970.
[15] W. Hackbusch. *Theorie und Numerik elliptischer Differentialgleichungen.* Heidelberg: Springer Spektrum, 4th revised edition edition, 2017.
[16] J. Kadlec. On the regularity of the solution of the Poisson problem on a domain with boundary locally similar to the boundary of a convex open set. *Czech. Math. J.*, 14:386–393, 1964.

[17] T. Kato. Estimation of iterated matrices, with application to the von Neumann condition. *Numer. Math.*, 2:22–29, 1960.
[18] O. A. Ladyzhenskaya. *The mathematical theory of viscous incompressible flow*. Revised English edition. Translated from the Russian by Richard A. Silverman. Gordon and Breach Science Publishers, New York-London, 1963.
[19] C. Le Bris, F. Legoll, and F. Madiot. Stabilisation de problèmes non coercifs via une méthode numérique utilisant la mesure invariante. *C. R., Math., Acad. Sci. Paris*, 354(8):799–803, 2016.
[20] J. Lindenstrauss and L. Tzafriri. *Classical Banach spaces. I*. Springer-Verlag, Berlin-New York, 1977. Sequence spaces, Ergebnisse der Mathematik und ihrer Grenzgebiete, Vol. 92.
[21] J.-L. Lions and E. Magenes. *Problèmes aux limites non homogènes et applications. Vol. 1*. Travaux et Recherches Mathématiques, No. 17. Dunod, Paris, 1968.
[22] W. V. Petryshyn. On projectional-solvability and the Fredholm alternative for equations involving linear A-proper operators. *Arch. Rational Mech. Anal.*, 30:270–284, 1968.
[23] W. V. Petryshyn. On the approximation-solvability of equations involving A-proper and psuedo-A-proper mappings. *Bull. Amer. Math. Soc.*, 81:223–312, 1975.
[24] A. Pietsch. *History of Banach Spaces and Linear Operators*. Birkhäuser, Basel, 2007.
[25] S. Prössdorf and B. Silbermann. *Numerical analysis for integral and related operator equations*, volume 52 of *Operator Theory: Advances and Applications*. Birkhäuser Verlag, Basel, 1991.
[26] C. J. Read. Different forms of the approximation property. Typed manuscript, Leeds, 1986.
[27] A. H. Schatz. An observation concerning Ritz-Galerkin methods with indefinite bilinear forms. *Mathematics of Computation*, 28(128):959–962, 1974.
[28] A. H. Schatz and J. Wang. Some new error estimates for ritz-galerkin methods with minimal regularity assumptions. *Mathematics of Computation*, 65(213):19–27, 1996.
[29] A. Stern. Banach space projections and Petrov-Galerkin estimates. *Numer. Math.*, 130(1):125–133, 2015.
[30] S. J. Szarek. A Banach space without a basis which has the bounded approximation property. *Acta Math.*, 159:81–98, 1987.
[31] J. Xu and L. Zikatanov. Some observations on Babuška and Brezzi theories. *Numer. Math.*, 94(1):195–202, 2003.
[32] E. Zeidler. *Nonlinear functional analysis and its applications. II/A*. Springer-Verlag, New York, 1990. Linear monotone operators, Translated from the German by the author and Leo F. Boron.

Wolfgang Arendt, Institute of Applied Analysis, University of Ulm. Helmholtzstr. 18, D-89069 Ulm (Germany)
E-mail address: wolfgang.arendt@uni-ulm.de

Isabelle Chalendar, Université Paris-Est, LAMA, (UMR 8050), UPEM, UPEC, CNRS, F-77454, Marne-la-Valle (France)
E-mail address: isabelle.chalendar@u-pem.fr

Robert Eymard, Université Paris-Est, LAMA, (UMR 8050), UPEM, UPEC, CNRS, F-77454, Marne-la-Valle (France)
E-mail address: robert.eymard@u-pem.fr