IMA Journal of Numerical Analysis (2014) Page 1 of 34
doi:10.1093/imanum/drnxxx

A Petrov-Galerkin discretization with optimal test space of a mild-weak formulation of convection-diffusion equations in mixed form

DIRK BROERSEN† AND ROB STEVENSON‡
Korteweg-de Vries Institute for Mathematics, University of Amsterdam, P.O. Box 94248, 1090 GE Amsterdam, The Netherlands

[Received on *some date*; revised on *another date*]

Motivated by the Discontinous Petrov-Galerkin method from [*Numer. Methods Partial Differential Equations*, 27 (2011), 70–105] by Demkowicz and Gopalakrishnan, we study a variational formulation of second order elliptic equations in mixed form, that is obtained by piecewise integrating *one* of the two equations in the system w.r.t. a partition of the domain into mesh cells. We apply a Petrov-Galerkin discretization with optimal test functions, or equivalently, minimize the residual in the natural norm associated to the variational form. These optimal test functions can be found by solving local problems. Well-posedness, uniformly in the partition, and optimal error estimates are demonstrated.
In the second part of the paper, the application to convection-diffusion problems is studied. The available freedom in the variational formulation and in its optimal Petrov-Galerkin discretization is used to construct a method that allows a (smooth) passing to a converging method in the convective limit, being a necessary condition to retain convergence and having a bound on the cost for a vanishing diffusion. The theoretical findings are illustrated by several numerical results.

Keywords: Petrov-Galerkin discretization, convection-diffusion, optimal test space, least-squares method, mixed formulation, finite elements

1. Introduction

On a domain $\Omega \subset \mathbb{R}^n$, we consider the boundary value problem

$$\begin{cases} -\operatorname{div}\mathbf{A}\nabla u + \mathbf{b}\cdot\nabla u + \gamma u = f & \text{on } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

where $\mathbf{A}$ is positive definite, and $\mathbf{b}$ and γ are such that the standard variational formulation of this problem on $H^1_0(\Omega)\times H^1_0(\Omega)$ is well-posed. It is well-known that Galerkin discretizations of this variational problem give unsatisfactory results in case of dominating convection. We will study a well-posed variational formulation of the mixed formulation

$$\begin{cases} \boldsymbol{\sigma} - \mathbf{A}_2\nabla u = 0 & \text{on } \Omega, \\ -\operatorname{div}\mathbf{A}_1\boldsymbol{\sigma} + \mathbf{b}\cdot\nabla u + \gamma u = f & \text{on } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.2}$$

where $\mathbf{A} = \mathbf{A}_1\mathbf{A}_2$, and consider *Petrov-Galerkin* discretizations of it, where we take an *optimal test space* (Demkowicz & Gopalakrishnan (2011b)).

†Email:D.Broersen@uva.nl
‡Corresponding author. Email: R.P.Stevenson@uva.nl

For an abstract variational problem of finding $u \in U$ such that $b(u,v) = f(v)$ $(v \in V)$, given a trial space $U_h \subset U$, the optimal test space is $V_h = R^{-1}BU_h$, where $(Bu)(v) := b(u,v)$, and $R : V \to V'$ is the Riesz map. The resulting Petrov-Galerkin solution minimizes the residual in V' over the space U_h, and so, for boundedly invertible B, it yields a quasi-best approximation to the solution in U from U_h. The application of this approach for solving convection-diffusion problems can already be found in Barrett & Morton (1984).

For a variational formulation of a boundary value problem, and V being an L_2-space, the optimal test space V_h is found by simply applying the differential operator in strong form to U_h. For other V, finding the optimal test space V_h amounts to solving a symmetric, bounded and coercive variational problem on $V \times V$ for any basis function from a basis for U_h, or to solving a sufficiently accurate Galerkin discretization of such a problem. A main contribution from Demkowicz & Gopalakrishnan (2011b) is the idea to consider variational formulations where V is a "broken space", so that the variational problems that determine V_h are local problems.

In Demkowicz & Gopalakrishnan (2011b), also U is taken to be a broken space which, however, is not essential, although convenient when one aims at applying hp-fem. Regardless whether the variational problems for the test functions are solved exactly, or approximately using a Galerkin discretization, the Petrov-Galerkin solution is found by solving a symmetric positive definite matrix-vector problem.

In the current paper, we study a variational formulation of (1.2) obtained by piecewise integrating the *second* equation by parts w.r.t. a partition of the domain Ω into mesh cells. This introduces the "flux", being the normal component of $u\mathbf{b} - \mathbf{A}_1\boldsymbol{\sigma}$ on the skeleton, as a third *independent* variable. This skeleton may or may not include parts of $\partial\Omega$. We will call our formulation a *mild-weak* variational formulation. It can be considered as intermediate between the *mild* formulation, where neither equation is integrated by parts, and the *ultra-weak* formulation, where both first and second equation are piecewise integrated by parts, giving rise to an additional fourth independent variable, called "trace". Applying a Petrov-Galerkin discretization with optimal test space, the mild or ultra-weak formulations result in the common first order least squares method, or in the Discontinuous Petrov-Galerkin method with optimal test space, DPG method for short, that was introduced in Demkowicz & Gopalakrishnan (2011b).

A reason for us to develop a modification of the DPG method is that with the ultra-weak formulation, both u and $\boldsymbol{\sigma}$ are sought in L_2-spaces. Consequently, assuming finite element spaces of sufficiently high order, in order to obtain an a priori error bound of say $\mathcal{O}(h^k)$ for u in $L_2(\Omega)$, besides the natural condition $u \in H^k(\Omega)$, it is needed that $\boldsymbol{\sigma} \in H^k(\Omega)^n$, and so $u \in H^{k+1}(\Omega)$.

With the mild-weak formulation, u is sought in $H^1_0(\Omega)$, and $\boldsymbol{\sigma}$ in $L_2(\Omega)^n$. We show well-posedness of this formulation, and with that, optimal error estimates for the Petrov-Galerkin discretization with optimal test space. For obtaining an a priori error bound $\mathcal{O}(h^k)$ for u in $H^1_0(\Omega)$, it suffices to have $u \in H^{k+1}(\Omega)$ and $f \in H^k(\Omega)$. The last, additional, condition is needed to guarantee an error bound $\mathcal{O}(h^k)$ for the third variable, being the flux. Being a condition on the right-hand side, however, it is usually harmless. By demonstrating approximation properties of the optimal test space, duality arguments even give optimal error estimates for $(\boldsymbol{\sigma}, u)$ in the space $(H^1(\Omega)^n)' \times L_2(\Omega)$.

Although we already briefly mentioned *convection-dominated* problems, the discussion so far refers to the problem (1.1) for *fixed* $\mathbf{A}$, $\mathbf{b}$, and γ. When we apply the aforementioned three methods –i.e., the standard first order least squares, the DPG method, and the method from the current paper– to the convection-dominated problem defined by $\mathbf{A} = \varepsilon\mathrm{Id}$, $\mathbf{b} \neq 0$ fixed, and $\gamma = 0$, for small ε the results are much better than with standard Galerkin applied to the non-mixed variational formulation. The reason is that all these three methods minimize the residual in some norm.

On the other hand, as our numerical results with uniform meshes show, also with these methods, for small ε, initially, i.e., with relatively large mesh-sizes, there is hardly no reduction of the error in u in $L_2(\Omega)$, and some oscillations are visible. The explanation is that in the limit $\varepsilon = 0$, the operator associated to the bilinear form has an unbounded inverse, which has the consequence that for small $\varepsilon > 0$, some components of the difference of the solution and an approximation from U_h hardly contribute to the residual, and therefore are hardly reduced in the least squares minimization.

In Demkowicz & Heuer (2011), this problem is tackled by equipping V with the problem dependent "optimal test norm", defined such that a residual measured in the resulting norm on V' is equal to the standard, problem independent norm on U of the error. It turns out that the price to be paid for taking this norm on V is that the variational problems on V, that determine the optimal test functions, become increasingly close to singular when $\varepsilon \downarrow 0$, and therefore are more and more difficult to solve with a sufficient accuracy. Modified methods were proposed that aim at finding a compromise between obtaining a best approximation in a nearly ε-independent norm, and getting well-conditioned variational problems for the test functions. In Chan *et al.* (2012), it was proposed to modify the boundary condition at the inflow boundary to ensure that solutions of the dual problem have no boundary layers.

Inspired by Cohen *et al.* (2012), the approach that we investigate is based on the observation that to avoid a numerical solution method looses convergence or becomes increasingly more costly when $\varepsilon \downarrow 0$, a *necessary* condition is that the scheme is well-defined and convergent in the limit $\varepsilon = 0$. To satisfy this condition, we use the available freedom in the Petrov-Galerkin discretization with optimal test space of the mild-weak formulation by factorizing $\mathbf{A} = \varepsilon \mathrm{Id} = \mathbf{A}_1\mathbf{A}_2$ such that both factors vanish for $\varepsilon = 0$; by excluding the outflow boundary from the skeleton, being the domain of definition of the flux; by equipping the test space V with an ε-dependent norm; and finally by making the trial space for the flux ε-dependent. For $\varepsilon = 0$ and a quasi-uniform mesh with mesh size h, the error in u in $L_2(\Omega)$ is shown to be $\mathcal{O}(h^{\frac{1}{2}})$, which is the best that is possible since our piecewise polynomial trial space is in $H^1_0(\Omega)$, whereas the solution generally does not vanish at the boundary outside its inflow part.

To verify the stability of the resulting method for convection dominated convection-diffusion problems, we performed numerical experiments in one and two dimensions using uniform meshes. The method is, however, not restricted to such meshes, and (much) better results can be expected with a proper local refinement in the layers.

A comparison in one dimension with the standard first order least squares method and the DPG method shows that the new method performs much better (we have not yet compared to several variants of the DPG method recently introduced in Demkowicz & Heuer (2011); Chan *et al.* (2012)). With our method, in one and two dimensional examples, for $h \gtrsim \varepsilon$ we observed an error in u in $L_2(\Omega)$ of order $h^{\frac{1}{2}}$. With a piecewise polynomial trial space in $H^1_0(\Omega)$, such an error is the best that is generally possible for solutions that exhibit boundary or internal layers. For any fixed $\varepsilon > 0$, the aforementioned optimal asymptotic error estimates for $h \downarrow 0$ apply.

In one dimension, the optimal test functions could be determined analytically. In two-dimensions, and for piecewise linear or quadratic trial functions, in those cases where the test functions could not be found analytically, we replaced them by Galerkin approximations from the space of piecewise cubics. We solved the symmetric, positive definite matrix-vector problem that defines the Petrov-Galerkin solution using the direct built-in `matlab` solver. With the new method we did not encounter any instabilities due to ill-conditioning.

Finally, we note that there is a vast literature on various classes of numerical methods for solving convection-diffusion problems. The aim of this work is to contribute to the development of Petrov-Galerkin methods with optimal test spaces, or equivalently, to least squares methods. Since these meth-

ods minimize the residual over the trial space in some norm, they have inherent stability properties, and so have the potential to yield near-best approximations with respect to meshes that do not accurately resolve the layers (other than with, e.g., methods based on Shishkin meshes). Since forming a least squares functional essentially means doubling the order of the equation, and so squaring its condition number, a common approach is to apply this technique on a reformulation of the equation as a first order system.

This paper is organized as follows: In Section 2, a relation is established between Petrov-Galerkin discretizations with optimal test functions and least squares methods.

In Section 3, we present the mild, ultra-weak and the new mild-weak variational formulations of second order elliptic boundary value problems in mixed form. We show that the mild-weak variational formulation is well-posed, and therefore gives rise to optimal error estimates in the "energy" space.

In Section 4, using duality arguments, we demonstrate optimal error estimates in a weaker norm.

In Section 5, we discuss the application to convection-dominated convection diffusion problems. We present numerical results in one dimension, that show that a straightforward application of the Petrov-Galerkin discretizations with optimal test spaces of the three variational formulations of the mixed system do not yield satisfactory results for a near-vanishing diffusion.

We study a variational formulation of the pure convection problem obtained by piecewise integrating the equation by parts w.r.t. a partition of Ω into mesh cells, and show that this formulation is well-posed. We then construct a Petrov-Galerkin discretization with optimal test space of the mild-weak variational formulation of convection-diffusion problem, that in the convective limit, becomes such a discretization of this variational formulation of the pure convection problem. We present various numerical experiments with solutions that have boundary and internal layers, which demonstrate the stability of the resulting numerical solution method.

A summary and brief outlook is given in Section 6.

2. Petrov-Galerkin with optimal test spaces, and least-squares approximations

For some real Hilbert spaces U and V, a bilinear form $b : U \times V \to \mathbb{R}$, and with $(Bu)(v) := b(u,v)$, let $B : U \to V'$ be *homeomorphism onto its range*, i.e.,

$$\|Bu\|_{V'} \eqsim \|u\|_U \quad (u \in U). \tag{2.1}$$

Here and in the remainder of this work, by $C \lesssim D$ we will mean that C can be bounded by a multiple of D, independently of parameters which C and D may depend on. Obviously, $C \gtrsim D$ is defined as $D \lesssim C$, and $C \eqsim D$ as $C \lesssim D$ and $C \gtrsim D$.

In the application that is central in this work, $\Im B$ will be equal to V', so that $B : U \to V'$ is boundedly invertible. In this section, we include the possibility that $\Im B \subsetneq V'$ without additional difficulty.

With $R \in \mathscr{B}(V,V')$ being the *Riesz map*, i.e., $(Rv)(w) = \langle v,w\rangle_V$ $(w \in V)$, we define $T = R^{-1}B \in \mathscr{B}(U,V)$. It satisfies

$$\langle Tu,v\rangle_V = b(u,v) \quad (u \in U, v \in V). \tag{2.2}$$

Given a closed linear *trial space* $U_h \subset U$, following Demkowicz & Gopalakrishnan (2011b) we set the *optimal test space*

$$V_h := \Im(T|_{U_h}),$$

and, for given $f \in V'$, consider the *Petrov-Galerkin* problem of finding $u_h \in U_h$ such that

$$b(u_h,v_h) = f(v_h) \quad (v_h \in V_h). \tag{2.3}$$

In the following proposition it is shown that solving the Petrov-Galerkin problem with optimal test space, the *optimal Petrov-Galerkin* problem for short, is equal to the *least squares problem* of minimizing the residual in V'.

PROPOSITION 2.1 It holds that $u_h = \arg\min_{\bar{u}_h \in U_h} \|f - B\bar{u}_h\|_{V'}$.

Proof. For any $u_h, w_h \in U_h$,

$$\begin{aligned}\langle f - Bu_h, Bw_h\rangle_{V'} &= \langle R^{-1}(f - Bu_h), R^{-1}Bw_h\rangle_V \\ &= (f - Bu_h)(R^{-1}Bw_h) = f(v_h) - b(u_h, v_h).\end{aligned}$$

where $v_h := R^{-1}Bw_h$. Note that u_h minimizes the residual when the left-hand side vanishes for any w_h, whereas it solves the Petrov-Galerkin problem when the right-hand side vanishes for any $v_h \in V_h$. □

Note that thanks to (2.1), the least squares problem, and so the optimal Petrov-Galerkin problem have a unique solution.

Equipping U with *energy-norm*

$$\|\cdot\|_E := \|B\cdot\|_{V'},$$

we infer that u_h is the *best approximation* w.r.t. $\|\cdot\|_E$ from U_h to

$$u_{\mathrm{ls}} := \arg\min_{u \in U} \|f - Bu\|_{V'},$$

and thus a quasi-best approximation w.r.t. $\|\cdot\|_U$. Indeed, $\|f - Bu\|_{V'}^2 = \|f - Bu_{\mathrm{ls}}\|_{V'}^2 + \|Bu_{\mathrm{ls}} - Bu\|_{V'}^2$ for any $u \in U$, shows that $u_h = \arg\min_{\bar{u}_h \in U} \|Bu_{\mathrm{ls}} - B\bar{u}_h\|_{V'}$. Clearly, $u_{\mathrm{ls}} = B^{-1}f$ when $\Im B = V'$.

In special cases only, as when b corresponds to a boundary value problem and V is an L_2-space, one can expect to be able to determine the optimal test space exactly. Therefore, let $\tilde{V}_h \subset V$ be a sufficiently large closed subspace such that in any case

$$\forall 0 \neq w_h \in U_h, \exists \tilde{v}_h \in \tilde{V}_h \text{ with } b(w_h, \tilde{v}_h) \neq 0, \tag{2.4}$$

which is satisfied for $\tilde{V}_h = V$ thanks to (2.1). With $R_h \in \mathscr{B}(\tilde{V}_h, \tilde{V}_h')$ defined by $(R_h\tilde{v}_h)(\tilde{w}_h) = \langle \tilde{v}_h, \tilde{w}_h\rangle_V$ ($\tilde{w}_h \in \tilde{V}_h$), we set $T_h = R_h^{-1}B \in \mathscr{B}(U, \tilde{V}_h)$, i.e.,

$$\langle T_h u, \tilde{v}_h\rangle_V = b(u, \tilde{v}_h) \quad (u \in U, \tilde{v}_h \in \tilde{V}_h).$$

Note that $T_h|_{U_h}$ is injective by (2.4). The solution $T_h u$ is the Galerkin approximation from the trial space $\tilde{V}_h \subset V$ to the solution Tu of the bounded, symmetric and coercive variational problem (2.2) on $V \times V$.

The Petrov-Galerkin problem with an *approximately optimal test space* reads as finding $\hat{u}_h \in U_h$ such that

$$b(\hat{u}_h, v_h) = f(v_h) \quad (v_h \in \Im(T_h|_{U_h})). \tag{2.5}$$

Writing $v_h = T_h w_h$, we have $b(\hat{u}_h, v_h) = \langle T_h\hat{u}_h, T_h w_h\rangle_V$, and so, by the injectivity of $T_h|_{U_h}$, we conclude that (2.5) has a unique solution, and moreover, that this variational problem is symmetric and coercive. Taking $\tilde{V}_h = V$, in particular this holds true for (2.3).

When $\dim U_h < \infty$, and $\{\phi_i : i \in I\}$ is a basis for U_h, a basis for $\Im T_h|_{U_h}$ is given by $\{T_h\phi_i : i \in I\}$.

A sufficient condition on $\tilde{V}_h$, dependent on U_h, such that also the Petrov-Galerkin solution with approximately optimal test space gives a quasi-best approximation to u_{ls} in U is given in (Gopalakrishnan & Qiu, 2012, Thm. 2.1).

REMARK 2.1 The Petrov-Galerkin problem (2.5) with the approximately optimal test space can equivalently be written as a symmetric saddle-point system of finding $(\hat{u}_h,\hat{y}_h)\in U_h\times\tilde{V}_h$ that solves

$$\begin{cases} \langle\hat{y}_h,\tilde{v}_h\rangle_V + b(\hat{u}_h,\tilde{v}_h) &= f(\tilde{v}_h) \quad (\tilde{v}_h\in\tilde{V}_h),\\ b(w_h,\hat{y}_h) &= 0 \quad (w_h\in U_h). \end{cases}$$

This is the point of view taken in Cohen *et al.* (2012). To see this, note that for $\tilde{v}_h\in\tilde{V}_h$, and $\hat{u}_h\in U_h$, $f(\tilde{v}_h)-b(\hat{u}_h,\tilde{v}_h)=\langle R_h^{-1}(f-B\hat{u}_h),\tilde{v}_h\rangle_V$, so that the first equation in the saddle point system is equivalent to $\hat{y}_h=R_h^{-1}(f-B\hat{u}_h)$. With this equality, the second equation $0=b(w_h,\hat{y}_h)=\langle R_h^{-1}Bw_h,\hat{y}_h\rangle_V$ $(w_h\in U_h)$ is equivalent to $0=\langle R_h^{-1}(f-B\hat{u}_h),R_h^{-1}Bw_h\rangle_V=(f-B\hat{u}_h)(R_h^{-1}Bw_h)$ $(w_h\in U_h)$, or to $b(\hat{u}_h,v_h)=f(v_h)$ $(v_h\in\mathfrak{T}_h|_{U_h})$.

Writing the above $\hat{y}_h$ as y_h in the case that $\tilde{V}_h=V$, in (Cohen *et al.*, 2012, Lemma 3.3) it was shown that if $\tilde{V}_h$ is chosen to be sufficiently large so that for some $\delta<2$, $\|y_h-\hat{y}_h\|_V\leqslant\delta\|\hat{y}_h\|_V$, then

$$\|u_{\rm ls}-\hat{u}_h\|_E+\|y_h-\hat{y}_h\|_V\leqslant 4(1-\delta/2)^{-2}\|u_{\rm ls}-u_h\|_E.$$

Moreover, for a particular V it is demonstrated how to control $\|y_h-\hat{y}_h\|_V$ using an a posteriori error estimator and adaptivity.

3. Variational formulations of second order elliptic boundary value problems in mixed form

3.1 *A mild-weak variational formulation*

For some bounded Lipschitz domain $\Omega\subset\mathbb{R}^n$, a symmetric $\mathbf{A}\in L_\infty(\Omega)^{n\times n}$ with $\mathbf{A}(\cdot)\gtrsim\mathrm{Id}$ a.e., $\mathbf{b}\in L_\infty(\Omega)^n$ and $\gamma\in L_\infty(\Omega)$, we consider the boundary value problem

$$\begin{cases} -\operatorname{div}\mathbf{A}\nabla u+\mathbf{b}\cdot\nabla u+\gamma u=f & \text{on } \Omega,\\ u=0 & \text{on } \partial\Omega. \end{cases} \tag{3.1}$$

We assume that $\mathbf{b}$ and γ are such that for

$$\begin{cases} (Lu)(v):=\displaystyle\int_\Omega \mathbf{A}\nabla u\cdot\nabla v+(\mathbf{b}\cdot\nabla u+\gamma u)v,\\ L:H_0^1(\Omega)\to H^{-1}(\Omega) \text{ is boundedly invertible.} \end{cases} \tag{3.2}$$

Factorizing $\mathbf{A}=\mathbf{A}_1\mathbf{A}_2$, where $\mathbf{A}_1,\mathbf{A}_2\in L_\infty(\Omega)^{n\times n}$, and introducing $\boldsymbol{\sigma}=\mathbf{A}_2\nabla u$, our problem in *mixed form* reads as

$$\begin{cases} \boldsymbol{\sigma}-\mathbf{A}_2\nabla u=0 & \text{on } \Omega,\\ -\operatorname{div}\mathbf{A}_1\boldsymbol{\sigma}+\mathbf{b}\cdot\nabla u+\gamma u=f & \text{on } \Omega,\\ u=0 & \text{on } \partial\Omega. \end{cases} \tag{3.3}$$

REMARK 3.1 Obvious choices are $\mathbf{A}_1=\mathbf{A}$ or $\mathbf{A}_2=\mathbf{A}$. For convection dominated convection-diffusion problems, which will be discussed in Section 5, it will be relevant to consider a different factorization.

For any h from an index set of mesh parameters, let Ω_h be a collection of disjoint open Lipschitz domains such that $\bar{\Omega}=\cup_{K\in\Omega_h}\bar{K}$. For any $K_1\neq K_2\in\Omega_h$ with $\mathrm{meas}_{n-1}(\bar{K}_1\cap\bar{K}_2)>0$, we fix $\mathbf{n}$ on $\partial K_1\cap\partial K_2$ to be the outward unit normal $\mathbf{n}_K$ on ∂K for K being either K_1 or K_2. By setting $\mathbf{n}$ to be the outward unit normal on $\partial\Omega$, this defines $\mathbf{n}$ on $\cup_{K\in\Omega_h}\partial K$ a.e.

For Γ_+ being the union of $\partial K\cap\partial\Omega$ for all K in some subset of Ω_h, we set the skeleton

$$\partial\Omega_h^\circ=\cup_{K\in\Omega_h}\partial K\setminus\Gamma_+. \tag{3.4}$$

REMARK 3.2 Currently the choice of Γ_+ is *arbitrary*. Canonical choices are $\Gamma_+ = \emptyset$ (as in Demkowicz & Gopalakrishnan (2011a,b)), or $\Gamma_+ = \partial\Omega$. For convection dominated convection-diffusion problems, it will turn out to be relevant to choose Γ_+ as the the outflow boundary, so that $\partial\Omega_h^\circ \cap \partial\Omega$ is the complement of the outflow boundary.

We are going to derive a variational formulation of the mixed problem, where the *second* equation of (3.3) will be integrated by parts on each "element" $K \in \Omega_h$ individually. Note that $\mathrm{div} \in \mathscr{B}(L_\infty(K)^n, W_1^1(K)')$ and

$$\int_K w \,\mathrm{div}\,\mathbf{b} = -\int_K \mathbf{b}\cdot\nabla w + \int_{\partial K} w\,\mathbf{b}\cdot\mathbf{n}_K \quad (w \in W_1^1(K)), \tag{3.5}$$

so that

$$\int_K v\mathbf{b}\cdot\nabla u = \int_K -u\mathbf{b}\cdot\nabla v - uv\,\mathrm{div}\,\mathbf{b} + \int_{\partial K} uv\mathbf{b}\cdot\mathbf{n}_K \quad (u,v \in H^1(K)).$$

In case $\mathbf{A}_1\boldsymbol{\sigma} \in H(\mathrm{div};K)$, we have

$$-\int_K v\,\mathrm{div}\,\mathbf{A}_1\boldsymbol{\sigma} = \int_K \mathbf{A}_1\boldsymbol{\sigma}\cdot\nabla v - \int_{\partial K} v\mathbf{A}_1\boldsymbol{\sigma}\cdot\mathbf{n}_K \quad (v \in H^1(K)). \tag{3.6}$$

By summing these relations over $K \in \Omega_h$, setting div_h and ∇_h by $(\mathrm{div}_h\,\mathbf{b})|_K = \mathrm{div}\,\mathbf{b}$ and $(\nabla_h v)|_K = \nabla v$ ($K \in \Omega_h$), and reading $(u\mathbf{b} - \mathbf{A}_1\boldsymbol{\sigma})|_{\partial\Omega_h^\circ}\cdot\mathbf{n}$ as an additional *independent variable* θ, we end up with the following *mild-weak* variational problem:

$$\begin{array}{l}
\text{With } U := L_2(\Omega)^n \times H_0^1(\Omega) \times H^{-\frac{1}{2}}(\partial\Omega_h^\circ),\ V := L_2(\Omega)^n \times H^1_{0,\Gamma_+}(\Omega_h), \\
\text{given } f \in H^1_{0,\Gamma_+}(\Omega_h)',\ \text{find } (\boldsymbol{\sigma},u,\theta) \in U \text{ such that for all } (\boldsymbol{\tau},v) \in V, \\
\begin{array}{lcl}
b(\boldsymbol{\sigma},u,\theta,\boldsymbol{\tau},v) & := & \displaystyle\int_\Omega (\boldsymbol{\sigma} - \mathbf{A}_2\nabla u)\cdot\boldsymbol{\tau} + (\mathbf{A}_1\boldsymbol{\sigma} - u\mathbf{b})\cdot\nabla_h v + (\gamma - \mathrm{div}_h\,\mathbf{b})uv \\
& & \displaystyle + \int_{\partial\Omega_h^\circ} [\![v]\!]\theta \\
& = & f(v).
\end{array}
\end{array} \tag{3.7}$$

Here, for $x \in \partial K \cap \partial K'$, and with $\mathbf{n}$ pointing into K',

$$[\![v]\!](x) := v|_K(x) - v|_{K'}(x),$$

and $[\![v]\!](x) := v_K(x)$ for $x \in \partial\Omega \cap \partial K$;

$$H^1_{0,\Gamma_+}(\Omega_h) := \{v \in L_2(\Omega) : v|_K \in H^1_{0,\partial K \cap \Gamma_+}(K)\,(K \in \Omega_h)\}, \tag{3.8}$$

equipped with the "broken" norm $\|v\|^2_{H^1(\Omega_h)} := \sum_{K\in\Omega_h} \|v|_K\|^2_{H^1(K)}$; and

$$H^{-\frac{1}{2}}(\partial\Omega_h^\circ) := \{\mathbf{q}|_{\partial\Omega_h^\circ}\cdot\mathbf{n} : \mathbf{q} \in H(\mathrm{div};\Omega)\}, \tag{3.9}$$

equipped with quotient norm

$$\|\theta\|_{H^{-\frac{1}{2}}(\partial\Omega_h^\circ)} := \inf\{\|\mathbf{q}\|_{H(\mathrm{div};\Omega)} : \mathbf{q} \in H(\mathrm{div};\Omega),\ \theta = \mathbf{q}|_{\partial\Omega_h^\circ}\cdot\mathbf{n}\}.$$

Here, for some bounded Lipschitz domain $\Upsilon \subset \mathbb{R}^n$, and a measurable $\Xi \subset \partial\Upsilon$, we set $H^1_{0,\Xi}(\Upsilon) := \{u \in H^1(\Upsilon) : u = 0 \text{ on } \Xi\}$. The above mapping $\mathbf{q} \mapsto \mathbf{q}|_{\partial\Omega_h^\circ} \cdot \mathbf{n}$ is given a precise meaning as the unique extension to $H(\mathrm{div};\Omega)$ of $(\mathscr{D}(\bar{\Omega})^n, \|\cdot\|_{H(\mathrm{div};\Omega)}) \to H^1_{0,\Gamma_+}(\Omega_h)' : \mathbf{q} \mapsto (v \mapsto \int_{\partial\Omega_h^\circ} [\![v]\!]\mathbf{q}\cdot\mathbf{n})$, that, as we will see (in (3.15)), is bounded.

REMARK 3.3 Although U, V and b depend on h, we suppress this in our notation.

REMARK 3.4 For the solution $(\boldsymbol{\sigma}, u, \theta)$, it holds that $\boldsymbol{\sigma} = \mathbf{A}_2\nabla u$. If $f \in L_2(\Omega)$, then on each $K \in \Omega_h$, by definition of a weak divergence, we have $\mathrm{div}\,\mathbf{A}_1\boldsymbol{\sigma} = \mathbf{b}\cdot\nabla u + \gamma u - f \in L_2(K)$, so that by an application of (3.6) in the reversed direction, we infer that $\theta = (u\mathbf{b} - \mathbf{A}_1\boldsymbol{\sigma})|_{\partial\Omega_h^\circ}\cdot\mathbf{n}$.

REMARK 3.5 Since $u \in H^1_0(\Omega)$ in the variational formulation (3.7), there was no strict need to integrate the term $\int_\Omega v\mathbf{b}\cdot\nabla u$ by parts. For the application to convection dominated convection-diffusion problems, this integration by parts is useful, since in the natural "limit" variational formulation of the pure convection problem, the space for u will be $L_2(\Omega)$.

REMARK 3.6 The idea to introduce the "flux" $(u\mathbf{b} - \mathbf{A}_1\boldsymbol{\sigma})|_{\partial\Omega_h^\circ}\cdot\mathbf{n}$ as an independent variable, rather than, as with a common discontinuous Galerkin method, in a discretized setting to replace it on each interface by some average of u and $\boldsymbol{\sigma}$ from both sides of this interface was introduced in Bottasso *et al.* (2002). In Demkowicz & Gopalakrishnan (2011b,a), this idea was combined with the introduction of optimal test spaces. Actually, in both Bottasso *et al.* (2002) and Demkowicz & Gopalakrishnan (2011b,a), the system is considered where both equations from (3.3) are integrated by parts. This so-called ultra-weak formulation will be considered in the next subsection.

In Subsection 3.3, we will show that the bilinear form $b : U \times V \to \mathbb{R}$ defines a boundedly invertible operator. This means that given a closed trial space $U_h \subset U$, we can run the Petrov-Galerkin method with optimal test space $V_h = \mathfrak{I}(T|_{U_h})$.

Writing

$$(\boldsymbol{\tau}, v) = T(\boldsymbol{\sigma}, u, \theta)$$

and $v = (v_K)_{K\in\Omega_h}$, we have the explicit expression

$$\boldsymbol{\tau} = \boldsymbol{\sigma} - \mathbf{A}_2\nabla u, \tag{3.10}$$

whereas for each $K \in \Omega_h$, $v_K \in H^1_{0,\partial K\cap\Gamma_+}(K)$ solves

$$\langle v_K, \hat{v}\rangle_{H^1(K)} = \int_K (\mathbf{A}_1\boldsymbol{\sigma} - u\mathbf{b})\cdot\nabla\hat{v} + (\gamma - \mathrm{div}\,\mathbf{b})u\hat{v} + \int_{\partial K} \mathbf{n}_K^\top\mathbf{n}\,\hat{v}\theta \tag{3.11}$$

($\hat{v} \in H^1_{0,\partial K\cap\Gamma_+}(K)$). Note that $\mathbf{n}_K^\top\mathbf{n}$ is ± 1. So in the common situation that U_h is spanned by a local basis, i.e., the support of each basis function extends to a uniformly bounded number of elements $K \in \Omega_h$, a local basis of V_h can be found by solving $\mathscr{O}(\#\Omega_h)$ of the above independent local problems on the individual elements. Here the essential point is that v is sought in the "broken space" $H^1_{0,\Gamma_+}(\Omega_h)$ as a consequence of the application of integration by parts on the individual elements when setting up the variational formulation.

3.2 *Mild and ultra-weak variational formulations*

The mild-weak variational formulation (3.7) of the mixed problem (3.3) is intermediate between the variational form where *neither* of the equations is integrated by parts and the one where *both* equations

are piecewise integrated by parts. The first one reads as

$$\left\{\begin{array}{lll} \text{With } U_1 := H(\mathrm{div};\Omega)\times H^1_0(\Omega),\, V_1 := L_2(\Omega)^n \times L_2(\Omega), & & \\ \text{given } f\in L_2(\Omega),\ \text{find } (\boldsymbol{\sigma},u)\in U_1 \text{ such that for all } (\boldsymbol{\tau},v)\in V_1, & & \\ b_1(\boldsymbol{\sigma},u,\boldsymbol{\tau},v) & := & \displaystyle\int_\Omega (\boldsymbol{\sigma}-\mathbf{A}\nabla u)\cdot\boldsymbol{\tau} + v(-\operatorname{div}\boldsymbol{\sigma}+\mathbf{b}\cdot\nabla u+\gamma u) \\ & = & f(v). \end{array}\right. \tag{3.12}$$

As shown in (Stevenson, 2013, (proof of) Thm. 3.1), under the assumptions that we made on $\mathbf{A}$, $\mathbf{b}$, and γ, with $(B_1(\boldsymbol{\sigma},u))(\boldsymbol{\tau},v) := b_1(\boldsymbol{\sigma},u,\boldsymbol{\tau},v)$, $B_1 : U_1 \to V_1'$ is boundedly invertible. The solution of the Petrov Galerkin discretization of this *mild* variational formulation with trial space $U_h \subset U$ and optimal test space solves the common *first order least squares problem*

$$\operatorname*{arg\,min}_{(\sigma_h,u_h)\in U_h} \|\mathbf{A}\nabla u_h - \sigma_h\|^2_{L_2(\Omega)^n} + \|f+\operatorname{div}\sigma_h - \mathbf{b}\cdot\nabla u_h - \gamma u_h\|^2_{L_2(\Omega)^n}. \tag{3.13}$$

Note that $T(\boldsymbol{\sigma},u) = (\boldsymbol{\sigma}-\mathbf{A}\nabla u, -\operatorname{div}\boldsymbol{\sigma}+\mathbf{b}\cdot\nabla u+\gamma u)$, which corresponds to simply applying the operator in strong form.

Following Bottasso *et al.* (2002); Demkowicz & Gopalakrishnan (2011b,a), we give the second alternative formulation, known as the *ultra-weak* formulation, where we restrict to $\mathbf{b}$ with $\operatorname{div}\mathbf{b}=0$ and $\gamma = 0$. With $\partial\Omega_h = \cup_{K\in\Omega_h}\partial K$, i.e., $\partial\Omega_h = \partial\Omega_h^\circ$ taking $\Gamma_+ = \emptyset$, it reads as

$$\left\{\begin{array}{lll} \text{With } U_2 := L_2(\Omega)^n \times L_2(\Omega) \times H^{\frac12}_{00}(\partial\Omega_h)\times H^{-\frac12}(\partial\Omega_h), & & \\ V_2 := H(\mathrm{div};\Omega_h)\times H^1(\Omega_h),\ \text{given } f\in H^1(\Omega_h)',\ \text{find } (\boldsymbol{\sigma},u,\rho,\theta)\in U_2 \text{ such that} & & \\ b_2(\boldsymbol{\sigma},u,\rho,\theta,\boldsymbol{\tau},v) & := & \displaystyle\int_\Omega \mathbf{A}^{-1}\boldsymbol{\sigma}\cdot\boldsymbol{\tau} + u\operatorname{div}_h\boldsymbol{\tau} + (\boldsymbol{\sigma}-u\mathbf{b})\cdot\nabla_h v \\ & & \displaystyle + \int_{\partial\Omega_h} [\![v]\!]\theta - [\![\boldsymbol{\tau}\cdot\mathbf{n}]\!]\rho \\ & = & f(v) \qquad\qquad ((\boldsymbol{\tau},v)\in V_2), \end{array}\right. \tag{3.14}$$

where ρ and θ replace the "trace" $u|_{\partial\Omega_h}$ and "flux" $(u\mathbf{b}-\boldsymbol{\sigma})|_{\partial\Omega_h}\cdot\mathbf{n}$, which are not defined on the full function spaces $L_2(\Omega)$ and $L_2(\Omega)^n$ for u and $\boldsymbol{\sigma}$. Here

$$H(\mathrm{div};\Omega_h) := \{\boldsymbol{\tau}\in L_2(\Omega)^n : \operatorname{div}\boldsymbol{\tau}|_K \in H(\mathrm{div};K)\,(K\in\Omega_h)\},$$

equipped with the "broken" norm $\|\boldsymbol{\tau}\|^2_{H(\mathrm{div};\Omega_h)} := \sum_{K\in\Omega_h}\|\boldsymbol{\tau}|_K\|^2_{H(\mathrm{div};K)}$; and

$$H^{\frac12}_{00}(\partial\Omega_h) := \{u|_{\partial\Omega_h} : u\in H^1_0(\Omega)\},$$

equipped with quotient norm

$$\|\rho\|_{H^{\frac12}_{00}(\partial\Omega_h)} := \inf\{\|u\|_{H^1(\Omega)} : u\in H^1_0(\Omega),\, \rho = u|_{\partial\Omega_h}\}.$$

As shown in Demkowicz & Gopalakrishnan (2011a), under the conditions we made on $\mathbf{A}$ and for divergence-free $\mathbf{b}\in L_\infty(\Omega)^n$, with $(B_2(\boldsymbol{\sigma},u,\rho,\theta))(\boldsymbol{\tau},v) := b_2(\boldsymbol{\sigma},u,\rho,\theta,\boldsymbol{\tau},v)$, $B_2 : U_2 \to V_2'$ is boundedly invertible, uniformly in h.

Although it can be expected that results can be generalized to $\gamma \neq 0$ and $\operatorname{div}\mathbf{b} \neq 0$, a non-divergence free $\mathbf{b}$ requires some care since, only assuming that $\mathbf{b} \in L_\infty(\Omega)^n$, a term $\int_K uv \operatorname{div}\mathbf{b}$ only makes sense when $uv \in W_1^1(K)$.

Let us consider a Petrov-Galerkin discretization of the ultra-weak formulation with finite element trial spaces w.r.t. Ω_h and an optimal test space. Then, assuming a quasi-uniform Ω_h with mesh-size h, in order to obtain an error in U_2 of $\mathcal{O}(h^k)$, it is needed that $\boldsymbol{\sigma} \in H^k(\Omega)^n$, $u \in H^k(\Omega)$, and, for approximating ρ and θ, $u \in H^{k+1}(\Omega)$, $u\mathbf{b} - \boldsymbol{\sigma} \in H^k(\Omega)^n$, and $\operatorname{div}(u\mathbf{b} - \boldsymbol{\sigma}) \in H^k(\Omega)$. So, to approximate u in $L_2(\Omega)$ with an error $\mathcal{O}(h^k)$, it does not suffice that $u \in H^k(\Omega)$, but it is needed that $u \in H^{k+1}(\Omega)$. The avoidance of such an additional regularity condition was one reason to consider the mild-weak variational formulation.

A similar problem seems to arise with the common first order least squares formulation. To approximate the solution with finite element spaces w.r.t. a quasi-uniform partition with mesh-size h such that the error in U_1 is $\mathcal{O}(h^k)$, it is needed, assuming Raviart-Thomas type spaces for $\boldsymbol{\sigma}$, that $\boldsymbol{\sigma} \in H^k(\Omega)^n$, $\operatorname{div}\boldsymbol{\sigma} \in H^k(\Omega)$, and $u \in H^{k+1}(\Omega)$. In view of the relations $\boldsymbol{\sigma} = \mathbf{A}\nabla u$ and $\operatorname{div}\mathbf{A}\nabla u = \mathbf{b}\cdot\nabla u + \gamma u - f$, however, here the additional condition $\operatorname{div}\boldsymbol{\sigma} \in H^k(\Omega)$, needed to approximate u in $H_0^1(\Omega)$ with an error $\mathcal{O}(h^k)$, is already satisfied under the additional smoothness condition $f \in H^k(\Omega)$, which, being a smoothness condition on the right-hand side, is usually harmless.

REMARK 3.7 This sheds some other light on the discussion in Bramble *et al.* (1997), where the undesirability of the additional smoothness requirement on $\boldsymbol{\sigma}$ was used as one reason to consider the variational form in (3.12) for $(\boldsymbol{\sigma}, u, \boldsymbol{\tau}, v) \in L_2(\Omega)^n \times H_0^1(\Omega) \times L_2(\Omega)^n \times H^{-1}(\Omega)$.

Finally, we note that since, other than with the other two variational formulations, in the ultra-weak formulation both $\boldsymbol{\sigma}$ and u are sought in L_2-spaces, it allows for a convenient application with nonconforming partitions.

3.3 *Well-posedness*

In this subsection, we prove the following result for the mild-weak variational formulation.

THEOREM 3.1 With $(B(\boldsymbol{\sigma}, u, \theta))(\boldsymbol{\tau}, v) := b(\boldsymbol{\sigma}, u, \theta, \boldsymbol{\tau}, v)$ from (3.7), it holds that $B : U \to V'$ is boundedly invertible with $\sup_h \max(\|B\|_{U\to V'}, \|B^{-1}\|_{V'\to U}) < \infty$.

Our proof is inspired by a corresponding proof from Demkowicz & Gopalakrishnan (2011a) (see also Causin & Sacco (2005)) for the ultra-weak formulation. Here we also allow $\operatorname{div}\mathbf{b} \neq 0$, $\gamma \neq 0$, and dimensions $n \notin \{2, 3\}$. Moreover, we include the possibility that parts of $\partial\Omega$ are excluded from the skeleton $\partial\Omega_h^\circ$, i.e., we allow $\Gamma_+ \neq \emptyset$ in (3.4), which will show up to be useful for convection dominated problems.

We will make use of the following well-known consequence of the *closed range theorem*.

LEMMA 3.1 For reflexive Banach spaces X and Y, let $G : X \to Y'$ be linear. Then G is boundedly invertible if and only if

(i). G is bounded,

(ii). $\rho := \inf_{0\neq y\in Y} \sup_{0\neq x\in X} \frac{(Gx)(y)}{\|x\|_X \|y\|_Y} > 0$,

(iii). $\forall 0 \neq x \in X, \exists y \in Y$, with $(Gx)(y) \neq 0$,

and $\|G^{-1}\|_{Y'\to X} = \frac{1}{\rho}$.

(ii) Gives G' injective and $\ran G'$ closed, so G surj. Reflexivity not needed

Since G being boundedly invertible is equivalent to G' being boundedly invertible, the roles of X and Y in (ii) and (iii) can be interchanged.

To show Theorem 3.1 we start with some preparations. First we will show that for $v \in H^1_{0,\Gamma_+}(\Omega_h)$, the jump $[\![v]\!]$, being a function on the skeleton $\partial\Omega_h^\circ$, is an element of $(H^{-\frac{1}{2}}(\partial\Omega_h^\circ))'$. In particular, using that for $v \in H^1_0(\Omega)$ this jump is zero, we obtain the following result.

THEOREM 3.2 For $v \in H^1_{0,\Gamma_+}(\Omega_h)$, it holds that $[\![v]\!] \in (H^{-\frac{1}{2}}(\partial\Omega_h^\circ))'$, and

$$\|[\![v]\!]\|_{(H^{-\frac{1}{2}}(\partial\Omega_h^\circ))'} \eqsim \inf_{w\in H^1_0(\Omega)} \|v-w\|_{H^1(\Omega_h)} \quad (v \in H^1_{0,\Gamma_+}(\Omega_h)).$$

Proof. For $v \in H^1_{0,\Gamma_+}(\Omega_h)$, $\mathbf{q} \in H(\mathrm{div};\Omega)$, we have

$$\int_{\partial\Omega_h^\circ} [\![v]\!]\mathbf{q}\cdot\mathbf{n} = \sum_{K\in\Omega_h}\int_K \nabla v\cdot\mathbf{q} + v\,\mathrm{div}\,\mathbf{q} \lesssim \|v\|_{H^1(\Omega_h)}\|\mathbf{q}\|_{H(\mathrm{div};\Omega)}, \tag{3.15}$$

showing that $\|[\![v]\!]\|_{(H^{-\frac{1}{2}}(\partial\Omega_h^\circ))'} \lesssim \|v\|_{H^1(\Omega_h)}$.

Since for $w \in H^1_0(\Omega)$ and $\mathbf{q} \in H(\mathrm{div};\Omega)$, it holds that $\int_\Omega \nabla w\cdot\mathbf{q} + w\,\mathrm{div}\,\mathbf{q} = 0$, it follows that $\|[\![w]\!]\|_{(H^{-\frac{1}{2}}(\partial\Omega_h^\circ))'} = 0$. We infer that for $v \in H^1_{0,\Gamma_+}(\Omega_h)$, $\|[\![v]\!]\|_{(H^{-\frac{1}{2}}(\partial\Omega_h^\circ))'} \lesssim \inf_{w\in H^1_0(\Omega)}\|v-w\|_{H^1(\Omega_h)}$.

Given $v \in H^1_{0,\Gamma_+}(\Omega_h)$, let $w \in H^1_0(\Omega)$ be the solution of $\int_\Omega \nabla w\cdot\nabla\phi = \int_\Omega \nabla_h v\cdot\nabla\phi$ $(\phi \in H^1_0(\Omega))$. Define $\boldsymbol{\tau} := \nabla_h(v-w)$. Then $\mathrm{div}\,\boldsymbol{\tau} = 0$, and so

$$\begin{aligned}\|\boldsymbol{\tau}\|^2_{L_2(\Omega)^n} &= \int_\Omega \nabla_h(v-w)\cdot\boldsymbol{\tau} = \sum_{K\in\Omega_h}\int_{\partial K}(v-w)\boldsymbol{\tau}\cdot\mathbf{n}_K = \int_{\partial\Omega_h^\circ}[\![v]\!]\boldsymbol{\tau}\cdot\mathbf{n}\\ &\leqslant \frac{\int_{\partial\Omega_h^\circ}[\![v]\!]\boldsymbol{\tau}\cdot\mathbf{n}}{\|\boldsymbol{\tau}\|_{H(\mathrm{div};\Omega)}}\|\boldsymbol{\tau}\|_{H(\mathrm{div};\Omega)} \leqslant \|[\![v]\!]\|_{(H^{-\frac{1}{2}}(\partial\Omega_h^\circ))'}\|\boldsymbol{\tau}\|_{H(\mathrm{div};\Omega)},\end{aligned}$$

or $\|\nabla_h(v-w)\|_{L_2(\Omega)^n} \leqslant \|[\![v]\!]\|_{(H^{-\frac{1}{2}}(\partial\Omega_h^\circ))'}$.

Since $\|v-w\|_{L_2(\Omega)} \leqslant \|\nabla_h(v-w)\|_{L_2(\Omega)} + \|[\![v]\!]\|_{(H^{-\frac{1}{2}}(\partial\Omega_h^\circ))'}$ by Lemma 3.2 stated below, the proof is completed. □

LEMMA 3.2 (Poincaré-type inequality, (Demkowicz & Gopalakrishnan, 2011a, Lemma 4.2)) It holds that

$$\|v\|_{L_2(\Omega)} \lesssim \|\nabla_h v\|_{L_2(\Omega)} + \|[\![v]\!]\|_{(H^{-\frac{1}{2}}(\partial\Omega_h^\circ))'} \quad (v \in H^1_{0,\Gamma_+}(\Omega_h)).$$

Proof. For the reader's convenience, we recall the proof from Demkowicz & Gopalakrishnan (2011a). Given $v \in H^1_{0,\Gamma_+}(\Omega_h)$, let $w \in H^1_0(\Omega)$ solve $\int_\Omega \nabla w\cdot\nabla\phi = \int_\Omega v\phi$ $(\phi \in H^1_0(\Omega))$. Then $\|\nabla w\|^2_{L_2(\Omega)} \leqslant \|v\|_{L_2(\Omega)}\|w\|_{L_2(\Omega)}$, and so $\|\nabla w\|_{L_2(\Omega)} \lesssim \|v\|_{L_2(\Omega)}$ by an application of the Poincaré inequality on $H^1_0(\Omega)$.

From

$$\begin{aligned}\|v\|^2_{L_2(\Omega)} &= \int_\Omega -v\Delta w = \int_\Omega \nabla w\cdot\nabla_h v - \int_{\partial\Omega_h^\circ}[\![v]\!]\frac{\partial w}{\partial\mathbf{n}}\\ &\leqslant \|\nabla w\|_{L_2(\Omega)}\|\nabla_h v\|_{L_2(\Omega)} - \frac{\int_{\partial\Omega_h^\circ}[\![v]\!]\frac{\partial w}{\partial\mathbf{n}}}{\|\nabla w\|_{H(\mathrm{div};\Omega)}}\|\nabla w\|_{H(\mathrm{div};\Omega)}\\ &\leqslant \|\nabla w\|_{L_2(\Omega)}\|\nabla_h v\|_{L_2(\Omega)} + \|[\![v]\!]\|_{(H^{-\frac{1}{2}}(\partial\Omega_h^\circ))'}\|\nabla w\|_{H(\mathrm{div};\Omega)},\end{aligned}$$

$\|\nabla w\|_{H(\mathrm{div};\Omega)} = \sqrt{\|\nabla w\|^2_{L_2(\Omega)^n} + \|v\|^2_{L_2(\Omega)}}$, and $\|\nabla w\|_{L_2(\Omega)} \lesssim \|v\|_{L_2(\Omega)}$, the proof follows. □

A direct consequence of Theorem 3.2 is that $\|[\![v]\!]\|_{(H^{-\frac{1}{2}}(\partial\Omega_h^\circ))'} \lesssim \|v\|_{H^1(\Omega_h)}$ $(v \in H^1_{0,\Gamma_+}(\Omega_h))$. Using in addition that $\mathbf{A}_1, \mathbf{A}_2 \in L_\infty(\Omega)^{n\times n}$, $\gamma \in L_\infty(\Omega)$, and $\mathbf{b} \in L_\infty(\Omega)^n$, the latter showing that $|\int_K \mathrm{div}\,\mathbf{b}\, uv| \lesssim \|uv\|_{W^1_1(K)} \lesssim \|u\|_{H^1(K)}\|v\|_{H^1(K)}$ (cf. (3.5)), it follows that $B : U \to V'$ is bounded, i.e., condition (i) of Lemma 3.1 applied to $G = B$ is satisfied, and, moreover, that it holds true *uniformly in h*.

To show condition (iii) of Lemma 3.1, we will need a characterization of the dual of a certain trace space. As an introduction, we recall some facts about quotient spaces.

For a normed linear space V, and a closed subspace M, the quotient space V/M is equipped with $\|v\|_{V/M} = \inf_{\{\tilde v \in V : v - \tilde v \in M\}} \|\tilde v\|_H$. If V is a Banach (Hilbert) space, then so is V/M. With the annihilator $M^\circ := \{f \in V' : f(M) = \{0\}\}$, being a closed subspace of V', we have $(V/M)' \simeq M^\circ$.

For a linear space W, let $G \in \mathscr{B}(V,W)$, so that $\ker G$ is closed. From $G : V/\ker G \to \mathfrak{I}G$ being invertible, $|||Gv||| := \|v\|_{V/\ker G} = \inf_{\{\tilde v \in V : G\tilde v = Gv\}} \|\tilde v\|_V$ defines a norm on $\mathfrak{I}G$, and $(\mathfrak{I}G, |||\cdot|||) \simeq V/\ker G$. From $\|Gv\|_W \leqslant \|G\|_{\mathscr{B}(V,W)}|||Gv|||$, we have $(\mathfrak{I}G, |||\cdot|||) \hookrightarrow W$.

We are going to apply these facts in the following situation. Let $\Upsilon \subset \mathbb{R}^n$ be a bounded Lipschitz domain, and $\Xi \subset \partial\Upsilon$ with $|\Xi| > 0$. Setting $\Xi^c = \partial\Upsilon \setminus \Xi$, the condition on Υ ensures that the trace mapping $v \mapsto v|_\Xi \in \mathscr{B}(H^1_{0,\Xi^c}(\Upsilon), L_2(\Xi))$. Its kernel is the space $H^1_0(\Upsilon)$. We define $H^{\frac{1}{2}}_{00}(\Xi) := \{v|_\Xi : v \in H^1_{0,\Xi}(\Upsilon)\}$, equipped with

$$\|v|_\Xi\|_{H^{\frac{1}{2}}_{00}(\Xi)} := \inf_{\{\tilde v \in H^1_{0,\Xi^c}(\Upsilon) : \tilde v|_\Xi = v|_\Xi\}} \|\tilde v\|_{H^1(\Upsilon)}.$$

i.e., $H^{\frac{1}{2}}_{00}(\Xi) \simeq H^1_{0,\Xi^c}(\Upsilon)/H^1_0(\Upsilon)$ $(\hookrightarrow L_2(\Xi)$, with the embedding being even dense), and so $(H^{\frac{1}{2}}_{00}(\Xi))' \simeq H^1_0(\Upsilon)^\circ$ $(\subset H^1_{0,\Xi^c}(\Upsilon)')$.

LEMMA 3.3 For $\Upsilon \subset \mathbb{R}^n$ being a bounded Lipschitz domain, and $\Xi \subset \partial\Upsilon$ with $|\Xi| > 0$, we have $(H^{\frac{1}{2}}_{00}(\Xi))' = \{\mathbf{q}\cdot\mathbf{n}|_\Xi : \mathbf{q} \in H(\mathrm{div};\Upsilon)\}$, and

$$\|\mathbf{q}\cdot\mathbf{n}|_\Xi\|_{(H^{\frac{1}{2}}_{00}(\Xi))'} = \inf_{\{\tilde{\boldsymbol{q}} \in H(\mathrm{div};\Upsilon) : \tilde{\boldsymbol{q}}\cdot\mathbf{n}|_\Xi = \mathbf{q}\cdot\mathbf{n}|_\Xi\}} \|\tilde{\boldsymbol{q}}\|_{H(\mathrm{div};\Upsilon)}, \tag{3.16}$$

i.e., $(H^{\frac{1}{2}}_{00}(\Xi))' \simeq H(\mathrm{div};\Upsilon)/H_{0,\Xi}(\mathrm{div};\Upsilon)$, where $H_{0,\Xi}(\mathrm{div};\Upsilon) := \{\mathbf{q} \in H(\mathrm{div};\Upsilon) : \mathbf{q}\cdot\mathbf{n}|_\Xi = 0\}$.

Proof. Given $f \in (H^{\frac{1}{2}}_{00}(\Xi))'$, define $u \in H^1_{0,\Xi^c}(\Upsilon) \subset H^1(\Upsilon)$ by

$$\int_\Upsilon \nabla u\cdot\nabla v + uv = f(v|_\Xi) \quad (v \in H^1_{0,\Xi^c}(\Upsilon)). \tag{3.17}$$

Then $\|u\|_{H^1(\Upsilon)} = \|f\|_{(H^{\frac{1}{2}}_{00}(\Xi))'}$. Since $f(v|_\Xi)$ vanishes for $v \in H^1_0(\Upsilon)$, setting $\mathbf{q} = \nabla u$, we find $\mathrm{div}\,\mathbf{q} = u$, and so

$$\int_\Upsilon \mathbf{q}\cdot\nabla v + v\,\mathrm{div}\,\mathbf{q} = f(v|_\Xi) \quad (v \in H^1_{0,\Xi^c}(\Upsilon)),$$

or $\mathbf{q}\cdot\mathbf{n}|_\Xi = f$, and $\|\mathbf{q}\|_{H(\mathrm{div};\Upsilon)} = \|f\|_{(H^{\frac{1}{2}}_{00}(\Xi))'}$.

Given $\mathbf{q} \in H(\mathrm{div};\Upsilon)$, $f := v \mapsto \int_\Xi \mathbf{q}\cdot\mathbf{n}v = \int_\Upsilon \mathbf{q}\cdot\nabla v + v\,\mathrm{div}\,\mathbf{q} \in (H^1_{0,\Xi^c}(\Upsilon))'$, and it vanishes on $H^1_0(\Upsilon)$, i.e., $f \in (H^{\frac{1}{2}}_{00}(\Xi))'$. From $|f(v)| \leqslant \|v\|_{H^1(\Upsilon)}\|\mathbf{q}\|_{H(\mathrm{div};\Upsilon)}$, i.e., $\|f\|_{(H^{\frac{1}{2}}_{00}(\Xi))'} \leqslant \|\mathbf{q}\|_{H(\mathrm{div};\Upsilon)}$, the proof is completed. □

REMARK 3.8 Let $H^{\frac{1}{2}}(\Xi) := \{v|_\Xi : v \in H^1(\Upsilon)\}$, equipped with

$$\|v|_\Xi\|_{H^{\frac{1}{2}}(\Xi)} := \inf_{\{\tilde{v}\in H^1(\Upsilon):\tilde{v}|_\Xi = v|_\Xi\}} \|\tilde{v}\|_{H^1(\Upsilon)}.$$

i.e., $H^{\frac{1}{2}}(\Xi) \simeq H^1(\Upsilon)/H^1_{0,\Xi}(\Upsilon)$. Then, in a similar way, one finds that $H^{-\frac{1}{2}}(\Xi) := (H^{\frac{1}{2}}(\Xi))' = \{\mathbf{q}\cdot\mathbf{n}|_\Xi : \mathbf{q} \in H_{0,\Xi^c}(\mathrm{div};\Upsilon)\}$, and

$$\|\mathbf{q}\cdot\mathbf{n}|_\Xi\|_{H^{-\frac{1}{2}}(\Xi)} = \inf_{\{\tilde{\boldsymbol{q}}\in H_{0,\Xi^c}(\mathrm{div};\Upsilon):\tilde{\boldsymbol{q}}\cdot\mathbf{n}|_\Xi = \mathbf{q}\cdot\mathbf{n}|_\Xi\}} \|\tilde{\boldsymbol{q}}\|_{H(\mathrm{div};\Upsilon)}, \tag{3.18}$$

i.e., $H^{-\frac{1}{2}}(\Xi) \simeq H_{0,\Xi^c}(\mathrm{div};\Upsilon)/H_0(\mathrm{div};\Upsilon)$. A comparison of (3.16) with (3.18) confirms that $H^{-\frac{1}{2}}(\Xi) \hookrightarrow (H^{\frac{1}{2}}_{00}(\Xi))'$, being a consequence of $H^{\frac{1}{2}}_{00}(\Xi) \hookrightarrow H^{\frac{1}{2}}(\Xi)$.

It is known that the embedding $H^{\frac{1}{2}}_{00}(\Xi) \hookrightarrow H^{\frac{1}{2}}(\Xi)$ is strict (see e.g. (Grisvard, 1985, Corollary 1.4.4.5)), i.e., there is a sequence $(u_n)_{n\in\mathbb{N}} \subset H^{\frac{1}{2}}_{00}(\Xi)$ with $\|u_n\|_{H^{\frac{1}{2}}(\Xi)} = 1$ and $\lim_{n\to\infty}\|u_n\|_{H^{\frac{1}{2}}_{00}(\Xi)} = \infty$. Consequently, also the embedding $H^{-\frac{1}{2}}(\Xi) \hookrightarrow (H^{\frac{1}{2}}_{00}(\Xi))'$ is strict.

Now we are ready to show condition (iii) of Lemma 3.1. Let $(\boldsymbol{\sigma},u,\theta) \in U$ with $b(\boldsymbol{\sigma},u,\theta,\boldsymbol{\tau},v) = 0$ for all $(\tau,v) \in V$. Then $\boldsymbol{\sigma} = \mathbf{A}_2\nabla u$. By taking $v \in H^1_0(\Omega)$, from $-\int_\Omega u\mathbf{b}\cdot\nabla v + \mathrm{div}_h\,\mathbf{b}uv = \int_\Omega v\mathbf{b}\cdot\nabla u - \sum_{K\in\Omega_h}\int_{\partial K} uv\mathbf{b}\cdot\mathbf{n}_k = \int_\Omega v\mathbf{b}\cdot\nabla u$, it follows that $\int_\Omega \mathbf{A}\nabla u\cdot\nabla v + (\mathbf{b}\cdot\nabla u + \gamma u)v = 0$, or, by (3.2), that $u = 0$, and so also $\boldsymbol{\sigma} = 0$.

Writing $\theta = \mathbf{q}|_{\partial\Omega_h^\circ}\cdot\mathbf{n}$ for $\mathbf{q} \in H(\mathrm{div};\Omega)$, from $0 = \int_{\partial\Omega_h^\circ}[\![v]\!]\mathbf{q}\cdot\mathbf{n} = \sum_{K\in\Omega_h}\int_K \nabla v\cdot\mathbf{q} + v\,\mathrm{div}\,\mathbf{q}$ for all $v \in H^1_{0,\Gamma_+}(\Omega_h)$, it follows that for any $K \in \Omega_h$,

$$0 = \int_K \nabla v\cdot\mathbf{q} + v\,\mathrm{div}\,\mathbf{q} = \int_{\partial K\setminus\Gamma_+} v\mathbf{q}\cdot\mathbf{n}_K \quad (v \in H^1_{0,\partial K\cap\Gamma_+}(K)).$$

Since $v|_{\partial K\setminus\partial\Gamma_+}$ runs over $H^{\frac{1}{2}}_{00}(\partial K\setminus\Gamma_+)$, and $\mathbf{q}|_K \in H(\mathrm{div};K)$, Lemma 3.3 shows that $\mathbf{q}\cdot\mathbf{n}_K$ vanishes on $\partial K\setminus\Gamma_+$, and thus that $\theta = 0$. We conclude that condition (iii) of Lemma 3.1 is satisfied.

REMARK 3.9 We just showed that for any $0 \neq \theta \in H^{-\frac{1}{2}}(\partial\Omega_h^\circ)$, there is a $v \in H^1_{0,\Gamma_+}(\Omega_h)$ with $\int_{\partial\Omega_h^\circ}[\![v]\!]\theta \neq 0$. As a consequence of Theorem 3.1 (whose proof still has to be completed), we even have that $\inf_{0\neq\theta\in H^{-\frac{1}{2}}(\partial\Omega_h^\circ)} \sup_{0\neq v\in H^1_{0,\Gamma_+}(\Omega_h)} \frac{\int_{\partial\Omega_h^\circ}[\![v]\!]\theta}{\|\theta\|_{H^{-\frac{1}{2}}(\partial\Omega_h^\circ)}\|v\|_{H^1(\Omega_h)}} > 0$. Since, on the other hand, as a consequence of Theorem 3.2, for any $v \in H^1_{0,\Gamma_+}(\Omega_h)\setminus H^1_0(\Omega)$, there exists a $\theta \in H^{-\frac{1}{2}}(\partial\Omega_h^\circ)$ with $\int_{\partial\Omega_h^\circ}[\![v]\!]\theta \neq 0$, we conclude that we have the following relation between two function spaces on the skeleton $\partial\Omega_h^\circ$:

$$(H^1_{0,\Gamma_+}(\Omega_h)/H^1_0(\Omega))' \simeq H(\mathrm{div};\Omega)/\{\mathbf{q}\in H(\mathrm{div};\Omega) : \mathbf{q}|_{\partial\Omega_h^\circ}\cdot\mathbf{n} = 0\}. \tag{3.19}$$

(So also for $|\Gamma_+| > 0$, i.e., when a non-neglectable part of $\partial\Omega$ is excluded from the skeleton, no boundary conditions should be included in $H(\mathrm{div};\Omega)$, actually contradicting (Ben Belgacem, 1999, Prop. 2.1).) Note that (3.19) extends upon Theorem 3.2.

To verify the remaining condition (ii) of Lemma 3.1, we need the following auxiliary result.

LEMMA 3.4 The linear mappings

$$L_2(\Omega)^n \times H^1_0(\Omega) \to (L_2(\Omega)^n \times H^1_0(\Omega))' :$$
$$\begin{cases} (\boldsymbol{\sigma},u) \to \big[(\boldsymbol{\tau},v) \mapsto \int_\Omega (\boldsymbol{\sigma} - \mathbf{A}_2\nabla u)\cdot\boldsymbol{\tau} + \mathbf{A}_1\boldsymbol{\sigma}\cdot\nabla v + (\mathbf{b}\cdot\nabla u + \gamma u)v\big] \\ (\boldsymbol{\tau},v) \to \big[(\boldsymbol{\sigma},u) \mapsto \int_\Omega \boldsymbol{\sigma}\cdot\boldsymbol{\tau} - \mathbf{A}_1\boldsymbol{\sigma}\cdot\nabla v + \mathbf{A}_2\nabla u\cdot\boldsymbol{\tau} + (\mathbf{b}\cdot\nabla u + \gamma u)v\big] \end{cases}$$

are boundedly invertible.

Proof. Clearly both mappings are bounded. Given $\mathbf{f} \in L_2(\Omega)^n$ and $g \in H^{-1}(\Omega)$, the problems of finding $\boldsymbol{\sigma} \in L_2(\Omega)^n$ and $u \in H^1_0(\Omega)$ such that

$$\int_\Omega (\boldsymbol{\sigma} - \mathbf{A}_2\nabla u)\cdot\boldsymbol{\tau} + \mathbf{A}_1\boldsymbol{\sigma}\cdot\nabla v + (\mathbf{b}\cdot\nabla u + \gamma u)v = \mathbf{f}(\boldsymbol{\tau}) + g(v) \quad (\boldsymbol{\tau} \in L_2(\Omega)^n,\, v \in H^1_0(\Omega)),$$

or that of finding $\boldsymbol{\tau} \in L_2(\Omega)^n$ and $v \in H^1_0(\Omega)$ such that

$$\int_\Omega \boldsymbol{\sigma}\cdot\boldsymbol{\tau} - \mathbf{A}_1\boldsymbol{\sigma}\cdot\nabla v + \mathbf{A}_2\nabla u\cdot\boldsymbol{\tau} + (\mathbf{b}\cdot\nabla u + \gamma u)v = \mathbf{f}(\boldsymbol{\sigma}) + g(u) \quad (\boldsymbol{\sigma} \in L_2(\Omega)^n,\, u \in H^1_0(\Omega)),$$

are equivalent to solving

$$\begin{aligned} &\boldsymbol{\sigma} - \mathbf{A}_2\nabla u = \mathbf{f}, \quad (Lu)(v) = g(v) - \mathbf{f}(\mathbf{A}_1^\top\nabla v) \quad (v \in H^1_0(\Omega)), \quad \text{or} \\ &\boldsymbol{\tau} - \mathbf{A}_1^\top\nabla v = \mathbf{f}, \quad (L'v)(u) = g(u) - \mathbf{f}(\mathbf{A}_2\nabla u) \quad (u \in H^1_0(\Omega)), \end{aligned}$$

respectively. The boundedness of $L^{-1} : H^{-1}(\Omega) \to H^1_0(\Omega)$ ((3.2)), and thus of $L'^{-1} : H^{-1}(\Omega) \to H^1_0(\Omega)$, shows that both problems have a unique solution with $\|\boldsymbol{\sigma}\|_{L_2(\Omega)^n} + \|u\|_{H^1(\Omega)} \lesssim \|g\|_{H^{-1}(\Omega)} + \|\mathbf{f}\|_{L_2(\Omega)^n}$, or $\|\boldsymbol{\tau}\|_{L_2(\Omega)^n} + \|v\|_{H^1(\Omega)} \lesssim \|g\|_{H^{-1}(\Omega)} + \|\mathbf{f}\|_{L_2(\Omega)^n}$. □

Having verified the conditions (i), uniform in h, and (iii) of Lemma 3.1, the proof of Theorem 3.1 is completed by the following result, being condition (ii) of Lemma 3.1, uniform in h.

THEOREM 3.3 One has $\inf_h \inf_{0\neq(\boldsymbol{\tau},v)\in V} \sup_{0\neq(\boldsymbol{\sigma},u,\theta)\in U} \dfrac{b(\boldsymbol{\sigma},u,\theta,\boldsymbol{\tau},v)}{\|(\boldsymbol{\sigma},u,\theta)\|_U\|(\boldsymbol{\tau},v)\|_V} > 0.$

Proof. Let $0 \neq (\boldsymbol{\tau},v) \in L_2(\Omega)^n \times H^1_{0,\Gamma_+}(\Omega_h)$ be given. We solve $(\boldsymbol{\tau}_1,v_1) \in L_2(\Omega)^n \times H^1_0(\Omega)$ from

$$\begin{cases} \int_\Omega \boldsymbol{\sigma}\cdot\boldsymbol{\tau}_1 - \mathbf{A}_1\boldsymbol{\sigma}\cdot\nabla v_1 = -\int_\Omega \mathbf{A}_1\boldsymbol{\sigma}\cdot\nabla_h v, \\ \int_\Omega \mathbf{A}_2\nabla u\cdot\boldsymbol{\tau}_1 + (\mathbf{b}\cdot\nabla u + \gamma u)v_1 = \int_\Omega (\gamma - \operatorname{div}_h\mathbf{b})vu - u\mathbf{b}\cdot\nabla_h v. \end{cases} \tag{3.20}$$

$(\boldsymbol{\sigma} \in L_2(\Omega)^n,\, u \in H^1_0(\Omega))$. The estimates $|\int_\Omega \mathbf{A}_1\boldsymbol{\sigma}\cdot\nabla_h v| \lesssim \|v\|_{H^1(\Omega_h)}\|\boldsymbol{\sigma}\|_{L_2(\Omega)^n}$ and $|\int_\Omega (\gamma - \operatorname{div}_h\mathbf{b})vu - u\mathbf{b}\cdot\nabla_h v| \lesssim \|v\|_{H^1(\Omega_h)}\|u\|_{H^1(\Omega)}$, i.e., the boundedness of the functionals at the right-hand sides, together with Lemma 3.4 (second mapping) show that there exists a (unique) solution $(\boldsymbol{\tau}_1,v_1)$ with $\|\boldsymbol{\tau}_1\|_{L_2(\Omega)^n} + \|v_1\|_{H^1(\Omega)} \lesssim \|v\|_{H^1(\Omega_h)}$, and so $\|\boldsymbol{\tau}_1\|_{L_2(\Omega)^n} + \|v_1 - v\|_{H^1(\Omega_h)} \lesssim \|v\|_{H^1(\Omega_h)}$.

From

$$\begin{aligned} -\int_\Omega vu\operatorname{div}_h\mathbf{b} &= \sum_{K\in\Omega_h}\int_K \mathbf{b}\cdot(u\nabla v + v\nabla u) - \int_{\partial K} vu\mathbf{b}\cdot\mathbf{n}_k \\ &= \int_\Omega \mathbf{b}\cdot(u\nabla_h v + v\nabla u) - \int_{\partial\Omega_h^\circ} [\![v]\!]u\mathbf{b}\cdot\mathbf{n}, \end{aligned}$$

and $[\![v_1]\!] = 0$, it follows that an equivalent formulation of (3.20) is

$$\begin{cases} \displaystyle\int_\Omega \mathbf{A}_1\boldsymbol{\sigma}\cdot\nabla_h(v-v_1) - \boldsymbol{\sigma}\cdot\boldsymbol{\tau}_1 = 0 \quad (\boldsymbol{\sigma}\in L_2(\Omega)^n), \\ \displaystyle\int_\Omega \mathbf{A}_2\nabla u\cdot\boldsymbol{\tau}_1 + (v_1-v)(\mathbf{b}\cdot\nabla u+\gamma u) + \int_{\partial\Omega_h^\circ}[\![v-v_1]\!]u\mathbf{b}\cdot\mathbf{n} = 0 \quad (u\in H_0^1(\Omega)), \end{cases}$$

so that $v_1 = v$ and $\boldsymbol{\tau}_1 = 0$ when $v\in H_0^1(\Omega)$. By Theorem 3.2, we conclude that

$$\begin{aligned} \|\boldsymbol{\tau}_1\|_{L_2(\Omega)^n} + \|v_1-v\|_{H^1(\Omega_h)} &\lesssim \inf_{w\in H_0^1(\Omega)} \|v-w\|_{H^1(\Omega_h)} \eqsim \|[\![v]\!]\|_{(H^{-\frac12}(\partial\Omega_h^\circ))'}, \quad \text{and so} \\ \|\boldsymbol{\tau}\|_{L_2(\Omega)^n} + \|v\|_{H^1(\Omega_h)} &\lesssim \|\boldsymbol{\tau}-\boldsymbol{\tau}_1\|_{L_2(\Omega)^n} + \|v_1\|_{H^1(\Omega)} + \|[\![v]\!]\|_{(H^{-\frac12}(\partial\Omega_h^\circ))'}. \end{aligned} \tag{3.21}$$

Substituting the equations from (3.20) in the definition of b shows that for all $(\boldsymbol{\sigma},u,\theta)\in U$,

$$b(\boldsymbol{\sigma},u,\theta,\boldsymbol{\tau},v) = \int_\Omega(\boldsymbol{\sigma}-\mathbf{A}_2\nabla u)\cdot(\boldsymbol{\tau}-\boldsymbol{\tau}_1) + \mathbf{A}_1\boldsymbol{\sigma}\cdot\nabla v_1 + (\mathbf{b}\cdot\nabla u+\gamma u)v_1 + \int_{\partial\Omega_h^\circ}[\![v]\!]\theta.$$

From Lemma 3.4 (first mapping) and Lemma 3.1, it follows that there exists an absolute constant $\rho>0$, and $(\boldsymbol{\sigma},u)\in L_2(\Omega)^n\times H_0^1(\Omega)$ with $\|\boldsymbol{\sigma}\|^2_{L_2(\Omega)^n} + \|u\|^2_{H^1(\Omega)} = \|\boldsymbol{\tau}-\boldsymbol{\tau}_1\|^2_{L_2(\Omega)^n} + \|v_1\|^2_{H^1(\Omega)}$ and

$$\rho[\|\boldsymbol{\tau}-\boldsymbol{\tau}_1\|^2_{L_2(\Omega)^n} + \|v_1\|^2_{H^1(\Omega)}] \leqslant \int_\Omega(\boldsymbol{\sigma}-\mathbf{A}_2\nabla u)\cdot(\boldsymbol{\tau}-\boldsymbol{\tau}_1) + \mathbf{A}_1\boldsymbol{\sigma}\cdot\nabla v_1 + (\mathbf{b}\cdot\nabla u+\gamma u)v_1.$$

By definition of a dual space, there exists a $\theta\in H^{-\frac12}(\partial\Omega_h^\circ)$ with $\|\theta\|_{H^{-\frac12}(\partial\Omega_h^\circ)} = \|[\![v]\!]\|_{(H^{-\frac12}(\partial\Omega_h^\circ))'}$ and $\|[\![v]\!]\|^2_{(H^{-\frac12}(\partial\Omega_h^\circ))'} = \int_{\partial\Omega_h^\circ}[\![v]\!]\theta$. We conclude that

$$\begin{aligned} &\min(\rho,1)\sqrt{\|\boldsymbol{\tau}-\boldsymbol{\tau}_1\|^2_{L_2(\Omega)^n} + \|v_1\|^2_{H^1(\Omega)} + \|[\![v]\!]\|^2_{(H^{-\frac12}(\partial\Omega_h^\circ))'}} \\ &\qquad\leqslant \frac{b(\boldsymbol{\sigma},u,\theta,\boldsymbol{\tau},v)}{\sqrt{\|\boldsymbol{\sigma}\|^2_{L_2(\Omega)^n} + \|u\|^2_{H^1(\Omega)} + \|\theta\|^2_{H^{-\frac12}(\partial\Omega_h^\circ)}}}, \end{aligned}$$

which, together with (3.21), completes the proof. □

3.4 *Error estimates*

Given $f\in H^1_{0,\Gamma_+}(\Omega_h)'$, let $(\boldsymbol{\sigma},u,\theta)\in U = L_2(\Omega)^n\times H_0^1(\Omega)\times H^{-\frac12}(\partial\Omega_h^\circ)$ denote the exact solution of the mild-weak formulation (3.7), and for some closed trial space $U_h\subset U$, let $(\boldsymbol{\sigma}_h,u_h,\theta_h)\in U_h$ denote its Petrov-Galerkin approximation with optimal test space V_h.

We will use the notation

$$W_p^s(\operatorname{div};\Upsilon) = \{\mathbf{v}\in W_p^s(\Upsilon)^n : \operatorname{div}\mathbf{v}\in W_p^s(\Upsilon)\},$$

equipped with norm $\|\mathbf{v}\|_{W_p^s(\operatorname{div};\Upsilon)} = \big(\|\mathbf{v}\|^p_{W_p^s(\Upsilon)^n} + \|\operatorname{div}\mathbf{v}\|^p_{W_p^s(\Upsilon)}\big)^{\frac1p}$, with the usual adaptation for $p=\infty$. The space $W_2^s(\operatorname{div};\Upsilon)$ will be denoted as $H^s(\operatorname{div};\Upsilon)$, so that in particular $H^0(\operatorname{div};\Upsilon) = H(\operatorname{div};\Upsilon)$.

In the following we think of $\Sigma_h\subset L_2(\Omega)^n$ and $P_h\subset H_0^1(\Omega)$ being common finite element spaces w.r.t. Ω_h of order k and $k+1$, respectively, and $\mathbf{Q}_h\subset H(\operatorname{div};\Omega)$, typically being a Raviart-Thomas space w.r.t. Ω_h of order k. We assume that Ω_h is quasi-uniform, shape-regular, with mesh-size h.

PROPOSITION 3.4 For a $k \in \mathbb{N} = \{1,2,\dots\}$, let $\gamma \in W^k_\infty(\Omega)$, $\mathbf{b} \in W^k_\infty(\mathrm{div};\Omega)$, and $\mathbf{A}_1, \mathbf{A}_2 \in W^k_\infty(\Omega)^{n\times n}$. Let $f \in H^k(\Omega)$ and $u \in H^{k+1}(\Omega)$. Let $\Sigma_h \subset L_2(\Omega)^n$, $P_h \subset H^1_0(\Omega)$, and $\mathbf{Q}_h \subset H(\mathrm{div};\Omega)$ with

$$\inf_{\hat{\boldsymbol{\sigma}}_h \in \Sigma_h} \|\hat{\boldsymbol{\sigma}} - \hat{\boldsymbol{\sigma}}_h\|_{L_2(\Omega)^n} \lesssim h^k \|\hat{\boldsymbol{\sigma}}\|_{H^k(\Omega)^n} \quad (\hat{\boldsymbol{\sigma}} \in H^k(\Omega)^n), \tag{3.22}$$

$$\inf_{\hat{u}_h \in P_h} \|\hat{u} - \hat{u}_h\|_{H^1(\Omega)} \lesssim h^k \|\hat{u}\|_{H^{k+1}(\Omega)} \quad (\hat{u} \in H^{k+1}(\Omega) \cap H^1_0(\Omega)), \tag{3.23}$$

$$\inf_{\mathbf{q}_h \in \mathbf{Q}_h} \|\mathbf{q} - \mathbf{q}_h\|_{H(\mathrm{div};\Omega)} \lesssim h^k \|\mathbf{q}\|_{H^k(\mathrm{div};\Omega)} \quad (\mathbf{q} \in H^k(\mathrm{div};\Omega)). \tag{3.24}$$

Then with $U_h = \Sigma_h \times P_h \times \{\mathbf{q}_h|_{\partial\Omega_h^\circ} \cdot \mathbf{n} : \mathbf{q}_h \in \mathbf{Q}_h\}$, the Petrov-Galerkin approximation with optimal test space V_h satisfies

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{L_2(\Omega)^n} + \|u - u_h\|_{H^1(\Omega)} + \|\theta - \theta_h\|_{H^{-\frac{1}{2}}(\partial\Omega_h^\circ)} \lesssim h^k \big[\|u\|_{H^{k+1}(\Omega)} + \|f\|_{H^k(\Omega)}\big].$$

Proof. As shown in Proposition 2.1, the Petrov-Galerkin solution $(\boldsymbol{\sigma}_h, u_h, \theta_h)$ with optimal test space V_h is the best approximation from U_h to $(\boldsymbol{\sigma}, u, \theta)$ in energy norm $\|\cdot\|_E = \|B\cdot\|_{V'}$. Since, as shown in Theorem 3.1, $B : U \to V'$ is boundedly invertible, uniformly in h, we infer that, up to some constant multiple, this Petrov-Galerkin solution realizes the smallest error from U_h w.r.t $\|\cdot\|_U$.

We have $\boldsymbol{\sigma} = \mathbf{A}_2 \nabla u$, and, from $f \in L_2(\Omega)$, $\theta = (u\mathbf{b} - \mathbf{A}_1\boldsymbol{\sigma})|_{\partial\Omega_h^\circ} \cdot \mathbf{n}$ and $-\operatorname{div} \mathbf{A}_1\boldsymbol{\sigma} + \mathbf{b}\cdot\nabla u + \gamma u = f$ (cf. Remark 3.4).

From $\mathbf{A}_1, \mathbf{A}_2 \in W^k_\infty(\Omega)^{n\times n}$, $\mathbf{b} \in W^k_\infty(\Omega)^n$, we have $\|\boldsymbol{\sigma}\|_{H^k(\Omega)^n} \lesssim \|u\|_{H^{k+1}(\Omega)}$ and $\|u\mathbf{b} - \mathbf{A}_1\boldsymbol{\sigma}\|_{H^k(\Omega)^n} \lesssim \|u\|_{H^{k+1}(\Omega)}$. By $\operatorname{div}(u\mathbf{b} - \mathbf{A}_1\boldsymbol{\sigma}) = \mathbf{b}\cdot\nabla u + u\operatorname{div}\mathbf{b} + f - \mathbf{b}\cdot\nabla u - \gamma u = u\operatorname{div}\mathbf{b} + f - \gamma u$, and $\operatorname{div}\mathbf{b}, \gamma \in W^k_\infty(\Omega)$, we have $\|\operatorname{div}(u\mathbf{b} - \mathbf{A}_1\boldsymbol{\sigma})\|_{H^k(\Omega)} \lesssim \|u\|_{H^k(\Omega)} + \|f\|_{H^k(\Omega)}$. The proof is completed by the approximation properties of the spaces Σ_h, P_h, and $\mathbf{Q}_h$. □

Apart from the additional smoothness condition on f, that is usually harmless, note that the regularity condition on u is the mildest one that can be expected in view of the convergence order that is realized.

Although, to establish well-posedness of the continuous variational problem (3.7), it was needed to treat $\theta = (u\mathbf{b} - \mathbf{A}_1\boldsymbol{\sigma})|_{\partial\Omega_h^\circ} \cdot \mathbf{n}$ as an independent variable, it is *not* needed to do so with its Petrov-Galerkin discretization:

PROPOSITION 3.5 For a $k \in \mathbb{N} = \{1,2,\dots\}$, let $\gamma \in W^k_\infty(\Omega)$, $\mathbf{b} \in W^k_\infty(\mathrm{div};\Omega)$, $\mathbf{A}_1^{-1} \in L_\infty(\Omega)^{n\times n}$, and $\mathbf{A} \in W^k_\infty(\Omega)^{n\times n}$. Let $f \in H^k(\Omega)$ and $u \in H^{k+1}(\Omega)$. Let $P_h \subset H^1_0(\Omega)$ and $\mathbf{Q}_h \subset H(\mathrm{div};\Omega)$ satisfy (3.23) and (3.24), respectively.

Then the Petrov-Galerkin approximation from

$$U_h := \{(\mathbf{A}_1^{-1}\hat{\boldsymbol{\sigma}}_h, \hat{u}_h, (\hat{u}_h\mathbf{b} - \hat{\boldsymbol{\sigma}}_h)|_{\partial\Omega_h^\circ} \cdot \mathbf{n}) : (\hat{\boldsymbol{\sigma}}_h, \hat{u}_h) \in \mathbf{Q}_h \times P_h\},$$

with optimal test space V_h, satisfies

$$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{L_2(\Omega)^n} + \|u - u_h\|_{H^1(\Omega)} + \|\theta - \theta_h\|_{H^{-\frac{1}{2}}(\partial\Omega_h^\circ)} \lesssim h^k \big[\|u\|_{H^{k+1}(\Omega)} + \|f\|_{H^k(\Omega)}\big].$$

Proof. The proof follows from $\theta = (u\mathbf{b} - \mathbf{A}_1\boldsymbol{\sigma})|_{\partial\Omega_h^\circ}\cdot\mathbf{n}$, and so

$$\begin{aligned}
&\inf_{(\hat{\boldsymbol{\sigma}}_h,\hat{u}_h,\hat{\theta}_h)\in U_h} \|\boldsymbol{\sigma}-\mathbf{A}_1^{-1}\hat{\boldsymbol{\sigma}}_h\|_{L_2(\Omega)^n} + \|u-\hat{u}_h\|_{H^1(\Omega)} + \|\theta-\hat{\theta}_h\|_{H^{-\frac{1}{2}}(\partial\Omega_h^\circ)}\\
&\quad\leqslant \inf_{(\hat{\boldsymbol{\sigma}}_h,\hat{u}_h)\in\mathbf{Q}_h\times P_h} \|\mathbf{A}_1^{-1}\|_{L_\infty(\Omega)^{n\times}}\|\mathbf{A}_1\boldsymbol{\sigma}-\hat{\boldsymbol{\sigma}}_h\|_{L_2(\Omega)^n} + \|u-\hat{u}_h\|_{H^1(\Omega)}\\
&\qquad\qquad + \|(u-\hat{u}_h)\mathbf{b}\|_{H(\mathrm{div};\Omega)} + \|\mathbf{A}_1\boldsymbol{\sigma}-\hat{\boldsymbol{\sigma}}_h\|_{H(\mathrm{div};\Omega)}\\
&\quad\lesssim h^k\big[\|u\|_{H^{k+1}(\Omega)} + \|f\|_{H^k(\Omega)}\big],
\end{aligned}$$

similarly to the proof of Proposition 3.4, by using $\mathbf{A}_1\boldsymbol{\sigma} = \mathbf{A}\nabla u$ and $\mathrm{div}\,\mathbf{A}_1\boldsymbol{\sigma} = \mathbf{b}\cdot\nabla u + \gamma u - f$. □

Note that a price that has to be paid for the avoidance of an independent discrete flux variable is that the space $\mathbf{Q}_h$ for $\boldsymbol{\sigma}_h$ used in Proposition 3.5 is richer than that is needed for the approximation of $\boldsymbol{\sigma}$ itself.

REMARK 3.10 If, instead of $\mathbf{A}_1^{-1}\in L_\infty(\Omega)^{n\times n}$ and $\mathbf{A}\in W_\infty^k(\Omega)^{n\times n}$, we impose that $\mathbf{A}_1,\mathbf{A}_2\in W_\infty^k(\Omega)^{n\times n}$ and $\mathbf{A}_1$ be scalar valued, so that multiplication with $\mathbf{A}_1$ maps $H(\mathrm{div};\Omega)$ into itself, then the statement of Proposition 3.5 is also valid with U_h reading as $\{(\hat{\boldsymbol{\sigma}}_h,\hat{u}_h,(\hat{u}_h\mathbf{b}-\mathbf{A}_1\hat{\boldsymbol{\sigma}}_h)|_{\partial\Omega_h^\circ}\cdot\mathbf{n}) : (\hat{\boldsymbol{\sigma}}_h,\hat{u}_h)\in\mathbf{Q}_h\times P_h\}$.

4. Higher order rates in a weaker norm

For the optimal Petrov-Galerkin discretization of the mild-weak formulation (3.7), we demonstrate higher order rates in a weaker norm by applying a *duality argument*. The common ingredients are regularity of the adjoint equation (Lemma 4.1), for which we will need $H^2(\Omega)$-regularity of the adjoint of the original boundary value problem (3.1), and an approximation property of the optimal test space $V_h = \Im(T|_{U_h})$. For the latter we will use $H^2(\Omega)$-regularity of the primal version of the original boundary value problem (Lemma 4.2).

LEMMA 4.1 Let $\mathbf{A}_1,\mathbf{A}_2\in W_\infty^1(\Omega)^{n\times n}$ and $(L')^{-1}: L_2(\Omega)\to H^2(\Omega)\cap H_0^1(\Omega)$ be bounded. Then for $\boldsymbol{\psi}\in H^1(\Omega)^n$ and $\phi\in L_2(\Omega)$, the solution $(\boldsymbol{\tau},v)\in V = L_2(\Omega)^n\times H^1_{0,\Gamma_+}(\Omega_h)$ of

$$b(\boldsymbol{\sigma},u,\theta,\boldsymbol{\tau},v) = \int_\Omega \boldsymbol{\psi}\cdot\boldsymbol{\sigma} + \phi u \quad ((\boldsymbol{\sigma},u,\theta)\in U = L_2(\Omega)^n\times H_0^1(\Omega)\times H^{-\frac{1}{2}}(\partial\Omega_h^\circ))$$

is in $H^1(\Omega)^n\times(H^2(\Omega)\cap H_0^1(\Omega))$, and

$$\|\boldsymbol{\tau}\|_{H^1(\Omega)^n} + \|v\|_{H^2(\Omega)} \lesssim \|\boldsymbol{\psi}\|_{H^1(\Omega)^n} + \|\phi\|_{L_2(\Omega)}.$$

Proof. The system of equations reads as

$$\begin{cases}
\int_\Omega -\mathbf{A}_2\nabla u\cdot\boldsymbol{\tau} - u\mathbf{b}\cdot\nabla_h v + (\gamma-\mathrm{div}_h\,\mathbf{b})uv = \int_\Omega \phi u & (u\in H_0^1(\Omega)),\\
\int_\Omega \boldsymbol{\sigma}\cdot\boldsymbol{\tau} + \mathbf{A}_1\boldsymbol{\sigma}\cdot\nabla_h v = \int_\Omega \boldsymbol{\psi}\cdot\boldsymbol{\sigma} & (\boldsymbol{\sigma}\in L_2(\Omega)^n),\\
\int_{\partial\Omega_h^\circ}[\![v]\!]\theta = 0 & (\theta\in H^{-\frac{1}{2}}(\partial\Omega_h^\circ)),
\end{cases}$$

and so $\boldsymbol{\tau} = \boldsymbol{\psi} - \mathbf{A}_1^\top\nabla_h v$. By substituting this, and by using that $\int_\Omega -uv\,\mathrm{div}_h\,\mathbf{b} = \int_\Omega \mathbf{b}\cdot(v\nabla u + u\nabla_h v) - \int_{\partial\Omega_h^\circ}[\![v]\!]u\mathbf{b}\cdot\mathbf{n}$, we obtain the equivalent system

$$\begin{cases}
\int_\Omega \mathbf{A}\nabla u\cdot\nabla_h v + v\mathbf{b}\cdot\nabla u + \gamma uv = \int_\Omega \phi u + \mathbf{A}_2\nabla u\cdot\boldsymbol{\psi} & (u\in H_0^1(\Omega)),\\
\int_{\partial\Omega_h^\circ}[\![v]\!]\theta = 0 & (\theta\in H^{-\frac{1}{2}}(\partial\Omega_h^\circ)).
\end{cases}$$

By our assumption on $\mathbf{A}_2$, $\int_\Omega \mathbf{A}_2\nabla u\cdot\boldsymbol{\psi} = -\int_\Omega u\operatorname{div}\mathbf{A}_2^\top\boldsymbol{\psi}$, and $\|\phi-\operatorname{div}\mathbf{A}_2^\top\boldsymbol{\psi}\|_{L_2(\Omega)}\lesssim\|\phi\|_{L_2(\Omega)}+\|\boldsymbol{\psi}\|_{H^1(\Omega)}$. We conclude that the solution of the last system is given by $v=(L')^{-1}(\phi-\operatorname{div}\mathbf{A}_2^\top\boldsymbol{\psi})$ which completes the proof. □

Knowing that our mapping $B:U\to V'$ is boundedly invertible, the corresponding mapping $T:U\to V$ is boundedly invertible. In the next lemma, we will show that T is *regular*, in the sense that T^{-1} maps, boundedly, a subspace of smooth functions of V into a subspace of smooth functions of U.

LEMMA 4.2 Let $\mathbf{A}_1,\mathbf{A}_2\in W^1_\infty(\Omega)^{n\times n}$, $\mathbf{b}\in W^1_\infty(\operatorname{div};\Omega)$, and $\gamma\in W^1_\infty(\Omega)$, and let $L^{-1}:L_2(\Omega)\to H^2(\Omega)\cap H^1_0(\Omega)$ be bounded. Then for $(\boldsymbol{\tau},v)\in H^1(\Omega)^n\times H^2(\Omega)$, the solution $(\sigma,u,\theta)\in U=L_2(\Omega)^n\times H^1_0(\Omega)\times H^{-\frac12}(\partial\Omega_h^\circ)$ of

$$b(\sigma,u,\theta,\hat{\boldsymbol{\tau}},\hat v)=\langle\boldsymbol{\tau},\hat{\boldsymbol{\tau}}\rangle_{L_2(\Omega)^n}+\langle v,\hat v\rangle_{H^1(\Omega_h)}\quad((\hat{\boldsymbol{\tau}},\hat v)\in L_2(\Omega)^n\times H^1_{0,\Gamma_+}(\Omega_h)) \tag{4.1}$$

satisfies $\sigma\in H^1(\Omega)^n$, $u\in H^2(\Omega)\cap H^1_0(\Omega)$, and $\theta=\mathbf{q}|_{\partial\Omega_h^\circ}\cdot\mathbf{n}$ for some $\mathbf{q}\in H^1(\operatorname{div};\Omega)$, with

$$\|\sigma\|_{H^1(\Omega)^n}+\|u\|_{H^2(\Omega)}+\|\mathbf{q}\|_{H^1(\operatorname{div};\Omega)}\lesssim\|\boldsymbol{\tau}\|_{H^1(\Omega)^n}+\|v\|_{H^2(\Omega)}. \tag{4.2}$$

REMARK 4.1 If, additionally, $v=0$ on Γ_+, then $v\in H^1_{0,\Gamma_+}(\Omega_h)$, and so $(\sigma,u,\theta)=T^{-1}(\boldsymbol{\tau},v)$.

Proof. Let u solve $Lu=\operatorname{div}\mathbf{A}_1\boldsymbol{\tau}+v-\Delta v$, $\sigma:=\mathbf{A}_2\nabla u+\boldsymbol{\tau}$, and $\mathbf{q}:=u\mathbf{b}-\mathbf{A}\nabla u-\mathbf{A}_1\boldsymbol{\tau}+\nabla v$. Then from the assumptions, we have $u\in H^2(\Omega)\cap H^1_0(\Omega)$, $\sigma\in H^1(\Omega)^n$, $\mathbf{q}\in H^1(\Omega)^n$, and $\operatorname{div}\mathbf{q}=u\operatorname{div}\mathbf{b}+\mathbf{b}\cdot\nabla u-\operatorname{div}\mathbf{A}\nabla u-\operatorname{div}\mathbf{A}_1\boldsymbol{\tau}+\Delta v=u\operatorname{div}\mathbf{b}+v-\gamma u\in H^1(\Omega)$, and (4.2) is valid.

The definitions of u and $\theta=\mathbf{q}|_{\partial\Omega_h^\circ}\cdot\mathbf{n}$ show that

$$\begin{aligned}\int_\Omega(-\operatorname{div}\mathbf{A}\nabla u+\mathbf{b}\cdot\nabla u+\gamma u)\hat v+\int_{\partial\Omega_h^\circ}[\![\hat v]\!](\theta+(\mathbf{A}\nabla u+\mathbf{A}_1\boldsymbol{\tau}-u\mathbf{b}-\nabla v)\cdot\mathbf{n})\\ =\int_\Omega(\operatorname{div}\mathbf{A}_1\boldsymbol{\tau}+v-\Delta v)\hat v\quad(\hat v\in H^1_{0,\Gamma_+}(\Omega_h)),\end{aligned}$$

or

$$\begin{aligned}\int_\Omega-\operatorname{div}(\mathbf{A}\nabla u+\mathbf{A}_1\boldsymbol{\tau}-u\mathbf{b})\hat v+(\gamma-\operatorname{div}_h\mathbf{b})\hat v+\int_{\partial\Omega_h^\circ}[\![\hat v]\!](\theta+(\mathbf{A}\nabla u+\mathbf{A}_1\boldsymbol{\tau}-u\mathbf{b})\cdot\mathbf{n})\\ =\int_\Omega(v-\Delta v)\hat v+\int_{\partial\Omega_h^\circ}\frac{\partial v}{\partial\mathbf{n}}[\![\hat v]\!]\quad(\hat v\in H^1_{0,\Gamma_+}(\Omega_h)).\end{aligned}$$

By applying integration by parts at both sides, we arrive at

$$\int_\Omega(\mathbf{A}\nabla u+\mathbf{A}_1\boldsymbol{\tau}-u\mathbf{b})\cdot\nabla_h\hat v+(\gamma-\operatorname{div}_h\mathbf{b})\hat v+\int_{\partial\Omega_h^\circ}[\![\hat v]\!]\theta=\langle v,\hat v\rangle_{H^1(\Omega_h)}\ (\hat v\in H^1_{0,\Gamma_+}(\Omega_h)),$$

or, by definition of σ,

$$\int_\Omega(\sigma-\mathbf{A}_2\nabla u)\cdot\hat{\boldsymbol{\tau}}+(\mathbf{A}_1\sigma-u\mathbf{b})\cdot\nabla_h\hat v+(\gamma-\operatorname{div}_h\mathbf{b})\hat v+\int_{\partial\Omega_h^\circ}[\![\hat v]\!]\theta=\langle v,\hat v\rangle_{H^1(\Omega_h)}+\langle\boldsymbol{\tau},\hat{\boldsymbol{\tau}}\rangle_{L_2(\Omega)^n}$$

$((\hat{\boldsymbol{\tau}},\hat v)\in L_2(\Omega)^n\times H^1_{0,\Gamma_+}(\Omega_h))$, which is (4.1). □

As a corollary, we will see that approximation properties of U_h give rise to approximation properties of the optimal test space V_h.

COROLLARY 4.1 Let $U_h \subset U$ be such that for $\sigma \in H^1(\Omega)^n$, $u \in H^2(\Omega) \cap H_0^1(\Omega)$, $\theta = \mathbf{q}|_{\partial\Omega_h^\circ} \cdot \mathbf{n}$ with $\mathbf{q} \in H^1(\mathrm{div};\Omega)$,

$$\inf_{(\sigma_h,u_h,\theta_h)\in U_h} \|\sigma-\sigma_h\|_{L_2(\Omega)^n} + \|u-u_h\|_{H^1(\Omega)} + \|\theta-\theta_h\|_{H^{-\frac{1}{2}}(\partial\Omega_h^\circ)} \lesssim h[\|\sigma\|_{H^1(\Omega)^n} + \|u\|_{H^2(\Omega)} + \|\mathbf{q}\|_{H^1(\mathrm{div};\Omega)}].$$

Then, under the conditions of Lemma 4.2, for $(\boldsymbol{\tau}, v) \in H^1(\Omega)^n \times (H^2(\Omega) \cap H^1_{0,\Gamma_+}(\Omega_h))$,

$$\inf_{(\boldsymbol{\tau}_h,v_h)\in V_h} \|\boldsymbol{\tau}-\boldsymbol{\tau}_h\|_{L_2(\Omega)^n} + \|v-v_h\|_{H^1(\Omega_h)} \lesssim h[\|\boldsymbol{\tau}\|_{H^1(\Omega)^n} + \|v\|_{H^2(\Omega)}].$$

Proof. Let $(\sigma, u, \theta) = T^{-1}(\boldsymbol{\tau}, v)$, and $\theta = \mathbf{q}|_{\partial\Omega_h^\circ} \cdot \mathbf{n}$ as in Lemma 4.2. Then, from $V_h = \Im(T|_{U_h})$,

$$\inf_{(\boldsymbol{\tau}_h,v_h)\in V_h} \|(\boldsymbol{\tau}, v) - (\boldsymbol{\tau}_h, v_h)\|_V \eqsim \inf_{(\sigma_h,u_h,\theta_h)\in U_h} \|(\sigma, u, \theta) - (\sigma_h, u_h, \theta_h)\|_U$$
$$\lesssim h[\|\sigma\|_{H^1(\Omega)^n} + \|u\|_{H^2(\Omega)} + \|\mathbf{q}\|_{H^1(\mathrm{div};\Omega)}] \lesssim h[\|\boldsymbol{\tau}\|_{H^1(\Omega)^n} + \|v\|_{H^2(\Omega)}],$$

by Lemma 4.2. □

Having established regularity of the adjoint equation and approximation properties of the optimal test space V_h, we are ready to derive improved error estimates in weaker norms by applying a duality argument.

THEOREM 4.1 Let $\mathbf{A}_1, \mathbf{A}_2 \in W^1_\infty(\Omega)^{n\times n}$, $\mathbf{b} \in W^1_\infty(\mathrm{div};\Omega)$, and $\gamma \in W^1_\infty(\Omega)$. Let $L^{-1}, (L')^{-1} : L_2(\Omega) \to H^2(\Omega) \cap H_0^1(\Omega)$ be bounded, and let $U_h \subset U$ be such that for $\sigma \in H^1(\Omega)^n$, $u \in H^2(\Omega) \cap H_0^1(\Omega)$, $\theta = \mathbf{q}|_{\partial\Omega_h^\circ} \cdot \mathbf{n}$ with $\mathbf{q} \in H^1(\mathrm{div};\Omega)$,

$$\inf_{(\sigma_h,u_h,\theta_h)\in U_h} \|\sigma-\sigma_h\|_{L_2(\Omega)^n} + \|u-u_h\|_{H^1(\Omega)} + \|\theta-\theta_h\|_{H^{-\frac{1}{2}}(\partial\Omega_h^\circ)} \lesssim h[\|\sigma\|_{H^1(\Omega)^n} + \|u\|_{H^2(\Omega)} + \|\mathbf{q}\|_{H^1(\mathrm{div};\Omega)}]. \tag{4.3}$$

Then for $(\boldsymbol{\sigma}, u, \theta) \in L_2(\Omega)^n \times H_0^1(\Omega) \times H^{-\frac{1}{2}}(\partial\Omega_h^\circ)$ being the exact solution of (3.7), and $(\boldsymbol{\sigma}_h, u_h, \theta_h) \in U_h$ its Petrov-Galerkin approximation with optimal test space V_h, we have

$$\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{(H^1(\Omega)^n)'} + \|u-u_h\|_{L_2(\Omega)} \lesssim h[\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{L_2(\Omega)^n} + \|u-u_h\|_{H^1(\Omega)} + \|\theta-\theta_h\|_{H^{-\frac{1}{2}}(\partial\Omega_h^\circ)}]$$

Proof. Given $\boldsymbol{\psi} \in H^1(\Omega)^n$ and $\phi \in L_2(\Omega)$, let $(\boldsymbol{\tau}, v) \in V = L_2(\Omega)^n \times H^1_{0,\Gamma_+}(\Omega_h)$ denote the solution of

$$b(\hat{\boldsymbol{\sigma}}, \hat{u}, \hat{\theta}, \boldsymbol{\tau}, v) = \int_\Omega \boldsymbol{\psi}\cdot\hat{\boldsymbol{\sigma}} + \phi\hat{u} \quad ((\hat{\boldsymbol{\sigma}}, \hat{u}, \hat{\theta}) \in U = L_2(\Omega)^n \times H_0^1(\Omega) \times H^{-\frac{1}{2}}(\partial\Omega_h^\circ)). \tag{4.4}$$

Then

$$\begin{aligned}
&\int_\Omega \boldsymbol{\psi}\cdot(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h) + \phi(u-u_h) = b(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h, u-u_h, \theta-\theta_h, \boldsymbol{\tau}, v)\\
&= \inf_{(\boldsymbol{\tau}_h,v_h)\in V_h} b(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h, u-u_h, \theta-\theta_h, \boldsymbol{\tau}-\boldsymbol{\tau}_h, v-v_h)\\
&\lesssim [\|\boldsymbol{\sigma}-\boldsymbol{\sigma}_h\|_{L_2(\Omega)^n} + \|u-u_h\|_{H^1(\Omega)} + \|\theta-\theta_h\|_{H^{-\frac{1}{2}}(\partial\Omega_h^\circ)}] \times h[\|\boldsymbol{\psi}\|_{H^1(\Omega)^n} + \|\phi\|_{L_2(\Omega)}],
\end{aligned}$$

by applications of Corollary 4.1 and Lemma 4.1, from which the result follows. □

The assumption (4.3) corresponds to the approximation assumptions on U_h from Proposition 3.4 for $k = 1$.

REMARK 4.2 If (4.3) would be needed only for $\theta = (u\mathbf{b} - \mathbf{A}_1\sigma)|_{\partial\Omega_h^\circ} \cdot \mathbf{n}$, then, alternatively, the space U_h from Proposition 3.5 or Remark 3.10 could be applied. Considering the proof of Corollary 4.1, however, for $(\boldsymbol{\tau}, v) \in H^1(\Omega)^n \times H^2(\Omega)$ and $(\sigma, u, \theta) = T^{-1}(\boldsymbol{\tau}, v)$, the case $\theta = (u\mathbf{b} - \mathbf{A}_1\sigma)|_{\partial\Omega_h^\circ} \cdot \mathbf{n}$ corresponds to $(\nabla v)|_{\partial\Omega_h^\circ} \cdot \mathbf{n} = 0$, which can only be expected when $v = 0$. In the proof of Theorem 4.1, only for $\phi = \operatorname{div} \mathbf{A}_2^\top \boldsymbol{\psi}$, the v-component of the solution of (4.4) is zero. With this restriction on ϕ, this proof does not yield useful estimates for $\boldsymbol{\sigma} - \boldsymbol{\sigma}_h$ or $u - u_h$. We conclude that for the space U_h from Proposition 3.5, i.e., without an independent discrete flux variable, we have not established improved error estimates in norms weaker than that on U.

5. Convection dominated convection-diffusion problem

In the remainder of this paper, we consider (3.1) with $\mathbf{A} = \mathbf{A}(\varepsilon) = \varepsilon \operatorname{Id}$ for $\varepsilon > 0$, and $\gamma = 0$, i.e.,

$$\begin{cases} -\varepsilon \Delta u + \mathbf{b} \cdot \nabla u = f & \text{on } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{5.1}$$

5.1 *Numerical test in one dimension*

We tested the optimal Petrov-Galerkin discretizations of the three variational formulations of the mixed system, –i.e., mild, mild-weak and ultra-weak,– as well as the Galerkin discretization of the standard variational formulation of the non-mixed system, all for the one-dimensional equation

$$\begin{cases} -\varepsilon u'' + u' = f & \text{on } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

where $f(x) = x$, and $\Omega_h = \{((i-1)h, ih) : 1 \leqslant i \leqslant h^{-1} =: n \in \mathbb{N}\}$. One directly verifies that $u(x) = \frac{1}{2}x^2 + \varepsilon x + (\frac{1}{2} + \varepsilon)(e^{x/\varepsilon} - 1)/(1 - e^{1/\varepsilon})$, which has a "layer" at the outflow boundary $x = 1$.

For the mild-weak formulation, we took $\Gamma_+ = \emptyset$, and so $\partial\Omega_h^\circ = \{ih : 0 \leqslant i \leqslant n\}$, $\mathbf{A}_2(\varepsilon) = \mathbf{A}(\varepsilon) = \varepsilon$, and thus $\mathbf{A}_1(\varepsilon) = 1$, discontinuous piecewise linears for $\boldsymbol{\sigma}_h$, and continuous piecewise linears, zero at $\{0, 1\}$, for u_h, and $\theta_h \in \{g : \partial\Omega_h^\circ \to \mathbb{R}\} \simeq \mathbb{R}^{n+1}$. We replaced the standard squared norm $\sum_{i=1}^n \|v\|^2_{H^1((i-1)h,ih)}$ on $H^1(\Omega_h)$ by the uniformly equivalent one $\sum_{i=1}^n |v|^2_{H^1((i-1)h,ih)} + h|v(ih^-)|^2$, with which, in this one-dimensional setting, the optimal test functions could be determined analytically from (3.11).

For the ultra-weak formulation, we used discontinuous piecewise linears for $\boldsymbol{\sigma}_h$ and u_h, $\theta_h \in \mathbb{R}^{n+1}$, and $\rho_h \in \mathbb{R}^{n-1}$. For determining the optimal test functions, we used the norms on the broken $H(\operatorname{div}; \Omega_h)$ and $H^1(\Omega_h)$ spaces as in (Demkowicz & Gopalakrishnan, 2011b, (4.8)) that allow to find the optimal test functions analytically.

Finally, for the Galerkin method, we took continuous piecewise linear trial functions.

In Figure 1, the $L_2(0,1)$-errors in u_h vs. $1/h$ are given for $\varepsilon = 10^{-4}$. The curves for the mild and mild-weak Petrov-Galerkin solutions are indistinguishable, although the solutions are actually not equal. For small h, the ultra-weak Petrov-Galerkin solution suffers from instabilities caused by very ill-conditioned linear systems.

Exact and approximate solutions u and u_h for $h = \frac{1}{16}$ and $\varepsilon = 10^{-4}$ are shown in Figure 2. The left picture in Figure 2 confirms the familiar fact that for relatively large mesh-sizes, the Galerkin approximations are unstable. Since the Petrov-Galerkin discretizations with optimal test-spaces minimize the residual in some norm, as expected they are more stable. On the other hand, initially, i.e., for relatively

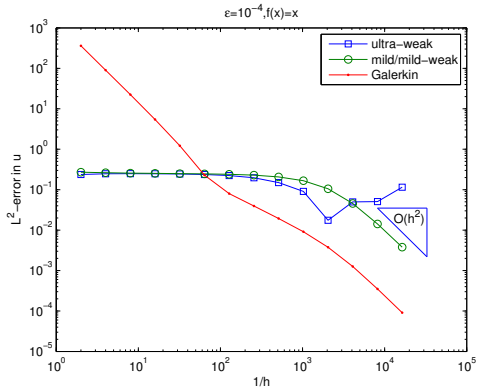

FIG. 1. $L_2(0,1)$-error in u_h vs. $1/h$ in the Galerkin, and in the Petrov-Galerkin approximations (mild/mild-weak, and ultra-weak) for the one-dimensional convection-diffusion equation with $\varepsilon = 10^{-4}$.

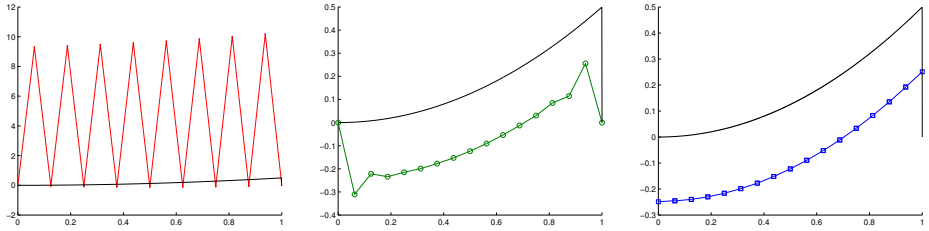

FIG. 2. Exact solution u and the Galerkin (left), mild/mild-weak, and ultra-weak Petrov-Galerkin approximations u_h for $h = \frac{1}{16}$ and $\varepsilon = 10^{-4}$.

large h, also the Petrov-Galerkin discretizations do not yield quasi-optimal approximations from the trial space.

The latter can be understood by noticing that the Petrov-Galerkin methods minimize the error over the trial space in $\|B\cdot\|_{V'}$, where $B : U \to V'$ is the operator associated to the bilinear form. From the variational problem not being well-posed for $\varepsilon = 0$, we infer that $\lim_{\varepsilon\downarrow 0} \inf_{0\neq u\in U} \|Bu\|_{V'}/\|u\|_U = 0$. Consequently, for small ε, some components of the difference between the solution and an approximation from the trial space U_h, typically near-constants, are hardly measured in $\|B\cdot\|_{V'}$, and therefore they will hardly be reduced in the least squares minimization, cf. Cohen *et al.* (2012). This results in Petrov-Galerkin approximations that have some oscillations at out- *and* inflow boundaries. These oscillations, however, are much smaller than with the Galerkin solution.

A way to improve results for the Petrov-Galerkin discretizations is to change the norm on V. Since for any fixed $\varepsilon > 0$, $B : U \to V'$, and so $B' : V \to U'$ are boundedly invertible, $\|B'\cdot\|_{U'}$ defines a norm on V. Equipping V' with the corresponding dual norm, one infers that $\|B\cdot\|_{V'} = \|\cdot\|_U$, so that an optimal Petrov-Galerkin method yields the best approximation from the trial space to u w.r.t. the original (ε-independent) norm $\|\cdot\|_U$ on U. For this reason, the norm $\|B'\cdot\|_{U'}$ on V is called the optimal test norm. Such an approach has been investigated in a somewhat different context in Dahmen *et al.* (2012); Cohen

et al. (2012), and for the ultra-weak formulation in Demkowicz & Heuer (2011).

Unfortunately, it turns out that for the resulting optimal Petrov-Galerkin discretization of the ultra-weak formulation, the variational problems on V, that determine the optimal test functions, are singularly perturbed ones with solutions that exhibit boundary layers, which for $\varepsilon \downarrow 0$ are increasingly difficult to resolve with a sufficient accuracy. Therefore, modified methods are proposed that aim at finding a good compromise between obtaining a best approximation in a nearly ε-independent norm, and getting easy-to-solve variational problems for the test functions. In Chan *et al.* (2012), additionally it was proposed to modify the boundary condition at the inflow boundary to ensure that solutions of the dual problem have no boundary layers.

Inspired by Cohen *et al.* (2012), the approach that we will investigate for the convection-dominated case is based on the observation that to avoid that a numerical solution method looses convergence or becomes increasingly more costly when $\varepsilon \downarrow 0$, a *necessary* condition is that the scheme is well-defined and convergent in the limit case $\varepsilon = 0$. Since the latter requires that this limit problem is well-posed, in the next subsection we start with studying variational formulations of the pure-convection problem.

5.2 *Limit problem: Pure convection*

We are searching for a variational formulation of the convection-diffusion problem in mixed form that is also well-defined in the limit case $\varepsilon = 0$. This can only be expected when (5.1) for $\varepsilon = 0$ allows a well-posed variational formulation, which is only possible when for $\varepsilon = 0$ the homogeneous Dirichlet boundary conditions are restricted to the inflow boundary. We set

$$\Gamma_- := \{\mathbf{x} \in \partial\Omega : \mathbf{b}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) < 0\}, \quad \Gamma_+ := \{\mathbf{x} \in \partial\Omega : \mathbf{b}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) > 0\}, \tag{5.2}$$

and with that fix the skeleton $\partial\Omega_h^\circ = \cup_{K\in\Omega_h} \partial K \backslash \Gamma_+$.

Canonical variational formulations of the convection problem are finding u, zero on Γ_-, such that

$$\int_\Omega v\mathbf{b} \cdot \nabla u = \int_\Omega fv$$

for all test functions v, or finding u such that

$$-\int_\Omega u \operatorname{div}(v\mathbf{b}) = \int_\Omega fv$$

for all test functions v that vanish at Γ_+. To arrive at the second formulation by using integration by parts, we used that $\int_{\partial\Omega\setminus\Gamma_+} uv\mathbf{b}\cdot\mathbf{n}$ vanishes because of the Dirichlet boundary condition on Γ_- –which with this formulation is a *natural* one–, and by $\mathbf{b}\cdot\mathbf{n} = 0$ on $\partial\Omega \setminus (\Gamma_+ \cup \Gamma_-)$.

Relevant Hilbert spaces for these variational formulations are

$$H(\mathbf{b};\Omega) := \{u \in L_2(\Omega) : \mathbf{b}\cdot\nabla u \in L_2(\Omega)\},$$

equipped with $\|u\|^2_{H(\mathbf{b};\Omega)} := \|u\|^2_{L_2(\Omega)} + \|\mathbf{b}\cdot\nabla u\|^2_{L_2(\Omega)}$, its closed subspace

$$H_{0,\Gamma_-}(\mathbf{b};\Omega) := \operatorname{clos}_{H(\mathbf{b};\Omega)}\{u \in C(\bar\Omega) \cap H(\mathbf{b};\Omega) : u = 0 \text{ on } \Gamma_-\},$$

and $H_{0,\Gamma_+}(\mathbf{b};\Omega)$ defined analogously.

In this subsection, we assume that $\mathbf{b} \in W_\infty(\operatorname{div};\Omega)$, and that

$$H_{0,\Gamma_-}(\mathbf{b};\Omega) \to L_2(\Omega) : u \mapsto \mathbf{b}\cdot\nabla u \quad \text{is boundedly invertible,} \tag{5.3}$$

$$H_{0,\Gamma_+}(\mathbf{b};\Omega) \to L_2(\Omega) : v \mapsto -\operatorname{div} v\mathbf{b} \quad \text{is boundedly invertible.} \tag{5.4}$$

These latter two assumptions are readily verified for non-zero, constant $\mathbf{b}$, but are not satisfied for any vector field $\mathbf{b}$, as when flow curves associated to $\pm\mathbf{b}$ do not reach the boundary. The assumptions (5.3) and (5.4) mean that the first or second variational form associated to the convection problem is well-posed over $H_{0,\Gamma_-}(\mathbf{b};\Omega)\times L_2(\Omega)$ or $L_2(\Omega)\times H_{0,\Gamma_+}(\mathbf{b};\Omega)$, respectively. Sufficient conditions on $\mathbf{b}$ and Γ_- for (5.3) or (5.4) to be valid can be found in De Sterck *et al.* (2004).

Piecewise integration by parts of the convection equation leads to the following problem

$$\left\{\begin{array}{l} \text{With } U^0 := L_2(\Omega)\times H(\mathbf{b};\partial\Omega_h^\circ),\ V^0 := H_{0,\Gamma_+}(\mathbf{b};\Omega_h),\\ \text{given } f\in H_{0,\Gamma_+}(\mathbf{b};\Omega_h)',\text{find } (u^0,\theta^0)\in U^0 \text{ such that for all } v\in V^0,\\ b(u^0,\theta^0,v) := \displaystyle\int_\Omega -u^0(\mathbf{b}\cdot\nabla_h v - v\,\mathrm{div}\,\mathbf{b}) + \int_{\partial\Omega_h^\circ}[\![v]\!]\theta^0 = f(v). \end{array}\right. \tag{5.5}$$

Here $H_{0,\Gamma_+}(\mathbf{b};\Omega_h)$ is the closure in $\{v\in L_2(\Omega):\mathbf{b}\cdot\nabla_h v\in L_2(\Omega)\}$ –equipped with squared "broken" norm $\|v\|^2_{H(\mathbf{b};\Omega_h)} := \|v\|^2_{L_2(\Omega)} + \|\mathbf{b}\cdot\nabla_h v\|^2_{L_2(\Omega)}$– of its subspace consisting of the functions that additionally are piecewise continuous w.r.t. $\bar\Omega = \cup_{K\in\Omega_h}\bar K$ and vanish at Γ_+; and

$$H(\mathbf{b};\partial\Omega_h^\circ) := \{w\mathbf{b}|_{\partial\Omega_h^\circ}\cdot\mathbf{n} : w\in H_{0,\Gamma_-}(\mathbf{b};\Omega)\},$$

equipped with quotient norm

$$\|\theta\|_{H(\mathbf{b};\partial\Omega_h^\circ)} := \inf\{\|w\|_{H(\mathbf{b};\Omega)} : \theta = w\mathbf{b}|_{\partial\Omega_h^\circ}\cdot\mathbf{n},\, w\in H_{0,\Gamma_-}(\mathbf{b};\Omega)\}.$$

Note that if $f\in L_2(\Omega)$, or, equivalently, $u^0\in H_{0,\Gamma_-}(\mathbf{b};\Omega)$, then

$$\theta^0 = u^0\mathbf{b}|_{\partial\Omega_h^\circ}\cdot\mathbf{n}. \tag{5.6}$$

THEOREM 5.1 With $(B(u,\theta))(v) := b(u,\theta,v)$, it holds that $B: U^0\to {V^0}'$ is boundedly invertible with $\sup_h \max(\|B\|_{U^0\to {V^0}'}, \|B^{-1}\|_{{V^0}'\to U^0}) < \infty$.

For proving this theorem, we need the following result (cf. Thm. 3.2):

LEMMA 5.1 For $v\in H_{0,\Gamma_+}(\mathbf{b};\Omega_h)$, it holds that $[\![v]\!]\in (H(\mathbf{b};\partial\Omega_h^\circ))'$, and

$$\|[\![v]\!]\|_{H(\mathbf{b};\partial\Omega_h^\circ)'} \eqsim \inf_{z\in H_{0,\Gamma_+}(\mathbf{b};\Omega)}\|v-z\|_{H(\mathbf{b};\Omega_h)} \quad (v\in H_{0,\Gamma_+}(\mathbf{b};\Omega_h)).$$

Proof. For $v\in H_{0,\Gamma_+}(\mathbf{b};\Omega_h)$, $w\in H_{0,\Gamma_-}(\mathbf{b};\Omega)$, we have

$$\int_{\partial\Omega_h^\circ}[\![v]\!]w\mathbf{b}\cdot\mathbf{n} = \sum_{K\in\Omega_h}\int_K \nabla v\cdot\mathbf{b}w + v(\mathbf{b}\cdot\nabla w + w\,\mathrm{div}\,\mathbf{b}) \lesssim \|v\|_{H(\mathbf{b};\Omega_h)}\|w\|_{H(\mathbf{b};\Omega)}, \tag{5.7}$$

showing that $\|[\![v]\!]\|_{H(\mathbf{b};\partial\Omega_h^\circ)'}\lesssim\|v\|_{H(\mathbf{b};\Omega_h)}$. Since for $z\in H_{0,\Gamma_+}(\mathbf{b};\Omega)$, and $w\in H_{0,\Gamma_-}(\mathbf{b};\Omega)$, $\int_\Omega \nabla z\cdot\mathbf{b}w + z(\mathbf{b}\cdot\nabla w + w\,\mathrm{div}\,\mathbf{b}) = 0$, it follows that $\|[\![z]\!]\|_{H(\mathbf{b};\partial\Omega_h^\circ)'} = 0$. We infer that for $v\in H_{0,\Gamma_+}(\mathbf{b};\Omega_h)$, $\|[\![v]\!]\|_{H(\mathbf{b};\partial\Omega_h^\circ)'}\lesssim\inf_{z\in H_{0,\Gamma_+}(\mathbf{b};\Omega)}\|v-z\|_{H(\mathbf{b};\Omega_h)}$.

Given $v\in H_{0,\Gamma_+}(\mathbf{b};\Omega_h)$, let $z\in H_{0,\Gamma_+}(\mathbf{b};\Omega)$ be the solution of $\mathrm{div}\, z\mathbf{b} = \mathrm{div}_h\, v\mathbf{b}$ whose existence is guaranteed by (5.4). From $0 = \mathrm{div}_h(v-z)\mathbf{b} = (v-z)\,\mathrm{div}\,\mathbf{b} + \mathbf{b}\cdot\nabla_h(v-z)$ and $\mathbf{b}\in W_\infty(\mathrm{div};\Omega)$, we have $\|\mathbf{b}\cdot\nabla_h(v-z)\|_{L_2(\Omega)}\lesssim\|v-z\|_{L_2(\Omega)}$.

By (5.3), there exists a $w \in H_{0,\Gamma_-}(\mathbf{b};\Omega)$ with $\mathbf{b}\cdot\nabla w = v - z$ and $\|w\|_{H(\mathbf{b};\Omega)} \lesssim \|v-z\|_{L_2(\Omega)}$. From the definitions of z and w, we have

$$\begin{aligned}\|v-z\|^2_{L_2(\Omega)} &= \int_\Omega (v-z)\mathbf{b}\cdot\nabla w = \sum_{K\in\Omega_h}\int_K (v-z)\mathbf{b}\cdot\nabla w \\ &= \sum_{K\in\Omega_h}\int_K \operatorname{div}(v-z)w\mathbf{b} = \sum_{K\in\Omega_h}\int_{\partial K}(v-z)w\mathbf{b}\cdot\mathbf{n}_K = \int_{\partial\Omega_h^\circ}[\![v]\!]w\mathbf{b}\cdot\mathbf{n} \\ &\leqslant \|[\![v]\!]\|_{H(\mathbf{b};\partial\Omega_h^\circ)'}\|w\|_{H(\mathbf{b};\Omega)} \lesssim \|[\![v]\!]\|_{H(\mathbf{b};\partial\Omega_h^\circ)'}\|v-z\|_{L_2(\Omega)},\end{aligned}$$

or $\|v-z\|_{L_2(\Omega)} \lesssim \|[\![v]\!]\|_{H(\mathbf{b};\partial\Omega_h^\circ)'}$, which completes the proof. □

Proof of Theorem 5.1. The boundedness of B, uniformly in h, follows easily from (5.7).

Now let $(u,\theta)\in U^0$ be such that $b(u,\theta,v)=0$ for all $v\in H_{0,\Gamma_+}(\mathbf{b};\Omega_h)$. Considering all v from the subspace $H_{0,\Gamma_+}(\mathbf{b};\Omega)$ shows that $u=0$ by (5.4). By now considering all v with $\operatorname{supp} v \subset K \in \Omega_h$, we infer that $\theta|_{\partial K\setminus\Gamma_+}=0$, and so $\theta=0$.

Finally, let $v\in H_{0,\Gamma_+}(\mathbf{b};\Omega_h)$ be given. By (5.4), there exists a $v_1\in H_{0,\Gamma_+}(\mathbf{b};\Omega)$ with $\operatorname{div} v_1\mathbf{b} = \operatorname{div}_h v_1\mathbf{b}$, and $\|v_1\|_{H(\mathbf{b};\Omega)}\lesssim\|v\|_{H(\mathbf{b};\Omega_h)}$, and so $\|v_1-v\|_{H(\mathbf{b};\Omega_h)}\lesssim\|v\|_{H(\mathbf{b};\Omega_h)}$. Moreover, we have $v_1=v$ when $v\in H_{0,\Gamma_+}(\mathbf{b};\Omega)$, and so

$$\|v_1-v\|_{H(\mathbf{b};\Omega_h)} \lesssim \inf_{z\in H_{0,\Gamma_+}(\mathbf{b};\Omega)}\|v-z\|_{H(\mathbf{b};\Omega_h)} \lesssim \|[\![v]\!]\|_{H(\mathbf{b};\partial\Omega_h^\circ)'},$$

by Lemma 5.1.

From (5.4) and Lemma 3.1, we deduce that there exists a $u\in L_2(\Omega)$ with $\|u\|_{L_2(\Omega)} = \|v_1\|_{H(\mathbf{b};\Omega)}$ and

$$\|v_1\|^2_{H(\mathbf{b};\Omega)} \lesssim -\int_\Omega u\operatorname{div} v_1\mathbf{b} = -\int_\Omega u\operatorname{div}_h v\mathbf{b}.$$

By definition of a dual norm, there exists a $\theta\in H(\mathbf{b};\partial\Omega_h^\circ)$ with $\|\theta\|^2_{H(\mathbf{b};\partial\Omega_h^\circ)} = \|[\![v]\!]\|^2_{H(\mathbf{b};\partial\Omega_h^\circ)'} = \int_{\partial\Omega_h^\circ}[\![v]\!]\theta$. We conclude that

$$\|v\|_{H(\mathbf{b};\Omega_h)} \lesssim \sqrt{\|v_1\|^2_{H(\mathbf{b};\Omega)} + \|[\![v]\!]\|^2_{H(\mathbf{b};\partial\Omega_h^\circ)'}} \lesssim \frac{b(u,\theta,v)}{\sqrt{\|u\|^2_{L_2(\Omega)} + \|\theta\|^2_{H(\mathbf{b};\partial\Omega_h^\circ)}}}.$$

Having verified all three conditions of Lemma 3.1 for the operator B, where (i) and (ii) hold uniformly in h, the proof is completed. □

As with the diffusion problem discussed in Sect. 3.1, given a closed trial space $U_h^0\subset U^0$, we can run a Petrov-Galerkin discretization of the pure convection problem (5.5) with optimal test space $V_h^0 = \mathfrak{T}T|_{U_h^0}$. Here $T: U^0\to V^0 = H_{0,\Gamma^+}(\mathbf{b};\Omega_h)$ is defined by $v=(v_K)_{K\in\Omega_h} = T(u^0,\theta^0)$, with $v_K\in H_{0,\partial K\cap\Gamma_+}(\mathbf{b};K) := \{z\in L_2(K) : \mathbf{b}\cdot\nabla z\in L_2(K),\, z=0 \text{ on } \partial K\cap\Gamma_+\}$ being the solution of

$$\langle v_K,\hat v\rangle_{L_2(K)} + \langle \mathbf{b}\cdot\nabla v_K, \mathbf{b}\cdot\nabla\hat v\rangle_{L_2(K)} = \int_K -u^0(\mathbf{b}\cdot\nabla\hat v - \hat v\operatorname{div}\mathbf{b}) + \int_{\partial K}\mathbf{n}_K^\top\mathbf{n}\,\hat v\theta^0$$

$(\hat v\in H_{0,\partial K\cap\Gamma_+}(\mathbf{b};K))$.

As a direct consequence of Proposition 2.1, Theorem 3.21, the definition of $H(\mathbf{b},\partial\Omega_h^\circ)$, and (5.6), we have the following error estimate:

COROLLARY 5.1 For a $k \in \mathbb{N}$, let the solution of (5.5) satisfy $u^0 \in H^{k+1}(\Omega)$. Let $P_h \subset L_2(\Omega)$, and $W_h \subset H_{0,\Gamma_-}(\mathbf{b};\Omega)$ with

$$\begin{aligned} &\inf_{\hat{u}_h \in P_h} \|\hat{u}-\hat{u}_h\|_{L_2(\Omega)} \lesssim h^k \|\hat{u}\|_{H^k(\Omega)} \quad (\hat{u} \in H^k(\Omega) \cap H^1_{0,\Gamma_-}(\Omega)), \\ &\inf_{w_h \in W_h} \|w-w_h\|_{H(\mathbf{b};\Omega)} \lesssim h^k \|w\|_{H^{k+1}(\Omega)} \quad (w \in H^{k+1}(\Omega) \cap H^1_{0,\Gamma_-}(\Omega)). \end{aligned} \tag{5.8}$$

Then the Petrov-Galerkin approximation (u_h^0, θ_h^0) from $U_h^0 = \{(\hat{u}_h, w_h\mathbf{b}|_{\partial\Omega_h^\circ} \cdot \mathbf{n}) : (\hat{u}_h, w_h) \in P_h \times W_h\}$ with optimal test space V_h^0 satisfies

$$\|u^0 - u_h^0\|_{L_2(\Omega)} + \|\theta^0 - \theta_h^0\|_{H(\mathbf{b};\partial\Omega_h^\circ)} \lesssim h^k \|u^0\|_{H^{k+1}(\Omega)}.$$

Clearly, (5.8) is satisfied when $\inf_{w_h \in W_h} \|w-w_h\|_{H^1(\Omega)} \lesssim h^k \|w\|_{H^{k+1}(\Omega)}$ for $w \in H^{k+1}(\Omega) \cap H^1_{0,\Gamma_-}(\Omega)$. In view of this estimate, thinking of P_h and W_h being common finite element spaces, the order of W_h should be one higher than the order of P_h.

Under conditions on W_h, it might be possible to approximate u^0 in $H_{0,\Gamma_-}(\mathbf{b};\Omega)$ from W_h, and so θ_0 in $H(\mathbf{b};\partial\Omega_h^\circ)$ from $\{w_h\mathbf{b}|_{\partial\Omega_h^\circ} \cdot \mathbf{n} : w_h \in W_h\}$, with an error of $\mathscr{O}(h^k)$ under the relaxed conditions $u^0 \in H^k(\Omega)$ and $f = \mathbf{b}\cdot\nabla u^0 \in H^k(\Omega)$, replacing the additional smoothness condition $u^0 \in H^{k+1}(\Omega)$.

5.3 *A Petrov-Galerkin discretization that allows passing to the convective limit*

We consider the mild-weak variational formulation (3.7) for the convection-diffusion problem, and assume that $\mathbf{b} \in W_\infty(\mathrm{div};\Omega)$. In view of the limit case $\varepsilon = 0$ analyzed in §5.2, we now fix

$$\Gamma_- := \{\mathbf{x} \in \partial\Omega : \mathbf{b}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) < 0\}, \quad \Gamma_+ := \{\mathbf{x} \in \partial\Omega : \mathbf{b}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) > 0\},$$

that latter which determines the skeleton $\partial\Omega_h^\circ = \cup_{K\in\Omega_h} \partial K \backslash \Gamma_+$, and so the function spaces $H^{-\frac{1}{2}}(\partial\Omega_h^\circ)$, $H(\mathbf{b};\partial\Omega_h^\circ)$ on the skeleton, and their duals, as well as the "broken" spaces $H^1_{0,\Gamma_+}(\Omega_h)$ and $H_{0,\Gamma_+}(\mathbf{b};\Omega_h)$.

With a factorization of $\mathbf{A}(\varepsilon) = \varepsilon\,\mathrm{Id}$ as $\mathbf{A}(\varepsilon) = \mathbf{A}_1(\varepsilon)\mathbf{A}_2(\varepsilon)$, the mild-weak variational problem reads as finding

$$\begin{cases} (\boldsymbol{\sigma}, u, \theta) \in U = L_2(\Omega)^n \times H^1_0(\Omega) \times H^{-\frac{1}{2}}(\partial\Omega_h^\circ), \text{ such that} \\ \displaystyle\int_\Omega (\boldsymbol{\sigma} - \mathbf{A}_2(\varepsilon)\nabla u)\cdot\boldsymbol{\tau} + (\mathbf{A}_1(\varepsilon)\boldsymbol{\sigma} - u\mathbf{b})\cdot\nabla_h v - uv\,\mathrm{div}\,\mathbf{b} + \int_{\partial\Omega_h^\circ} [\![v]\!]\theta = f(v) \\ \text{for all } (\boldsymbol{\tau}, v) \in V = L_2(\Omega)^n \times H^1_{0,\Gamma_+}(\Omega_h). \end{cases} \tag{5.9}$$

We will select $\mathbf{A}_1(\varepsilon)$ and $\mathbf{A}_2(\varepsilon)$ in such a way that they both vanish for $\varepsilon = 0$. Then (5.9) for $\varepsilon = 0$ reads as a well-posed decoupled system of equations $\boldsymbol{\sigma} = 0$, and the variational formulation (5.5) of the pure convection problem. Recalling the definitions $U^0 = L_2(\Omega) \times H(\mathbf{b};\partial\Omega_h^\circ)$ and $V^0 = H_{0,\Gamma_+}(\mathbf{b};\Omega_h)$ for the spaces in (5.5) for (u,θ) and v, respectively, in order to achieve this we will equip the test space $H^1_{0,\Gamma_+}(\Omega_h)$ with an ε-dependent norm that for $\varepsilon = 0$ reduces to the norm on $H_{0,\Gamma_+}(\mathbf{b};\Omega_h)$.

For an optimal Petrov-Galerkin discretization, it remains to specify trial spaces. Since $H^1_0(\Omega) \hookrightarrow L_2(\Omega)$ is dense, a trial space for u that is suitable for $\varepsilon > 0$, is also applicable for $\varepsilon = 0$. Since for $w \in H(\mathbf{b};\Omega)$, one has $w\mathbf{b} \in H(\mathrm{div};\Omega)$ with $\|w\mathbf{b}\|_{H(\mathrm{div};\Omega)} \lesssim \|w\|_{H(\mathbf{b};\Omega)}$, it holds that $H(\mathbf{b};\partial\Omega_h^\circ) \hookrightarrow H^{-\frac{1}{2}}(\partial\Omega_h^\circ)$. So a trial space for θ that is applicable for $\varepsilon = 0$, may also be applied for $\varepsilon > 0$. Since,

however, $H(\mathbf{b};\partial\Omega_h^\circ) \hookrightarrow H^{-\frac{1}{2}}(\partial\Omega_h^\circ)$ is not dense, such a choice cannot be expected to give rise to a convergent scheme for $\varepsilon > 0$. Therefore, using that for $f \in L_2(\Omega)$, it holds that $\theta = (u\mathbf{b} - \mathbf{A}_1(\varepsilon)\boldsymbol{\sigma})|_{\partial\Omega_h^\circ} \cdot \mathbf{n})$, we will approximate θ by a linear combination of an element from the trial space for $\varepsilon = 0$ and $\mathbf{A}_1(\varepsilon)\hat{\boldsymbol{\sigma}}_h|_{\partial\Omega_h^\circ} \cdot \mathbf{n}$ with $\hat{\boldsymbol{\sigma}}_h$ from the trial space for $\boldsymbol{\sigma}$. This latter construction is described precisely in the following proposition.

PROPOSITION 5.2 Let $H^1_{0,\Gamma_+}(\Omega_h)$ be equipped with squared norm

$$\mu(\varepsilon)\|\nabla_h v\|^2_{L_2(\Omega)} + \|v\|^2_{H(\mathbf{b};\Omega_h)}. \tag{5.10}$$

where $\mu(\varepsilon) > 0$ for $\varepsilon > 0$. Let $\mathbf{A}_1(\varepsilon) \in W^1_\infty(\Omega)$ be scalar valued. For some $\mathbf{Q}_h \subset H(\mathrm{div};\Omega)$ and $U_h^0 \subset H^1_0(\Omega) \times H(\mathbf{b};\partial\Omega_h^\circ)$, let the trial space

$$\boxed{U_h(\varepsilon) = \{(\hat{\boldsymbol{\sigma}}_h, \hat{u}_h, \hat{\theta}_h - \mathbf{A}_1(\varepsilon)\hat{\boldsymbol{\sigma}}_h|_{\partial\Omega_h^\circ} \cdot \mathbf{n}) : \hat{\boldsymbol{\sigma}}_h \in \mathbf{Q}_h, (\hat{u}_h, \hat{\theta}_h) \in U_h^0\}}. \tag{5.11}$$

Then $U_h(\varepsilon) \subset U$ and $U_h^0 \subset U^0$, defined in (5.5). When furthermore $\mathbf{A}_1(0) = \mathbf{A}_2(0) = 0$ and $\mu(0) = 0$, then the Petrov-Galerkin solution $(\boldsymbol{\sigma}_h, u_h, \theta_h) \in U_h(\varepsilon)$ with corresponding optimal test space $V_h(\varepsilon)$ of (5.9) also exists uniquely for $\varepsilon = 0$ (for $\varepsilon = 0$, reading V as $L_2(\Omega)^n \times H_{0,\Gamma_+}(\mathbf{b};\Omega_h)$). For $\varepsilon = 0$, it satisfies $\boldsymbol{\sigma}_h = 0$, whereas (u_h, θ_h) is the Petrov-Galerkin solution with optimal test space V_h^0 of the pure convection problem (5.5) with trial space U_h^0.

Proof. Already we know that $H(\mathbf{b};\partial\Omega_h^\circ) \hookrightarrow H^{-\frac{1}{2}}(\partial\Omega_h^\circ)$. Since multiplication with $\mathbf{A}_1(\varepsilon)$ maps $H(\mathrm{div};\Omega)$ into itself, we infer that $U_h(\varepsilon) \subset U$. From $H^1_0(\Omega) \subset L_2(\Omega)$, one has $U_h^0 \subset U^0$.

Let $\varepsilon = 0$. Since *both* $\mathbf{A}_1(0)$ and $\mathbf{A}_2(0)$ are zero, the equations for $\boldsymbol{\sigma}_h$ and (u_h, θ_h) are fully decoupled. The optimal test space $\mathfrak{I}T|_{U_h(0)}$ is the Cartesian product of $\mathbf{Q}_h$ and, by the selection of the norm on V and $\mu(0) = 0$, the optimal test space V_h^0 of the Petrov-Galerkin discretization of the pure convection problem with trial space U_h^0, giving the solution $(\boldsymbol{\sigma}_h, u_h, \theta_h)$ as stated. □

We recall that the optimal test space $V_h(\varepsilon)$ is $\mathfrak{I}T|_{U_h(\varepsilon)}$, with $(\boldsymbol{\tau}, v) = T(\boldsymbol{\sigma}, u, \theta)$ given by $\boldsymbol{\tau} = \boldsymbol{\sigma} - \mathbf{A}_2(\varepsilon)\nabla u$, and $v = (v_K)|_{K\in\Omega_h}$, where $v_K \in H^1_{0,\partial K\cap\Gamma_+}(K)$ solves

$$\begin{aligned}&\mu(\varepsilon)\langle\nabla v_K, \nabla\hat{v}\rangle_{L_2(K)^n} + \langle v_K, \hat{v}\rangle_{L_2(K)} + \langle\mathbf{b}\cdot\nabla v_K, \mathbf{b}\cdot\nabla\hat{v}\rangle_{L_2(K)^n}\\ &= \int_K (\mathbf{A}_1(\varepsilon)\boldsymbol{\sigma} - u\mathbf{b})\cdot\nabla\hat{v} - u\hat{v}\,\mathrm{div}\,\mathbf{b} + \int_{\partial K\setminus\Gamma_+} \mathbf{n}_K^\top\mathbf{n}\hat{v}\theta \quad (\hat{v} \in H^1_{0,\partial K\cap\Gamma_+}(K)),\end{aligned} \tag{5.12}$$

where, for $\varepsilon = 0$, $H^1_{0,\partial K\cap\Gamma_+}(K)$ should read as $H_{0,\partial K\cap\Gamma_+}(\mathbf{b};K)$ (cf. (3.11)). For $\hat{\boldsymbol{\sigma}}_h \in \mathbf{Q}_h$, and $(\hat{u}_h, \hat{\theta}_h) \in U_h^0$, substituting $(\boldsymbol{\sigma}, u, \theta) = (\hat{\boldsymbol{\sigma}}_h, \hat{u}_h, \hat{\theta}_h - \mathbf{A}_1(\varepsilon)\hat{\boldsymbol{\sigma}}_h|_{\partial\Omega_h^\circ} \cdot \mathbf{n})$ in the right-hand side of (5.12), by applying integration by parts it reads as

$$\int_K -\hat{u}_h\mathbf{b}\cdot\nabla\hat{v} - \hat{u}_h\hat{v}\,\mathrm{div}\,\mathbf{b} - \hat{v}\,\mathrm{div}\,\mathbf{A}_1(\varepsilon)\hat{\boldsymbol{\sigma}}_h + \int_{\partial K} \mathbf{n}_K^\top\mathbf{n}\hat{v}\hat{\theta}_h.$$

It defines a uniformly in $\varepsilon \in [0,1]$ bounded linear functional on $H_{0,\partial K\cap\Gamma_+}(\mathbf{b};K)$, and so on $H^1_{0,\partial K\cap\Gamma_+}(K)$ equipped with squared norm $\mu(\varepsilon)\|\nabla\cdot\|^2_{L_2(K)^n} + \|\cdot\|^2_{H(\mathbf{b};K)}$. Since consequently (5.12) is well-posed *uniformly* in $\varepsilon \in [0,1]$, we do not expect (nor observe) solutions that exhibit boundary layers.

Let us discuss the selection in Proposition 5.2 of $\mathbf{Q}_h$ and U_h^0, and so of $U_h(\varepsilon)$, in relation to error estimates for quasi-uniform meshes. Let $\mathbf{Q}_h$ be a space that satisfies (3.24), e.g., a Raviart-Thomas space

w.r.t. Ω_h of order k. Let $W_h \subset H^1_{0,\Gamma_-}(\Omega)$, $P_h \subset H^1_0(\Omega)$ be finite element spaces w.r.t. Ω_h of orders $\ell \geqslant k$ and k, respectively, with $P_h \subset W_h$. With $I_h : W_h \to P_h$ being the interpolant w.r.t. the nodal variables of P_h, we set $R_h := \mathfrak{I}(I - I_h)$. We take

$$U_h^0 = \{(\hat{u}_h, (\hat{u}_h + r_h)\mathbf{b}|_{\partial\Omega_h^\circ} \cdot \mathbf{n}) : \hat{u}_h \in P_h,\, r_h \in R_h\}, \tag{5.13}$$

so that

$$\boxed{U_h(\varepsilon) = \{\big(\hat{\boldsymbol{\sigma}}_h, \hat{u}_h, ((\hat{u}_h + r_h)\mathbf{b} - \mathbf{A}_1(\varepsilon)\hat{\boldsymbol{\sigma}}_h)|_{\partial\Omega_h^\circ} \cdot \mathbf{n}\big) : (\hat{\boldsymbol{\sigma}}_h, \hat{u}_h, r_h) \in \mathbf{Q}_h \times P_h \times R_h\}}. \tag{5.14}$$

REMARK 5.1 In most of our numerical experiments, we will take $P_h = W_h \cap H^1_0(\Omega)$, in which case $\operatorname{supp} r_h \subset \cup_{\{K \in \Omega_h : \partial K \cap (\partial\Omega \setminus \Gamma_-) \neq \emptyset\}} \overline{K}$ for any $r_h \in R_h$.

For $n = 1$, and with the obvious choice of the nodal variables, it holds that $R_h|_{\Omega_h^\circ} = \{0\}$, so that the trial space $U_h(\varepsilon)$ is equal to that from Remark 3.10.

The space $U_h(\varepsilon)$ from (5.14) includes that from Remark 3.10. Since furthermore for any *fixed* $\varepsilon > 0$, the norm defined in (5.10) is equivalent to the standard norm on V, the optimal error estimates from that remark apply.

In order to obtain favorable results as function of simultaneously the discretisation *and* ε, it is necessary that for $\varepsilon = 0$ the Petrov-Galerkin approximations are converging for $h \to 0$ to $(u^0, \theta^0) \in U^0 = L_2(\Omega) \times H(\mathbf{b}; \partial\Omega_h^\circ)$, being the solution of the pure convection problem.

For $Q_h : H^1(\Omega) \to W_h$ being, say, the Scott-Zhang quasi-interpolator (Scott & Zhang (1990)), we have

$$\|(I - Q_h)w\|_{L_2(\Omega)} + h|(I - Q_h)w|_{H^1(\Omega)} \lesssim h^s \|w\|_{H^s(\Omega)}$$

($s \in [1, \ell] \setminus \mathbb{N} + \{\frac{1}{2}\}$, $w \in H^s(\Omega) \cap H^1_{0,\Gamma_-}(\Omega)$), and by the $H^1(\Omega)$-stability of Q_h,

$$\|(I - I_h)Q_h w\|_{L_2(\Omega)} \lesssim h^{\frac{1}{2}} \|Q_h w\|_{H^1(\Omega)} \lesssim h^{\frac{1}{2}} \|w\|_{H^1(\Omega)} \quad (w \in H^1_{0,\Gamma_-}(\Omega)). \tag{5.15}$$

Therefore, selecting $\hat{u}_h = I_h Q_h u^0$, and $r_h = (I - I_h)Q_h u^0$, using $\theta^0 = u^0 \mathbf{b}|_{\partial\Omega_h^\circ} \cdot \mathbf{n}$, we infer that for the $\varepsilon = 0$ case

$$\begin{aligned} \|(u^0, \theta^0) - (u_h, \theta_h)\|_{U^0} &\lesssim \inf_{\hat{u}_h \in P_h,\, r_h \in R_h} \|(u^0, \theta^0) - (\hat{u}_h, (\hat{u}_h + r_h)\mathbf{b}|_{\partial\Omega_h^\circ} \cdot \mathbf{n})\|_{U^0} \\ &\lesssim \|(I - I_h Q_h)u^0\|_{L_2(\Omega)} + \|(I - Q_h)u^0\|_{H(\mathbf{b};\Omega)} \\ &\leqslant \|(I - Q_h)u^0\|_{L_2(\Omega)} + \|(I - I_h)Q_h u^0\|_{L_2(\Omega)} + |(I - Q_h)u^0|_{H^1(\Omega)} \\ &\lesssim h^{\frac{1}{2}} \|u^0\|_{H^s(\Omega)} \end{aligned} \tag{5.16}$$

when $u^0 \in H^s(\Omega) \cap H^1_{0,\Gamma_-}(\Omega)$ for some $s > \frac{3}{2}$.

REMARK 5.2 The reduced rate $\frac{1}{2}$ in (5.15) and, consequently, in (5.16), is due to the mismatch between the boundary conditions generally satisfied by u^0, and those that are incorporated in P_h in view of the application for $\varepsilon > 0$. In these circumstances, this rate is the best that generally can be expected. As follows from Corollary 5.1, without this mismatch, the error in the space U^0 would be $\mathcal{O}(h^{\min(\ell-1,k)})$ assuming that $u^0 \in H^{\min(\ell,k+1)}(\Omega)$.

Because for $\varepsilon \downarrow 0$ the exact solution u converges in $L_2(\Omega)$ to that of the pure convection problem, for $\varepsilon > 0$, and h being relatively large compared to ε, an error in u_h as for $\varepsilon = 0$ can be expected, i.e., $\sim h^{\frac{1}{2}}$ in $L_2(\Omega)$. This is also what we will observe in our numerical experiments. Note, again, that generally this is the best that can be realized with continuous piecewise polynomial approximation, zero at $\partial\Omega$, w.r.t. a quasi-uniform partition, and better results can only be achieved by a proper local refinement towards $\partial\Omega \setminus \Gamma_-$).

REMARK 5.3 All above considerations concerning the selection of U_h^0, and with that of $U_h(\varepsilon)$, would equally well apply to the choice $U_h^0 = \{(\hat{u}_h, w_h\mathbf{b}|_{\partial\Omega_h^\circ}\cdot\mathbf{n}) : \hat{u}_h \in P_h, w_h \in W_h\}$. The space U_h^0 from (5.13) is, however, of lower dimension, and, moreover, it turns out that the use of $\hat{u}_h$ for approximating simultaneously u and the flux θ has the effect that undesirable oscillations near the outflow boundary are damped.

Finally in this subsection, we discuss a relation between our method and the common first order least squares method (3.13), in the case of $\mathbf{Q}_h$ being the lowest order Raviart-Thomas space w.r.t. Ω_h, W_h being the space of continuous piecewise linears w.r.t. Ω_h, zero at Γ_-, $P_h = W_h \cap H_0^1(\Omega)$, and, for each ε, $\mathbf{A}_1(\varepsilon)$ being a constant. For $(\hat{\boldsymbol{\sigma}}_h, \hat{u}_h, \hat{\theta}_h) \in U_h(\varepsilon)$, and $\big(\boldsymbol{\tau}, (v_K)_{K\in\Omega_h}\big) = T(\hat{\boldsymbol{\sigma}}_h, \hat{u}_h, \hat{\theta}_h)$, as always we have $\boldsymbol{\tau} = \hat{\boldsymbol{\sigma}}_h - \mathbf{A}_2(\varepsilon)\nabla\hat{u}_h$, cf. (3.10).

For $r_h = 0$ in (5.14), i.e., $\hat{\theta}_h = (\hat{u}_h\mathbf{b} - \mathbf{A}_1(\varepsilon)\hat{\boldsymbol{\sigma}}_h)|_{\partial\Omega_h^\circ}\cdot\mathbf{n}$, a "reversed" integration by parts shows that

$$b(\hat{\boldsymbol{\sigma}}_h, \hat{u}_h, \hat{\theta}_h, \boldsymbol{\tau}, v) = \int_\Omega (\hat{\boldsymbol{\sigma}}_h - \mathbf{A}_2(\varepsilon)\nabla\hat{u}_h)\cdot\boldsymbol{\tau} + (\mathbf{b}\cdot\nabla\hat{u}_h - \operatorname{div}\mathbf{A}_1(\varepsilon)\hat{\boldsymbol{\sigma}}_h)v,$$

so that $v_K \in H^1_{0,\partial K\cap\Gamma_+}(K)$ solves

$$\begin{aligned}\mu(\varepsilon)\langle\nabla v_K, \nabla\hat{v}\rangle_{L_2(K)} + \langle v_K, \hat{v}\rangle_{L_2(K)} + \langle\mathbf{b}\cdot\nabla v_K, \mathbf{b}\cdot\nabla\hat{v}\rangle_{L_2(K)} \\ = \int_K (\mathbf{b}\cdot\nabla\hat{u}_h - \operatorname{div}\mathbf{A}_1(\varepsilon)\hat{\boldsymbol{\sigma}}_h)\hat{v} \quad (\hat{v} \in H^1_{0,\partial K\cap\Gamma_+}(K)).\end{aligned} \tag{5.17}$$

By our assumptions on $\mathbf{Q}_h$ and P_h, and from $\mathbf{A}_1(\varepsilon)$ being a constant, $(\mathbf{b}\cdot\nabla\hat{u}_h - \operatorname{div}\mathbf{A}_1(\varepsilon)\hat{\boldsymbol{\sigma}}_h)|_K$ is a constant. We conclude that for K with $\partial K\cap\Gamma_+ = \emptyset$, we have the explicit expression $v_K = (\mathbf{b}\cdot\nabla\hat{u}_h - \operatorname{div}\mathbf{A}_1(\varepsilon)\hat{\boldsymbol{\sigma}}_h)|_K$. For the remaining K along Γ_+, the local boundary value problem (5.17) has to be (approximately) solved.

For $(\hat{\boldsymbol{\sigma}}_h, \hat{u}_h, \hat{\theta}_h) = (0, 0, r_h\mathbf{b}|_{\partial\Omega_h^\circ}\cdot\mathbf{n})$ for some $r_h \in R_h$, we have $v_K = 0$ when $\partial K\cap(\partial\Omega\setminus\Gamma_-) \neq \emptyset$, whereas for the remaining K, $v_K \in H^1_{0,\partial K\cap\Gamma_+}(K)$ has to be (approximately) solved from

$$\begin{aligned}\mu(\varepsilon)\langle\nabla v_K, \nabla\hat{v}\rangle_{L_2(K)} + \langle v_K, \hat{v}\rangle_{L_2(K)} + \langle\mathbf{b}\cdot\nabla v_K, \mathbf{b}\cdot\nabla\hat{v}\rangle_{L_2(K)} \\ = \int_{\partial K} \mathbf{n}_K^\top\mathbf{n}\hat{v}r_h\mathbf{b}|_{\partial\Omega_h^\circ}\cdot\mathbf{n} \quad (\hat{v} \in H^1_{0,\partial K\cap\Gamma_+}(K)).\end{aligned} \tag{5.18}$$

We conclude that if it would not be that for K with $\partial K\cap\Gamma_+ \neq \emptyset$, v_K form (5.17) is unequal to $(\mathbf{b}\cdot\nabla\hat{u}_h - \operatorname{div}\mathbf{A}_1(\varepsilon)\hat{\boldsymbol{\sigma}}_h)|_K$, and, for $n > 1$, additional trial and so test functions, defined by (5.18), supported near $\partial\Omega\setminus\Gamma_-$ are added, then the first two components of the solution $(\boldsymbol{\sigma}_h, u_h, \theta_h)$ of our Petrov-Galerkin method with lowest order trial space and corresponding optimal test space would be equal to $\arg\min_{(\hat{\boldsymbol{\sigma}}_h,\hat{u}_h)\in\mathbf{Q}_h\times P_h} \|\hat{\boldsymbol{\sigma}}_h - \mathbf{A}_2(\varepsilon)\nabla\hat{u}_h\|^2_{L_2(\Omega)^n} + \|\mathbf{b}\cdot\nabla\hat{u}_h - \operatorname{div}\mathbf{A}_1(\varepsilon)\hat{\boldsymbol{\sigma}}_h\|^2_{L_2(\Omega)}$. That is, up to a harmless scaling of $\boldsymbol{\sigma}_h$, they would be the solution of the common first order least squares method (3.13) with trial space $\mathbf{Q}_h \times P_h$.

5.4 *Numerical results*

We applied the optimal Petrov-Galerkin method to the mild-weak variational formulation (3.7) of the convection-diffusion problem (5.1) on $\Omega = (0,1)$ or $\Omega = (0,1)^2$, with uniform partitions Ω_h into subintervals of length h, or into isosceles right angled triangles with legs of length h and hypothenuses parallel to the vector $(1,1)$, respectively.

As motivated in Subsection 5.3, we selected the following options: We took Γ_- and Γ_+ as the in- and outflow boundary, which fixes the definitions of the last factors of both $U = L_2(\Omega)^n \times H^1_0(\Omega) \times H^{-\frac{1}{2}}(\partial\Omega_h^\circ)$ and $V = L_2(\Omega)^n \times H_{0,\Gamma_+}(\Omega_h)$. We equipped $H_{0,\Gamma_+}(\Omega_h)$ with the ε-dependent norm from (5.10), where $\mu(\varepsilon) = \varepsilon$. To set up the mixed formulation, we factorized $\mathbf{A}(\varepsilon) = \varepsilon I = \mathbf{A}_1(\varepsilon)\mathbf{A}_2(\varepsilon)$, where $\mathbf{A}_1(0) = \mathbf{A}_2(0) = 0$. In particular, we took $\mathbf{A}_1(\varepsilon) = \varepsilon^{1/3} I$, $\mathbf{A}_2(\varepsilon) = \varepsilon^{2/3} I$ for $n = 1$, and $\mathbf{A}_1(\varepsilon) = \mathbf{A}_2(\varepsilon) = \varepsilon^{1/2} I$ for $n = 2$, respectively. These choices for $\mathbf{A}_i(\varepsilon)$ turned out to give approximately the best results.

Finally, we selected the trial space as in (5.14), where here we took $\mathbf{Q}_h$ to be the lowest order Raviart-Thomas space w.r.t. Ω_h for $n = 2$, or the space of continuous piecewise linears w.r.t Ω_h for $n = 1$, respectively, and P_h to be the space of continuous piecewise linears w.r.t. Ω_h, zero at $\partial\Omega$. In all but the last experiment, we took W_h to be the space of continuous piecewise linears w.r.t. Ω_h, zero at $\partial\Omega_-$. Note that the free variables in the resulting flux space are all supported near the complement of the inflow boundary. In the last experiment we will take W_h to be the space of continuous piecewise quadratics w.r.t. Ω_h, zero at $\partial\Omega_-$, meaning that for the approximation of the flux, we will add an additional free quadratic bubble, multiplied with $\mathbf{b}\cdot\mathbf{n}$, associated to each interior edge.

With these choices we satisfy the conditions formulated in §5.3 for having an optimal Petrov-Galerkin solver that also applies in the limit case of the pure convection problem. The latter can be viewed as a necessary condition for having a robust solver, i.e., a solver that does not loose convergence or becomes increasingly costly for $\varepsilon \downarrow 0$.

The solutions of the boundary value problems (5.12) on the elements $K \in \Omega_h$, that determine the optimal test functions, could be solved analytically for $n = 1$. As shown at the end of the previous subsection, for $n = 2$ and W_h being the space of continuous piecewise linears, the boundary value problems (5.12) could be solved analytically for any K with $\partial K \cap (\partial\Omega \setminus \Gamma_-) \neq \emptyset$. For the remaining K, the solutions of these boundary value problems were replaced by their Galerkin approximations from the space of cubics on K that vanish at $\partial K \cap \Gamma_+$. For W_h being the space of continuous piecewise quadratics, the solutions of the boundary value problems for the optimal test functions corresponding to the additional quadratic bubbles were replaced by their Galerkin approximations from again the space of cubics on K that, when $K \cap \Gamma_+ \neq \emptyset$, vanish at $\partial K \cap \Gamma_+$.

Both the problems that define the optimal test functions, and the resulting Petrov-Galerkin discretizations were solved with the built-in `matlab` solver. In doing so, we never encountered instabilities due to ill-conditioning.

For $n = 1$, and with $\mathbf{b} = 1$, and $f(x) = x$, in Figure 3 the $L_2(0,1)$-errors in u_h vs. $1/h$ are given for various ε, and the exact and approximate solutions u and u_h are shown for $\varepsilon = 10^{-4}$ and $h = \frac{1}{16}$. Note the large improvement compared to the results from Figure 2. For any *fixed* $\varepsilon > 0$, the error in u_h in $L_2(0,1)$ appears to be $\mathcal{O}(h^2)$. Note that in the current setting of not having an independent flux variable, we do not have a proof of this optimal error estimate in $L_2(0,1)$, cf. Remark 4.2.

For $n = 2$, we considered three test problems. In the first problem, $\mathbf{b} = [2\ 1]^\top$, and the right-hand side f is prescribed such that the exact solution is

$$u(x,y) = [x + (e^{b_1 x/\varepsilon} - 1)/(1 - e^{b_1/\varepsilon})] \cdot [y + (e^{b_1 y/\varepsilon} - 1)/(1 - e^{b_2/\varepsilon})], \tag{5.19}$$

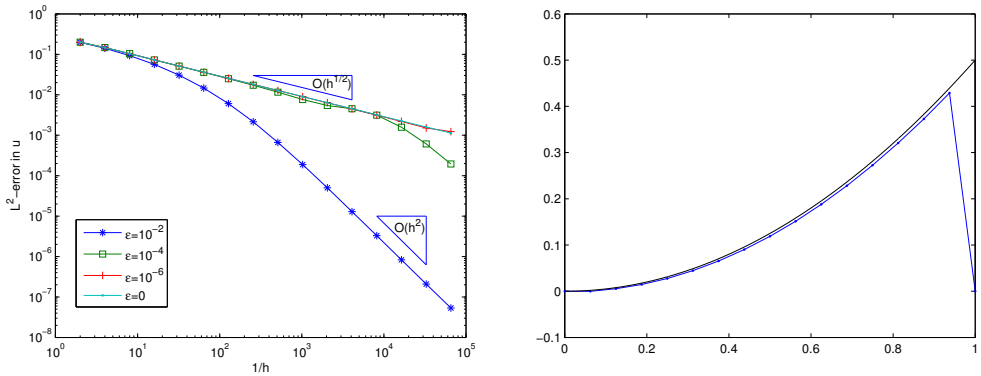

FIG. 3. $L_2(0,1)$-error in u_h vs. $1/h$ in the "robust" optimal Petrov-Galerkin approximation of the mild-weak variational formulation of the one-dimensional convection-diffusion equation for various ε (left); and the exact and approximate solutions u and u_h for $\varepsilon = 10^{-4}$ and $h = \frac{1}{16}$ (right).

which has typical boundary layers at the top and right outflow boundaries. In Figure 4, the $L_2((0,1)^2)$-errors in u_h vs. $1/h$ are given for various ε, and the approximate solution u_h for $\varepsilon = 10^{-6}$ and $h = \frac{1}{16}$ is shown. As in the one-dimensional case, the boundary layer is captured inside the elements that have

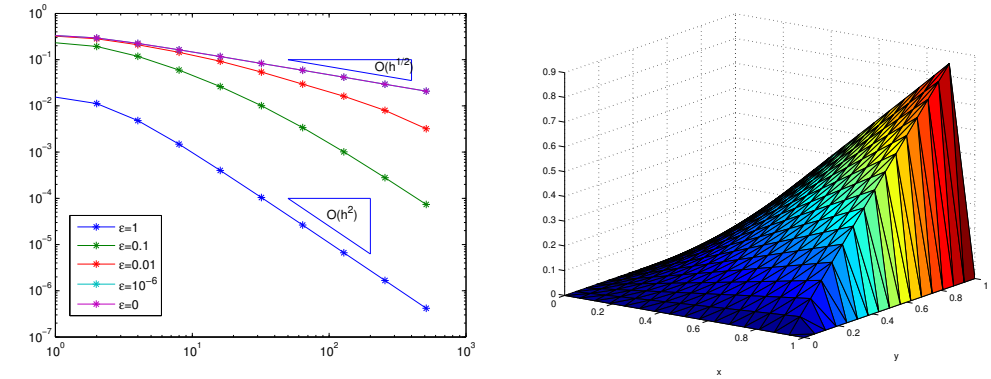

FIG. 4. $L_2((0,1)^2)$-error in u_h vs. $1/h$ for the "robust" optimal Petrov-Galerkin approximation of the mild-weak variational formulation of the two-dimensional convection-diffusion equation with $\mathbf{b} = [2\ 1]^\top$, f such that u is as in (5.19), and various ε (left); and the approximate solution u_h for $\varepsilon = 10^{-6}$ and $h = \frac{1}{16}$ (right).

non-empty intersection with the outflow boundary, and no oscillations occur.

In the second two-dimensional problem, $\mathbf{b} = [1\ 1]^\top$, and $f(x,y) = 1-x$ for $y > x$, and $f(x,y) = 0$ for $y < x$. The exact solution for $\varepsilon = 0$ is $u^0(x,y) = x - \frac{1}{2}x^2$ for $y > x$ and $u^0(x,y) = 0$ otherwise. The solution has an internal layer that is aligned with the grid and, for $\varepsilon > 0$, boundary layers at the outflow boundary. In Figure 5, the approximate solution u_h for $\varepsilon = 10^{-6}$ and $h = \frac{1}{16}$ is shown.

The behavior at the outflow boundary is similar as with the previous test problem, and the internal

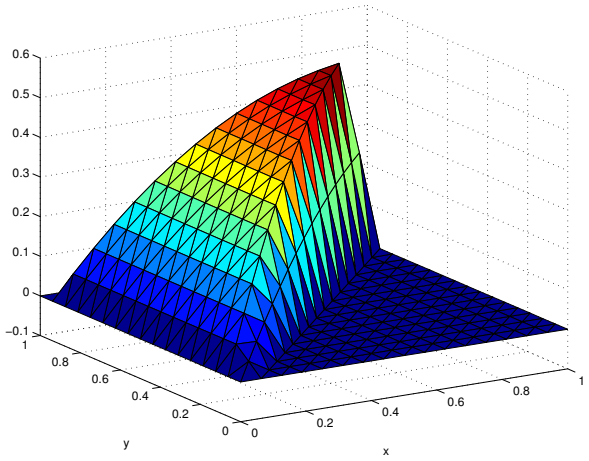

FIG. 5. The approximate solution u_h for $h = \frac{1}{16}$ for the "robust" optimal Petrov-Galerkin approximation of the mild-weak variational formulation of the two-dimensional convection-diffusion equation for $\varepsilon = 10^{-6}$, $\mathbf{b} = [1\ 1]^\top$, and $f(x,y) = 1-x$ for $y > x$, and $f(x,y) = 0$ for $y < x$.

layer is captured inside a strip that has a width of two elements.

In the third and last two-dimensional problem, $\mathbf{b} = [2\ 1]^\top$ and $f(x,y) = 1-x$ for $y > \frac{1}{2}x + \frac{1}{4}$, and $f(x,y) = 0$ otherwise. The exact solution for $\varepsilon = 0$ is $u^0(x,y) = \frac{1}{2}x - \frac{1}{4}x^2$ for $y > \frac{1}{2}x + \frac{1}{4}$ and $u^0(x,y) = 0$ otherwise. The solution has an internal layer that is not aligned with the grid and, for $\varepsilon > 0$, boundary layers at the outflow boundary. In Figure 6, the approximate solution u_h for $\varepsilon = 10^{-6}$ and $h = \frac{1}{16}$ is shown. Again, the behavior at the outflow boundary is similar as in the first two-dimensional test problem. In this example, however, we observe a "smearing" of the internal layer, as well as, unlike as with the other examples, some under or overshoot at both sides of this layer.

For this reason, we repeated the experiment where we replaced the space W_h of continuous piecewise linears by the space of continuous piecewise quadratics, meaning that we enriched the trial space for the flux by a quadratic bubble for each interior edge. In Figure 6, the resulting approximate solution u_h for $\varepsilon = 10^{-6}$ and $h = \frac{1}{32}$ is shown. In this case, the "numerical layer" has been sharpened to a width of approximately 5-6 elements, but the amplitude of the under and overshoot has not been reduced. It seems also not to be reduced by further mesh refinements, and the same under and overshoots are visible for $\varepsilon = 0$.

We infer that this oscillation is not caused by an unstable discretization, but that it is an instance of a Gibbs-type phenomenon, caused by the approximation of a discontinuous function by continuous piecewise linears. A similar oscillation can already be observed with the *best* L_2 -approximation of a smooth function on $\mathbb{R}$ that has a discontinuity at a non-dyadic point by continuous piecewise linears w.r.t. uniform partitions generated by dyadic refinements.

The oscillation along the internal layer in Figure 7 is directed perpendicularly to the flow direction $\mathbf{b}$. Potential other Gibbs type oscillations, e.g. at the outflow boundary, were prevented by the simultaneous use of u_h as approximation for u, and as an ingredient in the approximation for the flux θ, cf. Remark 5.3.

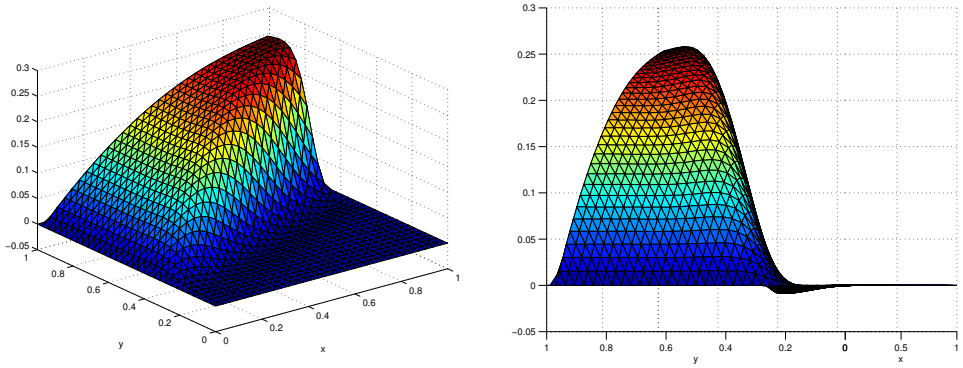

FIG. 6. The approximate solution u_h from two perspectives for $h = \frac{1}{32}$ for the "robust" optimal Petrov-Galerkin approximation of the mild-weak variational formulation of the two-dimensional convection-diffusion equation for $\varepsilon = 10^{-6}$, $\mathbf{b} = [2\ 1]^\top$, and $f(x,y) = 1-x$ for $y > \frac{1}{2}x + \frac{1}{4}$, and $f(x,y) = 0$ otherwise.

Oscillations in the approximation for θ in the direction of the flow are penalized by the norm on the flux space that for $\varepsilon = 0$ reads as the norm on $H(\mathbf{b}; \partial\Omega_h^\circ)$.

Finally, for comparison we repeated the second and third two-dimensional tests with the optimal Petrov-Galerkin method of the mild variational formulation of the mixed system, i.e., we solved the usual first order least squares problem (3.13) with trial space $\mathbf{Q}_h \times P_h$, i.e., the lowest order Raviart-Thomas space for $\boldsymbol{\sigma}_h$, and the space of continuous piecewise linears, zero at $\partial\Omega$, for u_h. The u_h-component of the solutions are illustrated in Figure 8.

6. Conclusion

We studied a mild-weak variational formulation of second order elliptic boundary value problems in mixed form constructed by piecewise integrating by parts one of the two equations in the system w.r.t. a partition of the domain into cells. It was shown that the variational formulation is well-posed, uniformly in the partition. We applied a Petrov-Galerkin discretization with optimal test space, or equivalently, minimized the residual over a given trial space. The required optimal test functions can be found by solving local boundary problems on the individual cells. Optimal error estimates in the natural norm were demonstrated, as well as, using duality arguments, in a weaker norm. Other than with the ultra-weak formulation studied in Demkowicz & Gopalakrishnan (2011b), which inspired this work, these optimal error estimates are valid under minimal regularity assumptions on the solution together with an additional smoothness assumption on the right-hand side, which is usually harmless.

In the second part of this paper, we applied our optimal Petrov-Galerkin method to convection-dominated convection-diffusion problems. Although for such problems least squares methods are more stable than the standard Galerkin discretization, not necessarily they give satisfactory results for small diffusion, as we illustrated with some one-dimensional numerical experiments. Generally, the operator associated to the variational form has an unbounded inverse in the convective limit, meaning that for small diffusion some error components are hardly measured in the residual, and so are hardly reduced in the least squares minimization.

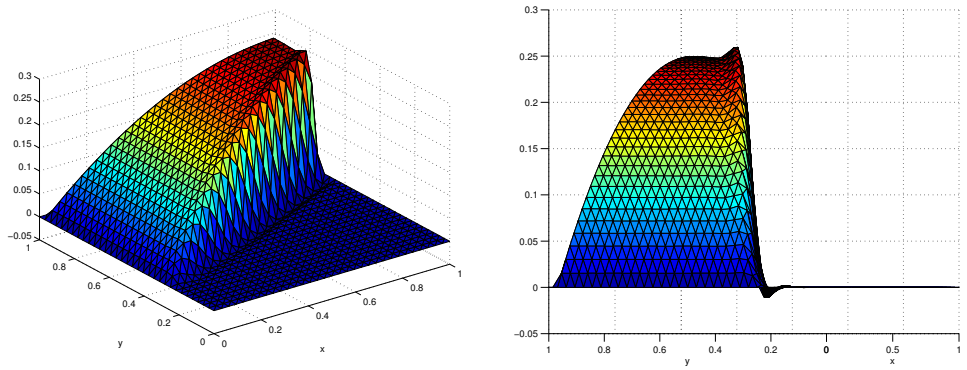

FIG. 7. Illustration as in Figure 6, but now with the trial space for the flux enriched with quadratic bubbles.

We used the available freedom in the mild-weak variational formulation, and in its optimal Petrov-Galerkin discretization, to construct a discretization that in the convective limit is an optimal Petrov-Galerkin discretization of a well-posed variational formulation of the convection problem. Numerical results show that the method performs very well for convection dominated problems.

For any fixed diffusion term, the optimal error estimates apply, whereas for the pure convection problem we showed an error estimate that is suboptimal, due to the essential boundary conditions incorporated in our trial space at the whole of the boundary. Error estimates that are explicit in both the mesh-size and the diffusion term are still lacking.

The approximation of internal or boundary layers with *continuous* finite elements has the disadvantage that oscillations easily occur, even if one could realize best L_2 approximations. For this reason, we plan to investigate whether a similar approach can be applied with discontinuous elements. Other open research topics include a posteriori error estimators and adaptivity.

REFERENCES

BARRETT, J. W. & MORTON, K. W. (1984) Approximate symmetrization and Petrov-Galerkin methods for diffusion-convection problems. *Comput. Methods Appl. Mech. Engrg.*, **45**.

BEN BELGACEM, F. (1999) The mortar finite element method with Lagrange multipliers. *Numer. Math.*, **84**, 173–197.

BOTTASSO, C., MICHELETTI, S. & SACCO, R. (2002) The discontinuous Petrov-Galerkin method for elliptic problems. *Comput. Methods Appl. Mech. Engrg.*, **191**, 3391–3409.

BRAMBLE, J. H., LAZAROV, R. D. & PASCIAK, J. E. (1997) A least-squares approach based on a discrete minus one inner product for first order systems. *Math. Comp.*, **66**, 935–955.

CAUSIN, P. & SACCO, R. (2005) A discontinuous Petrov-Galerkin method with Lagrangian multipliers for second order elliptic problems. *SIAM J. Numer. Anal.*, **43**, 280–302 (electronic).

CHAN, J., HEUER, N., BUI-THANH, T. & DEMKOWICZ, L. (2012) Robust DPG method for convection-dominated diffusion problems II: a natural inflow condition. *ICES Report* 12-21. University of Texas at Austin.

COHEN, A., DAHMEN, W. & WELPER, G. (2012) Adaptivity and variational stabilization for convection-diffusion equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, **46**, 1247–1273.

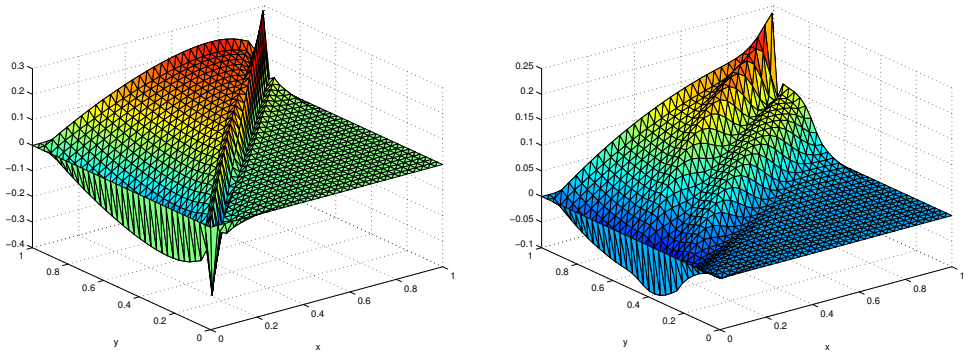

FIG. 8. The solution u_h for $h = \frac{1}{32}$ of the usual first order least squares formulation of the two-dimensional convection-diffusion equation for $\varepsilon = 10^{-6}$ for either $\mathbf{b} = [1\ 1]^\top$, and $f(x,y) = 1-x$ for $y > x$, and $f(x,y) = 0$ for $y < x$ (left), cf. Figure 5; or $\mathbf{b} = [2\ 1]^\top$, and $f(x,y) = 1-x$ for $y > \frac{1}{2}x + \frac{1}{4}$, and $f(x,y) = 0$ otherwise (right), cf. Figures 6 and 7.

DAHMEN, W., HUANG, C., SCHWAB, C. & WELPER, G. (2012) Adaptive Petrov-Galerkin methods for first order transport equations. *SIAM J. Numer. Anal.*, **50**, 2420–2445.

DE STERCK, H., MANTEUFFEL, T., MCCORMICK, S. & OLSON, L. (2004) Least-squares finite element methods and algebraic multigrid solvers for linear hyperbolic PDEs. *SIAM J. Sci. Comput.*, **26**, 31–54.

DEMKOWICZ, L. & GOPALAKRISHNAN, J. (2011a) Analysis of the DPG method for the Poisson equation. *SIAM J. Numer. Anal.*, **49**, 1788–1809.

DEMKOWICZ, L. & GOPALAKRISHNAN, J. (2011b) A class of discontinuous Petrov-Galerkin methods. II. Optimal test functions. *Numer. Methods Partial Differential Equations*, **27**, 70–105.

DEMKOWICZ, L. & HEUER, N. (2011) Robust DPG method for convection-dominated diffusion problems. *ICES Report* 11-33. University of Texas at Austin.

GOPALAKRISHNAN, J. & QIU, W. (2012) An analysis of the practical DPG method. *Technical Report*. To appear in *Math. Comp.*

GRISVARD, P. (1985) *Elliptic problems in nonsmooth domains*. Monographs and Studies in Mathematics, vol. 24. Boston, MA: Pitman (Advanced Publishing Program), pp. xiv+410.

SCOTT, L. R. & ZHANG, S. (1990) Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Math. Comp.*, **54**, 483–493.

STEVENSON, R. (2013) First-order system least squares with inhomogeneous boundary conditions. *IMA. J. Numer. Anal.*