Research Article

Wolfgang Dahmen and Rob P. Stevenson*

Adaptive Strategies for Transport Equations

Abstract: This paper is concerned with a posteriori error bounds for linear transport equations and related questions of contriving corresponding adaptive solution strategies in the context of Discontinuous Petrov Galerkin schemes. After indicating our motivation for this investigation in a wider context the first major part of the paper is devoted to the derivation and analysis of a posteriori error bounds that, under mild conditions on variable convection fields, are efficient and, modulo a data-oscillation term, reliable. In particular, it is shown that these error estimators are computed at a cost that stays uniformly proportional to the problem size. The remaining part of the paper is then concerned with the question whether typical bulk criteria known from adaptive strategies for elliptic problems entail a fixed error reduction rate also in the context of transport equations. This turns out to be significantly more difficult than for elliptic problems and at this point we can give a complete affirmative answer for a single spatial dimension. For the general multidimensional case we provide partial results which we find of interest in their own right. An essential distinction from known concepts is that global arguments enter the issue of error reduction. An important ingredient of the underlying analysis, which is perhaps interesting in its own right, is to relate the derived error indicators to the residuals that naturally arise in related least squares formulations. This reveals a close interrelation between both settings regarding error reduction in the context of adaptive refinements.

Keywords: Discontinuous Petrov Galerkin Formulation of Transport Equations, Optimal and Near-Optimal Test Spaces, Stability

MSC 2010: 65N12, 65N30, 35A15, 35F05

1 Introduction

Motivation and Goals. Adaptive solution concepts form an important component in strategies for ever advancing computational frontiers by generating discretizations whose solutions have a desired quality (e.g. in terms of accuracy) at the expense of a possibly small problem size, viz. number of degrees of freedom. Guaranteeing a certain performance and certifying the solution quality poses intrinsic mathematical challenges that have triggered numerous investigations.

It is fair to say that the most workable starting point for an adaptive method is a *variational formulation* of the problem at hand that allows one to relate errors – involving the unknown solution – to residuals – involving only known quantities. A little wrinkle lies in the fact that these residuals have to be typically evaluated in *dual norms* that are not straightforward to compute. A first important goal is therefore (A) to evaluate or approximate these residual quantities in a tight fashion, see e.g. the fundamental work of Verfürth [10]. By *tight* we mean in what follows that modulo a data oscillation term the *a posteriori bounds* are reliable as wells as efficient, i.e., up to moderate constant multiples provide upper as well as lower bounds for the error plus

*Corresponding author: Rob P. Stevenson, Korteweg-de Vries Institute for Mathematics, University of Amsterdam, P.O. Box 94248, 1090 GE Amsterdam, The Netherlands, e-mail: r.p.stevenson@uva.nl
Wolfgang Dahmen, Mathematics Department, University of South Carolina, Columbia, SC 29208, USA, e-mail: dahmen@math.sc.edu

data oscillation. This by itself is important since it allows one to quantify the solution accuracy for a given discretization without a priori knowledge about the solution such as norms of its derivatives. Aside from minimizing the size of discrete problems for a given target accuracy via adaptive strategies based on such error bounds, the availability of certified bounds is essential in a *nested iteration* context which is sometimes the only viable strategy for obtaining quantifiable results within a given computational budget.

As part of an adaptive strategy a second, often mathematically even more demanding goal (B) is to contrive a suitable *mesh refinement strategy* derived from the a posteriori residual quantities and to understand its convergence and complexity properties. The first step in this regard is to show that each step of such a refinement does decrease the current error by a fixed factor. In many works on adaptive methods this last issue is often ignored or taken for granted when using a "plausible" refinement strategy based on a posteriori indicators. However, in the context of highly convection dominated convection diffusion problems it is shown in [5] that an error reduction can be delayed until full resolution of boundary layers is established, despite the fact that robust efficient and reliable error estimators are used.

Once a fixed error reduction rate is established one then estimates in a second step the increase of degrees of freedom caused by the refinement.

Background. Both steps (A) and (B) are so far best understood for problems of elliptic type and their close relatives, see e.g. [1, 8, 9]. By this we mean, in particular, variational formulations involving *isotropic* function spaces that are essentially independent of problem parameters. Moreover, these variational formulations appear more or less in a natural way and lead to problems that are well conditioned (on the continuous infinite-dimensional level) in a sense to be made precise later. This luxury is lost abruptly already when dealing with simple linear transport equations. Our particular interest in the seemingly simple model of first order steady state linear transport equations stems from the following points. First, classical techniques for transport equations do typically not come with tight a posteriori error bounds, let alone a rigorously founded adaptive solution strategy. Second, linear transport equations form a core constituent of important kinetic models whose treatment would benefit from the availability of tight a posteriori error bounds because they would warrant a rigorous control of nested source term iterations avoiding the inversion of large linear systems which are densely populated due to global scattering operators. Last but not least, linear transport equations can be viewed as a limit case of convection dominated convection diffusion equations. Thus, appropriate variational formulations are instructive for the singularly perturbed versions as well. We are content here with the time-independent formulations since corresponding variational formulations would immediately offer space-time formulations for the time dependent case where initial conditions enter as "inflow-boundary conditions".

The classical footing for rigorous a posteriori bounds is a variational formulation of the underlying (infinite-dimensional) problem for which the induced operator is an isomorphism from the trial space onto the dual of the test space. This means errors in the trial metric are equivalent to residuals in a dual test-norm which at least in principle contains only known quantities and hence is amendable to a numerical evaluation. For transport equations, the lack of *any* diffusion is well known to cause standard Galerkin formulations being extremely ill-conditioned. This results in notoriously unstable schemes which precludes the availability of obvious tight lower and upper a posteriori error bounds. Instead, suitable variational formulations that could give rise to tight residual a posteriori bounds need to be *unsymmetric*, i.e., trial and test metrics differ from each other. In this regard the Discontinuous Petrov Galerkin (DPG) concept offers a promising framework to accommodate problem classes that are not satisfactorily treated by conventional schemes, i.e., they help identifying and numerically accessing suitable pairs of trial and test spaces. A concise discussion of DPG methods involves two stages: first, in contrast to ordinary DG methods it is important to start from a *mesh dependent infinite-dimensional* variational formulation which has to be shown to be uniformly inf-sup stable with respect to the underlying meshes. The proper choice of function spaces for the bulk as well as skeleton quantities is crucial. Second, the optimal test spaces that inherit for a given *finite-dimensional* trial space the stability of the infinite-dimensional problem are not practical. A computational version requires replacing local infinite-dimensional test-search spaces by finite-dimensional ones whose size, however, determines the computational cost. There are to our knowledge only a few results guaranteeing uniform

fully discrete stability. In the DPG context, this concerns on the one hand problems of elliptic type and their close relatives in the sense that the involved functions spaces are isotropic [4, 7]. On the other hand, we have studied in [2] an essentially different problem class, namely first order linear transport problems. There, we have proposed a *fully discrete* DPG scheme for linear transport equations with variable convection fields that is shown to be *uniformly inf-sup stable* with respect to hierarchies of shape regular meshes. The perhaps most noteworthy obstruction encountered in this context is the fact that the involved function spaces are *anisotropic* and depend on the convection field in an essential way. This means, for instance, that when perturbing the convection field the test spaces not only vary with respect to the norms but even as sets. This affects, in particular, the issue of data oscillation. Therefore, the case of variable convection fields is rather delicate and requires a very careful organization of perturbation arguments, see [2]. The present work builds on the findings in [2].

Objectives, Results, Layout of the Paper. The central objectives of this paper concern both goals (A) and (B) for linear first order transport equations. In Section 2 we briefly recall the basic DPG concepts the remaining developments will be based upon. This includes the notion of *projected optimal test spaces* as well as the principal elements of error estimation with the aid of *lifted residuals*.

In Section 3 we detail the ingredients of the transport problem and recall from [2] a corresponding DPG scheme. The level of technicality observed there is in our opinion unavoidable and stems from the three stages of the DPG concept mentioned above. To ease accessibility of the material and fix notation we recall from [2] some relevant results which the subsequent discussion will build upon.

Section 4 is devoted to goal (A) the derivation of *efficient and reliable* (in brief "tight") a posteriori error bounds. DPG schemes are often perceived as providing "natural" local error indicators ready to use for adaptive refinements. Of course, once the uniform well-posedness of the infinite-dimensional DPG formulation has been established the error in the trial metric is indeed equivalent to a Riesz-lifted residual which is in fact a sum of local terms. However, in exactly the same way as for optimal test-functions, these quantities require solving local *infinite-dimensional* Galerkin problems. Again, one has to develop a practical variant using appropriate finite-dimensional test-search spaces. To ensure a proper complexity scaling, these spaces should again have a fixed uniformly bounded finite dimension. An improper choice of such test-search spaces could result in gross under-estimation of the actual error. Thus, the central issue here is to rigorously ensure that the so called "practical" versions using localized test-search spaces of fixed finite dimension do actually capture the true infinite-dimensional residual well enough to quantitatively reflect the error plus a data oscillation term. This is done in Section 4 for *variable* convection fields under the same moderate regularity conditions as used for the uniform inf-sup stability. Again, a central issue here is a very subtle perturbation strategy that is eventually able to cope with the essential dependence of the test spaces on the convection field and the fact that the perturbations are only meaningful on the finite-dimensional level.

Finally, in Section 5 we address goal (B). As indicated earlier, the situation differs in essential ways from the key mechanisms that work for elliptic problems. A key obstruction, shared with least-squares methods for other problem types, is the fact that the error indicators do not explicitly contain any power of the local mesh-size. Hence, it is now far from obvious that a fixed local refinement actually reduces the error indicator or the error itself. This means establishing a *fixed error reduction* being guaranteed by a concrete refinement strategy becomes the main issue. In fact, we anticipate that, once error reduction is in place the analysis of the overall complexity will then follow again along more established paths. Therefore, we concentrate in Section 5 on error reduction. The main tools are carefully exploiting what may be called "Petrov–Galerkin orthogonality", and local piecewise polynomial approximations. The central focus point emerging from related attempts, however, is the fact that the tight a posteriori error indicators are actual equivalent to an entirely *mesh-free* indicator of least squares type. In fact, this latter indicator may be viewed as a certain "limit" of the DPG-indicators resulting from different approximate Riesz-lifts. This connection is in our opinion of interest in its own right. Using these concepts, we rigorously prove that refinement strategies based on a standard bulk criterion imply error-reduction in a single spatial dimension. For several space dimensions we formulate an analogous result for collections of marked cells which in certain cases are enriched in downstream direction. The necessity of such enrichments is, however, open.

Since the focus of this work is on revealing the intrinsic theoretical mechanisms, we dispense with numerical tests but hope that our findings offer new insight and will prove useful for eventually extending the current state of the art. We present in Section 6 some concluding remarks addressing, in particular, the relation between DPG and least squares schemes.

We sometimes write $a \lesssim b$ to express that a can be bounded by a fixed constant multiple of b where the multiplicative factor is independent of the relevant parameters a and b may depend on. Likewise $a \eqsim b$ means that both $a \lesssim b$ and $b \lesssim a$ hold.

2 Abstract Setting and Preliminary Observations

Transport dominated problems are prominent instances where *symmetric* variational formulations – trial and test space coincide – fail to provide well-conditioned problems already on the continuous level. This section serves two purposes. First, we briefly recap some preliminaries about *unsymmetric* Petrov–Galerkin formulations which, in particular, *Discontinuous Petrov Galerkin* (DPG) schemes are based upon. Second, we collect some general basic facts that will be used later in the a posteriori error analysis.

2.1 Petrov–Galerkin Formulation with Projected Optimal Test Spaces

Let $\mathbb{U}, \mathbb{V}$ be Hilbert spaces and $b : \mathbb{U} \times \mathbb{V} \to \mathbb{R}$ a continuous bilinear form, i.e., $|b(u; v)| \le C_b \|u\|_{\mathbb{U}} \|v\|_{\mathbb{V}}$, $u \in \mathbb{U}$, $v \in \mathbb{V}$. This means that $(\mathcal{B}u)(v) := b(u; v)$ induces a bounded linear operator from $\mathbb{U}$ to $\mathbb{V}'$, the normed dual of $\mathbb{V}$, endowed as usual with the norm $\|z\|_{\mathbb{V}'} := \sup_{v \in \mathbb{V}: \|v\|_{\mathbb{V}} = 1} z(v)$. Moreover, let us assume that $\mathcal{B}$ is an isomorphism which we express by writing $\mathcal{B} \in \mathcal{L}\text{is}(\mathbb{U}, \mathbb{V}')$. It is well known that this latter property is equivalent to the validity of the inf-sup conditions

$$\inf_{u \in \mathbb{U}} \sup_{v \in \mathbb{V}} \frac{b(u; v)}{\|u\|_{\mathbb{U}} \|v\|_{\mathbb{V}}} \ge \gamma, \quad \inf_{v \in \mathbb{V}} \sup_{u \in \mathbb{U}} \frac{b(u; v)}{\|u\|_{\mathbb{U}} \|v\|_{\mathbb{V}}} \ge \gamma, \tag{2.1}$$

for some $\gamma > 0$. One consequence of the entailed stability is the relation

$$C_b^{-1} \|f - \mathcal{B}w\|_{\mathbb{V}'} \le \|u^{\text{ex}} - w\|_{\mathbb{U}} \le \gamma^{-1} \|f - \mathcal{B}w\|_{\mathbb{V}'}, \quad w \in \mathbb{U}, \tag{2.2}$$

where $u^{\text{ex}} = \mathcal{B}^{-1} f$ is the exact solution of the problem: find $u \in \mathbb{U}$ such that

$$b(u; v) = f(v), \quad v \in \mathbb{V}. \tag{2.3}$$

Clearly, (2.2) is a natural starting point for deriving a posteriori bounds. The tightness of such bounds depends on the condition (number)

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

of problem (2.3) which can equivalently be expressed as the operator equation $\mathcal{B}u = f$.

When trying to approximate u^{ex} by some element in a finite-dimensional trial space $\mathbb{U}^\delta \subset \mathbb{U}$ ("δ" refers to "discrete") the choice of the test space becomes a central issue. A by now well-established mechanism is to choose a so called *test search space* $\bar{\mathbb{V}}^\delta \subseteq \mathbb{V}$ of dimension typically larger than $\dim \mathbb{U}^\delta$, for which

$$\bar{\gamma}^\delta := \inf_{0 \ne u \in \mathbb{U}^\delta} \sup_{0 \ne v \in \bar{\mathbb{V}}^\delta} \frac{b(u; v)}{\|u\|_{\mathbb{U}} \|v\|_{\mathbb{V}}} > 0. \tag{2.4}$$

Clearly, $\bar{\mathbb{V}}^\delta = \mathbb{V}$ would yield $\bar{\gamma}^\delta = \gamma$ so that the size of $\bar{\mathbb{V}}^\delta$ can be viewed as the "invested stabilization". Defining then the *trial-to-test map* $t^\delta = t^\delta(\bar{\mathbb{V}}^\delta) \in \mathcal{L}(\mathbb{U}, \bar{\mathbb{V}}^\delta)$ by

$$\langle t^\delta u, v \rangle_{\mathbb{V}} = b(u; v) \quad (v \in \bar{\mathbb{V}}^\delta), \tag{2.5}$$

we obtain that the function $t^\delta u$ is the $\mathbb{V}$-orthogonal projection onto $\bar{\mathbb{V}}^\delta$ of the *optimal test function* $R^{-1} \mathcal{B} u$,

where $R^{-1} : \mathbb{V}' \to \mathbb{V}$ is the inverse Riesz map (or Riesz lift). The space

$$\mathbb{V}^\delta = \mathbb{V}^\delta(\mathbb{U}^\delta, \bar{\mathbb{V}}^\delta) := \operatorname{ran} t^\delta|_{\mathbb{U}^\delta}$$

is called *projected optimal test space* because $t^\delta u$ is the $\mathbb{V}$-orthogonal projection onto $\bar{\mathbb{V}}^\delta$ of the *optimal test function*. Also note that

$$\frac{b(u; t^\delta u)}{\|t^\delta u\|_{\mathbb{V}}} = \|t^\delta u\|_{\mathbb{V}} = \sup_{0 \neq v \in \bar{\mathbb{V}}^\delta} \frac{b(u, v)}{\|v\|_{\mathbb{V}}},$$

so $\mathbb{V}^\delta$ gives the same inf-sup constant as $\bar{\mathbb{V}}^\delta$. Once (2.4) has been established for $\bar{\mathbb{V}}^\delta$, the problem of finding the *Petrov–Galerkin solution* $u^\delta = u^\delta(f, \mathbb{U}^\delta, \mathbb{V}^\delta) \in \mathbb{U}^\delta$ of

$$b(u^\delta; v) = f(v) \quad (v \in \mathbb{V}^\delta) \tag{2.6}$$

is for any $f \in \mathbb{V}'$ well posed. Moreover, the solution of (2.6) yields, up to a factor $C_b/\bar{\gamma}^\delta$ (bounding $\kappa_{\mathbb{U}, \mathbb{V}'}(\mathcal{B})$) the best approximation to $\mathcal{B}^{-1} f$ from $\mathbb{U}^\delta$. Here and below we use the superscript δ to refer to a discretization or better finite-dimensional problems.

In summary, it would of course be highly desirable to guarantee uniform stability in δ, i.e., $\gamma^\delta \geq \underline{\gamma} > 0$ in (2.4), while keeping the computational work proportional to the dimension $\dim \mathbb{U}^\delta$ of the trial spaces, viz. the number of degrees of freedom. This requires a uniform bound for the test-search spaces of the form $\dim \bar{\mathbb{V}}^\delta \lesssim \dim \mathbb{U}^\delta$. In [2] this has been shown for linear transport problems with variable convection fields which the present work will heavily build on, see also Section 3.

2.2 Error Estimation

The accuracy of the Petrov–Galerkin solution $u^\delta \in \mathbb{U}^\delta$ is, in view of (2.2), estimated from below and above by the residual $f - \mathcal{B} u^\delta$ in $\mathbb{V}'$ whose evaluation would require computing the *supremizer*

$$\langle R(u^\delta; f), v \rangle_{\mathbb{V}} = b(u^\delta; v) - f(v), \quad v \in \mathbb{V},$$

since $\|R(u^\delta; f)\|_{\mathbb{V}} = \|f - \mathcal{B} u^\delta\|_{\mathbb{V}'}$. We refer to $R(u^\delta; f)$ as a *lifted residual*. The exact computation of $R(u^\delta; f)$ is, of course, not possible. However, to obtain a quantity that is at least uniformly proportional to $\|R(u^\delta; f)\|_{\mathbb{V}}$ one can proceed as in (2.5).

To that end, let us first suppose that f is contained in a finite-dimensional subspace $\mathbb{F}^\delta$ of $\mathbb{V}'$ with $\dim \mathbb{F}^\delta \approx \dim \mathbb{U}^\delta$. Now let $\bar{\bar{\mathbb{V}}}^\delta \subset \mathbb{V}$ be a closed subspace, that we call the *lifted residual search space*, such that

$$\bar{\bar{\gamma}}^\delta := \inf_{\{(u,f) \in \mathbb{U}^\delta \times \mathbb{F}^\delta \,:\, \mathcal{B} u \neq f\}} \sup_{0 \neq v \in \bar{\bar{\mathbb{V}}}^\delta} \frac{b(u; v) - f(v)}{\|u - \mathcal{B}^{-1} f\|_{\mathbb{U}} \|v\|_{\mathbb{V}}} > 0. \tag{2.7}$$

In analogy to (2.5) we then define $R^\delta = R^\delta(\bar{\bar{\mathbb{V}}}^\delta) : \mathbb{U} \times \mathbb{V}' \to \bar{\bar{\mathbb{V}}}^\delta$ by

$$\langle R^\delta(u; f), v \rangle_{\mathbb{V}} = b(u; v) - f(v) = b(u - \mathcal{B}^{-1} f; v) \quad (v \in \bar{\bar{\mathbb{V}}}^\delta). \tag{2.8}$$

We call $R^\delta(u; f)$ the *projected lifted residual* since it is the $\mathbb{V}$-orthogonal projection of the exact lifted residual (2.8) onto $\bar{\bar{\mathbb{V}}}^\delta$. For $(u, f) \in \mathbb{U}^\delta \times \mathbb{F}^\delta$, it holds that

$$\bar{\bar{\gamma}}^\delta \|u - \mathcal{B}^{-1} f\|_{\mathbb{U}} \leq \|R^\delta(u; f)\|_{\mathbb{V}} \leq \|\mathcal{B}\|_{\mathcal{L}(\mathbb{U}, \mathbb{V}')} \|u - \mathcal{B}^{-1} f\|_{\mathbb{U}}. \tag{2.9}$$

Thus the quantities $\|R^\delta(u; f)\|_{\mathbb{V}}$ provide computable upper and lower bounds for the error $\|u - \mathcal{B}^{-1} f\|_{\mathbb{U}}$ incurred by an approximation $u \in \mathbb{U}^\delta$ to the exact solution $u^{\text{ex}} = \mathcal{B}^{-1} f$.

Regarding stable DPG formulations of the transport problem, Section 4 is devoted to identifying suitable lifted residual search spaces $\bar{\bar{\mathbb{V}}}^\delta$ for which (2.7) will be shown to hold, *uniformly* in δ. In order to do so, just as for f we will need that the coefficients of transport problem belong to certain finite-dimensional spaces with dimensions proportional to $\dim \mathbb{U}^\delta$. Consequently, for general data, i.e. right-hand side f as well as convection and reaction coefficients, the lower bound in (2.9) will be valid modulo a data oscillation term that measures the distance between this data and their best approximations from the aforementioned finite-dimensional spaces.

2.3 Towards Error Reduction

By replacing both $\bar{\mathbb{V}}^\delta$ and $\bar{\bar{\mathbb{V}}}^\delta$ by their sum $\bar{\mathbb{V}}^\delta + \bar{\bar{\mathbb{V}}}^\delta$, from here on we will assume that $\bar{\bar{\mathbb{V}}}^\delta = \bar{\mathbb{V}}^\delta$. Then the relation

$$R^\delta(u_1; f) - R^\delta(u_2; f) = t^\delta(u_1 - u_2), \quad f \in \mathbb{V}', \ u_1, u_2 \in \mathbb{U},$$

follows directly from the definitions of R^δ and t^δ.

For the Petrov–Galerkin solution $u^\delta = u^\delta(f, \mathbb{U}^\delta, \mathbb{V}^\delta) \in \mathbb{U}^\delta$, *Petrov–Galerkin orthogonality*

$$\langle R^\delta(u^\delta; f), t^\delta(\mathbb{U}^\delta) \rangle_{\mathbb{V}} = 0$$

yields for *any* $u \in \mathbb{U}^\delta$,

$$\|R^\delta(u^\delta; f)\|_{\mathbb{V}}^2 = \|R^\delta(u; f)\|_{\mathbb{V}}^2 - \|t^\delta(u - u^\delta)\|_{\mathbb{V}}^2. \tag{2.10}$$

Remark 2.1. In particular, u^δ minimizes $\|R^\delta(\,\cdot\,; f)\|_{\mathbb{V}}$ over $\mathbb{U}^\delta$.

3 A Variational Formulation of the Transport Equation with Broken Test and Trial Spaces

For the convenience of the reader and to fix notation we briefly recall in this section the results from [2] to ensure the validity of the stability relations (2.1) and (2.4) which all subsequent developments will be based upon.

3.1 Transport Equation

We adhere to the setting considered in [2, Section 2] and let $\Omega \subset \mathbb{R}^n$ be a bounded polytopal domain, $\mathbf{b} \in L_\infty(\text{div}; \Omega)$, and $c \in L_\infty(\Omega)$. Here we set $L_\infty(\text{div}; \Omega) := W_\infty^0(\text{div}; \Omega)$, where $\mathbf{b} \in W_\infty^k(\text{div}; \Omega)$ means that both div $\mathbf{b}$ and each $\mathbf{b}_i$ belong to $W_\infty^k(\Omega)$. As usual the outflow/inflow boundary $\Gamma_\pm$ is the closure of all those points on $\partial\Omega$ for which the outward unit normal $\mathbf{n}$ is well defined and $\pm\mathbf{n} \cdot \mathbf{b} > 0$ while $\Gamma_0 = \partial\Omega \setminus (\Gamma_- \cup \Gamma_+)$ stands for the characteristic boundary. We consider the transport equation

$$\begin{cases} \mathbf{b} \cdot \nabla u + cu = f & \text{on } \Omega, \\ u = g & \text{on } \Gamma_-. \end{cases} \tag{3.1}$$

To explain in which sense u is to solve (3.1), the space $H(\mathbf{b}; \Omega) := \{u \in L_2(\Omega) : \mathbf{b} \cdot \nabla u \in L_2(\Omega)\}$, equipped with the norm $\|u\|_{H(\mathbf{b};\Omega)}^2 := \|u\|_{L_2(\Omega)}^2 + \|\mathbf{b} \cdot \nabla u\|_{L_2(\Omega)}^2$, plays a crucial role. More precisely, we need to work with the closed subspaces $H_{0,\Gamma_\pm}(\mathbf{b}; \Omega)$ obtained by taking the closure of smooth functions vanishing on $\Gamma_\pm$, respectively, under the norm $\|\cdot\|_{H(\mathbf{b};\Omega)}$. In fact, for $g = 0$ a first canonical variational formulation of (3.1) is to find $u \in H_{0,\Gamma_-}(\mathbf{b}; \Omega)$ such that

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

holds for all smooth test functions $v \in C^\infty(\bar{\Omega})$. Alternatively, after integration by parts one looks for $u \in L_2(\Omega)$ such that

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

holds for all $v \in H_{0,\Gamma_+}(\mathbf{b}; \Omega)$, where now the inflow boundary condition enters as a natural boundary condition. The second summand on the right-hand side vanishes of course for $g = 0$ which is the case we will focus on for convenience in what follows, see the discussion in [2].

Accordingly, these formulations induce bounded operators

$$\mathcal{B} : u \mapsto \mathbf{b} \cdot \nabla u + cu \in \mathcal{L}(H_{0,\Gamma_-}(\mathbf{b}; \Omega), L_2(\Omega)),$$

$$\mathcal{B}^* : v \mapsto cv - \text{div}\, v\mathbf{b} \in \mathcal{L}(H_{0,\Gamma_+}(\mathbf{b}; \Omega), L_2(\Omega)).$$

We stress that $\mathcal{B}^*$ is the *formal* adjoint of $\mathcal{B}$. In fact, the "true" adjoint $\mathcal{B}'$ would have to be considered as an element of $\mathcal{L}(L_2(\Omega), H_{0,\Gamma_-}(\mathbf{b};\Omega)')$. Moreover, $\mathcal{B}^*$ is the "true" adjoint of the transport operator considered as a mapping in $\mathcal{L}(L_2(\Omega), H_{0,\Gamma_+}(\mathbf{b};\Omega)')$. In view of these distinctions $\mathcal{B}$ and $\mathcal{B}^*$ may in general have different properties in terms of invertibility.

Since we do not strive for identifying the weakest possible assumptions on the problem parameters under which both mappings are invertible, we adopt this in what follows as an *assumption*

$$\mathcal{B} \in \mathcal{L}\mathrm{is}(H_{0,\Gamma_-}(\mathbf{b};\Omega), L_2(\Omega)), \tag{3.2}$$

$$\mathcal{B}^* \in \mathcal{L}\mathrm{is}(H_{0,\Gamma_+}(\mathbf{b};\Omega), L_2(\Omega)), \tag{3.3}$$

where $\mathcal{L}\mathrm{is}(\mathbb{U}, \mathbb{V})$ denotes the space of linear isomorphisms from $\mathbb{U}$ onto $\mathbb{V}$ and refer to, e.g., [2, 6] for concrete conditions on the problem parameters under which these assumptions are valid. Assumption (3.2) is essential for the stability of the subsequent DPG scheme. Finally, we note that the true adjoint of $\mathcal{B}^*$, in turn, belongs to $\mathcal{L}(L_2(\Omega), H_{0,\Gamma_+}(\mathbf{b};\Omega)')$ and can be viewed as an extension of $\mathcal{B}$ to $L_2(\Omega)$.

3.2 DPG Formulation of (3.1)

For a polyhedral Ω let $\mathbb{T}$ denote an (infinite) family of partitions $\mathcal{T}$ of $\bar{\Omega}$ into essentially disjoint closed n-simplices that can be created from an initial partition $\mathcal{T}_\perp$ by a repeated application of a refinement rule to individual n-simplices which splits them into two or more subsimplices. For $\mathcal{T}, \tilde{\mathcal{T}} \in \mathbb{T}$, we write $\mathcal{T} \preceq \tilde{\mathcal{T}}$ when $\tilde{\mathcal{T}}$ is a refinement of $\mathcal{T}$. We write $\mathcal{T} \prec \tilde{\mathcal{T}}$ when $\mathcal{T} \preceq \tilde{\mathcal{T}}$ and $\mathcal{T} \neq \tilde{\mathcal{T}}$. For a n-simplex K, let

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

denote its shape-parameter. With $\mathfrak{T}$ denoting the set of all n-simplices in any partition $\mathcal{T} \in \mathbb{T}$, we assume that these simplices (or briefly $\mathfrak{T}$) are (is) uniformly *shape regular* in the sense that

$$\varrho := \sup_{K \in \mathfrak{T}} \varrho_K < \infty.$$

For each $K \in \mathfrak{T}$, we split its boundary into characteristic and in- and outflow boundaries, i.e.,

$$\partial K = \partial K_0 \cup \partial K_+ \cup \partial K_-,$$

and, for $\mathcal{T} \in \mathbb{T}$, denote by $\partial \mathcal{T} := \bigcup_{K \in \mathcal{T}} \partial K \setminus \partial K_0$ the *mesh skeleton*, i.e., the union of the non-characteristic boundary portions of the elements.

Denoting by $\nabla_{\mathcal{T}}$ the piecewise gradient operator, we consider the "broken" counterpart to $H(\mathbf{b};\Omega)$

$$H(\mathbf{b};\mathcal{T}) = \{v \in L_2(\Omega) : \mathbf{b} \cdot \nabla_{\mathcal{T}} v \in L_2(\Omega)\},$$

equipped with squared "broken" norm $\|v\|^2_{H(\mathbf{b};\mathcal{T})} := \|v\|^2_{L_2(\Omega)} + \|\mathbf{b} \cdot \nabla_{\mathcal{T}} v\|^2_{L_2(\Omega)}$, and view the quantities living on the skeleton as elements of the space

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

equipped with quotient norm

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

For $\mathcal{T} \in \mathbb{T}$, a *piecewise* integration-by-parts of the transport equation (3.1) leads to the following "mesh-dependent" (but otherwise *"continuous" infinite-dimensional*) variational formulation:

Problem 3.1. For $\mathbb{U}_{\mathcal{T}} := L_2(\Omega) \times H_{0,\Gamma_-}(\mathbf{b};\partial\mathcal{T})$, $\mathbb{V}_{\mathcal{T}} := H(\mathbf{b};\mathcal{T})$, given $f \in \mathbb{V}'_{\mathcal{T}}$, find the solution

$$(u_{\mathcal{T}}, \theta_{\mathcal{T}}) = (u_{\mathcal{T}}(f), \theta_{\mathcal{T}}(f)) \in \mathbb{U}_{\mathcal{T}}$$

that, for all $v \in \mathbb{V}_{\mathcal{T}}$, satisfies

$$b_{\mathcal{T}}(u_{\mathcal{T}}, \theta_{\mathcal{T}}; v) := \int_{\Omega} (cv - \mathbf{b} \cdot \nabla_{\mathcal{T}} v - v\,\mathrm{div}\,\mathbf{b})u_{\mathcal{T}}\,d\mathbf{x} + \int_{\partial\mathcal{T}} [\![v\mathbf{b}]\!]\theta_{\mathcal{T}}\,d\mathbf{s} = f(v).$$

Here $\int_{\partial\mathcal{T}} [\![v\mathbf{b}]\!]\theta_{\mathcal{T}}\,d\mathbf{s}$ should read as the unique extension to a bounded bilinear form on $H_{0,\Gamma_-}(\mathbf{b};\partial\mathcal{T}) \times \mathbb{V}_{\mathcal{T}}$ (cf. [2, Lemma. 3.4]) of the integral over $\partial\mathcal{T}$ of the product of $[\![v\mathbf{b}]\!]$ and $\theta_{\mathcal{T}}$, where for smooth v and $x \in \partial K \cap \partial K'$, $[\![v\mathbf{b}]\!](x) := (v\mathbf{b}|_K \cdot \mathbf{n}_K)(x) + (v\mathbf{b}|_{K'} \cdot \mathbf{n}_{K'})(x)$, and $[\![v\mathbf{b}]\!](x) := (v\mathbf{b}|_K \cdot \mathbf{n}_K)(x)$ for $x \in \partial\Omega \cap \partial K$. Note the introduction of the notation $(u_{\mathcal{T}}, \theta_{\mathcal{T}})$ for the exact solution of this variational problem.

In the following, 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 following result roughly says that Problem 3.1 is uniformly inf-sup stable whenever the operators $\mathcal{B}$, $\mathcal{B}^*$ are isomorphisms on the respective function space pairs.

Theorem 3.2 ([2, Theorem 3.1]). *Assume that $\mathbf{b} \in L_\infty(\operatorname{div};\Omega)$, $c \in L_\infty(\Omega)$ and that conditions* (3.2), (3.3) *hold. Then, defining $\mathcal{B}_{\mathcal{T}} : \mathbb{U}_{\mathcal{T}} \to \mathbb{V}'_{\mathcal{T}}$ by $(\mathcal{B}_{\mathcal{T}}(u, \theta))(v) := b_{\mathcal{T}}(u, \theta; v)$, one has $\mathcal{B}_{\mathcal{T}} \in \mathcal{L}\mathrm{is}(\mathbb{U}_{\mathcal{T}}, \mathbb{V}'_{\mathcal{T}})$ with*

$$\|\mathcal{B}_{\mathcal{T}}\|_{\mathcal{L}(\mathbb{U}_{\mathcal{T}}, \mathbb{V}'_{\mathcal{T}})} \le 2 + \|\operatorname{div}\mathbf{b}\| + \|c - \operatorname{div}\mathbf{b}\|,$$

$$\|\mathcal{B}_{\mathcal{T}}^{-1}\|_{\mathcal{L}(\mathbb{V}'_{\mathcal{T}}, \mathbb{U}_{\mathcal{T}})} \le \sqrt{\|\mathcal{B}^{-*}\|^2 + \tilde{C}_{\mathcal{B}}^2},$$

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

The additional independent variable $\theta_{\mathcal{T}}$ introduced in the mesh-dependent variational formulation replaces the trace $u_{\mathcal{T}}|_{\partial\mathcal{T}}$ which generally is not defined for $u_{\mathcal{T}} \in L_2(\Omega)$. If $f \in L_2(\Omega)$, however, or, equivalently, $u_{\mathcal{T}} \in H_{0,\Gamma_-}(\mathbf{b};\Omega)$, *then* a reversed integration by parts shows that

$$u_{\mathcal{T}} = u^{\mathrm{ex}} = u^{\mathrm{ex}}(f) := \mathcal{B}^{-1}f, \qquad \theta_{\mathcal{T}} = u^{\mathrm{ex}}|_{\partial\mathcal{T}}.$$

3.3 Petrov–Galerkin

For any $\mathcal{T} \in \mathbb{T}$, let $\mathcal{T}_s \in \mathbb{T}$ be a *refinement* of $\mathcal{T}$. We set

$$\sigma := \sup_{\mathcal{T}\in\mathbb{T}} \max_{K'\in\mathcal{T}} \left(\max_{\{K\in\mathcal{T}_s : K \subset K'\}} \frac{\operatorname{diam}(K)}{\operatorname{diam}(K')}, \operatorname{diam}(K') \right), \tag{3.4}$$

which later will be assumed to be sufficiently small. We also require that

$$\inf_{\mathcal{T}\in\mathbb{T}} \min_{K'\in\mathcal{T}} \min_{\{K\in\mathcal{T}_s : K \subset K'\}} \frac{\operatorname{diam}(K)}{\operatorname{diam}(K')} \gtrsim \sigma. \tag{3.5}$$

This means that we will assume that any partition $\mathcal{T} \in \mathbb{T}$ is sufficiently fine, and, what is more important, that $\mathcal{T}_s \in \mathbb{T}$ is a refinement of $\mathcal{T}$ such that the *subgrid refinement factor* (or sometimes called subgrid refinement depth) $\frac{1}{\sigma}$ when going from any $\mathcal{T}$ to $\mathcal{T}_s$ is sufficiently large. In addition to the conditions from Theorem 3.2, we assume henceforth

$$\mathbf{b}|_K \in W_\infty^1(\operatorname{div}; K), \quad c|_K \in W_\infty^1(K) \ (K \in \mathcal{T}_s), \quad |\mathbf{b}|^{-1} \in L_\infty(\Omega).$$

Under these assumptions, we have the following result:

Theorem 3.3 ([2, Theorem 4.8]). *Selecting, for some fixed degrees $m_w \ge 1$, and m_u,*

$$\mathbb{U}_{\mathcal{T}}^{\delta} := \prod_{K'\in\mathcal{T}} \mathcal{P}_{m_u}(K') \times \left(H_{0,\Gamma_-}(\mathbf{b};\Omega) \cap \prod_{K'\in\mathcal{T}} \mathcal{P}_{m_w}(K') \right)\Big|_{\partial\mathcal{T}_s} \subset \mathbb{U}_{\mathcal{T}_s},$$

$$\bar{\mathbb{V}}_{\mathcal{T}_s}^{\delta} := \prod_{K\in\mathcal{T}_s} \mathcal{P}_{m_v}(K) \subset \mathbb{V}_{\mathcal{T}_s},$$

where $m_v \ge \max(m_u, m_w) + 1$, for $\sigma > 0$ small enough it holds that

$$\inf_{\mathcal{T}\in\mathbb{T}} \inf_{0\neq(u,\theta)\in\mathbb{U}_{\mathcal{T}}^{\delta}} \sup_{0\neq v\in\bar{\mathbb{V}}_{\mathcal{T}_s}^{\delta}} \frac{b_{\mathcal{T}_s}(u, \theta; v)}{\|(u, \theta)\|_{\mathbb{U}_{\mathcal{T}_s}} \|v\|_{\mathbb{V}_{\mathcal{T}_s}}} > 0,$$

only dependent on (upper bounds for) m_u, m_w, ϱ, $\||\mathbf{b}|^{-1}\|_{L_\infty(\Omega)}$, $\|\mathcal{B}^{-1}\|_{\mathcal{L}(L_2(\Omega),H_{0,\Gamma_-}(\mathbf{b};\Omega))}$, $\sup_{K\in\mathcal{T}_s} \|\mathbf{b}|_K\|_{W_\infty^1(\operatorname{div};K)}$, and $\sup_{K\in\mathcal{T}_s} \|c|_K\|_{W_\infty^1(K)}$.[1]

---

[1] In the theorem in [2] the last two expressions read as $\|\mathbf{b}\|_{W_\infty^1(\operatorname{div};\Omega)}$ and $\|c\|_{W_\infty^1(\Omega)}$, but an inspection of the proof shows that they can be replaced by the current ones.

Consequently, as we have seen in Section 2.1, the *Petrov–Galerkin solution* $(u_{\mathcal{T}}^{\delta}, \theta_{\mathcal{T}}^{\delta}) \in \mathbb{U}_{\mathcal{T}}^{\delta} \subset \mathbb{U}_{\mathcal{T}_s}$ of

$$b_{\mathcal{T}_s}(u_{\mathcal{T}}^{\delta}, \theta_{\mathcal{T}}^{\delta}; v) = f(v) \quad (v \in \operatorname{ran} t^{\delta}|_{\mathbb{U}_{\mathcal{T}}^{\delta}}),$$

where

$$\langle t^{\delta}(u, \theta), v \rangle_{\mathbb{V}_{\mathcal{T}_s}} = b_{\mathcal{T}_s}(u, \theta; v) \quad (v \in \bar{\mathbb{V}}_{\mathcal{T}_s}^{\delta}),$$

is a near-best approximation to $(u_{\mathcal{T}_s}, \theta_{\mathcal{T}_s}) = \mathcal{B}_{\mathcal{T}_s}^{-1} f \in \mathbb{U}_{\mathcal{T}_s}$ from $\mathbb{U}_{\mathcal{T}}^{\delta}$.

Since the above stability is ensured by "some" fixed subgrid-refinement depth, the computational work for computing the test-basis functions remains uniformly proportional to the dimension of the trial space and in this sense scales optimally. While the actual depth is hard to quantify precisely the experiments considered in [2] actually suggest that one or even no additional refinement suffice in these examples.

We emphasize that although the bilinear form $b_{\mathcal{T}_s}$ corresponds to the variational formulation of the transport problem obtained by applying a piecewise integration by parts with respect to the "*fine*" partition $\mathcal{T}_s$, and the test search space $\bar{\mathbb{V}}_{\mathcal{T}_s}^{\delta}$ consists of piecewise polynomials with respect to $\mathcal{T}_s$ too, the applied *trial* space consists of pairs of functions that are piecewise polynomial with respect to the "*coarse*" partition $\mathcal{T}$, or that are restrictions of such functions to $\partial \mathcal{T}_s$, respectively.

Remark 3.4. Actually, in [2] we established a slightly stronger inf-sup condition. Defining

$$\check{\mathbb{U}}_{\mathcal{T}}^{\delta} := \prod_{K' \in \mathcal{T}} \mathcal{P}_{m_u}(K') \times \left(H_{0,\Gamma_-}(\mathbf{b}; \Omega) \cap \prod_{K' \in \mathcal{T}} \mathcal{P}_{m_w}(K') \right) \subset L_2(\Omega) \times H_{0,\Gamma_-}(\mathbf{b}; \Omega) =: \check{\mathbb{U}},$$

any $(u, \theta) \in \mathbb{U}_{\mathcal{T}}^{\delta} \subset \mathbb{U}_{\mathcal{T}_s}$ is of the form $(u, w|_{\partial \mathcal{T}_s})$ for some $(u, w) \in \check{\mathbb{U}}_{\mathcal{T}}^{\delta}$. In [2] it was shown that

$$\inf_{\mathcal{T} \in \mathbb{T}} \inf_{0 \neq (u,w) \in \check{\mathbb{U}}_{\mathcal{T}}^{\delta}} \sup_{0 \neq v \in \bar{\mathbb{V}}_{\mathcal{T}_s}^{\delta}} \frac{b_{\mathcal{T}_s}(u, w|_{\partial \mathcal{T}_s}; v)}{\|(u, w)\|_{\check{\mathbb{U}}} \|v\|_{\mathbb{V}_{\mathcal{T}_s}}} > 0, \tag{3.6}$$

which implies Theorem 3.3 because of $\|(u, w)\|_{\check{\mathbb{U}}} \geq \|(u, w|_{\partial \mathcal{T}_s})\|_{\mathbb{U}_{\mathcal{T}_s}}$.

Knowing (3.6), the uniform boundedness of $\|\mathcal{B}_{\mathcal{T}_s}\|_{\mathcal{L}(\mathbb{U}_{\mathcal{T}_s}, \mathbb{V}_{\mathcal{T}_s}')}$ shows that $\|(u, w)\|_{\check{\mathbb{U}}} \approx \|(u, w|_{\partial \mathcal{T}_s})\|_{\mathbb{U}_{\mathcal{T}_s}}$ on $\check{\mathbb{U}}_{\mathcal{T}}^{\delta}$. In particular, this means that $(u, w|_{\partial \mathcal{T}_s})$ determines $(u, w) \in \check{\mathbb{U}}_{\mathcal{T}}^{\delta}$ uniquely, so that equally well we can speak of the *Petrov–Galerkin solution* $(u_{\mathcal{T}}^{\delta}, w_{\mathcal{T}}^{\delta}) \in \check{\mathbb{U}}_{\mathcal{T}}^{\delta}$ of

$$b_{\mathcal{T}_s}(u_{\mathcal{T}}^{\delta}, w_{\mathcal{T}}^{\delta}|_{\mathcal{T}_s}; v) = f(v) \quad (v \in \operatorname{ran} t^{\delta}|_{\check{\mathbb{U}}_{\mathcal{T}}^{\delta}}), \tag{3.7}$$

where, of course, $t^{\delta}(u, w) := t^{\delta}(u, w|_{\partial \mathcal{T}_s})$.

Remark 3.5. The trial spaces $\check{\mathbb{U}}_{\mathcal{T}}^{\delta}$ are *nested* whenever the underlying partitions are nested. This plays an important role for conceiving adaptive strategies.

Remark 3.6. Since a polynomial of degree ≥ 3 is not uniquely determined by its values on the boundary of a triangle, the inf-sup stability (3.6) can apparently only hold for $m_w \geq 3$ when $\mathcal{T}_s$ is a true refinement of $\mathcal{T}$.

In the latter formulation involving the lifted version w of the skeleton quantity θ, the scheme provides two approximations for the solution of the transport problem, namely $u_{\mathcal{T}}^{\delta} \in L_2(\Omega)$ and a second one $w_{\mathcal{T}}^{\delta} \in H(\mathbf{b}; \Omega)$.

Remark 3.7. For a function in $\prod_{K' \in \mathcal{T}} \mathcal{P}_{m_w}(K')$ to be in $H(\mathbf{b}; \Omega)$, it has to be continuous at any intersection of an in- and outflow face of any $K' \in \mathcal{T}$. To realize this condition, an obvious approach is to consider in the definition of $\check{\mathbb{U}}_{\mathcal{T}}^{\delta}$ or $\mathbb{U}_{\mathcal{T}}^{\delta}$ the space $H_{0,\Gamma_-}(\mathbf{b}; \Omega) \cap C(\Omega) \cap \prod_{K' \in \mathcal{T}} \mathcal{P}_{m_w}(K')$ instead of $H_{0,\Gamma_-}(\mathbf{b}; \Omega) \cap \prod_{K' \in \mathcal{T}} \mathcal{P}_{m_w}(K')$. Obviously with this modification, Theorem 3.3 and Remark 3.4 remain valid, and so does the whole further exposition.

4 A Posteriori Error Estimation

The central goal in this section is to establish the validity of (2.7) for locally uniformly finite-dimensional test search spaces of the same form as used in Theorem 3.3. We will be able to do so modulo a data oscillation

term. The principal difficulty lies in an intrinsic sensitivity of essential problem metrics with respect to perturbations in the *convection field*. To exploit the fact that we can identify optimal test spaces for locally constant convection comes at the price of an elaborate perturbation analysis to be carried out in this section. In fact, it involves two levels of perturbation, namely passing from general data $\mathbf{b}$, c, f to piecewise polynomial ones, and then to piecewise constant $\mathbf{b}$ on a subgrid. The passage to piecewise polynomial data is accounted for by *data oscillation* terms. The piecewise polynomial structure of the data with respect to the (coarser) trial grid $\mathfrak{T}$, in turn, is needed to control the effect of the reduction to piecewise constant convection on the discrete level.

4.1 Main Results

Theorem 4.1. *Assume* (3.2), *and let* $f \in L_2(\Omega)$. *For* $\mathfrak{T} \in \mathbb{T}$, *assume that for* $K' \in \mathfrak{T}$, *one has* $\mathbf{b}|_{K'} \in W^1_\infty(K')^n$, $c|_{K'} \in W^1_\infty(K')$, *and let* $\tilde{\mathbf{b}}$, $\tilde{c}$, $\tilde{f}$ *denote the best piecewise polynomial approximations to* $\mathbf{b}$, c, f *of degrees* $m_\mathbf{b}$, m_c, *and* m_f *with respect to* $\mathfrak{T}$ *in* $L_\infty(\Omega)^n$-, $L_\infty(\Omega)$-, *or* $L_2(\Omega)$-norm, respectively. Let

$$\mathrm{osc}_\mathfrak{T}(\mathbf{b}, c, f) := \max\Big(\|f - \tilde{f}\|_{L_2(\Omega)}, \big(\|c - \tilde{c}\|_{L_\infty(\Omega)}, \max_{K' \in \mathfrak{T}} \mathrm{diam}(K')^{-1} \|\|\mathbf{b} - \tilde{\mathbf{b}}\|_{L_\infty(K')} \big) \|f\|_{L_2(\Omega)} \Big)$$

and

$$m_v \geq \max(m_w + \max(m_c, 1, m_\mathbf{b} - 1), m_u + \max(m_c, 1), m_f). \tag{4.1}$$

Then, with $\breve{\mathbb{U}}^\delta_\mathfrak{T}$ *and* $\bar{\mathbb{V}}^\delta_{\mathfrak{T}_s}$ *as defined before, for fixed sufficiently small* $\sigma > 0$ *in* (3.4), *and for any* $(u, w) \in \breve{\mathbb{U}}^\delta_\mathfrak{T}$ *for which* $\max(\|u\|_{L_2(\Omega)}, \|w\|_{L_2(\Omega)}) \lesssim \|f\|_{L_2(\Omega)}$ *(which, on account of* (3.6), *is valid for the Petrov–Galerkin solution), it holds that*

$$\|R^\delta_{\mathfrak{T}_s}\|_{\mathbb{V}_{\mathfrak{T}_s}} \lesssim \|(u^{\mathrm{ex}}, u^{\mathrm{ex}}) - (u, w)\|_{\breve{\mathbb{U}}} \lesssim \|R^\delta_{\mathfrak{T}_s}\|_{\mathbb{V}_{\mathfrak{T}_s}} + \mathrm{osc}_\mathfrak{T}(\mathbf{b}, c, f), \tag{4.2}$$

where $R^\delta_{\mathfrak{T}_s} \in \bar{\mathbb{V}}^\delta_{\mathfrak{T}_s}$ *is defined by*

$$\langle R^\delta_{\mathfrak{T}_s}, v \rangle_{\mathbb{V}_{\mathfrak{T}_s}} = b_{\mathfrak{T}_s}(u, w|_{\partial\mathfrak{T}_s}; v) - \int_\Omega fv \, d\mathbf{x} \quad (v \in \bar{\mathbb{V}}^\delta_{\mathfrak{T}_s}) \tag{4.3}$$

(cf. (2.8)). *The constants absorbed by the* $\lesssim$-*symbols in* (4.2) *depend only on the polynomial degrees and on (upper bounds for)* ϱ, $\|\mathbf{b}|^{-1}\|_{L_\infty(\Omega)}$, $\sup_{K' \in \mathfrak{T}} \|\mathbf{b}\|_{W^1_\infty(K')^n}$, $\sup_{K' \in \mathfrak{T}} \|c\|_{W^1_\infty(K')}$, *and* $\|\mathcal{B}^{-1}\|_{\mathcal{L}(L_2(\Omega), H_{0,\Gamma_-}(\mathbf{b};\Omega))}$.

Note that for given degrees m_u and m_w, then, for sufficiently large $m_\mathbf{b}$, m_c, and m_f (and thus m_v) and piecewise smooth $\mathbf{b}$, c and f, $\mathrm{osc}_\mathfrak{T}(\mathbf{b}, c, f)$ can be reduced at a better rate in terms of $\#\mathfrak{T}$ than generally can be expected for $\|(u^{\mathrm{ex}}, u^{\mathrm{ex}}) - (u^\delta_\mathfrak{T}, w^\delta_\mathfrak{T})\|_{\breve{\mathbb{U}}}$.

The proof of Theorem 4.1 will be based on the following proposition.

Proposition 4.2. *In the situation of Theorem 4.1, let*

$$\tilde{b}_{\mathfrak{T}_s}(u, w; v) := \sum_{K \in \mathfrak{T}_s} \tilde{b}_K(u, w; v),$$

where

$$\tilde{b}_K(u, w; v) := \int_K (\tilde{c}u + \tilde{\mathbf{b}} \cdot \nabla w)v + (w - u)(v \, \mathrm{div}\, \tilde{\mathbf{b}} + \tilde{\mathbf{b}} \cdot \nabla v) \, d\mathbf{x}.$$

Then, for any $(u, w, \tilde{f}) \in \mathbb{D}_\mathfrak{T} := \prod_{K' \in \mathfrak{T}} \mathcal{P}_{m_u}(K') \times \mathcal{P}_{m_w}(K') \times \mathcal{P}_{m_f}(K')$, *it holds that*

$$\underbrace{\|w - u\|_{L_2(\Omega)} + \|\tilde{\mathbf{b}} \cdot \nabla_\mathfrak{T} w + \tilde{c}w - \tilde{f}\|_{L_2(\Omega)}}_{=: \mathcal{E}_\mathfrak{T}(u, w, \tilde{f})} \lesssim \sup_{0 \neq v \in \bar{\mathbb{V}}^\delta_{\mathfrak{T}_s}} \frac{\tilde{b}_{\mathfrak{T}_s}(u, w; v) - \int_\Omega \tilde{f}v \, d\mathbf{x}}{\|v\|_{\mathbb{V}_{\mathfrak{T}_s}}}, \tag{4.4}$$

only dependent on the polynomial degrees and on (upper bounds for) ϱ, $\|\mathbf{b}|^{-1}\|_{L_\infty(\Omega)}$, $\sup_{K' \in \mathfrak{T}} \|\mathbf{b}\|_{W^1_\infty(K')^n}$, *and* $\sup_{K' \in \mathfrak{T}} \|c\|_{W^1_\infty(K')}$.

Remark 4.3. In a strict sense the quantities $R^\delta_{\mathfrak{T}_s}$, defined in (4.3) can, for general coefficients $\mathbf{b}$, c, not be computed *exactly*. Under the presumption that the accuracy of quadrature can be adjusted, this issue is usu-

ally neglected, as we did in (4.2) above. Since quadrature is in essence based on replacing the integrand by a local polynomial approximation, a natural way of incorporating this issue here is to work with the analogous projected lifted residuals with respect to the perturbed data

$$\langle \tilde{R}^\delta_{\mathcal{T}_s}, v \rangle_{\mathbb{V}_{\mathcal{T}_s}} = \tilde{b}_{\mathcal{T}_s}(u, w|_{\partial \mathcal{T}_s}; v) - \int_\Omega \tilde{f} v \, d\mathbf{x} \quad (v \in \bar{\mathbb{V}}^\delta_{\mathcal{T}_s}),$$

which can be computed exactly. Under the assumptions of Theorem 4.1 one then obtains the following estimates:

$$\|\tilde{R}^\delta_{\mathcal{T}_s}\|_{\mathbb{V}_{\mathcal{T}_s}} \lesssim \|(u^{\mathrm{ex}}, u^{\mathrm{ex}}) - (u, w)\|_{\check{\mathbb{U}}} + \mathrm{osc}_{\mathcal{T}}(\mathbf{b}, c, f) \lesssim \|\tilde{R}^\delta_{\mathcal{T}_s}\|_{\mathbb{V}_{\mathcal{T}_s}} + \mathrm{osc}_{\mathcal{T}}(\mathbf{b}, c, f). \tag{4.5}$$

Sections 4.2–4.4 will be devoted to the proof of Proposition 4.2. In the course of these developments it will be seen that the residual $\mathcal{E}_{\mathcal{T}}(u, w, \tilde{f})$ is actually equivalent to $\|\tilde{R}^\delta_{\mathcal{T}_s}\|_{\mathbb{V}_{\mathcal{T}_s}}$ and may therefore also be used as error indicator.

Assuming for the moment the validity of Proposition 4.2, we can give the proof of Theorem 4.1 and Remark 4.3.

Proof of Theorem 4.1. Applications of the triangle inequality show that

$$\|(u^{\mathrm{ex}}, u^{\mathrm{ex}}) - (u, w)\|_{\check{\mathbb{U}}} \approx \|u - w\|_{L_2(\Omega)} + \|u^{\mathrm{ex}} - w\|_{H(\mathbf{b};\Omega)},$$

and it holds that

$$\|u^{\mathrm{ex}} - w\|_{H(\mathbf{b};\Omega)} \le \|\mathcal{B}^{-1}\|_{\mathcal{L}(L_2(\Omega), H_{0,\Gamma_-}(\mathbf{b};\Omega))} \|\mathcal{B} w - f\|_{L_2(\Omega)}.$$

By using the inverse inequality on piecewise polynomials of degree m_w, and $\|w\|_{L_2(\Omega)} \lesssim \|f\|_{L_2(\Omega)}$, we infer that

$$\left| \|\mathcal{B} w - f\|_{L_2(\Omega)} - \|\tilde{\mathbf{b}} \cdot \nabla_{\mathcal{T}} w + \tilde{c} w - \tilde{f}\|_{L_2(\Omega)} \right| \lesssim \mathrm{osc}_{\mathcal{T}}(\mathbf{b}, c, f).$$

An application of Proposition 4.2 gives

$$\|w - u\|_{L_2(\Omega)} + \|\tilde{\mathbf{b}} \cdot \nabla_{\mathcal{T}} w + \tilde{c} w - \tilde{f}\|_{L_2(\Omega)} \lesssim \sup_{0 \ne v \in \bar{\mathbb{V}}^\delta_{\mathcal{T}_s}} \frac{\tilde{b}_{\mathcal{T}_s}(u, w; v) - \int_\Omega \tilde{f} v \, d\mathbf{x}}{\|v\|_{\mathbb{V}_{\mathcal{T}_s}}}.$$

We show next that the right-hand side deviates from the analogous unperturbed quantity only by $\mathrm{osc}_{\mathcal{T}}(\mathbf{b}, c, f)$. To that end, it holds that

$$\left| \int_\Omega f v \, d\mathbf{x} - \int_\Omega \tilde{f} v \, d\mathbf{x} \right| \le \|f - \tilde{f}\|_{L_2(\Omega)} \|v\|_{L_2(\Omega)},$$

and

$$\left| b_{\mathcal{T}_s}(u, w|_{\partial \mathcal{T}_s}; v) - \tilde{b}_{\mathcal{T}_s}(u, w; v) \right| = \left| \sum_{K \in \mathcal{T}_s} \int_K ((c - \tilde{c}) u + (\mathbf{b} - \tilde{\mathbf{b}}) \cdot \nabla w) v + (w - u)(v \operatorname{div}(\mathbf{b} - \tilde{\mathbf{b}}) + (\mathbf{b} - \tilde{\mathbf{b}}) \cdot \nabla v) \, dx \right|$$

$$\lesssim \mathrm{osc}_{\mathcal{T}}(u, w, f) \|v\|_{L_2(\Omega)},$$

where we used that for $K' \in \mathcal{T}$,

$$\|\mathbf{b} - \tilde{\mathbf{b}}\|_{W^1_\infty(K')^n} \lesssim \operatorname{diam}(K')^{-1} \|\mathbf{b} - \tilde{\mathbf{b}}\|_{L_\infty(K')^n} \tag{4.6}$$

(cf., e.g., [3] for the argument); $\|\nabla v\|_{L_2(K)^n} \lesssim \operatorname{diam}(K)^{-1} \|v\|_{L_2(K)}$ for $K \in \mathcal{T}_s$; and $\operatorname{diam}(K)^{-1} \lesssim \operatorname{diam}(K')^{-1}$ for $K \subset K'$ by (3.5). We conclude that

$$\left| \sup_{0 \ne v \in \bar{\mathbb{V}}^\delta_{\mathcal{T}_s}} \frac{\tilde{b}_{\mathcal{T}_s}(u, w; v) - \int_\Omega \tilde{f} v \, d\mathbf{x}}{\|v\|_{\mathbb{V}_{\mathcal{T}_s}}} - \sup_{0 \ne v \in \bar{\mathbb{V}}^\delta_{\mathcal{T}_s}} \frac{b_{\mathcal{T}_s}(u, w; v) - \int_\Omega f v \, d\mathbf{x}}{\|v\|_{\mathbb{V}_{\mathcal{T}_s}}} \right| \lesssim \mathrm{osc}_{\mathcal{T}}(\mathbf{b}, c, f).$$

From

$$\|R^\delta_{\mathcal{T}_s}\|_{\mathbb{V}_{\mathcal{T}_s}} = \sup_{0 \ne v \in \bar{\mathbb{V}}^\delta_{\mathcal{T}_s}} \frac{b_{\mathcal{T}_s}(u, w; v) - \int_\Omega f v \, d\mathbf{x}}{\|v\|_{\mathbb{V}_{\mathcal{T}_s}}} \lesssim \|(u^{\mathrm{ex}}, u^{\mathrm{ex}}) - (u, w)\|_{\check{\mathbb{U}}},$$

the assertion of Theorem 4.1 follows. The above argument also shows that $\|\tilde{R}^\delta_{\mathcal{T}_s} - R^\delta_{\mathcal{T}_s}\|_{\mathbb{V}_{\mathcal{T}_s}} \lesssim \mathrm{osc}_{\mathcal{T}}(\mathbf{b}, c, f)$ which confirms (4.5) as well. □

Remark 4.4. Estimate (4.6) together with $\|\mathbf{b} - \tilde{\mathbf{b}}\|_{L_\infty(K')^n} \lesssim \mathrm{diam}(K')|\mathbf{b}|_{W^1_\infty(K')^n}$ implies that

$$|\tilde{\mathbf{b}}|_{W^1_\infty(K')^n} \lesssim |\mathbf{b}|_{W^1_\infty(K')^n}$$

which, as the analogous $|\tilde{c}|_{W^1_\infty(K')} \lesssim |c|_{W^1_\infty(K')}$, will be often used in the following.

4.2 Lifting Modified Residuals

As in [2] the verification of uniform inf-sup stability (4.4) relies on judiciously perturbing exact Riesz lifts corresponding to certain perturbed bilinear forms. To describe this, given $\mathcal{T} \in \mathbb{T}$, let we define for $K \in \mathcal{T}_s \cup \mathcal{T}$, the *constants*

$$\check{\mathbf{b}}_K := |K|^{-1} \int_K \tilde{\mathbf{b}}\, d\mathbf{x}, \quad \check{d}_K := |K|^{-1} \int_K \mathrm{div}\, \tilde{\mathbf{b}}\, d\mathbf{x}$$

and for $\mathcal{T} \in \mathbb{T}$, let $\check{\mathbf{b}} \in L_\infty(\Omega)^n$ be given by

$$\check{\mathbf{b}}|_K := \check{\mathbf{b}}_K \quad (K \in \mathcal{T}_s).$$

On $\check{\mathbb{U}}^\delta_{\mathcal{T}} \times \bar{\mathbb{V}}^\delta_{\mathcal{T}_s}$ we introduce yet another *modified bilinear form*

$$\check{b}_{\mathcal{T}_s}(u, w; v) := \sum_{K \in \mathcal{T}_s} \check{b}_K(u, w; v), \tag{4.7}$$

where the summands $\check{b}_K(u, w; v)$ are defined by

$$\check{b}_K(u, w; v) := \int_K (\check{\mathbf{b}} \cdot \nabla u + \tilde{c}u + \check{d}_K(w - u))v\, d\mathbf{x} + \int_{\partial K} \check{\mathbf{b}}_K \cdot \mathbf{n}_K (w - u)v\, d\mathbf{s}. \tag{4.8}$$

Note that $\check{\mathbf{b}}$ and $\check{d}$ are piecewise constant with respect to $\mathcal{T}_s$, whereas $\tilde{\mathbf{b}}$ and $\tilde{c}$ are piecewise polynomial with respect to $\mathcal{T}$. This form is only introduced for analysis purposes since, as it turns out, it allows us to determine local lifted residuals exactly. Their use requires then yet another layer of perturbation arguments.

Remark 4.5. The particular form of the modified bilinear form (4.7)–(4.8), in particular the integrand in the boundary integral over ∂K, is to ensure that $v \mapsto \check{b}_K(u, w; v)$ is in $H(\check{\mathbf{b}}_K; K)'$.

The proof of Proposition 4.2 is based on the following steps:

(I) We will construct an $\check{R} = \check{R}_{\mathcal{T}_s}(u, w; \tilde{f}) \in \bar{\mathbb{V}}^\delta_{\mathcal{T}_s}$ such that

$$\check{b}_{\mathcal{T}_s}(u, w; \check{R}) - \int_\Omega \tilde{f}\check{R}\, d\mathbf{x} \gtrsim \mathcal{E}_{\mathcal{T}}(u, w, \tilde{f})\|\check{R}\|_{\mathbb{V}_{\mathcal{T}_s}},$$

of course, uniformly in $\mathcal{T} \in \mathbb{T}$ and $(u, w, \tilde{f}) \in \mathbb{D}_{\mathcal{T}}$.

(II) Starting from the simple decomposition

$$\tilde{b}_{\mathcal{T}_s}(u, w; \check{R}) - \int_\Omega \tilde{f}\check{R}\, d\mathbf{x} = \check{b}_{\mathcal{T}_s}(u, w; \check{R}) - \int_\Omega \tilde{f}\check{R}\, d\mathbf{x} + \tilde{b}_{\mathcal{T}_s}(u, w; \check{R}) - \check{b}_{\mathcal{T}_s}(u, w; \check{R}),$$

we will show for the second summand that

$$|\tilde{b}_{\mathcal{T}_s}(u, w; \check{R}) - \check{b}_{\mathcal{T}_s}(u, w; \check{R})| \le \delta \mathcal{E}_{\mathcal{T}}(u, w, \tilde{f})\|\check{R}\|_{\mathbb{V}_{\mathcal{T}_s}}$$

holds for a sufficiently small $\delta > 0$, depending on the inf-sup constant for the first summand.

As the construction of the modified bilinear form $\check{b}_{\mathcal{T}_s}$ from $\tilde{b}_{\mathcal{T}_s}$ builds on the approximation of $\tilde{\mathbf{b}}$ by $\check{\mathbf{b}}$, the space $H(\check{\mathbf{b}}; \mathcal{T}_s) = \prod_{K \in \mathcal{T}_s} H(\check{\mathbf{b}}_K; K)$, equipped with the corresponding product norm $\|\cdot\|_{H(\check{\mathbf{b}};\mathcal{T}_s)}$, will play its role as a space "nearby" $\mathbb{V}_{\mathcal{T}_s} = H(\mathbf{b}; \mathcal{T}_s)$. In the next proposition, we equip $H(\check{\mathbf{b}}_K; K)$ with an equivalent Hilbertian norm that, as we will see, gives rise to a local Riesz lift $H(\check{\mathbf{b}}_K; K)' \to H(\check{\mathbf{b}}_K; K)$ of the residual of the modified bilinear form that can be determined *explicitly*.

Proposition 4.6 ([2, Remark 4.5]). *For* $\operatorname{diam}(K) \le |\check{\mathbf{b}}_K|$, *and with* $r(\mathbf{s})$ *denoting the distance from* $\mathbf{s} \in \partial K_-$ *to* ∂K_+ *along* $\check{\mathbf{b}}_K$, *the scalar product*

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

gives rise to a (uniform) equivalent norm $\vert\!\vert\!\vert \cdot \vert\!\vert\!\vert_{H(\check{\mathbf{b}}_K;K)}$ *on* $H(\check{\mathbf{b}}_K; K)$.

The corresponding global versions read

$$\langle\!\langle \cdot, \cdot \rangle\!\rangle_{H(\check{\mathbf{b}};\mathcal{T}_s)} = \sum_{K \in \mathcal{T}_s} \langle\!\langle \cdot|_K, \cdot|_K \rangle\!\rangle_{H(\check{\mathbf{b}}_K;K)},$$

and so $\vert\!\vert\!\vert \cdot \vert\!\vert\!\vert_{H(\check{\mathbf{b}};\mathcal{T}_s)} := \sqrt{\sum_{K \in \mathcal{T}_s} \vert\!\vert\!\vert \cdot|_K \vert\!\vert\!\vert^2_{H(\check{\mathbf{b}}_K;K)}}$.

For the next observation it is convenient to use the shorthand notations

$$\mu := w - u, \quad \lambda := \partial_{\check{\mathbf{b}}} w + \check{c} w - \tilde{f}, \quad \gamma := \lambda - (\partial_{\check{\mathbf{b}}} \mu + \check{c} \mu + \check{d}_K \mu),$$

so that, in particular, $\gamma = \partial_{\check{\mathbf{b}}} u + \check{c} u - \tilde{f} - \check{d}_K(w - u)$. Note also that

$$\mathcal{E}_{\mathcal{T}}(u, w, \tilde{f})^2 \approx \sum_{K \in \mathcal{T}_s} \|\mu\|^2_{L_2(K)} + \|\lambda\|^2_{L_2(K)}.$$

For smooth u, w, and $\tilde{f}$ on K, the solution $\check{R}_K = \check{R}_K(u, w; f) \in H(\check{\mathbf{b}}_K; K)$ of the variational problem

$$\langle\!\langle \check{R}_K, v \rangle\!\rangle_{H(\check{\mathbf{b}}_K;K)} = \check{b}_K(u, w; v) - \int_K \tilde{f} v \, d\mathbf{x} \quad (v \in H(\check{\mathbf{b}}_K; K)), \tag{4.9}$$

is the (strong) solution of

$$\begin{cases} -\partial^2_{\check{\mathbf{b}}_K} \check{R}_K = \gamma & \text{on } K, \\ \partial_{\check{\mathbf{b}}_K} \check{R}_K - r|\check{\mathbf{b}}_K|^{-1} \check{R}_K = \mu & \text{on } \partial K_-, \\ \partial_{\check{\mathbf{b}}_K} \check{R}_K = \mu & \text{on } \partial K_+. \end{cases}$$

This $\check{R}_K$ is the exact Riesz lift of the local modified residual

$$v \mapsto \check{b}_K(u, w|_{\partial K}, v) - \int_K \tilde{f} v \, d\mathbf{x} \in H(\check{\mathbf{b}}_K; K)',$$

with $H(\check{\mathbf{b}}_K; K)$ being equipped with $\langle\!\langle \cdot, \cdot \rangle\!\rangle_{H(\check{\mathbf{b}}_K;K)}$.

To identify next $\check{R}_K$ exactly, let $(x_1, \ldots, x_n)$ denote Cartesian coordinates on K with the first basis vector being equal to $\frac{\check{\mathbf{b}}_K}{|\check{\mathbf{b}}_K|}$. 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.
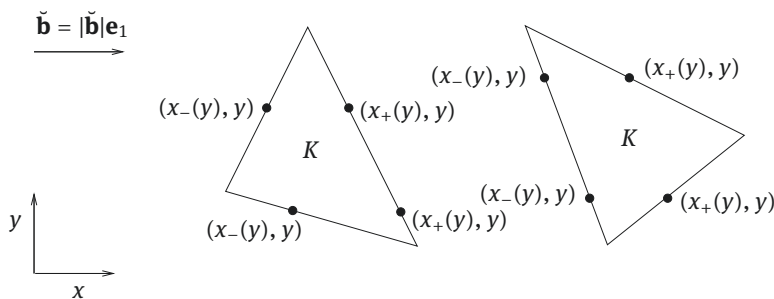

Figure 1: $x_\pm$ on a triangle K with two (left) or one (right) inflow boundaries. The enclosing triangle $\bar{K}$ and $\bar{x}_-$ will get their meaning in Section 4.3.

The solution $\check{R}_K$ reads then as

$$\check{R}_K(x, \mathbf{y}) = - |\check{\mathbf{b}}_K|^{-2} \int_{x_-(\mathbf{y})}^{x} \int_{x_-(\mathbf{y})}^{z} \gamma(q, \mathbf{y}) \, dq \, dz + \left(|\check{\mathbf{b}}_K|^{-1} \mu(x_+(\mathbf{y}), \mathbf{y}) + |\check{\mathbf{b}}_K|^{-2} \int_{x_-(\mathbf{y})}^{x_+(\mathbf{y})} \gamma(q, \mathbf{y}) \, dq \right)(x - x_-(\mathbf{y}))$$

$$+ \frac{\int_{x_-(\mathbf{y})}^{x_+(\mathbf{y})} (\partial_{\check{\mathbf{b}}_K} \mu + \gamma)(q, \mathbf{y}) \, dq}{x_+(\mathbf{y}) - x_-(\mathbf{y})},$$

and is seen to be piecewise polynomial over K when γ, μ are polynomial over K.

4.3 Approximate Lifted Residuals

Next we define an approximation $\breve{R}_K$ to $\check{R}_K$ by discarding higher order terms. Whereas, for polynomial u, w, $\tilde{f}$, $\tilde{\mathbf{b}}$, and $\tilde{c}$ on K, $\check{R}_K$ is only *piecewise* polynomial with respect to a partition of K into subsimplices (indicated by the dotted lines in Figure 1) that depends on the field $\check{\mathbf{b}}_K$, $\breve{R}_K$ will always be *polynomial* on K.

The reason for introducing $\breve{R}_K$ is that $\check{\mathbf{b}}^\perp \cdot \nabla \check{R}_K$ can be arbitrarily large, which would not allow us to perform Step (II) on page 12 of our proof. This is caused by the fact that the subdivision of K into the aforementioned subsimplices can have arbitrarily small angles, and thus impedes a useful application of the inverse (or Bernstein) inequality to $\check{R}_K$.

To define $\breve{R}_K$, first we construct a *polyhedral* set $\bar{K}$ that contains K as follows. The number of inflow faces of K is between 1 and $n - 1$, where n is the spatial dimension. Let F be the inflow face whose normal makes the smallest angle with $\check{\mathbf{b}}_K$, 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., $\bar{F} := \{\mathbf{x} \in H_F : \{\mathbf{x} + t\check{\mathbf{b}}_K : t \in \mathbb{R}\} \cap K \neq \emptyset\}$, is an $(n - 1)$-dimensional polyhedron containing F. 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 $\partial\bar{K}_- := \bar{F}$, and $K \subseteq \bar{K}$ with equality if and only if K has only *one* inflow face, namely $\partial K_- = F$.

For $\mathbf{x} = (x, \mathbf{y}) \in \bar{K} \supset K$, let $\mathbf{x} \mapsto \bar{x}_-(\mathbf{y}) \in \mathcal{P}_1(K)$ 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

$$\text{diam}(\bar{K}) \lesssim \text{diam}(K), \tag{4.10}$$

$$|\bar{x}_-|_{W^1_\infty(\bar{K})} \lesssim 1, \tag{4.11}$$

where both constants depend only on (an upper bound for) ϱ_K.

We define the *approximate lifted local residual* $\breve{R}_K = \breve{R}_K(u, w; f) \in \mathcal{P}_{m_v}(K)$ (cf. (4.1)) by

$$\breve{R}_K(x, \mathbf{y}) := |\check{\mathbf{b}}_K|^{-1} \mu(\bar{x}_-(\mathbf{y}), \mathbf{y})(x - \bar{x}_-(\mathbf{y})) + (\lambda - (\tilde{c} + \check{d}_K)\mu)(\bar{x}_-(\mathbf{y}), \mathbf{y}). \tag{4.12}$$

Note that $\partial_{\check{\mathbf{b}}_K} \breve{R}_K = \mu(\bar{x}_-(\mathbf{y}), \mathbf{y})$.

The following lemmas show how $\breve{R}_K$ relates on the one hand to the exact Riesz lift $\check{R}_K$ and on the other hand to the "residuals" $\mu = w - u$, $\lambda = \partial_{\tilde{\mathbf{b}}} w + \tilde{c} w - \tilde{f}$ on K.

Lemma 4.7. *For* $\text{diam}(K) \leq |\check{\mathbf{b}}_K|$, *it holds that*

$$\|\check{R}_K - \breve{R}_K\|_{H(\check{\mathbf{b}}_K; K)} \lesssim |\check{\mathbf{b}}_K|^{-1} \text{diam}(K)(\|\mu\|_{H(\check{\mathbf{b}}_K; \bar{K})} + \|\lambda\|_{H(\check{\mathbf{b}}_K; \bar{K})}) + \text{diam}(K)\|\mu\|_{H^1(K)},$$

with a constant depending only on (upper bounds for) $\|c\|_{W^1_\infty(K)}$, $|\mathbf{b}|_{W^1_\infty(K)^n}$, *and* ϱ_K.

Proof. We write $\check{R}_K - \breve{R}_K$ as

$$|\check{\mathbf{b}}_K|^{-2}\left((x - x_-(\mathbf{y})) \int_{x_-(\mathbf{y})}^{x_+(\mathbf{y})} \gamma(q, \mathbf{y}) \, dq - \int_{x_-(\mathbf{y})}^{x} \int_{x_-(\mathbf{y})}^{z} \gamma(q, \mathbf{y}) \, dq \right) \tag{4.13}$$

$$+ |\check{\mathbf{b}}_K|^{-1}(\mu(x_+(\mathbf{y}), \mathbf{y})(x - x_-(\mathbf{y})) - \mu(\bar{x}_-(\mathbf{y}), \mathbf{y})(x - \bar{x}_-(\mathbf{y}))) \tag{4.14}$$

$$+ \frac{\int_{x_-(\mathbf{y})}^{x_+(\mathbf{y})} (\partial_{\check{\mathbf{b}}_K} \mu + \gamma)(q, \mathbf{y}) \, dq}{x_+(\mathbf{y}) - x_-(\mathbf{y})} - (\lambda - (\tilde{c} + \check{d}_K)\mu)(\bar{x}_-(\mathbf{y}), \mathbf{y}). \tag{4.15}$$

Writing $\mu(x_+(\mathbf{y}),\mathbf{y})$ as $\mu(x,\mathbf{y}) + |\check{\mathbf{b}}_K|^{-1}\int_x^{x_+(\mathbf{y})}\partial_{\check{\mathbf{b}}_K}\mu(q,\mathbf{y})\,dq$, and similarly for $\mu(\bar{x}_-(\mathbf{y}),\mathbf{y})$, and using that $\mathrm{diam}(K)\le|\check{\mathbf{b}}_K|$, one infers that the $L_2(K)$-norm of (4.14) is

$$\lesssim |\check{\mathbf{b}}_K|^{-1}\,\mathrm{diam}(\bar{K})\|\mu\|_{L_2(K)} + |\check{\mathbf{b}}_K|^{-2}\,\mathrm{diam}(\bar{K})^2\|\partial_{\check{\mathbf{b}}_K}\mu\|_{L_2(\bar{K})}$$
$$\lesssim |\check{\mathbf{b}}_K|^{-1}\,\mathrm{diam}(K)\|\mu\|_{H(\check{\mathbf{b}}_K;\bar{K})},$$

with a constant only depending on ϱ_K.

The $L_2(K)$-norm of (4.13) in turn is

$$\lesssim |\check{\mathbf{b}}_K|^{-2}\,\mathrm{diam}(K)^2\|\gamma\|_{L_2(K)}$$
$$\lesssim |\check{\mathbf{b}}_K|^{-2}\,\mathrm{diam}(K)^2\big(\|\lambda\|_{L_2(K)} + (\|\tilde{c}\|_{L_\infty(K)} + |\check{d}_K|)\|\mu\|_{L_2(K)} + \||\tilde{\mathbf{b}}\|_{L_\infty(K)}|\mu|_{H^1(K)}\big)$$
$$\lesssim |\check{\mathbf{b}}_K|^{-1}\,\mathrm{diam}(K)\big(\|\lambda\|_{L_2(K)} + \|\mu\|_{L_2(K)}\big) + \mathrm{diam}(K)|\mu|_{H^1(K)},$$

with a constant depending only on (upper bounds for) $\|\tilde{c}\|_{L_\infty(K)}\lesssim\|c\|_{L_\infty(K)}$, and $\|\mathrm{div}\,\tilde{\mathbf{b}}\|_{L_2(K)}\lesssim|\tilde{\mathbf{b}}|_{W^1_\infty(K)^n}\lesssim|\mathbf{b}|_{W^1_\infty(K)^n}$, where we have used that $\||\tilde{\mathbf{b}}-\check{\mathbf{b}}_K\|_{L_\infty(K)}\lesssim\mathrm{diam}(K)|\tilde{\mathbf{b}}|_{W^1_\infty(K)^n}$ and $\mathrm{diam}(K)\le|\check{\mathbf{b}}_K|$.

Using that $\partial_{\check{\mathbf{b}}_K}\mu + \gamma = \lambda - (\tilde{c}+\check{d}_K)\mu + (\check{\mathbf{b}}_K - \tilde{\mathbf{b}})\cdot\nabla\mu$, we find that the $L_2(K)$-norm of (4.15) is bounded by a constant multiple of

$$|\check{\mathbf{b}}_K|^{-1}\,\mathrm{diam}(\bar{K})\|\partial_{\check{\mathbf{b}}_K}(\lambda - (\tilde{c}+\check{d}_K)\mu)\|_{L_2(\bar{K})} + |\tilde{\mathbf{b}}|_{W^1_\infty(K)^n}\,\mathrm{diam}(K)|\mu|_{H^1(K)}$$
$$\lesssim |\check{\mathbf{b}}_K|^{-1}\,\mathrm{diam}(K)\big(\|\partial_{\check{\mathbf{b}}_K}\lambda\|_{L_2(\bar{K})} + \|\mu\|_{H(\check{\mathbf{b}}_K;\bar{K})}\big) + \mathrm{diam}(K)\|\mu\|_{H^1(K)},$$

only dependent on (upper bounds for) $\|\tilde{c}\|_{W^1_\infty(K)}\lesssim\|c\|_{W^1_\infty(K)}$, $|\mathbf{b}|_{W^1_\infty(K)^n}$, and ϱ_K.

Next, we write

$$\partial_{\check{\mathbf{b}}_K}(\check{R}_K(x,\mathbf{y}) - \check{R}_K(x,\mathbf{y})) = \mu(x_+(\mathbf{y}),\mathbf{y}) + |\check{\mathbf{b}}_K|^{-1}\int_x^{x_+(\mathbf{y})}\gamma(q,\mathbf{y})\,dq - \mu(\bar{x}_-(\mathbf{y}),\mathbf{y}).$$

Its $L_2(K)$-norm is

$$\lesssim |\check{\mathbf{b}}_K|^{-1}\,\mathrm{diam}(K)\big(\|\partial_{\check{\mathbf{b}}_K}\mu\|_{L_2(\bar{K})} + \|\gamma\|_{L_2(K)}\big)$$
$$\lesssim |\check{\mathbf{b}}_K|^{-1}\,\mathrm{diam}(K)\big(\|\mu\|_{H(\check{\mathbf{b}}_K;\bar{K})} + \|\lambda\|_{L_2(K)}\big),$$

only dependent on (upper bounds for) $\|c\|_{L_\infty(K)}$, $\|\mathbf{b}\|_{W^1_\infty(K)^n}$, and ϱ_K. By collecting the derived upper bounds, the proof is completed. □

We end this subsection with another technical lemma which will play a key role to prove Step (I) on page 12. In fact, using that λ and μ are piecewise polynomial on $\mathcal{T}$, inverse inequalities will allow us to show that the terms involving first order derivatives can be kept small relative to the other ones by choosing the subgrid depth sufficiently large. Then the next lemma in conjunction with the previous Lemma 4.7 already hints at the fact that $\mathcal{E}_{\mathcal{T}}(u,w,\tilde{f})$ provides a lower bound for $\|\check{R}\|_{H(\check{\mathbf{b}};\mathcal{T}_s)}$. It then remains to switch to the correct norm to establish Step (I), see Corollary 4.9 below.

Lemma 4.8. *For $\mathrm{diam}(K)\le|\check{\mathbf{b}}_K|$, it holds that*

$$\|\check{R}_K\|^2_{H(\check{\mathbf{b}}_K;K)} + \mathrm{diam}(K)^2\big(|\mu|^2_{H^1(\bar{K})} + |\lambda|^2_{H^1(\bar{K})}\big) \gtrsim \|\lambda\|^2_{L_2(K)} + \|\mu\|^2_{L_2(K)},$$

where the constant depends only on (upper bounds for) $\|c\|_{L_\infty(\Omega)}$, $|\tilde{\mathbf{b}}|_{W^1_\infty(\Omega)^n}$, and ϱ_K.

Proof. By $\mathrm{diam}(\bar{K})\le\mathrm{diam}(K)\le|\check{\mathbf{b}}_K|$, similarly as in the proof of Lemma 4.7, one infers that

$$\|\check{R}_K - \lambda\|_{L_2(K)} \lesssim \|\mu\|_{L_2(K)} + \mathrm{diam}(K)\big(|\mu|_{H^1(\bar{K})} + |\lambda|_{H^1(\bar{K})}\big)$$

and

$$\|\partial_{\check{\mathbf{b}}_K}\check{R}_K - \mu\|_{L_2(K)} \lesssim \mathrm{diam}(K)|\mu|_{H^1(\bar{K})},$$

with constants depending only on (upper bounds for) $\|c\|_{L_\infty(\Omega)}$, $|\tilde{\mathbf{b}}|_{W^1_\infty(\Omega)^n}$, and ϱ_K.

By two applications of Young's inequality, we infer that

$$\|\check{R}_K\|^2_{L_2(K)} + \|\partial_{\check{\mathbf{b}}_K}\check{R}_K\|^2_{L_2(K)} \geq (1-\eta)\|\lambda\|^2_{L_2(K)} - (\eta^{-1}-1)\|\check{R}_K - \lambda\|^2_{L_2(K)} + \frac{1}{2}\|\mu\|^2_{L_2(K)} - (2-1)\|\partial_{\check{\mathbf{b}}_K}\check{R}_K - \mu\|^2_{L_2(K)}.$$

By selecting the constant $\eta \in (0, 1)$ sufficiently close to 1, the proof is completed. □

4.4 Proof of Proposition 4.2

So far we have not used that u, w, $\tilde{f}$, $\mathbf{b}$, and $\tilde{c}$ are piecewise polynomial with respect to $\mathcal{T}$, whilst $\check{b}_{\mathcal{T}_s}(\cdot, \cdot)$ is a "broken" bilinear form with respect to a sufficiently refined partition $\mathcal{T}_s$, and furthermore that $|\mathbf{b}|^{-1} \in L_\infty(\Omega)$. These facts are going to be used in the following.

Setting

$$D := \sup_{K \in \mathfrak{T}} \sup_{0 \neq \mathbf{b} \in W^1_\infty(K)^n} \frac{\||\check{\mathbf{b}}_K - \mathbf{b}|\|_{L_\infty(K)}}{\mathrm{diam}(K)|\mathbf{b}|_{W^1_\infty(K)^n}} \ (< \infty),$$

let

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

be such that for $\sigma \in (0, \bar{\sigma}]$ and $\mathcal{T} \in \mathbb{T}$, $\mathcal{T}_s$ is sufficiently fine to ensure that

$$\mathrm{diam}(K)\||\mathbf{b}|^{-1}\|_{L_\infty(K)} \max\left(1, D|\mathbf{b}|_{W^1_\infty(K)^n}\right) \leq \frac{1}{2} \quad (K \in \mathcal{T}_s). \tag{4.16}$$

Then for any $K \in \mathcal{T}_s$, we have

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

where we have used (4.16).

For $K \in \mathfrak{T}$, and $k \geq \ell \in \mathbb{N}_0$, we will make repeated use of the *inverse inequality*

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

where the constant depends only on m, ϱ_K, and k.

Corollary 4.9. *We define $\check{R}$ and $\check{\check{R}}$ by $\check{R}|_K := \check{R}_K$ and $\check{\check{R}}|_K := \check{\check{R}}_K$ for $K \in \mathcal{T}_s$. Then one has for $(u, w, f) \in \mathbb{D}_{\mathcal{T}}$, $\sigma \in (0, \bar{\sigma}]$ that*
(i) $\|\check{R} - \check{\check{R}}\|_{H(\check{\mathbf{b}}; \mathcal{T}_s)} \lesssim \sigma\mathcal{E}_{\mathcal{T}}(u, w, \tilde{f})$,
(ii) $\|\check{\check{R}}\|_{H(\check{\mathbf{b}}; \mathcal{T}_s)} \gtrsim \mathcal{E}_{\mathcal{T}}(u, w, \tilde{f})$, provided that $\sigma \in (0, \sigma_0]$ with $\sigma_0 \in (0, \bar{\sigma}]$ being sufficiently small.
Both constants hidden in the $\lesssim$ and $\gtrsim$ symbols, and the upper bound for σ_0 depend only on the quantities mentioned in the statement of Proposition 4.2.

Proof. For $K' \in \mathfrak{T}$ and $p \in \mathcal{P}_m(K')$, we have that

$$\sum_{\{K \in \mathcal{T}_s : K \subset K'\}} |p|^2_{H^1(\bar{K})} \approx \sum_{\{K \in \mathcal{T}_s : K \subset K'\}} |p|^2_{H^1(K)} = |p|^2_{H^1(K')} \lesssim \mathrm{diam}(K')^{-2}\|p\|^2_{L_2(K')}, \tag{4.17}$$

with a constant depending on ϱ and m. By applying this type of estimate to λ and μ, preceded by an application of Lemma 4.7 whilst using $|\check{\mathbf{b}}_K|^{-1} \leq 2\||\mathbf{b}|^{-1}\|_{L_\infty(\Omega)}$ and $|\check{\mathbf{b}}_K| \leq \|\tilde{\mathbf{b}}\|_{L_\infty(K)} \leq 2\|\mathbf{b}\|_{L_\infty(\Omega)}$, we obtain

$$\|\check{R} - \check{\check{R}}\|_{H(\check{\mathbf{b}}; \mathcal{T}_s)} \lesssim \sigma\mathcal{E}_{\mathcal{T}}(u, w, \tilde{f}).$$

By summing the result of Lemma 4.8 over $K \in \mathcal{T}_s$ and applying (4.17) with $p = \mu$ and $p = \lambda$, we infer that for σ small enough, $\|\check{\check{R}}\|_{H(\check{\mathbf{b}}; \mathcal{T}_s)} \gtrsim \mathcal{E}_{\mathcal{T}}(u, w, \tilde{f})$. □

The next proposition is almost Step (I) on page 12, except that we still have to replace $\|\check{\check{R}}\|_{H(\check{\mathbf{b}}; \Omega)}$ by $\|\check{\check{R}}\|_{\mathbb{V}_{\mathcal{T}_s}}$, which will be done using the subsequent Lemma 4.11 (b).

Proposition 4.10. *There exist a $\kappa > 0$ and a $\sigma_1 \in (0, \sigma_0]$, that depend only on the quantities mentioned in the statement of Proposition 4.2, such that for $\sigma \in (0, \sigma_1]$, and any $(u, w, f) \in \mathbb{D}_{\mathcal{T}}$,*

$$\breve{b}_{\mathcal{T}_s}(u, w; \breve{R}) - \int_\Omega f \breve{R}\, d\mathbf{x} \geq \kappa \mathcal{E}_{\mathcal{T}}(u, w, \tilde{f}) \|\breve{R}\|_{H(\breve{\mathbf{b}}; \mathcal{T}_s)}.$$

Proof. With $\breve{R}|_K := \breve{R}_K(u, w; f)$, its definition in (4.9) shows that

$$\breve{b}_{\mathcal{T}_s}(u, w; \breve{R}) - \int_\Omega f \breve{R}\, d\mathbf{x} = \sum_{K \in \mathcal{T}_s} \langle\!\langle \breve{R}_K, \breve{R}_K \rangle\!\rangle_{H(\breve{\mathbf{b}}_K; K)}.$$

Thanks to the equivalence of norms from Proposition 4.6, an application of Corollary 4.9((i)) shows that

$$\left| \sum_{K \in \mathcal{T}_s} \langle\!\langle \breve{R}_K - \breve{\breve{R}}_K, \breve{R}_K \rangle\!\rangle_{H(\breve{\mathbf{b}}_K; K)} \right| \lesssim \sigma \mathcal{E}_{\mathcal{T}}(u, w, \tilde{f}) \|\breve{R}\|_{H(\breve{\mathbf{b}}; \mathcal{T}_s)}.$$

For σ being sufficiently small, an application of Corollary 4.9 (ii) shows that

$$\|\breve{\breve{R}}\|^2_{H(\breve{\mathbf{b}}; \mathcal{T}_s)} \approx \|\breve{R}\|^2_{H(\breve{\mathbf{b}}; \mathcal{T}_s)} \gtrsim \mathcal{E}_{\mathcal{T}}(u, w, \tilde{f}) \|\breve{R}\|_{H(\breve{\mathbf{b}}; \mathcal{T}_s)},$$

by which the proof is easily completed. □

Lemma 4.11. *For $(u, w, f) \in \mathbb{D}_{\mathcal{T}}$, $\sigma \in (0, \sigma_0]$, it holds that*
(a) $\sum_{K \in \mathcal{T}_s} \operatorname{diam}(K)^2 \|\breve{R}_K\|^2_{H^1(K)} \lesssim \sigma^2 \|\breve{R}\|^2_{H(\breve{\mathbf{b}}; \Omega)}$,
(b) $\left| \|\breve{R}\|_{H(\breve{\mathbf{b}}; \mathcal{T}_s)} - \|\breve{R}\|_{\mathbb{V}_{\mathcal{T}_s}} \right| \lesssim \sigma \|\breve{R}\|_{H(\breve{\mathbf{b}}; \mathcal{T}_s)}$,
depending only on the quantities mentioned in the statement of Proposition 4.2.

Proof. (a) For $K \in \mathcal{T}_s$, we split $\breve{R}_K = \breve{R}_{K,1} + \breve{R}_{K,2} + \breve{R}_{K,3}$ defined by

$$\breve{R}_{1,K}(x, \mathbf{y}) := |\breve{\mathbf{b}}_K|^{-1} \mu(\bar{x}_-(\mathbf{y}), \mathbf{y})(x - \bar{x}_-(\mathbf{y})),$$

$$\breve{R}_{2,K}(x, \mathbf{y}) := (\lambda - (\tilde{c} + \breve{d}_{K'})\mu)(\bar{x}_-(\mathbf{y}), \mathbf{y}),$$

$$\breve{R}_{3,K}(x, \mathbf{y}) := (\breve{d}_{K'} - \breve{d}_K)\mu(\bar{x}_-(\mathbf{y}), \mathbf{y}),$$

where $K' \in \mathcal{T}$ is such that $K \subset K'$. Correspondingly, we split $\breve{R} = \breve{R}_1 + \breve{R}_2 + \breve{R}_3$.

Since $\breve{R}_{1,K}$ vanishes on $\partial \bar{K}_-$, an application of Poincaré's inequality on each streamline following $\breve{\mathbf{b}}_K$ shows that (cf. possibly [2, Proposition 4.3])

$$\|\breve{R}_{1,K}\|_{L_2(\bar{K})} \lesssim |\breve{\mathbf{b}}_K|^{-1} \operatorname{diam}(\bar{K}) \|\partial_{\breve{\mathbf{b}}_K} \breve{R}_{1,K}\|_{L_2(\bar{K})}.$$

From the fact that $\breve{R}_{1,K}$ is polynomial, $\operatorname{diam}(\bar{K}) \lesssim \operatorname{diam}(K)$, and $\partial_{\breve{\mathbf{b}}_K} \breve{R}_{1,K} = \partial_{\breve{\mathbf{b}}_K} \breve{R}_K$, by an application of the inverse inequality we obtain

$$\sum_{K \in \mathcal{T}_s} \operatorname{diam}(K)^2 \|\breve{R}_{1,K}\|^2_{H^1(K)} \lesssim \sum_{K \in \mathcal{T}_s} |\breve{\mathbf{b}}_K|^{-2} \operatorname{diam}(K)^2 \|\partial_{\breve{\mathbf{b}}_K} \breve{R}_K\|^2_{L_2(K)} \lesssim \sigma^2 \|\breve{R}\|^2_{H(\breve{\mathbf{b}}; \Omega)}.$$

Recalling from (4.11) that $|\bar{x}_-|_{W^1_\infty(K)} \lesssim 1$, and since $\lambda - (\tilde{c} + \breve{d}_{K'})\mu$ is polynomial on K, we have

$$\|(x, \mathbf{y}) \mapsto (\lambda - (\tilde{c} + \breve{d}_{K'})\mu)(\bar{x}_-(\mathbf{y}), \mathbf{y})\|_{H^1(K)} \lesssim \|\lambda - (\tilde{c} + \breve{d}_{K'})\mu\|_{H^1(K)}.$$

Now using that for $\mathcal{T} \ni K' \supset K$, $\lambda - (\tilde{c} + \breve{d}_{K'})\mu$ is polynomial on K', an application of (4.17) shows that

$$\sum_{K \in \mathcal{T}_s} \operatorname{diam}(K)^2 \|\breve{R}_{2,K}\|^2_{H^1(K)} \lesssim \sigma^2 \mathcal{E}_{\mathcal{T}}(u, w, \tilde{f})^2 \lesssim \sigma^2 \|\breve{R}\|^2_{H(\breve{\mathbf{b}}; \mathcal{T}_s)},$$

where the last inequality follows from Corollary 4.9 (ii).

Again by $|\bar{x}_-|_{W^1_\infty(K)} \lesssim 1$, we have $\|\breve{R}_{3,K}\|_{H^1(K)} \lesssim |\breve{\mathbf{b}}|_{W^1_\infty(K')} \|\mu\|_{H^1(K)}$, which together with (4.17) yields that

$$\sum_{K \in \mathcal{T}_s} \operatorname{diam}(K)^2 \|\breve{R}_{3,K}\|^2_{H^1(K)} \lesssim \sum_{K \in \mathcal{T}_s} \operatorname{diam}(K)^2 \|\mu\|^2_{L_2(K)} \lesssim \sigma^2 \|\breve{R}\|^2_{H(\breve{\mathbf{b}}; \mathcal{T}_s)}.$$

by Corollary 4.9 (ii), which completes the proof of (a).

(b) From the triangle inequality, and $\|\|\breve{\mathbf{b}}_K - \mathbf{b}\|\|_{L_\infty(K)} \le D \operatorname{diam}(K)|\mathbf{b}|_{W_\infty^1(K)^n}$, we infer that

$$\left| \|\breve{R}\|_{H(\breve{\mathbf{b}};\mathcal{T}_s)} - \|\breve{R}\|_{\mathbb{V}_{\mathcal{T}_s}} \right| \lesssim \sqrt{\sum_{K \in \mathcal{T}_s} \operatorname{diam}(K)^2 |\breve{R}_K|_{H^1(K)}^2} \lesssim \sigma \|\breve{R}\|_{H(\breve{\mathbf{b}};\Omega)}$$

by (a), which is (b). $\qquad\qquad\square$

Proposition 4.10 together with Lemma 4.11 (b) complete the proof of

$$\breve{b}_{\mathcal{T}_s}(u, w; \breve{R}) - \int_\Omega f \breve{R} \, d\mathbf{x} \ge \kappa \mathcal{E}_{\mathcal{T}}(u, w, \tilde{f}) \|\breve{R}\|_{\mathbb{V}_{\mathcal{T}_s}}$$

for sufficiently small $\sigma > 0$, being Step (I) in our proof of Proposition 4.2.

Step (II) is implied by the next result when we use that $\|u - w\|_{L_2(\Omega)} \le \mathcal{E}_{\mathcal{T}}(u, w, \tilde{f})$.

Proposition 4.12. *For $(u, w, f) \in \mathbb{D}_{\mathcal{T}}$, $\sigma \in (0, \sigma_0]$ sufficiently small, it holds that*

$$|\tilde{b}_{\mathcal{T}_s}(u, w; \breve{R}) - \breve{b}_{\mathcal{T}_s}(u, w; \breve{R})| \lesssim \sigma \|u - w\|_{L_2(\Omega)} \|\breve{R}\|_{\mathbb{V}_{\mathcal{T}_s}}.$$

Both the upper bound for σ and the constant hidden in the $\lesssim$-symbol depend only on the quantities mentioned in the statement of Proposition 4.2.

Proof. For $K \in \mathcal{T}_s$ and sufficiently smooth u, w, and v, it holds that

$$\tilde{b}_K(u, w; v) = \int_K (\tilde{c} u + \tilde{\mathbf{b}} \cdot \nabla u) v \, d\mathbf{x} + \int_{\partial K} \tilde{\mathbf{b}} \cdot \mathbf{n}_K (w - u) v \, ds,$$

and so

$$\tilde{b}_K(u, w; v) - \breve{b}_K(u, w; v) = - \int_K \breve{d}_K (w - u) v \, d\mathbf{x} + \int_{\partial K} (\tilde{\mathbf{b}} - \breve{\mathbf{b}}_K) \cdot \mathbf{n}_K (w - u) v \, ds.$$

With $z := (w - u)v$, and $\bar{z} := |K|^{-1} \int_K z \, d\mathbf{x}$, recalling that $\breve{d}_K = |K|^{-1} \int_K \operatorname{div} \tilde{\mathbf{b}} \, d\mathbf{x}$, an application of the trace theorem shows that

$$\left| - \int_K \breve{d}_K z \, d\mathbf{x} + \int_{\partial K} (\tilde{\mathbf{b}} - \breve{\mathbf{b}}_K) \cdot \mathbf{n}_K z \, ds \right| = \left| - \int_K \breve{d}_K (z - \bar{z}) \, d\mathbf{x} + \int_{\partial K} (\tilde{\mathbf{b}} - \breve{\mathbf{b}}_K) \cdot \mathbf{n}_K (z - \bar{z}) \, ds \right|$$

$$\lesssim \|\operatorname{div} \tilde{\mathbf{b}}\|_{L_\infty(K)} \|z - \bar{z}\|_{L_1(K)} + \|\|\tilde{\mathbf{b}} - \breve{\mathbf{b}}_K\|\|_{L_\infty(K)} |z|_{W_1^1(K)}$$

$$\lesssim \operatorname{diam}(K) |z|_{W_1^1(K)}$$

$$\lesssim \operatorname{diam}(K) \left(\|w - u\|_{L_2(K)} |v|_{H^1(K)} + \|v\|_{L_2(K)} |w - u|_{H^1(K)} \right).$$

By substituting $v = \breve{R}_K$, summing over $K \in \mathcal{T}_s$, and applying the Cauchy–Schwarz inequality, we find that

$$|\tilde{b}_{\mathcal{T}_s}(u, w; \breve{R}) - \breve{b}_{\mathcal{T}_s}(u, w; \breve{R})| \lesssim \|w - u\|_{L_2(\Omega)} \sqrt{\sum_{K \in \mathcal{T}_s} \operatorname{diam}(K)^2 |\breve{R}|_{H^1(K)}^2} + \|\breve{R}\|_{L_2(\Omega)} \sqrt{\sum_{K \in \mathcal{T}_s} \operatorname{diam}(K)^2 |u - w|_{H^1(K)}^2}$$

$$\lesssim \sigma \|u - w\|_{L_2(\Omega)} \|\breve{R}\|_{H(\breve{\mathbf{b}};\mathcal{T}_s)}$$

where we have applied (4.17) and Lemma 4.11 (a). Finally, for sufficiently small σ, in the last expression $\|\breve{R}\|_{H(\breve{\mathbf{b}};\mathcal{T}_s)}$ can be replaced in view of Lemma 4.11 (b) by $\|\breve{R}\|_{\mathbb{V}_{\mathcal{T}_s}}$. $\qquad\square$

Since we have performed Steps (I)–(II) on page 12, the proof of Proposition 4.2 is complete.

5 Effective Mark and Refinement Strategy for an Adaptive DPG Method

The key common ingredient of an adaptive solution strategy for a PDE is a collection of local error indicators associated with the current partition $\mathcal{T}$ underlying the discretization. While an individual indicator does

not characterize the actual local error the accumulation of all indicators is equivalent to the global current approximation error. Based on the error indicators one contrives a *marking strategy* which identifies a subset $\mathcal{M} \subset \mathcal{T}$ of *marked cells* to be refined in the subsequent adaptive step. The perhaps most prominent marking strategy is based on a *bulk criterion*, sometimes called "Dörfler Marking" where one collects (a possibly small number of) cells for which the accumulated combined indicators capture at least a given fixed portion of the global a posteriori error bound. While this is usually perceived as a heuristically very plausible strategy, a rigorous convergence and complexity analysis is actually quite intricate. It typically comes in two stages, namely establishing first that such a strategy reduces the current error by a fixed ratio, and second to estimate the number of new degrees of freedom incurred by the refinement step. This paradigm has been studied extensively and is by now well understood for problems of elliptic type where the dominating effect is diffusion. The first step of error reduction hinges on (near-)*Galerkin orthogonality* and is greatly helped by the fact that the common residual based error indicators contain as an explicit factor a power of the respective cell diameter. Thus, a refinement does decrease the indicators.

In the current scenario of transport equations the situation looks similar at the first glance. Using $(u, w) \in \check{\mathbb{U}}_{\mathcal{T}}^{\delta}$ as primal unknowns, we have a hierarchy of *nested* trial spaces at hand, see Remark 3.5. Due to the product structure of the test search spaces we have computable local error indicators associated with the current discretization whose sum is, thanks to Theorem 4.1, modulo data oscillation uniformly equivalent to the error in the trial metric. This suggests using a similar bulk criterion in a *mark-and-refine framework* to drive adaptive refinements which is, in fact our choice in the subsequent discussion.

A closer look reveals, however, some essential distinctions which may actually nourish some doubts about whether such strategies work in a transport problem just as well as in a diffusion problem. The error indicators in the form of projected lifted residuals depend of course on the mesh defining the DPG scheme but they do *not* contain any local mesh size factor that ensures a decay under refinement. In contrast to the usual way of analyzing residual based a posteriori error estimators we are able to deduce a fixed error reduction rate only when starting from a Petrov–Galerkin solution using what one may call *Petrov–Galerkin orthogonality* in place of Galerkin orthogonality. Moreover, there is actually an *infinite* family of equivalent a posteriori bounds obtained for *any* refinement of the current partition arising from different mesh-dependent Riesz liftings. A key observation, which we heavily exploit and which may actually be of interest in its own right, is the interrelation of these error indicators with yet another completely *mesh-independent* variant representing the residual for a least squares formulation.

As indicated by these comments the crucial issue for adaptivity in the context of transport equations is the *effectivity* of a given mark and refinement strategy in the sense of a guaranteed error reduction rate. The basic structure of a subsequent complexity analysis can instead be expected to be less problem specific. We therefore confine the subsequent discussion entirely to the issue of effectivity which we are currently only able to fully establish in one spatial dimension $n = 1$.

For $n > 1$ we will employ a *downstream enriched* refinement strategy where additionally cells downwind from the marked cells are refined as well. Our derivation of effectivity in this case will partly be based on a conjecture.

5.1 Setting and Results

In view of the already considerable level of technicality we confine the subsequent discussion to the case of a *constant* convection field **b**, and a piecewise constant reaction coefficient c with respect to the current partition $\mathcal{T}$ for the trial space. In an adaptive setting the latter means that necessarily c is piecewise constant with respect to the initial partition $\mathcal{T}_{\perp}$, i.e., we always assume that

$$\mathbf{b}(x) \equiv \mathbf{b}, \quad c = (c_{K'})_{K' \in \mathcal{T}_{\perp}} \in \mathcal{P}_0(\mathcal{T}_{\perp}). \tag{5.1}$$

Given $(u, w) \in \check{\mathbb{U}}_{\mathcal{T}}^{\delta}$ and $f \in \mathbb{F}_{\mathcal{T}}^{\delta}$, from (4.3) recall the definition of the *projected lifted residual*

$$R_{\mathcal{T}_s}^{\delta} = R_{\mathcal{T}_s}^{\delta}(u, w; f) = (R_K^{\delta})_{K \in \mathcal{T}_s} \in \bar{\mathbb{V}}_{\mathcal{T}_s}^{\delta} \subset \mathbb{V}_{\mathcal{T}_s}.$$

For a collection of marked cells $\mathcal{M} \subset \mathcal{T}$, we set

$$\mathcal{T}_s(\mathcal{M}) = \left\{ K \in \mathcal{T}_s : K \subset \bigcup_{K' \in \mathcal{M}} K' \right\}$$

for the corresponding portion of the test-subgrid with the convention $\mathcal{T}_s = \mathcal{T}_s(\mathcal{T})$. We use the notation $R^\delta_{\mathcal{T}_s(\mathcal{M})}$ to denote $(R^\delta_K)_{K \in \mathcal{T}_s(\mathcal{M})}$.

Aside from a partition $\mathcal{T}$ and its refinement $\mathcal{T}_s$, we consider a refined partition $\tilde{\mathcal{T}}$ with companion refinement $\tilde{\mathcal{T}}_s$. Note that $\breve{U}^\delta_{\mathcal{T}} \subset \breve{U}^\delta_{\tilde{\mathcal{T}}}$, $\bar{V}^\delta_{\mathcal{T}_s} \subset \bar{V}^\delta_{\tilde{\mathcal{T}}_s}$, and $\mathbb{F}^\delta_{\mathcal{T}} \subset \mathbb{F}^\delta_{\tilde{\mathcal{T}}}$, see Remark 3.5.

Since in the current setting $\mathrm{osc}_{\mathcal{T}}(\mathbf{b}, c, f) = 0$, according to (4.2) one has for any $(u, w) \in \breve{U}^\delta_{\mathcal{T}}$,

$$\|R^\delta_{\mathcal{T}_s}(u, w; f)\|_{H(\mathbf{b};\mathcal{T}_s)} \eqsim \|(u, w) - (u^{\mathrm{ex}}, u^{\mathrm{ex}})\|_{\breve{U}}, \tag{5.2}$$

errors are thus uniformly equivalent to sums of computable local quantities that suggest themselves as error indicators.

Definition 5.1. For $r \in \mathbb{N}$ and $v \in (0, 1)$, we say that a strategy of marking $\mathcal{M} \subset \mathcal{T}$ is (r, v)-*effective* when for $\tilde{\mathcal{T}} = \tilde{\mathcal{T}}(\mathcal{T}, \mathcal{M}, r) \in \mathbb{T}$, obtained from $\mathcal{T}$ by r refinements of each $K' \in \mathcal{M}$, and for $n > 1$, of each $K' \in \mathcal{T}$ with $K' \cap \bigcup_{K'' \in \mathcal{M}, t \geq 0} K'' + t\mathbf{b} \neq \emptyset$, it holds that

$$\|R^\delta_{\tilde{\mathcal{T}}_s}(u^\delta_{\tilde{\mathcal{T}}}, w^\delta_{\tilde{\mathcal{T}}}; f)\|_{H(\mathbf{b};\tilde{\mathcal{T}}_s)} \leq v \|R^\delta_{\mathcal{T}_s}(u^\delta_{\mathcal{T}}, w^\delta_{\mathcal{T}}; f)\|_{H(\mathbf{b};\mathcal{T}_s)},$$

where $(u^\delta_{\mathcal{T}}, w^\delta_{\mathcal{T}})$, $(u^\delta_{\tilde{\mathcal{T}}}, w^\delta_{\tilde{\mathcal{T}}})$ are the Petrov–Galerkin solutions of (3.7) from $\breve{U}^\delta_{\mathcal{T}}$, $\breve{U}^\delta_{\tilde{\mathcal{T}}}$, respectively.

Note that only for $n > 1$ the refinement includes a *downstream enrichment* comprised of those cells that are intersected by rays in direction $\mathbf{b}$ emanating from cells in $\mathcal{M}$.

Remark 5.2. A repeated application, starting from some initial partition, of mark followed by the downwind enriched refinement strategy, described in Definition 5.1, ensures that no mesh can ever become coarser in the down-stream direction.

Of course, by (5.2), (r, v)-effectiveness translates for some $v' \in (0, 1)$ into error decay for the solutions

$$\|(u^{\mathrm{ex}}, u^{\mathrm{ex}}) - (u^\delta_{\tilde{\mathcal{T}}}, w^\delta_{\tilde{\mathcal{T}}})\|_{U} \leq v' \|(u^{\mathrm{ex}}, u^{\mathrm{ex}}) - (u^\delta_{\mathcal{T}}, w^\delta_{\mathcal{T}})\|_{U},$$

where now $\tilde{\mathcal{T}}$ is to be understood as the result of possibly several but uniformly bounded finite number of refinements of the above type.

As indicated earlier, our goal is to prove effectiveness for a marking strategy based on a bulk-criterion. To make this precise for some $\vartheta \in (0, 1]$, $(u, w) \in \breve{U}^\delta_{\mathcal{T}}$, we let

$$\mathcal{M} = \mathcal{M}((u, w), \vartheta) \subseteq \mathcal{T} \quad \text{be such that} \quad \|R^\delta_{\mathcal{T}_s(\mathcal{M})}(u, w; f)\|_{H(\mathbf{b};\mathcal{T}_s(\mathcal{M}))} \geq \vartheta \|R^\delta_{\mathcal{T}_s}(u, w; f)\|_{H(\mathbf{b};\mathcal{T}_s)}. \tag{5.3}$$

We are currently able to fully establish effectivity of the standard bulk chasing strategy based on refining just cells in $\mathcal{M}$ given by (5.3), only in the one-dimensional case.

Theorem 5.3. *We adopt the assumptions of Theorem 4.1 with the additional assumption $m_w \leq m_u + 1$, and the specifications (5.1) of $\mathbf{b}$ and c. Then, for $n = 1$ and σ sufficiently small there exist $r \in \mathbb{N}$, $v = v(\vartheta) < 1$ such that the marking strategy based on (5.3) is (r, v)-effective for $\mathcal{T} \in \mathbb{T}$, $f \in \mathbb{F}^\delta_{\mathcal{T}}$.*

Under the forthcoming Conjecture 5.17, the same result holds true for $n > 1$ (thus with the downwind enriched refinement strategy).

The remainder of this section is to develop the conceptual ingredients entering results of the above type.

A first natural ingredient for proving Theorem 5.3 seems to be Petrov–Galerkin orthogonality (2.10)

$$\|R^\delta_{\tilde{\mathcal{T}}_s}(u^\delta_{\tilde{\mathcal{T}}}, w^\delta_{\tilde{\mathcal{T}}}; f)\|^2_{H(\mathbf{b};\tilde{\mathcal{T}}_s)} = \|R^\delta_{\tilde{\mathcal{T}}_s}(u^\delta_{\mathcal{T}}, w^\delta_{\mathcal{T}}; f)\|^2_{H(\mathbf{b};\tilde{\mathcal{T}}_s)} - \|t^\delta_{\tilde{\mathcal{T}}_s}(u^\delta_{\mathcal{T}} - u^\delta_{\tilde{\mathcal{T}}}, w^\delta_{\mathcal{T}} - w^\delta_{\tilde{\mathcal{T}}})\|^2_{H(\mathbf{b};\tilde{\mathcal{T}}_s)} \tag{5.4}$$

in combination with a proof of

$$\|t^\delta_{\tilde{\mathcal{T}}_s}(u^\delta_{\mathcal{T}} - u^\delta_{\tilde{\mathcal{T}}}, w^\delta_{\mathcal{T}} - w^\delta_{\tilde{\mathcal{T}}})\|_{H(\mathbf{b};\tilde{\mathcal{T}}_s)} \gtrsim \|R^\delta_{\tilde{\mathcal{T}}_s}(u^\delta_{\mathcal{T}}, w^\delta_{\mathcal{T}}; f)\|_{H(\mathbf{b};\tilde{\mathcal{T}}_s)}.$$

A complication, however, is the presence of the "wrong" mesh-dependent lifting $R^\delta_{\tilde{\mathcal{T}}_s}(u^\delta_{\mathcal{T}}, w^\delta_{\mathcal{T}};f)$ instead of $R^\delta_{\mathcal{T}_s}(u^\delta_{\mathcal{T}}, w^\delta_{\mathcal{T}};f)$ in the first term on the right-hand side of (5.4). To tackle this problem, in the next subsection we construct mesh-independent error indicators. (The appearance of the "wrong" norm $\|\cdot\|_{H(\mathbf{b};\tilde{\mathcal{T}}_s)}$ instead of $\|\cdot\|_{H(\mathbf{b};\mathcal{T}_s)}$ does not cause any problems because $\|R^\delta_{\mathcal{T}_s}(u^\delta_{\mathcal{T}}, w^\delta_{\mathcal{T}};f)\|_{H(\mathbf{b};\tilde{\mathcal{T}}_s)} = \|R^\delta_{\mathcal{T}_s}(u^\delta_{\mathcal{T}}, w^\delta_{\mathcal{T}};f)\|_{H(\mathbf{b};\mathcal{T}_s)}$.)

5.2 A Mesh-Independent Error Indicator and Related Least Squares Problems

In the light of the remarks at the end of the previous subsection we quantify next the interrelation of various equivalent error indicators arising from different liftings as well as from different equivalent inner products. A pivotal role is played by the following "domain-additive" quantity. For any subdomain $\Omega' \subseteq \Omega$ we introduce

$$\eta^2_{\Omega'}(u, w; f) := \|u - w\|^2_{L_2(\Omega')} + \|\partial_\mathbf{b} w + cu - f\|^2_{L_2(\Omega')}. \qquad (5.5)$$

Accordingly, for a collection $\mathcal{O}$ of subdomains, we define

$$\eta^2_{\mathcal{O}}(u, w; f) = \sum_{\Omega' \in \mathcal{O}} \eta^2_{\Omega'}(u, w; f) = \eta^2_{\bigcup\{\Omega' \in \mathcal{O}\}}(u, w; f).$$

Note that for $f \in L_2(\Omega)$ and $\Omega' = \Omega$ both components (u, w) of the minimizer of (5.5) over $L_2(\Omega) \times H(\mathbf{b}; \Omega)$ agree with the minimizer $w \in H(\mathbf{b}; \Omega)$ of the *least squares* functional

$$\|\mathcal{B}w - f\|^2_{L_2(\Omega)} = \|\partial_\mathbf{b} w + cw - f\|^2_{L_2(\Omega')},$$

see the comment in Section 5.3.1 below.

As indicated above it will be crucial to relate these mesh-independent quantities to the following quantities each of which being useful for different purposes: Besides the projected lifted residual from (4.3), recall first the definitions of the *lifted residual*

$$R_{\mathcal{T}_s} = R_{\mathcal{T}_s}(u, w; f) = (R_K)_{K \in \mathcal{T}_s} \in \mathbb{V}_{\mathcal{T}_s} = H(\mathbf{b}; \mathcal{T}_s),$$

determined by $\langle R_{\mathcal{T}_s}, v \rangle_{H(\mathbf{b},\mathcal{T}_s)} = b_{\mathcal{T}_s}(u, w|_{\partial\mathcal{T}_s}; v) - \int_\Omega fv\, d\mathbf{x}$ ($v \in \mathbb{V}_{\mathcal{T}_s}$). In a similar spirit as in the analysis of test functions we need to make use of the *lifted modified residual* (cf. (4.9))

$$\check{R}_{\mathcal{T}_s} = \check{R}_{\mathcal{T}_s}(u, w; f) = (\check{R}_K)_{K \in \mathcal{T}_s} \in \mathbb{V}_{\mathcal{T}_s},$$

and the piecewise polynomial *approximate lifted modified residual* (cf. (4.12))

$$\check{\check{R}}_{\mathcal{T}_s} = \check{\check{R}}_{\mathcal{T}_s}(u, w; f) = (\check{\check{R}}_K)_{K \in \mathcal{T}_s} \in \bar{\mathbb{V}}^\delta_{\mathcal{T}_s}.$$

In the current setting of $\mathbf{b}$ being a constant, and so $d_K \equiv 0$ and $b_{\mathcal{T}_s} = \check{b}_{\mathcal{T}_s}$, the lifted residual and the lifted modified residual differ only in the sense that $\check{R}_K$ is the lift of the local residual with respect to the alternative inner product $\langle\!\langle \cdot, \cdot \rangle\!\rangle_{H(\mathbf{b};K)}$ on $H(\mathbf{b}; K)$.

The advantage of the latter quantity is its simple explicit analytic expression from which one can actually see the connection with (5.5) as the "limit case" with respect to increasing subgrid depth. In fact, for $K' \in \mathcal{T}$ we will show that

$$\|\check{\check{R}}_{\mathcal{T}_s(K')}\|^2_{H(\mathbf{b};\mathcal{T}_s(K'))} \to \eta^2_{K'}(u, w; f)$$

for the subgrid-depth $\frac{1}{\sigma}$ of the test-search spaces tending to ∞. Since $\check{\check{R}}_{\mathcal{T}_s}$ is constructed as a piecewise polynomial approximation for $\check{R}_{\mathcal{T}_s}$, we also have that

$$\|\check{\check{R}}_{\mathcal{T}_s(K')}\|^2_{H(\mathbf{b};\mathcal{T}_s(K'))} \to \|\check{R}_{\mathcal{T}_s(K')}\|^2_{H(\mathbf{b};\mathcal{T}_s(K'))} \quad \text{when } \frac{1}{\sigma} \to \infty.$$

As we will see, the norms $\|\cdot\|_{H(\mathbf{b};K)}$ and $\|\!|\cdot|\!\|_{H(\mathbf{b};K)}$ on $H(\mathbf{b}; K)$ are not only equivalent but even converge to each other when $\frac{1}{\sigma} \to \infty$, which will yield

$$\|\check{R}_{\mathcal{T}_s(K')}\|^2_{H(\mathbf{b};\mathcal{T}_s(K'))} \to \|R_{\mathcal{T}_s(K')}\|^2_{H(\mathbf{b};\mathcal{T}_s(K'))} \quad \text{when } \frac{1}{\sigma} \to \infty.$$

Finally, since R_K^δ is the best approximation to R_K from $\mathcal{P}_{m_v}(K)$, we have that

$$\|R_K - R_K^\delta\|_{H(\mathbf{b};K)} \le \|R_K - \check{R}_K\|_{H(\mathbf{b};K)} \to 0 \quad \text{when } \frac{1}{\sigma} \to \infty.$$

The details of this roadmap are as follows.

Proposition 5.4. *For $v \in H(\mathbf{b};K)$, we have*

$$\left| \|v\|_{H(\mathbf{b};K)}^2 - \|\!|v\|\!|_{H(\mathbf{b};K)}^2 \right| \le |\mathbf{b}|^{-1} \operatorname{diam}(K) \|v\|_{H(\mathbf{b};K)}^2.$$

Proof. Without loss of generality, we consider the case that $\frac{\mathbf{b}}{|\mathbf{b}|} = \vec{e}_1$. Given $\check{x} \in K$, let $\mathbf{s}(\mathbf{t})$ denote the projection of $\check{x}$ on ∂K_- (∂K_+) along the x_1-direction. Applying

$$\left| h z(0) - \int_0^h z(x)\, dx \right| = \left| -\int_0^h \int_0^x z'(y)\, dy\, dx \right| \le h \int_0^h |z'(y)|\, dy,$$

we find that

$$\left| r(\mathbf{s}) v(\mathbf{s})^2 - \int_{s_1}^{t_1} v(\mathbf{x})^2\, dx_1 \right| \le \frac{2 r(\mathbf{s})}{|\mathbf{b}|} \int_{s_1}^{t_1} |v(\mathbf{x}) \partial_\mathbf{b} v(\mathbf{x})|\, dx_1.$$

Integrating now this estimate over $x_2, \dots, x_n$, using that $d\mathbf{s} = \frac{|\mathbf{b}|}{|\mathbf{b} \cdot \mathbf{n}_K(\mathbf{s})|} dx_2 \dots dx_n$, and finally applying Cauchy–Schwarz' inequalities confirms the claim. □

As a consequence, lifted and modified lifted residuals become closer with increasing subgrid depth.

Corollary 5.5. *For $((u,w), f) \in \check{\mathbb{U}}_{\mathcal{J}}^\delta \times \mathbb{F}_{\mathcal{J}}^\delta$ and $K \in \mathcal{T}_s$, we have*

$$\|R_K - \check{R}_K\|_{H(\mathbf{b};K)} \lesssim |\mathbf{b}|^{-\frac{1}{2}} \operatorname{diam}(K)^{\frac{1}{2}} \|\check{R}_K\|_{H(\mathbf{b};K)}.$$

Proof. Inside this proof we drop the subscript $H(\mathbf{b};K)$ from the norms and inner products. Note that for any $v \in H(\mathbf{b};K)$, it holds by definition that $\langle\!\langle \check{R}_K, v \rangle\!\rangle = \langle R_K, v \rangle$.

With

$$\tau := \sup_{0 \ne v \in H(\mathbf{b};K)} \left| \frac{\|\!|v\|\!|^2}{\|v\|^2} - 1 \right| \ (\lesssim |\mathbf{b}|^{-1} \operatorname{diam}(K)),$$

we find that

$$\left| \langle R_K, R_K - \check{R}_K \rangle \right| = \left| \|R_K\|^2 - \|\check{R}_K\|^2 \right| \le \tau \|R_K\|^2. \tag{5.6}$$

From

$$\|R_K\|^2 = \sup_{0 \ne v \in H(\mathbf{b};K)} \frac{\langle R_K, v \rangle^2}{\|v\|^2} = \sup_{0 \ne v \in H(\mathbf{b};K)} \frac{\langle\!\langle \check{R}_K, v \rangle\!\rangle^2}{\|v\|^2} = \sup_{0 \ne v \in H(\mathbf{b};K)} \frac{\langle\!\langle \check{R}_K, v \rangle\!\rangle^2}{\|\!|v\|\!|^2} \frac{\|\!|v\|\!|^2}{\|v\|^2},$$

and

$$\sup_{0 \ne v \in H(\mathbf{b};K)} \frac{\langle\!\langle \check{R}_K, v \rangle\!\rangle^2}{\|\!|v\|\!|^2} = \|\!|\check{R}_K\|\!|^2, \quad \frac{\|\!|v\|\!|^2}{\|v\|^2} \in [1-\tau, 1+\tau],$$

we infer that $\left| \|R_K\|^2 - \|\!|\check{R}_K\|\!|^2 \right| \le \tau \|\check{R}_K\|^2$. Now from $\langle \check{R}_K, R_K - \check{R}_K \rangle = \|\check{R}_K\|^2 - \|\!|\check{R}_K\|\!|^2$ and (5.6), we arrive at $\|R_K - \check{R}_K\|^2 \le \tau(\|\!|\check{R}_K\|\!|^2 + \|R_K\|^2) \le \tau(2+\tau)\|\check{R}_K\|^2$, which gives the result. □

Corollary 5.5 is one of the ingredients to prove the mutual closeness of the various error indicators.

Proposition 5.6. *For $(u,w) \in \check{\mathbb{U}}_{\mathcal{J}}^\delta, f \in \mathbb{F}_{\mathcal{J}}^\delta, K' \in \mathcal{T}$, we have*

$$\sum_{K \in \mathcal{T}_s(K')} \|R_K^\delta - (\partial_\mathbf{b} w + cu - f)\|_{L_2(K)}^2 + \|\partial_\mathbf{b} R_K^\delta - (w-u)\|_{L_2(K)}^2 \lesssim \sigma^2 \eta_{K'}^2(u,w;f), \tag{5.7}$$

and

$$\left| \|R_{\mathcal{T}_s(K')}^\delta(u,w;f)\|_{H(\mathbf{b};\mathcal{T}_s(K'))}^2 - \eta_{K'}^2(u,w;f) \right| \lesssim \sigma\, \eta_{K'}^2(u,w;f),$$

only dependent on the involved polynomial degrees, and on (upper bounds for) $|\mathbf{b}|^{-1}$, $\|c\|_{L_\infty(K')}$ and ϱ.

Proof. Applications of the triangle inequality show that

$$\sum_{K \in \mathcal{T}_s(K')} \|R_K^\delta - (\partial_{\mathbf{b}} w + cu - f)\|_{L_2(K)}^2 + \|\partial_{\mathbf{b}} R_K^\delta - (w - u)\|_{L_2(K)}^2$$

$$\leq 2\|R_{\mathcal{T}_s(K')}^\delta - \breve{R}_{\mathcal{T}_s(K')}\|_{H(\mathbf{b};\mathcal{T}_s(K'))}^2 + 2\sum_{K \in \mathcal{T}_s(K')} \|\breve{R}_K - (\partial_{\mathbf{b}} w + cu - f)\|_{L_2(K)}^2 + \|\partial_{\mathbf{b}} \breve{R}_K - (w - u)\|_{L_2(K)}^2. \quad (5.8)$$

We estimate now the terms on the right-hand side of (5.8). For each $K \in \mathcal{T}_s(K')$, from $\breve{R}_K \in \mathcal{P}_{m_\nu}(K)$ and R_K^δ being the $H(\mathbf{b}, K)$-orthogonal projection of R_K onto $\mathcal{P}_{m_\nu}(K)$, we have

$$\|R_K^\delta - \breve{R}_K\|_{H(\mathbf{b};K)} \leq \|R_K^\delta - R_K\|_{H(\mathbf{b};K)} + \|R_K - \breve{R}_K\|_{H(\mathbf{b};K)}$$

$$\leq 2\|R_K - \breve{R}_K\|_{H(\mathbf{b};K)}$$

$$\leq 2\|R_K - \breve{R}_K\|_{H(\mathbf{b};K)} + 2\|\breve{R}_K - \breve{R}_K\|_{H(\mathbf{b};K)},$$

which yields

$$\|R_{\mathcal{T}_s(K')}^\delta - \breve{R}_{\mathcal{T}_s(K')}\|_{H(\mathbf{b};\mathcal{T}_s(K'))}^2 \leq 8\|R_{\mathcal{T}_s(K')} - \breve{R}_{\mathcal{T}_s(K')}\|_{H(\mathbf{b};\mathcal{T}_s(K'))}^2 + 8\|\breve{R}_{\mathcal{T}_s(K')} - \breve{R}_{\mathcal{T}_s(K')}\|_{H(\mathbf{b};\mathcal{T}_s(K'))}^2. \quad (5.9)$$

Using that for $K \in \mathcal{T}_s(K')$, $\mathrm{diam}(K) \leq \sigma\,\mathrm{diam}(K') \leq \sigma^2$, an application of Corollary 5.5 shows that

$$\|R_{\mathcal{T}_s(K')} - \breve{R}_{\mathcal{T}_s(K')}\|_{H(\mathbf{b};\mathcal{T}_s(K'))}^2 \lesssim |\mathbf{b}|^{-1}\sigma^2\|\breve{R}_{\mathcal{T}_s(K')}\|_{H(\mathbf{b};\mathcal{T}_s(K'))}^2. \quad (5.10)$$

Lemma 4.7 shows that for $K \in \mathcal{T}_s(K')$,

$$\|\breve{R}_K - \breve{R}_K\|_{H(\mathbf{b};K)} \lesssim \mathrm{diam}(K)(\|u - w\|_{H^1(K)} + \|\partial_{\mathbf{b}} w + cw - f\|_{H^1(K)}),$$

dependent on (upper bounds for) ϱ, $|\mathbf{b}|^{-1}$, and $\|c\|_{L_\infty(K')}$. Squaring, summing over $K \subset K'$, and using inverse inequalities yields

$$\|\breve{R}_{\mathcal{T}_s(K')} - \breve{R}_{\mathcal{T}_s(K')}\|_{H(\mathbf{b};\mathcal{T}_s(K'))}^2 \lesssim \sigma^2\eta_{K'}^2(u, w; f). \quad (5.11)$$

It remains to estimate the terms in the sum in the right-hand side of (5.8). For each $K \in \mathcal{T}_s(K')$, we have

$$\|\breve{R}_K - (\partial_{\mathbf{b}} w + cu - f)\|_{L_2(K)} \lesssim |\mathbf{b}|^{-1}\,\mathrm{diam}(K)\{\|w - u\|_{L_2(K)} + |\mathbf{b}|^{-1}\,\mathrm{diam}(K)\|\partial_{\mathbf{b}}(w - u)\|_{L_2(K)}\}$$

$$+ |\mathbf{b}|^{-1}\,\mathrm{diam}(K)\|\partial_{\mathbf{b}}(\partial_{\mathbf{b}} w + cu - f)\|_{L_2(K)},$$

by applications of Poincaré's inequality in the streamline direction (cf. the second paragraph in the proof of Lemma 4.11). Similarly

$$\|\partial_{\mathbf{b}} \breve{R}_K - (w - u)\|_{L_2(K)} \lesssim |\mathbf{b}|^{-1}\,\mathrm{diam}(K)\|\partial_{\mathbf{b}}(w - u)\|_{L_2(K)}.$$

Squaring and summing over $K \in \mathcal{T}_s(K')$, and using inverse estimates yields

$$\sum_{K \in \mathcal{T}_s(K')} \|\breve{R}_K - (\partial_{\mathbf{b}} w + cu - f)\|_{L_2(K)}^2 + \|\partial_{\mathbf{b}} \breve{R}_K - (w - u)\|_{L_2(K)}^2 \lesssim \sigma^2\eta_{K'}^2(u, w; f), \quad (5.12)$$

only dependent on (upper bounds for) ϱ, $|\mathbf{b}|^{-1}$ and the involved polynomial degrees.

By combining (5.8)–(5.12), one infers (5.7).

Now using that for vectors $\vec{a}, \vec{b}$,

$$\left| \|\vec{a}\|^2 - \|\vec{b}\|^2 \right| \leq \|\vec{a} - \vec{b}\|\|\vec{a} + \vec{b}\| \leq \|\vec{a} - \vec{b}\|(2\|\vec{b}\| + \|\vec{a} - \vec{b}\|) \leq \|\vec{b}\|^2 \frac{\|\vec{a} - \vec{b}\|}{\|\vec{b}\|}\left(2 + \frac{\|\vec{a} - \vec{b}\|}{\|\vec{b}\|}\right)$$

and, when $\vec{a}$ is of the form $(\|f_i\|)_i$ and $\vec{b} = (\|g_i\|)_i$, furthermore

$$\frac{\|\vec{a} - \vec{b}\|}{\|\vec{b}\|} \leq \frac{\sqrt{\sum_i \|f_i - g_i\|^2}}{\sqrt{\sum_i \|g_i\|^2}},$$

from (5.7) we conclude that

$$\left| \sum_{K \subset K'} \|R_K^\delta\|_{L_2(K)}^2 + \|\partial_{\mathbf{b}} R_K^\delta\|_{L_2(K)}^2 - \sum_{K \subset K'} \|\partial_{\mathbf{b}} w + cu - f\|_{L_2(K)}^2 + \|w - u\|_{L_2(K)}^2 \right|$$

$$\lesssim \sigma^2 \sum_{K \subset K'} \|\partial_{\mathbf{b}} w + cu - f\|_{L_2(K)}^2 + \|w - u\|_{L_2(K)}^2.$$

which, in compact notation, is the second statement to be proven. $\qquad\square$

5.3 A Companion Mesh-Independent Least Squares Formulation of the Transport Problem

For $(u, w) \in \breve{U}$, it holds that

$$
\begin{aligned}
\eta_\Omega^2(u, w; 0) &= \|\partial_\mathbf{b} w + cu\|_{L_2(\Omega)}^2 + \|u - w\|_{L_2(\Omega)}^2 \gtrsim \|\partial_\mathbf{b} w + cw\|_{L_2(\Omega)}^2 + \|u - w\|_{L_2(\Omega)}^2 \\
&\gtrsim \|\mathcal{B}^{-1}\|_{\mathcal{L}(H_{0,\Gamma_-}(\mathbf{b};\Omega), L_2(\Omega))}^{-2} \|w\|_{H(\mathbf{b};\Omega)}^2 + \|u - w\|_{L_2(\Omega)}^2 \\
&\gtrsim \|(u, w)\|_{\breve{U}}^2 \gtrsim \eta_\Omega^2(u, w; 0).
\end{aligned}
$$

Therefore, for $f \in L_2(\Omega)$ and any closed subspace of $\breve{U}$, the problem of minimizing $\eta_\Omega^2(\cdot, \cdot; f)$ over that subspace is well posed.

Proposition 5.7. *For $\mathcal{T} \in \mathbb{T}$, let*

$$
(\bar{u}_\mathcal{T}^\delta, \bar{w}_\mathcal{T}^\delta) := \underset{(u,w) \in \breve{U}_\mathcal{T}^\delta}{\arg\min} \, \eta_\Omega^2(u, w; f). \tag{5.13}
$$

Then, for σ small enough, it holds that

$$
\|(u_\mathcal{T}^\delta, w_\mathcal{T}^\delta) - (\bar{u}_\mathcal{T}^\delta, \bar{w}_\mathcal{T}^\delta)\|_{\breve{U}}^2 \lesssim \sigma \|(u^{\text{ex}}, u^{\text{ex}}) - (u_\mathcal{T}^\delta, w_\mathcal{T}^\delta)\|_{\breve{U}}^2,
$$

where $(u_\mathcal{T}^\delta, w_\mathcal{T}^\delta) \in \breve{U}_\mathcal{T}^\delta$ is the Petrov–Galerkin solution of (3.7)

Proof. "Galerkin orthogonality" shows that for any $(u, w) \in \breve{U}_\mathcal{T}^\delta$,

$$
\eta_\Omega^2(u, w; f) - \eta_\Omega^2(\bar{u}_\mathcal{T}^\delta, \bar{w}_\mathcal{T}^\delta; f) = \eta^2(u - \bar{u}_\mathcal{T}^\delta, w - \bar{w}_\mathcal{T}^\delta; 0) \approx \|(u, w) - (\bar{u}_\mathcal{T}^\delta, \bar{w}_\mathcal{T}^\delta)\|_{\breve{U}}^2. \tag{5.14}
$$

Since $(u_\mathcal{T}^\delta, w_\mathcal{T}^\delta)$ minimizes $\|R_{\mathcal{T}_s}^\delta(u, w; f)\|_{H(\mathbf{b};\mathcal{T}_s)}^2$ over $(u, w) \in \breve{U}_\mathcal{T}^\delta$, two applications of Proposition 5.6 show that for some $|\xi_1|, |\xi_2| \lesssim \sigma$,

$$
(1 + \xi_1)\eta_\Omega^2(u_\mathcal{T}^\delta, w_\mathcal{T}^\delta; f) = \|R_{\mathcal{T}_s}^\delta(u_\mathcal{T}^\delta, w_\mathcal{T}^\delta; f)\|_{H(\mathbf{b};\mathcal{T}_s)}^2 \leq \|R_{\mathcal{T}_s}^\delta(\bar{u}_\mathcal{T}^\delta, \bar{w}_\mathcal{T}^\delta; f)\|_{H(\mathbf{b};\mathcal{T}_s)}^2 = (1 + \xi_2)\eta_\Omega^2(\bar{u}_\mathcal{T}^\delta, \bar{w}_\mathcal{T}^\delta; f),
$$

which, together with (5.14), shows that for σ small enough,

$$
\|(u_\mathcal{T}^\delta, w_\mathcal{T}^\delta) - (\bar{u}_\mathcal{T}^\delta, \bar{w}_\mathcal{T}^\delta)\|_{\breve{U}}^2 \lesssim \sigma \eta_\Omega^2(u_\mathcal{T}^\delta, w_\mathcal{T}^\delta; f) \approx \sigma \|(u^{\text{ex}}, u^{\text{ex}}) - (u_\mathcal{T}^\delta, w_\mathcal{T}^\delta)\|_{\breve{U}}^2. \qquad \square
$$

In complete analogy we can define effectivity of a mark-and-refine strategy for the least squares scheme (5.13) based on a bulk criterion for the quantities η_K, denoting the collection of correspondingly marked cells by $\tilde{\mathcal{M}} = \tilde{\mathcal{M}}((\bar{u}_\mathcal{T}^\delta, \bar{w}_\mathcal{T}^\delta), \vartheta)$.

Proposition 5.8. *For sufficiently small σ, (r, v)-effectivity of the above refinement strategy for the DPG scheme is equivalent to (r, v)-effectivity of the analogous strategy with the same ϑ for the least squares estimator.*

Proof. Using Proposition 5.7, stability of both estimators shows that for any $\mathcal{M} \subset \mathcal{T}$,

$$
\left.
\begin{aligned}
\left| \|R_{\mathcal{T}_s(\mathcal{M})}^\delta(\bar{u}_\mathcal{T}^\delta, \bar{w}_\mathcal{T}^\delta; f)\|_{H(\mathbf{b};\mathcal{T}_s(\mathcal{M}))} - R_{\mathcal{T}_s(\mathcal{M})}^\delta(u_\mathcal{T}^\delta, w_\mathcal{T}^\delta; f)\|_{H(\mathbf{b};\mathcal{T}_s(\mathcal{M}))} \right| \\
\left| \eta_\mathcal{M}(\bar{u}_\mathcal{T}^\delta, \bar{w}_\mathcal{T}^\delta; f) - \eta_\mathcal{M}(u_\mathcal{T}^\delta, w_\mathcal{T}^\delta; f) \right|
\end{aligned}
\right\} \lesssim \sqrt{\sigma} \eta_\Omega(u_\mathcal{T}^\delta, w_\mathcal{T}^\delta; f).
$$

Now let $\mathcal{M} \subset \mathcal{T}$ be such that $\|R_{\mathcal{T}_s(\mathcal{M})}^\delta(u_\mathcal{T}^\delta, w_\mathcal{T}^\delta; f)\|_{H(\mathbf{b};\mathcal{T}_s(\mathcal{M}))} \geq \vartheta \|R_{\mathcal{T}_s}^\delta(u_\mathcal{T}^\delta, w_\mathcal{T}^\delta; f)\|_{H(\mathbf{b};\mathcal{T}_s)}$. Then elementary operations using Propositions 5.6 and 5.7 show the existence of a $|\xi| \lesssim \sqrt{\sigma}$, and thus for σ small enough, $|\xi| \leq \frac{1}{2}$, with $\eta_\mathcal{M}(\bar{u}_\mathcal{T}^\delta, \bar{w}_\mathcal{T}^\delta; f) \geq \vartheta(1 + \xi)\eta_\Omega(\bar{u}_\mathcal{T}^\delta, \bar{w}_\mathcal{T}^\delta; f)$. Now, *if* the latter implies that for some $v = v(\vartheta) < 1$, and with the refined mesh $\tilde{\mathcal{T}} = \tilde{\mathcal{T}}(\mathcal{T}, \mathcal{M}, r)$ from Definition 5.1, it holds that $\eta_\Omega(\bar{u}_{\tilde{\mathcal{T}}}^\delta, \bar{w}_{\tilde{\mathcal{T}}}^\delta; f) \leq v \eta_\Omega(\bar{u}_\mathcal{T}^\delta, \bar{w}_\mathcal{T}^\delta; f)$, *then* we have that for some $|\xi_1|, |\xi_2|, |\xi_3| \leq \sigma$,

$$
\begin{aligned}
\|R_{\tilde{\mathcal{T}}_s}^\delta(u_\mathcal{T}^\delta, w_\mathcal{T}^\delta; f)\|_{H(\mathbf{b};\tilde{\mathcal{T}}_s)} &\leq \|R_{\tilde{\mathcal{T}}_s}^\delta(\bar{u}_\mathcal{T}^\delta, \bar{w}_\mathcal{T}^\delta; f)\|_{H(\mathbf{b};\tilde{\mathcal{T}}_s)} = \eta_\Omega(\bar{u}_{\tilde{\mathcal{T}}}^\delta, \bar{w}_{\tilde{\mathcal{T}}}^\delta; f)(1 + \xi_1) \\
&\leq v \eta_\Omega(\bar{u}_\mathcal{T}^\delta, \bar{w}_\mathcal{T}^\delta; f)(1 + \xi_1) = v \eta_\Omega(u_\mathcal{T}^\delta, w_\mathcal{T}^\delta; f)(1 + \xi_1)(1 + \sqrt{\xi_2}) \\
&= v \|R_{\mathcal{T}_s}^\delta(u_\mathcal{T}^\delta, w_\mathcal{T}^\delta; f)\|_{H(\mathbf{b};\mathcal{T}_s)}(1 + \xi_1)(1 + \sqrt{\xi_2})(1 + \xi_3),
\end{aligned}
$$

showing for σ small enough the result of Theorem 5.3.

Applying the above arguments with interchanged roles of $\|R_{\mathcal{T}_s}^\delta(\cdot, \cdot; f)\|_{H(\mathbf{b};\mathcal{T}_s)}$ and $\eta_\Omega(\cdot, f)$ and choosing σ small enough, the claim of Remark 5.8 follows. $\square$

In view of Proposition 5.8, the proof of Theorem 5.3 is complete once we establish the following equivalent result.

Theorem 5.9. *Let* **b** *and c be as in Theorem 5.3, and, to control* $\sup_{K\in\mathcal{T}\in\mathbb{T}}\mathrm{diam}(K)$*, let σ be sufficiently small. Then, for all* $\vartheta \in (0,1]$*, there exist* $r \in \mathbb{N}$*,* $\nu = \nu(\vartheta) < 1$ *with the following property: whenever for* $\mathcal{T} \in \mathbb{T}$ *and* $f \in \mathbb{F}_\mathcal{T}^\delta$*, the set of marked elements* $\bar{\mathcal{M}} = \bar{\mathcal{M}}((\bar{u}_\mathcal{T}^\delta, \bar{w}_\mathcal{T}^\delta), \vartheta) \subseteq \mathcal{T}$ *is such that*

$$\eta_{\bar{\mathcal{M}}}(\bar{u}_\mathcal{T}^\delta, \bar{w}_\mathcal{T}^\delta; f) \geq \vartheta \eta_\Omega(\bar{u}_\mathcal{T}^\delta, \bar{w}_\mathcal{T}^\delta; f), \tag{5.15}$$

then for the refinement $\tilde{\mathcal{T}} = \tilde{\mathcal{T}}(\mathcal{T}, \bar{\mathcal{M}}, r)$ *according to Definition 5.1, it follows that*

$$\eta_\Omega(\bar{u}_{\tilde{\mathcal{T}}}^\delta, \bar{w}_{\tilde{\mathcal{T}}}^\delta; f) \leq \nu \eta_\Omega(\bar{u}_\mathcal{T}^\delta, \bar{w}_\mathcal{T}^\delta; f). \tag{5.16}$$

The remainder of this section is devoted to the proof of Theorem 5.9. We are going to show that for some constants $\vartheta' > 0$ and $\nu' < 1$, thus independent of $\mathcal{T}$ (subject to σ being sufficiently small), for $\bar{\mathcal{M}}$ as in (5.15) there exists an $\bar{\bar{\mathcal{M}}} \subset \bar{\mathcal{M}}$ with

$$\eta_{\bar{\bar{\mathcal{M}}}}(\bar{u}_\mathcal{T}^\delta, \bar{w}_\mathcal{T}^\delta; f) \geq \vartheta' \eta_{\bar{\mathcal{M}}}(\bar{u}_\mathcal{T}^\delta, \bar{w}_\mathcal{T}^\delta; f), \tag{5.17}$$

and that for any $K' \in \bar{\bar{\mathcal{M}}}$,

$$\inf_{\{(u,w)\in\check{\mathbb{U}}_\mathcal{T}^\delta\,:\,\mathrm{supp}\,u,\,\mathrm{supp}\,w\subset K'\}} \eta_{K'}(\bar{u}_\mathcal{T}^\delta - u, \bar{w}_\mathcal{T}^\delta - w; f) \leq \nu' \eta_{K'}(\bar{u}_\mathcal{T}^\delta, \bar{w}_\mathcal{T}^\delta; f). \tag{5.18}$$

In other words, for the cells in $\bar{\bar{\mathcal{M}}}$ one can correct the current approximation *cell-wise* to reduce the corresponding error indicator. An elementary calculation shows that then these two properties imply (5.16) with constant $\nu := \sqrt{(\vartheta\vartheta')^2(\nu')^2 + 1 - (\vartheta\vartheta')^2} < 1$.

5.3.1 Reduction of the Local Mesh-Independent Error Indicator

In this subsection we work towards the verification of (5.18) for those $K' \in \bar{\mathcal{M}}$ that satisfy certain conditions. Then in the following two subsections, for two possible scenarios we will construct subsets $\bar{\bar{\mathcal{M}}} \subset \bar{\mathcal{M}}$ of K' that satisfy these conditions, and for which (5.17) is satisfied. This will then prove Theorem 5.9 and hence Theorem 5.3.

We recall that the reaction coefficient c is assumed to be a non-negative constant over each $K' \in \mathcal{T}$. We introduce the shorthand notations

$$g := \partial_\mathbf{b}\bar{w}_\mathcal{T}^\delta + c\bar{u}_\mathcal{T}^\delta - f, \quad e := \bar{u}_\mathcal{T}^\delta - \bar{w}_\mathcal{T}^\delta, \tag{5.19}$$

so that

$$\eta_{K'}^2(\bar{u}_\mathcal{T}^\delta - u, \bar{w}_\mathcal{T}^\delta - w; f) = \|e - (u-w)\|_{L_2(K')}^2 + \|g - (\partial_\mathbf{b}w + cu)\|_{L_2(K')}^2.$$

Fixing $\beta \in (0, \frac{1}{4})$, we refer to the $K' \in \bar{\mathcal{M}}$ for which

$$\frac{\|e + cg\|_{L_2(K')}^2}{\|g\|_{L_2(K')}^2 + \|e\|_{L_2(K')}^2} \geq \beta \quad \text{(Type-(I))},$$

as Type-(I) and for the remaining ones as Type-(II). Accordingly, we decompose $\bar{\mathcal{M}}$ into the Type-(I) and Type-(II) elements writing $\bar{\mathcal{M}} = \bar{\mathcal{M}}_\mathrm{I} \dot\cup \bar{\mathcal{M}}_\mathrm{II}$.

Type-(I) Elements. We start with showing that for $K' \in \bar{\mathcal{M}}_\mathrm{I}$, (5.18) can be already established by a correction of the u-component.

Lemma 5.10. *Assume that* $\|\cdot\|$ *is induced by the inner product* $\langle\cdot,\cdot\rangle$ *of some Hilbert space H and let* $g, e \in H$ *be arbitrary but fixed. For any scalar c and* $u \in H$ *let*

$$Q(u) := \|e - u\|^2 + \|g - cu\|^2.$$

Then

$$u_{\min} := \arg\min_{u \in H} Q(u) = \frac{e + cg}{1 + c^2},$$

$$Q(u) - Q(u_{\min}) = (1 + c^2)\|u - u_{\min}\|_H^2,$$

$$\|u_{\min}\|^2 \le \frac{Q(0)}{1 + c^2},$$

$$Q(u_{\min}) = \left(1 - \frac{\|e + cg\|^2}{(1 + c^2)(\|g\|^2 + \|e\|^2)}\right)Q(0).$$

Proof. The first two statements follow from

$$Q(u + h) - Q(u) = 2\langle h, (c^2 + 1)u - (e + cg)\rangle + (1 + c^2)\|h\|^2.$$

The third statement is a consequence of

$$\|u_{\min}\| = \|\frac{e + cg}{1 + c^2}\| \le \frac{1}{1 + c^2}\|e\| + \frac{c}{1 + c^2}\|g\| \le \frac{\sqrt{\|e\|^2 + \|g\|^2}}{\sqrt{1 + c^2}} = \frac{Q(0)^{\frac{1}{2}}}{\sqrt{1 + c^2}}.$$

The last statement follows from

$$\frac{Q(0) - Q(u_{\min})}{Q(0)} = \frac{(1 + c^2)\|\frac{e+cg}{1+c^2}\|^2}{\|g\|^2 + \|e\|^2}. \qquad \square$$

Corollary 5.11. *For r sufficiently large, only dependent on the polynomial degrees m_u, m_w and m_f, and on an upper bound for $|c_{K'}|$, for all $K' \in \bar{\mathcal{M}}_I$ it holds that*

$$\inf_{\{(u,0) \in \breve{\mathbb{U}}_{\bar{\mathcal{J}}}^\delta \,:\, \mathrm{supp}\, u \subset K'\}} \eta_{K'}^2(\bar{u}_{\mathcal{J}}^\delta - u, \bar{w}_{\mathcal{J}}^\delta; f) \le \left(1 - \frac{\beta}{2(1 + c_{K'}^2)}\right)\eta_{K'}^2(\bar{u}_{\mathcal{J}}^\delta, \bar{w}_{\mathcal{J}}^\delta; f).$$

Proof. Lemma 5.10 says that $u_{\min} = \frac{e + c_{K'}g}{1 + c_{K'}^2}$ minimizes $Q(u) := \eta_{K'}^2(\bar{u}_{\mathcal{J}}^\delta - u, \bar{w}_{\mathcal{J}}^\delta; f)$ over $L_2(K')$, and that

$$Q(u_{\min}) \le \left(1 - \frac{\beta}{1 + c_{K'}^2}\right)Q(0).$$

The function $u_{\min}$ is a polynomial on K' and can therefore be approximated with relative accuracy $\sqrt{\beta/(2(1 + c_{K'}^2))}$ by a piecewise polynomial $\tilde{u}$ on a sufficiently refined mesh. This follows from the usual combination of direct and inverse estimates. The proof is completed by

$$Q(\tilde{u}) - Q(u_{\min}) = (1 + c_{K'}^2)\|\tilde{u} - u_{\min}\|_{L_2(K')}^2 \le \frac{\beta}{2\|u_{\min}\|_{L_2(K')}^2} \le \frac{\beta}{2(1 + c_{K'}^2)}Q(0)$$

by applications of the statements from Lemma 5.10. $\qquad \square$

Type-(II) Elements. It remains to discuss $K' \in \bar{\mathcal{M}}_{II}$. For those elements we need to find suitable corrections for the component $\bar{w}_{\mathcal{J}}^\delta$ – in brief the w-component.

We will search for a $(0, w) \in \breve{\mathbb{U}}_{\bar{\mathcal{J}}}^\delta$ with supp $w \subset K'$ such that

$$\|g - \partial_{\mathbf{b}} w\|_{L_2(K')}^2 < \|g\|_{L_2(K')}^2.$$

In order to show that this reduction is not lost by a similar increase by the replacement of $\|e\|_{L_2(K')}^2$ by $\|e - w\|_{L_2(K')}^2$, we will make use of the fact that for $K' \in \bar{\mathcal{M}}_{II}$, the term $\|e\|_{L_2(K')}$ is controlled by a multiple of $\|g\|_{L_2(K')}$ depending only on $\|c\|_{L_\infty(\Omega)}$:

Lemma 5.12. *For $K' \in \bar{\mathcal{M}}_{II}$, it holds that*

$$\omega_{K'} = \omega_{K'}(e, g) := \frac{\|e\|_{L_2(K')}}{\|g\|_{L_2(K')}} < 2|c_{K'}| + 1, \tag{5.20}$$

and thus

$$\eta_{K'}^2(\bar{u}_{\mathcal{J}}^\delta, \bar{w}_{\mathcal{J}}^\delta; f)^2 \le ((2|c_{K'}| + 1)^2 + 1)\|g\|_{L_2(K')}^2.$$

Proof. Recall that $K' \in \bar{\mathcal{M}}_{\mathrm{II}}$ means that

$$\frac{\|e + cg\|^2_{L_2(K')}}{\|g\|^2_{L_2(K')} + \|e\|^2_{L_2(K')}} < \beta, \tag{5.21}$$

so that in particular $g \neq 0$. Substituting $\|e\|_{L_2(K')} = \omega_{K'} \|g\|_{L_2(K')}$, (5.21) implies

$$\left| \omega_{K'} - |c_{K'}| \right| < \sqrt{\beta(1 + \omega_{K'}{}^2)},$$

which gives

$$\omega_{K'} < |c_{K'}| + \sqrt{\beta} + \sqrt{\beta}\,\omega_{K'} < |c_{K'}| + \frac{1}{2} + \frac{1}{2}\omega_{K'}$$

by our assumption that $\beta < \frac{1}{4}$. This confirms the first and so the second claim. □

Our argument for handling Type-(II) elements requires the following further preparations. For every $\mathbf{s} \in \partial K'_-$ let as before $r(\mathbf{s})$ denote length of the line segment emanating from $\mathbf{s} \in \partial K'_-$ and ending in $\partial K'_+$. One observes then that a function Q on K' can be written as

$$Q = \partial_{\mathbf{b}} z, \quad z|_{\partial K'_- \cup \partial K'_+} = 0, \tag{5.22}$$

if and only if each of its *line averages in direction* $\mathbf{b}^\circ := \frac{\mathbf{b}}{|\mathbf{b}|}$ vanishes, i.e.,

$$A_{\mathbf{s}}(Q) := r(\mathbf{s})^{-1} \int_0^{r(\mathbf{s})} Q(\mathbf{s} + t\mathbf{b}^\circ)\, dt = 0, \quad \mathbf{s} \in \partial K_-.$$

In fact, then

$$z(\mathbf{s} + t\mathbf{b}^\circ) := |\mathbf{b}|^{-1} \int_0^t Q(\mathbf{s} + t'\mathbf{b}^\circ)\, dt'$$

satisfies (5.22).

For g as in (5.19), the function $G = G(g)$, defined on each $K' \in \mathcal{T}$ by

$$G(\mathbf{x}) = A_{\mathbf{s}}(g) \quad \text{for } \mathbf{x} = \mathbf{s} + t\mathbf{b}^\circ, \ \mathbf{s} \in \partial K_-, \ t \in [0, r(s)], \tag{5.23}$$

is obviously constant along $\mathbf{b}$ and

$$A_{\mathbf{s}}(g - G) = 0 \quad \text{for every } \mathbf{s} \in \partial K'_-.$$

Hence, for z_g, defined by

$$z_g(\mathbf{s} + t\mathbf{b}^\circ) := |\mathbf{b}|^{-1} \int_0^t (g - G)(\mathbf{s} + t'\mathbf{b}^\circ)\, dt' \quad \text{for } t \in [0, r(s)],$$

we have

$$g - \partial_{\mathbf{b}} z_g = G.$$

Thanks to $g - G \perp_{L_2(K')} G$, we have

$$\|G\|^2_{L_2(K')} = \|g\|^2_{L_2(K')} - \|g - G\|^2_{L_2(K')},$$

and so in particular $\|G\|_{L_2(K')} \leq \|g\|_{L_2(K')}$.

Under the condition that $\|G\|_{L_2(K')} < \|g\|_{L_2(K')}$, one infers from $\|z_g\|_{L_2(K')} \lesssim |\mathbf{b}|^{-1} \mathrm{diam}\, K' \|g - G\|_{L_2(K')}$ by Poincaré's inequality, in combination with (5.20) that for $\mathrm{diam}\, K'$ being sufficiently small,

$$\eta^2_\Omega(\bar{u}^\delta_{\mathcal{T}}, \bar{w}^\delta_{\mathcal{T}} - z_g; f) < \eta^2_\Omega(\bar{u}^\delta_{\mathcal{T}}, \bar{w}^\delta_{\mathcal{T}}; f).$$

When proceeding to the natural next step to approximate z_g with functions of type $(0, w) \in \breve{\mathbb{U}}^\delta_{\mathcal{T}}$ with supp $w \subset K'$, a difficulty is that z_g is continuous piecewise polynomial with respect to a partition of $\breve{K}'$ into

subsimplices that can have arbitrarily bad aspects ratios. To tackle this problem, we first approximate z_g by an "isotropic" function $\tilde{z}_g$ for which $\|g - \partial_{\mathbf{b}}\tilde{z}_g\|_{L_2(K')}$ is at most slightly larger than $\|g - \partial_{\mathbf{b}}z_g\|_{L_2(K')}$:

Lemma 5.13. *Let*

$$\alpha_{K'} = \alpha_{K'}(g) := \frac{\|G\|_{L_2(K')}}{\|g\|_{L_2(K')}} < 1.$$

Then there exists a $\tilde{z}_g \in H_0^1(K') \cap H^s(K')$ such that for any $s < \frac{3}{2}$,

$$|\tilde{z}_g|_{H^s(K')} \lesssim (\operatorname{diam} K')^{-s}\|\tilde{z}_g\|_{L_2(K')} \tag{5.24}$$

(depending on upperbounds for $\alpha_{K'}$ and $\varrho_{K'}$), and

$$\|g - \partial_{\mathbf{b}}\tilde{z}_g\|_{L_2(K')} \leq \frac{1 + \alpha_{K'}}{2}\|g\|_{L_2(K')}.$$

Proof. For $n = 1$, $\tilde{z}_g = z_g$ satisfies the conditions. Now let $n > 1$. Let $\rho \in C^\infty$ with $0 \leq \rho \leq 1$, $\rho(x) = 0$ for $x \leq \frac{1}{2}$, and $\rho(x) = 1$ for $x \geq 1$, and let $\rho_\eta(x) := \rho(x/\eta)$.

We are going to construct a modification of z_g that is zero on subsimplices that have very bad aspect ratios. With $F_1, \ldots, F_{n+1}$ denoting the faces of K', for $1 \leq i \leq n + 1$ let $\mathbf{d}_{F_i}$ be the orthogonal projection of the inward pointing normal to F_i onto the plane $\mathbf{b}^\perp$. For each i, we choose a Cartesian coordinate system $\mathbf{y}^{(i)} = T^{(i)}\mathbf{x} + \mathbf{z}^{(i)}$ such that the first coordinate direction is $\frac{\mathbf{d}_{F_i}}{|\mathbf{d}_{F_i}|}$, the origin equals one of the vertices of F_i, and all other vertices of F_i have a non-negative first component. Now for some $\varepsilon > 0$, we define $\tilde{z}_g$ by

$$\partial_{\mathbf{b}}\tilde{z}_g = (g - G)\prod_{i=1}^{n+1}\rho_{\varepsilon\operatorname{diam}K'}((T^{(i)}\cdot + \mathbf{z}^{(i)})_1), \quad \tilde{z}_g|_{\partial K' \cup \partial K'_+} = 0.$$

Since $(T^{(i)}(\cdot + t\mathbf{b}))_1 = (T^{(i)}(\cdot))_1$, $A_{\mathbf{s}}(\partial_{\mathbf{b}}\tilde{z}_g) = 0$ and the function $\tilde{z}_g$ is well defined.

Since $\mathbf{x} \mapsto \prod_{i=1}^{n+1}\rho_{\varepsilon\operatorname{diam}K'}((T^{(i)}\mathbf{x})_1)$ vanishes on all subsimplices that have very bad aspect ratios (relative to ε) in the partition of K' with respect to which z_g is a continuous piecewise polynomial, a homogeneity argument shows that $\tilde{z}_g$ satisfies (5.24), with a constant depending on ε. Moreover, also $\tilde{z}_g$ vanishes on a possible characteristic boundary of K'.

Writing

$$g - \partial_{\mathbf{b}}\tilde{z}_g = G + \left(1 - \prod_{i=1}^{n+1}\rho_{\varepsilon\operatorname{diam}K'}((T^{(i)}\cdot)_1)\right)(g - G),$$

and using that $\|G\|_{L_2(K')} = \alpha_{K'}\|g\|_{L_2(K')}$, and

$$\left\|\left(1 - \prod_{i=1}^{n+1}\rho_{\varepsilon\operatorname{diam}K'}((T^{(i)}\cdot)_1)\right)(g - G)\right\|_{L_2(K')} \leq \left\|\left(1 - \prod_{i=1}^{n+1}\rho_{\varepsilon\operatorname{diam}K'}((T^{(i)}\cdot)_1)\right)\right\|_{L_2(K')}\|g - G\|_{L_\infty(K')}$$

$$\leq \sqrt{\varepsilon|K'|}\,2\|g\|_{L_\infty(K')} \lesssim \sqrt{\varepsilon}\,\|g\|_{L_2(K')},$$

which holds again by a homogeneity argument, the proof is completed by taking ε sufficiently small, dependent on $\alpha_{K'}$. $\square$

Corollary 5.14. *For $K' \in \bar{\mathcal{M}}_{\text{II}}$ let $\alpha_{K'} < 1$. Then for σ sufficiently small, and r sufficiently large, only dependent on upperbounds for m_u, m_w, m_f, ϱ, $|\mathbf{b}|^{-1}$, $\alpha_{K'}$, σ, and $|c_{K'}|$, it holds that*

$$\inf_{\{w\,:\,\operatorname{supp}w\subset K',\,(0,w)\in\breve{\mathbb{U}}_{\tilde{\mathcal{J}}}^\delta\}}\eta_{K'}^2(\bar{u}_{\mathcal{J}}^\delta, \bar{w}_{\mathcal{J}}^\delta - w; f) \leq \left(\frac{1}{2} + \frac{1}{2}\frac{\frac{1+\alpha_{K'}}{2} + |c_{K'}| + 1}{1 + |c_{K'}| + 1}\right)\eta_{K'}^2(\bar{u}_{\mathcal{J}}^\delta, \bar{w}_{\mathcal{J}}^\delta; f).$$

Proof. Let

$$\breve{\sigma} = \breve{\sigma}(r) := \max_{\{K\in\tilde{\mathcal{J}}:K\subset K'\}}\frac{\operatorname{diam}K}{\operatorname{diam}K'}.$$

By taking w with $(0, w) \in \breve{\mathbb{U}}_{\tilde{\mathcal{J}}}^\delta$ to be the Scott–Zhang interpolant of $\tilde{z}_g$ from Lemma 5.13, for $s \in (1, \frac{3}{2})$ we have

$$\|\tilde{z}_g - w\|_{L_2(K')} + |\mathbf{b}|^{-1}\breve{\sigma}\operatorname{diam}K'\|\partial_{\mathbf{b}}(\tilde{z}_g - w)\|_{L_2(K')} \lesssim (\breve{\sigma}\operatorname{diam}K')^s|\tilde{z}_g|_{H^s(K')}$$

$$\lesssim \breve{\sigma}^s\|\tilde{z}_g\|_{L_2(K')} \lesssim \breve{\sigma}^s|\mathbf{b}|^{-1}\operatorname{diam}K'\|\partial_{\mathbf{b}}\tilde{z}_g\|_{L_2(K')}$$

$$\lesssim \breve{\sigma}^s|\mathbf{b}|^{-1}\operatorname{diam}K'\|g\|_{L_2(K')},$$

where we used Poincaré's inequality. We obtain that

$$\|g - \partial_{\mathbf{b}} w\|_{L_2(K')} \leq \|g - \partial_{\mathbf{b}} \tilde{z}_g\|_{L_2(K')} + \|\partial_{\mathbf{b}}(\tilde{z}_g - w)\|_{L_2(K')} \leq \left(\frac{1 + \alpha_{K'}}{2} + \breve{\sigma}^{s-1} \right) \|g\|_{L_2(K')},$$

and

$$\|e + w\|_{L_2(K')} - \|e\|_{L_2(K')} \leq \|\tilde{z}_g\|_{L_2(K')} + \|\tilde{z}_g - w\|_{L_2(K')} \lesssim (|\mathbf{b}|^{-1} \operatorname{diam} K' + \breve{\sigma}^s |\mathbf{b}|^{-1} \operatorname{diam} K') \|g\|_{L_2(K')}.$$

Recalling that $\max_{K' \in \mathcal{T}} \operatorname{diam} K' \leq \sigma$, $\eta_{K'}^2(\bar{u}_{\mathcal{T}}^\delta, \bar{w}_{\mathcal{T}}^\delta; f) = \|g\|_{L_2(K')}^2 + \|e\|_{L_2(K')}^2$, and $\omega_{K'} = \frac{\|e\|_{L_2(K')}}{\|g\|_{L_2(K')}} \leq 2|c_{K'}| + 1$, the assertion follows. $\qquad\square$

In summary, for $K' \in \bar{\mathcal{M}}_{\mathrm{I}}$ completely local u-corrections on refinements of fixed depth suffice to reduce $\eta_{K'}$ by a constant factor $v' < 1$. For $K' \in \bar{\mathcal{M}}_{\mathrm{II}}$ an analogous statement, this time by means of a *local w-correction*, holds *provided* that there exists a constant $\alpha < 1$ such that

$$\alpha_{K'} = \frac{\|G\|_{L_2(K')}}{\|g\|_{L_2(K')}} = \frac{\sqrt{\langle G, g \rangle_{L_2(K')}}}{\|g\|_{L_2(K')}} \leq \alpha. \tag{5.25}$$

5.4 Selection of $\bar{\bar{\mathcal{M}}} \subset \bar{\mathcal{M}}$ That Satisfy Both (5.18) and (5.17)

In case

$$\eta_{\bar{\mathcal{M}}_{\mathrm{II}}}(\bar{u}_{\mathcal{T}}^\delta, \bar{w}_{\mathcal{T}}^\delta; f)^2 < \eta_{\bar{\mathcal{M}}_{\mathrm{I}}}(\bar{u}_{\mathcal{T}}^\delta, \bar{w}_{\mathcal{T}}^\delta; f)^2, \tag{5.26}$$

equation (5.17) is valid with $\bar{\bar{\mathcal{M}}} = \bar{\mathcal{M}}_{\mathrm{I}}$ and $\vartheta' = \frac{1}{2}\sqrt{2}$, whereas (5.18) follows from the reduction of the $\eta_{K'}$ for $K' \in \bar{\mathcal{M}}_{\mathrm{I}}$ by Corollary 5.11. We conclude that Theorem 5.9 is valid for both $n = 1$ and $n > 1$ (even without the additional downwind refinements described in Definition 5.1).

It remains to investigate the case where (5.26) does not hold. It is only for this case that we have to establish (5.25) for sufficiently many $K' \in \bar{\mathcal{M}}_{\mathrm{II}}$. It will require 'global' arguments, already announced in the abstract, that make use of the fact that $(\bar{u}_{\mathcal{T}}^\delta, \bar{w}_{\mathcal{T}}^\delta)$ is the minimizer of $\eta_\Omega^2(u, w; f)$ over $\breve{\mathcal{U}}_{\mathcal{T}}^\delta$.

Lemma 5.15. *Suppose there exists a constant $\alpha < 1$ such that validity of*

$$\eta_{\bar{\mathcal{M}}_{\mathrm{II}}}(\bar{u}_{\mathcal{T}}^\delta, \bar{w}_{\mathcal{T}}^\delta; f)^2 \geq \eta_{\bar{\mathcal{M}}_{\mathrm{I}}}(\bar{u}_{\mathcal{T}}^\delta, \bar{w}_{\mathcal{T}}^\delta; f)^2 \tag{5.27}$$

implies

$$\sum_{K' \in \bar{\mathcal{M}}_{\mathrm{II}}} \|G\|_{L_2(K')}^2 \leq \alpha^2 \sum_{K' \in \bar{\mathcal{M}}_{\mathrm{II}}} \|g\|_{L_2(K')}^2. \tag{5.28}$$

Then Theorem 5.9 is valid.

Proof. In view of the discussion preceding this lemma, it suffices to verify (5.17) and (5.18) for some $\bar{\bar{\mathcal{M}}} \subset \bar{\mathcal{M}}$ for the case that (5.27) holds. By the hypothesis of this lemma, (5.28) is then also valid. We define

$$\bar{\bar{\mathcal{M}}} := \left\{ K' \in \bar{\mathcal{M}}_{\mathrm{II}} : \alpha_{K'} \leq \sqrt{\frac{1 + \alpha^2}{2}} \right\}$$

Then $\bar{\bar{\mathcal{M}}}$ satisfies (5.18) by Corollary 5.14, and it remains to verify that it satisfies (5.17).

Thanks to (5.27), we have

$$\eta_{\bar{\mathcal{M}}}(\bar{u}_{\mathcal{T}}^\delta, \bar{w}_{\mathcal{T}}^\delta; f)^2 \leq 2\eta_{\bar{\mathcal{M}}_{\mathrm{II}}}(\bar{u}_{\mathcal{T}}^\delta, \bar{w}_{\mathcal{T}}^\delta; f)^2,$$

whereas by Lemma 5.12, the right-hand side is bounded by a constant multiple of $\sum_{K' \in \bar{\mathcal{M}}_{\mathrm{II}}} \|g\|_{L_2(K')}^2$. The definition of $\bar{\bar{\mathcal{M}}}$ and (5.28) imply that

$$\sum_{K' \in \bar{\mathcal{M}}_{\mathrm{II}} \backslash \bar{\bar{\mathcal{M}}}} \|g\|_{L_2(K')}^2 < \frac{2}{1 + \alpha^2} \sum_{K' \in \bar{\mathcal{M}}_{\mathrm{II}} \backslash \bar{\bar{\mathcal{M}}}} \|G\|_{L_2(K')}^2 \leq \frac{2\alpha^2}{1 + \alpha^2} \sum_{K' \in \bar{\mathcal{M}}_{\mathrm{II}}} \|g\|_{L_2(K')}^2,$$

or, equivalently,

$$\sum_{K' \in \bar{\mathcal{M}}_{\mathrm{II}}} \|g\|_{L_2(K')}^2 < \frac{1 + \alpha^2}{1 - \alpha^2} \sum_{K' \in \bar{\bar{\mathcal{M}}}} \|g\|_{L_2(K')}^2.$$

The proof of (5.17) follows from $\sum_{K' \in \bar{\bar{\mathcal{M}}}} \|g\|_{L_2(K')}^2 \leq \eta_{\bar{\bar{\mathcal{M}}}}(\bar{u}_{\mathcal{T}}^\delta, \bar{w}_{\mathcal{T}}^\delta; f)^2.$ $\qquad\square$

5.5 Proof of Theorem 5.9 for $n = 1$

By Lemma 5.15 the proof of Theorem 5.9 for $n = 1$, and hence of Theorem 5.3, follows as soon as we have shown that (5.27) implies (5.28). To that end, consider the one-dimensional case $n = 1$, with $\Omega = (0, 1)$, $\mathbf{b} = 1$, and c piecewise constant.

Recalling that
$$g = (\bar{w}_{\mathcal{T}}^\delta)' + c\bar{u}_{\mathcal{T}}^\delta - f, \quad e = \bar{u}_{\mathcal{T}}^\delta - \bar{w}_{\mathcal{T}}^\delta,$$
the definition of $(\bar{u}_{\mathcal{T}}^\delta, \bar{w}_{\mathcal{T}}^\delta)$ as minimizer of $\eta_\Omega^2(\cdot, \cdot; f)$ over $\check{\mathbb{U}}_{\mathcal{T}}^\delta$ shows that
$$\langle u - w, e \rangle_{L_2(\Omega)} + \langle w' + cu, g \rangle_{L_2(\Omega)} = 0 \quad ((u, w) \in \check{\mathbb{U}}_{\mathcal{T}}^\delta),$$
or, equivalently,
$$e + cg \perp_{L_2(K')} \mathcal{P}_{m_u}(K') \quad (K' \in \mathcal{T}), \tag{5.29}$$
and
$$\int_\Omega gw' - we \, d\mathbf{x} = 0 \quad ((0, w) \in \check{\mathbb{U}}_{\mathcal{T}}^\delta). \tag{5.30}$$

Remark 5.16. When $m_u = m_w$, (5.29) says that $e = -cg$ which means that all cells are of Type-(II). In particular, when in addition $c = 0$, we obtain $\bar{u}_{\mathcal{T}}^\delta = \bar{w}_{\mathcal{T}}^\delta$.

For the piecewise constant function
$$F = F(G, \bar{\mathcal{M}}_{\mathrm{II}}) := \begin{cases} G|_{K'} & \text{on } K' \in \bar{\mathcal{M}}_{\mathrm{II}}, \\ 0 & \text{elsewhere}, \end{cases} \tag{5.31}$$
let z be the solution of
$$z' = -cz + F \quad \text{on } (0, 1), \qquad z(0) = 0, \tag{5.32}$$
i.e.,
$$z(x) = \int_0^x F(t)e^{-\int_t^x c(\tau)\,d\tau}\,dt.$$
Then
$$\max(\|z\|_{L_2(0,1)}, \|z'\|_{L_2(0,1)}) \lesssim \|F\|_{L_2(0,1)} \leq \sqrt{\sum_{K' \in \bar{\mathcal{M}}_{\mathrm{II}}} \|g\|_{L_2(K')}^2}.$$
Moreover, z is piecewise smooth with respect to $\mathcal{T}$, and $(z|_{K'})'' = -c|_{K'}(z|_{K'})'$ $(K' \in \mathcal{T})$.

Let $(0, w) \in \check{\mathbb{U}}_{\mathcal{T}}$ be defined by taking w as the continuous piecewise linear interpolant of z with respect to $\mathcal{T}$. We have that
$$\|z - w\|_{L_2(K')} \lesssim \mathrm{diam}(K')\|z'\|_{L_2(K')},$$
$$\|z' - w'\|_{L_2(K')} \lesssim \mathrm{diam}(K')\|z''\|_{L_2(K')} \lesssim |c|_{K'}|\,\mathrm{diam}(K')\|z'\|_{L_2(K')}.$$

Let us first assume that $c|_{K'} \neq 0$ for all $K' \in \mathcal{T}$. Using (5.30), the definition of F, (5.29), $m_w \leq m_u + 1$, $F|_{K'} \in P_0(K') \subset \mathcal{P}_{m_u}(K')$, and the definition of z, we obtain
$$\sum_{K' \in \bar{\mathcal{M}}_{\mathrm{II}}} \|G\|_{L_2(K')}^2 = \sum_{K' \in \bar{\mathcal{M}}_{\mathrm{II}}} \int_{K'} Gg \, d\mathbf{x} - \int_\Omega gw' - we \, d\mathbf{x}$$
$$= \sum_{K' \in \mathcal{T}} \int_{K'} Fg - gw' + we \, d\mathbf{x} = \sum_{K' \in \mathcal{T}} \int_{K'} \frac{e}{c}(w' - F + cw) \, d\mathbf{x}$$
$$= \sum_{K' \in \mathcal{T}} \int_{K'} \frac{e}{c}((w - z)' + c(w - z)) \, d\mathbf{x} = \sum_{K' \in \mathcal{T}} \int_{K'} \frac{e}{c}(w - z)' + e(w - z) \, d\mathbf{x}$$
$$\lesssim \max_{K' \in \mathcal{T}} \mathrm{diam}(K') \sum_{K' \in \mathcal{T}} \|e\|_{L_2(K')}\|z'\|_{L_2(K')} \leq \sigma \sqrt{\sum_{K' \in \mathcal{T}} \|e\|_{L_2(K')}^2} \|z'\|_{L_2(0,1)}$$
$$\lesssim \sigma \eta_\Omega(\bar{u}_{\mathcal{T}}^\delta, \bar{w}_{\mathcal{T}}^\delta; f)^2 \lesssim \sigma \sum_{K' \in \bar{\mathcal{M}}_{\mathrm{II}}} \|g\|_{L_2(K')}^2,$$
where the last inequality follows from (5.27) and Lemma 5.12.

Now consider the case that for one or more K', $c|_{K'} = 0$. Then on such a K', z is linear (or even constant when $K' \in \mathcal{T} \setminus \bar{\mathcal{M}}_{\mathrm{II}}$) and so coincides with w. Let $\bar{z}$ denote the average of z on K'. For such a K', from $e \perp \mathcal{P}_{m_u}(K')$ we estimate

$$\left| \int_{K'} Fg - gw' + we \, d\mathbf{x} \right| = \left| \int_{K'} ze \, d\mathbf{x} \right| = \left| \int_{K'} (z - \bar{z})e \, d\mathbf{x} \right|$$

$$\leq \mathrm{diam}(K') \|z'\|_{L_2(K')} \|e\|_{L_2(K')},$$

and arrive at the same conclusion. For $n = 1$ this completes the proof that, for σ sufficiently small, (5.27) implies (5.28), and thus of Theorem 5.9. Note that $\alpha > 0$ could even be stipulated as small as we wish. ☐

5.6 Theorem 5.9 for $n > 1$

The above reasoning for $n = 1$ does not seem to directly carry over to the multi-dimensional case. In fact, it is not clear how to approximate the solution z to the analog of (5.32) by a w-component in the current trial space, the difficulty being the non-smoothness of z in the directions orthogonal to $\mathbf{b}$.

To deal with this problem, for $n > 1$ we consider a downwind enriched refinement procedure as specified in Definition 5.1. Let us assume that nevertheless Theorem 5.9 does *not* hold. That is, there is a $\vartheta \in (0, 1]$ such that for any $v < 1$, $r \in \mathbb{N}$, there exist $\mathcal{T} \in \mathbb{T}$ and $f \in \mathbb{F}_{\mathcal{T}}^{\delta}$ with the property that for the marked cells $\bar{\mathcal{M}} = \bar{\mathcal{M}}((\bar{u}_{\mathcal{T}}^{\delta}, \bar{w}_{\mathcal{T}}^{\delta}), \vartheta)$ and refined triangulation $\tilde{\mathcal{T}} = \tilde{\mathcal{T}}(\mathcal{T}, \bar{\mathcal{M}}, r)$, one still has

$$\eta_\Omega(\bar{u}_{\tilde{\mathcal{T}}}^{\delta}, \bar{w}_{\tilde{\mathcal{T}}}^{\delta}; f) > v \eta_\Omega(\bar{u}_{\mathcal{T}}^{\delta}, \bar{w}_{\mathcal{T}}^{\delta}; f). \tag{5.33}$$

Splitting $\bar{\mathcal{M}} = \bar{\mathcal{M}}_{\mathrm{I}} \dot{\cup} \bar{\mathcal{M}}_{\mathrm{II}}$ as before, as we have seen in Section 5.4 for v sufficiently close to 1 and r sufficiently large, the case that $\eta_{\bar{\mathcal{M}}_{\mathrm{II}}}(\bar{u}_{\mathcal{T}}^{\delta}, \bar{w}_{\mathcal{T}}^{\delta}; f)^2 < \eta_{\bar{\mathcal{M}}_{\mathrm{I}}}(\bar{u}_{\mathcal{T}}^{\delta}, \bar{w}_{\mathcal{T}}^{\delta}; f)^2$ would, on account of Corollary 5.11, immediately lead to a contradiction.

So let us focus on the case that

$$\eta_{\bar{\mathcal{M}}_{\mathrm{II}}}(\bar{u}_{\mathcal{T}}^{\delta}, \bar{w}_{\mathcal{T}}^{\delta}; f)^2 \geq \eta_{\bar{\mathcal{M}}_{\mathrm{I}}}(\bar{u}_{\mathcal{T}}^{\delta}, \bar{w}_{\mathcal{T}}^{\delta}; f)^2. \tag{5.34}$$

Following the analysis of the previous Section 5.5, recall the definitions of $g = \partial_{\mathbf{b}} \bar{w}_{\mathcal{T}}^{\delta} + c\bar{u}_{\mathcal{T}}^{\delta} - f$, $e = \bar{u}_{\mathcal{T}}^{\delta} - \bar{w}_{\mathcal{T}}^{\delta}$, and that of G in (5.23) and F in (5.31). From the definition of bulk chasing, (5.34), and Lemma 5.12 we infer that

$$\eta_\Omega(\bar{u}_{\mathcal{T}}^{\delta}, \bar{w}_{\mathcal{T}}^{\delta}; f) \leq \frac{\sqrt{2}}{\vartheta} \eta_{\bigcup\{K' \in \bar{\mathcal{M}}_{\mathrm{II}}\}}(\bar{u}_{\mathcal{T}}^{\delta}, \bar{w}_{\mathcal{T}}^{\delta}; f)$$

$$\leq \frac{\sqrt{2}}{\vartheta} \sqrt{(2\|c\|_{L_\infty(\Omega)} + 1)^2 + 1} \|g\|_{L_2(\bigcup\{K' \in \bar{\mathcal{M}}_{\mathrm{II}}\})}. \tag{5.35}$$

Let us now define the quantities $\tilde{g}$, $\tilde{e}$, and $\tilde{G}$ in analogy to g, e, and G, but with respect to the least-squares solution $(\bar{u}_{\tilde{\mathcal{T}}}^{\delta}, \bar{w}_{\tilde{\mathcal{T}}}^{\delta}) \in \check{\mathbb{U}}_{\tilde{\mathcal{T}}}^{\delta}$ and the refined partition $\tilde{\mathcal{T}}$. The pair $(\bar{u}_{\tilde{\mathcal{T}}}^{\delta}, \bar{w}_{\tilde{\mathcal{T}}}^{\delta})$ being a minimizer of $\eta_\Omega^2(\cdot, \cdot; g)$ over $\check{\mathbb{U}}_{\tilde{\mathcal{T}}}^{\delta}$ is equivalent to

$$\tilde{e} + c\tilde{g} \perp_{L_2(\tilde{K})} \mathcal{P}_{m_u}(\tilde{K}) \ (\tilde{K} \in \tilde{\mathcal{T}}), \quad \int_\Omega \tilde{g}\partial_{\mathbf{b}} w - w\tilde{e} \, d\mathbf{x} = 0 \ ((0, w) \in \check{\mathbb{U}}_{\tilde{\mathcal{T}}}^{\delta}). \tag{5.36}$$

As shown next, the assumption that the error indicator has not been reduced much when passing to $\tilde{\mathcal{T}}$, implies that g, e must be very close to $\tilde{g}$, $\tilde{e}$, respectively. In fact, the orthogonality relation analogous to (5.4) reads as

$$\eta_\Omega^2(\bar{u}_{\tilde{\mathcal{T}}}^{\delta}, \bar{w}_{\tilde{\mathcal{T}}}^{\delta}; f) = \eta_\Omega^2(\bar{u}_{\mathcal{T}}^{\delta}, \bar{w}_{\mathcal{T}}^{\delta}; f) - \eta_\Omega^2(\bar{u}_{\tilde{\mathcal{T}}}^{\delta} - \bar{u}_{\mathcal{T}}^{\delta}, \bar{w}_{\tilde{\mathcal{T}}}^{\delta} - \bar{w}_{\mathcal{T}}^{\delta}; 0).$$

In combination with estimate (5.35) and our assumption (5.33), this shows that there exists a $\zeta = \zeta(v)$ with $\lim_{v\uparrow 1} \zeta(v) = 0$ such that

$$\|g - \tilde{g}\|_{L_2(\Omega)} \leq \zeta \|g\|_{L_2(\bigcup\{K' \in \bar{\mathcal{M}}_{\mathrm{II}}\})}, \quad \|e - \tilde{e}\|_{L_2(\Omega)} \leq \zeta \|g\|_{L_2(\bigcup\{K' \in \bar{\mathcal{M}}_{\mathrm{II}}\})}. \tag{5.37}$$

This fact together with an affirmative answer to the following conjecture will allow us to complete the proof of Theorem 5.9.

Conjecture 5.17. There exist constants

$$\xi < \left(\frac{\sqrt{2}}{9}\sqrt{(2\|c\|_{L_\infty(\Omega)} + 1)^2 + 1}\right)^{-1}$$

and $r \in \mathbb{N}$ such that there exists a $(0, \tilde{w}) \in \breve{\mathbb{U}}_{\tilde{\mathcal{T}}}^\delta$ with

$$\|\partial_\mathbf{b}\tilde{w} + c\tilde{w} - F\|_{L_2(\Omega)} \le \xi\|F\|_{L_2(\Omega)}, \quad \|\tilde{w}\|_{L_2(\Omega)} \lesssim \|F\|_{L_2(\Omega)}, \tag{5.38}$$

where $\tilde{w}$ vanishes outside the union of the cells of $\mathcal{T}$ that were refined in $\tilde{\mathcal{T}} = \tilde{\mathcal{T}}(\mathcal{T}, \bar{\mathcal{M}}, r)$.

We postpone supporting arguments for the validity of this conjecture and turn first, for r large enough and ν sufficiently close to 1, to verifying the hypothesis of Lemma 5.15. This lemma then asserts the validity of Theorem 5.9, which, for ν sufficiently close to 1, will contradict (5.33), thereby finishing the proof.

To that end, with $\tilde{w}$ from Conjecture 5.17, using (5.36) we write

$$\begin{aligned}
\|G\|_{L_2(\bigcup\{K' \in \bar{\mathcal{M}}_{\mathrm{II}}\})}^2 &= \sum_{K' \in \bar{\mathcal{M}}_{\mathrm{II}}} \int_{K'} Gg\, d\mathbf{x} = \sum_{K' \in \bar{\mathcal{M}}_{\mathrm{II}}} \int_{K'} G\tilde{g}\, d\mathbf{x} + \sum_{K' \in \bar{\mathcal{M}}_{\mathrm{II}}} \int_{K'} G(g - \tilde{g})\, d\mathbf{x} \\
&= \int_\Omega F\tilde{g}\, d\mathbf{x} - \int_\Omega \tilde{g}\partial_\mathbf{b}\tilde{w} - \tilde{w}\tilde{e}\, d\mathbf{x} + \sum_{K' \in \bar{\mathcal{M}}_{\mathrm{II}}} \int_{K'} G(g - \tilde{g})\, d\mathbf{x} \\
&= -\int_\Omega (\partial_\mathbf{b}\tilde{w} + c\tilde{w} - F)\tilde{g}\, d\mathbf{x} + \sum_{K' \in \bar{\mathcal{M}}_{\mathrm{II}}} \int_{K'} G(g - \tilde{g})\, d\mathbf{x} + \int_\Omega \tilde{w}(\tilde{e} + c\tilde{g})\, d\mathbf{x}.
\end{aligned}$$

The first and second term on the right can be bounded by

$$\left|\int_\Omega (\partial_\mathbf{b}\tilde{w} + c\tilde{w} - F)\tilde{g}\, d\mathbf{x}\right| \le \xi\|F\|_{L_2(\Omega)}(1 + \zeta)\|g\|_{L_2(\Omega)}$$

$$\le \xi\|G\|_{L_2(\bigcup\{K' \in \bar{\mathcal{M}}_{\mathrm{II}}\})}(1 + \zeta)\frac{\sqrt{2}}{9}\sqrt{(2\|c\|_{L_\infty(\Omega)} + 1)^2 + 1}\,\|g\|_{L_2(\bigcup\{K' \in \bar{\mathcal{M}}_{\mathrm{II}}\})}, \tag{5.39}$$

where we have used (5.37) and (5.35), and

$$\left|\sum_{K' \in \bar{\mathcal{M}}_{\mathrm{II}}} \int_{K'} G(g - \tilde{g})\, d\mathbf{x}\right| \le \zeta\|G\|_{L_2(\bigcup\{K' \in \bar{\mathcal{M}}_{\mathrm{II}}\})}\|g\|_{L_2(\bigcup\{K' \in \bar{\mathcal{M}}_{\mathrm{II}}\})}, \tag{5.40}$$

respectively.

To proceed let $Q_{\tilde{\mathcal{T}}}$ denote the $L_2(\Omega)$-orthogonal projector onto $\prod_{\tilde{K} \in \tilde{\mathcal{T}}} \mathcal{P}_{m_u}(\tilde{K})$, using (5.36) for the third term, we write

$$\begin{aligned}
\int_\Omega \tilde{w}(\tilde{e} + c\tilde{g})\, d\mathbf{x} &= \int_\Omega ((I - Q_{\tilde{\mathcal{T}}})\tilde{w})(\tilde{e} + c\tilde{g})\, d\mathbf{x} \\
&= \int_\Omega ((I - Q_{\tilde{\mathcal{T}}})\tilde{w})(\tilde{e} - e + c(\tilde{g} - g))\, d\mathbf{x} + \int_\Omega \tilde{w}((I - Q_{\tilde{\mathcal{T}}})(e + cg))\, d\mathbf{x}.
\end{aligned}$$

Thanks to the estimates in (5.37), the first term at the right can be bounded by a constant multiple of $\zeta\|G\|_{L_2(\bigcup\{K' \in \bar{\mathcal{M}}_{\mathrm{II}}\})}\|g\|_{L_2(\bigcup\{K' \in \bar{\mathcal{M}}_{\mathrm{II}}\})}$. We use next that w vanishes outside the union of the cells of $\mathcal{T}$ which have been refined in $\tilde{\mathcal{T}} = \tilde{\mathcal{T}}(\mathcal{T}, \bar{\mathcal{M}}, r)$, and that e and g are piecewise polynomial with respect to $\mathcal{T}$. Moreover, by Remark 5.2, all cells in the support of $\tilde{w}$ are (at least) rth refinements of cells in $\mathcal{T}$ underlying e and g. Hence, the usual combination of direct and inverse estimates shows then that the second term can be bounded by $\|G\|_{L_2(\bigcup\{K' \in \bar{\mathcal{M}}_{\mathrm{II}}\})}\eta(r)\|g\|_{L_2(\bigcup\{K' \in \bar{\mathcal{M}}_{\mathrm{II}}\})}$, where η as a function of r, tends to zero as $r \to \infty$. For any constant

$$\alpha \in \left(\xi\frac{\sqrt{2}}{9}\sqrt{(2\|c\|_{L_\infty(\Omega)} + 1)^2 + 1}, 1\right),$$

the combination of these latter results, (5.39), and (5.40) shows that for r large enough and ν sufficiently close to 1, $\|G\|_{L_2(\bigcup\{K' \in \bar{\mathcal{M}}_{\mathrm{II}}\})} \le \alpha\|g\|_{L_2(\bigcup\{K' \in \bar{\mathcal{M}}_{\mathrm{II}}\})}$, which by Lemma 5.15, for ν sufficiently close to 1, contradicts (5.33), as required. $\qquad\square$

Let us close this section with some brief comments on Conjecture 5.17. First, as mentioned earlier, by Remark 5.2, the downwind enrichment in the refinement strategy makes sure that the correction $\bar{w}$ is constructed on (an essentially uniform) refined mesh. This certainly helps a relation like (5.38) to be possible and actually motivated the inclusion of the downwind enrichments. Moreover, the conjecture asks "only" for a fixed relative accuracy ξ where ξ need not be arbitrarily small. Given that the data are piecewise polynomials (which are actually piecewise constants in stream direction), this does not seem to ask for too much.

On the other hand, since we can neither limit a priori the number of polynomial pieces in F nor their position relative to the direction of $\mathbf{b}$, an argument does not seem to be straightforward. In fact, whereas we can represent the exact solution of $\partial_{\mathbf{b}} z + cz = F$ with zero inflow conditions explicitly along characteristics ensuring sufficient smoothness in this direction, smoothness in cross-flow direction does not seem to be easy to control. Nevertheless, the overall variation in cross-flow direction is still highly restrained for data of the type F.

Finally, we would like to stress that a possibly $\mathcal{T}$-dependent r such that (5.38) holds true always exists. By the above arguments this immediately translates into a statement on error reduction based on such a (variable) refinement depth.

6 Concluding Remarks

We have established reliability and efficiency of computable local error indicators for DPG discretizations of linear transport equations with variable convection and reaction coefficients. For constant (with respect to the spatial variables) convection fields, arising for instance in kinetic models, we have determined refinement strategies based on the a posteriori error indicators which are guaranteed to give rise to a fixed error reduction rate. The latter results make essential use of a tight interrelation of the DPG scheme with certain least squares formulations providing insight of its own right. In particular, error reduction for one scheme implies the same for the other one. To our knowledge the issue of error reduction for least squares methods even for the classical elliptic case is largely open. In that sense the present results mark some progress in this regard as well.

On the other hand, in view of these findings one may raise the question as to why not using the seemingly simpler least-squares scheme instead of the DPG scheme. However, giving up on the simple interpretation of the w-component as a second approximation to the exact solution in a stronger norm when $f \in L_2(\Omega)$, the DPG scheme still provides a meaningful approximate solution $u_{\mathcal{T}}^{\delta}$ in $L_2(\Omega)$ to the transport equation even when f is less regular. But also for L_2-data f, in the least squares formulation errors are measured solely in a norm that depends in a very sensitive way directly on the convection field. In the variable convection case the corresponding space varies essentially (even as a set) under perturbations of this convection field. Therefore, at this point Proposition 5.7 serves primarily as a theoretical tool.

Among other things a prize for using the interrelation between DPG and least squares formulations is a remaining lack of quantification of the error reduction results manifesting itself in two ways: the subgrid depth needed to establish efficiency and reliability of the computable error indicators, similar to establishing uniform inf-sup stability of the pairs of trial- and test-spaces, is not precisely specified. As indicated by earlier numerical results in [2] any attempt along the given lines to quantify the subgrid-depth would still be over pessimistic. The same is expected to be true for the refinement depth r associated with the marked cells. These issues call for further research in this area.

Finally, the refinement strategies that can be shown to guarantee a fixed error reduction involve for several spatial variables so far a certain downstream enrichment of the marked cells in combination with a conjecture. It is open whether this enrichment is in general necessary which would establish an essential difference from the univariate case where it is not necessary.

References

[1] P. Binev, W. Dahmen and R. DeVore, Adaptive finite element methods with convergence rates, *Numer. Math.* **97** (2004), 219–268.

[2] D. Broersen, W. Dahmen and R. P. Stevenson, On the stability of DPG formulations of transport equations, *Math. Comp.* **87** (2018), no. 311, 1051–1082.

[3] J. H. Bramble and J. Xu, Some estimates for a weighted L^2 projection, *Math. Comp.* **56** (1991), 463–476.

[4] 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.

[5] 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.

[6] 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.

[7] J. Gopalakrishnan and W. Qiu, An analysis of the practical DPG method, *Math. Comp.* **83** (2014), no. 286, 537–552.

[8] R. H. Nochetto, K. G. Siebert and A. Veeser, Theory of adaptive finite element methods: An introduction, in: *Multiscale, Nonlinear and Adaptive Approximation*, Springer, Berlin (2009), 409–542.

[9] R. P. Stevenson, Optimality of a standard adaptive finite element method, *Found. Comput. Math.* **7** (2007), no. 2, 245–269.

[10] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley–Teubner, Chichester, 1996.