MATHEMATICS OF COMPUTATION
Volume 87, Number 311, May 2018, Pages 1051–1082
http://dx.doi.org/10.1090/mcom/3242
Article electronically published on September 7, 2017

ON THE STABILITY OF DPG FORMULATIONS OF TRANSPORT EQUATIONS

D. BROERSEN, W. DAHMEN, AND R. P. STEVENSON

Abstract. In this paper we formulate and analyze a Discontinuous Petrov-Galerkin formulation of linear transport equations with variable convection fields. We show that a corresponding *infinite dimensional* mesh-dependent variational formulation, in which besides the principal field its trace on the mesh skeleton is also an unknown, is uniformly stable with respect to the mesh, where the test space is a certain product space over the underlying domain partition.

Our main result then states the following. For piecewise polynomial trial spaces of degree m, we show under mild assumptions on the convection field that piecewise polynomial test spaces of degree $m+1$ over a refinement of the primal partition with uniformly bounded refinement depth give rise to uniformly (with respect to the mesh size) stable Petrov-Galerkin discretizations. The partitions are required to be shape regular but need not be quasi-uniform. An important startup ingredient is that for a constant convection field one can identify the exact optimal test functions with respect to a suitably modified but uniformly equivalent broken test space norm as piecewise polynomials. These test functions are then varied towards simpler and stably computable near-optimal test functions for which the above result is derived via a perturbation analysis. We conclude indicating some consequences of the results that will be treated in forthcoming work.

1. Introduction

There has been a recent vibrant development of the so-called *Discontinuous Petrov-Galerkin* (DPG) method, initiated and developed mainly by L. Demkowicz and J. Gopalakrishnan; see e.g. [DG11, GQ14]. The general underlying methodology aims, in particular, at an improved treatment of problem classes that are, roughly speaking, much less understood than classical second order elliptic problems. Of course, "improved" leaves much room for interpretation, but for us, predominant aspects are the following:

(i) Ideally, even though the original problem may be nonsymmetric or indefinite, the arising system matrices are symmetric positive definite and sparse,

Received by the editor October 7, 2015 and, in revised form, September 6, 2016 and November 2, 2016.

2010 *Mathematics Subject Classification.* Primary 65N12, 65N30, 35A15, 35F05.

Key words and phrases. Discontinuous Petrov-Galerkin formulation of transport equations, optimal and near-optimal test spaces, stability.

The first author was supported by the Netherlands Organization for Scientific Research (NWO) under contract no. 613.001.109.

The second author was supported in part by the DFG SFB-Transregio 40, by the DFG Research Group 1779, and the Excellence Initiative of the German Federal and State Governments.

so that one has a chance to keep the computational complexity proportional to the problem size.

(ii) Ideally, the method is based on a DG-type variational formulation that establishes a tight relation between errors and residuals.

We emphasize that we mean in (ii) the *outer* residual, i.e., the residual in a full infinite dimensional space where it is well defined. Once a suitable topology for this space is identified such a residual can be used as a rigorous foundation for deriving error indicators that could steer adaptive techniques. Being able to do this beyond the class of elliptic problems is a major motivation for this paper. Specifically, the central objective of this paper is to discuss (i) and (ii) for a class of *linear transport equations* with possibly *variable* convection field.

We explain next the relevance of (i), (ii) for us in more detail, relate our findings to the state of the art, and lay out the objectives of the present work.

1.1. **Conceptual background and motivation.** Both issues (i), (ii) above rely crucially on the notion of *optimal test bases*. The key underlying idea is easily described in an abstract framework and has been presented in the literature in different variants for different purposes [BM84, BS15, DG11, GQ14, DSMMO04, DHSW12, DPW14]. To explain this let $\mathbb{U}, \mathbb{V}$ denote Hilbert spaces over $\mathbb{R}$, endowed with norms $\|\cdot\|_{\mathbb{U}}, \|\cdot\|_{\mathbb{V}}$, respectively, and assume that $b(\cdot,\cdot) : \mathbb{U}\times\mathbb{V}\to\mathbb{R}$ is a continuous bilinear form. Given $f\in\mathbb{V}'$, the normed dual of $\mathbb{V}$, endowed with the norm

$$\|w\|_{\mathbb{V}'} := \sup_{v\in\mathbb{V}} \frac{|w(v)|}{\|v\|_{\mathbb{V}}},$$

consider the variational problem

$$b(u,v) = f(v), \quad v\in\mathbb{V}. \tag{1.1}$$

Since the form $b(\cdot,\cdot)$ is continuous, i.e.,

$$\|\mathcal{B}\| := \sup_{\|v\|_{\mathbb{V}}\le 1} \sup_{\|w\|_{\mathbb{U}}\le 1} b(w,v) < \infty,$$

the operator $\mathcal{B}:\mathbb{U}\to\mathbb{V}'$, defined by $(\mathcal{B}w)(v) = b(w,v)$, $w\in\mathbb{U}, v\in\mathbb{V}$, is continuous and (1.1) is equivalent to the operator equation

$$\mathcal{B}u = f. \tag{1.2}$$

Its unique solvability is well known to be equivalent to the validity of the inf-sup conditions

$$\inf_{w\in\mathbb{U}}\sup_{v\in\mathbb{V}} \frac{b(w,v)}{\|w\|_{\mathbb{U}}\|v\|_{\mathbb{V}}} \ge \beta, \quad \inf_{v\in\mathbb{V}}\sup_{w\in\mathbb{U}} \frac{b(w,v)}{\|w\|_{\mathbb{U}}\|v\|_{\mathbb{V}}} \ge \beta, \tag{1.3}$$

for some positive β, i.e., $\mathcal{B}\in\mathcal{L}\mathrm{is}(\mathbb{U},\mathbb{V}')$, where $\mathcal{L}\mathrm{is}(\mathbb{X},\mathbb{Y})$ denotes the collection of norm-isomorphisms from a Hilbert space $\mathbb{X}$ onto a Hilbert space $\mathbb{Y}$. Moreover, denoting by $\mathcal{L}(\mathbb{X},\mathbb{Y})$ the space of bounded linear operators from the normed linear space $\mathbb{X}$ to the normed linear space $\mathbb{Y}$, it is well known that $\|\mathcal{B}^{-1}\|_{\mathcal{L}(\mathbb{V}',\mathbb{U})}\le\beta^{-1}$. Thus, the *condition number* of $\mathcal{B}\in\mathcal{L}\mathrm{is}(\mathbb{U},\mathbb{V}')$,

$$\kappa_{\mathbb{U},\mathbb{V}'}(\mathcal{B}) := \|\mathcal{B}\|_{\mathcal{L}(\mathbb{U},\mathbb{V}')}\|\mathcal{B}^{-1}\|_{\mathcal{L}(\mathbb{V}',\mathbb{U})},$$

satisfies

$$\kappa_{\mathbb{U},\mathbb{V}'}(\mathcal{B}) \le \|\mathcal{B}\|/\beta,$$

i.e., the smaller $\|\mathcal{B}\|$ and the larger β, the better. In particular, since in these terms $\|w\|_{\mathbb{U}} \leq \beta^{-1}\|\mathcal{B}w\|_{\mathbb{V}'}$, $\|\mathcal{B}w\|_{\mathbb{V}'} \leq \|\mathcal{B}\|\|w\|_{\mathbb{U}}$, we have for any approximation $\bar{u}$ to the solution u of (1.1) the error-residual relation

$$\|\mathcal{B}\|^{-1}\|f - \mathcal{B}\bar{u}\|_{\mathbb{V}'} \leq \|u - \bar{u}\|_{\mathbb{U}} \leq \beta^{-1}\|f - \mathcal{B}\bar{u}\|_{\mathbb{V}'}. \tag{1.4}$$

Of course, the larger $\kappa_{\mathbb{U},\mathbb{V}'}(\mathcal{B})$ is the harder it is for a numerical method based on the above variational formulation to perform well. Moreover, the residual in $\mathbb{V}'$ then does not provide accurate information about the error in $\mathbb{U}$.

In general one may have to face two types of obstructions: first, $\kappa_{\mathbb{U},\mathbb{V}'}(\mathcal{B})$—although finite—could be very large. A typical example is a convection dominated convection diffusion problem for $\mathbb{U} = \mathbb{V} = H_0^1(\Omega)$. Fixing $\|\cdot\|_{\mathbb{U}}$ and appropriately varying $\|\cdot\|_{\mathbb{V}}$, or vice versa, may lead to a different variational formulation with a much smaller condition number, ideally even equal to one; see [DHSW12]. The prize to be paid is that one has to accept that trial and test space (already on the infinite dimensional level) are different. This is the second obstruction, namely having to deal with an *asymmetric* variational formulation, $\mathbb{U} \neq \mathbb{V}$, so that the uniform discrete stability of projected versions of (1.1) is no longer taken for granted even though the inf-sup constant β in (1.3) may be close to one.

The present paper is concerned with the second issue, starting with a well-conditioned infinite dimensional variational formulation and later for a class of transport equations. Then, given a (finite dimensional) *trial space* $\mathbb{U}^h \subset \mathbb{U}$ we wish to find a *test space* $\mathbb{T}^h \subset \mathbb{V}$ that inherits the stability (1.3) of the infinite dimensional problem (for a positive constant possibly smaller than β, but h-independent), and therefore deserves to be called (uniformly) (*near-*)*optimal*. To identify such a near-optimal test space, notice first that the *trial-to-test map* $\mathcal{T} \in \mathcal{L}\mathrm{is}(\mathbb{U}, \mathbb{V})$, defined by

$$\langle \mathcal{T}u, v\rangle_{\mathbb{V}} = b(u; v) \quad (u \in \mathbb{U},\ v \in \mathbb{V}), \tag{1.5}$$

yields the *supremizer* in the first relation of (1.3), i.e.,

$$\|\mathcal{T}u\|_{\mathbb{V}} = \sup_{v\in\mathbb{V}} \frac{b(u,v)}{\|v\|_{\mathbb{V}}}, \tag{1.6}$$

which means

$$\|\mathcal{T}u\|_{\mathbb{V}}^2 = b(u, \mathcal{T}u). \tag{1.7}$$

Therefore, the (truly) optimal test space for a given subspace $\mathbb{U}^h \subset \mathbb{U}$ is

$$\mathcal{T}(\mathbb{U}^h) = \{\mathcal{T}u^h : u^h \in \mathbb{U}^h\}, \tag{1.8}$$

in the sense that the Petrov-Galerkin scheme: find $u_h \in \mathbb{U}_h$ such that

$$b(u_h, v_h) = f(v_h), \quad v_h \in \mathcal{T}(\mathbb{U}^h), \tag{1.9}$$

is uniquely solvable and the corresponding finite dimensional operator has at most the same condition number as the infinite dimensional problem (1.1). Moreover, (1.9) is easily seen to form the normal equations for minimizing the residual $\|f - \mathcal{B}w\|_{\mathbb{V}'}$ over $\mathbb{U}^h$, i.e.,

$$u^h = \operatorname*{argmin}_{\bar{u}^h\in\mathbb{U}^h} \|f - \mathcal{B}\bar{u}^h\|_{\mathbb{V}'}. \tag{1.10}$$

Denoting by $\mathcal{R}_{\mathbb{U}} \in \mathcal{L}\mathrm{is}(\mathbb{U}, \mathbb{U}')$ the *Riesz map* defined by

$$\langle z, w\rangle_{\mathbb{U}} = (\mathcal{R}_{\mathbb{U}}z)(w), \quad z, w \in \mathbb{U}, \tag{1.11}$$

we have, of course, $\mathcal{T} = \mathcal{R}_{\mathbb{V}}^{-1}\mathcal{B} = \mathcal{R}_{\mathbb{V}'}\mathcal{B}$. Hence, the application of $\mathcal{T}$ amounts to solving an infinite dimensional Galerkin problem in $\mathbb{V}$. Thus, for each basis function $\phi \in \mathbb{U}^h$, finding the corresponding test-basis function $\psi = \mathcal{T}\phi$, would require solving an infinite dimensional variational problem, possibly even of the same complexity as the one for solving (1.1).

A natural idea propagated in many works (see e.g. [DG11, CDW12, DHSW12, BS15]) is to reduce this $\mathbb{V}$-projection to a finite dimensional subspace $\mathbb{V}^h \subset \mathbb{V}$ which we refer to as the *test-search-space*. Specifically, this amounts to replacing $\mathcal{T}$ by the mapping $\mathcal{T}^h = \mathcal{T}^{\mathbb{V}^h} \in \mathcal{L}(\mathbb{U}, \mathbb{V}^h)$, defined by

$$\langle \mathcal{T}^h u, v^h\rangle_{\mathbb{V}} = b(u; v^h) \quad (u \in \mathbb{U},\ v \in \mathbb{V}^h), \tag{1.12}$$

whose existence is guaranteed by Riesz' representation theorem. Given a closed linear trial space $\mathbb{U}^h \subset \mathbb{U}$, and denoting by $\mathcal{P}_{\mathbb{V}^h}$ the $\mathbb{V}$-orthogonal projection onto $\mathbb{V}^h$, defined by $\langle \mathcal{P}_{\mathbb{V}^h} v, z\rangle_{\mathbb{V}} = \langle v, z\rangle_{\mathbb{V}}$, $v \in \mathbb{V}$, $z \in \mathbb{V}^h$, we see that $\mathcal{T}^h = \mathcal{P}_{\mathbb{V}^h} \circ \mathcal{T}$. The range of its restriction to $\mathbb{U}^h$,

$$\mathcal{T}^h(\mathbb{U}^h) = (\mathcal{P}_{\mathbb{V}^h} \circ \mathcal{T})(\mathbb{U}^h),$$

known as the *projected optimal test space*, will now be used as test space in the *Petrov-Galerkin* problem of finding $\tilde{u}^h \in \mathbb{U}^h$ such that

$$b(\tilde{u}^h; v^h) = f(v^h) \quad (v^h \in \mathcal{T}^h(\mathbb{U}^h)). \tag{1.13}$$

Our key requirement on $\mathbb{V}^h$ is that

$$\gamma^h := \inf_{0 \neq w^h \in \mathbb{U}^h} \sup_{0 \neq v^h \in \mathbb{V}^h} \frac{b(w^h; v^h)}{\|w^h\|_{\mathbb{U}} \|v^h\|_{\mathbb{V}}} \geq \gamma > 0 \tag{1.14}$$

holds uniformly in h. Then the (projected optimal) test space $\mathcal{T}^h(\mathbb{U}^h)$ is near-optimal. In particular, the generalized Céa lemma shows that

$$\|u - \tilde{u}^h\|_{\mathbb{U}} \leq \frac{\|\mathcal{B}\|_{\mathcal{L}(\mathbb{U},\mathbb{V}')}}{\gamma^h} \inf_{w^h \in \mathbb{U}^h} \|u - w^h\|_{\mathbb{U}}; \tag{1.15}$$

see e.g. [GQ14, Thm. 2.1], [BS14, Prop. 2.3], [CDW12, DHSW12].

Recall that a necessary condition for realizing our initial objective (i) of linearly scaling computational complexity is that

$$\dim \mathbb{V}^h \eqsim \dim \mathbb{U}^h, \tag{1.16}$$

uniformly in h.

Note, however, that even when (1.16) holds, determining the corresponding projected optimal test space still requires solving for each basis function a discrete problem which, generally, has the same size as the corresponding Petrov-Galerkin problem itself.

Therefore a central objective is to keep also the cost for *computing* $\mathcal{T}^h(\mathbb{U}^h)$ under control, which is the primary focus of this paper. One strategy is to *localize* the computation of the projected optimal test functions. As advocated by Demkowicz and Gopalakrishnan in several of their works, this localization can be achieved by replacing the "original" formulation (1.1) from the start by a mesh-dependent *Discontinuous Petrov-Galerkin* formulation

$$b_h(U, v) = (\mathcal{B}_h U)(v) = f(v), \quad v \in \mathbb{V}; \tag{1.17}$$

see, e.g., [DG11]. Here, the "new" unknown U may now involve in addition to the original field u also a "skeleton-component" that lives in the union $\partial\Omega_h$ of cell interfaces of the underlying mesh Ω_h. For smooth solutions this skeleton-component agrees with the traces of u on $\partial\Omega_h$ but these traces may not a priori exist for all elements in the function space for u. Choosing now the (infinite dimensional) test space as a "broken" space

$$\mathbb{V} := \prod_{K\in\Omega_h} \mathbb{V}_K, \quad \|v\|_{\mathbb{V}}^2 := \sum_{K\in\Omega_h} \|v\|_{\mathbb{V}_K}^2, \tag{1.18}$$

the trial-to-test mapping $\mathcal{T} : \mathbb{U} \to \mathbb{V}$ indeed localizes, i.e., for $b_h(u,v) = \sum_{K\in\Omega_h} b_K(u,v)$ we have

$$\mathcal{T}u = \sum_{K\in\Omega_h} \mathcal{T}_K u, \quad \text{where} \quad \langle \mathcal{T}_K u, v\rangle_{\mathbb{V}_K} = b_K(u,v), \quad v \in \mathbb{V}_K. \tag{1.19}$$

One now faces two main issues:

(I) Imposing the structure (1.18) on the test space, it is not clear that the *infinite dimensional* (new) variational formulation (1.17) is well-posed. More precisely, one has to establish *uniform* inf-sup stability with respect to a given family of partitions Ω_h with decreasing mesh size parameter h.

(II) For a given finite dimensional trial space $\mathbb{U}^h$ associated with Ω_h, one still has to find a *finite dimensional* test search space

$$\mathbb{V}^h = \prod_{K\in\Omega_h} \mathbb{V}_K^h,$$

that satisfies (1.14).

Regarding our introductory issues (i) and (ii), realizing a linear scaling of the computational work for the uniformly stable Petrov-Galerkin problems one would need to assure that $\dim \mathbb{V}^h \lesssim \dim \mathbb{U}^h$, uniformly in h. This would be the case if one were able to assert that for some fixed $M \in \mathbb{N}$,

$$\dim\left(\mathbb{V}_K^h\right) \le M, \quad h \to 0, \tag{1.20}$$

suffices to warrant the desired uniform inf-sup stability, and as a consequence, the desired rigorous error-residual relation (1.4).

Cases where these desiderata have been rigorously established include the Poisson and linear elasticity problems (in [GQ14]), and the Maxwell equations (in [CDG16]). After having established (I), in those cases (II) (with (1.20)) was proven by constructing a suitable "Fortin" projector in "broken" H^1, $H(\mathrm{div})$, or $H(\mathrm{curl})$ spaces, respectively.

The topic of the current work is to obtain such results for a class of linear *transport equations* with a generally *variable convection* field $\mathbf{b}$. The main obstacle is then that the arising spaces $\mathbb{U}$ and $\mathbb{V}$ *depend* on $\mathbf{b}$. Any perturbation argument affecting the field $\mathbf{b}$, which is generally unavoidable, is therefore rather delicate. Related obstructions are expected to arise also in more general problems involving strongly dominating transport. Nevertheless, we will obtain results that are valid *uniformly* in the relative orientation of the (local) mesh and $\mathbf{b}$, and, of course, in the mesh itself. The proofs of these results and necessary prerequisites turn out to be quite elaborate.

Our motivation for investing in a rigorous stability analysis for transport equations stems in part from several envisaged applications that will be addressed in more detail in forthcoming work. This concerns, in particular, the design and analysis of rigorous adaptive methods for transport equations and, in fact, for a somewhat wider scope of problems where transport plays a dominant role such as kinetic models.

1.2. **Layout of the paper.** In Section 2 we formulate the first order linear transport equations treated in this paper. Section 3 is devoted to its variational formulation and the proof of its well-posedness, addressing the aforementioned issue (I). In Section 4 we derive and analyze optimal test functions along with their computable near-optimal counterparts culminating in the uniform stability of the DPG scheme, i.e., this section deals with issue (II).

In 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 $C \gtrsim D$.

2. Transport equation

For a bounded Lipschitz domain $\Omega \subset \mathbb{R}^n$, let $\mathbf{b} \in W_\infty^0(\operatorname{div};\Omega)$, i.e., $\mathbf{b} \in L_\infty(\Omega)^n$ with $\operatorname{div}\mathbf{b} \in L_\infty(\Omega)$. We set

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

equipped with the norm $\|u\|^2_{H(\mathbf{b};\Omega)} := \|u\|^2_{L_2(\Omega)} + \|\mathbf{b}\cdot\nabla u\|^2_{L_2(\Omega)}$.

In order to define the *characteristic, outflow, and inflow* boundary portions $\Gamma_0, \Gamma_+, \Gamma_- \subset \partial\Omega$, respectively, under the above assumptions on the velocity field $\mathbf{b}$ we use the (formal) integration-by-parts formula

$$\int_\Omega 2w\mathbf{b}\cdot\nabla w + w^2 \operatorname{div}\mathbf{b}\, d\mathbf{x} = \int_{\partial\Omega} w^2 \mathbf{b}\cdot\mathbf{n}\, d\mathbf{s},$$

to define the characteristic boundary Γ_0 as the largest measurable subset of $\partial\Omega$ such that the left-hand side vanishes for all $w \in H(\mathbf{b};\Omega)\cap C(\bar\Omega)$ that vanish on $\partial\Omega\setminus\Gamma_0$. Similarly, we set the outflow boundary Γ_+ as the largest measurable subset of $\partial\Omega\setminus\Gamma_0$ such that $\int_\Omega 2w\mathbf{b}\cdot\nabla w + w^2\operatorname{div}\mathbf{b}\, d\mathbf{x} \geq 0$ for all $w \in H(\mathbf{b};\Omega)\cap C(\bar\Omega)$ that vanish on $(\partial\Omega\setminus\Gamma_0)\setminus\Gamma_+$, and finally, we define the inflow boundary as $\Gamma_- = \partial\Omega\setminus(\Gamma_0\cup\Gamma_+)$. For continuous $\mathbf{b}$, it means that $\Gamma_0 := \{x \in \partial\Omega \colon \mathbf{b}(x)\cdot\mathbf{n}(x) = 0\}$ whenever $\mathbf{n}(x)$ is uniquely defined, and $\Gamma_\pm := \{x \in \partial\Omega \colon \pm\mathbf{b}(x)\cdot\mathbf{n}(x) > 0\}$.

For $\mathbf{b} \in W_\infty^0(\operatorname{div};\Omega)$ and $c \in L_\infty(\Omega)$, we consider the transport equation for finding $u : \Omega \to \mathbb{R}$ such that, for given $f \colon \Omega \to \mathbb{R}$ and $g \colon \Gamma_- \to \mathbb{R}$, solves

$$\left\{\begin{aligned} \mathbf{b}\cdot\nabla u + cu &= f \quad \text{on } \Omega,\\ u &= g \quad \text{on } \Gamma_-. \end{aligned}\right. \tag{2.1}$$

When $g = 0$ a first canonical variational formulation of the transport problem reads: find u, zero at Γ_-, such that

$$\int_\Omega (\mathbf{b}\cdot\nabla u + cu)v\, d\mathbf{x} = \int_\Omega fv\, d\mathbf{x} \tag{2.2}$$

holds for all smooth test functions $v \in C^\infty(\bar\Omega)$. A second variant seeks u such that

$$\int_\Omega (cv - \operatorname{div} v\mathbf{b})u\, d\mathbf{x} = \int_\Omega fv - \int_{\Gamma_-} gv\mathbf{b}\cdot\mathbf{n}\, d\mathbf{x} \tag{2.3}$$

holds for all smooth test functions v that vanish on Γ_+. Note that in the second formulation, the Dirichlet boundary condition enters as a *natural* condition, and therefore this formulation applies equally well for an inhomogeneous boundary condition on Γ_-.

Applying Cauchy-Schwarz followed by taking closures, shows that the Hilbert spaces

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

are relevant for these variational formulations. In fact, the operators

$$\mathcal{B} := u \mapsto \mathbf{b}\cdot\nabla u + cu,$$

and its *formal* adjoint

$$\mathcal{B}^* := v \mapsto cv - \operatorname{div} v\mathbf{b},$$

are obviously continuous as mappings into $L_2(\Omega)$, i.e.,

$$\mathcal{B} \in \mathcal{L}(H_{0,\Gamma_-}(\mathbf{b};\Omega), L_2(\Omega)), \quad \mathcal{B}^* \in \mathcal{L}(H_{0,\Gamma_+}(\mathbf{b};\Omega), L_2(\Omega)).$$

In addition, we *assume* that

$$\mathcal{B} \in \mathcal{L}\mathrm{is}(H_{0,\Gamma_-}(\mathbf{b};\Omega), L_2(\Omega)), \tag{2.4}$$

$$\mathcal{B}^* \in \mathcal{L}\mathrm{is}(H_{0,\Gamma_+}(\mathbf{b};\Omega), L_2(\Omega)), \tag{2.5}$$

meaning that the first (for $g = 0$) or second variational form of the 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. These assumptions are readily verified for nonzero, constant $\mathbf{b}$, but are not necessarily satisfied for every vector field $\mathbf{b}$ as, for instance, when flow curves associated to $\pm\mathbf{b}$ do not reach the boundary. Sufficient conditions for both assumptions are $\mathbf{b} \in C^1(\bar{\Omega})$ with $\mathbf{b}(x) \neq 0$ for $x \in \bar{\Omega}$, or $c - \frac{1}{2}\operatorname{div}\mathbf{b} \geq \kappa > 0$ a.e. on Ω, for some constant κ; see [DHSW12, Remark 2.2].

For completeness, note that $\mathcal{B}^*$ should not be confused with the adjoint of $\mathcal{B}$ as a mapping from $L_2(\Omega)$ onto $H_{0,\Gamma_-}(\mathbf{b};\Omega)'$, which is an isomorphism if and only if $\mathcal{B}$ is.

3. A variational formulation of the transport equation with broken test and trial spaces

In order to allow us to eventually localize the determination of the optimal test functions we follow the approach introduced by Demkowicz and Gopalakrishnan [DG11] replacing (2.3) by a *Discontinuous Galerkin* formulation. We introduce first the relevant notation. For any h from an index of mesh parameters, let Ω_h be a collection of disjoint open Lipschitz domains ("elements") such that $\bar{\Omega} = \bigcup_{K\in\Omega_h}\bar{K}$. We will refer to such an Ω_h as a partition of Ω. For each $K \in \Omega_h$, we split its boundary into characteristic inflow and outflow boundaries, i.e., $\partial K = \partial K_0 \cup \partial K_+ \cup \partial K_-$, and denote by

$$\partial\Omega_h := \bigcup_{K\in\Omega_h} \partial K \setminus \partial K_0$$

the *mesh skeleton*, i.e., the union of the noncharacteristic boundary portions of the elements.

Let us first assume that $g = 0$ referring to Remark 3.6 for $g \neq 0$. Moreover, denoting by ∇_h the piecewise gradient operator, let us introduce the spaces $H(\mathbf{b};\Omega_h) = \{v \in L_2(\Omega)\colon \mathbf{b}\cdot\nabla_h v \in L_2(\Omega)\}$, equipped with squared "broken" norm $\|v\|^2_{H(\mathbf{b};\Omega_h)} := \|v\|^2_{L_2(\Omega)} + \|\mathbf{b}\cdot\nabla_h v\|^2_{L_2(\Omega)}$, and let

$$H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h) := \{w|_{\partial\Omega_h}\colon w \in H_{0,\Gamma_-}(\mathbf{b};\Omega)\},$$

equipped with quotient norm

$$\|\theta\|_{H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)} := \inf\{\|w\|_{H(\mathbf{b};\Omega)} \colon \theta = w|_{\partial\Omega_h},\, w \in H_{0,\Gamma_-}(\mathbf{b};\Omega)\}. \tag{3.1}$$

A standard *piecewise* integration-by-parts of the transport equation (2.1) leads to the following problem:

$$\begin{cases} \text{For } \mathbb{U} := L_2(\Omega) \times H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h),\ \mathbb{V} := H(\mathbf{b};\Omega_h), \\ \text{given } f \in H(\mathbf{b};\Omega_h)', \text{find } (u,\theta) \in \mathbb{U} \text{ such that for all } v \in \mathbb{V}, \\ b_h(u,\theta;v) := \displaystyle\int_\Omega (cv - \mathbf{b}\cdot\nabla_h v - v \operatorname{div}\mathbf{b}) u\, d\mathbf{x} + \int_{\partial\Omega_h} [\![v\mathbf{b}]\!]\theta\, d\mathbf{s} = f(v). \end{cases} \tag{3.2}$$

Here we define as usual for $x \in \partial K \cap \partial K'$,

$$[\![v\mathbf{b}]\!](x) := (v\mathbf{b}|_K \cdot \mathbf{n}_K)(x) + (v\mathbf{b}|_{K'} \cdot \mathbf{n}_{K'})(x),$$

and $[\![v]\!](x) := (v\mathbf{b}|_K \cdot \mathbf{n}_K)(x)$ for $x \in \partial\Omega \cap \partial K$, and set $\mathcal{B}_h : \mathbb{U} \to \mathbb{V}'$ by

$$(\mathcal{B}_h(u,\theta))(v) := b_h(u,\theta;v).$$

The additional independent variable θ replaces the trace $u|_{\partial\Omega_h}$ which is not defined for general $u \in L_2(\Omega)$. If $f \in L_2(\Omega)$ or, equivalently, $u \in H_{0,\Gamma_-}(\mathbf{b};\Omega)$, then a reversed integration by parts shows that indeed $\theta = u|_{\partial\Omega_h}$.

Well-posedness of the variational formulation (3.2) is demonstrated in the next theorem. It is an adaptation of [BS15, Thm. 5.1] where we employ here slightly different spaces $\mathbb{U}$ and $\mathbb{V}$, and where we exhibit explicit bounds on the norms of the operator and its inverse. In [BS15], the spaces were chosen such that both θ and v vanish on Γ_+. Also the transport equation here is more general since it may contain a reaction term. For convenience we include the proof.

Next, we abbreviate $\|\mathcal{B}^{-1}\|_{\mathcal{L}(L_2(\Omega),H_{0,\Gamma_-}(\mathbf{b};\Omega))}$, $\|(\mathcal{B}^*)^{-1}\|_{\mathcal{L}(L_2(\Omega),H_{0,\Gamma_+}(\mathbf{b};\Omega))}$, $\|\operatorname{div}\mathbf{b}\|_{L_\infty(\Omega)}$, $\|c\|_{L_\infty(\Omega)}$, and $\|c - \operatorname{div}\mathbf{b}\|_{L_\infty(\Omega)}$ as $\|\mathcal{B}^{-1}\|$, $\|\mathcal{B}^{-*}\|$, $\|\operatorname{div}\mathbf{b}\|$, $\|c\|$, and $\|c - \operatorname{div}\mathbf{b}\|$, respectively. The operators $\mathcal{B}$, $\mathcal{B}^*$, induced by the conforming formulations (2.2), (2.3), should not be confused with the operators $\mathcal{B}_h$ induced by the DPG formulation.

Theorem 3.1. *Assume that* $\mathbf{b} \in W_\infty^0(\operatorname{div};\Omega)$, $c \in L_\infty(\Omega)$ *and that conditions* (2.4), (2.5) *hold. Then one has* $\mathcal{B}_h \in \mathcal{L}\mathrm{is}(\mathbb{U},\mathbb{V}')$ *with*

$$\|\mathcal{B}_h\|_{\mathcal{L}(\mathbb{U},\mathbb{V}')} \le 2 + \|\operatorname{div}\mathbf{b}\| + \|c - \operatorname{div}\mathbf{b}\|,$$

$$\|\mathcal{B}_h^{-1}\|_{\mathcal{L}(\mathbb{V}',\mathbb{U})} \le \sqrt{\|\mathcal{B}^{-*}\|^2 + \tilde{C}_{\mathcal{B}}^2},$$

where $\tilde{C}_{\mathcal{B}} := (1 + \|\mathcal{B}^{-*}\|(1 + \|c - \operatorname{div}\mathbf{b}\|))\|\mathcal{B}^{-1}\|(\|c - \operatorname{div}\mathbf{b}\| + 1)$.

Remark 3.2. As the bilinear form b_h and the operator $\mathcal{B}_h$, obviously also the spaces $\mathbb{U}$ and $\mathbb{V}$, and the solution (u,θ) depend on h, but we supress these latter dependencies in the notation.

Remark 3.3. A consequence of Theorem 3.1 is that $H(\mathbf{b};\Omega_h) \to H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)'; v \mapsto [\![v\mathbf{b}]\!]$ is surjective.

Anticipating this latter fact, we can say that the following lemma, which is the first tool for proving Theorem 3.1, provides an equivalent norm for $H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)'$. In particular, it shows that $H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)' \simeq H(\mathbf{b};\Omega_h)/H_{0,\Gamma_+}(\mathbf{b};\Omega)$.

Lemma 3.4. *For $v \in H(\mathbf{b};\Omega_h)$, one has $[\![v\mathbf{b}]\!] \in (H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h))'$ with*

$$\big(\|\mathcal{B}^{-1}\|(\|c-\operatorname{div}\mathbf{b}\|+1)\big)^{-1} \leq \frac{\|[\![v\mathbf{b}]\!]\|_{H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)'}}{\inf_{z\in H_{0,\Gamma_+}(\mathbf{b};\Omega)}\|v-z\|_{H(\mathbf{b};\Omega_h)}} \leq 1+\|\operatorname{div}\mathbf{b}\|$$

$(v \in H(\mathbf{b};\Omega_h)\setminus H_{0,\Gamma_+}(\mathbf{b};\Omega))$.

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

$$\begin{aligned}\int_{\partial\Omega_h}[\![v\mathbf{b}]\!]w\,d\mathbf{s} &= \sum_{K\in\Omega_h}\int_K \nabla v\cdot\mathbf{b}w + v(\mathbf{b}\cdot\nabla w + w\operatorname{div}\mathbf{b})\,d\mathbf{x}\\ &\leq (1+\|\operatorname{div}\mathbf{b}\|)\|v\|_{H(\mathbf{b};\Omega_h)}\|w\|_{H(\mathbf{b};\Omega)},\end{aligned} \tag{3.3}$$

showing that $\|[\![v\mathbf{b}]\!]\|_{H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)'} \leq (1+\|\operatorname{div}\mathbf{b}\|)\|v\|_{H(\mathbf{b};\Omega_h)}$. Since for $z \in H_{0,\Gamma_+}(\mathbf{b};\Omega)$ and $w \in H_{0,\Gamma_-}(\mathbf{b};\Omega)$, $\int_\Omega \nabla z\cdot\mathbf{b}w + z(\mathbf{b}\cdot\nabla w + w\operatorname{div}\mathbf{b})\,d\mathbf{x} = 0$, it follows that $\|[\![z\mathbf{b}]\!]\|_{H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)'} = 0$. This shows that for $v \in H(\mathbf{b};\Omega_h)$, $\|[\![v\mathbf{b}]\!]\|_{H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)'} \leq (1+\|\operatorname{div}\mathbf{b}\|)\inf_{z\in H_{0,\Gamma_+}(\mathbf{b};\Omega)}\|v-z\|_{H(\mathbf{b};\Omega_h)}$.

To prove the converse estimate let div_h denote the piecewise divergence operator. Given $v \in H(\mathbf{b};\Omega_h)$, let $z \in H_{0,\Gamma_+}(\mathbf{b};\Omega)$ be the solution of

$$\mathcal{B}^*z = cz - \operatorname{div}(z\mathbf{b}) = cv - \operatorname{div}_h(v\mathbf{b}),$$

whose existence is guaranteed by (2.5). From

$$c(v-z) = \operatorname{div}_h\big((v-z)\mathbf{b}\big) = (v-z)\operatorname{div}\mathbf{b} + \mathbf{b}\cdot\nabla_h(v-z), \tag{3.4}$$

we derive that

$$\|\mathbf{b}\cdot\nabla_h(v-z)\|_{L_2(\Omega)} \leq (\|c-\operatorname{div}\mathbf{b}\|)\|v-z\|_{L_2(\Omega)}. \tag{3.5}$$

By (2.4), there exists a $w \in H_{0,\Gamma_-}(\mathbf{b};\Omega)$ such that $\mathcal{B}w = \mathbf{b}\cdot\nabla w + cw = v-z$ and

$$\|w\|_{H(\mathbf{b};\Omega)} \leq \|\mathcal{B}^{-1}\|\|v-z\|_{L_2(\Omega)}. \tag{3.6}$$

From the definitions of w and z, we have

$$\begin{aligned}\|v-z\|^2_{L_2(\Omega)} &= \int_\Omega (v-z)(\mathbf{b}\cdot\nabla w + cw)\,d\mathbf{x} = \sum_{K\in\Omega_h}\int_K (v-z)(\mathbf{b}\cdot\nabla w + cw)\,d\mathbf{x}\\ &= \sum_{K\in\Omega_h}\int_K \big(\operatorname{div}\big((z-v)\mathbf{b}\big) + c(v-z)\big)w\,d\mathbf{x} + \int_{\partial K}(v-z)w\mathbf{b}\cdot\mathbf{n}_K\,d\mathbf{s}\\ &= \int_{\partial\Omega_h}[\![v\mathbf{b}]\!]w\,d\mathbf{s},\end{aligned}$$

where we have used (3.4) in the last step. Thus, invoking (3.6), we have

$$\begin{aligned}\|v-z\|^2_{L_2(\Omega)} &\leq \|[\![v\mathbf{b}]\!]\|_{H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)'}\|w\|_{H(\mathbf{b};\Omega)}\\ &\leq \|[\![v\mathbf{b}]\!]\|_{H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)'}\|\mathcal{B}^{-1}\|\|v-z\|_{L_2(\Omega)}.\end{aligned}$$

In other words $\|v-z\|_{L_2(\Omega)} \leq \|\mathcal{B}^{-1}\|\|[\![v\mathbf{b}]\!]\|_{H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)'}$, which, in combination with (3.5), completes the proof. ☐

The second tool for the proof of Theorem 3.1 is the following well-known consequence of the *closed range theorem*.

Lemma 3.5. *For reflexive Banach spaces X and Y, let $G : X \to Y'$ be linear. Then $G \in \mathcal{L}\mathrm{is}(X, Y')$ if and only if:*

(i) $G \in \mathcal{L}(X, Y')$,
(ii) $\beta := \inf_{0\neq y\in Y} \sup_{0\neq x\in X} \frac{(Gx)(y)}{\|x\|_X\|y\|_Y} > 0$,
(iii) $\forall 0 \neq x \in X$, $\exists y \in Y$, *with* $(Gx)(y) \neq 0$.

Moreover, one has $\|G^{-1}\|_{\mathcal{L}(Y',X)} = \frac{1}{\beta}$.

Since $G \in \mathcal{L}\mathrm{is}(X, Y')$ is equivalent to $G' \in \mathcal{L}\mathrm{is}(X', Y)$, the roles of X and Y in (ii) and (iii) can be interchanged.

Proof of Theorem 3.1. The bound on $\|\mathcal{B}_h\|_{\mathcal{L}(U,V')}$ follows easily from (3.3).

We will establish the remaining claim with the aid of Lemma 3.5. To first verify (iii), let $(u, \theta) \in \mathbb{U}$ be such that $b_h(u, \theta; v) = 0$ for all $v \in H(\mathbf{b}; \Omega_h)$. Considering first all v from the subspace $H_{0,\Gamma_+}(\mathbf{b}; \Omega)$, (2.5) yields $u = 0$ because $\mathcal{B}$ agrees with $\mathcal{B}_h$ on this subspace. By considering now for any $K \in \Omega_h$ all v with $\operatorname{supp} v \subset K$, we infer that $\theta|_{\partial K} = 0$, and so $\theta = 0$.

Finally, let $v \in H(\mathbf{b}; \Omega_h)$ be given. By (2.5), there exists a $v_1 = v_1(v) \in H_{0,\Gamma_+}(\mathbf{b}; \Omega)$ with

$$cv_1 - \operatorname{div}(v_1\mathbf{b}) = cv - \operatorname{div}_h(v\mathbf{b}), \quad \|v_1\|_{H(\mathbf{b};\Omega)} \leq \|\mathcal{B}^{-*}\|\|cv - \operatorname{div}_h(v\mathbf{b})\|_{L_2(\Omega)}. \tag{3.7}$$

Thus $\|v_1\|_{H(\mathbf{b};\Omega)} \leq \|\mathcal{B}^{-*}\|(1 + \|c - \operatorname{div}\mathbf{b}\|)\|v\|_{H(\mathbf{b};\Omega_h)}$, which says that

$$\|v_1 - v\|_{H(\mathbf{b};\Omega_h)} \leq (1 + \|\mathcal{B}^{-*}\|(1 + \|c - \operatorname{div}\mathbf{b}\|))\|v\|_{H(\mathbf{b};\Omega_h)}. \tag{3.8}$$

Moreover, we have $v_1 = v$ when $v \in H_{0,\Gamma_+}(\mathbf{b}; \Omega)$, so that for any $z \in H_{0,\Gamma_+}(\mathbf{b}; \Omega)$ we have $v_1(v - z) - (v - z) = v_1(v) - v$ so that (3.8) actually gives

$$\begin{aligned} \|v_1 - v\|_{H(\mathbf{b};\Omega_h)} &\leq (1 + \|\mathcal{B}^{-*}\|(1 + \|c - \operatorname{div}\mathbf{b}\|)) \inf_{z\in H_{0,\Gamma_+}(\mathbf{b};\Omega)} \|v - z\|_{H(\mathbf{b};\Omega_h)} \\ &\leq \tilde{C}_{\mathcal{B}}\|[\![v\mathbf{b}]\!]\|_{H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)'} \end{aligned} \tag{3.9}$$

by an application of Lemma 3.4.

There exists a $\theta \in H_{0,\Gamma_-}(\mathbf{b}; \partial\Omega_h)$ with $\|[\![v\mathbf{b}]\!]\|_{H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)'} = \frac{\int_{\partial\Omega_h}[\![v\mathbf{b}]\!]\theta\, d\mathbf{s}}{\|\theta\|_{H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)}}$. By selecting $\|\theta\|_{H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)} = \tilde{C}_{\mathcal{B}}^{-1}\|v_1 - v\|_{H(\mathbf{b};\Omega_h)}$, and invoking (3.9), we have

$$\|\theta\|^2_{H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)} = \tilde{C}_{\mathcal{B}}^{-2}\|v_1 - v\|^2_{H(\mathbf{b};\Omega_h)} \leq \int_{\partial\Omega_h}[\![v\mathbf{b}]\!]\theta\, d\mathbf{s}. \tag{3.10}$$

Similarly, there exists a $u \in L_2(\Omega)$ with $\|\mathcal{B}^*v_1\|_{L_2(\Omega)} = \frac{\int_\Omega cuv_1 - u\operatorname{div}(v_1\mathbf{b})\, d\mathbf{x}}{\|u\|_{L_2(\Omega)}}$. By selecting $\|u\|_{L_2(\Omega)} = \|\mathcal{B}^{-*}\|^{-1}\|v_1\|_{H(\mathbf{b};\Omega)}$, and using the first relation in (3.7), we infer that

$$\|u\|^2_{L_2(\Omega)} = \|\mathcal{B}^{-*}\|^{-2}\|v_1\|^2_{H(\mathbf{b};\Omega)} \leq \int_\Omega cuv - u\operatorname{div}_h(v\mathbf{b})\, d\mathbf{x}. \tag{3.11}$$

The combination of (3.10) and (3.11) shows that

$$
\begin{aligned}
&(\|\mathcal{B}^{-*}\|^2 + \tilde{C}_{\mathcal{B}}^2)^{-\frac{1}{2}} \|v\|_{H(\mathbf{b};\Omega_h)} \\
&\leq (\|\mathcal{B}^{-*}\|^2 + \tilde{C}_{\mathcal{B}}^2)^{-\frac{1}{2}} (\|v_1\|_{H(\mathbf{b};\Omega)} + \|v_1 - v\|_{H(\mathbf{b};\Omega_h)}) \\
&\leq \sqrt{\|\mathcal{B}^{-*}\|^{-2}\|v_1\|^2_{H(\mathbf{b};\Omega)} + \tilde{C}_{\mathcal{B}}^{-2}\|v_1 - v\|^2_{H(\mathbf{b};\Omega_h)}} \\
&= \frac{\|\mathcal{B}^{-*}\|^{-2}\|v_1\|^2_{H(\mathbf{b};\Omega)} + \tilde{C}_{\mathcal{B}}^{-2}\|v_1 - v\|^2_{H(\mathbf{b};\Omega_h)}}{\sqrt{\|u\|^2_{L_2(\Omega)} + \|\theta\|^2_{H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)}}} \\
&\leq \frac{b(u,\theta;v)}{\sqrt{\|u\|^2_{L_2(\Omega)} + \|\theta\|^2_{H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)}}}.
\end{aligned}
$$

Invoking Lemma 3.5 completes the proof. □

Remark 3.6 (Inhomogenous boundary condition). The variational formulation (3.2) is not suited for an inhomogeneous boundary condition $u = g$ on Γ_-, because the homogeneous condition $u = 0$ on Γ_- has been incorporated in the space $H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)$ for the variable θ.

Therefore, for $g \neq 0$, let $\bar{g} \in H(\mathbf{b},\Omega)$ be an extension of g. Then with $\bar{u} := u - \bar{g}$, one may apply the variational formulation (3.2) to the transport equation

$$
\begin{cases}
\mathbf{b}\cdot\nabla\bar{u} + c\bar{u} = f - \mathbf{b}\cdot\nabla\bar{g} - c\bar{g} & \text{on } \Omega,\\
\bar{u} = 0 & \text{on } \Gamma_-,
\end{cases}
$$

which gives the problem of finding $(\bar{u},\bar{\theta}) \in \mathbb{U}$ such that for all $v \in \mathbb{V}$,

$$
\begin{aligned}
b_h(\bar{u},\bar{\theta};v) &= f(v) - \int_\Omega (\mathbf{b}\cdot\nabla\bar{g} + c\bar{g})v\,d\mathbf{x} \\
&= f(v) + \int_\Omega (\mathbf{b}\cdot\nabla_h v + v\operatorname{div}\mathbf{b} - cv)\bar{g}\,d\mathbf{x} - \int_{\partial\Omega_h} [\![v\mathbf{b}]\!]\bar{g}\,d\mathbf{s}.
\end{aligned}
$$

When $f \in L_2(\Omega)$, it holds that $\bar{\theta} = \bar{u}|_{\partial\Omega_h} = (u - \bar{g})|_{\partial\Omega_h}$.

Alternatively, using that only the space for θ is inappropriate for $g \neq 0$, by subtracting $\int_{\Omega_h} [\![v\mathbf{b}]\!]\bar{g}\,d\mathbf{s}$ from both sides of (3.2), and introducing $\bar{\theta} := \theta - \bar{g}|_{\partial\Omega_h}$, one arrives at the problem of finding $(u,\bar{\theta}) \in \mathbb{U}$ such that for all $v \in \mathbb{V}$,

$$
b_h(u,\bar{\theta};v) = f(v) - \int_{\partial\Omega_h} [\![v\mathbf{b}]\!]\bar{g}\,d\mathbf{s}.
$$

4. Optimal test functions

4.1. **Preliminary remarks and a roadmap.** Given a family of finite dimensional piecewise polynomial trial spaces $\mathbb{U}^h \subset \mathbb{U} = L_2(\Omega) \times H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)$, parametrized by the mesh size parameter h, we wish to construct a uniformly stable finite dimensional family of test search spaces $\mathbb{V}^h \subset \mathbb{V} = H(\mathbf{b};\Omega_h)$ which, due to the product structure of $\mathbb{V}$, have the form

$$
\mathbb{V}^h = \prod_{K\in\Omega_h} \mathbb{V}^h_K.
$$

By uniformly stable we mean of course that there exists a positive constant $\gamma > 0$ such that (1.14) holds for the present setting, i.e.,

$$\inf_{(u,\theta)\in\mathbb{U}^h} \sup_{v\in\mathbb{V}^h} \frac{b_h(u,\theta;v)}{\|(u,\theta)\|_{\mathbb{U}}\|v\|_{\mathbb{V}}} =: \gamma^h \geq \gamma \quad (h>0). \tag{4.1}$$

In view of (3.1), it suffices to establish inf-sup stability for a slightly modified formulation replacing the component $\theta \in H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)$ by a suitable "lifting" $w \in H_{0,\Gamma_-}(\mathbf{b};\Omega)$, i.e., $w|_{\partial\Omega_h} = \theta$, which we express by writing

$$b_h(u,w;v) := \sum_{K\in\Omega_h} b_K(u,w;v),$$

where

$$\begin{aligned} b_K(u,w;v) &:= \int_K (cv - \operatorname{div}(\mathbf{b}v))u d\mathbf{x} + \int_{\partial K} \mathbf{b}\cdot\mathbf{n}_K vw d\mathbf{s} \\ &= \int_K (c-\operatorname{div}\mathbf{b})vu + (w-u)\mathbf{b}\cdot\nabla v + v\mathbf{b}\cdot\nabla w + vw\operatorname{div}\mathbf{b}d\mathbf{x}. \end{aligned} \tag{4.2}$$

In consequence we endow $L_2(\Omega)\times H_{0,\Gamma_-}(\mathbf{b};\Omega)$, that with some abuse of notation we again denote as $\mathbb{U}$, with the norm

$$|\!|\!|(u,w)|\!|\!|_{\mathbb{U}}^2 := \|u\|_{L_2(\Omega)}^2 + \|w\|_{H(\mathbf{b};\Omega)}^2 \ \big(\geq \|(u,w|_{\partial\Omega_h})\|_{\mathbb{U}}^2\big), \tag{4.3}$$

and recall from (1.19) that the trial-to-test map $\mathcal{T}:\mathbb{U}\to\mathbb{V}$ has now also product form

$$\mathcal{T}(u,w) = \big(\mathcal{T}_K(u|_K, w|_{\partial K})\big)_{K\in\Omega_h},$$

where each local optimal test-function $t_K = t_K(u,w) := \mathcal{T}_K(u|_K, w|_{\partial K})$ is defined by

$$\langle t_K, v\rangle_{H(\mathbf{b};K)} = b_K(u,w;v) \quad (v\in H(\mathbf{b};K)). \tag{4.4}$$

Our goal is to identify stable formulations for variable fields $\mathbf{b}$ subject to the assumptions made earlier. For such general fields one cannot expect to find truly optimal test functions, but essentially we will be able to do so for *piecewise constant* fields. Therefore we will introduce a perturbed bilinear form

$$\breve{b}_h(u,w;v) := \sum_{K\in\Omega_h} \breve{b}_K(u,w;v), \tag{4.5}$$

where the summands $\breve{b}_K(u,w;v)$ are defined as follows. Suppose that $\underline{c}_K, \underline{\mathbf{b}}_K, \underline{d}_K$ are approximations on K to the fields $c-\operatorname{div}\mathbf{b}, \mathbf{b}, \operatorname{div}\mathbf{b}$, respectively. Then in accordance with (4.2), we set

$$\begin{aligned} \breve{b}_K(u,w;v) &:= \int_K \underline{c}_K vu + (w-u)\underline{\mathbf{b}}_K\cdot\nabla v + v\underline{\mathbf{b}}_K\cdot\nabla w + \underline{d}_K vw d\mathbf{x} \\ &= \int_K \big((\underline{c}_K + \operatorname{div}\underline{\mathbf{b}}_K)v - \operatorname{div}(\underline{\mathbf{b}}_K v)\big)u + (\underline{d}_K - \operatorname{div}\underline{\mathbf{b}}_K)wv\,d\mathbf{x} \\ &\quad + \int_{\partial K} \underline{\mathbf{b}}_K\cdot\mathbf{n}_K vw d\mathbf{s}. \end{aligned} \tag{4.6}$$

These approximations will be specified later in Section 4.5. At this point note that $\underline{d}_K$ vanishes when $\mathbf{b}$ is constant and will generally differ from $\operatorname{div}\underline{\mathbf{b}}_K$. Its effect is that, for $\underline{d}_K \neq \operatorname{div}\underline{\mathbf{b}}_K$, the corresponding (near-)optimal test functions no longer depend only on the traces $w|_{\partial\Omega_h}$.

Given such a perturbed form $\breve{b}_h$ and a finite dimensional (piecewise polynomial) trial space $\mathbb{U}^h \subset \mathbb{U}$, we then have to carry out two main tasks:

(i) For any $(u, w) \in \mathbb{U}^h$ we wish to find a $\breve{t} = \breve{t}(u, w; \breve{b}_h) \in \mathbb{V}$, preferably piecewise polynomial, such that $\breve{b}_h(u, w; \breve{t}) \gtrsim \|(u, w)\|_{\mathbb{U}} \|\breve{t}\|_{\mathbb{V}}$, of course, uniformly in h and in $(u, w) \in \mathbb{U}^h$.

(ii) Starting from the simple decomposition

$$b_h(u, w; \breve{t}) = \breve{b}_h(u, w; \breve{t}) + (b_h(u, w; \breve{t}) - \breve{b}_h(u, w; \breve{t})), \tag{4.7}$$

the choice of $\breve{t}$ allows us to handle the first summand. It then remains to show for the second summand that

$$|b_h(u, w; \breve{t}) - \breve{b}_h(u, w; \breve{t})| \leq \delta \|(u, w)\|_{\mathbb{U}} \|\breve{t}\|_{\mathbb{V}}, \tag{4.8}$$

holds for a sufficiently small $\delta > 0$, depending on the inf-sup constant for the first summand. Note that after having established (i)-(ii), *any* test search space

$$\mathbb{V}^h \supseteq \operatorname{span}\{\breve{t}(u, w; \breve{b}_h) \colon (u, w) \in \mathbb{U}^h\}$$

will be uniformly stable in the sense of (4.1).

Concerning (i), in Section 4.3 we will see that after equipping the test space with a different but equivalent norm, the trial-to-test map can be evaluated exactly. It turns out, however, that the resulting truly optimal test functions corresponding to $\breve{b}_h$ are possibly very sensitive to perturbations in the convection field. Therefore, in order to be able to simultaneously establish (ii), we will have to replace them by near-optimal test functions.

Another issue we will have to deal with is the following: If one has a bilinear form for which the corresponding operator, in the infinite dimensional setting, is boundedly invertible, then for given finite dimensional trial space, the corresponding optimal test space gives an inf-sup stable pair. The convection field corresponding to the perturbed bilinear form $\breve{b}_h$, however, will generally not be in $W^0_\infty(\operatorname{div}; \Omega)$, and so the theory about well-posedness in the infinite dimensional setting developed in Section 3 will not be applicable to this perturbed form. We will establish the inf-sup stability needed in (i) partly by direct calculations, and partly by invoking the well-posedness of the original bilinear form.

4.2. **Reduction to two-point boundary value problems.** Recall from (4.2)–(4.4) that the local optimal test functions are given by local variational problems of the form $\langle t_K, v\rangle_{H(\mathbf{b};K)} = b_K(u, w; v)$. To determine the structure of such test functions when the coefficients $c, \mathbf{b}$ are also approximated we consider the local variational problems

$$\langle t_K, v\rangle_{H(\mathbf{b};K)} = \int_K (cv - \operatorname{div}(\mathbf{b}v))u + dwv \, d\mathbf{x} + \int_{\partial K} \mathbf{b} \cdot \mathbf{n}_K vw \, d\mathbf{s}, \tag{4.9}$$

where the bilinear form on the right-hand side “extends” the one in (4.2) by including the additional term dwv. This allows us, in particular, to accommodate forms of the type (4.6). Thus, for the moment, the coefficients c, d stand for specifications used later to construct approximations $\breve{b}_h$ to the bilinear form b_h; see (4.28).

When $u|_K \in H(\mathbf{b}; K)$, as is the case when u is piecewise polynomial, integration by parts shows that $-\int_K \operatorname{div}(\mathbf{b}v)u \, d\mathbf{x} = \int_K v\partial_{\mathbf{b}} u \, d\mathbf{x} - \int_{\partial K} \mathbf{b} \cdot \mathbf{n}_K vu \, d\mathbf{s}$, and $\int_K \partial_{\mathbf{b}} t_K \partial_{\mathbf{b}} v \, d\mathbf{x} = -\int_K (\partial^2_{\mathbf{b}} t_K + \partial_{\mathbf{b}} t_K \operatorname{div} \mathbf{b}) v \, d\mathbf{x} + \int_{\partial K} \mathbf{b} \cdot \mathbf{n}_K v \partial_{\mathbf{b}} t_K \, d\mathbf{s}$, which reveals that the strong form of (4.9) is given by

$$\text{(4.10)} \qquad \begin{cases} -[\partial_{\mathbf{b}}^2 t_K + \partial_{\mathbf{b}} t_K \operatorname{div} \mathbf{b}] + t_K &= \partial_{\mathbf{b}} u + cu + dw \quad \text{in } K, \\ \partial_{\mathbf{b}} t_K &= w - u \quad \text{on } \partial K_+ \cup \partial K_-. \end{cases}$$

Using a transformation to characteristic coordinates defined by

$$\frac{d}{d\lambda}\chi(\lambda, \mathbf{s}) = \mathbf{b}(\chi(\lambda, \mathbf{s})), \quad \chi(0, \mathbf{s}) = \mathbf{s} \in K_-,$$

(4.10) can be viewed as a family of ordinary two-point boundary value problems. In fact, defining $\hat{t}_K := t_K \circ \chi$, $\hat{u} := u \circ \chi$, and denoting by $L(s) > 0$ the smallest number for which $\chi(L(\mathbf{s}), \mathbf{s}) \in K_+$, (4.10) takes the form

$$\begin{aligned} -[\tfrac{d^2 \hat{t}_K}{d\lambda^2} + \tfrac{d\hat{t}_K}{d\lambda}(\operatorname{div} \mathbf{b}) \circ \chi] + \check{t}_K &= \tfrac{d\hat{u}}{d\lambda} + c \circ \chi\, \hat{u} + (dw) \circ \chi \quad \text{in } (0, L(\mathbf{s})), \\ \tfrac{d\hat{t}_K}{d\lambda} &= w \circ \chi - \hat{u} \quad \text{at } \{0, L(\mathbf{s})\}, \end{aligned}$$

which, in principle, we can solve for each $\mathbf{s}$ at any desired accuracy and, for certain $\underline{\mathbf{b}}, u, w$ even exactly.

4.3. **(Piecewise) constant convection field.** A simple explicit representation of t_K can be obtained when $\mathbf{b}|_K = \underline{\mathbf{b}}$ is constant, K is polyhedral, and the restrictions of u, w, c and d to each K are polynomial. The characteristics are then straight lines and the local optimal test function t_K, can then be determined analytically. It fails, however, to be itself a piecewise polynomial. In order to arrive in this case at (piecewise) polynomial local optimal test functions we follow an idea from [DG11]. Namely, we equip $H(\underline{\mathbf{b}}; K)$ with an alternative, but equivalent Hilbertian norm. The key is the following simple lemma.

Lemma 4.1. *For $k \geq h > 0$, it holds that*

$$k^2 \|v'\|^2_{L_2(0,h)} + \|v\|^2_{L_2(0,h)} \eqsim k^2 \|v'\|^2_{L_2(0,h)} + h|v(0)|^2 \quad (v \in H^1(0,h)),$$

where the (hidden) constants are independent of h, $k \geq h$, and v.

Proof. First note that it is sufficient to prove the result for $k = h > 0$. Next, a homogeneity argument shows that it is sufficient to consider the case that $h = 1$. For this case, the statement follows from $\|v\|_{L_2(0,1)} \leq \|v - v(0)\|_{L_2(0,1)} + |v(0)| \lesssim \|v'\|_{L_2(0,1)} + |v(0)|$ by Friedrich's inequality, together with $|v(0)| \lesssim \|v\|_{H^1(0,1)}$ by Sobolev's inequality. ☐

Remark 4.2. Obviously, the condition $k \geq h$ can be replaced by $k \geq Ch$ for some $C > 0$. Since this constant would then propagate through essentially all subsequent developments combined with further unspecified constants, for convenience we keep $C = 1$.

Proposition 4.3. *Let $K \subset \mathbb{R}^n$ be a Lipschitz domain, and assume that $0 \neq \underline{\mathbf{b}} \in \mathbb{R}^n$ is a* constant. *Denoting by $r(\mathbf{s})$ the distance from $\mathbf{s} \in \partial K_-$ to ∂K_+ along $\underline{\mathbf{b}}$, one has for $q_K \geq |\underline{\mathbf{b}}|^{-1} \operatorname{diam}(K)$,*

$$\begin{aligned} &q_K^2 \|\partial_{\underline{\mathbf{b}}} v\|^2_{L_2(K)} + \|v\|^2_{L_2(K)} \\ &\eqsim q_K^2 \|\partial_{\underline{\mathbf{b}}} v\|^2_{L_2(K)} - \int_{\partial K_-} |v(\mathbf{s})|^2 \tfrac{\underline{\mathbf{b}} \cdot \mathbf{n}_K(\mathbf{s})}{|\underline{\mathbf{b}}|} r(\mathbf{s}) d\mathbf{s} \quad (v \in H(\underline{\mathbf{b}}; K)), \end{aligned}$$

where the constants are those from Lemma 4.1.

Proof. Obviously, it is sufficient to prove the statement for $q_K = |\underline{\mathbf{b}}|^{-1}\operatorname{diam}(K)$, so that $q_K^2\|\partial_{\underline{\mathbf{b}}} v\|_{L_2(K)}^2 = \operatorname{diam}(K)^2\|\partial_{\underline{\mathbf{b}}/|\underline{\mathbf{b}}|} v\|_{L_2(K)}^2$. Without loss of generality we may consider the case that $\underline{\mathbf{b}}/|\underline{\mathbf{b}}| = \mathbf{e}_1$. Given $x_2, \ldots, x_n$, let $\mathbf{s}$ denote the projection of $(x_2, \ldots, x_n)$ on ∂K_- along the x_1-direction. We apply Lemma 4.1 for the integration in the x_1-direction, where we use that for each $\mathbf{s}$ the quantity $r(\mathbf{s})$ plays the role of h in Lemma 4.1 while $\operatorname{diam}(K) \geq r(\mathbf{s})$ plays the role of k in Lemma 4.1. Integrating the result over $x_2, \ldots, x_n$ and using that $d\mathbf{s} = \frac{|\underline{\mathbf{b}}|}{|\underline{\mathbf{b}}\cdot\mathbf{n}_K(\mathbf{s})|} dx_2 \ldots dx_n$ and $\underline{\mathbf{b}} \cdot \mathbf{n}_K < 0$ on ∂K_-, confirms the claim. ☐

Remark 4.4. Proposition 4.3 can be generalized to nonconstant $\mathbf{b}$ by applying the coordinate transformation χ involving the characteristic coordinates. The constants absorbed by the equivalence symbol $\eqsim$ then depend also on the Jacobian of χ, and the length of the characteristic curve sections connecting the inflow and outflow boundaries; see also [DSMMO04].

The above lines of thought were already used in [DG11, (3.22)] where, however, the (necessary) factor $\frac{-\underline{\mathbf{b}}\cdot\mathbf{n}_K(\mathbf{s})}{|\underline{\mathbf{b}}|} r(\mathbf{s})$ in the integrand of the integral over ∂K_- is missing.

For later use we record the following consequence of Proposition 4.3.

Remark 4.5. For a *constant* $\underline{\mathbf{b}} \neq 0$, and

$$\operatorname{diam}(K) \leq |\underline{\mathbf{b}}|,$$

the scalar product

$$\langle\!\langle v, z\rangle\!\rangle_{K,\underline{\mathbf{b}}} := \langle \partial_{\underline{\mathbf{b}}} v, \partial_{\underline{\mathbf{b}}} z\rangle_{L_2(K)} - \int_{\partial K_-} v(\mathbf{s})z(\mathbf{s})\frac{\underline{\mathbf{b}}\cdot\mathbf{n}_K(\mathbf{s})}{|\underline{\mathbf{b}}|} r(\mathbf{s})d\mathbf{s}$$

gives rise to an *equivalent* norm on $H(\underline{\mathbf{b}}; K)$, so that this scalar product can be used to determine the local optimal test function.

Assuming that $u|_K \in H(\underline{\mathbf{b}}; K)$, the local optimal test function $t_K = \mathcal{T}_K(u|_K, w|_K)$ that results from replacing $\langle\, ,\, \rangle_{H(\underline{\mathbf{b}};K)}$ by $\langle\!\langle\, ,\, \rangle\!\rangle_{K,\underline{\mathbf{b}}}$ in (4.9), is the solution of

$$\left\{\begin{array}{rcll} -\partial_{\underline{\mathbf{b}}}^2 t_K &=& \partial_{\underline{\mathbf{b}}} u + cu + dw & \text{on } K, \\ \partial_{\underline{\mathbf{b}}} t_K - r|\underline{\mathbf{b}}|^{-1} t_K &=& w - u & \text{on } \partial K_-, \\ \partial_{\underline{\mathbf{b}}} t_K &=& w - u & \text{on } \partial K_+. \end{array}\right. \tag{4.11}$$

We consider the case of K being *convex*. By a rotation of coordinates, for solving (4.11) it is sufficient to consider the case of

$$\underline{\mathbf{b}} = |\underline{\mathbf{b}}|\mathbf{e}_1,$$

or, equivalently, to read $(x_1, \ldots, x_n)$ as Cartesian coordinates with the first basis vector being equal to $\underline{\mathbf{b}}/|\underline{\mathbf{b}}|$. For $\mathbf{x} = (x, \mathbf{y}) \in K$, let $x_\pm(\mathbf{y})$ be such that $(x_\pm(\mathbf{y}), \mathbf{y}) \in \partial K_\pm$; see Figure 1.

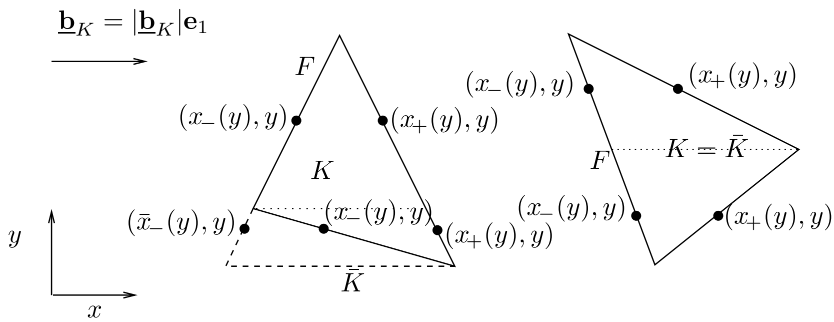

FIGURE 1. $x_\pm$ on a triangle K with two (left) or one (right) inflow faces (the meaning of the inflow face F, triangle $\bar{K} \supseteq K$, and the function $\bar{x}_-(\mathbf{y})$ will become clear in Section 4.4)

Furthermore, although not essential, we will think of c and d as being constant. Then the solution

$$t_K = \mathcal{T}_{K,\underline{\mathbf{b}},\underline{c},\underline{d}}(u|_K, w|_K)$$

of (4.11) is given by

(4.12)

$$\begin{aligned}
|\underline{\mathbf{b}}|\, t_K(x, \mathbf{y}) &= -|\underline{\mathbf{b}}|^{-1} \int_{x_-(\mathbf{y})}^{x} \int_{x_-(\mathbf{y})}^{z} (\partial_{\underline{\mathbf{b}}} u + \underline{c} u + \underline{d} w)(q, \mathbf{y}) dq dz \\
&+ \Big(w(x_+(\mathbf{y}), \mathbf{y}) - u(x_-(\mathbf{y}), \mathbf{y}) + |\underline{\mathbf{b}}|^{-1} \int_{x_-(\mathbf{y})}^{x_+(\mathbf{y})} (\underline{c} u + \underline{d} w)(q, \mathbf{y}) dq \Big) \Big(x - x_-(\mathbf{y}) \Big) \\
&+ |\underline{\mathbf{b}}|^2 \frac{w(x_+(\mathbf{y}), \mathbf{y}) - w(x_-(\mathbf{y}), \mathbf{y}) + |\underline{\mathbf{b}}|^{-1} \int_{x_-(\mathbf{y})}^{x_+(\mathbf{y})} (\underline{c} u + \underline{d} w)(q, \mathbf{y}) dq}{x_+(\mathbf{y}) - x_-(\mathbf{y})}.
\end{aligned}$$

For K being polyhedral, the function $\mathbf{y} \mapsto x_\pm(\mathbf{y})$ is continuous *piecewise* linear. Using that for any univariate polynomial p of degree $m \geq 1$, $(\alpha, \beta) \mapsto \frac{p(\beta)-p(\alpha)}{\beta-\alpha}$ is a bivariate polynomial of degree $m-1$, we infer that for u, w being polynomials on K, t_K is a continuous piecewise polynomial on K.

4.4. **A stability issue.** Unfortunately, depending on the angle between $\underline{\mathbf{b}}$ and a face, the derivatives of x_+ or x_- can be arbitrarily large. Therefore, generally the problem of determining an optimal test function is not stable regarding its dependence on $\underline{\mathbf{b}}$. Consequently, a serious impediment arises when the piecewise constant $\underline{\mathbf{b}}$ is an approximation to a variable field $\mathbf{b}$. As will be seen later (last statement of Lemma 4.10), when treating the second summand in (4.7), one eventually has to control the H^1-norm of the test functions via inverse estimates which requires controlling the derivatives of $x_\pm(\mathbf{y})$.

To tackle this problem, for K being an *n-simplex*, we define an approximation $\breve{t}_K$ to t_K by discarding higher order terms, which is stable as a function of $\underline{\mathbf{b}}$; whereas, for polynomial u and w, t_K is only *piecewise* polynomial w.r.t. a partition of K that depends on the field $\underline{\mathbf{b}}$, $\breve{t}_K$ will be *polynomial.*

To define $\breve{t}_K$, first we construct a *polyhedral* set $\bar{K}$ that contains K as follows. The number of inflow faces of K is between 1 and $n-1$. Let F be an inflow face whose inward pointing normal makes the smallest angle with $\underline{\mathbf{b}}$, and let v denote the vertex of K that does not belong to F. Finally let H_F denote the $(n-1)$-hyperplane containing F. The "shadow" of K on H_F, i.e.,

$$\bar{F} := \big\{\mathbf{x} \in H_F : \{\mathbf{x} + t\underline{\mathbf{b}} : t \in \mathbb{R}\} \cap K \neq \emptyset\big\}$$

is an $(n-1)$-dimensional polyhedron containing F and let $\bar{K}$ denote the convex hull of v and $\bar{F}$; cf. Figure 1 for $n = 2$. Then, by construction, $\bar{K}$ has only one inflow face $\bar{K}_- := \bar{F}$, and $K \subseteq \bar{K}$ with equality if and only if K has only one inflow face, namely $K_- = F$.

For $\mathbf{x} = (x, \mathbf{y}) \in \bar{K} \supset K$, let $\bar{x}_-(\mathbf{y})$ be the linear function with $(\bar{x}_-(\mathbf{y}), \mathbf{y}) \in \partial\bar{K}_-$, i.e., $\bar{x}(\mathbf{y})$ agrees with $x(\mathbf{y})$ on F. Then we have

$$\operatorname{diam}(\bar{K}) \lesssim \operatorname{diam}(K), \tag{4.13}$$

$$|\bar{x}_-|_{W^1_\infty(\bar{K})} \lesssim 1, \tag{4.14}$$

where both constants depend only on (an upper bound for) the shape regularity parameter

$$\varrho_K := \frac{\operatorname{diam}(K)}{\sup\{\operatorname{diam}(B) \colon B \text{ a ball in } K\}}.$$

For polynomials u and w on K, say of degree m, we define now the *local test function*

$$\breve{t}_K = \breve{t}_{K,\underline{\mathbf{b}},\underline{c},\underline{d}}(u|_K, w|_K) \in \mathcal{P}_{m+1}$$

by

$$\begin{aligned} |\underline{\mathbf{b}}|\, \breve{t}_K(x, \mathbf{y}) :=& \Big(w(\bar{x}_-(\mathbf{y}), \mathbf{y}) - u(\bar{x}_-(\mathbf{y}), \mathbf{y})\Big)\Big(x - \bar{x}_-(\mathbf{y})\Big) \\ &+ |\underline{\mathbf{b}}|(\partial_{\underline{\mathbf{b}}} w(\bar{x}_-(\mathbf{y}), \mathbf{y}) + \underline{c}u(\bar{x}_-(\mathbf{y}), \mathbf{y}) + \underline{d}w(\bar{x}_-(\mathbf{y}), \mathbf{y})). \end{aligned} \tag{4.15}$$

Since u and w are uniquely defined as polynomials on all of $\mathbb{R}^n$, the polynomial $\breve{t}_K$ is well-defined outside K and in particular on $\bar{K}$.

We will show that $\breve{t}_K$ deserves to be termed *near-optimal local test function* and as a first step we quantify the effect of the above modification.

Lemma 4.6. *Let $u|_K$ and $w|_K$ be polynomials of degree m. Then, with $t_K = t_{K,\underline{\mathbf{b}},\underline{c},\underline{d}}(u|_K, w|_K)$ as given in* (4.12)*, we have*

$$\|t_K - \breve{t}_K\|_{H(\underline{\mathbf{b}};K)} \lesssim |\underline{\mathbf{b}}|^{-1} \operatorname{diam}(K)\big[\|u\|_{L_2(K)} + \|w\|_{H(\underline{\mathbf{b}};K)} + \|\partial_{\underline{\mathbf{b}}} u\|_{L_2(K)} + \|\partial^2_{\underline{\mathbf{b}}} w\|_{L_2(K)}\big],$$

only dependent on upper bounds for m, $|\underline{c}|$, $|\underline{d}|$ and ϱ_K, and, as always, assuming that $\operatorname{diam}(K) \leq |\underline{\mathbf{b}}|$.

Proof. In view of the definitions of t_K and $\breve{t}_K$, we split their difference, as well as the difference of $\partial_{\underline{\mathbf{b}}} t_K$ and $\partial_{\underline{\mathbf{b}}} \breve{t}_K$, into a number of terms whose $L_2(K)$-norms we bound in a straightforward way. We start with the first task. It holds that

$$\begin{aligned} &\|(x, \mathbf{y}) \mapsto |\underline{\mathbf{b}}|^{-2} \int_{x_-(\mathbf{y})}^{x} \int_{x_-(\mathbf{y})}^{z} (\partial_{\underline{\mathbf{b}}} u + \underline{c}u + \underline{d}w)(q, \mathbf{y})\, dq dz\|_{L_2(K)} \\ &\quad \lesssim |\underline{\mathbf{b}}|^{-2} \operatorname{diam}(K)^2 \|\partial_{\underline{\mathbf{b}}} u + \underline{c}u + \underline{d}w\|_{L_2(K)} \end{aligned}$$

and

$$\|(x,\mathbf{y}) \mapsto |\underline{\mathbf{b}}|^{-2}(x - x_-(\mathbf{y})) \int_{x_-(\mathbf{y})}^{x_+(\mathbf{y})} (\underline{c}u + \underline{d}w)(q,\mathbf{y})dq\|_{L_2(K)} \lesssim |\underline{\mathbf{b}}|^{-2} \operatorname{diam}(K)^2 \|\underline{c}u + \underline{d}w\|_{L_2(K)}.$$

Writing

$$\begin{aligned}
&|\underline{\mathbf{b}}|^{-1}\big\{(w(x_+(\mathbf{y}),\mathbf{y}) - u(x_-(\mathbf{y}),\mathbf{y}))(x - x_-(\mathbf{y})) \\
&\qquad - (w(\bar{x}_-(\mathbf{y}),\mathbf{y}) - u(\bar{x}_-(\mathbf{y}),\mathbf{y}))(x - \bar{x}_-(\mathbf{y}))\big\} \\
&= |\underline{\mathbf{b}}|^{-1}\big(w(x,\mathbf{y}) - u(x,\mathbf{y})\big)\big(\bar{x}_-(\mathbf{y}) - x_-(\mathbf{y})\big) \\
&\quad + |\underline{\mathbf{b}}|^{-2}\Big(\int_x^{x_+(\mathbf{y})} \partial_{\underline{\mathbf{b}}} w(z,\mathbf{y})dz + \int_{x_-(\mathbf{y})}^{x} \partial_{\underline{\mathbf{b}}} u(z,\mathbf{y})dy\Big)\big(\bar{x}_-(\mathbf{y}) - x_-(\mathbf{y})\big) \\
&\quad + |\underline{\mathbf{b}}|^{-2}\Big(\int_{\bar{x}_-(\mathbf{y})}^{x_+(\mathbf{y})} \partial_{\underline{\mathbf{b}}} w(z,\mathbf{y})\,dz + \int_{x_-(\mathbf{y})}^{\bar{x}_-(\mathbf{y})} \partial_{\underline{\mathbf{b}}} u(z,\mathbf{y})\,dz\Big)\big(x - \bar{x}_-(\mathbf{y})\big),
\end{aligned}$$

the $L_2(K)$-norm of the expression on the first line on the right-hand side can be bounded by a constant multiple of

$$|\underline{\mathbf{b}}|^{-1} \operatorname{diam}(\bar{K})\big(\|w\|_{L_2(K)} + \|u\|_{L_2(K)}\big).$$

The terms on the second and third lines are bounded by constant multiples of

$$|\underline{\mathbf{b}}|^{-2} \operatorname{diam}(\bar{K})^2\big(\|\partial_{\underline{\mathbf{b}}} u\|_{L_2(\bar{K})} + \|\partial_{\underline{\mathbf{b}}} w\|_{L_2(\bar{K})}\big).$$

Proceeding to the difference of the last lines in (4.12), respectively (4.15), we have

$$\|(x,\mathbf{y}) \mapsto |\underline{\mathbf{b}}|\frac{w(x_+(\mathbf{y}),\mathbf{y}) - w(x_-(\mathbf{y}),\mathbf{y})}{x_+(\mathbf{y}) - x_-(\mathbf{y})} - \partial_{\underline{\mathbf{b}}} w(\bar{x}_-(\mathbf{y}),\mathbf{y})\|_{L_2(K)} \lesssim |\underline{\mathbf{b}}|^{-1} \operatorname{diam}(\bar{K})\|\partial_{\underline{\mathbf{b}}}^2 w\|_{L_2(\bar{K})}$$

and

$$\|(x,\mathbf{y}) \mapsto \frac{\int_{x_-(\mathbf{y})}^{x_+(\mathbf{y})} (\underline{c}u + \underline{d}w)(q,\mathbf{y})dq}{x_+(\mathbf{y}) - x_-(\mathbf{y})} - (\underline{c}u(\bar{x}_-(\mathbf{y}),\mathbf{y}) + \underline{d}w(\bar{x}_-(\mathbf{y}),\mathbf{y}))\|_{L_2(K)} \lesssim |\underline{\mathbf{b}}|^{-1} \operatorname{diam}(\bar{K})(|\underline{c}|\|\partial_{\underline{\mathbf{b}}} u\|_{L_2(\bar{K})} + |\underline{d}|\|\partial_{\underline{\mathbf{b}}} w\|_{L_2(\bar{K})}).$$

Second, we find the upper bound for the $L_2(K)$-norms for the different terms in $\partial_{\underline{\mathbf{b}}}(t_K - \breve{t}_K)$. Since

$$\begin{aligned}
\partial_{\underline{\mathbf{b}}} t_K(x,\mathbf{y}) = &- \Big[u(x,\mathbf{y}) - u(x_-(\mathbf{y}),\mathbf{y}) + |\underline{\mathbf{b}}|^{-1} \int_{x_-(\mathbf{y})}^{x} (\underline{c}u + \underline{d}w)(q,\mathbf{y})dq\Big] \\
&+ w(x_+(\mathbf{y}),\mathbf{y}) - u(x_-(\mathbf{y}),\mathbf{y}) + |\underline{\mathbf{b}}|^{-1} \int_{x_-(\mathbf{y})}^{x_+(\mathbf{y})} (\underline{c}u + \underline{d}w)(q,\mathbf{y})dq
\end{aligned}$$

and

$$\partial_{\underline{\mathbf{b}}} \breve{t}_K(x,\mathbf{y}) = w(\bar{x}_-(\mathbf{y}),\mathbf{y}) - u(\bar{x}_-(\mathbf{y}),\mathbf{y}),$$

we derive that

$$\|(x,\mathbf{y}) \mapsto u(x_-(\mathbf{y}),\mathbf{y}) - u(x,\mathbf{y})\|_{L_2(K)} \lesssim |\underline{\mathbf{b}}|^{-1}\operatorname{diam}(K)\|\partial_{\underline{\mathbf{b}}} u\|_{L_2(K)},$$

$$\begin{aligned}\|(x,\mathbf{y}) \mapsto |\underline{\mathbf{b}}|^{-1}\int_x^{x_+(\mathbf{y})} &(\underline{c}u + \underline{d}w)(q,\mathbf{y})dq\|_{L_2(K)} \\ &\lesssim |\underline{\mathbf{b}}|^{-1}\operatorname{diam}(K)(|\underline{c}|\|u\|_{L_2(K)} + |\underline{d}|\|w\|_{L_2(K)})\end{aligned}$$

and

$$\begin{aligned}\|(x,\mathbf{y}) \mapsto w(x_+(\mathbf{y}),\mathbf{y}) - w(\bar{x}_-(\mathbf{y}),\mathbf{y}) + u(\bar{x}_-(\mathbf{y}),\mathbf{y}) - u(x_-(\mathbf{y}),\mathbf{y})\|_{L_2(K)} \\ \lesssim |\underline{\mathbf{b}}|^{-1}\operatorname{diam}(\bar{K})(\|\partial_{\underline{\mathbf{b}}} w\|_{L_2(\bar{K})} + \|\partial_{\underline{\mathbf{b}}} u\|_{L_2(\bar{K})}).\end{aligned}$$

The proof is completed by collecting all upper bounds, by using that $\operatorname{diam}(\bar{K}) \lesssim \operatorname{diam}(K)$ ((4.13)), and that, for any polynomial p, $\|p\|_{L_2(\bar{K})} \lesssim \frac{|\bar{K}|}{|K|}\|p\|_{L_2(K)}$ with a constant depending only on its degree. ☐

As discussed earlier below (4.8), inf-sup stability of a perturbed bilinear form $\breve{b}_h$ with respect to a given piecewise polynomial trial space and corresponding test space based on (4.15) will be partly established by direct calculations. The next major step is given by the following lemma. Its result will be used to prove the forthcoming Lemma 4.10.

Lemma 4.7. *Let $u|_K$ and $w|_K$ be polynomials of degree m and assume that* $\operatorname{diam}(K) \leq |\underline{\mathbf{b}}|$. *For any $\varepsilon > 0$, one has*

$$\begin{aligned}&\|\breve{t}_K\|^2_{H(\underline{\mathbf{b}};K)} - \Big[\tfrac{1}{2+4|\underline{c}|^2}\|\partial_{\underline{\mathbf{b}}} w + (\underline{c}+\underline{d})w\|^2_{L_2(K)} + \tfrac{\varepsilon}{1+4\varepsilon}\|u\|^2_{L_2(K)} - \varepsilon\|w\|^2_{L_2(K)}\Big] \\ &\gtrsim -|\underline{\mathbf{b}}|^{-2}\operatorname{diam}(K)^2[\|u\|^2_{L_2(K)} + \|w\|^2_{H(\underline{\mathbf{b}};K)} + \|\partial_{\underline{\mathbf{b}}} u\|^2_{L_2(K)} + \|\partial^2_{\underline{\mathbf{b}}} w\|^2_{L_2(K)}],\end{aligned} \tag{4.16}$$

where the constant depends only on upper bounds for m, $|\underline{c}|$, $|\underline{d}|$ and ϱ_K.

Proof. By applying Young's inequality twice, in the form $\|\sigma\|^2 \geq (1-\eta)\|\tau\|^2 + (1-\eta^{-1})\|\sigma-\tau\|^2$ for $\eta \in (0,1)$, here for $\eta = \frac{1}{2}$, we have

$$\begin{aligned}\|\breve{t}_K\|^2_{H(\underline{\mathbf{b}};K)} =& \|\breve{t}_K\|^2_{L_2(K)} + \|\partial_{\underline{\mathbf{b}}}\breve{t}_K\|^2_{L_2(K)} \\ \geq& \tfrac{1}{2}\big[\|\partial_{\underline{\mathbf{b}}} w + \underline{c}u + \underline{d}w\|^2_{L_2(K)} + \|w-u\|^2_{L_2(K)}\big] \\ &- \big[\|\breve{t}_K - (\partial_{\underline{\mathbf{b}}} w + \underline{c}u + \underline{d}w)\|^2_{L_2(K)} + \|\partial_{\underline{\mathbf{b}}}\breve{t}_K - (w-u)\|^2_{L_2(K))}\big].\end{aligned}$$

The same arguments that were used in the proof of Lemma 4.6 show that

$$\begin{aligned}&\|\breve{t}_K - (\partial_{\underline{\mathbf{b}}} w + \underline{c}u + \underline{d}w)\|_{L_2(K)} \\ &\quad\lesssim |\underline{\mathbf{b}}|^{-1}\operatorname{diam}(K)\Big\{\|\partial^2_{\underline{\mathbf{b}}} w\|_{L_2(K)} + |\underline{c}|\|\partial_{\underline{\mathbf{b}}} u\|_{L_2(K)} + |\underline{d}|\|\partial_{\underline{\mathbf{b}}} w\|_{L_2(K)} \\ &\qquad + \|w\|_{L_2(K)} + \|u\|_{L_2(K)} + |\underline{\mathbf{b}}|^{-1}\operatorname{diam}(K)\big(\|\partial_{\underline{\mathbf{b}}} w\|_{L_2(K)} + \|\partial_{\underline{\mathbf{b}}} u\|_{L_2(K)}\big)\Big\}\end{aligned}$$

and

$$\|\partial_{\underline{\mathbf{b}}}\breve{t}_K - (w-u)\|_{L_2(K))} \lesssim |\underline{\mathbf{b}}|^{-1}\operatorname{diam}(K)[\|\partial_{\underline{\mathbf{b}}} w\|_{L_2(K)} + \|\partial_{\underline{\mathbf{b}}} u\|_{L_2(K)}].$$

Recalling that $\underline{c}$ is constant on K and taking $\eta = 1 - \frac{1}{1+2|\underline{c}|^2}$, two applications of Young's inequality provide

$$\begin{aligned}
&\|w-u\|^2_{L_2(K)} + \|\partial_{\underline{\mathbf{b}}} w + \underline{c}u + \underline{d}w\|^2_{L_2(K)} \\
&\geq \|w-u\|^2_{L_2(K)} + (1-\eta)\|\partial_{\underline{\mathbf{b}}} w + (\underline{c}+\underline{d})w\|^2_{L_2(K)} + (1-\tfrac{1}{\eta})\|\underline{c}(u-w)\|^2_{L_2(K)} \\
&= \tfrac{1}{2}\|w-u\|^2_{L_2(K)} + \tfrac{1}{1+2|\underline{c}|^2}\|\partial_{\underline{\mathbf{b}}} w + (\underline{c}+\underline{d})w\|^2_{L_2(K)} \\
&\geq \tfrac{2\varepsilon}{1+4\varepsilon}\|u\|^2_{L_2(K)} - 2\varepsilon\|w\|^2_{L_2(K)} + \tfrac{1}{1+2|\underline{c}|^2}\|\partial_{\underline{\mathbf{b}}} w + (\underline{c}+\underline{d})w\|^2_{L_2(K)},
\end{aligned}$$

with which the proof is easily completed. ☐

4.5. **The main result.** Let us fix

$$\mathbf{b} \in W^1_\infty(\mathrm{div};\Omega),\ c \in W^1_\infty(\Omega) \text{ such that (2.4) is valid, and } |\mathbf{b}|^{-1} \in L_\infty(\Omega), \tag{4.17}$$

and with that, for any partition Ω_h of Ω, the bilinear form b_h given in (3.2). For any $K \in \Omega_h$, we set

$$\underline{\mathbf{b}}_K := |K|^{-1}\int_K \mathbf{b}\,d\mathbf{x}, \quad \underline{c}_K := |K|^{-1}\int_K c - \operatorname{div}\mathbf{b}\,d\mathbf{x}, \quad \underline{d}_K := |K|^{-1}\int_K \operatorname{div}\mathbf{b}\,d\mathbf{x},$$

and define $\mathbf{b}_h \in L_\infty(\Omega)^n$, $c_h \in L_\infty(\Omega)$, and $d_h \in L_\infty(\Omega)$ by

$$\mathbf{b}_h|_K := \underline{\mathbf{b}}_K, \quad c_h|_K := \underline{c}_K, \quad d_h|_K := \underline{d}_K \quad (K \in \Omega_h), \tag{4.18}$$

with which we have defined the perturbed bilinear form $\breve{b}_h$ given in (4.5)–(4.6).

Our subsequent analysis of the terms on the right-hand side of (4.7) along the strategy outlined in Section 4.1 is guided by the following comments. First, note that generally $\mathbf{b}_h \notin W^0_\infty(\mathrm{div};\Omega)$, meaning that well-posedness of the corresponding variational form on the infinite dimensional level is not ensured. Indeed, since for $\phi \in C_0^\infty(\Omega)$, $\int_\Omega \phi \operatorname{div}\mathbf{b}_h\,d\mathbf{x} = \int_\Omega \phi \operatorname{div}_h \mathbf{b}_h\,d\mathbf{x} + \int_{\partial\Omega_h} [\![\mathbf{b}_h]\!]\phi$, and, unless $\mathbf{b}_h$ is constant over Ω, the right-hand side cannot be bounded by a multiple of $\|\phi\|_{L_1(\Omega)}$, we have $\operatorname{div}\mathbf{b}_h \notin L_\infty(\Omega)$. However, the perturbed form $\breve{b}_h$ is only applied to functions from finite dimensional spaces, which is also essential for treating the second summand in (4.7).

In this latter regard, another problem is that $\mathbf{b}_h$ is an approximation to $\mathbf{b}$ that is only *first order accurate.* In order to show that for a piecewise polynomial trial space, the second summand in (4.7) is sufficiently small relative to the first one, a central ingredient is to show that for a piecewise polynomial w_h, $\frac{\int_\Omega(\mathbf{b}-\mathbf{b}_h)\cdot\nabla w_h}{\|w\|_{H(\mathbf{b};\Omega_h)}}$ is sufficiently small. A combination of $\|\mathbf{b}-\mathbf{b}_h\|_{L_\infty(K)} \lesssim \operatorname{diam}(K)$ and the inverse inequality $|w|_{H^1(K)} \lesssim \operatorname{diam}(K)^{-1}\|w\|_{L_2(K)}$ shows only that this quotient is bounded.

We are going to solve this problem by considering trial spaces that are piecewise polynomial w.r.t. trial (macro-)partitions Ω_H, such that the ratio of the local mesh sizes h/H is less than some sufficiently small constant. This will allow us also to take care of those "higher order" terms in Lemma 4.6 which involve derivatives of u and w.

Specifically, let $\{\Omega_H : H \in \mathcal{I}\}$ be a family of partitions of a polyhedron $\Omega \subset \mathbb{R}^n$ into uniformly shape regular n-simplices, meaning that

$$\varrho := \sup_{H\in\mathcal{I}} \max_{K'\in\Omega_H} \varrho_{K'} < \infty. \tag{4.19}$$

For any $H \in \mathcal{I}$, let $\Omega_h = \Omega_{h(H)}$ be a refinement of Ω_H. We set

$$\sigma := \sup_{H\in\mathcal{I}} \max_{K'\in\Omega_H} \Big(\max_{\{K\in\Omega_h :\, K\subset K'\}} \frac{\operatorname{diam}(K)}{\operatorname{diam}(K')}, \operatorname{diam}(K')\Big), \tag{4.20}$$

which later will be assumed to be sufficiently small. This means that we will assume that any partition Ω_H is sufficiently fine, and that the (minimal) *subgrid refinement factor* when going from any Ω_H to Ω_h is sufficiently large. We consider only regular refinements Ω_h of Ω_H, in the sense that

$$\bar{\varrho} := \sup_{H\in\mathcal{I}} \max_{K\in\Omega_h} \varrho_K \lesssim \varrho, \tag{4.21}$$

uniformly in σ.

Given $u, w \in \prod_{K\in\Omega_h} \mathcal{P}_m(K)$, let $t = \mathcal{T}(u,w)$, $\breve{t} \in H(\mathbf{b}_h;\Omega_h)$ be defined for $K \in \Omega_h$ by

$$t|_K := \mathcal{T}_{K,\underline{\mathbf{b}}_K,\underline{c}_K,\underline{d}_K}(u|_K, w|_K),\ \breve{t}|_K := \breve{t}_K \in \mathcal{P}_{m+1}(K), \tag{4.22}$$

so that $t|_K$ is the optimal local test function defined in (4.12) corresponding to the approximate, constant coefficients $\underline{\mathbf{b}}_K$, $\underline{c}_K$, and $\underline{d}_K$, and the replacement of the standard scalar product on $H(\underline{\mathbf{b}};K)$ by $\langle\!\langle\, ,\, \rangle\!\rangle_{K,\underline{\mathbf{b}}}$; and $\breve{t}|_K$ is its polynomial approximation defined in (4.15).

We can now formulate the main result of this paper.

Theorem 4.8. *Assume the validity of* (4.19), (4.21), *and* (4.17). *Then there exists a* $\sigma_0 > 0$ *such that for* $0 < \sigma \le \sigma_0$ *(i.e., for sufficiently fine* Ω_H *and sufficiently large fixed subgrid refinement depth),*

$$(u,w) \in \mathbb{U}^H := \prod_{K\in\Omega_H} \mathcal{P}_m(K) \times H_{0,\Gamma_-}(\mathbf{b};\Omega) \cap \prod_{K\in\Omega_H} \mathcal{P}_m(K),$$

and with $\breve{t} = \breve{t}(u,w) \in \prod_{K\in\Omega_h} \mathcal{P}_{m+1}(K)$ *as defined in* (4.22), *for* b_h *defined in* (3.2) *it holds that*

$$b_h(u, w|_{\partial\Omega_h}; \breve{t}) \gtrsim |\!|\!|(u,w)|\!|\!|_{\mathbb{U}} \|\breve{t}\|_{\mathbb{V}},$$

where the constant depends only on (upper bounds for) m, ϱ, $\bar{\varrho}$, $\|\mathbf{b}\|_{W^1_\infty(\operatorname{div};\Omega)}$, $\||\mathbf{b}|^{-1}\|_{L_\infty(\Omega)}$, $\|c\|_{W^1_\infty(\Omega)}$, *and* $\|\mathcal{B}^{-1}\|_{\mathcal{L}(L_2(\Omega),H_{0,\Gamma_-}(\mathbf{b};\Omega))}$.

Remark 4.9. Theorem 4.8 guarantees inf-sup stability for a fixed, but otherwise unspecified subgrid refinement depth. In this latter aspect, our results are weaker than those from [GQ14, CDG16] for Poisson, elasticity and Maxwell problems, where the required dimension of the test search space could be made concrete.

Numerical experiments suggest that it is sufficient to take $h = H$.

The remainder of this section is devoted to the proof of Theorem 4.8. We begin with collecting some simple frequently needed technical preliminaries.

Obviously, we have

$$\|c_h\|_{L_\infty(\Omega)} \le \|c - \operatorname{div}\mathbf{b}\|_{L_\infty(\Omega)}, \quad \|d_h\|_{L_\infty(\Omega)} \le \|\operatorname{div}\mathbf{b}\|_{L_\infty(\Omega)}.$$

Moreover, for any n-simplex $K \subset \Omega$, it holds that

$$\begin{aligned} \|\underline{c}_K - (c - \operatorname{div}\mathbf{b})\|_{L_\infty(K)} &\lesssim \operatorname{diam}(K)|c - \operatorname{div}\mathbf{b}|_{W^1_\infty(K)}, \\ \|\underline{d}_K - \operatorname{div}\mathbf{b}\|_{L_\infty(K)} &\lesssim \operatorname{diam}(K)|\operatorname{div}\mathbf{b}|_{W^1_\infty(K)}, \\ \||\underline{\mathbf{b}}_K - \mathbf{b}|\|_{L_\infty(K)} &\le D\operatorname{diam}(K)|\mathbf{b}|_{W^1_\infty(K)^n}, \end{aligned} \tag{4.23}$$

where, as the constant in the first two inequalities, $D > 0$ is some constant depending only on n, which we name for use in (4.26) below.

In particular, we let

$$\bar{\sigma} > 0 \tag{4.24}$$

be such that for any $0 < \sigma \leq \bar{\sigma}$ and $H \in \mathcal{I}$, Ω_h is sufficiently fine to ensure that

$$\operatorname{diam}(K)\, \||\mathbf{b}|^{-1}\|_{L_\infty(K)} \max\left(1, D|\mathbf{b}|_{W^1_\infty(K)^n}\right) \leq \tfrac{1}{2} \quad (K \in \Omega_h). \tag{4.25}$$

Then for any $K \in \Omega_h$, we have

$$\begin{aligned} |\underline{\mathbf{b}}_K| &\geq \||\mathbf{b}|^{-1}\|^{-1}_{L_\infty(K)} - \||\underline{\mathbf{b}}_K - \mathbf{b}|\|_{L_\infty(K)} \\ &\geq \||\mathbf{b}|^{-1}\|^{-1}_{L_\infty(K)} - D \operatorname{diam}(K)|\mathbf{b}|_{W^1_\infty(K)^n} \\ &\geq \tfrac{1}{2}\||\mathbf{b}|^{-1}\|^{-1}_{L_\infty(K)} \geq \max\left(\tfrac{1}{2}\||\mathbf{b}|^{-1}\|^{-1}_{L_\infty(\Omega)}, \operatorname{diam}(K)\right), \end{aligned} \tag{4.26}$$

where we have used (4.25).

Finally, for $H \in \mathcal{I}$, $K \in \Omega_H \cup \Omega_h$, and $k \geq \ell \in \mathbb{N}_0$, we will make repeated use of the *inverse inequality*

$$|\cdot|_{H^k(K)} \lesssim \operatorname{diam}(K)^{-(k-\ell)}\|\cdot\|_{H^\ell(K)} \quad \text{on } \mathcal{P}_m(K),$$

where the constant depends only on m, ϱ, $\bar{\varrho}$, and k.

The main technical ingredients needed to prove Theorem 4.8 are collected in the following lemma.

Lemma 4.10. *Assume* (4.19), (4.21), *and* (4.17). *Then there exists a* $0 < \sigma_0 \leq \bar{\sigma}$ *(cf.* (4.24)*), such that for any* $\sigma \leq \sigma_0$*, one has for all* $(u, w) \in \prod_{K\in\Omega_H} \mathcal{P}_m(K) \times \prod_{K\in\Omega_H} \mathcal{P}_m(K) \cap H_{0,\Gamma_-}(\mathbf{b}; \Omega))$,

$$\begin{aligned} &\|\breve{t}\|_{H(\mathbf{b}_h;\Omega_h)} \gtrsim \|(u, w)\|_{\mathbb{U}}, \quad \|t - \breve{t}\|_{H(\mathbf{b}_h;\Omega_h)} \lesssim \sigma\|\breve{t}\|_{H(\mathbf{b}_h;\Omega_h)}, \\ &\sum_{K\in\Omega_h} \operatorname{diam}(K)^2\|\breve{t}|_K\|^2_{H^1(K)} \lesssim \sigma^2\|\breve{t}\|^2_{H(\mathbf{b}_h;\Omega_h)}, \end{aligned} \tag{4.27}$$

where the constants depend only on (upper bounds for) m, ϱ, $\bar{\varrho}$, $\|\mathbf{b}\|_{W^1_\infty(\operatorname{div};\Omega)}$, $\||\mathbf{b}|^{-1}\|_{L_\infty(\Omega)}$, $\|c\|_{W^1_\infty(\Omega)}$, *and* $\|B^{-1}\|_{\mathcal{L}(L_2(\Omega), H_{0,\Gamma_-}(\mathbf{b};\Omega))}$.

We defer the proof of this lemma to the end of this section and show first how it is used to complete the proof of Theorem 4.8 following steps (i) and (ii) announced in Section 4.1.

Proof of Theorem 4.8. For the selection of $\mathbf{b}_h$, c_h and d_h from (4.18), the perturbed bilinear form on $(L_2(\Omega) \times H(\mathbf{b}; \Omega)) \times H(\mathbf{b}_h; \Omega_h)$, first mentioned in (4.5)–(4.6), reads as

$$\begin{aligned} \breve{b}_h(u, w; v) &:= \int_\Omega (c_h v - \mathbf{b}_h \cdot \nabla_h v)u + d_h v w \, d\mathbf{x} + \int_{\partial\Omega_h} [\![v\mathbf{b}_h]\!] w \, d\mathbf{s} \\ &= \sum_{K\in\Omega_h} \int_K \underline{c}_K v u + (w - u)\underline{\mathbf{b}}_K \cdot \nabla v + v\underline{\mathbf{b}}_K \cdot \nabla w + \underline{d}_K v w \, d\mathbf{x}. \end{aligned} \tag{4.28}$$

Recall from (4.11) that the optimal test function t, defined in (4.22), was constructed such that

$$\sum_{K\in\Omega_h} \langle\!\langle t|_K, v|_K \rangle\!\rangle_{K,\underline{\mathbf{b}}_K} = \breve{b}_h(u, w; v) \quad (v \in H(\mathbf{b}_h; \Omega_h)).$$

Therefore, since for $\sigma \leq \bar{\sigma}$, $\operatorname{diam}(K) \leq |\underline{\mathbf{b}}_K|$ by (4.26), upon taking $\sigma_0 \leq \bar{\sigma}$, Proposition 4.3 applies and Remark 4.5 ensures that

$$\|t\|^2_{H(\mathbf{b}_h;\Omega_h)} \eqsim \sum_{K\in\Omega_h} \langle\!\langle t|_K, t|_K\rangle\!\rangle_{K,\underline{\mathbf{b}}_K} = \breve{b}_h(u,w;t). \tag{4.29}$$

For $(u,w) \in \mathbb{U}^H$, applying the inverse inequality in combination with (4.23), shows that $\|(\underline{\mathbf{b}}_K - \mathbf{b}|_K)\cdot\nabla\breve{t}|_K\|_{L_2(K)} \lesssim \|\breve{t}|_K\|_{L_2(K)}$ so that

$$\|\breve{t}\|_{H(\mathbf{b}_h;\Omega_h)} \eqsim \|\breve{t}\|_{\mathbb{V}}. \tag{4.30}$$

For $\sigma_0 > 0$ sufficiently small, the second inequality in Lemma 4.10 gives $\|\breve{t}\|_{H(\mathbf{b}_h;\Omega_h)} \eqsim \|t\|_{H(\mathbf{b}_h;\Omega_h)}$. We infer that

$$\breve{b}_h(u,w;t) \eqsim \|\breve{t}\|^2_{H(\mathbf{b}_h;\Omega_h)} \eqsim \|\breve{t}\|_{H(\mathbf{b}_h;\Omega_h)}\|\breve{t}\|_{\mathbb{V}} \gtrsim |\!|\!|(u,w)|\!|\!|_{\mathbb{U}}\|\breve{t}\|_{\mathbb{V}},$$

by the first inequality in Lemma 4.10.

Writing $\int_{\partial\Omega_h} [\![v\mathbf{b}_h]\!] w\,d\mathbf{s} = \sum_{K\in\Omega_h}\int_K w\underline{\mathbf{b}}_K\cdot\nabla v + v\mathbf{b}|_K\cdot\nabla w + (\underline{\mathbf{b}}_K - \mathbf{b}|_K)\cdot\nabla w\,d\mathbf{x}$, we infer that $\breve{b}_h$ is bounded on $\mathbb{U}^H \times H(\mathbf{b}_h;\Omega_h)$, uniformly in H. Consequently, we have

$$|\breve{b}_h(u,w;\breve{t}) - \breve{b}_h(u,w;t)| \lesssim |\!|\!|(u,w)|\!|\!|_{\mathbb{U}}\|t-\breve{t}\|_{H(\mathbf{b}_h;\Omega_h)} \lesssim \sigma|\!|\!|(u,w)|\!|\!|_{\mathbb{U}}\|\breve{t}\|_{\mathbb{V}}, \tag{4.31}$$

where we have again used the second inequality in Lemma 4.10 and (4.30). We conclude that for $\sigma \leq \sigma_0$ sufficiently small,

$$\breve{b}_h(u,w;\breve{t}) \gtrsim |\!|\!|(u,w)|\!|\!|_{\mathbb{U}}\|\breve{t}\|_{\mathbb{V}},$$

which is step (i) from Section 4.1.

As for step (ii), we have for $(u,w) \in \mathbb{U}$

$$b_h(u,w|_{\partial\Omega_h};v) := \sum_{K\in\Omega_h}\int_K (c - \operatorname{div}\mathbf{b})vu + (w-u)\mathbf{b}\cdot\nabla v + v\mathbf{b}\cdot\nabla w + vw\operatorname{div}\mathbf{b}\,d\mathbf{x}.$$

Applying (4.23) and subsequently the third inequality of (4.27) in Lemma 4.10, we obtain for $(u,w) \in \mathbb{U}^H$

$$\begin{aligned} &|b_h(u,w|_{\partial\Omega_h};\breve{t}) - \breve{b}_h(u,w;\breve{t})| \\ &\lesssim \sum_{K\in\Omega_h}\operatorname{diam}(K)\left[|\!|\!|(u,w)|\!|\!|_{\mathbb{U}}\|\breve{t}\|_{H^1(K)} + \|\breve{t}\|_{L_2(K)}\|w\|_{H^1(K)}\right] \\ &\lesssim |\!|\!|(u,w)|\!|\!|_{\mathbb{U}}\sqrt{\sum_{K\in\Omega_h}\operatorname{diam}(K)^2\|\breve{t}\|^2_{H^1(K)}} \\ &\qquad\qquad + \|\breve{t}\|_{L_2(\Omega)}\sigma\sqrt{\sum_{K'\in\Omega_H}\operatorname{diam}(K')^2\|w\|^2_{H^1(K')}} \\ &\lesssim \sigma|\!|\!|(u,w)|\!|\!|_{\mathbb{U}}\|\breve{t}\|_{H(\mathbf{b}_h;\Omega_h)} \lesssim \sigma|\!|\!|(u,w)|\!|\!|_{\mathbb{U}}\|\breve{t}\|_{\mathbb{V}}, \end{aligned} \tag{4.32}$$

where we have applied the inverse inequality to $w|_{K'}$ for $K' \in \Omega_H$ and, finally, (4.30). Estimate (4.32) is step (ii) from Section 4.1 which, together with step (i) completes the proof of Theorem 4.8. □

Proof of Lemma 4.10. To show the first inequality in (4.27), we will sum over $K \in \Omega_h$ the inequality (4.16) in Lemma 4.7. We start by showing below in (4.37) that the resulting right-hand side can be made small enough. To exploit that u and w

are piecewise polynomial w.r.t. the "coarse grid" Ω_H, we collect all $K \in \Omega_h$ that are contained in one $K' \in \Omega_H$.

To arrive at (4.37) we need, in particular, to get rid of the derivatives of u and to switch from $\|w\|_{H(\mathbf{b}_h;\Omega)}$ to $\|w\|_{H(\mathbf{b};\Omega)}$. To this end, an easy consequence of the third estimate in (4.23) is $\||\underline{\mathbf{b}}_K|\|_{L_\infty(K)} \lesssim \|\mathbf{b}\|_{W^1_\infty(K)^n}$ for $K \in \Omega_h$. Together with an application of the inverse inequality on $K' \in \Omega_H$, this shows that

$$\sum_{\{K\in\Omega_h\colon K\subset K'\}} \|\partial_{\underline{\mathbf{b}}_K} u\|^2_{L_2(K)} \lesssim |u|^2_{H^1(K')} \lesssim \operatorname{diam}(K')^{-2}\|u\|^2_{L_2(K')}, \tag{4.33}$$

with a constant depending on m, ϱ, and $\|\mathbf{b}\|_{W^1_\infty(\Omega)^n}$. Next, combining again the third inequality in (4.23) with an inverse estimate on $K'' \in \{K, K'\}$ yields

$$\begin{aligned}\|\partial_{\underline{\mathbf{b}}_{K''}} w\|^2_{L_2(K'')} &\leq 2\big\{\|\partial_{\mathbf{b}} w\|^2_{L_2(K'')} + \|(\mathbf{b}-\underline{\mathbf{b}}_{K''})\cdot\nabla w\|^2_{L_2(K'')}\big\}\\ &\lesssim \|w\|^2_{H(\mathbf{b};K'')},\end{aligned} \tag{4.34}$$

with a constant depending on ρ or $\bar{\rho}$, and on $m, D, \|\mathbf{b}\|_{W^1_\infty(\Omega)^n}$.

The terms $\|\partial^2_{\underline{\mathbf{b}}_K} w\|^2_{L_2(K)}$ require a little more care than $\|\partial_{\underline{\mathbf{b}}_K} u\|^2_{L_2(K)}$ since unlike u, $\partial_{\underline{\mathbf{b}}_K} w$ is generally not piecewise polynomial w.r.t. Ω_H. Therefore, we first use that for $\Omega_h \ni K \subset K' \in \Omega_H$,

$$\|\partial^2_{\underline{\mathbf{b}}_K} w\|^2_{L_2(K)} \leq 2\big\{\|\partial_{\underline{\mathbf{b}}_K}(\partial_{\underline{\mathbf{b}}_K} - \partial_{\underline{\mathbf{b}}_{K'}})w\|^2_{L_2(K)} + \|\partial_{\underline{\mathbf{b}}_K}\partial_{\underline{\mathbf{b}}_{K'}} w\|^2_{L_2(K)}\big\}.$$

For the second term on the right an application of (4.33) with u reading as $\partial_{\underline{\mathbf{b}}_{K'}} w$ shows that

$$\begin{aligned}\sum_{\{K\in\Omega_h\colon K\subset K'\}} \|\partial_{\underline{\mathbf{b}}_K}\partial_{\underline{\mathbf{b}}_{K'}} w\|^2_{L_2(K)} &\lesssim \operatorname{diam}(K')^{-2}\|\partial_{\underline{\mathbf{b}}_{K'}} w\|^2_{L_2(K')}\\ &\lesssim \operatorname{diam}(K')^{-2}\|w\|^2_{H(\mathbf{b};K')},\end{aligned} \tag{4.35}$$

where we have used (4.34) for $K'' = K'$. For the first term on the right we derive that

$$\begin{aligned}&\|\partial_{\underline{\mathbf{b}}_K}(\partial_{\underline{\mathbf{b}}_K} - \partial_{\underline{\mathbf{b}}_{K'}})w\|^2_{L_2(K)} = \|(\partial_{\underline{\mathbf{b}}_K} - \partial_{\underline{\mathbf{b}}_{K'}})\partial_{\underline{\mathbf{b}}_K} w\|^2_{L_2(K)}\\ &\lesssim \operatorname{diam}(K')^2|\partial_{\underline{\mathbf{b}}_K} w|^2_{H^1(K)} \lesssim \frac{\operatorname{diam}(K')^2}{\operatorname{diam}(K)^2}\|\partial_{\underline{\mathbf{b}}_K} w\|^2_{L_2(K)}\\ &\lesssim \frac{\operatorname{diam}(K')^2}{\operatorname{diam}(K)^2}\|w\|^2_{H(\mathbf{b};K)},\end{aligned} \tag{4.36}$$

where both (4.35) and (4.36) depend on m, ϱ, $\bar{\varrho}$, and $\|\mathbf{b}\|_{W^1_\infty(\Omega)^n}$.

By combining these four estimates (4.33), (4.34) for $K'' = K$, (4.35), (4.36), and using $|\underline{\mathbf{b}}_K|^{-1} \leq 2\||\mathbf{b}|^{-1}\|_{L_\infty(\Omega)}$ ((4.26)), $\operatorname{diam}(K) \leq \sigma \operatorname{diam}(K')$, and $\operatorname{diam}(K') \leq \sigma$, we infer that

$$\begin{aligned}\sum_{\{K\in\Omega_h\colon K\subset K'\}} \tfrac{\operatorname{diam}(K)^2}{|\underline{\mathbf{b}}_K|^2}\Big[\|u\|^2_{L_2(K)} + \|w\|^2_{H(\underline{\mathbf{b}}_K;K)} + \|\partial_{\underline{\mathbf{b}}_K} u\|^2_{L_2(K)} + \|\partial^2_{\underline{\mathbf{b}}_K} w\|^2_{L_2(K)}\Big]&\\ \lesssim \sigma^2\Big[\|u\|^2_{L_2(K')} + \|w\|^2_{H(\mathbf{b};K')}\Big],&\end{aligned}$$

and so

$$\begin{aligned}(4.37)\quad \sum_{K\in\Omega_h} \tfrac{\operatorname{diam}(K)^2}{|\underline{\mathbf{b}}_K|^2}\Big[\|u\|^2_{L_2(K)}+\|w\|^2_{H(\underline{\mathbf{b}}_K;K)}+\|\partial_{\underline{\mathbf{b}}_K}u\|^2_{L_2(K)}+\|\partial^2_{\underline{\mathbf{b}}_K}w\|^2_{L_2(K)}\Big]\\ \lesssim \sigma^2 |\!|\!|(u,w)|\!|\!|^2_{\mathbb{U}},\end{aligned}$$

where the constant depends on m, ϱ, $\bar{\varrho}$, $\|\mathbf{b}\|_{W^1_\infty(\Omega)^n}$, and $\||\mathbf{b}|^{-1}\|_{L_\infty(\Omega)}$.

To treat next the terms on the left-hand side of (4.16) analogous arguments, preceded by applications of the triangle inequality, show that

$$\begin{aligned}&\Big|\|\mathbf{b}_h\cdot\nabla_h w+(c_h+d_h)w\|_{L_2(\Omega)}-\|\mathbf{b}\cdot\nabla w+cw\|_{L_2(\Omega)}\Big|^2\\ &\leq \sum_{K'\in\Omega_H}\sum_{\{K\in\Omega_h:\,K\subset K'\}}\|(\underline{\mathbf{b}}_K-\mathbf{b})\cdot\nabla w+(\underline{c}_K+\underline{d}_K-c)w\|^2_{L_2(K)}\\ &\lesssim \sum_{K'\in\Omega_H}\sigma^2\|w\|^2_{L_2(K')}=\sigma^2\|w\|^2_{L_2(\Omega)},\end{aligned}$$

which, upon using $\big|\|f\|^2-\|g\|^2\big|\leq\big|\|f\|-\|g\|\big|\big(2\|g\|+\big|\|f\|-\|g\|\big|\big)$, yields

$$\begin{aligned}(4.38)\quad &\Big|\|\mathbf{b}_h\cdot\nabla_h w+(c_h+d_h)w\|^2_{L_2(\Omega)}-\|\mathbf{b}\cdot\nabla w+cw\|^2_{L_2(\Omega)}\Big|\\ &\lesssim \sigma\|w\|_{L_2(\Omega)}\big[(\|\mathbf{b}\cdot\nabla w+cw\|_{L_2(\Omega)}+\sigma\|w\|_{L_2(\Omega)}\big]\lesssim\sigma\|w\|^2_{H(\mathbf{b};\Omega)}\end{aligned}$$

dependent on m, ϱ, $\bar{\varrho}$, $\|\mathbf{b}\|_{W^1_\infty(\mathrm{div};\Omega)}$, and $\|c\|_{W^1_\infty(\Omega)}$.

Now by summing the inequality (4.16) in Lemma 4.7 over $K\in\Omega_h$, substituting the estimates (4.37) and (4.38), and using that

$$\|w\|_{H(\mathbf{b};\Omega)}\leq\|\mathcal{B}^{-1}\|_{\mathcal{L}(L_2(\Omega),H_{0,\Gamma_-}(\mathbf{b};\Omega))}\|\mathbf{b}\cdot\nabla w+cw\|_{L_2(\Omega)},$$

for any $\varepsilon>0$ we arrive at

$$\begin{aligned}&\|\breve{t}\|^2_{H(\mathbf{b}_h;\Omega_h)}-\Big[\tfrac{\|\mathcal{B}^{-1}\|^{-2}_{\mathcal{L}(L_2(\Omega),H_{0,\Gamma_-}(\mathbf{b};\Omega))}}{2+4\|c-\operatorname{div}\mathbf{b}\|^2_{L_\infty(\Omega)}}\|w\|^2_{H(\mathbf{b};\Omega)}+\tfrac{\varepsilon}{1+4\varepsilon}\|u\|^2_{L_2(\Omega)}-\varepsilon\|w\|^2_{L_2(\Omega)}\Big]\\ &\gtrsim-\sigma^2\|u\|^2_{L_2(\Omega)}-\sigma\|w\|^2_{H(\mathbf{b};\Omega)},\end{aligned}$$

with a constant depending on m, ϱ, $\|\mathbf{b}\|_{W^1_\infty(\mathrm{div};\Omega)}$, $\||\mathbf{b}|^{-1}\|_{L_\infty(\Omega)}$, and $\|c\|_{W^1_\infty(\Omega)}$. By selecting ε and, subsequently, σ_0 small enough, the proof of the first estimate in (4.27) is completed.

Lemma 4.6 in combination with (4.37) shows that

$$\|t-\breve{t}\|_{H(\mathbf{b}_h;\Omega_h)}\lesssim\sigma|\!|\!|(u,w)|\!|\!|_{\mathbb{U}}.$$

Now the second estimate follows from the first.

To prove the last estimate, we split $\breve{t}=\breve{t}_1+\breve{t}_2+\breve{t}_3$ (see (4.15)), where, for $K'\in\Omega_H$, $K\in\Omega_h$ with $K\subset K'$,

$$\begin{aligned}\breve{t}_1|_K(x,\mathbf{y})&:=|\underline{\mathbf{b}}_K|^{-1}\Big(w(\bar{x}_-(\mathbf{y}),\mathbf{y})-u(\bar{x}_-(\mathbf{y}),\mathbf{y})\Big)\Big(x-\bar{x}_-(\mathbf{y})\Big),\\ \breve{t}_2|_K(x,\mathbf{y})&:=\partial_{\underline{\mathbf{b}}_{K'}}w(\bar{x}_-(\mathbf{y}),\mathbf{y})+\underline{c}_K u(\bar{x}_-(\mathbf{y}),\mathbf{y})+\underline{d}_K w(\bar{x}_-(\mathbf{y}),\mathbf{y})\\ \breve{t}_3|_K(x,\mathbf{y})&:=(\underline{\mathbf{b}}_K-\underline{\mathbf{b}}_{K'})\cdot\nabla w(\bar{x}_-(\mathbf{y}),\mathbf{y}).\end{aligned}$$

Since $\breve{t}_1|_K \in \mathcal{P}_{m+1}(K)$ vanishes on $\partial\bar{K}_-$, the inverse inequality, Proposition 4.3 and (4.13) show that

$$\begin{aligned}\|\breve{t}_1|_K\|_{H^1(K)} &\lesssim \operatorname{diam}(K)^{-1}\|\breve{t}_1|_K\|_{L_2(K)} \leq \operatorname{diam}(K)^{-1}\|\breve{t}_1|_K\|_{L_2(\bar{K})} \\ &\lesssim \tfrac{\operatorname{diam}(\bar{K})}{\operatorname{diam}(K)}\|\partial_{\underline{\mathbf{b}}_K}\breve{t}_1|_K\|_{L_2(\bar{K})} \lesssim \|\partial_{\underline{\mathbf{b}}_K}\breve{t}|_K\|_{L_2(K)} \leq \|\breve{t}|_K\|_{H(\underline{\mathbf{b}}_K;K)},\end{aligned} \tag{4.39}$$

with a constant depending on $\bar{\varrho}$.

To treat $\breve{t}_2$, let $K \in \Omega_h$ and $p \in \mathcal{P}_m(K)$. Recalling from (4.14) that $|\bar{x}_-|_{W^1_\infty(K)} \lesssim 1$, we have

$$\|\mathbf{x} \mapsto p(\bar{x}_-(\mathbf{y}),\mathbf{y})\|_{H^1(K)} \lesssim |K|^{\frac{1}{2}}\|\mathbf{x} \mapsto p(\bar{x}_-(\mathbf{y}),\mathbf{y})\|_{W^1_\infty(K)} \lesssim |K|^{\frac{1}{2}}\|p\|_{W^1_\infty(K)},$$

also with a constant depending on $\bar{\varrho}$. Now consider a $p \in \mathcal{P}_m(K')$ for a $K' \in \Omega_H$. Then the combination of the previous result and the inverse inequality on K' show that

$$\begin{aligned}\sum_{\{K\in\Omega_h\colon K\subset K'\}} \operatorname{diam}(K)^2\|\mathbf{x} \mapsto p(\bar{x}_-(\mathbf{y}),\mathbf{y})\|^2_{H^1(K)} &\lesssim \sigma^2 \operatorname{diam}(K')^2|K'|\|\|p\|^2_{W^1_\infty(K')} \\ &\lesssim \sigma^2|K'|\|\|p\|^2_{L_\infty(K')} \lesssim \sigma^2\|p\|^2_{L_2(K')},\end{aligned}$$

dependent on ϱ, $\bar{\varrho}$, and m. By applying this to $\breve{t}_2$, we obtain

$$\begin{aligned}\sum_{K\in\Omega_h} \operatorname{diam}(K)^2\|\breve{t}_2|_K\|^2_{H^1(K)} &\lesssim \sigma^2\Big[\|u\|^2_{L_2(\Omega)} + \|w\|^2_{L_2(\Omega)} + \sum_{K'\in\Omega_H}\|\partial_{\underline{\mathbf{b}}_{K'}}w\|^2_{L_2(K')}\Big] \\ &\lesssim \sigma^2\|\!|(u,w)|\!\|^2_{\mathbb{U}} \lesssim \sigma^2\|\breve{t}\|^2_{H(\mathbf{b}_h;\Omega_h)},\end{aligned} \tag{4.40}$$

with a constant depending on ϱ, $\bar{\varrho}$, m, $\|c\|_{L_\infty(\Omega)}$, and $\|\mathbf{b}\|_{W^1_\infty(\operatorname{div};\Omega)}$, where we used (4.34) in the second but last step as well as the first inequality in (4.27) in the last step.

For $\Omega_h \ni K \subset K' \in \Omega_H$, using $\|\underline{\mathbf{b}}_K - \underline{\mathbf{b}}_{K'}\|_{L_\infty(K)} \lesssim \operatorname{diam}(K')|\mathbf{b}|_{W^1_\infty(K')^n}$ and (4.14), we estimate

$$\begin{aligned}\sum_{\{K\in\Omega_h\colon K\subset K'\}} \operatorname{diam}(K)^2\|\breve{t}_3|_K\|^2_{H^1(K)} &\leq \sum_{\{K\in\Omega_h\colon K\subset K'\}} \operatorname{diam}(K)^2|K|\|\breve{t}_3|_K\|^2_{W^1_\infty(K)} \\ &\lesssim \sum_{\{K\in\Omega_h\colon K\subset K'\}} \operatorname{diam}(K)^2|K|\operatorname{diam}(K')^2\|w|_{K'}\|^2_{W^2_\infty(K')} \\ &\lesssim \sum_{\{K\in\Omega_h\colon K\subset K'\}} \operatorname{diam}(K)^2|K|\operatorname{diam}(K')^{-2}\|w|_{K'}\|^2_{L_\infty(K')} \\ &\leq \sigma^2|K'|\|w|_{K'}\|^2_{L_\infty(K')} \lesssim \sigma^2\|w|_{K'}\|^2_{L_2(K')},\end{aligned}$$

with a constant depending on ϱ, $\bar{\varrho}$, m and $|\mathbf{b}|_{W^1_\infty(K')^n}$. Thus, we conclude that

$$\sum_{K\in\Omega_h} \operatorname{diam}(K)^2\|\breve{t}_3|_K\|^2_{H^1(K)} \lesssim \sigma^2\|w\|^2_{L_2(\Omega)} \lesssim \sigma^2\|\breve{t}\|^2_{H(\mathbf{b}_h;\Omega_h)} \tag{4.41}$$

using again the first inequality in (4.27) of this lemma. Combining (4.39), (4.40), and (4.41), completes the proof of the last claim of this lemma. ∎

5. Some numerical results

On $\Omega = (0,1)^2$, and for $\mathbf{b} \in W^1_\infty(\operatorname{div};\Omega)$ with $|\mathbf{b}|^{-1} \in L_\infty(\Omega)$ and $\mathbf{u} \mapsto \mathbf{b}\cdot\nabla\mathbf{u} \in \mathcal{L}\mathrm{is}(H_{0,\Gamma_-}(\mathbf{b};\Omega), L_2(\Omega))$, we consider the transport problem

$$\begin{cases} \mathbf{b}\cdot\nabla u = f & \text{on } \Omega, \\ u = 0 & \text{on } \Gamma_-. \end{cases}$$

We let Ω_H be a partition of Ω into uniformly shape regular triangles, and let Ω_h be the refinement of Ω_H by applying ℓ recursive red-refinements to each $K \in \Omega$, where $\ell \in \mathbb{N}_0$ is a fixed number. We let

$$\begin{aligned} \mathbb{U}^H &= \prod_{K\in\Omega_H} \mathcal{P}_{m-1}(K) \;\times\; \Big\{ w|_{\partial\Omega} \colon w \in C(\Omega) \cap \prod_{K\in\Omega_H} \mathcal{P}_m(K),\, w = 0 \text{ on } \Gamma_- \Big\}, \\ \mathbb{V}^h &= \prod_{K\in\Omega_h} \mathcal{P}_{m+1}(K). \end{aligned}$$

One infers that for $w|_{\partial\Omega_h}$ to belong to $H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)$, $w \in \prod_{K\in\Omega_H}\mathcal{P}_m(K)$ has to be continuous at any intersection of an inflow and outflow face of a $K \in \Omega_H$ (cf. [GMS15, Thm. 4.7]).

We solve $(u^H, \theta^H) \in \mathbb{U}^H$ from

$$\begin{aligned} b_h(u^H,\theta^H;v^h) &:= -\int_\Omega (\mathbf{b}\cdot\nabla_h v^h + v^h \operatorname{div}\mathbf{b}) u^H \, d\mathbf{x} + \int_{\partial\Omega_h} [\![v^h \mathbf{b}]\!] \theta^H \, d\mathbf{s} \\ &= f(v^h) \qquad (v^h \in \mathcal{T}^h(\mathbb{U}^H)), \end{aligned} \tag{5.1}$$

with $\mathcal{T}^h \in \mathcal{L}(L_2(\Omega)\times H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h), \mathbb{V}^h)$ being defined by

$$\langle \mathcal{T}^h(u,\theta), v^h\rangle_{H(\mathbf{b};\Omega_h)} = b_h(u,\theta;v^h) \quad (v^h \in \mathbb{V}^h). \tag{5.2}$$

Note that for (u,θ) running over an obvious localized basis for $\mathbb{U}^H$, finding each of the $\mathcal{T}^h(u,\theta)$ amounts to solving a fixed finite dimensional problem on a few mesh cells. Having determined such a basis for $\mathcal{T}^h(\mathbb{U}^H)$, the solution of (5.1) can be found by solving the sparse, symmetric positive definite system

$$\langle \mathcal{T}^h(u^H,\theta^H), \mathcal{T}^h(\tilde{u}^H,\tilde{\theta}^H)\rangle_{H(\mathbf{b};\Omega_h)} = f(\mathcal{T}^h(\tilde{u}^H,\tilde{\theta}^H)) \quad ((\tilde{u}^H,\tilde{\theta}^H) \in U^H).$$

As shown in Theorem 4.8, by taking a sufficiently large, but fixed ℓ,

$$\begin{aligned} \|u - u^H\|_{L_2(\Omega)} &+ \|\theta-\theta^H\|_{H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)} \\ &\lesssim \inf_{(\bar{u}^H,\bar{\theta}^H)\in\mathbb{U}^H} \big\{ \|u-\bar{u}^H\|_{L_2(\Omega)} + \|\theta-\bar{\theta}^H\|_{H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)} \big\}. \end{aligned} \tag{5.3}$$

In all our experiments, we only measure $\|u-u^H\|_{L_2(\Omega)}$, being the quantity of our main interest. We report on cases where $m = 1$, so on piecewise constant approximations for u and piecewise linear approximations for θ. It appears that in all these cases it is sufficient to take $\ell = 0$, i.e., $\Omega_h = \Omega_H$. Increasing ℓ leaves the numerical solutions essentially unchanged. This holds for true for $m = 1$, as well as in experiments that we performed where $m > 1$.

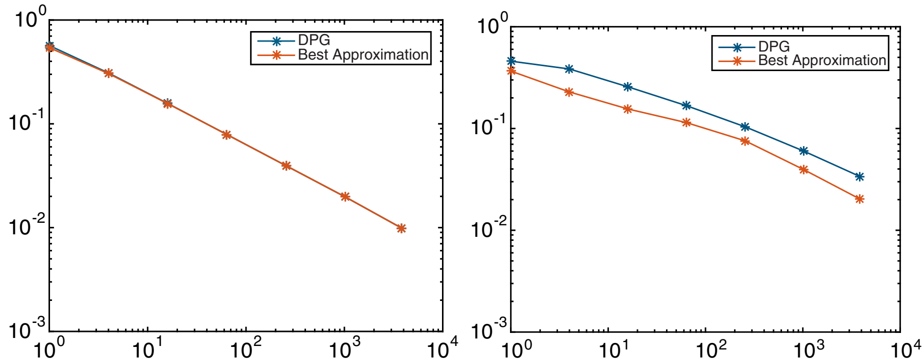

FIGURE 2. $L_2(\Omega)$-error in u_H and that in the best approximation versus $1/h^2$ ($\eqsim \dim \mathbb{U}_H$), for $f(\mathbf{x}) = 1 - x_1$, $\mathbf{b} = (1,1)^\top$ (left) and $\mathbf{b} = (1, \frac{1}{16})^\top$ (right).

In our first experiment, we take constant $\mathbf{b} = (b_1, b_2)^\top \in \mathbb{R}_{>0} \times \mathbb{R}_{\geq 0}$, Ω_H being a uniform partition of Ω into isosceles right-angled triangles with legs of length $H \in 2^{-\mathbb{N}_0}$ and hypothenuses parallel to the vector $(1,1)$, and $\Omega_h = \Omega_H$ so $\ell = 0$. We take $f(\mathbf{x}) = 1 - x_1$ so that the exact solution, given by

$$u(\mathbf{x}) = \begin{cases} \frac{x_1}{b_1} - \frac{x_1^2}{2b_1}, & -b_2 x_1 + b_1 x_2 \ \geq \ 0, \\ \frac{x_2}{b_2} - \frac{x_2(2b_2x_1 - b_1x_2)}{2b_2^2}, & -b_2 x_1 + b_1 x_2 \ < \ 0, \end{cases}$$

is continuous, piecewise quadratic, whose normal derivative over the line $\mathbf{x} \cdot \mathbf{b}^\perp = 0$ has a jump. The numerical results for various $\mathbf{b}$, illustrated in Figure 2 for $\mathbf{b} = (1,1)^\top$ and $\mathbf{b} = (1, 1/16)^\top$, show that $\|u - u^H\|_{L_2(\Omega)}$ is close to the error of best $L_2(\Omega)$-approximation from the space of piecewise constants.

In our second experiment, we change f into

$$f(\mathbf{x}) = \begin{cases} 1 - x_1, & -b_2 x_1 + b_1 x_2 \ \geq \ \frac{1}{4}, \\ 0, & -b_2 x_1 + b_1 x_2 \ < \ \frac{1}{4}, \end{cases} \tag{5.4}$$

so that the solution, given by

$$u(x_1, x_2) = \begin{cases} \frac{x_1}{b_1} - \frac{x_1^2}{2b_1}, & -b_2 x_1 + b_1 x_2 \ \geq \ \frac{1}{4}, \\ 0, & -b_2 x_1 + b_1 x_2 \ < \ \frac{1}{4}, \end{cases}$$

is piecewise quadratic with a discontinuity over the line $\mathbf{x} \cdot \mathbf{b}^\perp = \frac{1}{4}$.

When $h = 2^{-k}$ for $k \geq 2$ and $\mathbf{b} \in \{(1,0)^\top, (1,1)^\top\}$, then this discontinuity is over a grid line, and the right-hand side of (5.3) will be strongly dominated by the approximation error in θ, because the approximation error in u benefits from the discontinuous approximation. In this, rather special situation, the error $\|u - u_H\|_{L_2(\Omega)}$ might therefore be much larger than the error of best approximation in u. Unfortunately, this is indeed what happens as illustrated in Figure 3.

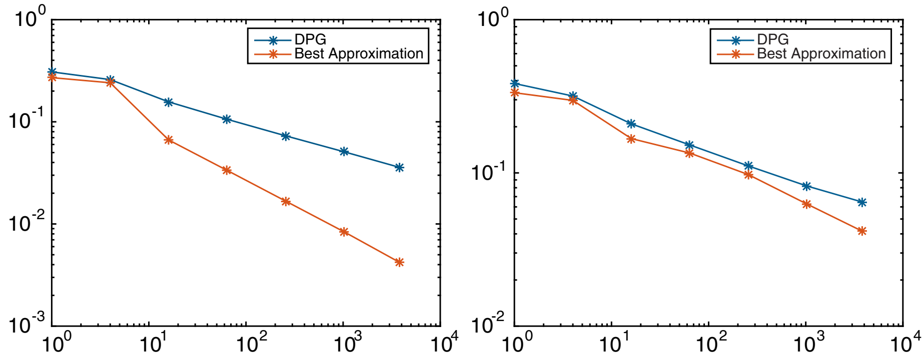

FIGURE 3. $L_2(\Omega)$-error in u_H and that in the best approximation versus $1/h^2$, for the discontinuous f from (5.4), $\mathbf{b} = (1,1)^\top$ (left) and $\mathbf{b} = (1, \frac{1}{16})^\top$ (right).

To deal with this difficulty, we replaced the trial space for θ by the space of discontinuous polynomials $\prod_e \mathcal{P}_1(e)$ with e running over all edges of the mesh skeleton $\partial\Omega_H$ without the inflow edges, and we determined the new test space $\mathcal{T}^h(\mathbb{U}^H)$ from (5.2) again with $\mathbb{V}^h = \prod_{K\in\Omega_h} \mathcal{P}_2(K)$. With this modification, the curve of the $L_2(\Omega)$-error in the resulting u_H is indistinguishable from that of the error in the best approximation. Since, as we have seen, $(\prod_e \mathcal{P}_1(e))|_{\partial\Omega_h} \not\subset H_{0,\Gamma_-}(\mathbf{b};\Omega)$, we are now dealing with a *nonconforming* DPG method.

Remark 5.1. This discontinuous trial space was already considered in the first paper [DG11] in which such DPG discretizations (there called DPG-A) for the transport problem were considered. Since instead of the correct space $H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)$, there the space $L_2(\partial\Omega_h)$ was considered as the space for the traces θ, the use of a nonconforming trial space was unintended. In [HKS14] one can find an analysis of a nonconforming DPG discretization for the Poisson problem. The analysis of the above nonconforming DPG discretization of the transport problem is open.

In our third experiment we took $\mathbf{b}(\mathbf{x}) = (x_2, -x_1)^\top$, $f = 0$, and the inhomogeneous boundary condition $u = g$ on Γ_-, where $g(x_1, 1) = 0$, and

$$g(0, x_2) = \begin{cases} 1, & x_2 \geq \frac{1}{4}, \\ 0, & x_2 < \frac{1}{4}. \end{cases}$$

To implement this inhomogeneous boundary condition, following the second approach discussed in Remark 3.6 we solved $(u^H, \theta^H) \in \mathbb{U}^H$ from

$$b_h(u^H, \theta^H; v^h) = -\int_{\partial\Omega_h} [\![v^h \mathbf{b}]\!] \bar{g}\, ds \qquad (v^h \in \mathcal{T}^h(\mathbb{U}^H)),$$

with $\bar{g} \in H(\mathbf{b};\Omega)$ being an extension of g. We took

$$\bar{g}(\mathbf{x}) = \begin{cases} 1, & |\mathbf{x}| \in [\frac{1}{4}, 1], \\ 0, & \text{elsewhere}, \end{cases}$$

which in this case equals the exact solution.

In this experiment, we employed an adaptive refinement strategy, that we implemented using the package iFEM by L. Chen ([C09]). By an application of Theorem 3.1 and Riesz' representation theorem, $r \in H(\mathbf{b};\Omega_h)$, defined by

$$\langle r, v\rangle_{H(\mathbf{b};\Omega_h)} = \int_{\partial\Omega_h} [\![v\mathbf{b}]\!] \bar{g}\, d\mathbf{s} - b_h(u^H, \theta^H; v) \quad (v \in H(\mathbf{b};\Omega_h)),$$

satisfies $\|r\|^2_{H(\mathbf{b};\Omega_h)} \eqsim \|u - u^H\|^2_{L_2(\Omega)} + \|\theta - \theta^H\|^2_{H_{0,\Gamma_-}(\mathbf{b};\partial\Omega_h)}$. We approximated r by the solution $\tilde{r} \in \mathbb{V}^h$ of

$$\langle \tilde{r}, v^h\rangle_{H(\mathbf{b};\Omega_h)} = \int_{\partial\Omega_h} [\![v^h\mathbf{b}]\!] \bar{g}\, d\mathbf{s} - b_h(u^H, \theta^H; v^h) \quad (v^h \in \mathbb{V}^h).$$

Based on the decomposition $\|\tilde{r}\|^2_{H(\mathbf{b};\Omega_h)} = \sum_{K\in\Omega^H} \|\tilde{r}\|^2_{H(\mathbf{b};K)}$, as local error indicators we used $\{\|\tilde{r}\|^2_{H(\mathbf{b};K)} : K \in \Omega^H\}$ to drive the common adaptive finite element method (AFEM) with Dörfler marking parameter $\vartheta = \frac{1}{2}$. Examples of a resulting mesh and approximate solution are given in Figure 4, and the $L_2(\Omega)$-errors vs. the number of unknowns are illustrated in Figure 5.

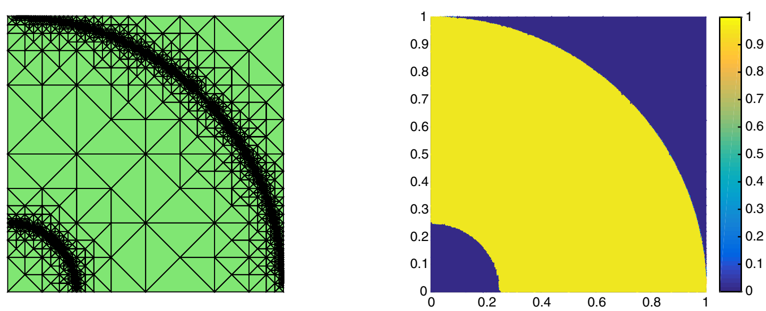

FIGURE 4. Mesh generated after some iterations (left) and the approximate solution (right) for the third experiment.

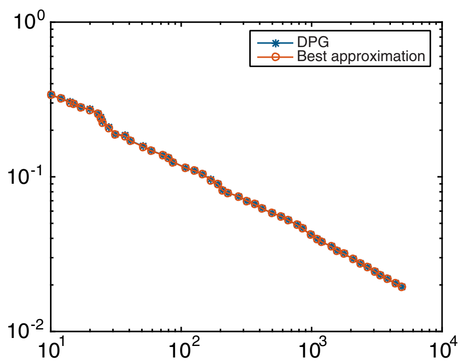

FIGURE 5. $L_2(\Omega)$-error in u_H and that in the best approximation versus the number of triangles in the mesh for the third experiment.

6. Conclusion

For a family of uniformly shape regular partitions $\Omega_H, H \in \mathcal{I}$ and given piecewise polynomial trial spaces on Ω_H and on the skeleton $\partial\Omega_H$, we have constructed *uniformly inf-sup stable* (with respect to $H \in \mathcal{I}$) DPG discretizations for linear transport equations with variable convection fields by associating with each cell $K' \in \Omega_H$ a *piecewise polynomial* test space $\mathbb{V}_{K'}$ on a subgrid $\Omega_h|_{K'}$ with the following properties. The polynomial degree of each $\mathbb{V}_{K'}$ exceeds the degree of the trial functions by one and the refinement depth of each subgrid $\Omega_h|_{K'}$ is uniformly bounded. The stability implies that the DPG scheme provides near-best approximations from the trial space as well as uniform error-residual relations that form essential prerequisits for a posteriori error control and adaptive refinement strategies; see (1.15), (1.4). The control of the polynomial degrees in the test space as well as that of the subgrid refinement depth entail an asymptotically optimal complexity scaling since the size of the linear systems stays proportional to the dimensions of the trial spaces. Several consequences of the findings in the present paper such as rigorous computable a posteriori error bounds or applications in more complex problem settings such as kinetic models will be addressed in forthcoming work.

References

[BM84] J. W. Barrett and K. W. Morton, *Approximate symmetrization and Petrov-Galerkin methods for diffusion-convection problems*, Comput. Methods Appl. Mech. Engrg. **45** (1984), no. 1-3, 97–122, DOI 10.1016/0045-7825(84)90152-X. MR759805

[BS14] D. Broersen and R. Stevenson, *A robust Petrov-Galerkin discretisation of convection-diffusion equations*, Comput. Math. Appl. **68** (2014), no. 11, 1605–1618, DOI 10.1016/j.camwa.2014.06.019. MR3279496

[BS15] D. Broersen and R. P. Stevenson, *A Petrov-Galerkin discretization with optimal test space of a mild-weak formulation of convection-diffusion equations in mixed form*, IMA J. Numer. Anal. **35** (2015), no. 1, 39–73, DOI 10.1093/imanum/dru003. MR3335195

[CDG16] C. Carstensen, L. Demkowicz, and J. Gopalakrishnan, *Breaking spaces and forms for the DPG method and applications including Maxwell equations*, Comput. Math. Appl. **72** (2016), no. 3, 494–522, DOI 10.1016/j.camwa.2016.05.004. MR3521055

[C09] L. Chen, *iFEM: An integrated finite element method package in MATLAB*, Technical Report, University of California at Irvine, 2009.

[CDW12] A. Cohen, W. Dahmen, and G. Welper, *Adaptivity and variational stabilization for convection-diffusion equations*, ESAIM Math. Model. Numer. Anal. **46** (2012), no. 5, 1247–1273, DOI 10.1051/m2an/2012003. MR2916380

[DG11] L. Demkowicz and J. Gopalakrishnan, *A class of discontinuous Petrov-Galerkin methods. II. Optimal test functions*, Numer. Methods Partial Differential Equations **27** (2011), no. 1, 70–105, DOI 10.1002/num.20640. MR2743600

[DHSW12] W. Dahmen, C. Huang, C. Schwab, and G. Welper, *Adaptive Petrov-Galerkin methods for first order transport equations*, SIAM J. Numer. Anal. **50** (2012), no. 5, 2420–2445, DOI 10.1137/110823158. MR3022225

[DPW14] W. Dahmen, C. Plesken, and G. Welper, *Double greedy algorithms: reduced basis methods for transport dominated problems*, ESAIM Math. Model. Numer. Anal. **48** (2014), no. 3, 623–663, DOI 10.1051/m2an/2013103. MR3177860

[DSMMO04] H. De Sterck, T. A. Manteuffel, S. F. McCormick, and L. Olson, *Least-squares finite element methods and algebraic multigrid solvers for linear hyperbolic PDEs*, SIAM J. Sci. Comput. **26** (2004), no. 1, 31–54, DOI 10.1137/S106482750240858X. MR2114333

[GMS15] J. Gopalakrishnan, P. Monk, and P. Sepúlveda, *A tent pitching scheme motivated by Friedrichs theory*, Comput. Math. Appl. **70** (2015), no. 5, 1114–1135, DOI 10.1016/j.camwa.2015.07.001. MR3378991

[GQ14] J. Gopalakrishnan and W. Qiu, *An analysis of the practical DPG method*, Math. Comp. **83** (2014), no. 286, 537–552, DOI 10.1090/S0025-5718-2013-02721-4. MR3143683

[HKS14] N. Heuer, M. Karkulik, and F.-J. Sayas, *Note on discontinuous trace approximation in the practical DPG method*, Comput. Math. Appl. **68** (2014), no. 11, 1562–1568, DOI 10.1016/j.camwa.2014.07.006. MR3279493

Korteweg-de Vries Institute for Mathematics, University of Amsterdam, P.O. Box 94248, 1090 GE Amsterdam, The Netherlands
E-mail address: dirkbroersen@gmail.com

Institut für Geometrie und Praktische Mathematik, RWTH Aachen, Templergraben 55, 52056 Aachen, Germany
E-mail address: wolfgang.anton.dahmen@googlemail.com

Korteweg-de Vries Institute for Mathematics, University of Amsterdam, P.O. Box 94248, 1090 GE Amsterdam, The Netherlands
E-mail address: r.p.stevenson@uva.nl