The DPG Method for the Convection–Reaction Problem, Revisited

Leszek Demkowicz[a], Nathan V. Roberts[b] and Judit Muñoz-Matute[a]

[a]Oden Institute, The University of Texas at Austin
[b]Sandia National Laboratories[1]

Abstract

We study both conforming and non-conforming versions of the practical DPG method for the convection-reaction problem. We determine that the most common approach for DPG stability analysis - construction of a local Fortin operator, is infeasible for the convection-reaction problem. We then develop a line of argument based on a direct proof of discrete stability; we find that employing a polynomial enrichment for the test space does not suffice for this purpose, motivating the introduction of a (two-element) subgrid mesh. The argument combines mathematical analysis with numerical experiments.

1 Introduction

The research on the Discontinuous Petrov-Galerkin (DPG) method started with the convection problem [11]. We began with a spectral Petrov-Galerkin method (one element case) based on the weak formulation (2.7). We considered only the case of pure convection with a constant advection vector b, and restricted ourselves to affine triangles in 2D. Under these assumptions, the outflow boundary consists of a single edge or two edges. The solution u is approximated with polynomials of order p, and the trace with polynomials of order $p+1$ defined edge-wise; i.e., in the case of two edges, the trace is discontinuous. The novelty of the formulation lied in the use of the weak formulation and a special optimal, *non-polynomial* test space including lifts of the trace space that are constant along the streamlines. We demonstrated that the use of such a test space guaranteed the same inf-sup constant as on the continuous level.

Under the assumption of a constant advection field, the spectral method was then extended to 2D triangular meshes using a marching strategy – from inflow to outflow boundary. The mesh is partitioned into layers $\Omega_{h,n}$, $n = 1, \ldots, N$. The first layer $\Omega_{h,1}$ consists of elements K with inflow boundary ∂K_- contained in the global inflow boundary Γ_-. The nth layer $\Omega_{h,n}$ contains all possible elements whose inflow edges are outflow edges for elements from the previous layers or are contained in Γ_-. The optimal Petrov-Galerkin spectral approximation is then applied to elements from one layer at a time. Traces computed in steps $1, \ldots, n$ provide inflow data for elements from $\Omega_{h,n+1}$. The method was proved to have optimal approximation properties, in both element size h and polynomial order p. In particular, contrary to the standard DG approach which guarantees to deliver only the suboptimal convergence rate $h^{p+1/2}$, the new method delivers the optimal h^{p+1} convergence rate.

The research on the DPG method for the convection problem was continued in [12]. We proposed

[1]Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This report is Sandia report number SAND2021-3435 R.

there the concept of optimal test functions based on inverting the Riesz operator element-wise. The (ideal) DPG method with optimal test functions was shown to be equivalent to a minimum-residual method where the residual is measured in an appropriate dual test norm. Consequently, the global stiffness matrix is symmetric (Hermitian for complex-valued problems) and positive-definite. Under the continued assumption of a constant advection field, we considered a special test inner product defined on each element K,

$$(v, \delta v)_{V(K)} := \int_K \mathrm{div}(bv)\,\mathrm{div}(b\delta v) + \int_{\partial K_+} v\delta v\,, \tag{1.1}$$

and made a number of important observations. First of all, with this special inner product, we were able to compute the *exact* optimal test functions explicitly. Secondly, for a single element case, the resulting optimal test space – the span of optimal test functions – coincided exactly with the test space used in [11]. Thirdly, we showed that, in 1D case and for a class of 2D meshes, both the original method from [11] and the new method were delivering exactly the same solution. Note that the marching scheme produces a lower-triangular global stiffness matrix whereas the new method produces a symmetric matrix. The equivalence between the two approaches was recently exploited in the context of DPG-based optimal time discretization strategies [23, **?**]. However, we emphasize that, for general 2D meshes, the two methods deliver slightly different results.

Finally, the ultimate DPG method proposed in [12] was based on the concept of the *approximate optimal test functions* determined by inverting an approximate Riesz operator defined on an enriched test space consisting of polynomials of order $p + 1$. The method seemed to deliver practically identical results with the two previous schemes, and was much more general and easier to implement, but proofs were missing.

The subject of the DPG method for the transport equation was more recently picked up by Broersen, Dahmen and Stevenson in [2]. The work contains an in-depth analysis of the advection-reaction problem with a *variable* advection vector. The authors start with the proof of well-posedness for the formulation with broken (product) test spaces. The proof is different from the general theory presented in [5]. The authors consider then an approximate bilinear form, replacing the variable advection vector with an element-wise constant approximation. A discrete inf-sup condition for the approximate bilinear form is shown by using a corrected version of inner product (1.1) [2]. The relation between the original and approximate bilinear forms is then studied culminating in the proof of the discrete inf-sup condition for an enriched (search) space obtained by refining the original element (of enriched order $p + 1$) a finite (unspecified) number of times. However, the authors mention that, in practice, no need for refining the test element has been observed. The result thus represents more of a 'sanity check' than an actual explanation of why the method works. Nevertheless, this is the first work in the DPG literature on problems with variable coefficients. The local constructions of Fortin operators [20, 5, 24, 15] are nothing else than 'sanity checks' as well, since both the required enrichment in order Δp and continuity constants for the Fortin operator are very pessimistic when compared with numerical experience[3] . A notable exception is [6] where a global stability analysis is

[2]The authors correctly observe that inner product (1.1) is not uniformly (in element size h) equivalent to the standard adjoint graph inner product.

[3]We use $\Delta p = 1$ in all practical computations.

presented.

The DPG method analyzed in [2] lays down the foundation for a general *anisotropic* refinement methodology for a class of transport problems developed in [7, 8]. For a model parametric transport problem in heterogeneous media in 2+1 dimensions, in [7] Dahmen et al. show that sparse tensorization of the scheme combining the anisotropic refinements in space with hierarchic collocation in ordinate space can overcome the curse of dimensionality when approximating averaged bulk quantities.

Finally, Dahmen and Stevenson come back to the subject in the their most recent contribution [8] devoted to the study of an automatic adaptive scheme for the convection-reaction problem based on the DPG discretization. Proving the convergence of adaptive scheme based on minimum-residual methods has been an open problem for several years[4] and, to our best knowledge, [8] presents the very first such analysis for a DPG method. The proof is complete in one space dimension and a multi-dimensional result holds under additional conjectures.

During the review process for this work, we learned about [3] which preceeds our work. The authors reverse the process of computing the optimal pairs of trial/test functions and, having selected test functions, compute the corresponding trial functions. This solves the problem of approximating the optimal test functions but brings up the issue of approximability. In particular, we have learned from [3] about reference [21] where the density of $C^\infty(\bar{\Omega})$ functions in the graph space can be found.

The present work attempts to explain why the simple 'practical DPG' method with the simple enriched polynomial space works. We combine analytical arguments with numerical experimentation at the single-element level. We start in Section 2 by establishing the well-posedness of the ultraweak variational formulation using the theory of closed operators. Well-posedness of the corresponding broken formulation is established in Section 3. In our proofs, we follow the general theory outlined in [5]. In Section 4 we take a break from theory and reproduce the numerical examples from [2]. Section 5 presents our first attempt to prove the convergence by constructing a local Fortin operator. Having recognized limitations of the local construction, we continue in Section 6 with an attempt to prove the global discrete stability directly. We consider first the case of a simple polynomial enriched space to convince ourselves that we stand no chance for proving discrete stability for this case. The analysis leads us back to composite test spaces considered in [11] and, in Section 7 we continue the global discrete stability analysis in the context of such test spaces. We conclude with final remarks in Section 8.

We apologize for perhaps a bit confusing order of the presentation which follows precisely the order of research that we have done. At first, we hoped that the construction of a local Fortin operator would be sufficient. When this turned out to be wrong, we pursued the proof of discrete stability with simple polynomial (enriched) space. When this turned out to be unlikely as well, we returned to the original idea of piecewise polynomial test spaces from [11]. In summary, our main contributions are:

1. Analysis of the well-posedness of the ultraweak formulation within the classical theory of unbounded

[4] [1] was the first breakthrough on the subject.

closed operators.

2. Analysis of the well-posedness of the broken formulation within the general framework presented in [5].

3. A failed attempt to construct a local Fortin operator. Sometimes a negative result can teach us something, too.

4. A proof of discrete stability for 2D triangular and quadrilateral meshes using *just* a two-subelement mesh for test functions. The proof has been reduced to numerical experiments for a master element and it is, indeed, more of a sketch of ideas than a formal proof. We hope, however, that it provides additional insight into the problem.

5. A stability analysis for the method with discontinuous traces.

2 Ultraweak Variational Formulation for the Convection-Reaction Problem

2.1 Closed Operators and the Ultraweak (UW) Variational Formulation.

We begin by recalling the general structure provided by the classical theory of unbounded closed operators applicable to differential operators. Let $\Omega \subset \mathbb{R}^n$ be a domain (an open and connected set). Consider a general closed operator [25] A,

$$L^2(\Omega) \supset D(A) \ni u \to Au \in L^2(\Omega)\,,$$

and its L^2-adjoint A^*,

$$L^2(\Omega) \supset D(A^*) \ni v \to A^*v \in L^2(\Omega)\,.$$

Recall that the necesary and sufficient condition for the adjoint to exist is that domain $D(A)$ is dense in $L^2(\Omega)$.

Strong variational formulation. Assuming $f \in L^2(\Omega)$, we consider the strong (trivial) variational formulation, equivalent to the strong statement of the problem:

$$\left\{\begin{array}{l} u \in D(A) \\ (Au,v) = (f,v) \quad v \in L^2(\Omega) \end{array}\right. \quad \Leftrightarrow \quad \left\{\begin{array}{ll} u \in D(A) & \\ Au = f & A\,:\,D(A) \to L^2(\Omega)\,. \end{array}\right.$$

Ultraweak (UW) variational formulation. The relation between operator A and its adjoint A^*,

$$(Au, v) = (u, A^* v) \qquad u \in D(A), v \in D(A^*)\,,$$

allows the introduction of the *ultraweak (UW) variational formulation*. Given $l \in D(A^*)'$, we consider:

$$\left\{ \begin{array}{l} u \in L^2(\Omega) \\ (u, A^* v) = l(v) \quad v \in D(A^*) \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} u \in L^2(\Omega) \\ \widetilde{A} u = l \qquad \widetilde{A} \,:\, L^2(\Omega) \to D(A^*)'\,. \end{array} \right. \tag{2.2}$$

In the formulations above, $D(A)$ and $D(A^*)$ are equipped with the graph norms:

$$\|u\|^2_{H_A} := \|u\|^2 + \|Au\|^2 \qquad \|v\|^2_{H_{A^*}} := \|v\|^2 + \|A^* v\|^2\,.$$

Note the difference between the original, closed operator A, and continuous operator A (denoted with the same symbol) corresponding to the strong variational formulation. Operator $\widetilde{A}$ is an extension of operator A but with different norms.

REMARK 1 If operator A represents a system of equations, computation of its adjoint requires relaxation (integration by parts) in all of the equations. Frequently, a selective relaxation of *some* of the equations leading to various *weak formulations* is possible. Hence the name of the UW formulation. For the convection-reaction operator, we are dealing with a single equation only, so weak and ultraweak formulation are identical. ∎

We recall now a version of the classical *Closed Range Theorem for Closed Operators*.

THEOREM 1

If operator A is bounded below,

$$\|Au\| \geq \alpha \|u\|, \qquad u \in D(A),\, \alpha > 0\,,$$

and its adjoint A^* is injective, then the adjoint A^* is also bounded below with the same constant,

$$\|A^* v\| \geq \alpha \|v\|, \qquad v \in D(A^*),\, \alpha > 0\,.$$

Consequently, for any $g \in L^2(\Omega)$, the adjoint problem:

$$\left\{ \begin{array}{l} v \in D(A^*) \\ A^* v = g\,, \end{array} \right.$$

is well-posed. ∎

THEOREM 2

Under the assumptions of Theorem 1, the UW formulation is well-posed. ∎

Proof:

- Inf-sup condition:
  Let $u \in L^2(\Omega)$, and w be the solution of the adjoint problem: $A^* w = u$. Then

$$\begin{array}{rcrl} \|w\|^2 & \leq & \alpha^{-2}\|u\|^2 & \Rightarrow \|w\|^2_{H_{A^*}} \leq (\alpha^{-2}+1)\,\|u\|^2 \Rightarrow \dfrac{1}{\|w\|_{H_{A^*}}} \geq (\alpha^{-2}+1)^{-\frac{1}{2}}\dfrac{1}{\|u\|} \\ \|A^* w\|^2 & = & \|u\|^2 & \end{array}$$

  and,

$$\sup_{v \in D(A^*)} \frac{|(u, A^* v)|}{\|v\|_{H_{A^*}}} \geq \frac{|(u, A^* w)|}{\|w\|_{H_{A^*}}} \geq (\alpha^{-2}+1)^{-\frac{1}{2}} \|u\| \,.$$

- The dual operator A^* is the *conjugate operator* of $\widetilde{A}$,

$$\begin{array}{ccc} L^2(\Omega) & \overset{\widetilde{A}}{\rightarrow} & (D(A^*))' \\ L^2(\Omega) & \overset{\widetilde{A}'=A^*}{\leftarrow} & D(A^*) \end{array}$$

  and $\mathcal{N}(A^*) = \{0\}$.

The result follows now from the Babuška-Nečas Theorem.

∎

Convection-reaction problem. We shall assume that both advection $b(x)$ and reaction coefficient $c(x)$ are piecewise smooth and satisfy additional assumptions to guarantee boundedness below of the operator, see Appendix A. In the sequel, we will also assume that the FE grid will always match the possible discontinuities; i.e., no discontinuities within elements are allowed. We also assume that $b(x)$ is globally $H(\mathrm{div})$-conforming, i.e., the normal component $b(x) \cdot n$ is continuous across inter-element boundaries. The convection-reaction problem under consideration is:

$$\left\{ \begin{array}{rll} b \cdot \boldsymbol{\nabla} u + cu &= f & \text{in } \Omega \\ u &= u_0 & \text{on } \Gamma_- \end{array} \right. , \tag{2.3}$$

where boundary $\Gamma = \partial\Omega$ is split (up to sets of measure zero) into three disjoint parts,

$$\begin{array}{rll} \Gamma_- & := \{x \in \Gamma : b_n(x) < 0\} & \text{(global inflow boundary),} \\ \Gamma_+ & := \{x \in \Gamma : b_n(x) > 0\} & \text{(global outflow boundary), and} \\ \Gamma_0 & := \{x \in \Gamma : b_n(x) = 0\} & \text{(global no-flow boundary).} \end{array}$$

Replacing the three disjoint parts with their (relative) interiors, we can assume that they are (relatively) open in Γ. Here $b_n = b_n(x) = b(x) \cdot n(x)$, with n being the outward unit normal vector on the boundary. Integration by parts leads to the formula for the *formal adjoint*,

$$\int_\Omega (\underbrace{b \cdot \boldsymbol{\nabla} u + cu}_{=:Au}) v = \int_\Omega u (\underbrace{-\operatorname{div}(bv) + cv}_{A^* v}) + \int_\Gamma b_n uv \,.$$

We study now the well-posedness of the problem within the outlined theory of closed operators. We begin by introducing the graph spaces for the operator A and its formal L^2-adjoint A^*. Observe that, with $u \in L^2(\Omega)$, the functional defined by

$$\phi \to \langle \mathrm{div}(bu), \phi \rangle := -\int_\Omega u\, b \cdot \boldsymbol{\nabla}\phi\,, \tag{2.4}$$

is continuous over $H_0^1(\Omega)$, i.e., it is an element of $H^{-1}(\Omega) := (H_0^1(\Omega))'$. Since

$$Au = b \cdot \boldsymbol{\nabla} u + cu = \mathrm{div}(bu) + \underbrace{(c - \mathrm{div}\, b)u}_{\in L^2(\Omega)} \quad \text{and} \quad A^*v = -\mathrm{div}(bv) + \underbrace{cv}_{\in L^2(\Omega)}\,,$$

asking for Au or A^*v to be an L^2-function is equivalent to asking for functional (2.4) (which is a-priori in $H^{-1}(\Omega)$) to be in $L^2(\Omega)$. Consequently, the graph spaces for operator A and its formal adjoint A^* are identical:

$$\begin{aligned} H_A(\Omega) &:= \{u \in L^2(\Omega) : Au \in L^2(\Omega)\} = \{u \in L^2(\Omega) : \mathrm{div}(bu) \in L^2(\Omega)\} \\ H_{A^*}(\Omega) &:= \{v \in L^2(\Omega) : A^*v \in L^2(\Omega)\} = \{v \in L^2(\Omega) : -\mathrm{div}(bu) \in L^2(\Omega)\} = H_A(\Omega)\,. \end{aligned}$$

REMARK 2 The assumption on piecewise smoothness of b, $\mathrm{div}\, b$ and c implies that $b, \mathrm{div}\, b, c \in L^\infty(\Omega)$, a necessary and sufficient condition for the functionals considered above to be in $H^{-1}(\Omega)$. In the context of classical arguments used in Appendix A and in the proof of Theorem 3, we prefer the more specific and yet sufficiently practical assumption on the coefficients. ∎

Lemma 1

The space of $C^\infty(\overline{\Omega})$ functions is dense in $H_A(\Omega)$ (in the graph norm).

$$\overline{C^\infty(\bar{\Omega})}^{H_A} = H_A(\Omega)\,. \tag{2.5}$$

∎

Proof: See [21]. ∎

Let $\phi_+(\phi_-)$ be a positive function defined on Γ, positive on $\Gamma_+(\Gamma_-)$ and zero on $\Gamma_-(\Gamma_+)$, such that it admits a $C^1(\bar{\Omega})$ extension. If Γ_- and Γ_+ are separated by a positive distance, we can assume $\phi_+(\phi_-)$ to be equal to unity on $\Gamma_+(\Gamma_-)$.

Lemma 2

There exist continuous trace operators,

$$\begin{aligned} \gamma_+ :&\quad H_A(\Omega) \to L^2_w(\Gamma_+) \\ \gamma_- :&\quad H_A(\Omega) \to L^2_w(\Gamma_-) \end{aligned}$$

where the weight

$$w = \begin{cases} \phi_+ |b_n| & \text{on } \Gamma_+ \\ \phi_- |b_n| & \text{on } \Gamma_- \end{cases} .$$

∎

Proof: The proof can be found in [22, 9, 18, 19], although in [22, 9, 19] specific blending functions are employed. For completeness, we reproduce the main argument. Let $u \in C^\infty(\overline{\Omega})$. We have:

$$\int_\Omega (b \cdot \boldsymbol{\nabla} u)\, u\, \phi_+ = \int_\Omega \phi_+\, b \cdot \boldsymbol{\nabla} \left(\frac{u^2}{2} \right) = - \int_\Omega \operatorname{div}(\phi_+\, b) \frac{u^2}{2} + \int_{\Gamma_+} \phi_+ b_n \frac{u^2}{2} .$$

Cauchy-Schwarz and Young's inequalities yield:

$$\begin{aligned} \|u\|^2_{L^2_w(\Gamma_+)} &\leq \| \operatorname{div}(\phi_+ b)\|_{L^\infty(\Omega)} \|u\|^2_{L^2(\Omega)} + 2 \|\phi_+\|_{L^\infty(\Omega)} \|b \cdot \boldsymbol{\nabla} u\|_{L^2(\Omega)} \|u\|_{L^2(\Omega)} \\ &\leq (\| \operatorname{div}(\phi_+ b)\|_{L^\infty(\Omega)} + \|\phi_+\|_{L^\infty(\Omega)}) \|u\|^2_{L^2(\Omega)} + \|\phi_+\|_{L^\infty(\Omega)} \|b \cdot \boldsymbol{\nabla} u\|^2_{L^2(\Omega)} . \end{aligned}$$

By the density argument (Lemma 1), the result extends to the graph space. The same technique is used for the inflow boundary. ∎

REMARK 3 We do not claim the trace operators to be surjective. Later on, we will introduce the trace spaces defined as the image of $H_A(\Omega)$ under the trace operators. ∎

The existence of the trace operator allows us now to define the domains on which operator A and its (so far formal) adjoint A^* are defined:

$$\begin{aligned} D(A) &:= \{u \in H_A(\Omega) : \gamma_- u = 0 \text{ on } \Gamma_-\}, \\ D(A^*) &:= \{v \in H_{A^*}(\Omega) : \gamma_+ v = 0 \text{ on } \Gamma_+\} . \end{aligned}$$

Lemma 3
Let $w = \phi_+ b_n$. Continuous functions on Γ_+ with compact support are dense in the weighted $L^2_w(\Gamma_+)$ space,

$$\overline{C_0(\Gamma_+)}^{L^2_w} = L^2_w(\Gamma_+) .$$

∎

Proof: Without losing generality, we assume that weight w may be zero only on the boundary of Γ_+. Let $u \in L^2_w(\Gamma_+)$. Then $w^{\frac{1}{2}} u \in L^2(\Gamma_+)$. By the density of $C_0(\Gamma_+)$ in $L^2(\Gamma_+)$, there exists a sequence $\psi_n \in C_0(\Gamma_+)$ converging to $w^{\frac{1}{2}} u$ in $L^2(\Gamma_+)$. But then $w^{-\frac{1}{2}} \psi_n \in C_0(\Gamma_+)$ as well, and $w^{-\frac{1}{2}} \psi_n$ converges to u in the weighted L^2 space. Indeed,

$$\int_{\Gamma_+} w (u - w^{-\frac{1}{2}} \psi_n)^2 = \int_{\Gamma_+} (w^{\frac{1}{2}} u - \psi_n)^2 \to 0 .$$

THEOREM 3
The closed operators A and A^* are adjoint to each other. ∎

Proof: We begin by observing that $D(A)$ is dense in $L^2(\Omega)$, a necessary and sufficient condition for the adjoint to exist. Indeed, $C_0^\infty(\Omega) \subset D(A)$, and $C_0^\infty(\Omega)$ is dense in $L^2(\Omega)$.

Step 1: We show first that

$$(Au, v) = (u, A^*v) \quad \forall\, u \in D(A), \forall\, v \in D(A^*)\,.$$

This, intuitively obvious result is actually quite delicate, due to the possibility of zero distance between the inflow and outflow boundaries. We begin by "cutting off" the inflow/outflow corners using the advection lines, see Fig. 1 for the illustration in the case of a simple square domain and constant advection field. Let $u \in D(A), v \in D(A^*)$, and let $u_k \in C^\infty(\bar\Omega), v_l \in C^\infty(\bar\Omega)$ be sequences converging to u, v in the graph norms H_A and H_{A^*}, respectively. We have,

$$\int_{\Omega_\epsilon} Au_k\, v_l - \int_{\Omega_\epsilon} u_k\, A^*v_l = \int_{\Gamma_-\cap\partial\Omega_\epsilon} b_n u_k v_l + \int_{\Gamma_+\cap\partial\Omega_\epsilon} b_n u_k v_l\,.$$

As for the truncated domain Ω_ϵ, the inflow and outflow boundaries are separated by a positive distance, the trace operators are continuous in the standard weighted L^2_w norm with weight $w = |b_n|$. Since

$$\|u\|_{H_A(\Omega_\epsilon)} \le \|u\|_{H_A(\Omega)} \quad \text{and} \quad \|v\|_{H_{A^*}(\Omega_\epsilon)} \le \|v\|_{H_{A^*}(\Omega)}\,,$$

convergence $\|u_k - u\|_{H_A(\Omega)} \to 0$, $\|v_l - v\|_{H_{A^*}(\Omega)}$ implies in the limit

$$\int_{\Omega_\epsilon} Au\, v - \int_{\Omega_\epsilon} u\, A^*v = 0\,.$$

Finally, passing with $\epsilon \to 0$ (Lebesgue Dominated Convergence Theorem), we obtain the desired result.

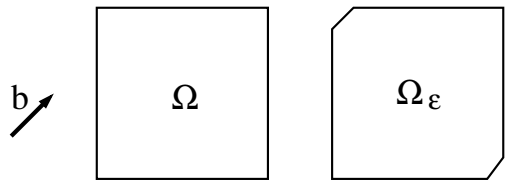

Figure 1: Cutting off inflow/outflow corners for a square domain in the case of a constant advection field.

Step 2: We show now that $D(A^*)$ is the *maximality* of functions for which the identity holds. By the same argument as above, we have

$$(Au, v) = (u, A^*v) + \int_{\Gamma_+} b_n uv \qquad u \in D(A),\, v \in H_{A^*}(\Omega)$$

where the boundary integral is understood in the Cauchy Principal-Value (PV) sense,

$$\int_{\Gamma_+} b_n uv = \lim_{\epsilon \to 0} \int_{\Gamma_+ \cap \partial\Omega_\epsilon} b_n uv \,.$$

Let now $v \in H_{A^*}(\Omega)$ be such that

$$(Au, v) = (u, A^*v) \quad \forall\, u \in D(A) \quad \Rightarrow \quad \int_{\Gamma_+} b_n uv = 0 \quad \forall\, u \in D(A)\,.$$

Let ϕ_+ be a smooth function from Lemma 2. We have

$$\int_{\Gamma_+} \phi_+ b_n v^2 = \int_{\Gamma_+} \phi_+ b_n (v - v_j) v + \int_{\Gamma_+} \phi_+ b_n v_j v,$$

where v_j is a sequence of $C_0(\Gamma_+)$ functions converging to v in the weighted L^2_w-norm with weight $w = \phi_+ b_n$. Using classical analysis based on characteristics (see Appendix A), we can claim existence of corresponding $u_j \in D(A)$ such that $\gamma_+ u_j = \phi_+ v_j$. We now have

$$\int_{\Gamma_+} \underbrace{\phi_+ b_n}_{=w} v^2 = \int_{\Gamma_+} \phi_+ b_n (v - v_j) v + \underbrace{\int_{\Gamma_+} b_n u_j v}_{=0}, \quad \text{and,}$$

$$\int_{\Gamma_+} w v^2 \le \|v - v_j\|_{L^2_w} \|v\|_{L^2_w} \to 0 \quad \text{as} \quad j \to \infty, \quad \Rightarrow \quad v = 0 \text{ in } L^2_w(\Gamma_+)\,.$$

∎

REMARK 4

- In what follows, we will need the integration by parts formula:

$$(Au, v) - (u, A^*v) = \int_\Gamma b_n uv \qquad u \in H_A(\Omega),\, v \in H_{A^*}(\Omega)\,. \tag{2.6}$$

  As in the proof above, the boundary integral is understood in the Cauchy PV sense. Indeed, we can claim

$$\int_{\Omega_\epsilon} Au\, v - \int_{\Omega_\epsilon} u\, A^*v = \int_{\Gamma \cap \partial\Omega_\epsilon} b_n uv\,.$$

  We use then the Lebesgue Theorem and pass to (a finite) limit on the left hand-side which proves that the limit on the right-hand side exists and it is finite as well. This was first observed in [22]. The Cauchy PV interpretation of the integral admits a singular behavior at inflow/outflow vertices provided a cancellation occurs [22]. Note, however, that if $u \in D(A)$ or $v \in D(A^*)$, this cannot happen as either the integral over the inflow or the outflow boundary vanishes.

- The explicit use of trace operators in the definitions of domains $D(A)$ and $D(A^*)$ is not necessary. We can define $D(A)$ first using the closure operation:

$$D(A) := \overline{\{u \in C^\infty(\bar{\Omega}) \,:\, u = 0 \text{ on } \Gamma_-\}}^{H_A(\Omega)} .$$

This is analogous, e.g., to defining $H^1_0(\Omega)$ as the closure of $C^\infty_0(\Omega)$ in the H^1-norm. It takes then non-trivial ground work to introduce the trace operator for $H^1(\Omega)$ and, eventually, characterize $H^1_0(\Omega)$ as a subspace of $H^1(\Omega)$ consisting of functions with vanishing trace. Once we have defined $D(A)$, we proceed with the definition of $D(A^*)$,

$$\begin{aligned} D(A^*) := &\quad \{v \in H_{A^*}(\Omega) \,:\, (Au, v) = (u, A^*v) \quad \forall\, u \in D(A)\} \\ = &\quad \{v \in H_{A^*}(\Omega) \,:\, (Au, v) = (u, A^*v) \quad \forall\, u \in C^\infty(\bar{\Omega}) \,:\, u = 0 \text{ on } \Gamma_-\} . \end{aligned}$$

where the last equality is a consequence of density of $C^\infty(\bar{\Omega}) \cap D(A)$ in $D(A)$ which has to be proved. This was the strategy used e.g. in [4, 13]. The explicit use of the trace operators is, however, considerably more intuitive, and their (even partial) knowledge is needed elsewhere, e.g., for identifying sufficient assumptions for initial conditions data to secure existence of finite energy extensions.

∎

Boundness below. In order to conclude the well-posedness of the strong and UW variational formulations for the convection-reaction problem, we still need to show that the operator A is bounded below, and its adjoint A^* is injective. This requires *additional* assumptions on the advection and reaction coefficients. A classical approach using characteristics is presented in Appendix A, and an alternative approach based on Friedrichs-like inequality is shown in Appendix B.

REMARK 5 In formulation (2.2), we test with test functions vanishing on the outflow boundary Γ_+. If we remove this assumption, and test with functions from the whole energy space $H_{A^*}(\Omega)$, we need to introduce an extra unknown – trace $\hat{u}$ that lives in the trace space $\gamma_+(D(A))$:

$$\hat{U} := \{\hat{u} \,:\, \exists\, u \in D(A) \,:\, \gamma_+ u = \hat{u}\} .$$

The corresponding formulation is then as follows:

$$\begin{cases} u \in L^2(\Omega),\ \hat{u} \in \hat{U} \\ (u, A^*v) + \langle b_n \hat{u}, v\rangle_{\Gamma_+} = l(v) \quad v \in H_{A^*}(\Omega) . \end{cases} \tag{2.7}$$

The problem is well-posed, and it constitutes a special case of the broken formulation discussed next. ∎

3 Broken Variational Formulation

As usual for the DPG method, we define now the mesh skeleton and introduce the broken test space.

Mesh skeleton.

$$\Gamma_h := \bigcup_{K\in\mathcal{T}_h} (\partial K - \partial K_0)$$

where ∂K_0 denotes the no-flow part of element boundary ∂K.

Broken test space.

$$V(\mathcal{T}_h) = \{v \in L^2(\Omega) : A_h^* v \in L^2(\Omega)\} \qquad \|v\|^2_{V(\mathcal{T}_h)} = \|v\|^2 + \|A_h^* v\|^2$$

where A_h^* is the operator A^* applied element-wise.

We will follow precisely [5] to establish the well-posedness of the broken variational formulation.

Lemma 4
(*Duality Lemma*)
Let v be the solution of the element variational Neumann problem,

$$\begin{cases} v \in H_{A^*}(K) \\ (A^*v, A^*\delta v)_K + (v, \delta v)_K = \displaystyle\int_{\partial K} b_n \hat{u}\, \delta v \qquad \delta v \in H_{A^*}(K)\,. \end{cases}$$

Then $w = -A^*v$ is the solution to the Dirichlet problem,

$$\begin{cases} w \in H_A(K),\ w = \hat{u} \text{ on } \partial K - \partial K_0 \\ (Aw, A\delta w)_K + (w, \delta w)_K = 0 \qquad \delta w \in H_{A^*(K)}\,. \end{cases}$$

and

$$\|w\|_{H_A(\Omega)} = \|v\|_{H_{A^*}(\Omega)}\,.$$

∎

Proof: Integrating by parts, we obtain,

$$(AA^*v, \delta v)_K + (v, \delta v)_K - \int_{\partial K} b_n A^* v\, \delta v = \int_{\partial K} b_n \hat{u} \delta v \quad \forall \delta v\,.$$

This leads to the equation,

$$AA^*v + v = 0\,,$$

accompanied by BC:

$$\underbrace{-A^*v}_{=w} = \hat{u}\,.$$

Applying operator $-A^*$ to the equation, we get

$$A^*A(\underbrace{-A^*v}_{=w}) + \underbrace{-A^*v}_{=w} = 0,$$

which gives the final result. ∎

REMARK 6 The proof above is semi-formal. It may be replaced with a precise argument using the duality theory [16]. ∎

Duality Lemma 4 leads to the energy setting for traces.

$$\hat{U} := \{\hat{u} = \{\hat{u}_K\} \in \prod_{K\in\mathcal{T}_h} \gamma_{\partial K} H_A(K) \ : \ \text{there exists } u \in H_A(\Omega) \text{ such that } \gamma_{\partial K} u|_K = \hat{u}_K\} \tag{3.8}$$

with the minimum energy extension norm,

$$\|\hat{u}\|^2_{\hat{U}} := \sum_{K\in\mathcal{T}_h} \|w_K\|^2_{H_A(K)}$$

where $w_K \in H_A(K)$ is the minimum energy extension of $\hat{u}_K$. Above, $\gamma_{\partial K}$ denotes the (element) trace operator taking $H_A(K)$ onto its image in the weighted $L^2_w(\partial K)$ space, with the appropriately defined weight as in Section 2.

Lemma 5
(Compatibility Lemma)
Let $v \in V(\mathcal{T}_h)$ be a broken test function. Then

$$v \in D(A^*) \quad \Leftrightarrow \quad \langle \hat{u}, v\rangle_{\Gamma_h} := \sum_{K\in\mathcal{T}_h} \int_{\partial K} b_n \hat{u}_K \, v = \sum_{K\in\mathcal{T}_h} \int_{\partial K} b_n u \, v = 0$$
$$\text{for every } \hat{u} \in \hat{U} \text{ such that the corresponding global } u \in D(A);\ i.e., \hat{u} = 0 \text{ on } \Gamma_- \,. \tag{3.9}$$

∎

Proof: The result is a consequence of density result (2.5).
($\Rightarrow$) The functional on the right represents a continuous functional in $\hat{u} \in \hat{U}$ and, by the definition of $\hat{U}$, in the global $u \in D(A)$ corresponding to $\hat{u}$. We define a distributional derivative $g = -\operatorname{div}(bv)$ by

$$\int_\Omega bv \operatorname{grad}\phi = \int_\Omega g\phi \qquad \forall \phi \in C_0^\infty(\Omega)\,.$$

Employing $\phi \in C_0^\infty(K)$, we conclude that $-\operatorname{div}(bv|_K) = g|_K \in L^2(K)$. Integrating by parts, we obtain:

$$\sum_{K\in\mathcal{T}_h} \int_{\partial K} b_n v\phi = 0 \qquad \forall \phi \in C_0^\infty(\Omega)\,.$$

This implies that, for a general $\phi \in D(A)$, the sum over the elements reduces to the global boundary integral,

$$\sum_{K\in\mathcal{T}_h} \int_{\partial K} b_n v\phi = \int_\Gamma b_n v\phi = \int_{\Gamma_-} b_n v\phi\,.$$

Let $\phi_n \in C^\infty(\bar{\Omega})$ be now a sequence converging to $u \in D(A)$ in the graph norm. By the continuity of the trace operator, the last integral vanishes.
($\Leftarrow$) Reverse the procedure. Apply first $u = \phi \in C_0^\infty(\Omega)$ to learn that $-\operatorname{div}(bv)$ is an L^2-function and, therefore, the broken v is actually in the unbroken adjoint graph energy space, $v \in H_{A^*}(\Omega)$. Then test with general $u = \phi$ to learn that v vanishes on the outflow boundary. ∎

Broken UW variational formulation.

$$\begin{cases} u \in L^2(\Omega) \\ \hat{u} \in \hat{U},\ \hat{u} = u_0 \text{ on } \Gamma_- \\ (u, A_h^* v) + \langle \hat{u}, v \rangle_{\Gamma_h} = l(v) \qquad v \in H_{A^*}(\mathcal{T}_h)\,, \end{cases} \tag{3.10}$$

where $l \in (H_{A^*}(\mathcal{T}_h))'$. The load of interest is

$$l(v) = (f, v)$$

with $f \in L^2(\Omega)$. The functional setting allows for adding additional boundary or interface integrals to the load.

THEOREM 4
The bilinear form in (3.10) satisfies the inf-sup condition with a mesh-independent constant,

$$\gamma_{\mathcal{T}_h} \geq \big(2 + 3(\alpha^{-2} + 1)\big)^{-\frac{1}{2}}\ .$$

The broken UW variational formulation, is well-posed. ∎

Proof: Let[5]

$$l(v) := (u, A_h^* v) + \langle \hat{u}, v \rangle_{\Gamma_h}$$

Testing with $v \in H_{A^*}(\Omega)$, using Lemma 5 and Theorem 2, we obtain

$$(\alpha^{-2} + 1)^{-\frac{1}{2}} \|u\| \leq \sup_{v \in D(A^*)} \frac{|(u, A^* v)|}{\|v\|_{H_{A^*}(\Omega)}} = \sup_{v \in D(A^*)} \frac{|(u, A^* v) + \langle \hat{u}, v \rangle_{\Gamma_h}|}{\|v\|_{H_{A^*}(\mathcal{T}_h)}} = \|l\|_{(H_{A^*}(\mathcal{T}_h))'}\,.$$

Rearranging terms,

$$\langle \hat{u}, v \rangle_{\Gamma_h} = l(v) - (u, A_h^* v)\,.$$

by the duality lemma, we obtain

$$\|\hat{u}\|_{\hat{U}} = \sup_{v \in H_{A^*}(\mathcal{T}_h))'} \frac{|\langle \hat{u}, v \rangle_{\Gamma_h}|}{\|v\|_{H_{A^*}(\mathcal{T}_h)}} \leq \|l\| + \|u\| \leq (1 + (\alpha^{-2} + 1)^{\frac{1}{2}}) \|l\|\,.$$

[5] We are overloading symbol $l(v)$. This $l(v)$ is merely a notational shortcut for the right-hand side.

This leads to the final estimate,

$$\|u\|^2 + \|\hat{u}\|^2_{\hat{U}} \leq \left(2 + 3(\alpha^{-2} + 1)\right) \|l\|^2 .$$

The conjugate operator is injective. Indeed, assume

$$(u, A_h^* v) + \langle \hat{u}, v \rangle_{\Gamma_h} = 0 \quad \forall u \in L^2(\Omega), \forall \hat{u} \in \hat{U} .$$

By Lemma 5,

$$\langle \hat{u}, v \rangle_{\Gamma_h} = 0 \quad \forall \hat{u} \in \hat{U} \quad \Rightarrow \quad v \in D(A^*) .$$

Choosing $u = A_h^* v = A^* v$, we obtain $A^* v = 0$ and the boundedness below of A^* implies $v = 0$. Thus, by the Babuška-Nečas Theorem, the problem is well-posed. ∎

REMARK 7 The continuity constant for the broken bilinear form, $M \leq \sqrt{2}$. Indeed,

$$|(u, A_h^* v) + \langle \hat{u}, v \rangle_{\Gamma_h}| \leq \|u\| \, \|v\|_{H_{A^*}(\mathcal{T}_h)} + \|\hat{u}\|_{\hat{U}} \, \|v\|_{H_{A^*}(\mathcal{T}_h)} \leq \sqrt{2} \left(\|u\|^2 + \|\hat{u}\|^2_{\hat{U}} \right)^{\frac{1}{2}} \, \|v\|_{H_{A^*}(\mathcal{T}_h)} .$$

The stability properties of the ideal DPG method are thus independent of the mesh. ∎

4 Results for Problems from [2]

We begin our numerical studies by reproducing results for examples from [2]. All numerical results presented here and in the rest of the paper, are obtained with the *practical DPG method* in which the exact test space is replaced with an *enriched test space* - polynomials of order $p + \Delta p$, where p is the order of polynomials for the discrete trial space.

Integration of load vectors for L^2-projection as well as projection-based interpolation was done with adaptive integration [10]. The same adaptive integration was used for computing the error. We report the relative L^2-error for the field, using two versions of DPG: DPGc using continuous traces, and DPGd using discontinuous traces, and compare it with the L^2-projection error. We start with a mesh consisting of two triangles shown in Fig. 2. As in [2], we use quadratic polynomials for traces and linear polynomials for the fields. The presented results were obtained using enriched test functions of order $p = 2$; compared with order $p = 1$ for the fields, the increment in order is $\Delta p = 1$. The first example deals with a smooth solution and two advection fields; the first advection field is aligned with the mesh, while the second is not. In the first case, shown on the left side of Fig. 3, the three curves overlap each other: both versions of DPG deliver exactly the L^2-projection. For the advection field skewed to the mesh, shown on the right side of Fig. 3, the DPGd method delivers slightly better results that DPGc, especially on coarse meshes.

The second example, shown in Fig. 4, deals with a discontinuous solution. Here, the DPGd method performs much better than DPGc. For the advection field aligned with the grid, DPGd results overlap with the L^2-projection whereas DPGc lags significantly behind with a much slower rate of convergence. For

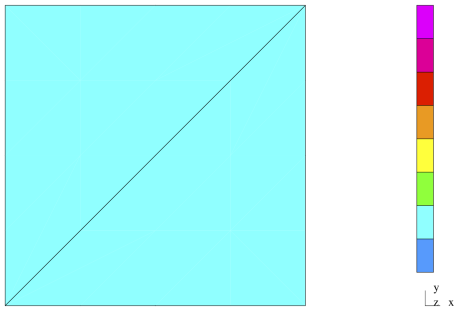

Figure 2: Initial mesh of two elements of second order. The scale on the right defines the polynomial order.

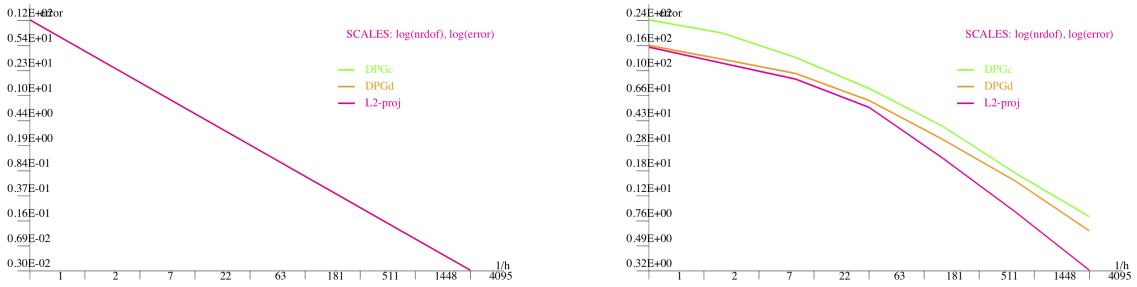

Figure 3: Example 1 from [2]: $f(x) = 1 - x_1$ and $b = (1, 1)$ (left), $b = (1, 1/16)$ (right). Comparison of DPG with continuous and discontinuous traces vs L^2-projection.

the advection field skewed to the mesh, the DPGd results no longer overlap with the L^2-projection but the results are significantly better than for DPGc.

In conclusion, our results are identical with those presented in [2].

REMARK 8 As usual, we use the logic of the exact sequence to fix the polynomial order for different variables. The reported results correspond to elements of order $p = 2$ and enriched spaces with $\Delta p = 1$, i.e. $r = 3$. This means that unknown u is discretized with polynomials from $\mathcal{P}^1$ and traces are discretized with polynomials of order $p = 2$, either continuous or discontinuous. This means that DPGd is using *more*

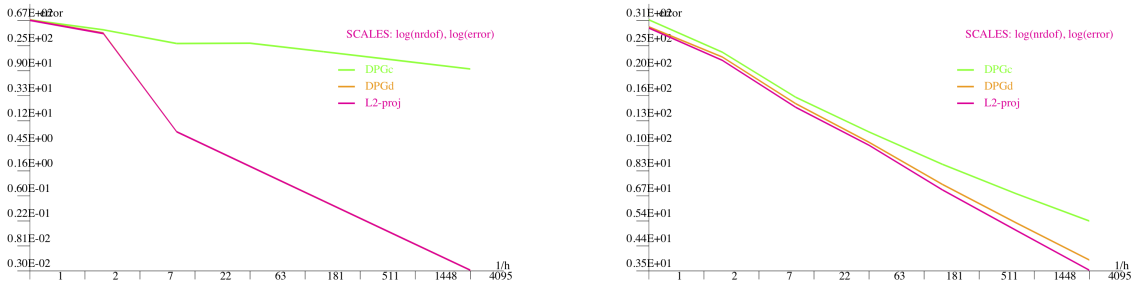

Figure 4: Example 2 from [2]: discontinuous f from (5.4), and $b = (1, 1)$ (left), $b = (1/1/16)$ (right). Comparison of DPG with continuous and discontinuous traces vs L^2-projection.

functions to approximate the traces than DPGc; the space of non-conforming traces contains the subspace of conforming ones (an observation critical for the subsequent convergence analysis). The default order for discontinuous polynomials on the skeleton is $p = 1$ (traces of Raviart-Thomas elements) and we had to increase it through a p-refinement of all edge nodes. An attempt to use lower-order traces resulted in a loss of the optimal rate of convergence. For $p = 1$ and some of the considered examples, the method did not converge at all. ∎

5 Local Fortin Operator for the Conforming DPG Method

In this section we pursue the local construction of a Fortin operator under the additional assumption that convection field b is constant element-wise. This, e.g., would be the case for 2D triangular meshes, and $b = \nabla \times w_h$ where w_h is any piece-wise linear, globally continuous function, or for 3D simplicial meshes with w_h coming from the lowest order Nédélec space. Intuitively, for sufficiently fine meshes, advection and reaction fields are 'almost constant' element-wise. In the end, the presented construction is just another 'sanity check'. The construction is based on ideas from [15]. For simplicity of presentation, we present the arguments in two space dimensions, $N = 2$.

Construction of the local Fortin operator is a standard step in concluding the discrete stability of the *practical DPG method* from the stability of the *ideal DPG method*. Our result in this section is negative: we find that we cannot construct the operator in a manner that is robust to the angle of elements relative to the convective direction. Because each step along the way is dictated to us by the requirements of the operator (which makes the construction somehow unique), we tentatively conclude that, for the convection-reaction problem, the discrete stability of the DPG method cannot be established by means of the local

Fortin operator.

Let K be an arbitrary triangular element. Without losing any further generality, we can assume that $b = (1, 0)$. We also assume a global bound on c, $c \leq c_{\max}$. The element contribution to the bilinear form is:

$$(u, -\partial_x v + cv)_K + \langle n_x \hat{u}, v\rangle_{\partial K} = (\partial_x u + cu, v)_K + \langle n_x(\hat{u} - u), v\rangle_{\partial K} .$$

The minimal element orthogonality conditions for the Fortin operator Π and case $c = 0$ are thus:

$$\begin{aligned} &(\psi, \Pi v - v)_K = 0 \qquad && \psi \in \partial_x(\mathcal{P}^{p-1}(K)) = \mathcal{P}^{p-2}(K), \\ &\langle n_x \phi, \Pi v - v\rangle_{\partial K} = 0 \qquad && \phi \in \mathcal{P}^p(K) . \end{aligned}$$

In order to simplify the presented analysis (for $c = 0$) and accommodate the more general case with $c \neq 0$, we will upgrade the first condition to polynomials of one order higher; we then require

$$\begin{aligned} &(\psi, \Pi v - v)_K = 0 \qquad && \psi \in \mathcal{P}^{p-1}(K), \\ &\langle n_x \phi, \Pi v - v\rangle_{\partial K} = 0 \qquad && \phi \in \mathcal{P}^p(K) . \end{aligned} \tag{5.11}$$

Taking $\psi = -\partial_x \chi$, $\chi \in \mathcal{P}^p(K)$, substituting into $(5.11)_1$, integrating by parts and utilizing $(5.11)_2$, we learn that

$$(\chi, \partial_x(\Pi v - v))_K = 0 \qquad \chi \in \mathcal{P}^p(K) . \tag{5.12}$$

This leads to the idea of defining $\partial_x \Pi v$ by L^2-projection,

$$\frac{1}{2}\|\partial_x(\Pi v - v)\|^2_{L^2(K)} \to \min_{\Pi v \in \mathcal{P}^r(K)} .$$

or, equivalently,

$$\begin{cases} \Pi v \in \mathcal{P}^r(K) \\ (\chi, \partial_x(\Pi v - v))_K = 0 \qquad \chi \in \mathcal{P}^{r-1}(K) . \end{cases} \tag{5.13}$$

In order to secure satisfaction of (5.12), we need to assume that $r - 1 \geq p$;i.e., $r \geq p + 1$. We have immediately,

$$\|\partial_x \Pi v\|_{L^2(K)} \leq \|\partial_x v\|_{L^2(K)} \leq \| - \partial_x v + cv\|_{L^2(K)} + c_{\max}\|v\|_{L^2(K)} \leq \sqrt{1 + c^2_{\max}}\|v\|_{H_{A^*}(K)} .$$

REMARK 9 Is the use of the L^2-projection necessary? Not really. Condition (5.12) implied by orthogonality conditions (5.11) is necessary. We need to complete it with additional conditions to make $\partial_x \Pi v$ unique, and we have to do it in such a way that $\partial_x \Pi v$ will depend *only* upon $\partial_x v$. Only then we get the right scaling properties[6] and can conclude

$$\|\partial_x \Pi v\|_{L^2(K)} \leq C\|\partial_x v\|_{L^2(K)} ,$$

with some h-independent constant C. The use of the L^2-projection is thus a natural choice but it is not necessary. ∎

[6]Recall derivation of interpolation error estimates for the exact sequence spaces and the need for the commutativity property.

p	1	2	3	4	5	6
r	2	3	4	6	8	10

Table 1: Minimal enriched order r resulting from the local construction of Fortin operator for different polynomial orders of discretization.

Having fixed $\partial_x \Pi v$, we still have not fixed Πv itself. More precisely, Πv has been defined so far up to polynomials that are independent of x, i.e. the subspace

$$\mathcal{P}^r_y(K) := \text{span}\{1, y, \dots, y^r\}, \qquad \dim \mathcal{P}^r_y(K) = r+1.$$

We are presented with the task of defining the undefined $\mathcal{P}^r_y(K)$-component of Πv in such a way that we satisfy orthogonality conditions (5.11). First of all, we claim that it is sufficient to satisfy only condition $(5.11)_1$. Indeed, integration by parts reveals that conditions $(5.11)_1$ and (5.12) imply $(5.11)_2$.

In order to estimate the minimum enriched order r, we begin with a simple counting argument comparing the number of additional orthogonality conditions that need to be satisfied with the number of remaining unknowns (dimension of $\mathcal{P}^r_y(K)$). We need to distinguish between the case: $p = 1, 2$ and case: $p \geq 3$.

For $p \geq 3$, the subspace of bubbles $\mathcal{P}^p_0(K)$ is non-empty. Using $\chi \in \mathcal{P}^p_0(K)$ in (5.13), and integrating by parts, we get,

$$(\partial_x \chi, \Pi v - v)_K = 0 \qquad \chi \in \mathcal{P}^p_0(K)\,.$$

The null space of linear transformation $\partial_x : \mathcal{P}^p_0(K) \to \mathcal{P}^{p-1}(K)$ is trivial which implies that

$$\dim \partial_x(\mathcal{P}^p_0(K)) = \dim \mathcal{P}^p_0(K) = \frac{(p-2)(p-1)}{2}\,.$$

As $\dim \mathcal{P}^{p-1}(K) = \frac{p(p+1)}{2}$, we are missing $\frac{p(p+1)}{2} - \frac{(p-2)(p-1)}{2} = 2p-1$ conditions. This results in the condition for the minimal enriched order r,

$$r+1 \geq 2p-1 \quad \Leftrightarrow \quad r \geq 2p-2\,.$$

For $p \leq 2$, the space of bubbles is trivial, so we need to satisfy:

$$r+1 \geq \dim \mathcal{P}^{p-1}(K) = \frac{p(p+1)}{2}\,.$$

Table 1 presents the resulting minimum value of enriched order r for different polynomial orders p. As we can see, except for low $p = 1, 2, 3$, the values are very pessimistic. We emphasize that they reflect only the deficiency of the local construction of the Fortin operator.

We complete now the definition of Πv by requesting the satisfaction of the orthogonality conditions. Consider first the case of $p > 2$ and $r = 2p-2$. In this case,

$$\dim \mathcal{P}^r_y(K) + \dim(\partial_x \mathcal{P}^p_0(K)) = \dim \mathcal{P}^{p-1}(K)\,.$$

The following lemma establishes the existence of a stable right-inverse of the derivative ∂_x, necessary for the formulatiom of problem (5.14) completing the definition of the Fortin operator.

Lemma 6

Let K be a rotated unit master triangle. There exists a continuous right-inverse of derivative ∂_x,

$$R : \mathcal{P}^p(K) \to \mathcal{P}^{p+1}(K), \quad \partial_x R\phi = \phi \quad \forall \phi \in \mathcal{P}^p(K)$$
$$\|R\phi\|_{L^2(K)} \leq \sqrt{2}\|\phi\|_{L^2(K)} .$$

∎

Proof: The triangle is illustrated in Fig.5. At least one of the bounds $x_{\min}(y), x_{\max}(y)$ is a linear (and

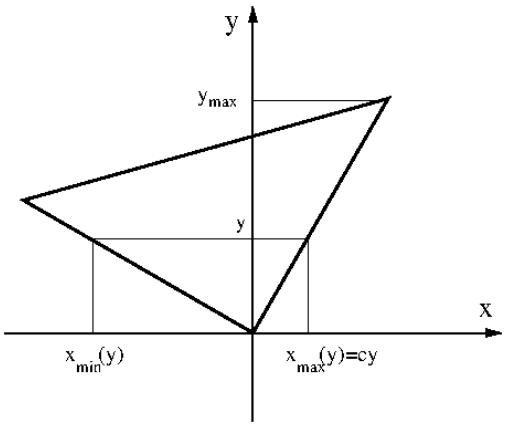

Figure 5: A rotated master triangle

not piece-wise linear) function of y. In the illustrated case, $x_{\max}(y) = cy$. We define the right-inverse as,

$$(R\phi)(x,y) = \int_x^{x_{\max}(y)=cy} \phi(s,y)\, ds .$$

Note that operator R is well-defined. Indeed, let $\phi(x,y) = x^\alpha y^\beta$. Then

$$\int_x^{cy} s^\alpha y^\beta \, ds = \frac{1}{\alpha+1} s^{\alpha+1}|_x^{cy} y^\beta = \frac{1}{\alpha+1}((cy)^{\alpha+1} - x^{\alpha+1})y^\beta$$

is indeed a polynomial of order $\alpha+\beta+1$. A standard estimation follows.

$$\begin{aligned}\int_0^{y_{\max}}\int_{x_{\min}(y)}^{x_{\max}(y)}\left|\int_x^{cy}\phi(s,y)\,ds\right|^2 dx\,dy &\leq \int_0^{y_{\max}}\int_{x_{\min}(y)}^{x_{\max}(y)}\int_x^{cy}|\phi(s,y)|^2\,ds(cy-x)\,dx\,dy\\ &\leq \int_0^{y_{\max}}\int_{x_{\min}(y)}^{x_{\max}(y)}\int_{x_{\min}(y)}^{x_{\max}(y)}|\phi(s,y)|^2\,ds(cy-x)\,dx\,dy\\ &\leq \int_0^{y_{\max}}\frac{(x_{\max}(y)-x_{\min}(y))^2}{2}\int_{x_{\min}(y)}^{x_{\max}(y)}|\phi(s,y)|^2\,ds\,dy\\ &\leq \sqrt{2}\int_0^{y_{\max}}\int_{x_{\min}(y)}^{x_{\max}(y)}|\phi(s,y)|^2\,ds\,dy.\end{aligned}$$

∎

Let now $v^r = R(\partial_x \Pi v) \in \mathcal{P}^r(K)$. We set up the following system of equations for component $v_y^r \in \mathcal{P}_y^r(K)$.

$$\left\{\begin{array}{l} v_y^r \in \mathcal{P}_y^r(K) \\ (\psi, v_y^r + v^r - v)_K = 0 \quad \psi \in \mathcal{P}^{p-1}(K)\,. \end{array}\right. \tag{5.14}$$

We introduce the discrete inf-sup constant corresponding to the bilinear form (5.14),

$$\alpha := \inf_{v_y^r \in \mathcal{P}_y^r(K)} \sup_{\psi \in \mathcal{P}^{p-1}(K))} \frac{(\psi, v_y^r)}{\|\psi\|_{L^2(K)}\,\|v_y^r\|_{L^2(K)}}. \tag{5.15}$$

This leads to the L^2-stability bound on the master element,

$$\begin{aligned}\|\hat{v}_y^r\|_{L^2(\hat{K})} &\leq \alpha^{-1}\|\hat{v}^r - \hat{v}\|_{L^2(\hat{K})}\\ &\leq \alpha^{-1}\left(\|\hat{v}^r\|_{L^2(\hat{K})} + \|\hat{v}\|_{L^2(\hat{K})}\right)\end{aligned}$$

and, consequently,

$$\begin{aligned}\|\hat{v}_y^r + \hat{v}^r\|_{L^2(\hat{K})} &\leq (\alpha^{-1}+1)\|\hat{v}^r\|_{L^2(\hat{K})} + \alpha^{-1}\|\hat{v}\|_{L^2(\hat{K})}\\ &\leq (\alpha^{-1}+1)\sqrt{2}\|\partial_\xi(\hat{\Pi}\hat{v})\|_{L^2(\hat{K})} + \alpha^{-1}\|\hat{v}\|_{L^2(\hat{K})}\,.\end{aligned}$$

A standard scaling argument yields then:

$$\begin{aligned}\|v_y^r + v^r\|_{L^2(K)} &\leq h\|\hat{v}_y^r + \hat{v}^r\|_{L^2(\hat{K})}\\ &\leq (\alpha^{-1}+1)\sqrt{2}h^2\|\partial_x v\|_{L^2(K)} + \alpha^{-1}\|v\|_{L^2(K)}\,.\end{aligned}$$

Above, as usual, $\hat{v}$ denotes the pullback of v to master element $\hat{K}$, and ξ denotes the first dimension on the master element. Supposing α is uniformly bounded away from 0 for all angles of rotation, this concludes the proof of boundedness of the Fortin operator in the $H_{A^*}(K)$-norm, with an h-independent continuity constant. The touchy issue with the presented construction is exactly the dependence of inf-sup constant α upon the orientation of the element with respect to the advection field. We will resort now to a numerical experiment to study this dependence.

Dependence of inf-sup constant α upon the orientation of the element. Computation of the inf-sup constant α translates into the determination of the smallest eigenvalue for the generalized eigenvalue problem:

$$B^T G^{-1} B u = \alpha^2 M u$$

where

$$\begin{aligned} G_{ij} &= \int_K \psi_i \psi_j \qquad i,j = 1,\ldots,\dim \mathcal{P}^r(K) \\ B_{jk} &= \int_K \psi_j y^k \qquad j = 1,\ldots,\dim \mathcal{P}^r(K),\ k = 1, r+1 \\ M_{ik} &= \int_K y^i y^k \qquad j,k = 1,\ldots,r+1 \end{aligned}$$

Fig. 6 presents value of α for element of order $p = 3$ and angle changing from 0 to 2π. As we can see, whenever one of the triangle edges becomes parallel to the x-axis, the constant becomes zero. In other words, matrix B becomes singular. The trouble is not with the zero values. This could be fixed by avoiding linear dependence of the equations. The trouble is with the degeneration of α as one of the edges approaches the x axis. Evidently, constant α is *not* uniformly (in angle) bounded away from zero. The result does not prove that the DPG method is unstable. It reflects the limitation of the local construction of the Fortin operator.

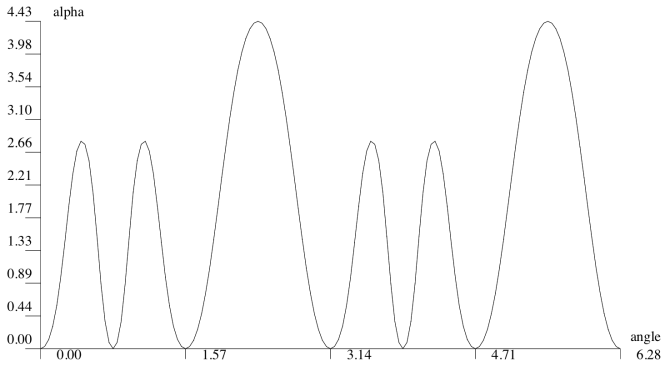

Figure 6: Value of inf-sup constant α for angles between 0 and 2π, for $p = 3$.

6 Discrete Stability

Having convinced ourselves about the limitations of the local Fortin operators, we will instead now attempt to prove global discrete stability directly. It is easy to show that discrete stability is equivalent to the existence of a *global Fortin operator*; see [17], as well as the reasoning in Appendix C. We will emulate the stability analysis for the broken test spaces on the continuous level, proving first that we control the L^2-norm of field u_h and continue with a stability and convergence analysis for traces at the end of Section 7.

Similarly to the analysis in [11], we divide the domain and the mesh into layers $\Omega_{h,1},\ldots,\Omega_{h,N}$ defined in a recursive way starting from the outflow boundary:

$$\begin{aligned}\Omega_{h,1} &:= \bigcup\{K\in\mathcal{T}_h\,:\,\partial K_+\subset\Gamma_+\}\\ \Omega_{h,n} &:= \bigcup\{K\in\mathcal{T}_h\,:\,\partial K_+\subset\Gamma_+\cup\Gamma_{h,-,n-1}\},\quad n=2,\ldots,N\end{aligned}$$

where $\Gamma_{h,-,n}$ denotes the inflow part of the boundary of $\Omega_{h,n}$.

Let u_h now be a discrete trial function, i.e., a piecewise polynomial of order $p-1$. We define a corresponding conforming (globally continuous) test function v_h, a piecewise polynomial of order $p+1$ as follows. We start with elements K from the first layer $\Omega_{h,1}$. For each element $K\subset\Omega_{h,1}$, we determine a function $v\in H_{A^*}(K)$ such that

$$A^*v = u_h,\quad v=0 \text{ on } \partial K_+\,.$$

Given the function v, we construct its element approximation v_h via the constrained minimization problem analyzed in Appendix C with weights w,a to be specified in a moment. We have

$$\int_K u_h^2 = \int_K u_h A^*v = \int_K u_h A^*v_h \tag{6.16}$$

and function v_h vanishes on the outflow boundary. We proceed next with elements from the second layer $\Omega_{h,2}$. For each element $K\subset\Omega_{h,2}$, we determine a function $v\in H_{A^*}(K)$ such that

$$A^*v = u_h,\quad v=v_h \text{ on } \partial K_+\,,$$

and replace it with its polynomial approximation v_h. Condition (6.16) stays satisfied, and the local construction of v_h implies that v_h is globally continuous. We continue by recursion, obtaining a globally continuous test function v_h that vanishes on Γ_+ and

$$\int_\Omega u_h^2 = \int_\Omega u_h A^*v_h\,.$$

In order to prove discrete stability, it is sufficient to show the existence of a mesh-independent constant C such that

$$\|A^*v_h\|\le C\|u_h\|\,.$$

Note that, with $v_h\in D(A^*)$, the adjoint norm and the adjoint graph norm of v_h are equivalent.

We will try to emulate the global stability analysis presented in Appendix B. For each element K from the last layer, $K \subset \Omega_{h,N}$,

$$\int_{\partial K_-} e^{2V}|b_n|v_h^2 + \int_K \frac{e^{2V}}{2a}(A^* v_h)^2 \leq \int_K \frac{e^{2V}}{a}(A^* v_h)^2 + \int_{\partial K_+} e^{2V} b_n v_h^2 \,, \tag{6.17}$$

and the following estimate holds,

$$\int_K \frac{e^{2V}}{a}(A^* v_h)^2 + \int_{\partial K_+} e^{2V} b_n v_h^2 \leq (1+\alpha_h^{-2}) \int_K \frac{e^{2V}}{a}(\underbrace{A^* v}_{u_h})^2 + \beta_h^{-2} \int_{\partial K_+} e^{2V} b_n \underbrace{v^2}_{=v_h^2},$$

where the element stability constants α_h, β_h depend upon the weights $w = e^{2V}$ and $a = |b|^2 + c - \operatorname{div} b$. This yields

$$\int_{\partial K_-} e^{2V}|b_n|v_h^2 + \int_K \frac{e^{2V}}{2a}(A^* v_h)^2 \leq (1+\alpha_h^{-2}) \int_K \frac{e^{2V}}{a} u_h^2 + \beta_h^{-2} \int_{\partial K_+} e^{2V} b_n v_h^2 \,.$$

An identical inequality holds for elements from layer $\Omega_{h,N-1}$ (with element-dependent stability constants). Let $K \subset \Omega_{h,N-1}$. Let $\beta_{h,\min}$ be the minimum stability constant for elements from layer $\Omega_{h,N}$ sharing (part of) the inflow boundary ∂K_-. Multiplying the estimate above with $\beta_{h,\min}^{-2}$, we get

$$\beta_{h,\min}^{-2} \int_{\partial K_+} e^{2V}|b_n|v_h^2 + \beta_{h,\min}^{-2} \int_K \frac{e^{2V}}{2a}(A^* v_h)^2 \leq (1+\alpha_h^{-2})\beta_{h,\min}^{-2} \int_K \frac{e^{2V}}{a} u_h^2 + \beta_{h,\min}^{-2}\beta_h^{-2} \int_{\partial K_+} e^{2V} b_n v_h^2 \,.$$

Note that $\beta_h \leq 1$. Summing up over all elements, we obtain the estimate

$$\int_\Omega \frac{e^{2V}}{2a}(A^* v_h)^2 = \sum_K \int_K \frac{e^{2V}}{2a}(A^* v_h)^2 \leq \sum_K (1+\alpha_h^{-2}) \prod \beta_{h,\min}^{-2} \int_K \frac{e^{2V}}{a} u_h^2,$$

where $\prod \beta_{h,\min}^{-2}$ on the right-hand side is the product of up to $N-1$ stability constants for elements 'flowing into' element K.

It becomes clear that the global stability argument requires constants β_h^{-2} to be bounded by $(1+Ch)$ where h is the element size, and constant α_h to be bounded away from zero. For quasi-uniform meshes $h \approx 1/N$ and the constant

$$(1+Ch)^N \approx \left(1 + \frac{C}{N}\right)^N \leq e^C$$

remains bounded as $h \to 0$ ($N \to \infty$).

We will limit ourselves to numerical experiments to study element stability constants α_h, β_h.

Computation of the inf-sup stability constant α_h. Let $\psi_i, i = 1, \ldots, N$ denote a basis for the polynomial space

$$\{v \in \mathcal{P}^{p+1}(K) \,:\, v = 0 \text{ on } \partial K_+\} \,,$$

p/h	1.0	0.1	0.01	0.001	0.0001
2	0.737	0.712	0.708	0.707	0.707
3	0.657	0.637	0.633	0.633	0.632
4	0.593	0.582	0.578	0.577	0.577
5	0.545	0.539	0.535	0.535	0.535
6	0.508	0.505	0.501	0.500	0.500
7	0.477	0.476	0.472	0.471	0.471
8	0.451	0.451	0.448	0.447	0.447

Table 2: Minimal (over angles) value of inf-sup constant α_h for different values of element size h and polynomial order p, for advection vector $b = (1, 0)$, and reaction coefficient $c = 1.0$.

and let $\phi_j, j = 1, \ldots, M$ be a basis for the polynomial space $\mathcal{P}^{p-1}(K)$. The computation of inf-sup constant α_h reduces to the solution of the generalized eigenvalue problem:

$$B^T G^{-1} B u = \alpha^2 M u$$

where

$$\begin{aligned} G_{ij} &= \int_K \frac{w}{a} A^* \psi_i \, : A^* \psi_j + \int_{\partial K_+} w b_n \psi_i \psi_j \qquad i,j = 1,\ldots,N \\ M_{ij} &= \int_K \phi_i \phi_j \qquad\qquad i,j = 1,\ldots,M \\ B_{ij} &= \int_K \phi_i A^* \psi_j \qquad\qquad i = 1,\ldots,M,\ j = 1,\ldots,N \end{aligned}$$

Table 2 presents numerical values of constant α_h for different values of polynomial order p and element size h. All values are the minimum values over rotation angles from the whole range of $[0, 2\pi)$. Clearly the inf-sup constant stays uniformly bounded away from zero and remains of order 1 in the whole range of polynomial orders p and element size h.

Computation of the inf-sup stability constant β_h. Let $\psi_i, i = 1, \ldots, N$ denote a basis for the polynomial space $\mathcal{P}^{p+1}(K)$ and let $\phi_j, j = 1, \ldots, M$ be a basis for the polynomial space $\mathcal{P}^p(\partial K_+)$. The computation of inf-sup constant β_h reduces to the solution of the generalized eigenvalue problem:

$$B^T G^{-1} B u = \alpha^2 M u$$

where

$$\begin{aligned} G_{ij} &= \int_K \frac{w}{a} A^* \psi_i \, A^* \psi_j + \int_{\partial K_+} w b_n \psi_i \psi_j \qquad i,j = 1,N \\ B_{ij} &= \int_{\partial K_+} w b_n \phi_i \psi_j \qquad\qquad i = 1,\ldots,M,\ j = 1,\ldots,N \\ M_{ij} &= \int_{\partial K_+} w b_n \phi_i \phi_j \qquad\qquad i,j = 1,\ldots M \end{aligned}$$

Fig. 7 presents results for the case of the rotated master triangle of order $p = 2$, advection vector $b = (1, 0)$ and reaction coefficient $c = 1$. For all triangles with just one outflow edge, the inf-sup constant is practically equal to one. Unfortunately, the results show a clear degeneration of stability for all triangles with two outflow edges. Consequently, our attempt to prove global discrete stability fails. A this point, we have

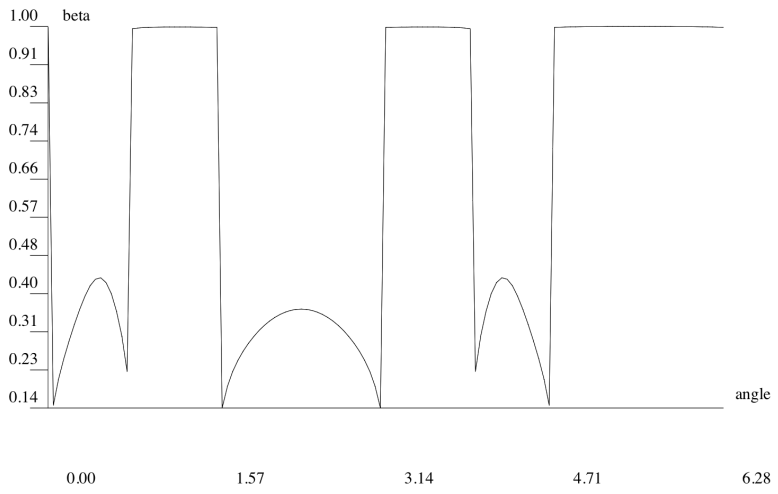

Figure 7: Values of the inf-sup constant β_h for the rotated unit triangle of order $p = 2$, $b = (1, 0), c = 1$, and test space constisting of polynomials of order $p + 1$.

convinced ourselves that the enriched polynomial space is insufficient, and in the next section, we turn to the original method from [11] using piecewise polynomial test spaces on each element constructed with a two subelement mesh.

7 Discrete Stability Analysis for the Composite Test Spaces

Construction of the composite enriched test space is illustrated in Fig. 8. In the illustrated case, the outflow boundary consists of two edges. The element is partitioned into two subelements using a line parallel to the advection vector in such a way that each of the subelements K_1, K_2 contains now only one outflow edge. The test space is then a standard H^1-conforming space of order $p + 1$ corresponding to the element sub-mesh.

Let K be an element from the last layer, $K \subset \Omega_{h,N}$. Let $A^* v = u_h, v = 0$ on ∂K_+, and let v_h be its

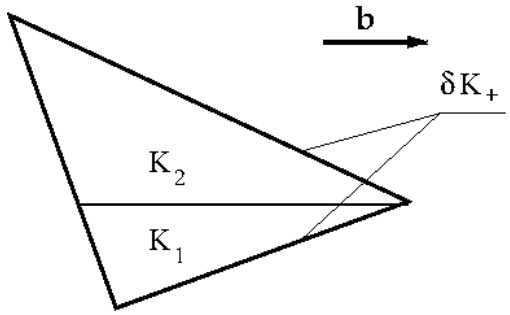

Figure 8: Construction of the piece-wise polynomial enriched test space.

(possibly) piecewise polynomial approximation discussed in Appendix C. We have,

$$\begin{aligned}\int_{\partial K_-} e^{2V}|b_n|v_h^2+\int_K e^{2V}av_h^2+\int_K \frac{e^{2V}}{a}|A_h^* v_h|^2 &\leq \int_K \frac{2e^{2V}}{a}|A_h^* v_h|^2+\int_{\partial K_+} e^{2V}b_n v_h^2 \\ &\leq (1+\alpha_h^{-2})\int_K \frac{2e^{2V}}{a}|\underbrace{A_h^* v}_{=u_h}|^2+\beta_h^{-2}\int_{\partial K_+} e^{2V}b_n \underbrace{v^2}_{=0},\end{aligned} \tag{7.18}$$

REMARK 10 Notice that, in contrast to the estimate (6.17), we have also included the L^2 term on the left-hand side. This is related to the fact that the global discrete (approximate) test function that we are constructing now is only *weakly conforming*; i.e. the jumps $[v_h]$ are orthogonal to discrete traces $\hat{u}_h$ but they do not vanish. We cannot use thus boundedness below of operator A^* to control the L^2-norm of v_h a-posteriori. ∎

Summing over all elements from the last layer, we obtain

$$\begin{aligned}\sum_{K\subset\Omega_{h,N}}\left\{\int_{\partial K_-} e^{2V}|b_n|v_h^2+\int_K e^{2V}av_h^2+\int_K \frac{e^{2V}}{a}|A_h^* v_h|^2\right\} &\leq \sum_{K\subset\Omega_{h,N}}\left\{\int_K \frac{2e^{2V}}{a}|A_h^* v_h|^2+\int_{\partial K_+} e^{2V}b_n v_h^2\right\} \\ &\leq (1+\alpha^{-2})\sum_{K\subset\Omega_{h,N}}\left\{\int_K \frac{2e^{2V}}{a}|u_h|^2\right\}\end{aligned}$$

where $\alpha>0$ is a lower bound for element inf-sup constants α_h. Consider now elements $K\subset\Omega_{h,N-1}$. The element outflow boundary ∂K_+ consists of edges e that are either contained in the global outflow boundary Γ_+, or are contained in the inflow boundary of elements belonging to the last layer $\Omega_{h,N}$. We consider solution v of the adjoint equation, $A^*v=u_h$, $v=v_0$ on ∂K_+ where

$$v_0=\begin{cases}0 & \text{if } \partial K_+\subset\Gamma_+\\ v_h^N & \text{otherwise}\end{cases}$$

with v_h^N denoting the discrete (approximate) test function from element(s) $K \subset \Omega_{h,N}$. We obtain

$$\sum_{K\subset\Omega_{h,N-1}} \left\{ \int_{\partial K_-} e^{2V}|b_n|v_h^2 \quad + \int_K e^{2V} a v_h^2 + \int_K \frac{e^{2V}}{a}|A_h^* v_h|^2 \right\} \leq \sum_{K\subset\Omega_{h,N-1}} \left\{ \int_K \frac{2e^{2V}}{a}|A_h^* v_h|^2 + \int_{\partial K_+} e^{2V} b_n v_h^2 \right\}$$

$$\leq \sum_{K\subset\Omega_{h,N-1}} \left\{ (1+\alpha^{-2}) \int_K \frac{2e^{2V}}{a} |\underbrace{A_h^* v}_{=u_h}|^2 + \beta^{-2} \int_{\partial K_+} e^{2V} b_n v^2 \right\}$$

where $\beta > 0$ is a lower bound for element inf-sup constants β_h.

We want now to add the two inequalities side-wise and cancel the first term in the first inequality with the last term in the second inequality (a telescoping effect). In order to do so, we have to premultiply the entire first inequality by factor β^{-2}. As discussed before, this leads to a multiplicative accumulation of constant β^{-2}. The product of such constants can be bounded by a mesh independent constant provided $\beta_h = 1 - O(h)$.

Conjecture. We postulate the following behavior of stability constants α_h, β_h under the assumption that weights e^{2V} and $\frac{2e^{2V}}{a}$ are uniformly bounded throughout the domain:

$$\beta_h \geq 1 - Ch, \quad \alpha_h > \alpha > 0\,,$$

for some generic, mesh-independent constants C, α.

THEOREM 5

Under the conjecture above, the discrete inf-sup condition holds:

$$\sup_{v_h \in V_h^0} \frac{\sum_K \int_K u_h A_h^* v_h}{\|v_h\|_{H_{A^*}}} \geq C\|u_h\|,$$

with a mesh-independent constant C. Above, V_h^0 stands for the subspace of weakly conforming broken test functions. ∎

We return now to the numerical experiments with constant β_h to support the conjecture. Table 3 presents the results. The constant stays very close to one, uniformly in the polynomial order, and it converges to one as $h \to 0$. At this point in the game, we realize that the enriched space order $p+1$ (implied by the local construction of the Fortin operator) may not be necessary and order p (same as for traces) may be sufficient. Table 4 presents the results for the lower order of the test space. The stability constants are worse but the overall trend of β_h converging to one with $h \to 0$ remains.

For the investigated case, we can actually easily prove the following lemma.

Lemma 7

Let b be a constant advection vector, reaction coefficient c and weight w be bounded from above, and weights

p/h	1.0	0.1	0.01	0.001	0.0001
2	0.99492667	0.99998878	0.99999998	0.99999999	0.99999999
3	0.99561802	0.99999021	0.99999998	0.99999999	0.99999999
4	0.99597293	0.99999096	0.99999999	0.99999999	0.99999999
5	0.99618684	0.99999143	0.99999999	0.99999999	0.99999999
6	0.99632916	0.99999174	0.99999999	0.99999999	0.99999999
7	0.99643040	0.99999197	0.99999999	0.99999999	0.99999999
8	0.99650598	0.99999214	0.99999999	0.99999999	0.99999999

Table 3: Composite test space of order $p+1$. Minimal (over angles) value of inf-sup constant β_h for different values of element size h and polynomial order p, for advection vector $b = (1, 0)$, and reaction coefficient $c = 1.0$, weights $a = w = 1$.

p/h	1.0	0.1	0.01	0.001	0.0001
2	0.86181281	0.97997501	0.99789133	0.99978799	0.99978799
3	0.87011447	0.98094156	0.99799185	0.99979809	0.99997979
4	0.87469774	0.98147798	0.99804769	0.99980370	0.99998035
5	0.87760601	0.98181914	0.99808322	0.99980727	0.99998071
6	0.87961588	0.98205525	0.99810782	0.99980974	0.99998096
7	0.88108800	0.98222835	0.99812586	0.99981155	0.99998114
8	0.88221274	0.98236070	0.99813965	0.99981293	0.99998128

Table 4: Composite test space of order p. Minimal (over angles) value of inf-sup constant β_h for different values of element size h and polynomial order p, for advection vector $b = (1, 0)$, and reaction coefficient $c = 1.0$, weights $a = w = 1$.

a, w from below:

$$|c(x)| \leq c_{\max} < \infty \quad 0 < w_{\min} \leq |w(x)| \leq w_{\max} < \infty \quad 0 < a_{\min} \leq a(x)$$

Then there exists a constant $C > 0$, such that $\beta_h \geq 1 - Ch$. ∎

Proof: Let w_h be a piece-wise polynomial of order p defined on element outflow boundary ∂K_+. Take $v_h = w_h$ on the outflow boundary ∂K_+ and lift it with constant values along the streamlines. The function lives in the composite test space. Moreover,

$$\|v_h\|_V^2 = \int_K \frac{w}{a} (\underbrace{b \cdot \boldsymbol{\nabla} v_h + c v_h}_{A_h^* v_h})^2 + \int_{\partial K_+} w b_n v_h^2 = \int_K \frac{w}{a} c^2 v_h^2 + \int_{\partial K_+} w b_n v_h^2 \leq \frac{c_{\max}^2}{a_{\min}} \int_K w v_h^2 + \int_{\partial K_+} w b_n v_h^2 \,.$$

The first term is of order h^2, the second one of order h. Finite dimensionality argument and the assumptions on the coefficients imply that there exists a constant C such that

$$\|v_h\|_V \leq (1 + Ch)^{1/2} \left(\int_{\partial K_+} w b_n v_h^2 \right)^{1/2} .$$

Consequently,

$$\beta_h \geq (1 + Ch)^{-1/2} \underbrace{\inf_{w_h} \sup_{v_h} \frac{\int_{\partial K_+} w b_n w_h v_h}{\|w_h\|_{w b_n}^2}}_{=1} .$$

But, for small h, $(1 + Ch)^{-1/2} \approx 1 - \frac{C}{2} h$, by a Taylor series argument. ∎

The second stability constant α_h already behaved 'nicely' for the polynomial test space. Increasing the test space to piecewise polynomials only makes it bigger.

Convergence of fields

Once we have established the stability for fields,

$$\gamma_h \|u_h\| \leq \sup_{v_h \in V_h^0} \frac{(u_h, A^* v_h)}{\|v_h\|_V} \tag{7.19}$$

where

$$V_h^0 := \{ v_h \in V_h \,:\, \langle \hat{w}_h, v_h \rangle_{\Gamma_h} = 0 \quad \forall \hat{w}_h \in \hat{U}_h \} \,,$$

we can easily show the convergence of the fields, for both conforming and non-conforming versions of the method. Let $\widetilde{V}_h \subset V_h$ denote the subspace of approximate optimal test functions.

$$
\begin{aligned}
\|u-u_h\| &\leq \|u-w_h\| + \|w_h-u_h\| \\
&\leq \|u-w_h\| + \gamma_h^{-1} \sup_{v_h \in V_h^0} \frac{(w_h-u_h, A^* v_h)}{\|v_h\|_V} && \text{(condition (7.19))} \\
&= \|u-w_h\| + \gamma_h^{-1} \sup_{v_h \in V_h^0} \frac{(w_h-u_h, A^* v_h) + \langle \hat{w}_h - \hat{u}_h, v_h\rangle_{\Gamma_h}}{\|v_h\|_V} && (\langle \hat{w}_h - \hat{u}_h, v_h\rangle_{\Gamma_h} = 0,\ v_h \in V_h^0) \\
&\leq \|u-w_h\| + \gamma_h^{-1} \sup_{v_h \in V_h} \frac{(w_h-u_h, A^* v_h) + \langle \hat{w}_h - \hat{u}_h, v_h\rangle_{\Gamma_h}}{\|v_h\|_V} && \text{(supremum taken over a bigger set)} \\
&= \|u-w_h\| + \gamma_h^{-1} \sup_{\widetilde{v}_h \in \widetilde{V}_h} \frac{(w_h-u_h, A^* \widetilde{v}_h) + \langle \hat{w}_h - \hat{u}_h, \widetilde{v}_h\rangle_{\Gamma_h}}{\|\widetilde{v}_h\|_V} && \text{(optimal test functions realize the sup)} \\
&\leq \|u-w_h\| + \gamma_h^{-1} \sup_{\widetilde{v}_h \in \widetilde{V}_h} \frac{(w_h-u, A^* \widetilde{v}_h) + \langle \hat{w}_h - \hat{u}, \widetilde{v}_h\rangle_{\Gamma_h}}{\|\widetilde{v}_h\|_V} && \text{(Galerkin orthogonality)} \\
&\leq (1+\gamma_h^{-1})\|u-w_h\| + \gamma_h^{-1} \sup_{v_h \in V_h} \frac{\langle \hat{w}_h - \hat{u}, v_h\rangle_{\Gamma_h}}{\|v_h\|_V}
\end{aligned}
$$

where $w_h, \hat{w}_h$ are, respectively, an arbitrary discrete field and trace. Note that, for the non-conforming version, the duality pairing has to be understood in the discrete sense:

$$
\langle \hat{w}_h, v_h\rangle_{\Gamma_h} = \sum_{K \in \mathcal{T}_h} \int_{\partial K} b_n \hat{w}_h v_h,
$$

and it makes sense only for discrete test functions v_h. Once we use the Galerkin orthogonality, it is replaced with the actual duality pairing, provided we assume that $\hat{w}_h$ comes from the conforming subspace $\hat{U}_h^c$ of space $\hat{U}_h$ of non-conforming traces. We can follow with the estimate,

$$
\sup_{v_h \in V_h} \frac{\langle \hat{w}_h - \hat{u}, v_h\rangle_{\Gamma_h}}{\|v_h\|_V} \leq \sup_{v \in V} \frac{\langle \hat{w}_h - \hat{u}, v\rangle_{\Gamma_h}}{\|v\|_V} = \|\hat{u} - \hat{w}_h\|_E,
$$

where $\|\cdot\|_E$ is the minimum energy extension norm. This leads to the a-priori error estimate:

$$
\|u-u_h\| \leq (1+\gamma_h^{-1}) \inf_{w_h \in U_h} \|u-w_h\| + \gamma_h^{-1} \inf_{\hat{w}_h \in \hat{U}_h^c} \|\hat{u}-\hat{w}_h\|_E \,.
$$

For the non-conforming version, given a sufficient regularity of exact trace $\hat{u}$, we can attempt to estimate the best approximation error in the discrete dual seminorm,

$$
\inf_{\hat{w}_h \in \hat{U}_h} \underbrace{\sum_K \sup_{v_h \in V_h(K)} \left(\frac{\int_{\partial K} b_n (\hat{u}-\hat{w}_h) v_h}{\|v_h\|_{V(K)}} \right)^2}_{|\hat{u}-\hat{w}_h|^2_{V_h'}} . \tag{7.20}
$$

We have more discrete traces $\hat{w}_h$ to approximate with, so the best approximation error should be smaller.

The results above show that the convergence of fields should not be affected by the loss of control of traces in the minimum energy extension norm discussed next. In order to verify the assertion, we

have run an example with a smooth solution $u = 1 + x^3 + y^3$ with a constant advection field $b = (1, 1.1), (1, 1.01), (1, 1.001), (1, 1.0001)$ and the degenerate case $b = (1, 1)$. Note that in the last case, the diagonal edges are excluded. The results are shown in Fig. 9. The convergence curves are sitting literally on top of each other.

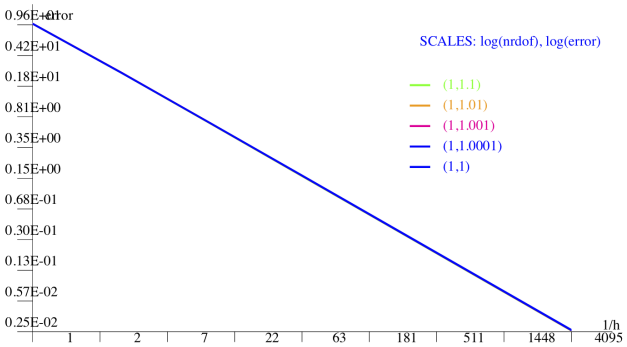

Figure 9: Smooth exact solution. $b = (1, 1.1), (1, 1.01), (1, 1.001), (1, 1.0001), (1, 1)$. Uniform h-refinements.

Convergence of traces

Can we proceed with the Brezzi argument to control traces? The discrete inf-sup constant of interest is defined as follows,

$$\sup_{v \in \mathcal{P}^{p+\Delta p}(K)} \frac{|\int_{\partial K} b_n uv|}{\|v\|_{H_{A^*}(K)}} \geq \delta \|u\|_E \quad u \in \mathcal{P}_c^p(\partial K), \tag{7.21}$$

where

$$\|v\|^2_{H_{A^*}(K)} = \int_K |A^* v|^2 + |v|^2$$

and $\|u\|_E$ is the minimum energy extension norm with respect to graph norm,

$$\|U\|^2_{H_A(K)} = \int_K |AU|^2 + |U|^2 .$$

We resort one more time to a numerical experiment. Fig. 10 presents values of constant δ for the unit triangle rotated by an angle $\alpha \in [0, 2\pi]$, $p = 2$, and enriched polynomial test space of order $p + 1$. The minimum energy extensions have been computed with polynomials of order $p + dp, dp = 5$. The constant

$p(\Delta p)/h$	1.0	0.1	0.01	0.001	0.0001
2(1)	0.84849	0.78929	0.7743	0.77281	0.77281
3(1)	0.36486	0.28960	0.2888	0.28878	0.28882
4(1)	5.8813E-2	5.6704E-4	5.5638E-8	5.5638E-8	5.5566E-10
4(2)	0.15936	0.10112	0.10377	0.10412	0.10414

Table 5: Dependence of constant δ for the right triangle rotated by angle $3\pi/8$ upon element size h for polynomial order $p = 2, 3, 4$.

degenerates to zero whenever one of the triangle edges becomes parallel to the advection vector. And the same results hold for element size $h = 0.1, 0.01, 0.001, 0.0001$. Clearly, to secure a robust convergence of traces, we have to impose a minimum angle condition on element edges with respect to the advection vector. Ideally, we should have used the composite test space in the computations but for, e.g., $\alpha \in (\pi/4, \pi/2)$ the triangle has just one outflow edge, i.e. the composite space reduces to the polynomial space and the constant still experiences the degeneration. Table 5 presents values of δ for a particular rotated right triangle as a

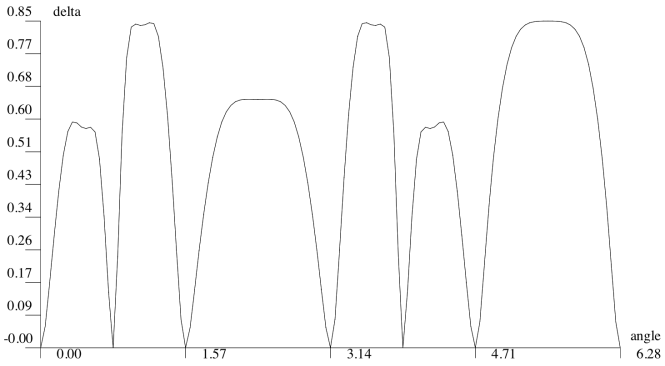

Figure 10: Constant δ for a rotated unit triangle, $b = (1, 0), c = 1$ and $p = 2, \Delta p = 1, dp = 5$.

function of element size h and polynomial order p. The first three rows present the values for the test space enrichment $\Delta p = 1$. For $p = 2, 3$, the constant is of order one and it converges with $h \to 0$. The values for $p = 4$ are much smaller and they deteriorate with $h \to 0$. We repeated the computations using $\Delta p = 2$, and the results became again h-independent although the constant got smaller again.

With the inf-sup constant δ_h in place, we can claim the convergence result for the conforming traces. This follows now directly from the Babuška - Brezzi Theorem. Let $\widetilde{V}_h \subset V_h$ denote the subspace of

approximate optimal test functions. We can reason as follows:

$$
\begin{aligned}
\|\hat{u}-\hat{u}_h\|_E &\leq \|\hat{u}-\hat{w}_h\|_E + \|\hat{w}_h-\hat{u}_h\|_E \\
&\leq \|\hat{u}-\hat{w}_h\|_E + \delta_h^{-1} \sup_{v_h\in V_h} \frac{\langle \hat{w}_h-\hat{u}_h, v_h\rangle_{\Gamma_h}}{\|v_h\|_V} \\
&\leq \|\hat{u}-\hat{w}_h\|_E + \delta_h^{-1} \sup_{v_h\in V_h} \frac{(w_h-u_h, A_h^* v_h) + \langle \hat{w}_h-\hat{u}_h, v_h\rangle_{\Gamma_h} - (w_h-u_h, A_h^* v_h)}{\|v_h\|_V} \\
&\leq \|\hat{u}-\hat{w}_h\|_E + \delta_h^{-1} \sup_{v_h\in V_h} \frac{(w_h-u_h, A_h^* v_h) + \langle \hat{w}_h-\hat{u}_h, v_h\rangle_{\Gamma_h}}{\|v_h\|_V} + \delta_h^{-1}\|w_h-u_h\| \\
&\leq \|\hat{u}-\hat{w}_h\|_E + \delta_h^{-1}(1+\gamma_h^{-1}) \sup_{v_h\in V_h} \frac{(w_h-u_h, A_h^* v_h) + \langle \hat{w}_h-\hat{u}_h, v_h\rangle_{\Gamma_h}}{\|v_h\|_V} \\
&= \|\hat{u}-\hat{w}_h\|_E + \delta_h^{-1}(1+\gamma_h^{-1}) \sup_{\widetilde{v}_h\in \widetilde{V_h}} \frac{(w_h-u_h, A_h^* \widetilde{v}_h) + \langle \hat{w}_h-\hat{u}_h, \widetilde{v}_h\rangle_{\Gamma_h}}{\|\widetilde{v}_h\|_V} \\
&\leq \|\hat{u}-\hat{w}_h\|_E + \delta_h^{-1}(1+\gamma_h^{-1}) \sup_{\widetilde{v}_h\in \widetilde{V_h}} \frac{(w_h-u, A_h^* \widetilde{v}_h) + \langle \hat{w}_h-\hat{u}, \widetilde{v}_h\rangle_{\Gamma_h}}{\|\widetilde{v}_h\|_V} \\
&\leq (1+\delta_h^{-1}(1+\gamma_h^{-1}))\|\hat{u}-\hat{w}_h\|_E + \delta_h^{-1}(1+\gamma_h^{-1})\|u-w_h\| \,.
\end{aligned}
$$

As $w_h, \hat{w}_h$ above are arbitrary functions, we obtain

$$
\|\hat{u}-\hat{u}_h\|_E \leq (1+\delta_h^{-1}(1+\gamma_h^{-1})) \inf_{\hat{w}_h} \|\hat{u}-\hat{w}_h\|_E + \delta_h^{-1}(1+\gamma_h^{-1}) \inf_{w_h} \|u-w_h\| \,.
$$

The result above holds for non-conforming traces as well, provided we replace the minimum energy extension norm with the discrete dual seminorm (7.20). Constant δ_h is then equal one by definition.

8 Conclusions

The present work presents several attempts to prove the convergence of the 'practical DPG' method for the convection-reaction problem proposed in [12]. The first in-depth analysis of the method was offered by Broersen, Dahmen and Stevenson in [2]. The authors were able to show the discrete stability provided the original enriched test space is refined 'sufficiently many times'. At the same time, they pointed out that numerical experiments do not reflect the need for the refinement. In the end, they did not modify the original method.

Our first attempt to prove the discrete stability through the construction of a local Fortin operator fails to show robustness in the orientation of elements with respect to the advection field. First of all, this type of analysis is done always under the assumption of element-wise constant material data. This assumption alone reduces the proposed methodology to a 'sanity check' and not a real proof for variable advection. Having convinced ourselves about the failure of the local construction, we proceeded with the global discrete stability analysis. Recall that global discrete stability is equivalent to the existence of a global Fortin operator. Indeed, we could rephrase the presented arguments in terms of constructing an appropriate Fortin operator. The presented argument is based on a combination of analytical stability arguments with numerical experiments done for a single element. 'Purists' will not recognize this as a proof, but this was the best we could

do at this point. The analytical argument is based on stability analysis presented in Appendix B that require additional assumptions on the advection field. We emphasize that the stability analysis using exponentials is not unique and can be reproduced under other assumptions on the advection field. We believe the ideas presented in this paper would still apply in such contexts. In the end, we use a 'marching strategy' to construct a discrete, weakly conforming test function that delivers stability of the fields. The marching strategy resembles ideas used in [11]. In the process of proving the discrete stability we arrived at the need of the composite test space, the idea that was explored in [11, 12]. In the end, the presented argument strongly indicates that using a simple enriched polynomial space may be insufficient for a robust method. We have not attempted yet to illustrate this conclusion with numerical experiments.

Once the stability and convergence of the fields is established (with no limitation on orientation of elements), we followed with stability and convergence arguments for traces, both conforming and non-conforming versions. In contrast to the fields, stability of traces (in the minimum energy extension norm), requires element edges to be uniformly bounded away from the streamlines of the advection field, an assumption very difficult to realize in practice. We emphasize, however, that this lack of robust stability for traces *does not* affect the convergence of fields. In conclusion, the numerically obtained traces should be used with caution or, better yet, not used at all.

References

[1] P. Bringmann and C. Carstensen. An adaptive least-squares FEM for the Stokes equations with optimal convergence rates. *Numer. Math.*, 135:59–492, 2017.

[2] D. Broersen, W. Dahmen, and R. P. Stevenson. On the stability of DPG formulations of transport equations. *Math. Comp.*, 87(311):1051–1082, May 2018.

[3] J. Brunken, K. Smetana, and K. Urban. (parametrized) first order transport equations: realization of optimally stable Petrov-Galerkin methods. *SIAM J. Sci. Comput.*, 41(1):A592–A621, 2019.

[4] T. Bui-Thanh, L. Demkowicz, and O. Ghattas. A unified discontinuous Petrov-Galerkin Method and its analysis for Friedrichs' systems. *SIAM J. Num. Anal.*, 51(4):1933–1958, 2013.

[5] C. Carstensen, L. Demkowicz, and J. Gopalakrishnan. Breaking spaces and forms for the DPG method and applications including Maxwell equations. *Comput. Math. Appl.*, 72(3):494–522, 2016.

[6] C. Carstensen and F. Hellwig. Low order discontinuous Petrov-Galerkin finite element methods for linear elasticity. *SIAM J. Num. Anal.*, 54(6):3388–3410, 2017.

[7] W. Dahmen, G. Kutyniok, W-Q. Lim, Ch. Schwab, and G. Welper. Adaptive anisotropic Petrov-Galerkin methods for first order transport equations. *J. Comput. Appl. Math.*, 340:191–220, 2018.

[8] W. Dahmen and R. P. Stevenson. Adaptive strategies for transport equations. *Comput. Methods Appl. Math.*, 19(3):431–464, 2019.

[9] H. De Sterck, T.A. Manteuffel, S.F. Mccormick, and L. Olson. Least-squares finite element methods and algebraic multigrid solvers for linear hyperbolic pdes.

[10] L. Demkowicz. *Computing with hp Finite Elements. I.One- and Two-Dimensional Elliptic and Maxwell Problems*. Chapman & Hall/CRC Press, Taylor and Francis, Boca Raton, October 2006.

[11] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. Part I: The transport equation. *Comput. Methods Appl. Mech. Engrg.*, 199(23-24):1558–1572, 2010. see also ICES Report 2009-12.

[12] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. Part II: Optimal test functions. *Numer. Meth. Part. D. E.*, 27:70–105, 2011. See also ICES Report 2009-16.

[13] L. Demkowicz, J. Gopalakrishnan, S. Nagaraj, and P. Sepúlveda. A spacetime DPG method for the Schrödinger equation. *SIAM J. Num. Anal.*, 55(4):1740–1759, 2017.

[14] L. Demkowicz and N. Heuer. Robust DPG method for convection-dominated diffusion problems. *SIAM J. Num. Anal*, 51:2514–2537, 2013. see also ICES Report 2011/13.

[15] L. Demkowicz and P. Zanotti. Construction of DPG Fortin operators revisited. *Comp. and Math. Appl.*, 80:2261–2271, 2020. Special Issue on Higher Order and Isogeometric Methods.

[16] I. Ekeland and R. Temam. *Convex Analysis and Variational Problems*. North Holland, Amsterdam, 1976.

[17] A. Ern and J-L Guermond. A converse to Fortin's lemma in Banach spaces. *C. R. Math. Acad. Sci. Paris*, 354(11):1092–1095, 2016.

[18] A. Ern and J.L. Guermond. Discontinuous Galerkin methods for Friedrichs' systems. I. General theory. *SIAM J. Numer. Anal.*, 44(2):753–778, July 2006.

[19] J. Gopalakrishnan, P. Monk, and P. Sepúlveda. A tent pitching scheme motivated by Friedrichs theory. *Comp. Math. Appl.*, 70:114–1135, 2015.

[20] J. Gopalakrishnan and W. Qiu. An analysis of the practical DPG method. *Math. Comp.*, 83(286):537–552, 2014.

[21] M. Jensen. *Discontinuous Galerkin Methods for Friedrichs Systems with Irregular Solutions*. PhD thesis, Corpus Christi College, University of Oxford, 2004.

[22] P. Joly. Some trace theorems in anisotropic Sobolev spaces. *SIAM J. Math. Anal.*, 23:799–819, May 1992.

[23] J. Munoz-Matute, D. Pardo, and L. Demkowicz. Equivalence between the DPG method and the Exponential Integrators for linear parabolic problems. *J. Comp. Physics*, 429(2):110016, 2021.

[24] S. Nagaraj, S. Petrides, and L. Demkowicz. Construction of DPG Fortin operators for second order problems. *Comput. Math. Appl.*, 74(8):1964–1980, 2017.

[25] J.T. Oden and L.F. Demkowicz. *Applied Functional Analysis for Science and Engineering*. Chapman & Hall/CRC Press, Boca Raton, 2018. Third edition.

A Boundness Below of the Convection-Reaction Operator

Clearly, we have to make some assumptions about coefficients b_i and c to guarantee boundedness below of the operator with some reasonable lower bound on boundedness below constant α. If both b and c converge to zero then $\alpha \to 0$ as well. The stability has to come either from convection or reaction[7].

Stability from reaction can be accessed quickly by testing with u. We get,

$$\int_\Omega b_i \frac{\partial u}{\partial x_i} u + cu^2 = \int_\Omega fu\,.$$

Integrating the first term by parts,

$$\int_\Omega b_i \frac{\partial u}{\partial x_i} u = \int_\Omega b_i \frac{\partial}{\partial x_i}\left(\frac{u^2}{2}\right) = -\int_\Omega \operatorname{div} b \frac{u^2}{2} + \int_{\Gamma_+} b_n \frac{u^2}{2}\,,$$

we obtain

$$\int_{\Gamma_+} b_n \frac{u^2}{2} + \int_\Omega (c - \frac{1}{2}\operatorname{div} b) u^2 = \int_\Omega fu\,.$$

Assuming

$$c - \frac{1}{2}\operatorname{div} b \geq \beta > 0\,,$$

we obtain

$$\beta \|u\|^2 \leq \|f\|\,\|u\| \quad \Rightarrow \quad \beta\|u\| \leq \|f\|\,.$$

This sufficient condition clearly does not cover the pure advection case with $\operatorname{div} b = 0$.

Stability from advection is a bit more difficult to analyze. We have to use a bit more sophisticated analysis based on characteristics and turning the advection-reaction problem into a family of ODEs.

[7] Or both, consistently with the meaning of the alternative.

Characteristics. Solutions of the system of first order nonlinear ODEs,

$$\frac{dx}{dt} = b(x(t))\,, \tag{A.22}$$

are called *characteristics*. We shall assume that the family of characteristics can be extended to a curvilinear system of coordinates (t, ξ) covering the whole domain Ω, see Fig. 11.

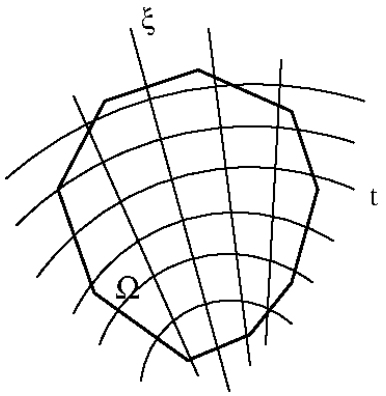

Figure 11: Characteristics.

Each characteristic originates on inflow boundary Γ_- and terminates on outflow boundary Γ_+. We shall assume the parametrization

$$x = x(t, \xi) \quad \xi \in O,\, x \in (t_-(\xi), t_+(\xi))$$

with the corresponding jacobian $j(x)$. If we furthermore assume that the system is orthogonal,

$$\frac{\partial x}{\partial \xi_i} \cdot \frac{\partial x}{\partial \xi_j} = \delta_{ij} \qquad \frac{\partial x}{\partial \xi_i} \cdot \frac{\partial x}{\partial t} = 0 \quad i, j = 1, 2 \qquad \text{(3D version)}\,,$$

the jacobian equals the magnitude of the advection vector:

$$\mathrm{jac}(x(t, \xi)) = |b(x(t, \xi)))|\,.$$

Implicit in the assumptions is that $b(x) \neq 0$ in domain Ω. If we assume that $b \in C^1(\bar{\Omega})$, the Weierstrass Theorem implies that we must have a positive lower bound on $|b(x)|$.

Convection-reaction problem with homogeneous BC.

$$\begin{cases} b \cdot \boldsymbol{\nabla} u + cu &= f \qquad \text{in } \Omega \\ u &= 0 \qquad \text{on } \Gamma_- \end{cases} \tag{A.23}$$

Let $\hat{u}(t) = u(x(t))$ where $x(t)$ is the characteristic. Then $u(x)$ satisfies (A.23) if and only if $\hat{u}(t)$ solves the initial-value ODE problem:

$$\begin{cases} \dfrac{d\hat{u}}{dt} + \hat{c}\hat{u} &= \hat{f} \qquad t \in (0, t_{x_0}) \\ \hat{u}(t_-) &= 0 \end{cases} \tag{A.24}$$

where $\hat{c}(t) := c(x(t))$, $\hat{f}(t) := f(x(t))$. The linear ODE with variable coefficients (A.24) admits a closed form solution,

$$\hat{u}(t) = \int_{t_-}^{t} e^{-(\hat{C}(t)-\hat{C}(s))} \hat{f}(s)\, ds \tag{A.25}$$

where $\hat{C}(t)$ is a primitive of $\hat{c}(t)$, i.e. $d\hat{C}/dt = \hat{c}$. Standard reasoning and Cauchy-Schwarz inequality lead to the estimate:

$$\begin{aligned}
\int_{t_-}^{t_+} |\hat{u}(t)|^2\, dt &= \int_{t_-}^{t_+} |\int_{t_-}^{t} e^{-(\hat{C}(t)-\hat{C}(s))} \hat{f}(s)\, ds|^2\, dt \\
&\le \int_{t_-}^{t_+} \int_{t_-}^{t} |e^{-(\hat{C}(t)-\hat{C}(s))}|^2\, ds \int_{t_-}^{t} |\hat{f}(s)|^2\, ds\, dt \\
&\le \underbrace{\int_{t_-}^{t_+} \int_{t_-}^{t} |e^{-(\hat{C}(t)-\hat{C}(s))}|^2\, ds\, dt}_{\text{stability constant}} \int_{t_-}^{t_+} |\hat{f}(s)|^2\, ds\,.
\end{aligned}$$

We need to work out now sufficient conditions on coefficients $b(x), c(x)$ to translate this estimate into a boundness below estimate for the convection-reaction operator.

Assumption on the reaction term. We shall assume that $c(x) \ge 0$. With this assumption,

$$\hat{C}(t) - \hat{C}(s) = \int_{s}^{t} c(x(\eta,\xi))\, d\eta \ge 0 \quad \Rightarrow \quad e^{-(\hat{C}(t)-\hat{C}(s))} \le 1,$$

and the stability constant can be estimated by

$$\int_{t_-}^{t_+} \int_{t_-}^{t} ds\, dt = \frac{(t_+ - t_-)^2}{2}\,.$$

Unfortunately, the 1D estimate does not translate immediately into the boundedness below condition as we have to include the jacobian in the computations.

$$\begin{aligned}
\int_{t_-}^{t_+} |\hat{u}(t)|^2 \underbrace{\mathrm{jac}(x(t,\xi))}_{=\widehat{\mathrm{jac}}(t)}\, dt &= \int_{t_-}^{t_+} |\int_{t_-}^{t} e^{-(\hat{C}(t)-\hat{C}(s))} \hat{f}(s)\, ds|^2\, \widehat{\mathrm{jac}}(t)\, dt \\
&\le \int_{t_-}^{t_+} \int_{t_-}^{t} |e^{-(\hat{C}(t)-\hat{C}(s))}|^2\, ds \int_{t_-}^{t} |\hat{f}(s)|^2\, ds\, \widehat{\mathrm{jac}}(t)\, dt \\
&\le \int_{t_-}^{t_+} (t-t_-) \int_{t_-}^{t} |\hat{f}(s)|^2\, ds\, \widehat{\mathrm{jac}}(t)\, dt && (c \ge 0) \\
&= \int_{t_-}^{t_+} \int_{s}^{t_+} (t-t_-)\, \widehat{\mathrm{jac}}(t)\, dt\, |\hat{f}(s)|^2\, ds && (\text{Fubini's Theorem})\,.
\end{aligned}$$

This leads to a rather technical assumption,

$$\int_{s}^{t_+} (t-t_-)\, \widehat{\mathrm{jac}}(t)\, dt \le C\, \widehat{\mathrm{jac}}(s) \qquad s < t < t_+,\, t_- < s < t_+ \quad C > 0\,.$$

The assumption is easily satisfied if we assume lower and upper bounds on the jacobian,

$$0 < b_{\min} \le \widehat{\mathrm{jac}} = |\hat{b}| \le b_{\max} < \infty\,. \tag{A.26}$$

This implies

$$\widehat{\mathrm{jac}}(t) \le \frac{b_{\max}}{b_{\min}}\,\widehat{\mathrm{jac}}(s),$$

from which it follows that

$$\int_s^{t_+} (t-t_-)\,\widehat{\mathrm{jac}}(t)\,dt \le \frac{b_{\max}}{b_{\min}}\,\widehat{\mathrm{jac}}(s)\int_s^{t_+}(t-t_-)\,dt \le \frac{b_{\max}}{b_{\min}}\,\frac{(t_+-t_-)^2}{2}\,\widehat{\mathrm{jac}}(s)\,.$$

B Appendix: Boundness Below of the Convection-Reaction Operator. Second Approach

The disadvantage of the analysis shown in Appendix A is that it does not cover the case when the stability (boundedness below) may come from the advection in one part of the domain, and from reaction in the rest of the domain. The approach presented here rectifies this deficiency in the case when the advection field has a scalar potential,

$$b(x) = \boldsymbol{\nabla} V(x)\,.$$

We shall also focus now on the adjoint operator,

$$A^* v = -\operatorname{div}(bv) + cv, \quad D(A^*) = \{v \in H_{A^*}(\Omega)\ : v = 0 \text{ on } \partial\Omega_+\}\,. \tag{B.27}$$

Actually, we will develop a more general stability estimate for any $v \in H_{A^*}(\Omega)$. Following [14], we start by introducing an auxiliary unknown,

$$w(x) := e^{V(x)} v(x) \quad \boldsymbol{\nabla} w = e^V b v + e^V \boldsymbol{\nabla} v\,.$$

Let $f := A^* v$. We have,

$$e^V f = e^V(-b\cdot\boldsymbol{\nabla} v + (c - \operatorname{div} b)v) = -\operatorname{div}(bw) + (|b|^2 + c)w\,.$$

Multiplying both sides with w and integrating over Ω, we obtain

$$-\int_\Omega \operatorname{div}(bw)w + \int_\Omega (|b|^2+c)w^2 = \int_\Omega e^V f w\,.$$

The first term is now integrated by parts,

$$-\int_\Omega \operatorname{div}(bw)w = \int_\Omega b\cdot\boldsymbol{\nabla}(\frac{w^2}{2}) - \int_\Gamma b_n w^2 = -\frac{1}{2}\int_\Gamma b_n w^2 - \frac{1}{2}\int_\Omega \operatorname{div} b\, w^2\,.$$

This gives:

$$-\frac{1}{2}\int_\Gamma b_n w^2 + \int_\Omega \underbrace{(|b|^2 + c - \frac{1}{2}\operatorname{div} b)}_{=:a} w^2 = \int_\Omega e^V f w\,.$$

Under an additional assumption that coefficient $a(x)$ is positive, we can use Young's inequality to estimate the right-hand side,

$$fw \leq \frac{a}{2}w^2 + \frac{e^{2V}}{2a}f^2\,.$$

This leads to the final estimate,

$$\frac{1}{2}\int_{\Gamma_-}|b_n|w^2 + \frac{1}{2}\int_\Omega aw^2 \leq \int_\Omega \frac{e^{2V}}{2a}f^2 + \frac{1}{2}\int_{\Gamma_+} b_n w^2\,.$$

In particular, for $v = 0$ on Γ_+, we obtain

$$\int_\Omega a\, e^{2V} v^2 \leq \int_\Omega \frac{e^{2V}}{a}f^2\,.$$

If a admits a lower bound

$$a_{\min} \leq a$$

and e_1, e_2 are lower and upper bounds for e^{2V}, we obtain

$$a_{\min}\, e_1 \int_\Omega v^2 \leq \int_\Omega a\, e^{2V} v^2 \leq \frac{e_2}{a_{\min}}\int_\Omega f^2\,.$$

This gives the final estimate for the boundedness below constant:

$$\frac{e_1}{e_2}a_{\min}^2 \int_\Omega v^2 \leq \int_\Omega f^2\,. \tag{B.28}$$

REMARK 11

1. For an incompressible advection field $\operatorname{div} b = \Delta V = 0$, $a = |b|^2 + c$. Thus both advection term $|b|^2$ and reaction coefficient c may degenerate to zero, as long as the sum stays bounded away from zero.

2. The assumption on coefficient a to be bounded below away from zero is not as restrictive as it may appear. We can redefine $w = e^{kV}v$ and obtain a modified equation for w:

   $$-\operatorname{div}(bw) + (k^2|b|^2 + c)w\,.$$

   As long as $\operatorname{div} b$ is bounded, we can select a sufficiently large multiplier k to ensure the positivity of coefficient a. Of course, a larger k will result in a larger bound e_1/e_2.

∎

C Appendix: A Local Stability Result

Let K be an element with its outflow boundary denoted by ∂K_+. Let $a(x)$ and $w(x)$ be positive weights defined on K. Consider the constrained minimization problem

$$\min_{v_h \in \mathcal{P}^{p+1}(K)} \frac{1}{2}\left\{\int_K \frac{w}{a}(A^*(v_h - v))^2 + \int_{\partial K_+} w b_n (v_h - v)^2\right\}$$

under the constraints:

$$\begin{aligned}\int_K \delta u_h\, A^*(v_h - v) &= 0 \qquad \forall\, \delta u_h \in \mathcal{P}^{p-1}(K)\,,\\ \int_{\partial K_+} wb_n\, \delta w_h (v_h - v) &= 0 \qquad \forall\, \delta w_h \in \mathcal{P}_c^{p+1}(\partial K_+)\,.\end{aligned}$$

The first constraint is a Fortin-like condition enabling the replacement of the exact test function v with its approximation v_h in our stability analysis. The second constraint enforces weak conformity of the global discrete test function; see Section 6 for details. The minimization problem is equivalent to the mixed problem:

$$\begin{cases} v_h \in \mathcal{P}^{p+1}(K),\, u_h \in \mathcal{P}^{p-1}(K),\, w_h \in \mathcal{P}_c^{p+1}(\partial K_+) & & \\ \displaystyle\int_K \frac{w}{a} A^* v_h A^* \delta v_h + \int_K u_h A^* \delta v_h + \int_{\partial K_+} wb_n (v_h + w_h)\delta v_h &= \displaystyle\int_K \frac{w}{a} A^* v A^* \delta v_h + \int_{\partial K_+} wb_n v \delta v_h & \delta v_h \in \mathcal{P}^{p+1}(K) \\ \displaystyle\int_K \delta u_h\, A^* v_h &= \displaystyle\int_K \delta u_h\, A^* v & \delta u_h \in \mathcal{P}^{p-1}(K) \\ \displaystyle\int_{\partial K_+} wb_n\, \delta w_h v_h &= \displaystyle\int_{\partial K_+} wb_n\, \delta w_h v & \delta w_h \in \mathcal{P}_c^{p+1}(\partial K_+)\,. \end{cases}$$

or, equivalently,

$$\begin{cases} v_h \in \mathcal{P}^{p+1}(K),\, u_h \in \mathcal{P}^{p-1}(K),\, w_h \in \mathcal{P}_c^{p+1}(\partial K_+) & & \\ \displaystyle\int_K \frac{w}{a} A^* v_h A^* \delta v_h + \int_K u_h A^* \delta v_h + \int_{\partial K_+} wb_n w_h \delta v_h &= \displaystyle\int_K \frac{w}{a} A^* v A^* \delta v_h & \delta v_h \in \mathcal{P}^{p+1}(K) \\ \displaystyle\int_K \delta u_h\, A^* v_h &= \displaystyle\int_K \delta u_h\, A^* v & \delta u_h \in \mathcal{P}^{p-1}(K) \\ \displaystyle\int_{\partial K_+} wb_n\, \delta w_h v_h &= \displaystyle\int_{\partial K_+} wb_n\, \delta w_h v & \delta w_h \in \mathcal{P}_c^{p+1}(\partial K_+)\,. \end{cases} \tag{C.29}$$

Let

$$\begin{aligned} V_h &:= \mathcal{P}^{p+1}(K)\\ V_{h,0} &:= \{v_h \in V_h \,:\, \int_{\partial K_+} wb_n\, \delta w_h v_h = 0 \quad \forall\, \delta w_h \in \mathcal{P}_c^{p+1}(\partial K_+)\}\\ V_{h,00} &:= \{v_h \in V_{h,0} \,:\, \int_K \delta u_h A^* v_h = 0 \quad \forall\, \delta u_h \in \mathcal{P}^{p-1}(K)\}\,. \end{aligned}$$

Note that, if the restriction of v to ∂K_+ is itself in space $\mathcal{P}_c^{p+1}(\partial K_+)$, then

$$V_{h,0} = \{v_h \in V_h \,:\, v_h = 0 \text{ on } \partial K_+\}\,.$$

Consider the corresponding decomposition of $v_h \in V_h$,

$$v_h = v_{h,00} + v_{h,0}^{\perp} + v_h^{\perp}\,, \qquad v_{h,00} \in V_{h,00},\, v_{h,0}^{\perp} \in V_{h,0}^{\perp},\, v_h^{\perp} \in V_h^{\perp}$$

where

$$\begin{aligned} V_{h,0}^{\perp} &:= \{v_{h,0} \in V_{h,0} \,:\, (v_{h,0}^{\perp}, \delta v_h)_V = 0 \qquad \forall\, \delta v_h \in V_{h,00}\}\\ V_h^{\perp} &:= \{v_h \in V_h \,:\, (v_h^{\perp}, \delta v_h)_V = 0 \qquad \forall\, \delta v_h \in V_{h,0}\} \end{aligned}$$

with the inner product:

$$(v,\delta v)_V = \int_K \frac{w}{a} A^* v A^* \delta v + \int_{\partial K_+} w b_n \, v \delta v \, .$$

Introduce norms for the Lagrange multiplier $u_h \in \mathcal{P}_h^{p-1}(K)$, $w_h \in \mathcal{P}_c^{p+1}(\partial K_+)$,

$$\|u_h\|_K^2 := \int_K u_h^2, \quad \|w_h\|_{\partial K_+}^2 := \int_{\partial K_+} w b_n w^2,$$

and consider the corresponding inf-sup constants,

$$\begin{aligned} \alpha_h &:= \inf_{u_h \in \mathcal{P}^{p-1}(K)} \sup_{v_h \in V_{h,0}} \frac{\int_\Omega u_h A^* v_h}{\|u_h\|_K \, \|v_h\|_V} \\ \beta_h &:= \inf_{w_h \in \mathcal{P}_c^{p+1}(\partial K_+)} \sup_{v_h \in V_h} \frac{\int_{\partial K_+} w b_n \, w_h v_h}{\|w_h\|_{\partial K_+} \, \|v_h\|_V} \end{aligned}$$

Lemma 8

The following estimate holds:

$$\|v_h\|_V^2 \leq (1+\alpha_h^{-2})\|A^* v\|^2 + \beta_h^{-2} \|v\|_{\partial K_+}^2 \, . \tag{C.30}$$

∎

Proof: Testing in (C.29)$_1$ with $\delta v_h = v_{h,00}$, we get

$$\|v_{h,00}\|_V = \|A^* v_{h,00}\|_{L^2(K)} \leq \|A^* v\|_{L^2(K)} \, .$$

By the Banach Closed Range Theorem,

$$\alpha_h = \inf_{u_h \in \mathcal{P}^{p-1}(K)} \sup_{v_h \in V_{h,0}} \frac{\int_\Omega u_h A^* v_h}{\|u_h\|_K \, \|v_h\|_V} = \inf_{v_{h,0}^\perp \in V_{h,0}^\perp} \sup_{u_h \in \mathcal{P}^{p-1}(K)} \frac{\int_\Omega u_h A^* v_h}{\|u_h\|_K \, \|v_{h,0}^\perp\|_V}.$$

This, along with (C.29)$_2$, yields the estimate:

$$\|v_{h,0}^\perp\|_V = \|A^* v_{h,0}^\perp\|_K \leq \alpha_h^{-1} \|A^* v\|_K \, .$$

Similarly,

$$\beta_h = \inf_{w_h \in \mathcal{P}_c^{p+1}(\partial K_+)} \sup_{v_h \in V_h} \frac{\int_{\partial K_+} w b_n \, w_h v_h}{\|w_h\|_{\partial K_+} \, \|v_h\|_V} = \inf_{v_h^\perp \in V_h^\perp} \sup_{w_h \in \mathcal{P}_c^{p+1}(\partial K_+)} \frac{\int_{\partial K_+} w b_n \, w_h v_h}{\|w_h\|_{\partial K_+} \, \|v_h^\perp\|_V}$$

along with (C.29)$_3$, implies:

$$\|v_h^\perp\|_V =\leq \beta_h^{-1} \|v\|_{\partial K_+} \, .$$

In conclusion,

$$\|v_h\|_V^2 = \|v_{h,00}\|_V^2 + \|v_{h,0}^\perp\|_V^2 + \|v_h^\perp\|_V^2 \leq (1+\alpha_h^{-2})\|A^* v\|^2 + \beta_h^{-2} \|v\|_{\partial K_+}^2 \, .$$

∎

D Appendix: Quadrilateral Elements

We present numerical experiments with the inf-sup constant β for a quadrilateral element. Following the results for a triangular element, we consider a composite enriched space defined by partitions shown in Fig. 12. In the case of a square element, only two outflow edges are possible (the first two cases shown in Fig. 12) but in the case of a general quadrilateral, the third case with three outflow edges may occur as well. Also, the element may be subdivided in just two triangles (case not shown). We have experimented

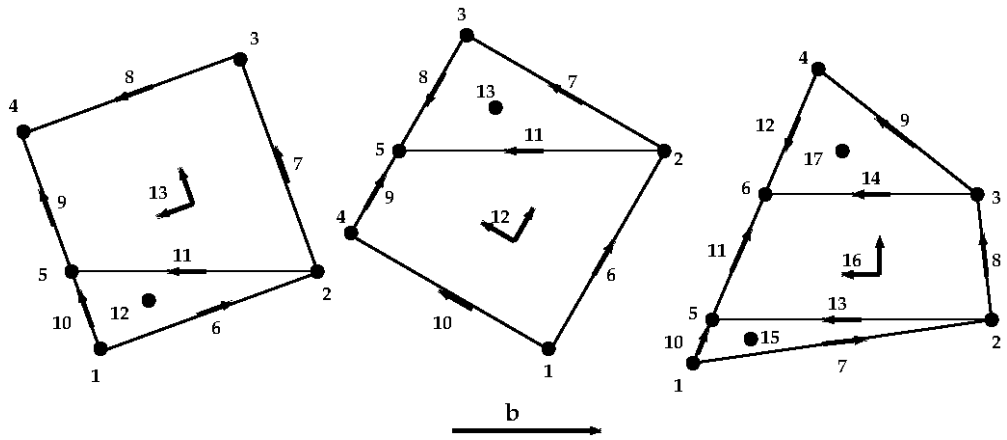

Figure 12: Partitions of a general quad element into triangular and quadrilateral subelements defining the enriched test space.

with the two elements shown in Fig. 13. Depending upon the rotation angle, the square element may be refined into a quad and a triangle, two triangles or may not be refined at all. The quadrilateral element may undergo any of the three refinements shown in Fig. 12. The minimum values (over all rotation angles) of the

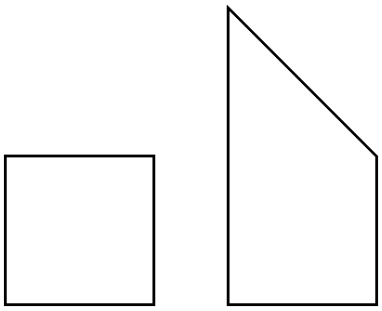

Figure 13: The considered quadrilateral elements.

inf-sup constant β_h for the two element shapes, $p = 2$ and different values of element size h are shown in Table 6. The constant stays beautifully close to one, and it converges to one as $h \to 0$. The same behavior is observed for elements of higher order p.

quad /h	1.0	0.1	0.01	0.001	0.0001
1	0.9949266746	0.9999887821	0.9999999876	0.9999999999	0.9999999999
2	0.9878674760	0.9999641904	0.9999999585	0.9999999999	0.9999999999

Table 6: Minimal (over angles) value of inf-sup constant β_h for different values of element size h for advection vector $b = (1, 0)$, reaction coefficient $c = 1.0$, and element shapes shown in Fig. 13.