Adaptive Petrov–Galerkin Methods for First Order Transport Equations

Wolfgang Dahmen, Chunyan Huang,
Christoph Schwab, Gerrit Welper

Bericht Nr. 321 **Januar 2011**

Key words: Linear transport problems, L_2–stable Petrov–Galerkin formulations, trace theorems, δ–proximality, adaptive refinement schemes, residual approximation, error reduction.

AMS subject classifications: Primary: 65N30, 65J15, 65N12, 65N15

Institut für Geometrie und Praktische Mathematik
RWTH Aachen

Templergraben 55, D–52056 Aachen (Germany)

This research was performed with the Priority Research Programme SPP1324 of the German Research Foundation (DFG) within the project ,,Anisotropic Adaptive Discretization Concepts". C. Schwab also acknowledges partial support by the European Research Council under grant AdG 247277.

ADAPTIVE PETROV-GALERKIN METHODS FOR FIRST ORDER TRANSPORT EQUATIONS

WOLFGANG DAHMEN, CHUNYAN HUANG, CHRISTOPH SCHWAB, GERRIT WELPER

Abstract. We propose a general framework for well posed variational formulations of linear unsymmetric operators, taking first order transport and evolution equations in bounded domains as primary orientation. We outline a general variational framework for stable discretizations of boundary value problems for these operators. To adaptively resolve anisotropic solution features such as propagating singularities the variational formulations should allow one, in particular, to employ as trial spaces directional representation systems. Since such systems are known to be stable in L_2 special emphasis is placed on L_2-stable formulations. The proposed stability concept is based on perturbations of certain "ideal" test spaces in Petrov-Galerkin formulations. We develop a general strategy for realizing corresponding schemes without actually computing excessively expensive test basis functions. Moreover, we develop adaptive solution concepts with provable error reduction. The results are illustrated by first numerical experiments.

AMS Subject Classification: Primary: 65N30,65J15, 65N12, 65N15

Key Words: Linear transport problems, L_2-stable Petrov-Galerkin formulations, trace theorems, δ-proximality, adaptive refinement schemes, residual approximation, error reduction.

1. Introduction

1.1. Background and Motivation.

The scope of (linear) operator equations

$$A_\circ u = f \tag{1.1}$$

for which currently rigorously founded *adaptive* solution concepts exist is signified by the validity of a *well posed variational formulation*. By this we mean that the bilinear form $a(v,w) := (Av,w)$, where $(\cdot,\cdot)$ is an L_2-like inner product and v, w are sufficiently regular functions, extends continuously to a pair of (separable) Hilbert spaces X, Y in such a way such that the corresponding continuous extension A of $A_\circ$ is a *norm isomorphism* from X onto Y', the normed dual of Y. Thus

$$a(u,v) = \langle f, v\rangle, \quad \forall\ v \in Y, \tag{1.2}$$

possesses for each $f \in Y'$ a unique solution $x \in X$. Well posedness is usually established by veryfing an *inf-sup condition* that will be taken up later in more detail.

The most common examples concern the *symmetric* case $X = Y$ such as Poisson-type elliptic boundary value problems, the Stokes system or boundary integral equations of potential type. In this case Galerkin discretizations suggest themselves. An immediate important consequence of well posedness is that the approximation error $\|u - u_h\|_X$ incurred by the Galerkin projection u_h of u onto any closed subspace $X_h \subset X$ equals up to a constant - the condition number $\|A\|_{X\to X'}\|A^{-1}\|_{X'\to X}$ of A - the best approximation error in X and is equivalent to the residual $\|Au_h - f\|_{X'}$ in the dual norm. In the general *unsymmetric* case, this amounts to ensuring that for any $v_h \in X_h$

$$\|A\|^{-1}_{X\to Y'}\|f - Av_h\|_{Y'} \le \|u - v_h\|_X \le \|A^{-1}\|_{Y'\to X}\|f - Av_h\|_{Y'}. \tag{1.3}$$

Date: January 24, 2011.

Key words and phrases. Transport equations, wavelets, shearlets, adaptivity, computational complexity, best N-term approximation, matrix compression.

This research was performed with the Priority Research Programme SPP1324 of the German Research Foundation (DFG) within the project "Anisotropic Adaptive Discretization Concepts". C. Schwab also acknowledges partial support by the European Research Council under grant AdG 247277.

Roughly speaking, one can distinguish two basic paradigms for adaptive solution concepts, both making crucial use of (1.3). In the context of finite element discretizations duality arguments often allow one to derive sharp upper and lower bounds of $\|f - Av_h\|_{X'}$ in terms of sums of local quantities that serve as error indicators for successive mesh refinements. The second paradigm concerns *wavelet or, more generally, frame discretizations*. In the symmetric case this amounts to employing a *Riesz-basis* for the "energy space" X or an X-stable frame. This together with the well posedness of the variational problem allows one to estimate $\|f - Av_h\|_{X'}$ by the the ℓ_2-norm of the wavelet representation of the residual. This latter quantity, in turn, can be estimated by exploiting knowledge about the data f, a posteriori information about the current approximation v_h, and *near sparsity* of the wavelet representation of the operator A, all properties that need to be established in each application, see e.g. [6, 7, 10, 13] or [11] and the references there for similar results with frames.

Much less seems to be known for the unsymmetric case, except for [22] where a space-time discretization for parabolic initial-boundary value problems is analyzed and optimality of the corresponding adaptive wavelet method in space-time tensorized wavelet bases is shown. In this case the use of a proper extension of the wavelet paradigm is essential since it allows one to employ sparse tensor product concepts.

At any rate, none of these approaches seem to cover problems that are governed by *transport phenomena*. At least two major obstructions arise in this context. First, while (except perhaps in [22]) the choice of X, needed for a well posed formulation, is more or less apparent, this seems to be less clear already for simple linear transport equations. Second, solutions typically exhibit strong *anisotropic* features such as shear layers or shock fronts. The metric entropy of compact sets of such functions suggest that "good" approximation methods should give rise to distortion rates that cannot be achieved by *isotropic* refinements corresponding to classical wavelet bases. Since the stability of discretizations based on anisotropic mesh refinements is not a straightforward matter, an interesting alternative is offered by recent developments centering on *directional representation systems* like curvelets (see [3]) or *shearlets* (see [17, 18, 20]). Best N-term approximations from such systems are known to resolve *wave front sets* - the prototype of anisotropic singularities - at a nearly optimal rate [16] when compared with the metric entropy of such classes.

However, the currently known directional representation systems do generally not form Riesz bases but merely *frames* for a specific function space, namely for $L_2(\mathbb{R}^d)$, $d = 2, 3$, [18]. Adaptive discretizations of operator equations using such representation systems must therefore take this fact into account.

Thus, in summary, in order to eventually employ directional representation systems in the spirit of [7, 11] for the adaptive resolution of transport problems, one faces two preparatory tasks:

(i) find a well posed variational formulation for such problems so that, in particular, errors are controlled by residuals;
(ii) in this variational formulation one should be able to actually *choose* the space X, in which the variational solution is sought, for instance, as an L_2-space.

The first objective of this paper is to address these two tasks. On one hand, as indicated above, (i) is an essential prerequisite for adaptive solution concepts. On the other hand, aside from this it is also crucial for the practical feasibility of so called Reduced Basis Methods where a greedy search for good reduced bases hinges on residual evaluations, see e.g. [21]. This aspect is relevant when dealing with parameter dependent problems, an issue to be taken up later again.

The importance of (ii) is that the employment of certain directional representation systems dictates to some extent the space in which we seek a weak solution. In that sense (ii) paves the way for subsequent developments using such systems which will be addressed in forthcoming work.

In Section 2 we discuss three examples that are to illustrate the subsequent general framework for deriving stable variational formulations. Our main focus is on two of those examples for transport problems.

In Section 3 we present an abstract framework for deriving well posed variational formulations that, in particular, covers the examples in the preceding section.

Of course, prescribing X, or more generally, *always* finding for a given operator equation a pair X, Y giving rise to a well posed variational formulation, comes at a price, namely that the test space Y and the corresponding dual norm $\|\cdot\|_{Y'}$ could be numerically difficult to handle. The key idea may be sketched as follows. Given any finite dimensional trial space $X_h \subset X$, one readily identifies an *ideal* test space $Y_h \subset Y$ which is, however, computationally infeasible, see Section 4. This test space is called ideal because the corresponding Petrov-Galerkin operator has condition number one. We then study in Section 4 which perturbations of the ideal test space are permitted so as to still give rise to uniformly stable Petrov-Galerkin discretizations. We formulate a principal condition, termed δ-*proximality*, which guarantees such stable discretizations. We emphasize that this is a condition on a *whole test space* Y_h^δ, not just on individual test functions.

Setting up a discrete system of equations by first computing a basis in the perturbed test space Y_h^δ and then assembling the corresponding stiffness matrix and load vector, would be way too expensive. One reason is that, for instance, when dealing with transport equations, such test basis functions would have global support. Therefore, we present in Section 5 a strategy for realizing such stable Petrov-Galerkin approximations *without* computing the individual test basis functions explicitly. This strategy allows us also to approximate the residual norm $\|f - Au_h\|_{Y'}$ in a reliable way which, in turn, is shown to lead to an adaptive refinement scheme. We identify concrete conditions under which this scheme can be shown to exhibit a constant error reduction per step. In Section 6 these findings are illustrated by first numerical experiments based on hierarchies of (discontinuous) piecewise polynomials. We conclude with a brief account of the key issues to be addressed in subsequent work in Section 7. In particular, a more detailed discussion and analysis of how the actually realize the key conditions like δ-proximality, as well as applications of the concepts to frame discretizations and parameter dependent problems will be given in a forthcoming work, [12].

After completion of this work we became aware of the work in [14] which is related to the present approach in that it also starts from similar principles of choosing an ideal variational setting. It seems to pursue though a different direction by computing "near-optimal" test functions, exploiting as much as possible the localization offered by a Discontinuous Galerkin context.

2. Examples

In this section we briefly discuss model problems that will be special cases of our general adaptive Petrov-Galerkin approach.

2.1. First Order Linear Transport Equations.

The following model problem will serve as the main conceptual guideline and will therefore be discussed in slightly more detail than the subsequent ones. By $D \subset \mathbb{R}^d, d > 1$, we shall always denote a bounded, polyhedral domain with Lipschitz boundary $\Gamma := \partial D$. We assume in addition that the boundary Γ consists of finitely many polyhedral faces again having Lipschitz boundaries. On Γ, the exterior unit normal $\vec{n}(x)$ exists for almost all $x \in \Gamma$. Moreover, we consider velocity fields $\vec{b}(x)$, $x \in D$, which for simplicity will always be assumed to be differentiable, i.e. $\vec{b}(x) \in C^1(\overline{D})^d$ although some of the statements remain valid under weaker assumptions. Likewise $c(x) \in C^0(\overline{D})$ will serve as the reaction term in the *first order transport equation*

$$A_\circ u := \vec{b} \cdot \nabla u + cu = f_\circ \text{ in } D\,, \tag{2.1}$$

$$u = g \text{ on } \Gamma_-\,, \tag{2.2}$$

which, at this point, should be interpreted in a pointwise sense, assuming for simplicity that $f_\circ, g$ are continuous in their respective domains. Here Γ_- is the *inflow boundary* in the following

partition of $\Gamma = \partial D$

$$\Gamma_- = \{x \in \partial D : \ \vec{b}(x) \cdot \vec{n}(x) < 0\} \quad \text{inflow boundary}, \tag{2.3}$$

$$\Gamma_+ = \{x \in \partial D : \ \vec{b}(x) \cdot \vec{n}(x) > 0\} \quad \text{outflow boundary}, \tag{2.4}$$

$$\Gamma_0 = \{x \in \partial D : \ \vec{b}(x) \cdot \vec{n}(x) = 0\} \quad \text{characteristic boundary}\,. \tag{2.5}$$

Note that $\Gamma_\pm$ are open, and Γ_0 is a closed subset of Γ. Denoting by $\dot\cup$ disjoint union of sets, we have

$$\Gamma = \overline{\Gamma_- \,\dot\cup\, \Gamma_0 \,\dot\cup\, \Gamma_+}\,. \tag{2.6}$$

Furthermore, to simplify the exposition we shall assume throughout this article that

$$c - \frac{1}{2}\nabla \cdot \vec{b} \geq \kappa > 0 \quad \text{in } D \tag{2.7}$$

holds.

2.1.1. *Variational Formulation.* As mentioned in the introduction, in order to employ eventually expansions in L_2-frames, we are interested in finding a variational formulation of (2.1) that allows us to approximate and represent solutions in $L_2(D)$. Using Green's Theorem for $u, v \in C^1(D) \cap C(\overline{D})$, say, yields with the formal adjoint

$$A_\circ^* v = -\vec{b} \cdot \nabla v + v(c - \nabla \cdot \vec{b}) \tag{2.8}$$

of $A_\circ$ the *Green identity*

$$(A_\circ w, v) = (w, A_\circ^* v) + \int_{\Gamma_-} wv(\vec{b} \cdot \vec{n})\, ds + \int_{\Gamma_+} wv(\vec{b} \cdot \vec{n})\, ds, \tag{2.9}$$

where $(\cdot,\cdot) := (\cdot,\cdot)_D$ will always denote the standard $L_2(D)$-inner product. Thus, defining $C^1_{\Gamma_\pm}(D) := \{v \in C^1(D) \cap C(\overline{D}) : v|_{\Gamma_\pm} = 0\}$, one has

$$(A_\circ w, v) = (w, A_\circ^* v), \quad w \in C^1_{\Gamma_-}, v \in C^1_{\Gamma_+}(D)\,. \tag{2.10}$$

Introducing the *graph norm*

$$\|v\|_{W(\vec{b},D)} := \left(\|v\|^2_{L_2(D)} + \int_D |\vec{b} \cdot \nabla v|^2\, dx\right)^{1/2}, \tag{2.11}$$

and noting that

$$\|A_\circ v\|_{L_2(D)} \leq \sqrt{2}\max\left\{1, \|c - \nabla \cdot \vec{b}\|_{L_\infty(D)}\right\}\|v\|_{W(\vec{b},D)}, \tag{2.12}$$

the right hand side of (2.10) still makes sense, even for $w \in L_2(D)$, as long as v belongs to the Hilbert space

$$W(\vec{b}, D) := \mathrm{clos}_{\|\cdot\|_{W(\vec{b},D)}}(C^1(D) \cap C(\overline{D})). \tag{2.13}$$

Obviously, one has

$$H^1(D) \subset W(\vec{b}, D) \subset L_2(D), \tag{2.14}$$

with strict inclusions.

Returning to (2.9), it remains to discuss the trace integrals on the right hand side. To this end, noting that $\omega := |\vec{b} \cdot \vec{n}|$ is positive on $\Gamma_\pm$, consider the weighted L_2-spaces $L_2(\Gamma_\pm, \omega)$, endowed with the norm

$$\|g\|^2_{L_2(\Gamma_\pm,\omega)} = \int_{\Gamma_\pm} |g|^2 \omega ds. \tag{2.15}$$

Thus, the boundary integral terms in (2.9) are well-defined whenever w, v possess meaningful restrictions to $L_2(\Gamma_\mp, \omega)$, respectively. Thinking of w to agree on Γ_- with "inflow boundary data" $g \in L_2(\Gamma_-, \omega)$, the trace integral over Γ_- is well defined provided that the test function v possesses a trace in $L_2(\Gamma_-, \omega)$. Moreover, the second "outflow trace integral" on the right hand side of (2.9)

vanishes when the test function is taken from the closed subspace $W_0(-\vec{b},D)\subset W(\vec{b},D)$, defined by

$$W_0(\mp\vec{b},D):=\mathrm{clos}_{\|\cdot\|_{W(\vec{b},D)}}\{v\in C^1(D)\cap C(\overline{D}),\ v\,|_{\Gamma_\pm}\equiv 0\}. \tag{2.16}$$

Note that $W(\vec{b},D)=W(-\vec{b},D)$ while for the subspaces $W_0(\pm\vec{b},D)$ the sign obviously matters.

Now going back to the integral over Γ_-, it should be noted that elements of $W(\vec{b},D)$ do *not* in general admit traces in $L_2(\Gamma_\pm,\omega)$ (cf. [1, Section 4]): restrictions of elements from $W(\vec{b},D)$ belong apriori only to $L_{2,loc}(\Gamma_\pm)$. However, if $v\in W(\vec{b},D)$ admits a trace in $L_2(\Gamma_+,\omega)$, as $v\in W_0(-\vec{b},D)$ does by definition, then $v|_{\Gamma_-}$ makes sense as an element of $L_2(\Gamma_-,\omega)$. More precisely, under the assumption (2.7), the following facts can be derived from the results in [1, Section 4].

Proposition 2.1. *Under the above hypotheses on D, $\vec{b}$, c assume furthermore that the field $\vec{b}$ has a C^1-extension to an open neighborhood of $\overline{D}$ and that (2.7) holds. Moreover, assume that $\partial\Gamma_-$ is piecewise smooth as well. Then there exist linear continuous mappings*

$$\gamma_\pm : W_0(\pm\vec{b},D)\mapsto L_2(\Gamma_\pm,\omega),\quad \omega:=|\vec{b}\cdot\vec{n}|, \tag{2.17}$$

i.e. there exists a constant C, depending only on $\vec{b},c,D$, with

$$\|\gamma_\pm(v)\|_{L_2(\Gamma_\pm,\omega)}\le C\|v\|_{W(\mp\vec{b},D)},\quad v\in W(\mp\vec{b},D). \tag{2.18}$$

Moreover, for $v\in C^1_{\Gamma_\pm}(D)$ one has $v\,|_{\Gamma_\pm}=\gamma_\pm(v)$.

Denoting now by A^* the continuous extension of $A^*_\circ$ to $W_0(-\vec{b},D)$, the above considerations suggest working with the bilinear form

$$a(w,v):=(w,A^*v):=\int_D w(-\vec{b}\cdot\nabla v+v(c-\nabla\cdot\vec{b}))\,dx, \tag{2.19}$$

which in view of (2.12) is trivially bounded on $L_2(D)\times W_0(-\vec{b},D)$.

Finally, let $(W_0(-\vec{b},D))'$ denote the normed dual of $W_0(-\vec{b},D)$ endowed with the dual norm

$$\|w\|_{(W_0(-\vec{b},D))'}:=\sup_{v\in W_0(-\vec{b},D)}\frac{{}_{(W_0(-\vec{b},D))'}\langle w,v\rangle_{W_0(-\vec{b},D)}}{\|v\|_{W_0(-\vec{b},D)}}$$

where ${}_{(W_0(-\vec{b},D))'}\langle w,v\rangle_{W_0(-\vec{b},D)}$ denotes the dual pairing obtained by continuous extension of the standard inner product for $L_2(D)$. Now let $A:=(A^*)^*$, i.e. $A:L_2(D)\to(W_0(-\vec{b},D))'$ is defined by

$${}_{(W_0(-\vec{b},D))'}\langle Aw,v\rangle_{W_0(-\vec{b},D)}=a(w,v),\quad\forall\ w\in L_2(D),\ v\in W_0(-\vec{b},D). \tag{2.20}$$

We can now present the L_2-stable variational formulation of (2.1),(2.2) which will be the basis for the adaptive Petrov-Galerkin discretization.

Theorem 2.2. *Under the above assumptions on $A_\circ$ and D, let $a(\cdot,\cdot)$ be defined by (2.19). Then, for any $f\in(W_0(\vec{b},D))'$ the variational transport problem*

$$a(u,v)=f(v),\quad\forall\ v\in W_0(-\vec{b},D), \tag{2.21}$$

has a unique variational solution $u\in L_2(D)$ and there exists a constant C, independent of f, such that

$$\|u\|_{L_2(D)}\le C\|f\|_{(W_0(\vec{b},D))'}. \tag{2.22}$$

That is, (2.21) is well-posed in the sense that $\|A^{-1}\|_{(W_0(\vec{b},D))'\to L_2(D)}\le C$ where A is given by (2.20). Moreover, for any $f_\circ\in L_2(D)$ and for any $g\in L_2(\Gamma_-,\omega)$ ($\omega:=|\vec{b}\cdot\vec{n}|$)

$$f(v):=(f,v)+\int_{\Gamma_-}g\gamma_-(v)|\vec{b}\cdot\vec{n}|ds, \tag{2.23}$$

belongs to $(W_0(\vec{b},D))'$ *and whenever the variational solution* u *of (2.21) for* f*, given by (2.23), belongs to* $C^1(D)\cap C(\overline{D})$ *then*

$$A_\circ u(x) = f_\circ(x), \quad \forall\, x\in D, \quad u(x) = g(x),\ \forall\, x\in\Gamma_-\ . \tag{2.24}$$

Remark 2.3. In the variational formulation (2.21), the Dirichlet boundary conditions in (2.24) appear as *natural boundary conditions*. This will allow *conforming discretizations* with *any* subspace $X_h\subset L^2(D)$ which does not have to accommodate essential boundary conditions. The variational formulation (2.21) is the analog of so-called *ultra-weak formulations* of second order, elliptic PDEs.

The linear functional f defined by (2.23) indeed belongs to $(W_0(-\vec{b},D))'$, even when $f_\circ$ is only assumed to belong to $(W_0(-\vec{b},D))'$. In fact, it immediately follows from (2.18) that

$$\begin{aligned} |f(v)| &\le \|f_\circ\|_{(W_0(-\vec{b},D))'}\|v\|_{W(\vec{b},D)} + \|g\|_{L_2(\Gamma_-,\omega)}\|\gamma_-(v)\|_{L_2(\Gamma_-,\omega)}\\ &\lesssim \big(\|f_\circ\|_{(W_0(-\vec{b},D))'} + \|g\|_{L_2(\Gamma_-,\omega)}\big)\|v\|_{W(\vec{b},D)}\end{aligned}$$

(recalling again that $W(\vec{b},D)=W(-\vec{b},D)$).

The claim (2.24) that a sufficiently regular variational solutions is also a classical solution of the original transport equation (2.1), (2.2) follows immediately from integration by parts. The rest of the proof of Theorem 2.2 will be given in Section 3.2.

Although we are mainly interested in the above L_2-formulation we conclude this section remarking that, arguing along similar lines and using again the trace theorems given in [1, Section 4], one could formulate a well posed variational formulation for the pair $X=W_0(\vec{b},D)$, $Y=L_2(D)$, where now, however, the "inflow" Dirichlet boundary conditions are imposed as essential conditions in the trial space X.

2.2. **Parametric Transport Problems.** In important variants of transport problems velocities appear as parameters as in Boltzmann equations, kinetic formulations of conservation laws, relaxation methods for conservation laws or when modeling radiative transfer. One faces particular challenges when trying to represent solutions as functions of the spatial variables as well as the involved parameters. In fact, such problems become *high dimensional* and particular measures have to be taken to deal with these fact, see [23, 24]. We believe that our approach offers particular advantages in this context. The simplest model for radiative transfer can be described as follows.

$$\begin{aligned} A_\circ u(x,\vec{s}) = \vec{s}\cdot\nabla u(x,\vec{s}) + \kappa(x)u(x,\vec{s}) &= f_\circ(x), \quad x\in D\subset\mathbb{R}^d,\ d=2,3,\\ u(x,\vec{s}) &= g(x,\vec{s}),\ x\in\Gamma_-(\vec{s}),\end{aligned} \tag{2.25}$$

where now the solution u depends also on the constant transport direction $\vec{s}$ which, however, varies over a set of directions $\mathcal{S}$. Thus, for instance, when $\mathcal{S}=S^2$, the unit 2−sphere, u is considered as a function of five variables, namely $d=3$ spatial variables and parameters from a two-dimensional set $\mathcal{S}$. Clearly, the in- and outflow boundary now depends on $\vec{s}$, i.e.

$$\Gamma_\pm(\vec{s}) := \{x\in\partial D : \mp\vec{s}\cdot\mathbf{n}(x) < 0\}, \qquad \vec{s}\in\mathcal{S}\ . \tag{2.26}$$

The absorption coefficient $\kappa\in L_\infty(D)$ will always be assumed to be nonnegative in D.

We wish to give now a variational formulation for (1.1) over $D\times\mathcal{S}$. To that end, let

$$\Gamma := \partial D\times\mathcal{S}, \quad \Gamma_\pm := \{(x,\vec{s})\in\Gamma : \mp\vec{s}\cdot\vec{n}(x) < 0\}, \quad \Gamma_0 := \Gamma\setminus(\Gamma_-\cup\Gamma_+), \tag{2.27}$$

and denote as before $(v,w) := (v,w)_{D\times\mathcal{S}} = \int_{D\times\mathcal{S}} v(x,\vec{s})w(x,\vec{s})dxd\vec{s}$ where, however, $d\vec{s}$ is for simplicity the normalized Haar measure on $\mathcal{S}$, i.e. $\int_{\mathcal{S}} d\vec{s}=1$. Following exactly the same lines as in Section 2.1, Fubini's and Green's Theorem yield (first for smooth u,v)

$$\begin{aligned}\int_{D\times\mathcal{S}}(\vec{s}\cdot\nabla u+\kappa u)v dxd\vec{s} &= \int_{\mathcal{S}}\int_D u(-\vec{s}\cdot\nabla v+\kappa v)dxd\vec{s} + \int_{\mathcal{S}}\int_{\partial D} uv(\vec{s}\cdot\vec{n})d\Gamma d\vec{s}\\ &=: (u,A_\circ^* v) + \int_{\Gamma_-} uv(\vec{s}\cdot\vec{n}(x))d\Gamma + \int_{\Gamma_+} uv(\vec{s}\cdot\vec{n}(x))d\Gamma.\end{aligned} \tag{2.28}$$

Defining similarly as in the preceding Section the Hilbert space

$$W(D\times\mathcal{S}) := \{v \in L_2(D\times\mathcal{S}) : \int_{\mathcal{S}\times D} |\vec{s}\cdot\nabla v|^2 dx d\vec{s} < \infty\} \tag{2.29}$$

(in the sense of distributions where the gradient ∇ always refers to the x-variable in D), endowed with the norm $\|v\|_{W(D\times\mathcal{S})}$ given by

$$\|v\|^2_{W(D\times\mathcal{S})} := \|v\|^2_{L_2(D\times\mathcal{S})} + \int_{\mathcal{S}\times D} |\vec{s}\cdot\nabla v|^2 dx d\vec{s}, \tag{2.30}$$

and denoting again the corresponding continuous extension of $A^*_\circ$ by A^*, the bilinear form $a(v,w) := (v, A^*w)$ is continuous on $L_2(D\times\mathcal{S})\times W(D\times\mathcal{S})$. Moreover, as in Section 2.1, a natural test space is $W_0^+(D\times\mathcal{S})$ where

$$W_0^\pm(D\times\mathcal{S}) := \mathrm{clos}_{\|\cdot\|_{W(D\times\mathcal{S})}}\{v \in C(\mathcal{S}, C^1(D)) : v|_{\Gamma_\pm} \equiv 0\} \tag{2.31}$$

which is again a Hilbert space under the norm $\|v\|_{W(D\times\mathcal{S})}$.

As before, (2.28) suggests the variational formulation: find $u \in L_2(D\times\mathcal{S})$ such that for $a(u,v) = (u, A^*v)$, any $g \in L_2(\Gamma_-,\omega)$, and any $f_\circ \in W_0^+(D\times\mathcal{S})'$

$$a(u,v) = f(v) := {}_{W_0^+(D\times\mathcal{S})'}\langle f_\circ, v\rangle_{W_0^+(D\times\mathcal{S})} + \int_{\Gamma_-} g\gamma_-(v)|\vec{s}\cdot\vec{n}|d\Gamma\ . \tag{2.32}$$

For (2.32) to be well-posed, we need to verify the continuity of the trace maps $\gamma_\pm : W_0^\mp(D\times\mathcal{S}) \to L_2(\Gamma_\pm,\omega)$, where $\omega := |\vec{s}\cdot\vec{n}|$. This, in turn, will imply that the functional f defined in (2.32) belongs to $W_0^+(D\times\mathcal{S})'$ and that $W_0^\pm(D\times\mathcal{S})$ in (2.31), equipped with the norm (2.30) are closed, linear subspaces of $W(D\times\mathcal{S})$. The continuity of γ_- is indeed ensured by the results in [1, 4, 5], so that (2.32) is well-defined. We are now ready to formulate the analogue to Theorem 2.2 whose proof is deferred to Section 3.2.

Theorem 2.4. *Let $D \subset \mathbb{R}^d$ be any bounded piecewise smooth Lipschitz domain and let $\kappa \in L_\infty(D)$, $\kappa \geq 0$. Then for any $f_\circ \in W_0^+(D\times\mathcal{S})'$ and any $g \in L_2(\Gamma_-,\omega)$ there exists a unique solution $u \in L_2(D\times\mathcal{S})$ of the variational problem (2.32) and*

$$\|u\|_{L_2(D\times\mathcal{S})} \leq C\|f\|_{W_0^+(D\times\mathcal{S})'}, \tag{2.33}$$

where C depends only on D, b and c. Moreover, the operator $A := (A^*)^*$ satisfies (1.3) for $X = L_2(D\times\mathcal{S})$, $Y = W_0^+(D\times\mathcal{S})$. The restriction of A to $C_0^\infty(D\times\mathcal{S})$ agrees with $A_\circ$ and when $u(\cdot,\vec{s})$ as a function of $x \in D$ belongs to $C(\mathcal{S}, C^1(D)\cap C(\overline{D}))$ it solves (2.25) pointwise for each $\vec{s} \in \mathcal{S}$.

Note that in this formulation once more the homogeneous boundary conditions are natural ones and therefore do not have to be accounted for in the trial spaces of Petrov-Galerkin discretizations. One should further note that the absorbtion coefficient κ need *not* stay bounded away from zero and may even vanish on sets of positive measure.

2.3. **Time Dependent Parametric Transport Problems.** The previous formulation directly extends to *nonstationary* parametric transport problems of the form

$$\begin{aligned}(\partial_t + \vec{s}\cdot\nabla + \kappa(x,t))u(x,t,\vec{s}) &= f_\circ(x,t), \quad (x,t) \in \hat{D} := D\times(0,T),\ \vec{s}\in\mathcal{S},\\ u(x,t,\vec{s}) &= g(x,t,\vec{s}),\ x \in \hat{\Gamma}_-(\vec{s}),\ \vec{s}\in\mathcal{S}\ .\end{aligned} \tag{2.34}$$

We append the variable t to the spatial variables, and assume again that $\kappa(x,t) \geq 0$, $(x,t) \in D_T$, a.e. This implies that now

$$\hat{\Gamma}_\pm(\vec{s}) = (\Gamma_\pm(\vec{s})\times(0,T))\cup(D\times\{\tau_\pm\}),\ \tau_- = 0,\ \tau_+ = T, \quad \vec{s}\in\mathcal{S}\ , \tag{2.35}$$

with $\Gamma_\mp(\vec{s})$ given by (2.26). Obviously, $\hat{\Gamma}_-(\vec{s})$ is that portion of the boundary $\hat{\Gamma} := \partial\hat{D}$ of the space-time cylinder $\hat{D} = D\times(0,T)$ for which the space-time "flow direction" $(\vec{s},1)$ is an inflow-direction.

An $L_2(D\times(0,T)\times\mathcal{S})$-stable variational formulation can now be immediately derived from the findings in the previous section. In fact, setting

$$\hat{x} := (x,t)\in \hat{D} = D\times(0,T),\quad \hat{\vec{s}} := (\vec{s},1)\in\hat{\mathcal{S}} := \mathcal{S}\times\{1\},\quad \hat{\nabla} := (\nabla,\partial_t), \tag{2.36}$$

(2.34) takes the form (2.25), the roles of $D,\mathcal{S},\vec{s},x$ being played by $\hat{D},\hat{\mathcal{S}},\hat{\vec{s}},\hat{x}$. The sets $\hat{\Gamma}_\pm(\vec{s})\subset\partial\hat{D}$ are exactly defined by (2.26) with respect to these substitutions. Looking again for the solution as a function of $(\hat{x},\vec{s}) = (x,t,\vec{s})$ (identifying $\mathcal{S}$ with $\mathcal{S}\times\{1\}$) and setting

$$\hat{\Gamma}_\pm := \{(\hat{x},\hat{\vec{s}}) : \hat{x}\in\hat{\Gamma}_\pm(\vec{s}),\ \vec{s}\in\mathcal{S}\} \tag{2.37}$$

the spaces $W(\hat{D}\times\hat{\mathcal{S}}), W_0^\pm(\hat{D}\times\hat{\mathcal{S}})$ are defined by (2.29), (2.31) with the replacements from (2.36). Note that, denoting by $\hat{\vec{n}}(\hat{x})$ the outward unit normal at $\hat{x}\in\partial\hat{D}$, we have

$$\hat{\vec{s}}\cdot\hat{\vec{n}}(\hat{x}) = \vec{s}\cdot\vec{n}(x) \text{ for } t\in(0,T),\quad \hat{\vec{s}}\cdot\hat{\vec{n}}((x,0)) = -1,\quad \hat{\vec{s}}\cdot\hat{\vec{n}}((x,T)) = 1,\ x\in D. \tag{2.38}$$

Therefore, defining for the operator $\hat{A}_\circ = (\partial_t + \vec{s}\cdot\nabla+\kappa)$ in (2.34) the formal adjoint $\hat{A}_\circ{}^*$ by

$$(u,\hat{A}_\circ{}^* v) = \int_{\hat{D}\times\mathcal{S}} u(x,t,\vec{s})(-\partial_t - \vec{s}\cdot\nabla+\kappa)v(x,t,\vec{s})dxdtd\vec{s}, \tag{2.39}$$

where $(\cdot,\cdot) = (\cdot,\cdot)_{\hat{D}\times\mathcal{S}}$, the identity (2.28) immediately yields in these terms

$$\begin{aligned}
&\int_D\int_0^T\int_{\mathcal{S}}\big((\partial_t+\vec{s}\cdot\nabla+\kappa)u\big)v dx\,dt\,d\vec{s}\\
&= (u,\hat{A}_\circ^* v) + \int_{\hat{\Gamma}_-} uv(\hat{\vec{s}}\cdot\hat{\vec{n}}))d\hat{\Gamma} + \int_{\hat{\Gamma}_+} uv(\hat{\vec{s}}\cdot\hat{\vec{n}}))d\hat{\Gamma}\\
&= (u,\hat{A}_\circ^* v) + \int_0^T\int_{\Gamma_-} uv(\vec{s}\cdot\vec{n}))d\Gamma dt + \int_0^T\int_{\Gamma_+} uv(\vec{s}\cdot\vec{n}))d\Gamma dt\\
&\quad - \int_{D\times\mathcal{S}} u(x,0,\vec{s})v(x,0,\vec{s})dxd\vec{s} + \int_{D\times\mathcal{S}} u(x,T,\vec{s})v(x,T,\vec{s})dxd\vec{s},
\end{aligned} \tag{2.40}$$

with $\Gamma_\pm$ from (2.27). Finally, let

$$\hat{\omega}(\hat{x},\vec{s}) := \begin{cases} |\vec{s}\cdot\vec{n}| & \text{when } \hat{x}\in\partial D\times(0,T);\\ 1 & \text{when } \hat{x} = (x,0),(x,T),\ x\in D.\end{cases}$$

Then, defining in analogy to (2.28) $\hat{a}(u,v) := (u,\hat{A}^* v)$, where $\hat{A}^*$ is the continuous extension of $\hat{A}_\circ^*$ to $W_0^+(\hat{D}\times\mathcal{S})$, (2.32) takes the form: for any $g\in L_2(\hat{\Gamma}_-,\hat{\omega})$, and any $f_\circ\in W_0^+(\hat{D}\times\hat{\mathcal{S}})'$ find $u\in L_2(D\times(0,T)\times\mathcal{S})$ such that

$$\begin{aligned}
\hat{a}(u,v) &= f(v) := {}_{W_0^+(\hat{D}\times\mathcal{S})'}\langle f,v\rangle_{W_0^+(\hat{D}\times\mathcal{S})} + \int_{\hat{\Gamma}_-} g\hat{\gamma}_-(v)d\hat{\Gamma}\\
&= {}_{W_0^+(\hat{D}\times\mathcal{S})'}\langle f,v\rangle_{W_0^+(\hat{D}\times\mathcal{S})} + \int_0^T\int_{\Gamma_-} g\hat{\gamma}_-(v)|\vec{s}\cdot\vec{n}|d\Gamma dt + \int_{D\times\mathcal{S}}\hat{\gamma}_-(v)(x,0,\vec{s})dxd\vec{s},
\end{aligned} \tag{2.41}$$

where again Γ_- is given by (2.27) and $\hat{\gamma}_-$ is the continuous trace map associated with $\hat{\Gamma}_-$.

Now we can state the following immediate consequence of Theorem 2.4.

Corollary 2.5. *Let $\hat{D}$ be defined by (2.36) where $D\subset R^d$ is any bounded piecewise smooth Lipschitz domain and let $\kappa\in L_\infty(\hat{D})$, $\kappa\geq 0$ a.e. in $\hat{D}$. Then for any $f_\circ\in W_0^+(\hat{D}\times\mathcal{S})'$ and any $g\in L_2(\hat{\Gamma}_-,\hat{\omega})$ there exists a unique $u\in L_2(\hat{D}\times\mathcal{S})$ of (2.41) and*

$$\|u\|_{L_2(\hat{D}\times\mathcal{S})}\leq C\|f\|_{W_0^+(\hat{D}\times\mathcal{S})'}, \tag{2.42}$$

where C depends only on D and T. Moreover, the operator $\hat{A}$ satisfies (1.3) for $X = L_2(\hat{D}\times\mathcal{S})$, $Y = W_0^+(\hat{D}\times\mathcal{S})$. On elements in $C_0^\infty(\hat{D}\times\mathcal{S})$ the operator $\hat{A}$ agrees with $\hat{A}_\circ$ and when $u\in C(\mathcal{S},C^1(\hat{D}))$ (and continuous data) it solves (2.34) for each $\vec{s}\in\mathcal{S}$.

3. Abstract Framework

The well posedness of the above variational formulations are consequences of the following general facts. We formulate them in a general abstract framework for two reasons. It identifies the key conditions for being able to *choose* pairs X, Y for well posed variational formulations, prescribing either X or Y. Second, it suggests an "ideal setting" for Petrov-Galerkin schemes from which numerically feasible versions can be derived in a systematic manner. A slightly different but related approach has been proposed in [9] for a more restricted scope of problems.

3.1. **The Basic Principle.**

Our starting point is the linear operator equation

$$A_\circ u = f_\circ, \tag{3.1}$$

on some spatial domain Ω where homogeneous side conditions are imposed as essential boundary conditions in the (classical) domain $D(A_\circ)$ of $A_\circ$. For the examples in Section 2, $\Omega \in \{D \times (0,T), D, D \times \mathcal{S}, D \times \mathcal{S} \times (0,T)\}$. In the distributional sense we may think of $A_\circ$ to act on a possibly larger Hilbert space $X \supseteq D(A_\circ)$, endowed with norm $\|\cdot\|_X$ with dense embedding, that is to host the solution of an associated abstract variational formulation of (3.1). Moreover, we shall assume that the dual pairing ${}_{X'}\langle w, v\rangle_X$ is induced by the inner product $(\cdot,\cdot)$ on some pivot Hilbert space H which we identify with its dual, thereby forming the Gelfand triple

$$D(A_\circ) \subseteq X \hookrightarrow H \hookrightarrow X', \tag{3.2}$$

again all embeddings being dense. In Section 2.3 we had $H = L_2(D \times (0,T))$ and $X = L_2(D \times \mathcal{S} \times (0,T))$. In Sections 2.1, 2.2 we had $X = H = L_2(D)$, $X = H = L_2(D \times \mathcal{S})$, respectively.

It is perhaps worth stressing that our analysis rests on two essential assumptions on the formal adjoint $A_\circ^*$ of $A_\circ$ given by $(A_\circ v, w) = (v, A_\circ^* w)$, for all $v, w \in C_0^\infty(\Omega)$:

(A*1): $A_\circ^*$ is injective on the dense subspace $D(A_\circ^*) \subseteq H$;
(A*2): the range $R(A_\circ^*)$ of $A_\circ^*$ is densely embedded in X'.

Now let

$$|\!|\!| v |\!|\!| := \|A_\circ^* v\|_{X'}, \tag{3.3}$$

which is a norm on $D(A_\circ^*)$. Hence

$$Y := \mathrm{clos}_{|\!|\!|\cdot|\!|\!|} D(A_\circ^*) \subseteq H \tag{3.4}$$

is a Hilbert space with inner product

$$\langle v, w\rangle_Y = \langle A^* v, A^* w\rangle_{X'}, \quad \|v\|_Y = |\!|\!| v |\!|\!| = \|A^* v\|_{X'}, \tag{3.5}$$

where $A^* : Y \to X'$ denotes the continuous extension of $A_\circ^*$ from $D(A_\circ^*)$ to Y.

We denote for any Gelfand triple $Z \hookrightarrow H \hookrightarrow Z'$ the corresponding inner products by $\langle\cdot,\cdot\rangle_Z$, $\langle\cdot,\cdot\rangle_{Z'}$ while ${}_Z\langle\cdot,\cdot\rangle_{Z'}$ (or briefly $\langle\cdot,\cdot\rangle$ when the roles of Z, Z' are clear from the context) denote the corresponding dual pairing.

As in the previous section we define by duality now $A : X \to Y'$ by

$${}_{Y'}\langle Aw, v\rangle_Y = {}_X\langle w, A^* v\rangle_{X'}, \quad \forall\, w \in X,\, v \in Y. \tag{3.6}$$

In what follows it will be convenient to employ the notion of Riesz map. For a given Hilbert space Z let us denote by $R_Z : Z \to Z'$ the Riesz map defined by

$$\langle v, w\rangle_Z = {}_Z\langle v, R_Z w\rangle_{Z'}, \quad \forall\, v, w \in Z. \tag{3.7}$$

so that, in particular,

$$\langle v, w\rangle_{Z'} = {}_{Z'}\langle v, R_Z^{-1} w\rangle_Z, \quad \forall\, v, w \in Z'. \tag{3.8}$$

One readily verifies that

$$\|R_Z w\|_{Z'} = \|w\|_Z. \tag{3.9}$$

Proposition 3.1. *The mappings* $A : X \to Y', \quad A^* : Y \to X'$ *are isometries, i.e.*

$$Y = A^{-*}X', \quad X = A^{-1}Y', \tag{3.10}$$

and

$$1 = \|A^*\|_{Y\to X'} = \|A\|_{X\to Y'} = \|A^{-1}\|_{Y'\to X} = \|A^{-*}\|_{X'\to Y}, \tag{3.11}$$

where we sometimes use the shorthand notation A^{-*} *for* $(A^*)^{-1}$.

Proof: First we infer from (3.3), (3.5) that $A^* \in \mathcal{L}(Y, X')$, the space of bounded linear operators from Y to X', and hence $A \in \mathcal{L}(X, Y')$. Again, by the definition of the graph norm (3.3) and by duality we conclude

$$1 = \|A^*\|_{Y\to X'} = \|A\|_{X\to Y'}. \tag{3.12}$$

Moreover, denseness of $D(A^*_\circ)$ in X' and injectivity of $A^*_\circ$ on $D(A^*_\circ)$ implies injectivity of the dense extension A^* of $A^*_\circ$ as a mapping from Y onto X'. By the "Bounded Inverse Theorem" we conclude that $A^{-*} : X' \to Y$ is also bounded. Therefore we infer that for $v, w \in D(A^*_\circ)$ (recalling that $R(A^*_\circ) \subseteq X'$ and the definition (3.5) of $\langle\cdot,\cdot\rangle_Y$)

$$\langle v, w\rangle_Y = \langle A^*v, A^*w\rangle_{X'} = {}_X\langle R_X^{-1}A^*v, A^*w\rangle_{X'} = {}_{Y'}\langle AR_X^{-1}A^*v, w\rangle_Y,$$

which, in view of (3.7), means that

$$R_Y = AR_X^{-1}A^*. \tag{3.13}$$

Thus, by (3.8),

$$\langle v, w\rangle_{Y'} = {}_{Y'}\langle v, A^{-*}R_XA^{-1}w\rangle_Y = {}_X\langle A^{-1}v, R_XA^{-1}w\rangle_{X'} = \langle A^{-1}v, A^{-1}w\rangle_X, \tag{3.14}$$

providing

$$\|v\|_{Y'} = \|A^{-1}v\|_X, \quad \forall\, v \in Y', \tag{3.15}$$

which means

$$\|A^{-1}\|_{Y'\to X} = 1. \tag{3.16}$$

Combining (3.16) with (3.12) and using duality confirms (3.11). □

As in the examples from Section 2.1 we define now the bilinear form

$$a(\cdot,\cdot) : X \times Y \to \mathbb{R}, \quad a(w, v) := {}_X\langle w, A^*v\rangle_{X'} = {}_{Y'}\langle Aw, v\rangle_Y, \quad w \in X,\, v \in Y, \tag{3.17}$$

to state the following immediate consequence of Proposition 3.1.

Corollary 3.2. *Assume (A*1) and (A*2). Then for* Y *defined through (3.4) the problem: given* $f \in Y'$ *find* $u \in X$ *such that*

$$a(u, v) = {}_{Y'}\langle f, v\rangle_Y \qquad \forall\, v \in Y. \tag{3.18}$$

is well posed and the induced operator $A = (A^*)^*$, *given by* ${}_{Y'}\langle Av, w\rangle_Y = a(v, w)$, $v \in X, w \in Y$ *satisfies*

$$\|A\|_{X\to Y'},\ \|A^{-1}\|_{Y'\to X} = 1. \tag{3.19}$$

Thus, in the topologies defined above, the variational formulation (3.18) of the transport problem is perfectly conditioned.

3.2. **Proof of Theorems 2.2, 2.4.** We begin with the proof of Theorem 2.2. First, applying integration by parts to (v, A^*v) for $0 \neq v \in C^1_{\Gamma_+} = D(A^*_\circ)$ (which is dense in $L_2(D)$) and adding both representations, the skew-symmetric terms cancel. This yields

$$2(v, A^*_\circ v) = \int_D (2c - \nabla\cdot\vec{b})|v|^2 dx - \int_{\Gamma_-} |v|^2 \vec{b}\cdot\vec{n} ds \geq \int_D (2c - \nabla\cdot\vec{b})|v|^2 dx \geq 2\kappa\|v\|^2_{L^2(D)}, \tag{3.20}$$

where we have used (2.7) in the last step. Thus, condition (A*1) holds for $X = X' = H = L_2(D)$. Moreover, (A*2) follows, under the given assumptions on $\vec{b}, c$, from the results in [1, Section 4]. Therefore, Proposition 3.1 applies and confirms that A^* is an isometry from Y, given by (3.4), onto $X' = X = L_2(D)$.

Next, observe that

$$Y = W_0(-\vec{b}, D). \tag{3.21}$$

In fact, the inclusion $W_0(-\vec{b}, D) \subseteq Y$ follows immediately from (2.12). The converse inclusion is a consequence of the following fact.

Proposition 3.3. *Whenever (2.7) holds, then* $\|\cdot\|_{W(\vec{b},D)}$ *and* $\|A^*\cdot\|_{L_2(D)}$ *are equivalent norms on* $W_0(-\vec{b}, D)$.

Proof: Abbreviating $c_0 := \|c - \nabla\cdot\vec{b}\|_{L_\infty(D)}$ and using the definition of A^*, we already know from (2.12)

$$\|A^*v\|_{L_2(D)} \leq \sqrt{2}\max\{1, c_0\}\|v\|_{W(\vec{b},D)}, \quad v \in W_0(-\vec{b}, D). \tag{3.22}$$

To establish the converse estimate we infer from (3.20) and (2.7) that

$$\|A^*v\|_{L_2(D)} = \sup_{w\in L_2(D)} \frac{(w, A^*v)}{\|w\|_{L_2(D)}} \geq \frac{(v, A^*v)}{\|v\|_{L_2(D)}} \geq \kappa\|v\|_{L_2(D)}. \tag{3.23}$$

On the other hand, by the triangle inequality

$$\begin{aligned} \|v\|^2_{W(\vec{b},D)} &\leq (\|A^*v\|_{L_2(D)} + c_0\|v\|_{L_2(D)})^2 + \|v\|^2_{L_2(D)} \\ &\leq 2\|A^*v\|^2_{L_2(D)} + (2c_0^2+1)\|v\|^2_{L_2(D)} \leq (2 + \kappa^{-2}(2c_0^2+1))\|A^*v\|^2_{L_2(D)}, \end{aligned} \tag{3.24}$$

where we have used (3.23) in the last step. □

The assertion of Theorem 2.2 is therefore an immediate consequence of Proposition 3.3 and (3.21) combined with Corollary 3.2. □

Remark 3.4. The above reasoning for showing that the operator A, defined by (2.20), is an isomorphism from $L_2(D)$ onto $(W_0(-\vec{b}, D))'$ can, of course, be interpreted as verifying a classical *inf-sup condition* for the bilinear form $a(\cdot,\cdot)$, given by (2.19). In fact, the continuity of the bilinear form $a(\cdot,\cdot) : L_2(D) \times W_0(-\vec{b}, D) \to \mathbb{R}$ follows from (2.12), the fact that

$$\forall 0 \neq v \in W_0(-\vec{b}, D) \quad \text{one has} \quad \sup_{w\in L_2(D)} a(w, v) > 0\,,$$

is an immediate consequence of (3.20) combined with a density argument for $w = v \in W_0(-\vec{b}, D) \subset L_2(D)$, while, denoting by $c^2_{A^*}$ the constant on the right hand side of (3.24), we infer from (3.11) that

$$\inf_{w\in L_2(D)} \sup_{v\in W_0(-\vec{b},D)} \frac{a(w, v)}{\|w\|_{L_2(D)}\|v\|_{W(\vec{b},D)}} \geq \inf_{w\in L_2(D)} \sup_{v\in W_0(-\vec{b},D)} \frac{a(w, v)}{\|w\|_{L_2(D)} c_{A^*}\|v\|_Y} = 1/c_{A^*} > 0\,. \tag{3.25}$$

which is the desired inf-sup condition.

Remark 3.5. Assumption (2.7) was used in establishing injectivity of $A_\circ^*$ (condition (A*1)) and in the coercivity estimate yielding the last inequality in (3.23). An inequality of the form

$$\|v\|_{L_2(D)} \le c\|A^*v\|_{L_2(D)}, \quad v \in W_0(-\vec{b}, D), \tag{3.26}$$

in turn, is the essential step in proving the norm equivalence in Proposition 3.3, see (3.24). Hence, the assertion of Theorem 2.2 remains valid *without* assumption (2.7), as long as the denseness of the range $R(A_\circ^*)$ in $L_2(D)$ (condition (A*2)) and injectivity of A^* on $W_0(-\vec{b}, D)$ (implying condition (A*1)) can be shown. In fact, the mapping A^* is trivially bounded in the norm $|\!|\!|v|\!|\!| := \|A^*v\|_{L_2(D)}$. By the *Bounded Inverse Theorem* $A^{-*} : R(A^*) \mapsto W_0(-\vec{b}, D)$ is then also bounded. This implies that

$$\forall v \in W_0(-\vec{b}; D): \quad \|v\|_{L_2(D)} \le \|A^{-*}\|_{R(A^*)\to W_0(-\vec{b},D)}\|A^*v\|_{L_2(D)}$$

which an inequality of the form (3.26) and one may argue as before in (3.24).

Concerning Theorem 2.4, choose

$$X = H = L_2(\Omega) = X', \tag{3.27}$$

for $\Omega = D \times \mathcal{S}$. Due to the fact that the flow field $\vec{s}$ is constant as a function of x, $A_\circ^*$, defined by (2.28), is clearly injective on its domain $D(A_\circ^*)$, containing $\{v \in C^1(\Omega) \cap C(\overline{\Omega}) : v|_{\Gamma_+} = 0\}$, which in turn is dense in $X = X' = H = L_2(\Omega)$. Since (2.25) possesses a classical solution for smooth data the range $R(A_\circ^*)$ is dense in $X' = X = L_2(\Omega)$. Thus, the conditions (A*1), (A*2) hold and Proposition 3.1 or Corollary 3.2 applies. To complete the proof we only need to verify the equivalence of the norms $\|A^* \cdot\|_{L_2(\Omega)}$ and $\|\cdot\|_{W(D\times\mathcal{S})}$ which would then again imply $Y = W_0^+(D \times \mathcal{S})$. The norm equivalence follows again as explained in Remark 3.5 by establishing an estimate of the type

$$\|v\|_{L_2(D\times\mathcal{S})} \le c\|A^*v\|_{L_2(D\times\mathcal{S})}, \quad v \in W_0^+(D \times \mathcal{S}),$$

using injectivity and the Bounded Inverse Theorem. □

3.3. **Ideal Petrov-Galerkin Discretization.** Suppose that $X_h \subset X$ is any subspace of X and observe that for any $u_h \in X_h$ and u the solution of (3.18) one has by (3.15) that

$$\|u - u_h\|_X = \|A(u - u_h)\|_{Y'} = \|Au_h - f\|_{Y'}, \tag{3.28}$$

which means

$$u_h = \mathrm{argmin}_{v_h\in X_h}\|u - v_h\|_X \quad \Leftrightarrow \quad u_h = \mathrm{argmin}_{v_h\in X_h}\|Av_h - f\|_{Y'}, \tag{3.29}$$

i.e. the best X-approximation to the solution u of (3.18) is given by the *least squares solution* of the residual in the Y'-norm. This latter problem is equivalent to a Petrov-Galerkin scheme.

Remark 3.6. u_h solves (3.29) if and only if u_h is the solution of

$$a(u_h, y_h) = {}_{Y'}\langle f, y_h\rangle_Y, \quad \forall\, y_h \in Y_h := A^{-*}R_X X_h. \tag{3.30}$$

Proof: The normal equations for the right hand extremal problem in (3.29) read

$${}_{Y'}\langle Au_h, Aw_h\rangle_{Y'} = {}_{Y'}\langle f, Aw_h\rangle_{Y'}, \quad \forall\, w_h \in X_h.$$

By the first relation in (3.14) we have

$${}_{Y'}\langle Au_h, Aw_h\rangle_{Y'} = {}_{Y'}\langle Au_h, A^{-*}R_X w_h\rangle_Y, \quad {}_{Y'}\langle f, Aw_h\rangle_{Y'} = {}_{Y'}\langle f, A^{-*}R_X w_h\rangle_Y$$

which is (3.30). □

Remark 3.7. Our subsequent discussion will be guided by the following interpretation of the above findings. The operator A_h defined by

$${}_{Y'}\langle A_h u_h, y_h\rangle_Y = a(u_h, y_h), \quad \forall\, u_h \in X_h,\ y_h \in Y_h = A^{-*}R_X X_h \subset Y, \tag{3.31}$$

is perfectly well conditioned, i.e. for every $h > 0$:

$$\|A_h\|_{X\to Y'} = 1 = \|A_h^{-1}\|_{Y'\to X}\ . \tag{3.32}$$

This implies that $(X_h, A^{-*}R_X X_h)$ would be an ideal Petrov-Galerkin pair for *any* choice of $X_h \subset X$.

4. Perturbing the Ideal Case

Of course, employing the ideal test space $Y_h := A^{-*}R_X X_h$ for numerical purposes is not feasible, in general. However, many stabilized Finite Element Methods for convection dominated problems could be viewed as employing test spaces whose structure mimics the ideal test space Y_h. The choice of approximation must warrant uniform stability. We address next the key question what kind of perturbations of the ideal test spaces Y_h preserve stability in a corresponding Petrov-Galerkin discretization.

4.1. A Stability Condition.

A natural idea, proposed in the context of Discontinuous Galerkin FE discretizations in [14] is to approximate *individually* ideal test basis functions. Since stability of the PG discretization depends on *subspaces* rather than individual basis elements, it is more appropriate to approximate *whole* subspaces. To this end, we assume that a one-parameter of finite dimensional trial spaces

$$\{X_h\}_{h>0} \subset X \tag{4.1}$$

is given, which are dense in X, i.e.

$$\forall h > 0: \quad N(h) = \dim X_h < \infty\,, \quad \text{and } \overline{\{X_h\}_{h>0}}^{\|\cdot\|_X} = X. \tag{4.2}$$

We have already seen in Section 3.3 that the corresponding ideal test spaces

$$Y_h := A^{-*}R_X X_h \tag{4.3}$$

form a family of test spaces which is dense in Y and for which the Petrov-Galerkin discretizations (3.30) are perfectly stable (see Remark 3.7). As said before, the Y_h are not numerically feasible. Instead we shall employ test spaces which are close to Y_h in the following sense:

Definition 4.1. For $\delta \in (0,1)$, a subspace $Y_h^\delta \subset Y$ with $\dim Y_h^\delta = N(h) = \dim X_h$ is called *δ-proximal* for X_h if

$$\forall\, 0 \neq y_h \in Y_h \;\exists\; \tilde{y}_h \in Y_h^\delta \text{ such that } \|y_h - \tilde{y}_h\|_Y \leq \delta \|y_h\|_Y. \tag{4.4}$$

In principle, δ in (4.4) may depend on h, for instance, tending to zero as the dimension of the X_h grows, thereby matching the ideal case in a better and better way. However, regarding computational efficiency, it will be preferable to work with a fixed δ. At any rate, the above proximality property will be seen to imply a *stability property*, even when δ is fixed independent of the dimensions $N(h)$. We emphasize that (4.4) is a *relative accuracy requirement*. Several ramifications of this fact will be discussed later.

Lemma 4.2. *For any subspace $X_h \subset X$ and any δ-proximal $Y_h^\delta \subset Y$ the bilinear form $a(v,z) := {}_{Y'}\langle Av, z\rangle_Y$ satisfies*

$$\inf_{v_h \in X_h} \sup_{z_h \in Y_h^\delta} \frac{a(v_h, z_h)}{\|v_h\|_X \|z_h\|_Y} \geq \frac{1-\delta}{1+\delta}, \tag{4.5}$$

so that for any $\delta < 1$ the pair of spaces X_h, Y_h^δ satisfy a discrete inf-sup condition uniformly in $h > 0$.

Proof: For $v_h \in X_h$, by (4.3) we choose $y_h := A^{-*}R_X v_h \in Y_h$ and recall that $\|v_h\|_X = \|y_h\|_Y$. We then pick a $\tilde{y}_h \in Y_h^\delta$ satisfying (4.4). Note that for this $\tilde{y}_h$ one has

$$(1-\delta)\|y_h\|_Y \leq \|\tilde{y}_h\|_Y \leq (1+\delta)\|y_h\|_Y. \tag{4.6}$$

Then we have

$$\begin{aligned} \frac{a(v_h, \tilde{y}_h)}{\|\tilde{y}_h\|_Y} &\geq \frac{a(v_h, y_h)}{\|\tilde{y}_h\|_Y} - \frac{\|v_h\|_X \|y_h - \tilde{y}_h\|_Y}{\|\tilde{y}_h\|_Y} = \frac{\|v_h\|_X^2 - \|v_h\|_X \|y_h - \tilde{y}_h\|_Y}{\|\tilde{y}_h\|_Y} \\ &= \frac{(1-\delta)\|y_h\|_Y^2}{\|\tilde{y}_h\|_Y} \geq \frac{1-\delta}{1+\delta}\|y_h\|_Y = \frac{1-\delta}{1+\delta}\|v_h\|_X, \end{aligned}$$

which is (4.5). □

Theorem 4.3. *For $a(\cdot,\cdot)$ as above the Petrov-Galerkin discretization*

$$\text{find } u_{h,\delta} \in X_h: \quad a(u_{h,\delta}, v_h) = \ell(v_h) \quad \forall v_h \in Y_h^\delta \tag{4.7}$$

where Y_h^δ is δ-proximal for X_h, is uniformly stable for $h > 0$. In particular, the operator A_h defined by

$$a(v_h, z_h) = {}_{Y'}\langle A_h v_h, z_h\rangle_Y, \quad \forall\ v_h \in X_h,\ z_h \in Y_h^\delta, \tag{4.8}$$

satisfies for $\delta < 1$

$$\|A_h\|_{X\to Y'} \le 1, \quad \|A_h^{-1}\|_{Y'\to X} \le \frac{1+\delta}{1-\delta}. \tag{4.9}$$

Moreover, the solution $u_{h,\delta}$ of (4.7) satisfies

$$\|u - u_{h,\delta}\|_X \le \frac{2}{1-\delta} \inf_{v_h\in X_h} \|u - v_h\|_X. \tag{4.10}$$

Proof: Since by Lemma 4.2

$$\|A_h v_h\|_{Y'} \ge \sup_{z_h\in Y_h^\delta} \frac{a(v_h, z_h)}{\|z_h\|_Y} \ge \frac{1-\delta}{1+\delta}\|v_h\|_X,$$

which means that A_h is injective. Since $\dim Y_h^\delta = \dim X_h$, A_h is also surjective and hence bijective and the second relation in (4.9) follows. Because of

$$\begin{aligned}\|A_h v_h\|_{Y'} &= \sup_{z\in Y} \frac{{}_{Y'}\langle A_h v_h, z\rangle_Y}{\|z\|_Y} = \sup_{z\in Y} \frac{a(v_h, z)}{\|z\|_Y} = \sup_{z\in Y} \frac{{}_X\langle v_h, A^* z\rangle_{X'}}{\|z\|_Y} \\ &\le \sup_{z\in Y} \frac{\|v_h\|_X \|A^* z\|_{X'}}{\|z\|_Y} = \sup_{z\in Y} \frac{\|v_h\|_X \|z\|_Y}{\|z\|_Y},\end{aligned}$$

where we have used (3.3) in the last step. This confirms also the first bound in (4.9). Finally, (4.10) follows from standard estimates (see also the proof of Lemma 5.3 below). □

4.2. **A Projection Approach.** The above stability result is a statement concerning *spaces* not about specific basis representations and resulting concrete linear systems. Of course, it remains to see how and at which computational cost δ-proximal subspaces Y_h^δ can be found for a given trial space X_h. We shall describe now *one* possible framework for the construction of δ-proximal subspaces, specifications of which will be discussed later.

To that end, we shall focus from now on the specific case

$$X = X' = L_2(\Omega), \qquad R_X = I, \quad (\cdot,\cdot) := \langle\cdot,\cdot\rangle_{L_2(\Omega)}. \tag{4.11}$$

Recall from (3.5) that this means, in particular,

$$\langle y, z\rangle_Y = (A^* y, A^* z), \quad \|y\|_Y = \|A^* y\|_{L_2(\Omega)}. \tag{4.12}$$

Suppose that $Z_h \subset Y$ is a finite dimensional auxiliary space which is associated with X_h and the ideal test space $Y_h = A^{-*}X_h$ in a way to be specified later. One should think of Z_h at this point to be large enough to approximate any $A^{-*}w_h$, $w_h \in X_h$, with sufficient relative accuracy in Y. In particular, one should therefore have $\dim Z_h \ge \dim X_h$. Let $P_h : Y \to Z_h$ be the Y-orthogonal projection given by

$$\langle P_h y, z_h\rangle_Y = \langle y, z_h\rangle_Y, \quad \forall\ z_h \in Z_h. \tag{4.13}$$

A natural candidate for a δ-proximal subspace for X_h is then the Y-orthogonal projection of the ideal test space Y_h into Z_h, i.e.

$$\tilde{Y}_h := P_h(Y_h) = P_h(A^{-*}X_h) \subset Z_h \subset Y. \tag{4.14}$$

Although the $A^{-*}w$, $w \in X_h$ are, of course, not known, the projections of these ideal test elements can be computed exactly. In fact, by definition of the Y-inner product (4.12), the projection $\tilde{y}_h = P_h y_h$ of $y_h = A^{-*}w_h \in Y_h$ is given by (4.12),

$$(A^*\tilde{y}_h, A^*z_h) = (w_h, A^*z_h), \quad \forall\ z_h \in Z_h. \tag{4.15}$$

Remark 4.4. We have for any $y_h := A^{-*}w_h \in Y_h$

$$\inf_{v_h \in \tilde{Y}_h} \|y_h - v_h\|_Y = \|y_h - P_h y_h\|_Y. \tag{4.16}$$

Moreover, $\tilde{Y}_h$ is δ-proximal for X_h if and only if

$$\|w_h - A^*P_hA^{-*}w_h\|_{L_2(\Omega)} \le \delta\|w_h\|_{L_2(\Omega)} \quad \forall\ w_h \in X_h. \tag{4.17}$$

Thus, in this case the δ-proximal subspace is the Y best-approximation of the ideal test space Y_h from some finite dimensional space Z_h.

It will be important later that $P_h\tilde{y}_h$ produces a Galerkin approximation to $(AA^*)^{-1}w_h$.

The above framework is not tied to specific choices of trial spaces, keeping the option of employing directional representation systems as well as finite element spaces. The present paper is to bring out the principal mechanisms. A more detailed analysis for specific examples and more elaborate numerical tests will be given in [12]. In our first numerical experiments below in Section 6, the spaces $X_h = \mathbb{P}_{p,\mathcal{T}_h}$ are comprised of piecewise polynomials of degree p on hierarchies of isotropically refined partitions $\mathcal{T}_h$ of $\Omega = D$, and the test spaces Z_h are simply H^1-conforming finite elements on partitions arising from a fixed number r of local refinements of the X_h-partition. In fact, $r = 1, 2$ turned out to suffice in all practical examples. One could equally well employ such piecewise polynomials on *anisotropic meshes* without affecting stability. Alternatively, other local enrichment strategies such as increasing the polynomial order or augmentation by so-called "bubble functions" are conceivable, see e.g. [14, 2].

5. Residual Evaluation and Adaptive Refinements

Suppose that we have determined for a hierarchy of trial spaces X_h corresponding δ-proximal test spaces Y_h^δ. A direct realization of the corresponding Petrov-Galerkin discretization would require

(a) Compute a basis of Y_h^δ, e.g. by $\theta_{h,i} = P_h(A^{-*}\phi_{h,i})$ when the $\phi_{h,i}$ form a basis for X_h;
(b) assemble the stiffness matrix $\mathbf{A}_h = \big(a(\phi_{h,j}, \theta_{h,i})\big)_{i,j\in\mathcal{I}_h}$;
(c) compute the load vector $\mathbf{f}_h = \big({}_{Y'}\langle f, \theta_{h,i}\rangle_Y\big)_{i\in\mathcal{I}_h}$.

Since $\dim Z_h$ is (at least) of the order of $\dim X_h$, the complexity of each projection in (a) is of the order of the problem size. For transport problems the $\theta_{h,i}$ are expected to have nonlocal support which also causes unacceptable computational cost in (b) and (c).

5.1. **The Basic Strategy.** In this section we propose a strategy that circumvents the above obstructions. We shall always assume that $X = X'$ equals the pivot space H, having mainly $H = L_2(\Omega)$ in mind as in Sections 2.1, 2.2. When $X \neq X'$ an additional Riesz map R_X would enter the subsequent considerations and we prefer to avoid this technical complication, having mainly the examples in Sections 2.1, 2.2 in mind. We continue to abbreviate $(\cdot,\cdot) = \langle\cdot,\cdot\rangle_X$.

To motivate our approach, in view of the mapping properties of A, see (3.19), Remark 3.1, we can write the operator equation (3.1) as a fixed point equation

$$u = u + A^{-1}(f - Au), \tag{5.1}$$

whose weak formulation reads

$$(u, v) = (u, v) + \left(A^{-1}\left(f - Au\right), v\right) = (u, v) + {}_{Y'}\langle f - Au, A^{-*}v\rangle_Y \quad \text{for all } v \in X.$$

However, now the ideal test functions $A^{-*}v$, $v \in X = L_2(\Omega)$ appear whose computation we wish to avoid. Therefore, write A^{-1} as a product of two factors one of which fits the above projection

approach, as we shall see soon below. In fact, writing $A^{-*} = (AA^*)^{-1}A$ and using the symmetry of $(AA^*)^{-1}$, the last identity becomes

$$(u,v) = (u,v) + {}_Y\langle (AA^*)^{-1}(f - Au), Av\rangle_{Y'} \quad \text{for all } v \in X, \tag{5.2}$$

since $(AA^*)^{-1}$ maps Y' onto Y.

Remark 5.1. The primary gain of the formulation (5.2) is that we have traded the inversion of A^* for *every* test function against a *single* application of $(AA^*)^{-1}$ to the residual data. By (3.13), $(AA^*)^{-1}$ is the inverse of the Riesz map $R_Y : Y \to Y'$. Since for any $g \in Y'$ the solution $r \in Y$ to $(AA^*)r = g$ is given by

$${}_{Y'}\langle g, z\rangle_Y = \langle (AA^*)r, z\rangle_Y = (A^*r, A^*z) \quad \forall\ z \in Y,$$

and noting that, by (3.13) and (3.7), ${}_{Y'}\langle g, z\rangle_Y = \langle R_Y^{-1}g, z\rangle_Y$, we see that approximate inversion of (AA^*) is realized by the projector P_h, defined in (4.13), see also Remark 4.4.

To actually solve (5.2), we consider the corresponding fixed point iteration in weak form

$$(u^{k+1}, v) = (u^k, v) + {}_Y\langle (AA^*)^{-1}(f - Au^k), Av\rangle_{Y'} \quad \text{for all } v \in X. \tag{5.3}$$

Introducing an auxiliary variable $\hat r^{k+1} := (AA^*)^{-1}(f - Au^k) \in Y$, (5.3) can be written as

$$\begin{aligned} (A^*\hat r^k, A^*z) &= {}_{Y'}\langle f - Au^k, z\rangle_Y && \text{for all } z \in Y, \\ (u^{k+1}, v) &= (u^k, v) + (A^*\hat r^{k+1}, v) && \text{for all } v \in X. \end{aligned} \tag{5.4}$$

This being still formulated in the infinite dimensional function spaces X, Y, we shall next discretize this iteration. From what we have learnt in previous sections, as soon as the primal trial space X_h is chosen, one should discretize $\hat r^k \in Y$ in some associated space Z_h that gives rise to δ-proximal test spaces, i.e. we assume that

$$\inf_{z_h \in Z_h} \|w_h - A^*z_h\|_X \le \delta \|w_h\|_X \qquad \text{for all } w_h \in X_h, \tag{5.5}$$

for a constant $\delta < 1$. Then the discretized iteration scheme is: given $u_h^k \in X_h$, find $\hat r_h^k \in Z_h$ and $u_h^{k+1} \in X_h$ such that

$$\begin{aligned} (A^*\hat r_h^k, A^*z_h) &= {}_{Y'}\langle f - Au_h^k, z_h\rangle_Y && \text{for all } z_h \in Z_h, \\ (u_h^{k+1}, v_h) &= (u_h^k, v_h) + (A^*\hat r_h^k, v_h) && \text{for all } v_h \in X_h. \end{aligned} \tag{5.6}$$

When X_h is a finite element space this iteration can be carried out by conventional finite element tools.

It will be helpful to record several interpretations of the first relation in (5.6), based on the frequently used Riesz relation

$$\|r\|_{Y'} = \|(AA^*)^{-1}r\|_Y, \quad r \in Y', \tag{5.7}$$

that will be applied especially to residuals $r = f - Av \in Y'$, see (3.7), (3.13).

Remark 5.2. Given $v_h \in X_h$, the solution $\hat r_h = \hat r_h(v_h)$ of $(A^*\hat r_h, A^*z_h) = {}_{Y'}\langle f - Av_h, z_h\rangle_Y$, $z_h \in Z_h$, solves

$$\hat r_h(v_h) = \operatorname{argmin}_{\phi \in Z_h} \|f - Av_h - (AA^*)\phi\|_{Y'} = \operatorname{argmin}_{\phi \in Z_h} \|(AA^*)^{-1}(f - Av_h) - \phi\|_Y. \tag{5.8}$$

In fact, the normal equation of the first minimization problem is easily seen to be $\langle AA^*\phi, AA^*z_h\rangle_{Y'} = \langle f - Av_h, AA^*z_h\rangle_{Y'}$, $z_h \in Z_h$, which by (4.12) and (3.14) is equivalent to $(A^*\phi, A^*z_h) = \langle f - Av_h, AA^*z_h\rangle_{Y'}$, $z_h \in Z_h$. Taking (3.8) and (3.13) into account, one obtains $\langle f - Av_h, AA^*z_h\rangle_{Y'} = {}_{Y'}\langle f - Av_h, z_h\rangle_Y$, which confirms (5.8).

In particular, it follows that $\hat r_h^k = P_h\hat r^k$ is the Y-orthogonal projection of $(AA^*)^{-1}(f - Au_h^k)$ on Z_h so that

$$\|\hat r_h^k\|_Y \le \|\hat r^k\|_Y, \tag{5.9}$$

which will be useful below.

The remainder of this Section is devoted to the following issues:

(i) Study the convergence of (5.6) for a fixed given X_h and an associated δ-proximal test space Y_h^δ.
(ii) Identify general conditions under which the iteration (5.6) can be intertwined with an adaptive refinement strategy.

5.2. **Convergence of the Iteration (5.6).** To analyze the convergence of the above iteration (5.6), let $u_{h,\delta}$ denote again the solution of the Petrov-Galerkin scheme

$$(u_{h,\delta}, A^*\tilde{y}_h) = {}_{Y'}\langle f, \tilde{y}_h\rangle_Y \quad \forall\ \tilde{y}_h \in \tilde{Y}_h := P_h(A^{-*}X_h), \tag{5.10}$$

where P_h is again the Y-orthogonal projection onto Z_h. Clearly, $u_{h,\delta}$ can be considered as a perturbation of the solution to the "ideal Petrov-Galerkin discretization

$$(u_h, A^*y_h) = {}_{Y'}\langle f, y_h\rangle_Y \quad \forall\ y_h \in Y_h := A^{-*}X_h, \tag{5.11}$$

see Remark 3.7, which incidentally is the *best L_2-approximation* u_h to u in X_h, i.e.

$$(u_h, v_h) = (u, v_h) \qquad \text{for all } v_h \in X_h.$$

More precisely, we already know that $u_{h,\delta}$ is a near-best approximation to u from X_h. In fact, (4.10) yields

$$\|u - u_{h,\delta}\|_X \le \frac{2}{1-\delta}\|u - u_h\|_X\ . \tag{5.12}$$

Morever, the deviation of $u_{h,\delta}$ from u_h can be bounded as follows.

Lemma 5.3. *Under the above assumptions one has*

$$\|u_h - u_{h,\delta}\|_X \le \delta\|f - Au_{h,\delta}\|_{Y'}. \tag{5.13}$$

This, in turn, implies

$$\|u_h - u_{h,\delta}\|_X \le \frac{\delta}{1-\delta}\|u - u_h\|_X. \tag{5.14}$$

Proof: We infer from (3.7), (3.13), (3.15) and the fact that u_h is the X-orthogonal projection of u onto X_h that, for every $v_h \in X_h$,

$$\begin{aligned}
(u_h - u_{h,\delta}, v_h) &= (u - u_{h,\delta}, v_h) = {}_{Y'}\langle A(u - u_{h,\delta}), Av_h\rangle_{Y'} \\
&= {}_{Y'}\langle f - Au_{h,\delta}, (AA^*)^{-1}Av_h\rangle_Y = {}_{Y'}\langle f - Au_{h,\delta}, A^{-*}v_h\rangle_{Y'} \\
&= {}_{Y'}\langle f - Au_{h,\delta}, (I - P_h)A^{-*}v_h\rangle_Y + {}_{Y'}\langle f - Au_{h,\delta}, P_hA^{-*}v_h\rangle_Y \\
&= {}_{Y'}\langle f - Au_{h,\delta}, (I - P_h)A^{-*}v_h\rangle_Y,
\end{aligned}$$

where we have used Petrov-Galerkin orthogonality of $u_{h,\delta}$ in the last step. Hence, we conclude upon using (5.5),

$$(u_h - u_{h,\delta}, v_h) \le \|f - Au_{h,\delta}\|_{Y'}\|(I - P_h)A^{-*}v_h\|_Y \le \delta\|f - Au_{h,\delta}\|_{Y'}\|v_h\|_X. \tag{5.15}$$

This estimate, combined with a duality argument in X_h, yields now

$$\|u_h - u_{h,\delta}\|_X = \sup_{v_h\in X_h}\frac{(u_h - u_{h,\delta}, v_h)}{\|v_h\|_X} \le \delta\|f - Au_{h,\delta}\|_{Y'}, \tag{5.16}$$

which is (5.13). As for the remaining part of the assertion, note that

$$\begin{aligned}
\|f - Au_{h,\delta}\|_{Y'} &\le \|f - Au_h\|_{Y'} + \|A(u_h - u_{h,\delta})\|_{Y'} \\
&= \|u - u_h\|_X + \|u_h - u_{h,\delta}\|_X.
\end{aligned} \tag{5.17}$$

Inserting this into (5.16) finishes the proof. □

Theorem 5.4. *Assume that* (5.5) *is satisfied and let* u_h, $u_{h,\delta}$ *be defined as above. Then the iterates generated by the scheme* (5.6) *converge to* $u_{h,\delta}$ *and*

$$\|u_{h,\delta} - u_h^{k+1}\|_X \leq \delta \|u_{h,\delta} - u_h^k\|_X, \quad k = 0, 1, 2, \ldots. \tag{5.18}$$

Moreover, the deviation of the iterates u_h^k *from the exact* L_2*-projection (viz. ideal Petrov-Galerkin projection)* u_h *of* u *can be estimated as follows.*

$$\|u_h - u_h^{k+1}\|_X \leq \|u - u_h\|_X + \delta \|u_h - u_h^k\|_X. \tag{5.19}$$

Remark 5.5. Note that as long as

$$\|u - u_h\|_X \leq \eta \|u_h - u_h^k\|_X \tag{5.20}$$

holds for some constant $\eta > 0$, one has

$$\|u_h - u_h^{k+1}\|_X \leq (\eta + \delta) \|u_h - u_h^k\|_X. \tag{5.21}$$

Thus, if $\eta + \delta < 1$, the iterates u_h^k tend to u_h with a fixed error reduction per step. However, in view of (5.18) and (5.14), condition (5.20) will be violated after a few iterations even for $\eta = 1$, say. In this case one obtains $\|u - u_h^k\|_X \leq \|u - u_h\|_X + \|u_h - u_h^k\|_X \leq 2\|u - u_h\|_X$, i.e. already after a few iterations the u_h^k are uniform near best approximations.

Proof of Theorem 5.4: Repeated use of the relations (3.7), (3.13), (3.15) as in the proof of Lemma 5.3 yields

$$\begin{aligned}
(u_{h,\delta} - u_h^{k+1}, v_h) &= (u_{h,\delta} - u_h^k, v_h) - {}_Y\langle \hat{r}_h^k, Av_h\rangle_{Y'} \\
&= {}_{Y'}\langle A(u_{h,\delta} - u_h^k), Av_h\rangle_{Y'} - {}_Y\langle \hat{r}_h^k, Av_h\rangle_{Y'} \\
&= {}_{Y'}\langle A(u_{h,\delta} - u_h^k), A^{-*}v_h\rangle_Y - \langle \hat{r}_h^k, A^{-*}v_h\rangle_Y.
\end{aligned}$$

Since

$$\langle \hat{r}_h^k, A^{-*}v_h\rangle_Y = \langle P_h\hat{r}_h^k, A^{-*}v_h\rangle_Y = \langle \hat{r}_h^k, P_hA^{-*}v_h\rangle_Y,$$

we obtain, upon also using the definition of $u_{h,\delta}$,

$$\begin{aligned}
(u_{h,\delta} - u_h^{k+1}, v_h) &= {}_{Y'}\langle A(u_{h,\delta} - u_h^k), A^{-*}v_h\rangle_Y - \langle \hat{r}_h^k, P_hA^{-*}v_h\rangle_Y \\
&= {}_{Y'}\langle A(u_{h,\delta} - u_h^k), P_hA^{-*}v_h\rangle_Y \\
&\quad + {}_{Y'}\langle A(u_{h,\delta} - u_h^k), (I - P_h)A^{-*}v_h\rangle_Y - \langle \hat{r}_h^k, P_hA^{-*}v_h\rangle_Y \\
&= {}_{Y'}\langle f - Au_h^k, P_hA^{-*}v_h\rangle_Y \\
&\quad + {}_{Y'}\langle A(u_{h,\delta} - u_h^k), (I - P_h)A^{-*}v_h\rangle_Y - \langle \hat{r}_h^k, P_hA^{-*}v_h\rangle_Y \\
&= \langle (AA^*)^{-1}(f - Au_h^k), P_hA^{-*}v_h\rangle_Y \\
&\quad + {}_{Y'}\langle A(u_{h,\delta} - u_h^k), (I - P_h)A^{-*}v_h\rangle_Y - \langle \hat{r}_h^k, P_hA^{-*}v_h\rangle_Y \\
&= \langle (AA^*)^{-1}(f - Au_h^k) - \hat{r}_h^k, P_hA^{-*}v_h\rangle_Y \\
&\quad + {}_{Y'}\langle A(u_{h,\delta} - u_h^k), (I - P_h)A^{-*}v_h\rangle_Y \\
&= {}_{Y'}\langle A(u_{h,\delta} - u_h^k), (I - P_h)A^{-*}v_h\rangle_Y \\
&\leq \|A(u_{h,\delta} - u_h^k)\|_{Y'} \|(I - P_h)A^{-*}v_h\|_Y \\
&\leq \|u_{h,\delta} - u_h^k\|_X \delta \|A^{-*}v_h\|_Y = \delta \|u_{h,\delta} - u_h^k\|_X \|v_h\|_X.
\end{aligned}$$

The assertion (5.18) follows now again by the duality argument (5.16).

As for (5.19), we argue similarly, using X-orthogonality of u_h,

$$\begin{aligned}
(u_h - u_h^{k+1}, v_h) &= (u, v_h) - (u_h^k, v_h) - {}_Y\langle (AA^*)^{-1}(f - Au_h^k), Av_h\rangle_{Y'} \\
&\quad + {}_Y\big\langle (AA^*)^{-1}(f - Au_h^k) - \hat{r}_h^k, Av_h\big\rangle_{Y'} \\
&= {}_Y\big\langle (AA^*)^{-1}(f - Au_h^k) - \hat{r}_h^k, Av_h\big\rangle_{Y'} \\
&\leq \big\|(AA^*)^{-1}(f - Au_h^k) - \hat{r}_h^k\big\|_Y \|Av_h\|_{Y'} \\
&\leq \|(AA^*)^{-1}(f - Au_h^k) - \hat{r}_h^k\|_Y \|v_h\|_X,
\end{aligned}$$

where we have used (3.15) in the last step. The first factor in the last bound can be estimated with the aid of the stability property (5.5) by

$$\begin{aligned}\inf_{\phi\in Z_h}\|(AA^*)^{-1}(f-Au_h^k)-\phi\|_Y &= \inf_{\phi\in Z_h}\|A^*(AA^*)^{-1}(f-Au_h^k)-A^*\phi\|_X\\ &= \inf_{\phi\in Z_h}\|u-u_h^k-A^*\phi\|_X\\ &\le \|u-u_h\|_X+\inf_{\phi\in Z_h}\|u_h-u_h^k-A^*\phi\|_X\\ &\le \|u-u_h\|_X+\delta\|u_h-u_h^k\|_X.\end{aligned}\tag{5.22}$$

Invoking again (5.16), finishes the proof. □

Remark 5.6. The main conclusion of Theorem 5.4 is that the iterates from (5.6) rapidly converge to the Petrov-Galerkin solution with respect to a δ-proximal test space for the given trial space, *without ever* computing a basis for the test space. Each iteration step requires the inversion of a symmetric positive definite system which is of the size of $\dim Z_h$ and in our numerical experiments below we shall have $\dim Z_h \sim \dim X_h$. We shall not discuss here concrete ways of performing these inversions as efficiently as possible but only remark that this cost can be expected to be far lower than the computation of individual test basis functions an corresponding load vector and matrix assemblations, see also the comments at the end of Section 5.3.

5.3. **An Adaptive Strategy.** Instead of driving the iterates u_h^k in (5.6) to the limit $u_{h,\delta}$ for a fixed X_h, we wish to explore next how to intertwine the iteration with an adaptive refinement strategy. The rationale is that after a few iterations for a fixed X_h one gets close to $u_{h,\delta}$ and hence uniformly close to the best L_2-approximation u_h in the current space. One would then want to infer how to expand X_h to some larger space $X_{h'}$ in such a way that the new error $\|u-u_{h'}\|_X$ reduces the previous one by a fixed factor. To accomplish that one has to estimate the (full infinite dimensional) residual $\|Au_h^k-f\|_{Y'}$. Here, as usual, the problem is the evaluation or estimation of the dual norm. This is where (5.7) will come into play which, in particular, means that

$$\|Au_h^k-f\|_{Y'} = \|(AA^*)^{-1}(Au_h^k-f)\|_Y.$$

Now recall that in the iteration scheme (5.6) the quantity $\hat r_h^k$, defined in (5.4), just approximates the term $(AA^*)^{-1}(Au_h^k-f)$ in the space Z_h. As we have argued before Z_h must have a finer resolution than X_h in order to guarantee the δ-proximality condition (5.5). In particular, the dimension of Z_h is larger than the dimension of X_h and of Y_h^δ (whose explicit computation is avoided above). Therefore, one could expect that $\hat r_h^k$ captures a significant portion of $(AA^*)^{-1}(Au_h^k-f)$, i.e.

$$\|A^*\hat r_h^k\|_X = \|\hat r_h^k\|_Y \approx \|Au_h^k-f\|_{Y'},$$

is a useful approximation of the size of the residual. On the other hand, since $\hat r_h^k$, involves a projection onto the finite dimensional space Z_h we have lost some information about f so that surely these two terms cannot be equivalent in general. They can only be comparable under some *assumptions on the data*. This problem arises in all a-posteriori error estimation based on duality and causes what is usually called "data oscillation terms" entering the a-posteriori bounds.

Therefore we shall make such assumptions on the data.

Assumption f: *We require a sightly stronger variant of* (5.5)*:*

$$\inf_{z_h\in Z_h}\|w_h-A^*z_h\|_X \le \delta\|w_h\|_X \qquad \text{for all } w_h\in X_h+A^{-1}F_h,$$

where $F_h\subset Y'$ is some subspace that is computationally accessible. This is equivalent to

$$\inf_{z_h\in Z_h}\|g_h-AA^*z_h\|_{Y'} \le \delta\|g_h\|_{Y'} \qquad \text{for all } g_h\in AX_h+F_h.\tag{5.23}$$

Moreover, we assume that for any given X_h and the orthogonal projection u_h of u onto X_h, there exists a refinement $X_{h'}$ and an associated $F_{h'}$ such that there exists an $\tilde f_h\in AX_{h'}+F_{h'}$ with

$$\|f-\tilde f_h\|_{Y'} \le \alpha\|u-u_h\|_X\tag{5.24}$$

for any fixed $\alpha \in (0,1)$. □

The validity of such a condition has to be verified for any concrete X_h and suitable choices of the auxiliary spaces Z_h. Again, here we are mainly interested in identifying the essential requirements in a general framework and will address this issue in more detail in [12]. In our first experiments below in Section 6 X_h is a space of locally refined piecewise polynomials over a mesh $\mathcal{T}_h$, Z_h is higher order conforming finite element space on a fixed local refinement of $\mathcal{T}_h$, and F_h is just taken as $\mathbb{P}_{0,\mathcal{T}_h}$.

Given the somewhat stronger proximality condition (5.23), one can indeed establish the desired equivalence.

Lemma 5.7. *Assume that* $v_h \in X_h$, $g_h \in F_h$ *and that condition* (5.23) *holds. Furthermore, let* $\hat{r}_h = \hat{r}_h(v_h, g_h)$ *solve the equation (see (5.6))*

$$(A^*\hat{r}_h, A^*z_h) = {}_{Y'}\langle g_h - Av_h, z_h\rangle_Y \qquad \text{for all } z_h \in Z_h.$$

Then we have

$$(1-\delta)\|Av_h - g_h\|_{Y'} \le \|\hat{r}_h(v_h,g_h)\|_Y \le \|Av_h - g_h\|_{Y'}. \tag{5.25}$$

Proof: The upper inequality follows already from (5.9). As for the lower inequality, invoking (5.8) in Remark 5.2, we have

$$\begin{aligned}\|Av_h - g_h\|_{Y'} &= \|(AA^*)^{-1}(Av_h - g_h)\|_Y \le \|\hat{r}_h\|_Y + \|\hat{r}_h - (AA^*)^{-1}(Av_h - g_h)\|_Y\\ &\le \|\hat{r}_h\|_Y + \inf_{\phi\in Z_h}\|\phi - (AA^*)^{-1}(Av_h - g_h)\|_Y\\ &\le \|\hat{r}_h\|_Y + \delta\|(AA^*)^{-1}(Av_h - g_h)\|_Y \le \|\hat{r}_h\|_Y + \delta\|Av_h - g_h\|_{Y'},\end{aligned}$$

providing $(1-\delta)\|Av_h - g_h\|_{Y'} \le \|\hat{r}_h\|_Y$, and proves the assertion. □

Note also that in the iteration scheme (5.6) the quantities $\hat{r}_h^k$ have to be computed anyway so that their use as an error estimator does not cause any additional cost. Now, one important property of the error indicator $\|\hat{r}_h\|_Y$ is that the norm can be localized to any given partition of Ω, so that one could extract local error indicators from $\|A^*\hat{r}_h\|_X = \|\hat{r}_h\|_{Y'}$. Instead of looking for large portions of $\hat{r}_h^k$ and then refine those portions of the mesh in some way, we shall use the information provided by $\hat{r}_h^k$ in the slightly different following way. Suppose for the moment that we have a method that computes an approximation r_H on an enlarged subspace X_H of X, i.e. $X_h \subset X_H \subset X$ such that

$$\|A^*\hat{r}_h - r_H\|_X \le \eta\|A^*\hat{r}_h\|_X, \tag{5.26}$$

where $\eta > 0$ is sufficiently small independent of h. At this point we shall formulate this as a key condition to be verified for a concrete variational formulation and an underlying family of trial spaces. Assuming now the validity of (5.26), we infer from (5.25), using again (5.8),

$$\begin{aligned}\|g_h - Av_h - Ar_H\|_{Y'} &\le \|g_h - Av_h - AA^*\hat{r}_h\|_{Y'} + \|AA^*\hat{r}_h - Ar_H\|_{Y'}\\ &= \inf_{\phi\in Z_h}\|g_h - Av_h - AA^*\phi\|_{Y'} + \|AA^*\hat{r}_h - Ar_H\|_{Y'}\\ &\le \delta\|g_h - Av_h\|_{Y'} + \eta\|\hat{r}_h\|_Y\\ &\le (\delta+\eta)\|g_h - Av_h\|_{Y'}.\end{aligned} \tag{5.27}$$

Thus for δ and η sufficiently small we achieve indeed a reduction of the residual and hence of the error. The following scheme specifies the relevant parameters that guarantee error reduction under assumptions (5.23) and (5.26).

Algorithm 1 Iteration scheme

1: (initialization) Fix the final error tolerance ϵ, a parameter $\rho \in (0,1)$; choose Initial spaces X_h, Z_h and $\bar{u} = 0$; set $e := \|f\|_{Y'} = \|u_h^0 - u\|_X$; choose $\alpha_1, \alpha_2 \in (0,1)$ such that for δ, η from (5.23) and (5.26), respectively

$$\Big(\alpha_2 + \Big(\frac{(\delta+\eta)(1+\alpha_2+(1-\delta)\alpha_1)}{1-\delta}\Big)\Big) \leq \rho;$$

let $K := \operatorname{argmin}\big\{k \in \mathbb{N} : \delta^k \left(\frac{(3-\delta)(\alpha_2+1)}{1-\delta}\right) \leq \alpha_1\big\}$.

2: **while** $e > \epsilon$ **do**
3: given $\bar{u} \in X_h$, $\tilde{f}_h \in AX_h + F_h$ with $\|u - \bar{u}\|_X \leq e$ and $\|f - \tilde{f}_h\|_{Y'} \leq \alpha_2 e$, set $u_h^0 := \bar{u}$
4: **for** $k = 0$ to K **do**
5: Solve
6: $(A^*\hat{r}_h^k, A^* z_h) = (\tilde{f}_h - Au_h^k, z_h)$
7: $(u_h^{k+1}, v_h) = (u_h^k, v_h) + (A^*\hat{r}_h^k, v_h)$
8: for all $z_h \in Z_h$ and $v_h \in X_h$
9: **end for**
10: Compute X_H and r_H s.t. $\|A^*\hat{r}_h^K - r_H\|_X \leq \eta \|A^*\hat{r}_h^K\|_X$
11: Compute $X_{h'}, F_{h'}$, $\tilde{f}_h \in AX_{h'} + F_{h'}$ such that $\|f - \tilde{f}_h\|_{Y'} \leq \alpha_2 \rho e$
12: Set $X_h \to X_h + X_{h'} + X_H$, $\rho e \to e$.
13: Choose Z_h according to the updated X_h.
14: Set $\bar{u} = u_h^K + r_H$ and go to 2:.
15: **end while**

Proposition 5.8. *Assume the validity of* **Assumption f**, *(5.23) and (5.26). Then* **Algorithm 1***outputs a trial space X_h and an approximate solution $\bar{u} \in X_h$ such that $\|u - \bar{u}\|_X \leq \epsilon$.*

Proof: We have to show only that for $\bar{u}$ in step 3: the element $u_h^K + r_H$ in step 14 satisfies $\|u - (u_h^K + r_H)\|_X \leq \rho e$. To that end, let u_h again denote the X-orthogonal projection of u to X_h. Let $\tilde{u}$ be the solution to $A\tilde{u} = \tilde{f}_h$ and let $\tilde{u}_{h,\delta}$ denote the Petrov-Galerkin solution of (5.10) with $\tilde{f}_h$ in place of f. We know from Theorem 5.4, (5.18), that

$$\begin{aligned}
\|\tilde{u}_{h,\delta} - u_h^K\|_X &\leq \delta^K \|\tilde{u}_{h,\delta} - \bar{u}\|_X \leq \delta^K \big(\|\tilde{u}_{h,\delta} - \tilde{u}\|_X + \|\tilde{u} - u\|_X + \|u - \bar{u}\|_X\big) \\
&\leq \delta^K \Big(\frac{2}{1-\delta}\|\tilde{u} - \tilde{u}_h\|_X + (\alpha_2+1)e\Big) \leq \delta^K \Big(\frac{2}{1-\delta}\|\tilde{u} - u_h\|_X + (\alpha_2+1)e\Big) \\
&\leq \delta^K \Big(\frac{2}{1-\delta}(\|\tilde{u} - u\|_X + \|u - u_h\|_X) + (\alpha_2+1)e\Big) \\
&\leq \delta^K \Big(\frac{3-\delta}{1-\delta}\Big)(\alpha_2+1)e \leq \alpha_1 e, \qquad (5.28)
\end{aligned}$$

where we have used (5.14) and the definition of K in step 1:. Then, by (5.24), (5.26) and (5.27), we obtain

$$\begin{aligned}
\|u - (u_h^K + r_H)\|_X &= \|f - A(u_h^K + r_H)\|_{Y'} \leq \alpha_2 e + \|\tilde{f}_h - A(u_h^K + r_H)\|_{Y'} \\
&\leq \alpha_2 e + (\delta+\eta)\|\tilde{f}_h - Au_h^K\|_{Y'} = \alpha_2 e + (\delta+\eta)\|\tilde{u} - u_h^K\|_X \qquad (5.29)
\end{aligned}$$

Now, by (5.28) and (5.14), we obtain

$$\begin{aligned}
\|\tilde{u} - u_h^K\|_X &\leq \|\tilde{u} - u_h\|_X + \|\tilde{u}_h - \tilde{u}_{h,\delta}\|_X + \|\tilde{u}_{h,\delta} - u_h^K\|_X \\
&\leq \Big(1 + \frac{\delta}{1-\delta}\Big)\|\tilde{u} - \tilde{u}_h\|_X + \alpha_1 e \leq \Big(1 + \frac{\delta}{1-\delta}\Big)\|\tilde{u} - u_h\|_X + \alpha_1 e \\
&\leq \Big(1 + \frac{\delta}{1-\delta}\Big)(\|\tilde{u} - u\|_X + \|u - u_h\|_X) + \alpha_1 e \\
&\leq \Big(1 + \frac{\delta}{1-\delta}\Big)(\alpha_2 e + e) + \alpha_1 e = \Big(\frac{1+\alpha_2+(1-\delta)\alpha_1}{1-\delta}\Big)e.
\end{aligned}$$

Inserting this in (5.29), yields

$$\|u-(u_h^K+r_H)\|_X \le \Big(\alpha_2+\Big(\frac{(\delta+\eta)(1+\alpha_2+(1-\delta)\alpha_1)}{1-\delta}\Big)\Big)e \le \rho e, \tag{5.30}$$

where we have used step 1:. □

Remark 5.9. Introducing the usual approximation spaces $\mathcal{A}^s$ consisting of all functions in X whose k-term approximation decays at least like k^{-s}, adjusting the tolerance $\rho<1$ appropriately and introducing a coarsening step after reducing the error to ρe, we can prove also asymptotically optimal complexity, provided that the number of degrees of freedom in X_h, X_H and in the corresponding Z_h stays uniformly proportional to $\dim X_h$.

It is perhaps worth pointing out that in step 5 of **Algorithm 1** one may not have to carry out K iterations but terminate the loop based on an aposteriori test. In fact, according to the above proof, we need to find the smallest k such that $\|\tilde{u}_{h,\delta}-u_h^K\|_X \le \alpha_1 e$. Here is a sketch of how to obtain such a termination criterion. The first step is to see what are the actual computational steps in 5 - 8 of **Algorithm 1**. To that end, suppose that

$$Z_h = \operatorname{span}\{\xi_k : k\in\hat{\mathcal{I}}_h\}$$

and consider the symmetric positive matrix $\mathbf{B}_h := \big((A^*\xi_k, A^*\xi_i)\big)_{i,k\in\hat{\mathcal{I}}_h}$. The first line in (5.6) is a residual computation. Given any $v_h=\sum_{j\in\mathcal{I}_h} v_j\phi_j$, with coefficient vector $\mathbf{v}_h$, it requires finding $\hat{r}_h=\sum_{k\in\hat{\mathcal{I}}_h}\hat{r}_k\xi_k$ with coefficient vector $\hat{\mathbf{r}}_h(v_h)=\hat{\mathbf{r}}_h$ satisfying

$$\mathbf{B}_h\hat{\mathbf{r}}_h = \hat{\mathbf{f}}_h-\mathbf{p}_h(v_h), \quad \text{where } \begin{cases} \hat{\mathbf{f}}_h & := \big({}_{Y'}\langle \tilde{f}_h,\xi_k\rangle_Y\big)_{k\in\hat{\mathcal{I}}_h},\\ \mathbf{p}_h(v_h) & := \big((v_h,A^*\xi_k)\big)_{k\in\hat{\mathcal{I}}_h}.\end{cases} \tag{5.31}$$

Hence the first line of (5.6) requires solving once a linear system of size comparable to $\dim(X_h)$. If Φ_h is orthonormal, the second line in (5.6) is a simple update of the coefficient vector

$$\mathbf{u}_h^{k+1} = \mathbf{u}^k + \mathbf{P}_h\hat{\mathbf{r}}_h(u_h^k), \tag{5.32}$$

where $\mathbf{P}_h$ is the $\#(\mathcal{I}_h)\times\#(\hat{\mathcal{I}}_h)$-matrix with rows $\mathbf{p}_h(\phi_i)^\top$, see (5.31), that, in particular, $\mathbf{p}_h(u_h^k)=\mathbf{P}^\top+\mathbf{u}_h^k$. (When Φ_h is not orthonormal one has to invert in addition a mass matrix.) Of course, combining (5.31) and (5.32) we can express (5.6) by a single relation

$$\mathbf{u}_h^{k+1} = \mathbf{u}_h^k + \mathbf{P}_h\mathbf{B}_h^{-1}(\hat{\mathbf{f}}_h-\mathbf{P}_h^\top\mathbf{u}_h^k)). \tag{5.33}$$

Remark 5.10. Defining, according to (4.13), the δ-proximal test functions $\tilde{\phi}_i := P_h\phi_i$, i.e.

$$(A^*\tilde{\phi}_i, A^*\xi_k) = (\phi_i, A^*\xi_k), \quad k\in\hat{\mathcal{I}}_h,$$

as well as the corresponding Galerkin matrix and load vectors,

$$\mathbf{A}_{h,\delta} := \big((\phi_j, A^*\tilde{\phi}_i)\big)_{i,j\in\mathcal{I}_h}, \quad \mathbf{f}_{h,\delta} := \big({}_{Y'}\langle f,\tilde{\phi}_i\rangle_Y\big)_{i\in\mathcal{I}_h}, \tag{5.34}$$

we have

$$\mathbf{P}_h\mathbf{B}_h^{-1}(\hat{\mathbf{f}}_h-\mathbf{P}_h^\top\mathbf{u}_h^k) = \mathbf{f}_{h,\delta}-\mathbf{A}_{h,\delta}\mathbf{u}_h^k. \tag{5.35}$$

Hence, (5.33) becomes

$$\mathbf{u}_h^{k+1} = \mathbf{u}_h^k + (\mathbf{f}_{h,\delta}-\mathbf{A}_{h,\delta}\mathbf{u}_h^k), \quad k\in\mathbb{N}_0. \tag{5.36}$$

Proof: Writing $\tilde{\phi}_i=\sum_{k\in\hat{\mathcal{I}}_h} q_k^i\xi_k$ with coefficient vector $\mathbf{q}_h^i$, we have $\mathbf{B}_h\mathbf{q}_h^i=\mathbf{p}_h(\phi_i)$. Thus, defining $\mathbf{Q}_h$ as the $\#(\mathcal{I}_h)\times\#(\hat{\mathcal{I}}_h)$- matrix whose rows are the $\mathbf{q}_h^i$ we have $\mathbf{B}_h\mathbf{Q}_h^\top=\mathbf{P}_h^\top$. Also, by definition, $\mathbf{f}_{h,\delta}=\mathbf{Q}_h\hat{\mathbf{f}}_h$. Therefore

$$\mathbf{P}_h\mathbf{B}_h^{-1}(\hat{\mathbf{f}}_h-\mathbf{P}_h^\top\mathbf{u}_h^k) = \mathbf{Q}_h\hat{\mathbf{f}}_h-\mathbf{Q}_h\mathbf{P}_h^\top\mathbf{u}_h^k = \mathbf{f}_{h,\delta}-\mathbf{Q}_h\mathbf{P}_h^\top\mathbf{u}_h^k.$$

Since

$$\mathbf{Q}_h\mathbf{P}_h^\top\mathbf{u}_h^k = \Big(\Big(u_h^k, \sum_{k\in\hat{\mathcal{I}}_h} A^*(\mathbf{q}_h^i)_k\xi_k\Big)\Big)_{i\in\mathcal{I}_h} = \big((u_h^k, A^*\tilde{\phi}_i)\big)_{i\in\mathcal{I}_h},$$

the assertion follows. □

Thus, each iteration requires (i) the computation of $\hat{\mathbf{f}}_h$; (ii) the application of $\mathbf{P}_h$ and of $\mathbf{P}_h^T$; (iii) as well as approximate solution of one linear system of size $\hat{N}_h := \dim Z_h$ with coefficient matrix $\mathbf{B}_h$. Whenever the basis functions ξ_k for Z_h have local supports the cost of (i) and (ii) is $O(\dim Z_h)$. An efficient inversion of the symmetric positive definite sparse matrix $\mathbf{B}_h$ is expected to cost $O(\dim Z_h)$ as well, so that the total cost of one step is expected to be $O(\dim Z_h)$ which presumably is $O(\dim X_h)$.

Remark 5.11. It should be emphasized that (5.15) offers a way to actually compute the array

$$({}_{Y'}\langle f - Au_h^k, \tilde{\phi}_i\rangle_Y)_{i\in\mathcal{I}} = \mathbf{f}_{h,\delta} - A_{h,\delta}\mathbf{u}_h^k$$

without computing the individual inner products with the test functions ϕ_i nor the test functions themselves, see Remark 5.6.

Next, in order to exploit the above discrete residuals we invoke some results from [12] which imply that for $\delta \le 1/3$ one has

$$\frac{1}{4}\|v - \tilde{u}_{h,\delta}\|_X^2 \quad \le \quad \|\mathbf{A}_{h,\delta}\mathbf{v} - \tilde{\mathbf{f}}_h\|_{\ell_2}^2 \le \frac{5}{4}\|Av - \tilde{f}_h\|_{Y'}^2.$$

Thus, $\|\mathbf{f}_{h,\delta} - A_{h,\delta}\mathbf{u}_h^k\|_{\ell_2}$ can be used to control the accuracy of $\|\tilde{u}_{h,\delta} - u_h^k\|_X$. In fact, Line 4 in **Algorithm 1** should be replaced by

$$\textbf{while } \|\mathbf{A}_{h,\delta}\mathbf{u}_h^k - \tilde{\mathbf{f}}_h\|_{\ell_2} > \alpha_1\epsilon/2,$$

instead of iterating until $k = K$.

Remark 5.12. In summary it is important to distinguish two issues. Adaptivity, which is an analytic task, takes place in the space X (typically an L_2-space). The approximate calculation of the infinite dimensional residuals $\|f - Au_h^k\|_{Y'}$ requires the inversion of certain symmetric positive definite systems (5.31), which we view as an algebraic task. Its complexity is not addressed here but deferred to [12] and will depend on the specific realizations.

Of course, the above framework leaves considerable room for variations. For instance, the auxiliary space Z_h used to approximate the residual error of the final subiteration u_h^K in step 10 of **Algorithm 1** could be chosen larger than the auxiliary spaces in the preceding iterations, since condition (5.23) could be more stringent than condition (5.5) for δ-proximality.

6. Numerical experiments

We illustrate the above framework with a first, admittedly preliminary numerical illustration for the test problem

$$Au = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \cdot \nabla u + u, \qquad f = \chi_{x_1 \ge x_2} + \frac{1}{2}\chi_{x_1 < x_2}. \tag{6.1}$$

Since many of the ingredients need to be further refined we do not address here any efficiency issues, e.g. concerning the inversion of the positive symmetric systems (5.31). More extensive numerical tests will be given in a forthcoming paper [12].

As a first set of experiments, addressing primarily quantitative aspects of δ-proximality in the context of the iteration scheme (5.6), we use uniform rectangular girds. Note that these grids are neither aligned to the flow direction $\vec{b}$ nor the the discontinuity of the right hand side. For X_h we take as a first example globally continuous piecewise bilinear polynomials, see Figure 1.

For the associated *auxiliary space* Z_h we always test the following simple receipe. It is always comprised of continuous piecewise polynomials of varying polynomial degree on a grid that results from at most one single refinement of the mesh for X_h. Recall, that for a given X_h the iterates converge to the Petrov-Galerkin solution at a speed depending on δ. To achieve discretization error accuracy one would need more iterations for increasing resolution in X_h. Therefore, we employ nested iteration and apply on each level only a fixed number K of iterations taking the approximate

n	$p=1, r=1, K=10$		$p=2, r=0, K=1$	
	$\|u-u_h\|_{L_2}$	estimate for δ	$\|u-u_h\|_{L_2}$	estimate for δ
9	0.0497599	0.623294	0.053896	3.66634e-08
25	0.0367748	0.693101	0.0388272	-
81	0.0279661	0.811951	0.0280033	-
289	0.0219017	0.901607	0.0200508	2.33678e-07
1089	0.017677	0.946074	0.0143055	7.27839e-08
4225	0.0145256	0.968126	0.0101684	1.58492e-06
16641	0.0120625	0.979499	0.00720102	-
66049	0.010075	0.985621	0.00510021	-

solution from the previous refinement level as an initial guess. This can be viewed as simplified variant of the adaptive scheme in **Algorithm** 5.3. Table 6 records exact errors produced by this scheme for different polynomial degrees p and additional dyadic refinement depth r, used for the auxiliary space Z_h, and for a given number K of iterations on each level. To test the δ-proximality achieved by the chosen auxiliary spaces Z_h we compute the value

$$\frac{\inf_{\phi\in Z_h}\|u_h-u_h^K-A^*\phi\|_X}{\|u_h-u_h^K\|_X} \tag{6.2}$$

which is a lower bound of the stability constant δ in (5.5). Note that this is the only way though how (5.22) enters in the proof of convergence (5.19). Here we find that piecewise bilinear spaces Z_h do not suffice to ensure δ-proximality, regardless of the resolution. This is reflected by Figure 1 showing that the convergence rate deviates form the optimal one. If we choose the degree $p=2$ for Z_h the estimate for δ already becomes very small, even when the same grids are used for X_h and Z_h, and only *a single* iteration step on each level is sufficient, due to the strong contraction. In the rows of Table 6 with a dash "-" the (square of the) numerator of (6.2) is negative. Due to its size being smaller than machine accuracy this not unexpected. Also in Figure 1 one sees that now the approximation error has the same rate as the best approximation of u. The same qualitative behavior can be observed when employing as trial spaces discontinuous piecewise bilinear functions $X_h=\mathcal{P}_{1,\mathcal{T}_h}$ and Z_h as above, see Table 6, although now the dimension of Z_h (due to global continuity) is hardly larger than that of X_h when using the same grids for both spaces. Figure 6 indicates that the discontinuous trial functions show less of a Gibbs phenomenon across the jump. The observed somewhat stronger overshoots seem to be unavoidable since the bilinears are too rigid when the jump crosses their support in a diagonal way. Of course, our setting would allow to just use piecewise constants across the discontinuity. One would also expect to benefit substantially from aligned anisotropic discretizations which will be addressed in forthcoming work.

Finally Figure (3) shows some very preliminary results for (a somewhat simplified version of) the adaptive Algorithm 5.3. Here we used discontinuous piecewise bilinear polynomials for X_h and continuous ones of degree $p=3$ for Z_h on a three times refined grid, $r=3$. Due to the locally strong variation of the grid size we cannot expect to use the same grids for both, X_h and Z_h, because the characteristic wake of a basis function in X_h supported on a high level cell would intersect much coarser cells so that the reduced accuracy might diminish the quality of the test space too much. Again the current implementation is very crude and far from optimized. Nevertheless, since the solution is only in the space BV which just fails to be in the Besov space $B_1^1(L_1(D))$, the optimal adaptive rate in an isotropic refinement setting is $N^{-1/2}$ in terms of the number N of degrees of freedom, which is not quite attained yet in this first experiment but already comes close to it even for the crude version of our algorithm. More extensive studies for a refined version of the algorithm and the discussion of more subtle choices for Z_h will be given in [12]. Of course, again anisotropic versions would be preferable giving rise to the optimal rate N^{-1}.

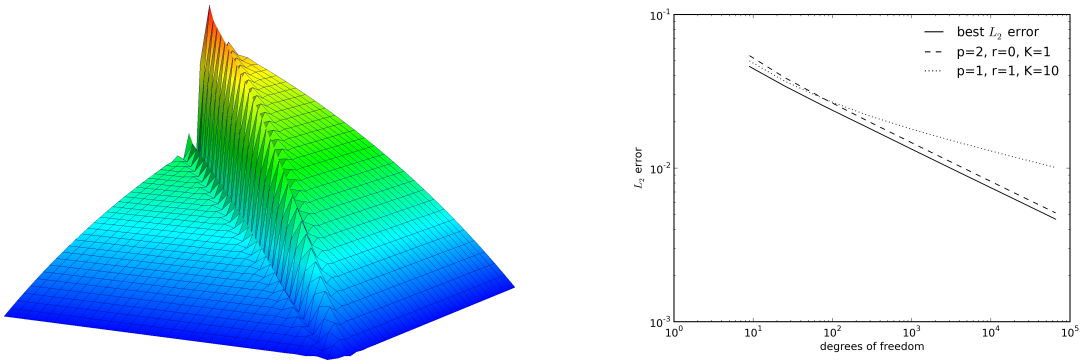

FIGURE 1. Left: solution for the model problem (6.1) on a uniform grid for the spaces $X_h = \mathcal{P}_{1,\mathcal{T}_h} \cap C(D)$ and $Z_h = \mathcal{P}_{2,\mathcal{T}_h} \cap C(D)$. Right: true errors for different test spaces Z_h, in doubly logarithmic scale. Here we vary the polynomial degree p, the additional refinements r and the number of iterations K. For reference also shown is the error of the best L_2 approximation to the solution u from X_h.

n	$p=1, r=1, K=10$		$p=2, r=0, K=1$	
	$\|u-u_h\|_{L_2}$	estimate for δ	$\|u-u_h\|_{L_2}$	estimate for δ
16	0.0631168	0.730086	0.0507563	4.2204e-08
64	0.0425758	0.859166	0.0351883	nan
256	0.0340799	0.911102	0.0248342	1.01583e-07
1024	0.0282526	0.934825	0.0175411	2.22848e-07
4096	0.0236602	0.946033	0.0124549	4.67827e-07
16384	0.0198675	0.951535	0.0088329	1.88817e-06
65536	0.0167049	0.954569	0.0062409	1.10327e-06
262144	0.0140478	0.956369	0.00441879	nan

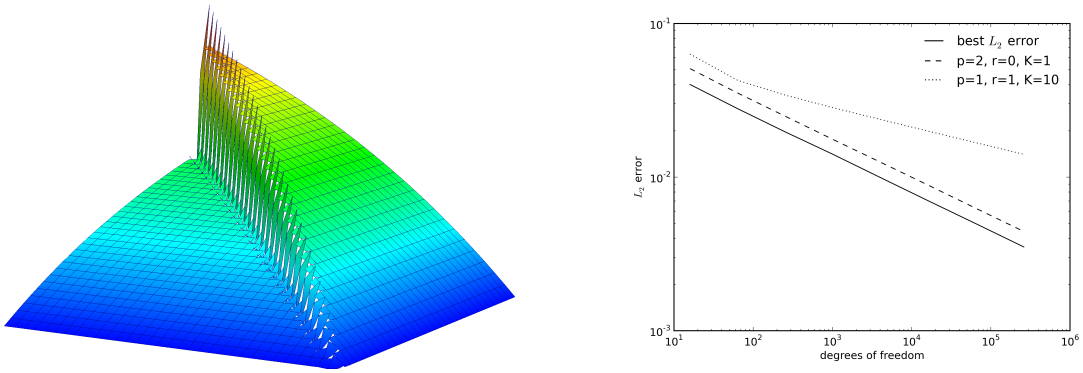

FIGURE 2. Left: solution for the model problem (6.1) on a uniform grid for the spaces $X_h = \mathcal{P}_{1,\mathcal{T}_h}$ and $Z_h = \mathcal{P}_{2,\mathcal{T}_h} \cap C(D)$. Right: doubly logarithmic plot of the true errors for different test spaces Z_h. Here we vary the polynomial degree p, the additional refinements r and the number of iterations K. For reference we included the error of the best L_2 approximation to the solution u from X_h.

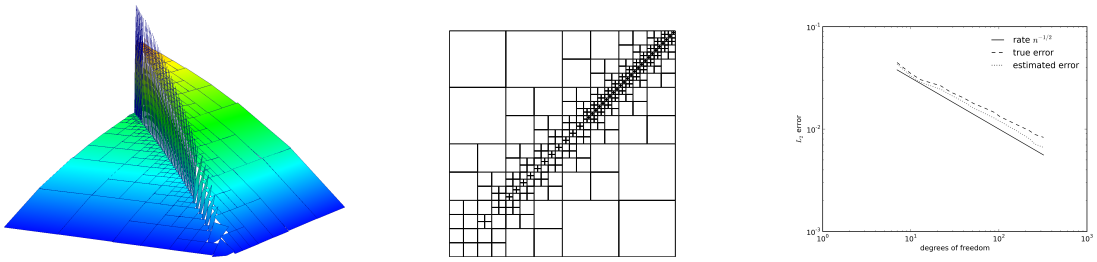

FIGURE 3. The 18th iterate of the adaptive solver and corresponding grid of model problem (6.1) with Algorithm 5.3 with $X_h = \mathcal{P}_{1,T_h}$, $Z_h = \mathcal{P}_{3,T_{h/8}}$ and $K = 9$. The rightmost plot shows a logarithmic plot of the true errors and estimated error. Shown for reference is the optimal rate $n^{-\frac{1}{2}}$ for the presently employed trial spaces.

7. Conclusions and Outlook

Mainly guided by first order linear transport equations we present a general framework for variational formulations with the following key features:

- The norm $\|\cdot\|_X$ - termed "energy norm" - with respect to which accuracy is to be measured, can be chosen within a certain specified regime.
- The resulting variational formulation can be viewed as an ideal (infinite dimensional) Petrov-Galerkin formulation where the test space is determined by the choice of the energy space.
- Although the ideal test metric is not practically feasible we have developed a perturbation approach that leads to practicable realizations and estimations of these residuals. Here we have concentrated first on identifying the essential conditions for the latter step to work, formulated in terms of the notion of δ-proximality leading to uniformly stable Petrov-Galerkin schemes. As a consequence, errors in the energy norm are uniformly comparable to residuals in the dual norm of the test space, see (5.17).
- The framework allows us to formulate concrete adaptive refinement schemes than can rigorously proven to converge, again under certain concrete conditions.

The Petrov-Galerkin formulations developed here cover in a natural way also more general problems like parametric transport problems or kinetic formulations. Resulting high-dimensional problems pose particular computational and analytical challenges. In particular, the present approach opens interesting perspectives for model reduction techniques such as greedy methods for constructing reduced bases since their numerical feasibility relies crucially on the estimation of errors by corresponding residuals. Moreover, the present L_2-formulation seems to offer particularly favorable features for sparse tensor discretizations for parametric transport problems as treated in [23, 24] in connection with least squares formulations. These aspects will be pursued in a forthcoming paper.

References

[1] C. Bardos, Problèmes aux limites pour les equations aux derivees partielles du premier ordre a coefficients reels; Theorèmes d'approximation; application a lequation de transport, Ann. Scient. Ec. Norm. Sup., 4° serie, 3(1970), 185–233.
[2] F. Brezzi, T.J.R. Hughes, L.D. Marini, A. Russo, E. Süli, A priori analysis of residual-free bubbles for advection-diffusion problems, SIAM J. Numer. Anal., 36(1999), 1933–1948.
[3] E.J. Cands, D. L. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise-C^2 singularities. Comm. Pure Appl. Math., 57(2002), 219–266.
[4] M. Cessenat, Théorèmes des traces pour des espaces des fonctions de la neutronique, C. R. Acad. Sc. Paris, t. 300, Série I, $n°$ 16 (1985), 89–92.
[5] M. Cessenat, Théorèmes des traces pour des espaces des fonctions de la neutronique, C. R. Acad. Sc. Paris, t. 299, Série I, $n°$ 3 (1984), 831–834.

[6] A. Cohen, W. Dahmen, R. DeVore, Adaptive wavelet methods for elliptic operator equations – Convergence rates, Math. Comp. 70 (2001), 27–75.
[7] A. Cohen, W. Dahmen, R. DeVore, Adaptive wavelet methods II - Beyond the elliptic case, IGPM Report, RWTH Aachen, Nov. 2000.
[8] A. Cohen, W. Dahmen, R. DeVore, Sparse Evaluation of Nonlinear Functionals of Multiscale Expansions, manuscript, Dec 2001.
[9] A. Cohen, W. Dahmen, G. Welper, Adaptivity and stabilization for convection diffusion problems, IGPM Report, January 2011, RWTH Aachen.
[10] S. Dahlke, W. Dahmen, K. Urban, Adaptive wavelet methods for saddle point problems – Convergence rates, SIAM J. Numer. Anal., 40 (No. 4) (2002), 1230–1262.
[11] S. Dahlke, M. Fornasier, T. Raasch, R. Stevenson, M. Werner, *Adaptive Frame Methods for Elliptic Operator Equations: The Steepest Descent Approach* , IMA J. Numer. Anal., **27**(4), 717-740 (2007).
[12] W. Dahmen, C. Huang, Ch. Schwab and G. Welper, Computational Aspects of Adaptive Multiresolution Petrov-Galerkin Discretizations of Transport Equations (in preparation).
[13] T. J. Dijkema, C. Schwab, R. Stevenson, *An adaptive wavelet method for solving high-dimensional elliptic PDEs*, Constr. Approx. **30**(3), 423-455 (2009)
[14] L. F. Demkowicz and J. Gopalakrishnan, A class of discontinuous Petrov-Galerkin Methods I: The transport equation, Comp. Meth. Appl. Mech. Engg. (2010).
[15] K. Grella and Ch. Schwab. Sparse tensor spherical harmonics approximation in radiative transfer. Technical report, SAM, ETH Zrich, 2010. `www.sam.math.ethz.ch/ reports`
[16] G. Kutyniok and D. Labate, *Resolution of the wavefront set using continuous shearlets*, Trans. AMS **361**(2009) 2719-2754.
[17] P. Kittipoom, G. Kutyniok, and W.-Q Lim, *Construction of compactly supported shearlet frames*, preprint.
[18] G. Kutyniok, J. Lemvig, and W.-Q Lim, Compactly Supported Shearlets, preprint.
[19] G. Kutyniok, W.-Q. Lim, Compactly Supported Shearlets are Optimally Sparse, Preprint AAG Dept. Mathematics, Univ. of Osnabruück (in review), (2010).
[20] W.-Q Lim, *The discrete shearlet transform: A new directional transform and compactly supported shearlet frames*, IEEE Trans. Image Proc. 19 (2010), 1166–1180.
[21] G. Rozza, D.B.P. Huynh, and A.T. Patera, *Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations application to transport and continuum mechanics*, Arch. Comput Method E, **15**(2008), 229–275.
[22] C. Schwab, R. Stevenson, Space-Time Adaptive Wavelet Methods for Parabolic Evolution Problems, Math. Comp. **78**, 1293-1318 (2009).
[23] G. Widmer, R. Hiptmair, and Ch. Schwab. Sparse adaptive finite elements for radiative transfer. *Journal of Computational Physics*, 227:60716105, 2008.
[24] Gisela Widmer. *Sparse Finite Elements for Radiative Transfer*. PhD thesis, ETH Zrich, 2009. No. 18420.