2010

A Class of Discontinuous Petrov–Galerkin Methods. Part I: The Transport Equation

Leszek Demkowicz
University of Texas at Austin

Jay Gopalakrishnan
Portland State University, gjay@pdx.edu

Citation Details

Demkowicz, Leszek and Gopalakrishnan, Jay, "A Class of Discontinuous Petrov–Galerkin Methods. Part I: The Transport Equation" (2010). *Mathematics and Statistics Faculty Publications and Presentations*. 55.
https://pdxscholar.library.pdx.edu/mth_fac/55

A CLASS OF DISCONTINUOUS PETROV-GALERKIN METHODS. PART I: THE TRANSPORT EQUATION

L. DEMKOWICZ AND J. GOPALAKRISHNAN

Abstract. Considering a simple model transport problem, we present a new finite element method. While the new method fits in the class of discontinuous Galerkin (DG) methods, it differs from standard DG and streamline diffusion methods, in that it uses a space of discontinuous trial functions tailored for stability. The new method, unlike the older approaches, yields optimal estimates for the primal variable in both the element size h and polynomial degree p, and outperforms the standard upwind DG method.

1. Introduction

We introduce a new Petrov-Galerkin method for advective problems. While it belongs in the class of discontinuous Galerkin (DG) methods, unlike the standard upwind DG method, we are able to prove *optimal h and p error estimates* in the L^2-norm for our discrete solution on general meshes (where, as usual, h is the mesh size and p is the polynomial degree). The method includes a separate outflux approximation on element interfaces and a space of non-standard test functions designed for stability.

The boundary value problem that is the subject of this paper is posed on a polyhedral domain Ω. Given f and g, we need to find a finite element approximation to the solution u of

$$\vec{\beta} \cdot \vec{\nabla}\, u = f \qquad \text{on } \Omega, \tag{1a}$$

$$u = g \qquad \text{on } \partial_{\text{in}}\Omega. \tag{1b}$$

We only consider the case of constant $\vec{\beta}$ in this paper (but extensions are possible, as mentioned in Section 5). The inflow boundary $\partial_{\text{in}}\Omega$ appearing in (1), is defined, letting $\vec{n}$ denote the unit outward normal, by

$$\partial_{\text{in}}\Omega = \{\vec{x} \in \partial\Omega : \vec{\beta} \cdot \vec{n}(\vec{x}) < 0\}, \tag{2}$$

i.e., $\partial_{\text{in}}\Omega$ denotes the global inflow boundary. While non-finite-element numerical techniques can be designed for this problem (e.g. the method of characteristics), we aim for finite elements because of its versatility in handling complicated domains as well as certain regular and singular perturbations of the above problem. A regular perturbation of (1) is

$$\vec{\beta} \cdot \vec{\nabla}\, u + \alpha(\vec{x})u = f. \tag{3}$$

A singular perturbation of (1) is obtained by the addition of a small viscosity term with second derivatives. This is harder to analyze. Within the domain of finite element methods for (1), there are two broad categories (see [11] for a review). One is the very popular

The work of the first author was supported by DOE through Predictive Engineering Science (PECOS) Center at ICES (PI: Bob Moser), and by a research contract with Boeing.

The second author was supported in part by the National Science Foundation under grants 0713833, 0619080, and an Oden fellowship at ICES.

streamline diffusion method [13] and its descendants. The other category is composed of DG methods. Since our contribution fits in the latter, we shall now review previous works in this category in detail.

The well known first papers proposing and analyzing the original DG method for (1) are [14, 15, 18]. To distinguish this method from our DG method, we will call the original DG method the "upwind DG method" and denote it by UDG, while we call ours the "discontinuous Petrov-Galerkin method" and denote it by DPG. It is proved in [15] that if u_h is the UDG approximation, then (for a fixed p) it satisfies $\|u - u_h\|_{L^2(\Omega)} \le O(h^{s-1})$ for some $s \le p + 1$ dictated by the regularity of the exact solution. This result was improved by [14] wherein it was shown that the rate of convergence is in fact $O(h^{s-1/2})$. In both cases convergence with respect to p was not studied. Even if we set aside the p-convergence issue, notice that both the results are suboptimal in h, as the best approximation error of the finite element space is $O(h^s)$.

For some special classes of meshes however, many authors have observed (and proved) the optimal rate of convergence of the UDG method [8, 19] with respect to h. Nonetheless, on general meshes, the suboptimal rate of convergence cannot be improved, as shown by a numerical example in [17] using a particular quasiuniform mesh and a smooth exact solution. To express the sentiment of many, we quote from [8] that "the mechanisms that induce the loss of $h^{1/2}$ in the order of convergence of the L^2-norm of the error are not very well known yet".

An hp analysis of the UDG scheme was first provided in [3]. They considered the regular perturbation (3) under the assumption that

$$0 < c_0 \le \alpha(\vec{x}) \qquad \forall \vec{x} \in \Omega \tag{4}$$

Because of this assumption, they are able to control the $L^2(\Omega)$-norm of the solution. They also introduced a stabilization parameter into the original upwind DG method. A few years later, the paper [12] extended the results of [3] in several directions, providing a unified theory for an hp version of the streamline diffusion method, as well as the upwind DG method. Their analysis did not assume (4), rather they let the advection vector $\vec{\beta}$ depend on $\vec{x}$ and assumed

$$0 < c_0 \le -\frac{1}{2} \nabla \cdot \beta(\vec{x}) + \alpha(\vec{x}) \qquad \forall \vec{x} \in \Omega, \tag{5}$$

as a consequence of which they have stability in $c_0\|u\|_{L^2(\Omega)}$. (Note that for the case we intend to study (1), the right hand side of (5) evaluates to zero, so (5) does not hold.) Both papers analyzed the stabilized version of the upwind DG method and both relied on the proper choice of the stabilization parameter. While the results of [12] are optimal in p, those of [3] are suboptimal in p. In all these works, the $L^2(\Omega)$-rate of convergence of the error with respect to h remained suboptimal by $h^{1/2}$. In contrast, our results do not exhibit this suboptimality, nor do we add any stabilization parameter. We believe our method is the first in the finite element (not just DG) family of methods for the transport equation which has provably optimal convergence rates on very general meshes.

The design of our method is guided by a generalization of Céa lemma due to Babuška [1, 5]. We only need a simple version of the result, which we now describe using the following notations (all our spaces are over $\mathbb{R}$): Let X, $X_h \subset X$, and V_h be Banach spaces and let $a_h(\cdot, \cdot)$ be a bilinear form on $X \times V_h$. Suppose the exact solution $U \in X$ satisfies and the discrete solution $U_h \in X_h$ satisfies

$$a_h(U - U_h, v_h) = 0 \quad \text{for all } v_h \in V_h. \tag{6}$$

If the bilinear form is continuous in the sense that there is a $C_1 > 0$ such that

$$a_h(w, v_h) \le C_1 \|w\|_X \|v_h\|_{V_h} \quad \text{for all } w \in X,\ v_h \in V_h \tag{7}$$

and also the inf-sup condition, i.e., there is a $C_2 > 0$ such that

$$C_2 \|w_h\|_X \le \sup_{v_h \in V_h} \frac{a_h(w_h, v_h)}{\|v_h\|_{V_h}} \quad \text{for all } w_h \in X_h, \tag{8}$$

then, as is well known, the following theorem can be formulated:

Theorem 1.1. *Under the above setting, we have the following error estimate:*

$$\|U - U_h\|_X \le (1 + \frac{C_1}{C_2}) \inf_{w_h \in X_h} \|U - w_h\|_X.$$

Proof. The argument is simple and standard:

$$\begin{aligned}
\|U - U_h\|_X &= \|U - w_h\|_X + \|w_h - U_h\|_X \\
&\le \|U - w_h\|_X + \frac{1}{C_2} \sup_{v_h \in V_h} \frac{a_h(w_h - U_h, v_h)}{\|v_h\|_{V_h}} && \text{by (8)}, \\
&\le \|U - w_h\|_X + \frac{1}{C_2} \sup_{v_h \in V_h} \frac{a_h(w_h - U, v_h)}{\|v_h\|_{V_h}} && \text{by (6)}, \\
&\le \|U - w_h\|_X + \frac{C_1}{C_2} \|w_h - U\|_X && \text{by (7)}.
\end{aligned}$$

This finishes the proof. □

Many refinements and improvements of such a theorem are known. But our purpose in going through the above simple argument is to clearly show that the test space need not have approximation properties. Hence, in designing Petrov-Galerkin methods, *while we must choose trial spaces with good approximation properties, we may design test spaces solely to obtain good stability properties.* This will be our guiding principle in designing our method. In fact, the test spaces we propose shortly can have discontinuities *inside* the mesh elements.

Many researchers have put the above principle to good use. In fact, even the abbreviation we use for our new method "DPG method", has been previously used [4, 6] for other methods. The theme in these works is the search for stable test spaces using bubbles or other polynomials. Our test space functions, in contrast, need not be polynomial on an element, and indeed, need not even be continuous. We are also not the first to consider such functions with discontinuities within a finite element. Such elements are routinely used in X-FEM and similar methods [2] for difficult simulations like crack propagation. However, we use discontinuities solely for stability purposes, and solely in test spaces. Our trial spaces, being standard polynomial spaces, possess provably good approximation properties.

Our method also introduces a new flux unknown on the element interfaces. This is in line with the recent developments on hybridized DG (HDG) methods [9]. HDG methods that extend the ideas in [9] to the case of convection can be found in recent works [10, 16]. These methods are constructed by defining a independent flux variable on the element interfaces which can solved for first, after which the internal variables can be locally solved for. While this can be thought of as akin to static condensation, additional advantages can be exploited, such as easy stabilization [16] using a penalty parameter. However, p-independent stability for such methods has not been proved yet. While we borrow the idea of letting the fluxes be independent variables in the design of our method, our method does not have stabilization parameters, and has p-independent stability.

We organize our presentation such that a spectral version of the method is first exhibited (in Section 2). Details regarding the new space of test functions and the stability estimates for the method on a single element are presented in that section. Section 3 then presents the composite method on a triangular mesh. Optimal L^2 error estimates are proved in Theorem 3.2 there. We conclude in Section 5 opining on important future directions. Proofs of a few technical estimates are gathered in Appendix A.

2. The spectral method on one element

We start by considering the one-element case to fix the ideas and study the element spaces. In other words, we let Ω be an N-simplex ($N = 2, 3$ are the cases of most practical importance) which we denote by K throughout this section. We now describe and analyze the DPG method on the single element K.

Define the outflow and inflow boundaries of K by

$$\partial_{\text{out}} K = \{\vec{x} \in \partial K : \vec{\beta} \cdot \vec{n} > 0\}, \tag{9a}$$

$$\partial_{\text{in}} K = \{\vec{x} \in \partial K : \vec{\beta} \cdot \vec{n} < 0\}. \tag{9b}$$

Here, and throughout the paper, we use $\vec{n}$ to generically denote the unit outward normal on the boundary of a polygonal domain (the domain will vary at different occurrences of $\vec{n}$, but will always be clear from the context, e.g., above it is K, while in (2) it was Ω). Note that we omitted the $\vec{\beta} \cdot \vec{n} = 0$ case in (9). This, and any concerns regarding computational inaccuracies in determination of strict inequalities of (9), will be allayed in § 2.3.

2.1. Petrov-Galerkin method using a new test space.

Before we describe the method, we need to introduce a new test space. For this, we need a coordinate system aligned with the flow – see Figure 1. Let $\vec{e}^{(1)}, \ldots, \vec{e}^{(N)}$, denote the standard unit vectors and let $\vec{e}_\beta$ denote the unit vector in the $\vec{e}_\beta$-direction. Completing $\{\vec{e}_\beta\}$ to form an orthonormal basis $\{\vec{e}_\beta, \vec{e}_\beta^{(2)}, \ldots \vec{e}_\beta^{(N)}\}$ for $\mathbb{R}^N$, we write the new coordinates as η_i, i.e.,

$$\begin{aligned} \vec{x} &= x_1 \vec{e}^{(1)} + \cdots + x_N \vec{e}^{(N)} \\ &= \eta_1 \vec{e}_\beta + \eta_2 \vec{e}_\beta^{(2)} + \cdots + \eta_N \vec{e}_\beta^{(N)}. \end{aligned} \tag{10}$$

Any function defined on $\partial_{\text{out}} K$ can be extended into K in such a way that the extension is constant along the $\vec{\beta}$-direction, i.e., the extension is independent of η_1. We call this the *extension from outflow boundary* and denote it by $\mathcal{E}_{\text{out}}$. We can think of the outflow surface as the graph of a function $\eta_1 = \mathcal{F}_{\text{out}}(\eta_2, \ldots, \eta_N)$. Then, a function on the outflow surface can be expressed (after eliminating the η_1 variable) as $\phi(\eta_2, \ldots, \eta_N)$. These coordinates make the expression for $\mathcal{E}_{\text{out}}\, \phi$ trivial, namely $\mathcal{E}_{\text{out}}\, \phi(\eta_1, \eta_2, \ldots, \eta_N) = \phi(\eta_2, \ldots, \eta_N)$.

Let us consider the question of computing a polynomial approximation of the transport solution u satisfying (1) with $\Omega = K$. The starting point in deriving the DPG method is

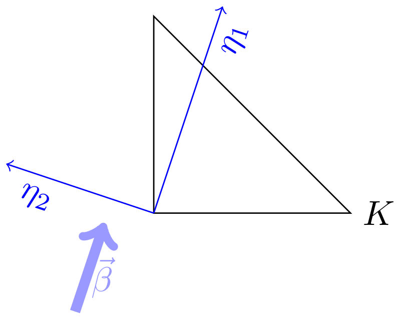


FIGURE 1. Streamline coordinates

to multiply (1a) by a test function v and integrate by parts:

$$\begin{aligned}(f,v) &= (\vec\beta\cdot\vec\nabla\,u,v)_K\\ &= -(u,\vec\beta\cdot\vec\nabla\,v)_K+\langle\vec\beta\cdot\vec n\,u,v\rangle_{\partial K}\\ &= -(u,\vec\beta\cdot\vec\nabla\,v)_K+\langle\phi,v\rangle_{\partial_{\text{out}}K}+\langle\vec\beta\cdot\vec n\,g,v\rangle_{\partial_{\text{in}}K}\end{aligned}\tag{11}$$

with $\phi=\vec\beta\cdot\vec n\,u$ on $\partial_{\text{out}}K$, the *outflux variable.*

The method finds an approximation to the pair (u,ϕ) using the spaces

$$\begin{aligned}M_\ell(\partial_{\text{out}}K) = \{\mu:\ \mu|_F\in P_\ell(F), &\text{for all faces } F\\ &\text{of } K \text{ contained in } \partial_{\text{out}}K\},\end{aligned}\tag{12a}$$

$$V_\ell(K) = \eta_1 P_\ell(K)+\mathcal{E}_{\text{out}}(M_{\ell+1}(\partial_{\text{out}}K)).\tag{12b}$$

Here, for any domain D, we use $P_\ell(D)$ to denote the space of polynomials of degree at most ℓ, and

$$\eta_1P_\ell(K)=\{\eta_1p_\ell:\ p_\ell\in P_\ell(K)\}.$$

Note that if $\partial_{\text{out}}K$ consists of a single face of K, then $V_\ell(K)$ consists of polynomials, but in general, $V_\ell(K)$ is a non-polynomial space, as it can have lines of discontinuity within K.

The *spectral DPG method*, motivated by (11), defines approximations $(u_p,\phi_{p+1})\in P_p(K)\times M_{p+1}(\partial_{\text{out}}K)$, satisfying

$$\begin{aligned}-(u_p,\vec\beta\cdot\vec\nabla\,v)_K+\langle\phi_{p+1},v\rangle_{\partial_{\text{out}}K}&\\ = (f,v)_K-\langle\vec\beta\cdot\vec n\,g,v\rangle_{\partial_{\text{in}}K},&\end{aligned}\tag{13}$$

for all $v\in V_p(K)$. For compact notation, we write

$$\begin{aligned}a(\,(u_p,\phi_{p+1}),v) &\overset{\text{def}}{=} -(u_p,\vec\beta\cdot\vec\nabla\,v)_K+\langle\phi_{p+1},v\rangle_{\partial_{\text{out}}K}\\ b(v) &\overset{\text{def}}{=} (f,v)_K-\langle\vec\beta\cdot\vec n\,g,v\rangle_{\partial_{\text{in}}K},\\ X_p(K) &\overset{\text{def}}{=} P_p(K)\times M_{p+1}(\partial_{\text{out}}K),\end{aligned}$$

so that (13) can be written as the problem of finding $(u_p,\phi_{p+1})\in X_p(K)$ such that

$$a(\,(u_p,\phi_{p+1}),v)=b(v)\qquad\forall v\in V_p(K).$$

This is a Petrov-Galerkin method because the test and trial spaces are different. The solvability of this method will be clear shortly, as soon as we study the new finite element space.

Proposition 2.1. *The component spaces $\eta_1 P_p(K)$ and $\mathcal{E}_{\text{out}}(M_{p+1}(\partial_{\text{out}}K))$ are linearly independent, so that $V_p(K)$ can be written as the direct sum*

$$V_p(K) = \eta_1 P_p(K) \oplus \mathcal{E}_{\text{out}}(M_{p+1}(\partial_{\text{out}}K)). \tag{14}$$

Proof. Suppose there is an $r \in P_p(K)$ and $\mu \in M_{p+1}(\partial_{\text{out}}K)$ such that

$$\eta_1 r = \mathcal{E}_{\text{out}}\, \mu,$$

then, letting ∂_{η_1} denote the distributional derivative with respect to the variable η_1, we have

$$\partial_{\eta_1}(\eta_1 r) = \partial_{\eta_1}(\mathcal{E}_{\text{out}}\, \mu) = 0.$$

Integrating from the $\eta_1 = 0$ plane (extending the polynomial r to all $\mathbb{R}^3$ if necessary) we find that

$$t\, r(t, \eta_2, \ldots, \eta_N) = \int_0^t \partial_{\eta_1}(\eta_1\, r(\eta_1, \ldots, \eta_N))\, ds = 0,$$

for any t, so $r \equiv 0$, which in turn implies that $\mu = 0$ as well. Thus the sum in (14) is indeed a direct sum. □

The degrees of freedom of $V_p(K)$ can be inferred from the next lemma.

Lemma 2.1. *Given any μ in $M_{p+1}(\partial_{\text{out}}K)$ and any w in $P_p(K)$, there is a unique v in $V_p(K)$ satisfying*

$$\vec{\beta} \cdot \vec{\nabla}\, v = w, \tag{15a}$$

$$v|_{\partial_{\text{out}}K} = \mu. \tag{15b}$$

Proof. We construct the required v as a sum of two functions v_1 and v_2, defined below. First set

$$v_1(\eta_1, \ldots, \eta_N) = \int_0^{\eta_1} \frac{1}{|\vec{\beta}|}\, w(s, \eta_2, \ldots, \eta_N)\, ds,$$

where $|\vec{\beta}|$ denotes the length of the vector $\vec{\beta}$. Note that v_2, when restricted to each face of $\partial_{\text{out}}K$ is a polynomial of degree at most $p+1$. Now let

$$v_2 = \mathcal{E}_{\text{out}}(\mu - v_1|_{\partial_{\text{out}}K}).$$

It is easy to verify that $v = v_1 + v_2$ satisfies both the equalities of (15). The uniqueness of v is also easy to see: If v satisfies (15) with $w = 0$ and $\mu = 0$, then integrating into K from the outflow boundary along the $\vec{\beta}$-direction, we find that $v \equiv 0$. □

One can now define a new finite element formally in the sense of Ciarlet [7] as follows. Let Σ denote the set of linear functionals ℓ_q, ℓ_η defined by

$$\begin{aligned} \ell_q(v) &= (\vec{\beta} \cdot \vec{\nabla}\, v, q)_K, && \text{for all } q \in P_p(K), \\ \ell_\eta(v) &= \langle v, \eta\rangle_F, && \text{for all } \eta \in P_{p+1}(F), \\ & && \text{for all faces } F \subseteq \partial_{\text{out}}K. \end{aligned}$$

Then, with Σ as the set of degrees of freedom (d.o.f), the geometry-space-d.o.f triple $(K, V_p(K), \Sigma)$ defines a unisolvent finite element because of Lemma 2.1. Although $V_p(K)$

may contain discontinuous functions, implementation of this finite element can proceed using standard finite element technology using the above d.o.fs.

Proposition 2.2. *There is a unique solution to the spectral DPG method* (13)*.*

Proof. The system (13) is square. This follows because the space of trial functions has dimension

$$\begin{aligned}\dim(X_p(K)) &= \dim(P_p(K)) + \dim(M_{p+1}(\partial_{\text{out}}K))\\ &= \dim(\eta_1 P_p(K)) + \dim(\mathcal{E}_{\text{out}}(M_{p+1}(\partial_{\text{out}}K)))\end{aligned}$$

which equals the dimension of the test space $V_p(K)$, due to Proposition 2.1.

Hence it suffices to show that if the right hand side of (13) vanishes, the only possible solution is trivial. But this is immediate from Lemma 2.1, by which we may choose v such that $\vec{\beta}\cdot\vec{\nabla}\,v = -u_p$ and $v|_{\partial_{\text{out}}K} = \phi_{p+1}$. □

Let us now outline an error analysis. We want to apply Theorem 1.1. The continuity of the form $a(\cdot,\cdot)$ is evident from

$$\begin{aligned} a(\,(u_p,\phi_{p+1}),v) &= -(u_p,\vec{\beta}\cdot\vec{\nabla}\,v)_K + \langle \phi_{p+1}, v\rangle_{\partial K_o}\\ &\le \|u_p\|_K\|\vec{\beta}\cdot\vec{\nabla}\,v\|_K + h_K^{-1/2}\|\phi_{p+1}\|_{\partial_{\text{out}}K} h_K^{1/2}\|v\|_{\partial_{\text{out}}K} \qquad (16)\\ &\le \left(\|u_p\|_K^2 + h_K^{-1}\|\phi_{p+1}\|^2_{\partial_{\text{out}}K}\right)^{1/2} \|\!|v|\!\|_K \end{aligned}$$

where we have chosen the test space norm

$$\|\!|v|\!\|_K = \left(\|\vec{\beta}\cdot\vec{\nabla}\,v\|_K^2 + h_K\|v\|^2_{\partial_{\text{out}}K}\right)^{1/2}.$$

The inf-sup condition also holds, because applying Lemma 2.1, we can find a v such that $-\vec{\beta}\cdot\vec{\nabla}\,v = u_p$ and $v|_{\partial_{\text{out}}K} = h_K^{-1}\phi_{p+1}$, so

$$\begin{aligned}\sup_{v\in V_p(K)} \frac{a(\,(u_p,\phi_{p+1}),v)}{\|\!|v|\!\|_K} &\ge \frac{\|u_p\|_K^2 + h_K^{-1}\|\phi_{p+1}\|^2_{\partial_{\text{out}}K}}{\|\!|v|\!\|_K}\\ &= \frac{\left(\|u_p\|_K^2 + h_K^{-1}\|\phi_{p+1}\|^2_{\partial_{\text{out}}K}\right)^{1/2}\|\!|v|\!\|_K}{\|\!|v|\!\|_K}.\end{aligned}$$

Thus applying Theorem 1.1, we obtain the following error estimate for the spectral method:

$$\begin{aligned}&\|u-u_p\|_K^2 + h_K^{-1}\|\phi-\phi_{p+1}\|_K^2\\ &\le 2\inf_{(w_p,\mu_{p+1})\in X_p(K)}\left(\|u-w_p\|_K^2 + h_K^{-1}\|\phi-\mu_{p+1}\|_K^2\right).\end{aligned}$$

Here $h_K = \operatorname{diam}(K)$. Reviewing the above argument, the introduction of factors h_K and h_K^{-1} in (16) may seem arbitrary, and indeed it is. While we chose these factors above for simplicity, later we will need to introduce different mesh dependent factors to obtain stability in more appropriate norms.

Although we used the inf-sup condition to motivate the design of our method, the way we used it above does not provide us with the best error estimate possible. It is possible to obtain better error estimates, and indeed to show that the error for each variable is decoupled. In fact, both variables possess superconvergence in the spectral case, as we see next.

Theorem 2.1. *Let $\mathit{\Pi}_W$ and $\mathit{\Pi}_M$ denote the L^2-orthogonal projections into $P_p(K)$ and $M_{p+1}(\partial_{\text{out}}K)$, respectively. Then*

$$\begin{aligned} \phi_{p+1} - \mathit{\Pi}_M \phi &= 0, \\ u_p - \mathit{\Pi}_W u &= 0. \end{aligned}$$

Proof. Subtracting (11) from (13), we find that

$$-(u_p - \mathit{\Pi}_M u, \vec{\beta} \cdot \vec{\nabla}\, v)_K + \langle \phi_{p+1} - \mathit{\Pi}_M \phi, v\rangle_{\partial_{\text{out}}K} = 0$$

for all v in $V_p(K)$. By Lemma 2.1, we can choose v in $V_p(K)$ so that $v|_{\partial_{\text{out}}K} = \mathcal{E}_{\text{out}}(\phi_{p+1} - \mathit{\Pi}_M \phi)$ and $\beta \cdot \vec{\nabla}\, v = \mathit{\Pi}_M u - u_p$. Hence the identities of the theorem follow. □

2.2. **Local stability estimates.** Although our error analysis of the spectral method is already completed (by Theorem 2.1), we want to obtain a few refined stability estimates for the solutions. This will be needed later, when we analyze the multi-element version of the method. To this end, we introduce the following *local solution operators*:

$$\begin{aligned} \mathcal{Q}_p &: L^2(\partial_{\text{in}}K) \longmapsto M_{p+1}(\partial_{\text{out}}K), \\ \mathcal{U}_p &: L^2(\partial_{\text{in}}K) \longmapsto P_p(K). \end{aligned}$$

Given ϕ in $L^2(\partial_{\text{in}}K)$, the local solutions $\mathcal{Q}_p\phi$ and $\mathcal{U}_p\phi$ are defined by the spectral method (13) with $f = 0$, element by element, as follows:

$$-(\mathcal{U}_p\phi, \vec{\beta} \cdot \vec{\nabla}\, v)_K + \langle \mathcal{Q}_p\phi, v\rangle_{\partial_{\text{out}}K} = \langle \phi, v\rangle_{\partial_{\text{in}}K} \tag{17}$$

for all $v \in V_p(K)$. Also define solution operators for element sources

$$\begin{aligned} \mathcal{Q}_p^f &: L^2(K) \longmapsto M_{p+1}(\partial_{\text{out}}K), \\ \mathcal{U}_p^f &: L^2(K) \longmapsto P_p(K) \end{aligned}$$

by

$$-(\mathcal{U}_p^f z, \vec{\beta} \cdot \vec{\nabla}\, v)_K + \langle \mathcal{Q}_p^f z, v\rangle_{\partial_{\text{out}}K} = (z, v)_K, \tag{18}$$

for all $v \in V_p(K)$ and any z in $L^2(K)$.

The next theorem is the main result of this subsection. Let $h_{\beta,K}$ denote the length of the longest line segment, in the $\vec{\beta}$-direction, contained in K, with one endpoint in $\partial_{\text{in}}K$ and the other in $\partial_{\text{out}}K$ (as marked in Fig. 10(b)). We denote by $\vec{n}_F$ a unit normal to a mesh face F. For any subcollection $\mathscr{F}_h$ of the set of faces of K where $\vec{\beta} \cdot \vec{n} \neq 0$, we define the following norms:

$$\|\psi\|^2_{\frac{1}{\beta},\mathscr{F}_h} = \sum_{F \subset \mathscr{F}_h} \frac{1}{|\vec{\beta} \cdot \vec{n}_F|}\, \|\psi\|^2_F, \tag{19}$$

$$\|v\|^2_{\beta,\mathscr{F}_h} = \sum_{F \subset \mathscr{F}_h} |\vec{\beta} \cdot \vec{n}_F|\; \|v\|^2_F. \tag{20}$$

Here the sums run over all mesh faces F contained in $\mathscr{F}_h$.

Theorem 2.2. *The local solution operators have the following stability estimates:*

$$\|\mathcal{Q}_p\phi\|^2_{\frac{1}{\beta},\partial_{\text{out}}K} \le \|\phi\|^2_{\frac{1}{\beta},\partial_{\text{in}}K}, \tag{21a}$$

$$\|\mathcal{U}_p\phi\|^2_K \le \frac{h_{\beta,K}}{|\vec{\beta}|}\|\phi\|^2_{\frac{1}{\beta},\partial_{\text{in}}K}, \tag{21b}$$

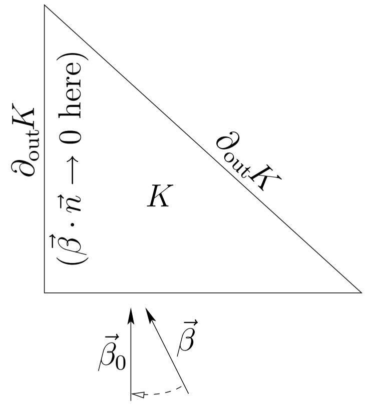


FIGURE 2. A outflow edge degenerating as $\vec{\beta} \to \vec{\beta}_0$

and

$$\|\mathcal{Q}_p^f z\|^2_{\frac{1}{\beta},\partial_{\text{out}}K} \le \frac{h_{\beta,K}}{|\vec{\beta}|}\|z\|^2_K, \tag{22a}$$

$$\|\mathcal{U}_p^f z\|^2_K \le \frac{h^2_{\beta,K}}{|\vec{\beta}|^2}\|z\|^2_K, \tag{22b}$$

for all $\phi \in L^2(\partial_{\text{in}}K)$ *and* z *in* $L^2(K)$.

A proof of the theorem is given in Appendix A.

2.3. **Remarks on the robustness of the outflux variable.** Notice that on faces where $\vec{\beta}\cdot\vec{n} = 0$, we have not yet defined any outflux approximation. Since the exact outflux (defined in (26)) vanishes on such faces, we set the discrete outflux ϕ_h to zero on those faces.

Computationally, the identification of inflow and outflow faces can only be as accurate as the round-off errors permit, i.e., the strict inequalities (9) can only be implemented up to round-off. This begs the question: On faces F where $\vec{\beta}\cdot\vec{n}_F$ is close to zero, is ϕ_p too sensitive to whether we set F to be an inflow or outflow face?

We now argue that it is not. For clarity, consider the situation of Figure 2 (which is entirely typical). The solution ϕ_p of (13) can be written as $\phi_p = \mathcal{Q}_p g + \mathcal{Q}_p^f f$. Then, by the estimates of Theorem 2.2, we have

$$\|\phi_p\|_{\frac{1}{\beta},\partial_{\text{out}}K} \le \|g\|_{\frac{1}{\beta},\partial_{\text{in}}K} + \frac{h^{1/2}_{\beta,K}}{|\vec{\beta}|^{1/2}}\|f\|_K.$$

This continues to hold as $\vec{\beta}\cdot\vec{n}$ approaches zero on an outflow face (as in Fig. 2). Note that the right hand side remains bounded during this limiting process. But since the left hand side norm contains the factor $|\vec{\beta}\cdot\vec{n}|^{-1/2}$ on the outflow face where $\vec{\beta}\cdot\vec{n}$ approaches zero, the solution ϕ_p must approach zero there.

It is this robustness in ϕ_p that led us to introduce it as the new 'outflux variable' to be solved for. The choice of such a variable is not traditional. In hybridized methods for elliptic problems [9] one usually sets the trace of the 'primal' variable u as a new unknown.

In this spirit, one could contemplate a method where one finds u_p and λ_p satisfying

$$-(u_p, \vec{\beta} \cdot \vec{\nabla}\, v)_K + \langle \vec{\beta} \cdot \vec{n}\, \lambda_p, v \rangle_{\partial_{\mathrm{out}} K} = (f, v)_K - \langle \vec{\beta} \cdot \vec{n}\, g, v \rangle_{\partial_{\mathrm{in}} K}, \tag{23}$$

instead of (13). Such methods have appeared in the literature – the standard UDG method can be recast into hybrid form with the upwind traces set as a new unknown λ_h, see e.g., the method in [10] for the purely convective case). However, an equation such as (23) can lead to matrices close to singularity in situations like in Figure 2. Indeed, the λ_h in (23) cannot be uniquely defined where $\vec{\beta} \cdot \vec{n} = 0$.

3. Approximations using a triangular mesh

In this section, we move to the multi-element case, restricting ourselves to the case of two space dimensions. We consider the composite DPG method on a mesh of the domain Ω, and a general polynomial degree p. Since the degree is uniformly set to p on all elements this is the "p-version" of the method.

We assume that the domain Ω is meshed by a geometrically conforming mesh of triangles, denoted by $\mathscr{T}_h$, and set $h = \max\{h_K : K \in \mathscr{T}_h\}$. Define the (discontinuous) finite element spaces

$$W_h = \{w : w|_K \in P_p(K), \quad \text{for all mesh elements } K \in \mathscr{T}_h\}, \tag{24a}$$

$$M_h = \{\, \mu : \mu|_F \in P_{p+1}(F), \quad \text{for all mesh edges } F \text{ not on } \partial_{\mathrm{in}}\Omega\}, \tag{24b}$$

$$V_h = \{\, v : v|_K \in V_p(K), \quad \text{for all mesh elements } K \in \mathscr{T}_h\}. \tag{24c}$$

Here $V_p(K)$ is the new finite element we introduced in Section 2.

3.1. **Definition of the composite method.**

We put together the spectral method on each element to get the following composite method on the mesh $\mathscr{T}_h$: Find $(u_h, \phi_h) \in X_h \overset{\mathrm{def}}{=} W_h \times M_h$ satisfying

$$a_h(\,(u_h, \phi_h), v_h) = (f, v_h)_\Omega - \langle \vec{\beta} \cdot \vec{n}\, g, v_h \rangle_{\partial_{\mathrm{in}}\Omega}, \tag{25}$$

for all $v_h \in V_h$, where

$$a_h(\,(u_h, \phi_h), v_h) = \sum_{K \in \mathscr{T}_h} \Big[\langle \phi_h, v_h \rangle_{\partial_{\mathrm{out}} K} - \langle \phi_h, v_h \rangle_{\partial_{\mathrm{in}} K \setminus \partial_{\mathrm{in}}\Omega} - (u_h, \vec{\beta} \cdot \vec{\nabla}\, v_h)_K \Big].$$

We call ϕ_h the *outflux approximation*, because on comparison with (11), we expect it to approximate

$$\phi \overset{\mathrm{def}}{=} \vec{\beta} u \cdot \vec{n}_+ \tag{26}$$

on every mesh edge F, where $\vec{n}_+$ is the unit normal on F with its direction chosen so that $\vec{\beta} \cdot \vec{n}_+ > 0$. On edges where $\vec{\beta} \cdot \vec{n} = 0$, we set $\phi_h = 0$. Equation (11) also shows that this ϕ together with the exact solution u, satisfies

$$a_h(\,(u, \phi), v) = (f, v)_\Omega - \langle \vec{\beta} \cdot \vec{n}\, g, v \rangle_{\partial_{\mathrm{in}}\Omega}, \tag{27}$$

for all v in $V = \{\nu \in L^2(\Omega) : \nu|_K \in V(K)$ for all mesh elements $K\}$, where $V(K)$ is as in (55). In other words, the DPG method is consistent. In order to simplify notation and write $\langle \phi_h, v_h\rangle_{\partial_{\rm in}K}$ for $\langle \phi_h, v_h\rangle_{\partial_{\rm in}K\setminus\partial_{\rm in}\Omega}$, we identify the space M_h with its extension by zero to $\partial_{\rm in}\Omega$, i.e., we identify M_h and

$$\{\mu :\ \mu|_F \in P_{p+1}(F) \text{ for all mesh edge } F \text{ and } \mu|_{\partial_{\rm in}\Omega} = 0\},$$

to be the same (cf. the definition in (24b)).

3.2. **Mesh dependent norms.** We collect the definitions of various mesh dependent norms here for ready reference.

First, we extend the definition of the norms $\|\psi\|^2_{\frac{1}{\beta},\mathscr{F}_h}$ and $\|v\|_{\beta,\mathscr{F}_h}$ to more general collections $\mathscr{F}_h$ of edges from the entire mesh (now not necessarily edges of one triangle) simply by letting the sums in (19) and (20) runs over all such edges in $\mathscr{F}_h$. As before, only edges with $\vec\beta\cdot\vec n \neq 0$ are picked. Using these extended definitions, we now define more norms on a subcollection of mesh elements $\mathscr{R}_h$, as follows:

$$\|v\|^2_{h,\beta,\mathscr{R}_h} = \sum_{K\in\mathscr{R}_h} h_{\beta,K}\,\|v\|^2_{\beta,\partial_{\rm out}K}\,, \tag{28}$$

$$\|\psi\|^2_{h,\frac{1}{\beta},\mathscr{R}_h} = \sum_{K\in\mathscr{R}_h} h_{\beta,K}\|\psi\|^2_{\frac{1}{\beta},\partial_{\rm out}K}, \tag{29}$$

$$\|\psi\|^2_{\frac{1}{h},\frac{1}{\beta},\mathscr{R}_h} = \sum_{K\in\mathscr{R}_h} \frac{1}{h_{\beta,K}}\|\psi\|^2_{\frac{1}{\beta},\partial_{\rm out}K}, \tag{30}$$

$$|\!|\!|\psi|\!|\!|^2_{\frac{1}{h},\frac{1}{\beta},\mathscr{R}_h} = \|\psi - \mathcal{Q}_p\psi\|^2_{\frac{1}{h},\frac{1}{\beta},\mathscr{R}_h}. \tag{31}$$

These norm notations have mnemonic subscripts of $h, 1/h, \beta$, or $1/\beta$ etc., which indicates a factor of $h_{\beta,K}, 1/h_{\beta,K}, |\vec\beta\cdot\vec n|$, or $1/|\vec\beta\cdot\vec n|$ (resp.) in the sums over mesh objects defining the norms. With these notations, the norm on the trial space is defined by

$$|\!|\!|(u,\phi)|\!|\!|^2_h = \sum_{K\in\mathscr{T}_h}\left(\|u - \mathcal{U}_p\phi\|^2_K + \|\psi - \mathcal{Q}_p\psi\|^2_{\frac{1}{h},\frac{1}{\beta},K}\right).$$

The norm on the test space is defined by

$$|\!|\!|v|\!|\!|^2_h = \sum_{K\in\mathscr{T}_h} \|\vec\beta\cdot\vec\nabla\, v\|^2_K + \|v\|^2_{h,\beta,K}.$$

It is obvious from the construction of $V_p(K)$ that the above is a norm on V_h. To show that $|\!|\!|(u,\phi)|\!|\!|_h$ is a norm on the test space, we first need to recall the following.

Lemma 3.1. *In any triangulation $\mathscr{T}_h$ of Ω, there is at least one triangle $K\in\mathscr{T}_h$ such that $\partial_{\rm in}K \subseteq \partial_{\rm in}\Omega$.*

Proof. See [15]. □

Proposition 3.1. *The functional $|\!|\!|(u,\phi)|\!|\!|_h$ is a norm on the space $X = L^2(\Omega)\times M$, where*

$$M = \{\mu :\ \mu|_F \in L^2(F) \text{ for all mesh edges } F \text{ and } \mu|_{\partial_{\rm in}\Omega} = 0\}.$$

Proof. Suppose $|\!|\!|(u,\phi)|\!|\!| = 0$. Then, in particular

$$\|\phi - \mathcal{Q}_p\phi\|_{\beta,\partial_{\rm out}K} = 0 \tag{32}$$

for all elements K. By Lemma 3.1, there is an element K with $\partial_{\text{in}}K \subseteq \partial_{\text{in}}\Omega$. On this element, $\phi|_{\partial_{\text{in}}K} = 0$, so $\mathcal{Q}_p\phi|_{\partial_{\text{out}}K} = 0$. Hence (32) implies that $\phi|_{\partial_{\text{out}}K} = 0$. Proceeding inductively over all elements we find that $\phi = 0$. This in turn implies that $u - \mathcal{U}_p\phi = u = 0$. □

3.3. **Preliminary error analysis.** It is possible to get stability and error estimates for the DPG method in the mesh dependent norm immediately. Solvability of the method follows as a particular consequence.

Theorem 3.1. *The bilinear form $a_h(\cdot,\cdot)$ satisfies the inf-sup and continuity conditions:*

$$\sup_{v_h\in V_h} \frac{a_h(\,(u_h,\phi_h),v_h)}{|\!|\!| v_h |\!|\!|_h} \ge |\!|\!| (u_h,\phi_h) |\!|\!|_h,$$

holds for all $u_h \in W_h$ and $\phi_h \in M_h$, and

$$a_h(\,(u,\phi),v_h) \le |\!|\!| (u,\phi) |\!|\!|_h |\!|\!| v_h |\!|\!|_h$$

for all $u \in L^2(\Omega), \phi \in M$, and $v_h \in V_h$. If (u,ϕ) and (u_h,ϕ_h) are the exact and discrete solutions, then

$$|\!|\!| (u-u_h,\phi-\phi_h) |\!|\!|_h^2 \le 2 \inf_{(w_h,\mu_h)\in W_h\times M_h} |\!|\!| (u-w_h,\phi-\mu_h) |\!|\!|_h^2.$$

Proof. We first rewrite (25) using the definition of the local solution operators in (17) as follows:

$$\begin{aligned} & a_h(\,(u_h,\phi_h),v_h) \\ &= \sum_{K\in\mathscr{T}_h} \big[-(u_h,\vec\beta\cdot\vec\nabla\, v_h)_K + \langle \phi_h, v_h\rangle_{\partial_{\text{out}}K}\big] - \langle \phi_h, v_h\rangle_{\partial_{\text{in}}K} \\ &= \sum_{K\in\mathscr{T}_h} -(u_h - \mathcal{U}_p\phi_h, \vec\beta\cdot\vec\nabla\, v_h)_K + \langle \phi_h - \mathcal{Q}_p\phi_h, v_h\rangle_{\partial_{\text{out}}K}. \end{aligned} \tag{33}$$

Let us prove the inf-sup condition. By Lemma 2.1, we can choose v_h element by element, such that

$$\begin{aligned} -\vec\beta\cdot\vec\nabla\, v_h\big|_K &= u_h - \mathcal{U}_p(\phi_h\big|_{\partial_{\text{in}}K}), \\ v_h\big|_{\partial_{\text{out}}K} &= \frac{(\phi_h - \mathcal{Q}_p(\phi_h\big|_{\partial_{\text{in}}K}))\big|_{\partial_{\text{out}}K}}{h_{\beta,K}|\vec\beta\cdot\vec n|}. \end{aligned}$$

Then

$$\begin{aligned} a_h(\,(u_h,\phi_h),v_h) =& \sum_{K\in\mathscr{T}_h} \|u_h - \mathcal{U}_p\phi_h\|_K^2 + \|\phi_h - \mathcal{Q}_p\phi_h\|^2_{\frac{1}{h},\frac{1}{\beta},K} \\ =& \ |\!|\!| (u_h,\phi_h) |\!|\!|_h \, |\!|\!| v_h |\!|\!|_h. \end{aligned}$$

This proves the inf-sup condition.

The proof of the continuity inequality also follows from (33) after applying the Cauchy-Schwarz inequality.

The error estimate then immediately follows from Theorem 1.1. □

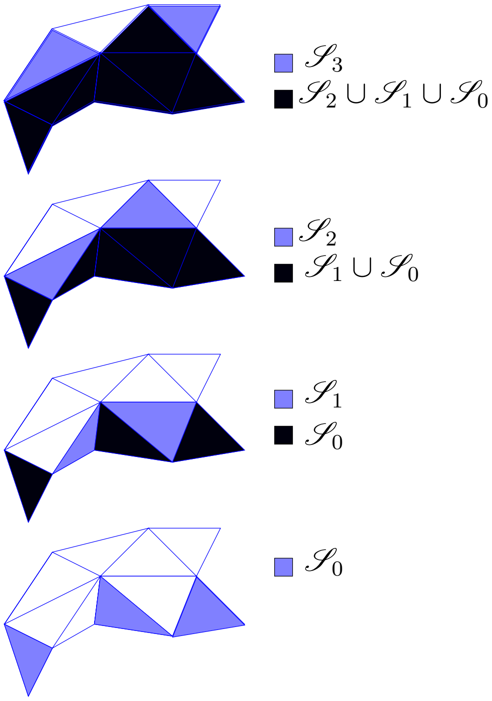

$\mathscr{S}_3$
$\mathscr{S}_2 \cup \mathscr{S}_1 \cup \mathscr{S}_0$
$\mathscr{S}_2$
$\mathscr{S}_1 \cup \mathscr{S}_0$
$\mathscr{S}_1$
$\mathscr{S}_0$
$\mathscr{S}_0$

FIGURE 3. A mesh being split into layers (flow is along the positive y-direction).

The above analysis is in the most "natural" norms suggested by the method itself. In fact, we have shown above that

$$\sup_{v_h \in V_h} \frac{a_h(\,(u_h, \phi_h), v_h)}{|||v_h|||_h} = |||(u_h, \phi_h)|||_h,$$

for all $u_h \in W_h$, $\phi_h \in M_h$. Nonetheless, convergence rates in standard norms (like L^2) are not obvious from the theorem. The remainder of this section is devoted to obtaining convergence rates in L^2-like norms.

3.4. **A Poincaré inequality.** As a first step to obtain stability and error estimates in L^2-norm, we prove a Poincaré inequality. Notice that the norm in (31) is applied to the difference $\psi\big|_{\partial_{\mathrm{out}}K} - \mathcal{Q}_p(\psi\big|_{\partial_{\mathrm{in}}K})$. This is the difference between the outflow values of ψ and its inflow values transported onto $\partial_{\mathrm{out}}K$. Discrete analogues of the Poincaré inequality bound L^2-like norms using sum of squares of differences (under some discrete analogue of a zero boundary condition). Therefore, it is perhaps not surprising that such a discrete Poincaré type estimate can also be found for the norm in (31) (recalling that any ψ in M_h satisfies a zero boundary condition on the inflow boundary).

To establish the inequality, we first need to define "layers" of mesh elements marching from $\partial_{\mathrm{in}}\Omega$. The first layer is defined by

$$\mathscr{S}_0 = \{K \in \mathscr{T}_h : \partial_{\mathrm{in}}K \subseteq \partial_{\mathrm{in}}\Omega\}.$$

By Lemma 3.1, $\mathscr{S}_0$ is not empty. Next, setting $\mathscr{T}_h^{(0)} = \mathscr{T}_h$, we recursively define, for all integers $\ell > 0$, the sets

$$\begin{aligned} \mathscr{T}_h^{(\ell)} &= \mathscr{T}_h^{(\ell-1)} \setminus \mathscr{S}_{\ell-1} \\ \mathscr{S}_\ell &= \{K \in \mathscr{T}_h^{(\ell)} : \partial_{\text{in}} K \subseteq \partial_{\text{in}} \Omega^{(\ell)}\}, \end{aligned}$$

where $\Omega^{(\ell)}$ is the domain formed by the union of all triangles of $\mathscr{T}_h^{(\ell)}$. Repeated application of Lemma 3.1 shows that $\mathscr{S}_\ell$ is nonempty for every ℓ, unless $\mathscr{T}_h^{(\ell)}$ is empty. So this process exhausts all elements of the mesh, at some finite value of ℓ, say $\ell = n$. The domain $\Lambda_\ell \overset{\text{def}}{=} \Omega \setminus \Omega^{(\ell)}$ is meshed by $\mathscr{S}_0 \cup \mathscr{S}_1 \cdots \cup \mathscr{S}_\ell$. Its outflow boundary is denoted by

$$\Gamma_\ell = \partial_{\text{out}} \Lambda_\ell.$$

All mesh edges, excluding those on $\partial_{\text{in}}\Omega$, and excluding those where $\vec{\beta} \cdot \vec{n} = 0$, are contained in Γ_ℓ for some $\ell \le n$. With these notations and observations, we impose a mild assumption on the mesh and prove the Poincaré inequality.

Assumption 3.1. Define the *layer width* by

$$d_\ell = \max\{h_{\beta,K} : \ K \in \mathscr{S}_\ell\}.$$

We assume that our meshes are such that

$$\sum_{\ell=0}^{n} d_\ell \le L_\Omega$$

where L_Ω is a fixed constant depending on Ω, but independent of the meshes and element sizes. We also assume that there is a fixed constant C_1 such that if K and K' are any two neighboring elements, then either the inequality

$$h_{\beta,K'} \le C_1 h_{\beta,K}$$

holds or it holds after exchanging K and K'.

Above and in the remainder, the letter C, with or without subscripts, denotes generic constants independent of elements K and polynomial degree p. Their value at different occurrences may vary.

Assumption 3.1 permits quite general meshes, including anisotropic refinements (e.g., to capture shock waves). Loosely speaking, the first part assumes that elements within each layer have about the same length in the flow direction. The second part assumes that element lengths in the flow direction for neighboring elements are comparable. Quasiuniform meshes obviously satisfy the assumption.

Theorem 3.2. *If Assumption 3.1 holds, then for all $\psi \in M_h$, we have*

$$\|\psi\|^2_{h,\frac{1}{\beta},\mathscr{T}_h} \le C_0 L_\Omega^2 \|\!|\psi|\!\|^2_{\frac{1}{h},\frac{1}{\beta},\mathscr{T}_h}.$$

(where C_0 is independent of $\mathscr{T}_h$ and p, and the norms are as in (29) *and* (31)*).*

Proof. The first step is a local inequality. On any triangle K,

$$\begin{aligned} \|\psi\|_{\frac{1}{\beta},\partial_{\text{out}}K} &\le \|\psi - \mathcal{Q}_p\psi\|_{\frac{1}{\beta},\partial_{\text{out}}K} + \|\mathcal{Q}_p\psi\|_{\frac{1}{\beta},\partial_{\text{out}}K} \\ &\le \|\psi - \mathcal{Q}_p\psi\|_{\frac{1}{\beta},\partial_{\text{out}}K} + \|\psi\|_{\frac{1}{\beta},\partial_{\text{in}}K}, \end{aligned}$$

where the last inequality was due to Theorem 2.2 (21). Consequently,

$$\begin{aligned}&\|\psi\|^2_{\frac{1}{\beta},\partial_{\rm out}K} - \|\psi\|^2_{\frac{1}{\beta},\partial_{\rm in}K}\\ &\le \big(\|\psi\|_{\frac{1}{\beta},\partial_{\rm out}K} + \|\psi\|_{\frac{1}{\beta},\partial_{\rm in}K}\big)\|\psi - \mathcal{Q}_p\psi\|_{\frac{1}{\beta},\partial_{\rm out}K}\end{aligned} \tag{34}$$

holds on every mesh element K.

The next step is to sum over $\mathscr{R}_\ell = \mathscr{S}_0 \cup \mathscr{S}_1 \cdots \cup \mathscr{S}_\ell$. Then, we have

$$\begin{aligned}&\sum_{K\in\mathscr{R}_\ell} \big(\|\psi\|^2_{\frac{1}{\beta},\partial_{\rm out}K} - \|\psi\|^2_{\frac{1}{\beta},\partial_{\rm in}K}\big)\\ &\le \sum_{K\in\mathscr{R}_\ell} \big(\|\psi\|_{\frac{1}{\beta},\partial_{\rm out}K} + \|\psi\|_{\frac{1}{\beta},\partial_{\rm in}K}\big)\|\psi - \mathcal{Q}_p\psi\|_{\frac{1}{\beta},\partial_{\rm out}K}.\end{aligned}$$

When rewriting the left hand side as a sum over the mesh edges, we observe that all contributions due to edges interior to Λ_ℓ cancel out. Furthermore, since $\psi \in M_h$, it vanishes on the inflow boundary $\partial_{\rm in}\Lambda_\ell \subseteq \partial_{\rm in}\Omega$. Therefore only contributions on the outflow boundary of Λ_ℓ, namely Γ_ℓ, remains, i.e.,

$$\begin{aligned}&\|\psi\|^2_{\frac{1}{\beta},\Gamma_\ell} \le\\ &\sum_{K\in\mathscr{R}_\ell} \big(\|\psi\|_{\frac{1}{\beta},\partial_{\rm out}K} + \|\psi\|_{\frac{1}{\beta},\partial_{\rm in}K}\big)\|\psi - \mathcal{Q}_p\psi\|_{\frac{1}{\beta},\partial_{\rm out}K}.\end{aligned} \tag{35}$$

We apply Cauchy-Schwarz inequality to the right hand side of (35) to get

$$\begin{aligned}\|\psi\|^2_{\frac{1}{\beta},\Gamma_\ell} \le &\bigg(\sum_{K\in\mathscr{R}_\ell} 2h_{\beta,K}\big(\|\psi\|^2_{\frac{1}{\beta},\partial_{\rm out}K} + \|\psi\|^2_{\frac{1}{\beta},\partial_{\rm in}K}\big)\bigg)^{1/2}\\ &\bigg(\sum_{K\in\mathscr{R}_\ell} \frac{1}{h_{\beta,K}}\|\psi - \mathcal{Q}_p\psi\|^2_{\frac{1}{\beta},\partial_{\rm out}K}\bigg)^{1/2}\end{aligned}$$

By Assumption 3.1, the lengths $h_{\beta,K}$ for neighboring elements are comparable, so that

$$\sum_{K\in\mathscr{R}_\ell} h_{\beta,K}\|\psi\|^2_{\frac{1}{\beta},\partial_{\rm in}K} \le C\sum_{K\in\mathscr{R}_\ell} h_{\beta,K}\|\psi\|^2_{\frac{1}{\beta},\partial_{\rm out}K}$$

which implies

$$\begin{aligned}\|\psi\|^2_{\frac{1}{\beta},\Gamma_\ell} &\le C\bigg(\sum_{K\in\mathscr{R}_\ell} h_{\beta,K}\|\psi\|^2_{\frac{1}{\beta},\partial_{\rm out}K}\bigg)^{1/2}\\ &\qquad\bigg(\sum_{K\in\mathscr{R}_\ell} \frac{1}{h_{\beta,K}}\|\psi - \mathcal{Q}_p\psi\|^2_{\frac{1}{\beta},\partial_{\rm out}K}\bigg)^{1/2}\\ &= C\|\psi\|_{h,\frac{1}{\beta},\mathscr{R}_\ell}|\!|\!|\psi|\!|\!|_{\frac{1}{h},\frac{1}{\beta},\mathscr{R}_\ell}.\end{aligned}$$

The final step involves multiplying by the layer width d_ℓ, and summing over ℓ,

$$\begin{aligned}\sum_{\ell=0}^{n} d_\ell\|\psi\|^2_{\frac{1}{\beta},\Gamma_\ell} &\le \sum_{\ell=0}^{n} d_\ell C\|\psi\|_{h,\frac{1}{\beta},\mathscr{R}_\ell}|\!|\!|\psi|\!|\!|_{\frac{1}{h},\frac{1}{\beta},\mathscr{R}_\ell}\\ &\le CL_\Omega\|\psi\|_{h,\frac{1}{\beta},\mathscr{T}_h}|\!|\!|\psi|\!|\!|_{\frac{1}{h},\frac{1}{\beta},\mathscr{T}_h},\end{aligned} \tag{36}$$

where we increased the norms over $\mathscr{R}_\ell$ to norms over the whole mesh $\mathscr{T}_h$. By the definition of d_ℓ, we know that

$$h_{\beta,K} \le d_\ell, \qquad \text{for all } K \in \mathscr{S}_\ell,$$

so we have

$$\begin{aligned}
\|\psi\|^2_{h,\frac{1}{\beta},\mathscr{T}_h} &= \sum_{\ell=0}^{n} \sum_{K\in\mathscr{S}_\ell} h_{\beta,K} \|\psi\|^2_{\frac{1}{\beta},\partial_{\mathrm{out}}K} \\
&\le \sum_{\ell=0}^{n} \sum_{K\in\mathscr{S}_\ell} d_\ell \|\psi\|^2_{\frac{1}{\beta},\partial_{\mathrm{out}}K} \\
&\le L_\Omega C \|\psi\|_{h,\frac{1}{\beta},\mathscr{T}_h} |\!|\!|\psi|\!|\!|_{\frac{1}{h},\frac{1}{\beta},\mathscr{T}_h},
\end{aligned}$$

where the last inequality is due to (36). This proves the result. □

3.5. L^2**-error estimates.** Now we prove an optimal error estimate for u_h in the standard $L^2(\Omega)$-norm, as well as error estimates for the outflux approximation ϕ_h in L^2-like norms on mesh edges. The estimates are with constants independent of the polynomial degree p and mesh $\mathscr{T}_h$.

Theorem 3.3. *Suppose (u,ϕ) is the exact solution, (u_h,ϕ_h) is the discrete solution. Then we have the following error estimates:*

$$|\!|\!|\phi - \phi_h|\!|\!|^2_{\frac{1}{h},\frac{1}{\beta},K} \le 3 \inf_{\psi_{p+1}\in M_h} \frac{1}{h_{\beta,K}} \|\phi - \psi_{p+1}\|^2_{\frac{1}{\beta},\partial K}. \tag{37}$$

If Assumption 3.1 holds, letting

$$\epsilon(\phi) \overset{\text{def}}{=} \inf_{\psi_h\in M_h} \left(\|\phi - \psi_h\|^2_{h,\frac{1}{\beta},\mathscr{T}_h} + C_0 L_\Omega^2 \|\phi - \psi_h\|^2_{\frac{1}{h},\frac{1}{\beta},\mathscr{T}_h} \right),$$

we have

$$\begin{aligned}
\|\phi - \phi_h\|^2_{h,\frac{1}{\beta},\mathscr{T}_h} &\le 2\,\epsilon(\phi) \qquad \textit{and} \qquad (38)\\
\|u - u_h\|^2_\Omega &\le \frac{4}{|\vec{\beta}|}\epsilon(\phi) + 2 \inf_{w_h\in W_h} \|u - w_h\|^2_\Omega. \qquad (39)
\end{aligned}$$

Proof. The first step of this proof consists of identifying the equations satisfied by the following discrete error functions:

$$\varepsilon_h^u = \varPi_W u - u_h, \qquad \varepsilon_h^\phi = \varPi_M \phi - \phi_h,$$

where $\mathit{\Pi}_W$ and $\mathit{\Pi}_M$ are L^2-orthogonal projections into W_h and M_h, respectively. Subtracting the exact and discrete equations,

$$\begin{aligned} 0 &= \sum_{K\in\mathscr{T}_h} \Big(-(u-u_h, \vec{\beta}\cdot\vec{\nabla}\, v_h)_K \\ &\qquad + \langle \phi-\phi_h, v_h\rangle_{\partial_{\rm out}K} - \langle \phi-\phi_h, v_h\rangle_{\partial_{\rm in}K}\Big) \\ &= \sum_{K\in\mathscr{T}_h} \Big(-(\mathit{\Pi}_W u-u_h, \vec{\beta}\cdot\vec{\nabla}\, v_h)_K \\ &\qquad + \langle \mathit{\Pi}_M\phi-\phi_h, v_h\rangle_{\partial_{\rm out}K} - \langle \phi-\phi_h, v_h\rangle_{\partial_{\rm in}K}\Big), \end{aligned}$$

where the introduction of $\mathit{\Pi}_W$ is possible as $\vec{\beta}\cdot\vec{\nabla}$ maps $V_p(K)$ into $P_p(K)$. Notice that in the last term we may not replace ϕ by $\mathit{\Pi}_M\phi$ as the inflow traces of v_h need not be polynomial. But we may add and subtract $\mathit{\Pi}_M\phi$ within it. Then, we can express the identity using the discrete error functions as follows:

$$a_h(\,(\varepsilon_h^u, \varepsilon_h^\phi), v_h) = \sum_{K\in\mathscr{T}_h} \langle \phi - \mathit{\Pi}_M\phi, v_h\rangle_{\partial_{\rm in}K},$$

or equivalently (see (33)),

$$\begin{aligned} -(\varepsilon_h^u - \mathcal{U}_p\varepsilon_h^\phi, \vec{\beta}\cdot\vec{\nabla}\, v_h)_K + \langle \varepsilon_h^\phi - \mathcal{Q}_p\varepsilon_h^\phi, v_h\rangle_{\partial_{\rm out}K} \\ = \langle \phi - \mathit{\Pi}_M\phi, v_h\rangle_{\partial_{\rm in}K}, \end{aligned} \tag{40}$$

on every mesh element K. In other words,

$$\varepsilon_h^u - \mathcal{U}_p\varepsilon_h^\phi = \mathcal{U}_p(\phi - \mathit{\Pi}_M\phi), \tag{41a}$$

$$\varepsilon_h^\phi - \mathcal{Q}_p\varepsilon_h^\phi = \mathcal{Q}_p(\phi - \mathit{\Pi}_M\phi). \tag{41b}$$

We shall use these identities to bound the errors.

The second step of this proof involves bounding ε_h^ϕ. By (41b),

$$\begin{aligned} \|\varepsilon_h^\phi - \mathcal{Q}_p\varepsilon_h^\phi\|_{\frac{1}{\beta},\partial_{\rm out}K} &= \|\mathcal{Q}_p(\phi - \mathit{\Pi}_M\phi)\|_{\frac{1}{\beta},\partial_{\rm out}K} \\ &\le \|\phi - \mathit{\Pi}_M\phi\|_{\frac{1}{\beta},\partial_{\rm in}K}, \end{aligned} \tag{42}$$

by the bound (21) on $\mathcal{Q}_p(\cdot)$ of Theorem 2.2. Therefore,

$$\begin{aligned} |\!|\!|\phi-\phi_h|\!|\!|_{\frac{1}{h},\frac{1}{\beta},K} &\le |\!|\!|\phi - \mathit{\Pi}_M\phi|\!|\!|_{\frac{1}{h},\frac{1}{\beta},K} + |\!|\!|\varepsilon_h^\phi|\!|\!|_{\frac{1}{h},\frac{1}{\beta},K} \\ &= h_{\beta,K}^{-1/2}\|(\phi - \mathit{\Pi}_M\phi) - \mathcal{Q}_p(\phi - \mathit{\Pi}_M\phi)\|_{\frac{1}{\beta},\partial_{\rm out}K} \\ &+ h_{\beta,K}^{-1/2}\|\varepsilon_h^\phi - \mathcal{Q}_p\varepsilon_h^\phi\|_{\frac{1}{\beta},\partial_{\rm out}K} \\ &\le h_{\beta,K}^{-1/2}\|\phi - \mathit{\Pi}_M\phi\|_{\frac{1}{\beta},\partial_{\rm out}K} \\ &+ 2h_{\beta,K}^{-1/2}\|\phi - \mathit{\Pi}_M\phi\|_{\frac{1}{\beta},\partial_{\rm in}K}. \end{aligned}$$

By the inequality of arithmetic and geometric means, this implies

$$|\!|\!|\phi-\phi_h|\!|\!|^2_{\frac{1}{h},\frac{1}{\beta},K} \le 3h_{\beta,K}^{-1}\|\phi - \mathit{\Pi}_M\phi\|^2_{\frac{1}{\beta},\partial_{\rm out}K\cup\partial_{\rm in}K},$$

which gives the first error estimate of the theorem.

The third step again involves a bound on ε_h^ϕ, but now in a weaker norm. We use the Poincaré inequality of Theorem 3.2:

$$\begin{aligned}\|\varepsilon_h^\phi\|^2_{h,\frac{1}{\beta},\mathscr{T}_h} &\le C_0 L_\Omega^2 |\!|\!|\varepsilon_h^\phi|\!|\!|^2_{\frac{1}{h},\frac{1}{\beta},\mathscr{T}_h} \\ &\le C_0 L_\Omega^2 \sum_{K\in\mathscr{T}_h} \frac{1}{h_{\beta,K}} \|\phi - \varPi_M \phi\|^2_{\beta,\partial_{\rm in}K},\end{aligned} \tag{43}$$

where we have also used (42) in the last step. Splitting $\phi - \phi_h = (\phi - \varPi_M\phi) + \varepsilon_h^\phi$ as before, and estimating, we prove the second estimate (38) of the theorem.

Now we come to the final step, where we obtain the error estimate for u_h, namely (39), as follows:

$$\begin{aligned}\|\varepsilon_h^u\|_K - \|\mathfrak{U}_p\varepsilon_h^\phi\|_K &\le \|\varepsilon_h^u - \mathfrak{U}_p\varepsilon_h^\phi\|_K && \text{by triangle inequality}\\ &= \|\mathfrak{U}_p(\phi - \varPi_M\phi)\|_K && \text{by (41a)}\\ &\le \frac{h_{\beta,K}^{1/2}}{|\vec\beta|^{1/2}}\|\phi - \varPi_M\phi\|_{\frac{1}{\beta},\partial_{\rm in}K} && \text{by Theorem 2.2.}\end{aligned}$$

Using Theorem 2.2 again, we have

$$\|\varepsilon_h^u\|_K \le \frac{h_{\beta,K}^{1/2}}{|\vec\beta|^{1/2}}\big(\|\varepsilon_h^\phi\|_{\frac{1}{\beta},\partial_{\rm in}K} + \|\phi - \varPi_M\phi\|_{\frac{1}{\beta},\partial_{\rm in}K}\big).$$

Summing over all elements and using (43), we obtain

$$\begin{aligned}\|\varepsilon_h^u\|_\Omega^2 &\le \frac{2}{|\vec\beta|}\big(\|\varepsilon_h^\phi\|^2_{h,\frac{1}{\beta},\mathscr{T}_h} + \sum_{K\in\mathscr{T}_h} h_{\beta,K}\,\|\phi - \varPi_M\phi\|^2_{\beta,\partial_{\rm in}K}\big)\\ &\le \frac{2}{|\vec\beta|}\sum_{K\in\mathscr{T}_h}\bigg(\frac{C_0L_\Omega^2}{h_{\beta,K}}\|\phi - \varPi_M\phi\|^2_{\beta,\partial_{\rm in}K}\\ &\qquad\qquad + h_{\beta,K}\,\|\phi - \varPi_M\phi\|^2_{\beta,\partial_{\rm in}K}\bigg),\end{aligned}$$

from which the estimate (39) follows. □

3.6. **Interior fluxes by postprocessing.** The advective flux is $\vec q = \vec\beta u$. A natural approximation to this is $\tilde q_h = \vec\beta u_h$, where u_h is the computed solution. However $\tilde q_h$ is not conservative. It is possible to easily adapt an idea of [8] to generate a conservative flux by a simple postprocessing scheme.

Our postprocessed flux is denoted by $\vec q_h$. Its restriction to each element K lies in the Raviart-Thomas space $P_{p+1}(K) + \vec x P_{p+1}(K)$ and is defined by

$$(\vec q_h, \vec r)_K = -(\vec\beta\, u_h, \vec r)_K, \qquad \forall \vec r \in P_p(K), \tag{44a}$$

$$\langle \vec q_h \cdot \vec n, \mu\rangle_F = \langle \vec\varphi_h, \mu\vec n\rangle_F, \qquad \forall \mu \in P_{p+1}(F), \tag{44b}$$

where (44b) is imposed on all edges F of K,

$$\vec{\varphi}_h = \begin{cases} \phi_h \vec{n}, & \text{on } \partial_{\text{out}} K, \\ -\phi_h \vec{n}, & \text{on } \partial_{\text{in}} K \setminus \partial_{\text{in}} \Omega, \\ \vec{\beta}\, g, & \text{on } \partial_{\text{in}} K \cap \partial_{\text{in}} \Omega, \\ 0, & \text{on faces where } \vec{\beta} \cdot \vec{n} = 0, \end{cases}$$

and, as before, $\vec{n}$ denotes the unit outward normal of K. Since (44) uses the well known degrees of freedom of the Raviart-Thomas space, it is easy to see that (44) uniquely defines a $\vec{q}_h$ in $H(\text{div}, \Omega)$.

Theorem 3.4. *Write* $u_\beta \overset{\text{def}}{=} \vec{\beta} \cdot \vec{\nabla}\, u = \nabla \cdot \vec{q}$ *and* $u_{\beta,h} \overset{\text{def}}{=} \nabla \cdot \vec{q}_h$. *Then,* $\vec{q}_h$ *is conservative, and*

$$u_{\beta,h} - \mathit{\Pi}_{p+1} u_\beta = 0 \tag{45}$$

where $\mathit{\Pi}_{p+1}$ *is the* $L^2(K)$*-orthogonal projection on* $P_{p+1}(K)$.

Proof. First of all, observe that $P_{p+1}(K) \subseteq V_p(K)$. Indeed, if z_{p+1} is any polynomial in $P_{p+1}(K)$, then finding a v in $V_p(K)$ as in Lemma 2.1 such that

$$\vec{\beta} \cdot \vec{\nabla}\, v = \vec{\beta} \cdot \vec{\nabla}\, z_{p+1}, \tag{46a}$$

$$v|_{\partial_{\text{out}} K} = z_{p+1}|_{\partial_{\text{out}} K}, \tag{46b}$$

we find from (46a) that $z_{p+1} - v$ must be a function that is constant along the $\vec{\beta}$-direction. But the same function vanishes on $\partial_{\text{out}} K$ by (46b), so it must vanish everywhere. Hence $z_{p+1} \equiv v \in V_p(K)$.

As a consequence, we have, for all z in $P_{p+1}(K)$,

$$\begin{aligned} &(\vec{q}_h, \vec{\nabla}\, z)_K + \langle \vec{q}_h \cdot \vec{n}, z \rangle_{\partial K} \\ &= -(u_h, \vec{\beta} \cdot \vec{\nabla}\, z)_K + \langle \phi_h, z \rangle_{\partial_{\text{out}} K} - \langle \phi_h, z \rangle_{\partial_{\text{in}} K} \\ &\quad + \langle \vec{\beta} \cdot \vec{n}\, g, z \rangle_{\partial_{\text{in}} K \cap \partial_{\text{in}} \Omega} && \text{by (44),} \\ &= (f, z)_K = (\vec{\beta} \cdot \vec{\nabla}\, u, z)_K && \text{by (25).} \end{aligned}$$

With $\vec{q} = \vec{\beta} u$, we can rewrite this as

$$(\nabla \cdot \vec{q}_h, z)_K = (\nabla \cdot \vec{q}, z), \quad \text{for all } z \in P_{p+1}(K), \tag{47}$$

which proves (45).

To show that $\vec{q}_h$ is conservative, let D be any subdomain formed by a union of some or all of the mesh elements. Then (47) applied with $z = 1$ implies

$$\int_{\partial D} \vec{q}_h \cdot \vec{n}\, ds = \int_{\partial D} \vec{q} \cdot \vec{n}\, ds,$$

where we have also used the $H(\text{div}, \Omega)$-conformity of $\vec{q}_h$. This shows that the net outward fluxes of $\vec{q}_h$ and $\vec{q}$ coincide on any such subdomain D. □

4. Numerical studies

We present a few small scale numerical studies to illustrate and confirm the theory. For more realistic and larger size simulations, we will need full *hp* adaptivity, which will be presented elsewhere.

4.1. **Implementation aspects.** In comparison with the UDG method, we do have more equations to solve. However, the reason our method remains competitive is that we are able to solve for the fluxes (ϕ_h) first and locally recover u_h afterwards. This results in a dimensional reduction that is quite attractive for high p. Indeed, with respect to p, the number of degrees of freedom for ϕ_h is $O(p)$, while the corresponding number for u_h is $O(p^2)$. This dimensional reduction is quite analogous to hybridized methods [9], where the so-called Lagrange multipliers are solved for first, and the remaining variables are recovered locally, element by element.

To state the matrix form of the method, we need bases for W_h, M_h, and V_h. For M_h, we set a basis consisting of functions supported only on one edge, while the bases for the other spaces consist of functions supported on only one triangle. Let Φ and U denote the vector of coefficients of ϕ_h and u_h in their basis expansion, respectively. To show how one solves for fluxes first, observe that the global finite element space V_h can be split into the outflow extensions $\mathcal{E}_{\text{out}}(M_h)$ plus a linearly independent remainder R_h. We enumerate the global basis for V_h such that the basis functions in $\mathcal{E}_{\text{out}}(M_h)$ come first. With v_h set to such test functions, (25) reduces to an equation involving only ϕ_h:

$$\sum_{K\in\mathscr{T}_h}\Big[\langle\phi_h, v_h\rangle_{\partial_{\text{out}}K} - \langle\phi_h, v_h\rangle_{\partial_{\text{in}}K\setminus\partial_{\text{in}}\Omega}\Big] = (f, v_h)_\Omega - \langle\vec\beta\cdot\vec n\, g, v_h\rangle_{\partial_{\text{in}}\Omega} \tag{48}$$

for all $v_h \in \mathcal{E}_{\text{out}}(M_h)$. As a consequence, the stiffness matrix of (25) is block triangular:

$$\begin{pmatrix} A & 0 \\ B & C \end{pmatrix}\begin{pmatrix} \Phi \\ U \end{pmatrix} = F \tag{49}$$

where A, B and C are sparse matrix blocks, and F is the corresponding load vector. This clearly shows that we can solve for Φ first and the result can then be used to compute u_h (or U) locally, element by element.

Additionally, note that in the decomposition $V_h = \mathcal{E}_{\text{out}}(M_h) \oplus R_h$, all the functions in V_h having discontinuities inside elements lie in the $\mathcal{E}_{\text{out}}(M_h)$-component. Therefore, when making the stiffness matrix of the bilinear form $a_h(\,(u_h,\phi_h), v_h)$, there is *no need to integrate basis functions with discontinuities* within the element. Indeed, as seen in (48), the term involving integration within elements, namely $(u_h, \vec\beta\cdot\vec\nabla\, v_h)$, vanishes for v_h in $\mathcal{E}_{\text{out}}(M_h)$, so there is no need to compute this integral.

Finally, suppose we enumerate the basis for M_h in such a way that functions on $\Gamma_{\ell-1}$ appear before those on Γ_ℓ (see the definition of the layers in §3.4). Then the matrix A in (49) is a square triangular matrix. In this case, Φ can be found fast using backsubstitution. Since this is the only globally coupled system to solve, the remaining (local) computations for u_h can be locally optimized (or parallelized).

4.2. **A one-dimensional example.** We begin with a simple one-dimensional example to show the enhanced accuracy and stability of the DPG method compared to the DG method. The model problem on $\Omega = (0,1)$ that we shall consider is

$$u' = f, \qquad u(0) = u_0,$$

with data $u_0 \in \mathbb{R}$ and $f \in L^2(0,1)$ set so that the exact solution is

$$u(x) = \arctan(\alpha(x - x_0))$$

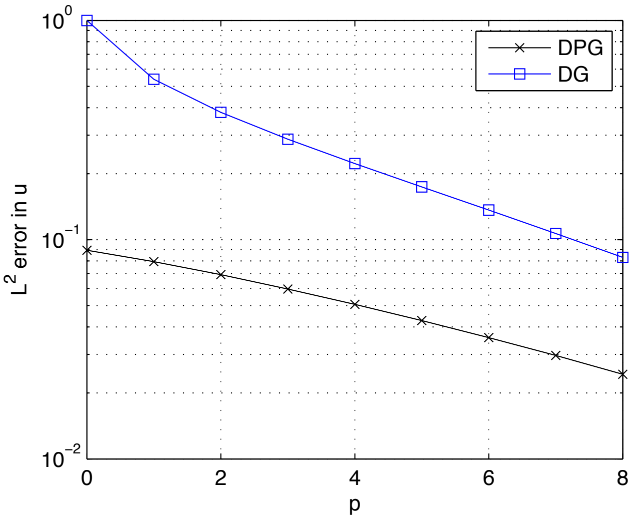


FIGURE 4. Convergence for the one-dimensional example

where $x_0 = 1$ and $\alpha = 100$. We use the analogue of our spectral DPG method (of Section 2) in one dimension, i.e., $K = (0, 1)$ and the test space is $V_p(K) = P_{p+1}(K)$.

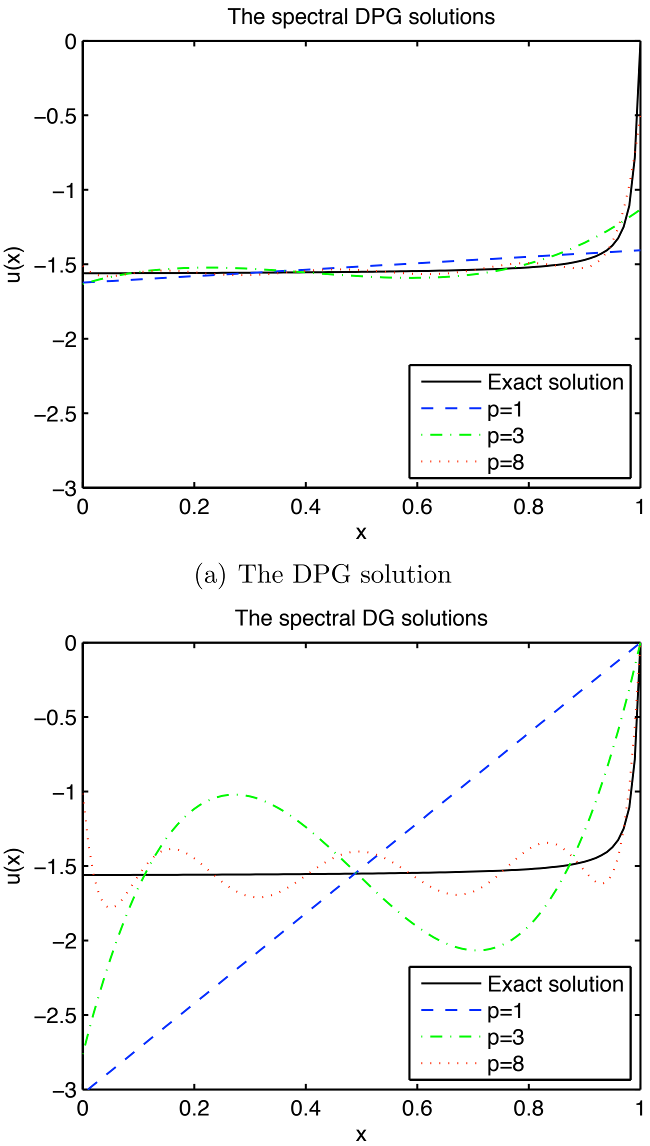
Figure 4 shows the error in u vs. p in a semilog plot. While both methods exhibit exponential convergence, the error for the DPG method is one order less than that for the DG method. The gap between the two convergence curves grows with increasing coefficient α that controls the "steepness" of the solution.

The different behavior of the two methods is further illustrated in Figures 5(a) and 5(b) that show the graphs of the approximate solutions for some values of p. Both methods deliver the exact value of the *flux* at $x = 1$. Contrary to the DG method, however, the DPG solution u within the element is "disconnected" from value at 1 and delivers the best approximation error in L^2-norm. This manifests itself with a less oscillatory behavior of the solution. Our numerics also confirmed Theorem 2.1, i.e., when we compared the DPG solution with the L^2-projection of the exact solution, we found the difference to be zero (of the order of round-off errors).

4.3. ***h* and *p* convergence to a solution with discontinuity.** Following [12], we consider a square domain $\Omega = (-1, 1)^2$ wherein an advection problem is set up so that $\vec{\beta} = (1, 9/10)$ and the exact solution is

$$u(x,y) = \begin{cases} \sin(\pi(x+1)^2/4\sin(\pi(y-9x/10)/2) \\ \qquad\qquad \text{for } -1 \le x \le 1,\ 9x/10x < y < 1, \\ e^{-5(x^2+(y-9x/10)^2)} \\ \qquad\qquad \text{for } -1 \le x \le 1,\ -1 \le y < 9x/10. \end{cases}$$

This u has a discontinuity along the line $y = 9x/10$. As in [12], we first mesh Ω by an $n \times n$ quadrilateral mesh aligned with the line of discontinuity (see [12, Fig. 10] for details). Since our method is based on triangular meshes ([12] only considered quadrilateral meshes), we further split each of the quadrilaterals into two triangles by connecting their diagonals of positive slope. Another difference with the experiment in [12] is that while [12] needed to include a stabilizing lower order reaction term for theoretical reasons, we do not need to,



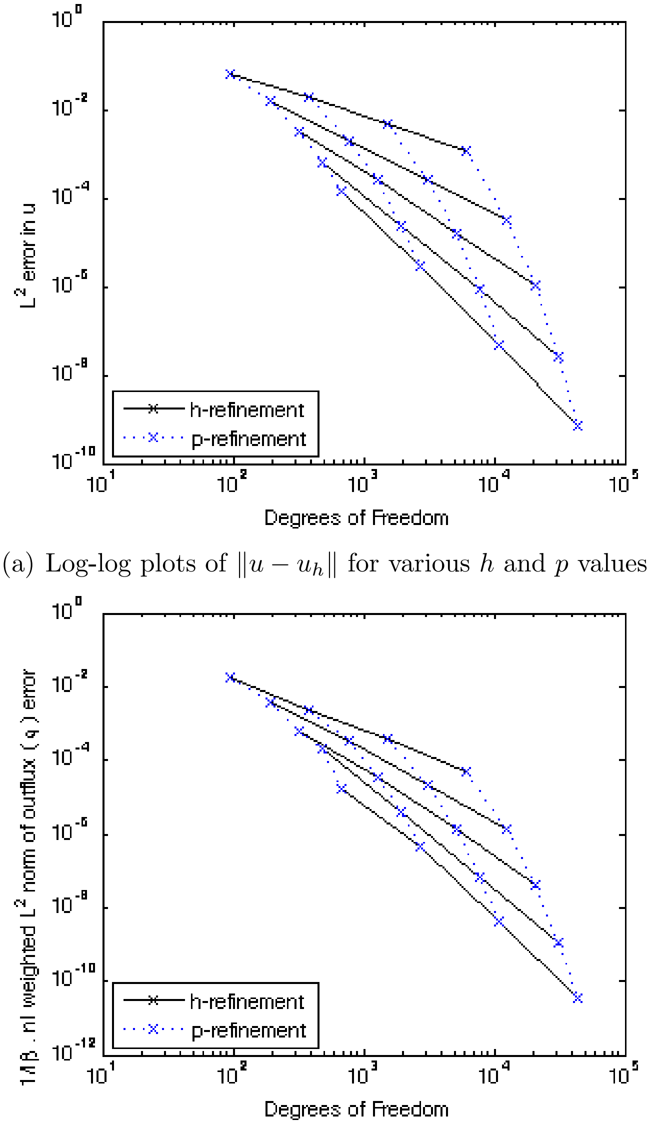
(a) The DPG solution

(b) The DG solution

FIGURE 5. The enhanced stability of DPG approximations, compared to the DG ones, in the case of a model transport problem with a layer at the right end

hence we solve the purely advective problem. We report the errors in u_h and ϕ_h for the cases $n = 5, 9, 17, 33$ and $p = 1, \ldots, 5$ in Figure 6.

It is clear from Figure 6 that the DPG method converged under h and p refinements as if the solution were infinitely regular (see Figure 6). We observed a similar behavior for the DG method as well. Such observations demonstrate the advantage of using discontinuous finite elements with meshes aligned with the solution discontinuity. For the standard DG method, this observation was made in [12], where DG is compared with the streamline diffusion method. While the streamline diffusion solution can also be improved by aligning meshes with shock lines [21], its convergence rate remains limited by the regularity of the



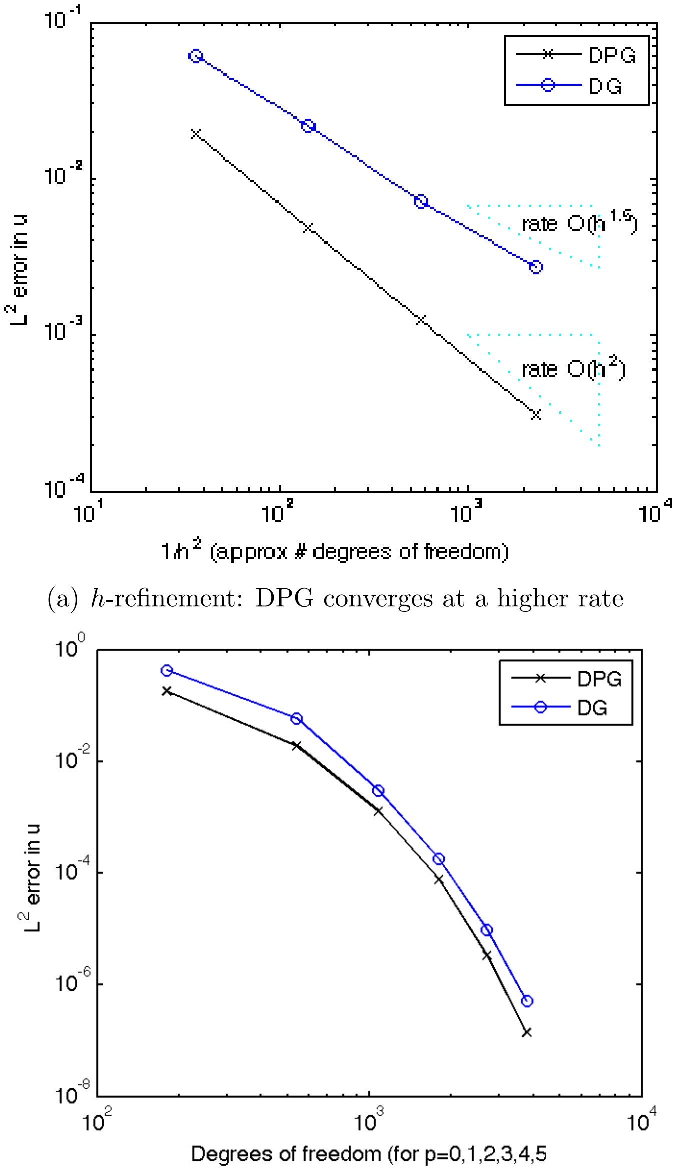
(a) Log-log plots of $\|u - u_h\|$ for various h and p values

(b) Log-log plots of flux errors $\|\phi - \phi_h\|_{\frac{1}{\beta},\mathscr{E}_h}$

FIGURE 6. The hp convergence of the DPG method (log-log plots) for the Houston-Schwab-Süli example [12]

solution, as observed in [12]. Since our purpose is a comparison of DG and DPG, we note here that the errors we observed for the DG method (not shown in the figure), as in § 4.2, remained higher than that of the DPG method. Figure 6 suggests that exponential rates of convergence can be obtained with the DPG method, even for discontinuous solutions, once an adaptive strategy to align the meshes with the shocks is implemented.

The data that was used to plot Figure 6 also shows the rate of the convergence. Figure 6(a) indicates that $\|u - u_h\|$ converges at the rate h^{p+1} (for $p = 1 \dots 5$). This is in accordance with Theorem 3.3. On the other hand, Figure 6(b) shows that the fluxes converge such that $\|\phi - \phi_h\|_{h,\frac{1}{\beta},\mathscr{T}_h}$ goes to zero at the rate of $h^{p+2.5}$, which is one order more than what we were able to prove in Theorem 3.3.

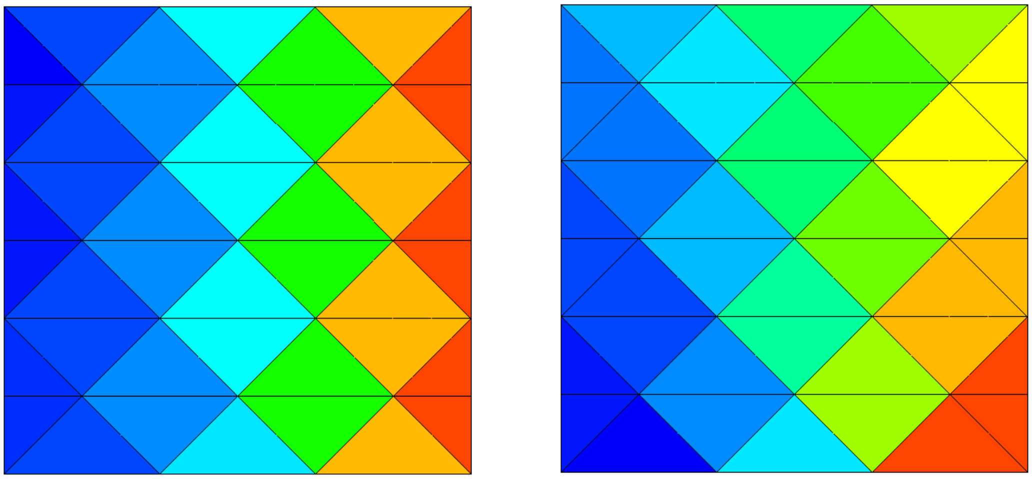


(a) h-refinement: DPG converges at a higher rate

(b) p-refinement: DPG gives lower error

Figure 7. The DPG method performs better than the DG method under both h and p refinements for the Peterson example [17]. Both plots are in a log-log scale.

4.4. **The Peterson example.** Next, we consider the well-known Peterson example [17]. Peterson's mesh of the unit square is a specific quasiuniform mesh of the type appearing in Figure 8, obtained from an $n \times n$ partition of the unit square, but with additional vertical lines through some mesh vertices such that the new lines divide the unit square into, say m, vertical strips (see [17] for details). On such meshes, we apply the UDG and DPG method to the problem $\partial u/\partial y = 0, u(x,0) = \sin 6x, x \in (0,1)$. The results in Figure 7 show that the DPG method delivers better results than the standard DG method. Figure 7(a) is obtained with $p = 1$ and four progressively finer meshes: first with $n = 6, m = 3$, second



(a) The DPG solution

(b) The DG solution

FIGURE 8. DPG solutions exhibit less crosswind diffusion than DG solutions. In this example, the flow is upward.

with $n = 12, m = 6$, third with $n = 24, m = 8$, and fourth with $n = 48$, $m = 16$. This allows us to study the case of h-refinement. On the other hand, Figure 7(b) is obtained by fixing the mesh with $n = 6$, $m = 3$, and increasing the degree p from 0 through 5. The results clearly demonstrate the well-known 0.5-order suboptimality of the DG method in the h-refinement case, but more interestingly, also show that the DPG method converges at the optimal rate.

We also present the solutions u_h obtained with both DG and DPG methods when the inflow data is $u(x, 0) = x^2$ in Fig. 8. We used piecewise constant approximations ($p = 0$). Observe that as the flow proceeds upwards, the values 'diffuse' horizontally for DG case, while there is little indication of such crosswind diffusion for the DPG method with the optimal test functions.

5. Concluding remarks

5.1. **Summary.** We proved optimal error estimates for u_h in h and p. Using the known best approximation estimates [20] for a quasiuniform mesh of meshsize h, we obtain as a corollary of Theorem 3.3 that

$$C\|u - u_h\|_\Omega \le \frac{h^{s_1}}{p^{s_1}}|u|_{H^{s_1}(\Omega)} + (\frac{1}{|\vec{\beta}|h^{1/2}})\frac{h^{s_2-1/2}}{p^{s_2-1/2}}|u|_{H^{s_2}(\Omega)} \tag{50}$$

with some C independent of h and p and any $0 \le s_1 \le p+1$ and $0 \le s_2 \le p+2$. If the solution is smooth, choosing $s_1 = p+1$ and $s_2 = p+2$, we obtain the optimal h-convergence rate of $O(h^{p+1})$. The p-convergence rate is also optimal.

Note however that there may be room for improvement in the techniques used in the current proof. In Theorem 3.3 and (50), note that the regularity on u that we need to obtain the optimal convergence rate is higher than expected. Furthermore, for the fluxes, we proved a rate of convergence that is suboptimal by one order, but numerical experiments

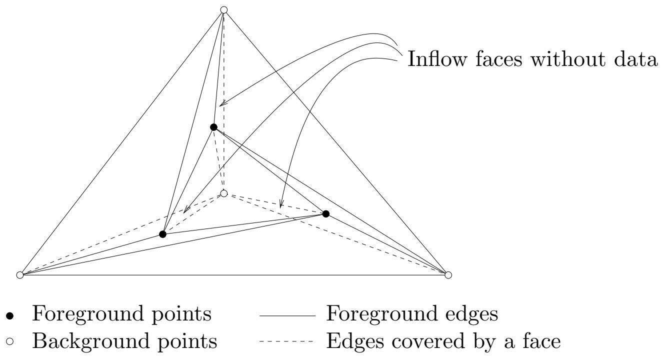


Figure 9. Top view of a 3D mesh. To obtain the 3D mesh, keep the background points on the plane of the paper and raise the foreground points one unit from the paper. The inflow direction is emanating from the paper.

indicate that the fluxes do converge at optimal order. A sharper analysis shedding light on these issues can be valuable.

In every numerical experiment we performed, without exception, the DPG method gave more accurate solutions than the standard DG method. Numerical experiments confirm the theoretically proved optimal rate of convergence for the primal variable u_h.

Of course, the full potential of the method is not indicated by an estimate like (50). Indeed, one can achieve *exponential* rate of convergence by constructing a variable-degree analogue where each element K is endowed with its own trial space of degree p_K. The fluxes must be chosen to be of degree one plus the maximum of the degrees from adjacent elements. A full study of the variable degree version, together with an *hp*-adaptive scheme which automatically selects h or p refinement in different locations, is postponed to a future work.

5.2. **The three-dimensional case.** The entire analysis of Section 2 holds in any dimension, including three. The *sole* reason we restricted ourselves to the two-dimensional case in Section 3 is that Lemma 3.1 does not hold for general tetrahedral meshes. This fact does not seem to be well known, so we indicate a counterexample in Figure 9. In the mesh shown therein, there is no tetrahedron whose inflow boundary is wholly contained in the global inflow boundary of the domain.

In situations like in Figure 9, we may choose to appropriately bisect the 'problematic' tetrahedra so that Lemma 3.1 does hold at each stage. Alternately, while generating tetrahedral meshes by advancing front meshing algorithms, we can ensure that elements at each front are such that Lemma 3.1 is satisfied. Notice that Lemma 3.1 is required not only for the analysis of our method, but also for practical implementation. Indeed, the process of splitting the mesh into layers $\mathscr{S}_\ell$ (see § 3.4) yields an enumeration of flux degrees of freedom that makes the matrix for ϕ_h's triangular, and hence suitable for fast solution by backsubstitution. Therefore any extra effort to obtain meshes of the required type is likely to pay off in the final analysis. (Note that Lemma 3.1 is needed to obtain triangular matrices even when using the standard UDG method.) These and related algorithmic issues will be studied in the future.

5.3. **Further studies.** We outline a few directions in which we are pursuing further studies. The singularly perturbed convection-diffusion problem is of great interest. It is easy to formulate a generalization of our method by combining with the mixed method. This is because the space for u_h in our method and in the mixed method is the same (one can proceed, for instance, along the same lines as in [10]). However, proving hp estimates with p independent constants for such a method does not appear to be an easy task.

The case of variable transport direction $\vec{\beta}$ can be handled by projecting $\vec{\beta}$ into an appropriate finite element space. In particular, a low order method can be immediately written down if we approximate $\vec{\beta}$ by its lowest-order Raviart-Thomas interpolant $\Pi_h^{RT}\vec{\beta}$. Then, whenever $\nabla \cdot \vec{\beta} = 0$, the approximant $\Pi_h^{RT}\vec{\beta}$ is a piecewise *constant* vector field with continuous interelement normal components. Hence, by simply modifying our test space $V_p(K)$ to align with $\Pi_h^{RT}\vec{\beta}$ within each element K, we obtain a low-order generalization of the DPG method to the case of variable $\vec{\beta}$ with minimal modifications. To obtain truly high-order methods for variable $\vec{\beta}$, our test space must take into account the variable advection.

This and other generalizations become clearer, once we view the method presented here as a particular case of a more general paradigm for constructing schemes, not only for the transport problem or its singular perturbation, but indeed for a much larger class of problems. The new paradigm is to find a space of *optimal test functions* tailored for any particular problem. It turns out that the framework of DG methods offers a unique opportunity to *locally* compute test functions that approach optimality in the sense of reproducing the exact stability properties of the boundary value problem at the discrete level. This approach and its connection to the space of test functions presented here will be clarified in a sequel.

Appendix A. Appendix: Proof of Theorem 2.2

This section is devoted to a proof of Theorem 2.2. Before we embark on this, we need to develop several intermediate results. Recall that $\mathcal{E}_{\text{out}}$ denotes the extension from the outflow boundary. Now, let us also define the *extension from inflow boundary*, denoted by $\mathcal{E}_{\text{in}}$ as follows. Given a function μ on $\partial_{\text{in}}K$, the function $\mathcal{E}_{\text{in}}\,\mu$ on K coincides with μ on $\partial_{\text{in}}K$ and is independent of η_1 (i.e., it is constant along the $\vec{\beta}$-direction).

Lemma A.1. *Suppose ϕ in $L^2(\partial_{\text{out}}K)$ and μ in $L^2(\partial_{\text{in}}K)$. Then*

$$\langle \vec{\beta}\cdot\vec{n}\ \mathcal{E}_{\text{out}}\,\phi,\ \mathcal{E}_{\text{in}}\,\mu\rangle_{\partial K} = 0. \tag{51}$$

Moreover, using the norm defined in (20)*, we have*

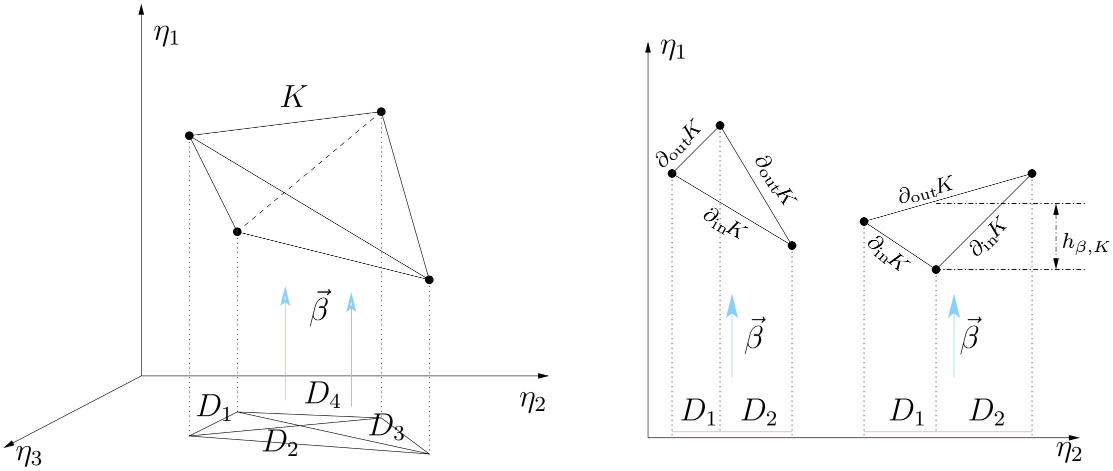
$$\|\mathcal{E}_{\text{in}}\,\phi\|_{\beta,\partial_{\text{out}}K} = \|\phi\|_{\beta,\partial_{\text{in}}K}\,. \tag{52}$$

Proof. By integration by parts, we have

$$\begin{aligned}\langle \vec{\beta}\cdot\vec{n}\,\mathcal{E}_{\text{out}}\,\phi, \mathcal{E}_{\text{in}}\,\mu\rangle_{\partial K} &= (\vec{\beta}\cdot\vec{\nabla}\,\mathcal{E}_{\text{out}}\,\phi, \mathcal{E}_{\text{in}}\,\mu)_K \\ &\quad + (\mathcal{E}_{\text{out}}\,\phi, \nabla\cdot(\vec{\beta}\,\mathcal{E}_{\text{in}}\,\mu))_K,\end{aligned}$$

and both terms on the right are zero. This proves (51).

Now let us prove (52). Recall the coordinate system aligned with the flow, defined in (10). Project K onto the $\eta_1 = 0$ hyperplane along the $\vec{\beta}$-direction to obtain a $N-1$ dimensional domain D, meshed with simplices D_j (see Fig. 10) in one less dimension. The flow field $\vec{\beta}$



(a) A tetrahedron projected to 4 triangles

(b) The case of triangles

FIGURE 10. The flow isomorphism

generates an isomorphism from D onto $\partial_{\rm in}K$ as well as $\partial_{\rm out}K$. We denote the former by $\mathcal{F}_{\rm in}$ and the latter by $\mathcal{F}_{\rm out}$.

The values of $\mathcal{E}_{\rm in}\,\mu$ on $\mathcal{F}_{\rm out}(D_j)$ are mapped over from $\mathcal{F}_{\rm in}(D_j)$. Since $\mathcal{F}_{\rm out}(D_j)$ is mapped to $\mathcal{F}_{\rm in}(D_j)$ by the map $\mathcal{F}_{\rm in}\circ\mathcal{F}_{\rm out}^{-1}$, which is a transformation whose Jacobian is constant on $\mathcal{F}_{\rm out}(D_j)$, integration via a change of variable shows that

$$\|\,\mathcal{E}_{\rm in}\,\mu\|^2_{\mathcal{F}_{\rm out}(D_j)} = \frac{|\,\mathcal{F}_{\rm out}(D_j)|}{|\,\mathcal{F}_{\rm in}(D_j)|}\,\|\mu\|^2_{\mathcal{F}_{\rm in}(D_j)}. \tag{53}$$

Now, consider the solid S_j formed by the faces $\mathcal{F}_{\rm out}(D_j)$, $\mathcal{F}_{\rm in}(D_j)$, filled by line segments parallel to $\vec\beta$. Since $\vec\beta\cdot\vec n = 0$ on all faces of this solid other than $\mathcal{F}_{\rm out}(D_j)$ and $\mathcal{F}_{\rm in}(D_j)$, by the divergence theorem, we find that

$$\begin{aligned}0 = \int_{S_j}\nabla\cdot\vec\beta\,dx = {} & |\vec\beta\cdot\vec n_{\mathcal{F}_{\rm out}(D_j)}|\,|\,\mathcal{F}_{\rm out}(D_j)| \\ & + |\vec\beta\cdot\vec n_{\mathcal{F}_{\rm in}(D_j)}|\,|\,\mathcal{F}_{\rm in}(D_j)|.\end{aligned}$$

Using this in (53), we find that

$$|\vec\beta\cdot\vec n_{\mathcal{F}_{\rm out}(D_j)}|\,\|\,\mathcal{E}_{\rm in}\,\mu\|^2_{\mathcal{F}_{\rm out}(D_j)} = |\vec\beta\cdot\vec n_{\mathcal{F}_{\rm in}(D_j)}|\,\|\mu\|^2_{\mathcal{F}_{\rm in}(D_j)}.$$

Summing over all j, we obtain (52). □

Now, let us define the exact local solution operators, in analogy with the discrete ones in (17). Namely, we define

$$\mathcal{Q} : L^2(\partial_{\rm in}K)\longmapsto L^2(\partial_{\rm out}K), \qquad \mathcal{U} : L^2(\partial_{\rm in}K)\longmapsto L^2(K),$$

as follows: Given ϕ in $L^2(\partial_{\rm in}K)$, the functions $\mathcal{Q}\phi$ and $\mathcal{U}\phi$ are defined by

$$-(\mathcal{U}\phi, \vec\beta\cdot\vec\nabla\,v)_K + \langle\mathcal{Q}\phi, v\rangle_{\partial_{\rm out}K} = \langle\phi, v\rangle_{\partial_{\rm in}K}, \tag{54}$$

for all v in

$$V(K) \overset{\rm def}{=} \{\nu\in L^2(K) : \beta\cdot\vec\nabla\,\nu\in L^2(K)\}. \tag{55}$$

The next result concerns the operators $\mathcal{Q}$ and $\mathcal{Q}_p$. Recall that $\mathit{\Pi}_M$ denotes the L^2-orthogonal projection onto M_h (it obviously acts face by face).

Lemma A.2. *Let $\phi \in L^2(\partial_{\text{out}}K)$. Using the norm defined in* (19)*, we have*

$$\|\mathcal{Q}\phi\|_{\frac{1}{\beta},\partial_{\text{out}}K} = \|\phi\|_{\frac{1}{\beta},\partial_{\text{in}}K} \tag{56}$$

Furthermore, the operator $\mathcal{Q}_p$ of (17) *is related to $\mathcal{Q}$ by*

$$\mathcal{Q}_p = \mathit{\Pi}_M \mathcal{Q}. \tag{57}$$

Proof. By (54), we know that

$$\langle \mathcal{Q}\phi, \mu\rangle_{\partial_{\text{out}}K} = \langle \phi, \mathcal{E}_{\text{out}}\,\mu\rangle_{\partial_{\text{in}}K} \qquad \forall \mu \in L^2(\partial_{\text{out}}K),$$

while by (51) of Lemma A.1, we know that

$$\begin{aligned}\langle \phi, \mathcal{E}_{\text{out}}\,\mu\rangle_{\partial_{\text{in}}K} &= \langle \vec{\beta}\cdot\vec{n}_{\text{in}}\frac{\phi}{\vec{\beta}\cdot\vec{n}_{\text{in}}}, \mathcal{E}_{\text{out}}\,\mu\rangle_{\partial_{\text{in}}K} \\ &= -\langle \vec{\beta}\cdot\vec{n}_{\text{out}}\,\mathcal{E}_{\text{in}}(\frac{\phi}{\vec{\beta}\cdot\vec{n}_{\text{in}}}), \mu\rangle_{\partial_{\text{out}}K},\end{aligned}$$

where $\vec{n}_{\text{in}}$ and $\vec{n}_{\text{out}}$ denote the unit outward normals on $\partial_{\text{in}}K$ and $\partial_{\text{out}}K$, respectively. Combining, we conclude that

$$\langle \mathcal{Q}\phi, \mu\rangle_{\partial_{\text{out}}K} = -\langle \vec{\beta}\cdot\vec{n}_{\text{out}}\,\mathcal{E}_{\text{in}}(\frac{\phi}{\vec{\beta}\cdot\vec{n}_{\text{in}}}), \mu\rangle_{\partial_{\text{out}}K}$$

for all $\mu \in L^2(\partial_{\text{out}}K)$. Thus, we find that

$$\mathcal{Q}\phi = -\vec{\beta}\cdot\vec{n}_{\text{out}}\,\mathcal{E}_{\text{in}}(\frac{\phi}{\vec{\beta}\cdot\vec{n}_{\text{in}}}).$$

In view of this, we can use (52) of Lemma A.1 with $\phi/\vec{\beta}\cdot\vec{n}_{\text{in}}$ in place of ϕ, to get

$$\begin{aligned}&\sum_{F_{\text{out}}\subseteq\partial_{\text{out}}K} |\vec{\beta}\cdot\vec{n}_{F_{\text{out}}}| \left\| \mathcal{E}_{\text{in}}(\frac{\phi}{\vec{\beta}\cdot\vec{n}_{\text{in}}})\right\|^2_{F_{\text{out}}} \\ &\qquad = \sum_{F_{\text{in}}\subseteq\partial_{\text{in}}K} |\vec{\beta}\cdot\vec{n}_{F_{\text{in}}}| \left\| \frac{\phi}{\vec{\beta}\cdot\vec{n}_{\text{in}}}\right\|^2_{F_{\text{in}}},\end{aligned}$$

or in other words,

$$\sum_{F_{\text{out}}\subseteq\partial_{\text{out}}K} \frac{1}{|\vec{\beta}\cdot\vec{n}_{F_{\text{out}}}|}\,\|\mathcal{Q}\phi\|^2_{F_{\text{out}}} = \sum_{F_{\text{in}}\subseteq\partial_{\text{in}}K} \frac{1}{|\vec{\beta}\cdot\vec{n}_{F_{\text{in}}}|}\,\|\phi\|^2_{F_{\text{in}}}.$$

Thus we have proved (56).

To prove (57), it suffices to observe that on each element, $\mathcal{Q}_p\phi$ is the discrete spectral approximation to the exact solution $\mathcal{Q}\phi$, hence by the superconvergence result of Theorem 2.1, we find that $\mathcal{Q}_p\phi - \mathit{\Pi}_M\mathcal{Q}\phi = 0$. □

We need two more lemmas before we proceed to obtain bounds for the local solution operators.

Lemma A.3. *Suppose w is in $P_p(K)$ and v is the function in $V_p(K)$ obtained via Lemma 2.1 satisfying*

$$v|_{\partial_{\rm out}K} = 0, \qquad \vec\beta \cdot \vec\nabla\, v = w. \tag{58}$$

Then

$$\|v\|^2_{\beta,\partial_{\rm in}K} \le \frac{1}{|\vec\beta|} h_{\beta,K} \|w\|^2_K, \tag{59}$$

$$\|v\|^2_K \;\le \frac{1}{|\vec\beta|^2} h^2_{\beta,K} \|w\|^2_K. \tag{60}$$

Proof. We use the notations in the proof of Lemma A.1. Let $\hat v$ denote the function obtained by mapping the values of $v|_{\partial_{\rm in}K}$ to D, i.e.,

$$\hat v(\eta_2, \ldots, \eta_N) = v(\mathcal{F}_{\rm in}(0, \eta_2, \ldots, \eta_N)).$$

Then, since $|\,\mathcal{F}_{\rm in}(D_j)|\,|\vec\beta \cdot \vec n_{\mathcal{F}_{\rm in}(D_j)}| = |D_j|\,|\vec\beta|$, we obtain by a change of variable that

$$\|v\|^2_{\mathcal{F}_{\rm in}(D_j)} = \frac{|\,\mathcal{F}_{\rm in}(D_j)|}{|D_j|} \|\hat v\|^2_{D_j} = \frac{|\vec\beta|}{|\vec\beta \cdot \vec n_{\mathcal{F}_{\rm in}(D_j)}|} \|\hat v\|^2_{D_j}. \tag{61}$$

Now, let $f_{\rm out}(\eta_2, \ldots, \eta_N)$ and $f_{\rm in}(\eta_2, \ldots, \eta_N)$ denote the η_1-components of $\mathcal{F}_{\rm out}(0, \eta_2, \ldots, \eta_N)$ and $\mathcal{F}_{\rm in}(0, \eta_2, \ldots, \eta_N)$, respectively. Then the inflow and outflow boundaries are the graphs of $f_{\rm in}$ and $f_{\rm out}$, respectively. By (58),

$$v(\eta_1, \ldots, \eta_N) = \int_{f_{\rm out}(\eta_2,\ldots,\eta_N)}^{\eta_1} \frac{1}{|\vec\beta|}\, w(s, \eta_2, \ldots, \eta_N)\, ds \tag{62}$$

When we evaluate this expression on the inflow boundary $\eta_1 = f_{\rm in}(\eta_2, \ldots, \eta_N)$, we obtain values of $\hat v$. Consequently, (61) implies that

$$\begin{aligned}
&\|v\|^2_{\beta,\partial_{\rm in}K} = \sum_j |\vec\beta| \int_{D_j} |\hat v|^2 \, d\eta_2 \ldots d\eta_N \\
&= |\vec\beta| \int_D \left| \int_{f_{\rm out}(\eta_2,\ldots,\eta_N)}^{f_{\rm in}(\eta_2,\ldots,\eta_N)} \frac{1}{|\vec\beta|}\, w(s, \eta_2, \ldots, \eta_N)\, ds \right|^2 d\eta_2 \ldots d\eta_N \\
&\le \frac{1}{|\vec\beta|} h_{\beta,K} \int_K |w|^2 \, d\eta_1 \ldots d\eta_N,
\end{aligned}$$

and the first estimate of the lemma follows.

For the second estimate, we again use (62).

$$\begin{aligned}\|v\|_K^2 = &\sum_j \int_{f_{\mathrm{out}}(\eta_2,\ldots,\eta_N)}^{f_{\mathrm{in}}(\eta_2,\ldots,\eta_N)} \int_{D_j} \\ &\qquad \left| \int_{f_{\mathrm{out}}(\eta_2,\ldots,\eta_N)}^{\eta_1} \frac{1}{|\vec{\beta}|}\, w(s,\eta_2,\ldots,\eta_N)\; ds \right|^2 d\eta_2 \ldots d\eta_N\; d\eta_1 \\ \le &\int_{f_{\mathrm{out}}(\eta_2,\ldots,\eta_N)}^{f_{\mathrm{in}}(\eta_2,\ldots,\eta_N)} \int_D \frac{h_{\beta,K}}{|\vec{\beta}|^2} \int_{f_{\mathrm{out}}(\eta_2,\ldots,\eta_N)}^{f_{\mathrm{in}}(\eta_2,\ldots,\eta_N)} \\ &\qquad |w(s,\eta_2,\ldots,\eta_N)|^2\; ds\, d\eta_2 \ldots d\eta_N\; d\eta_1 \\ \le &\frac{h_{\beta,K}^2}{|\vec{\beta}|^2} \int_K |w(s,\eta_2,\ldots,\eta_N)|^2\; ds\, d\eta_2 \ldots d\eta_N.\end{aligned}$$

This finishes the proof. □

Lemma A.4. *Suppose ψ is in $M_{p+1}(\partial_{\mathrm{out}}K)$ and v is the function in $V_p(K)$ satisfying*

$$v|_{\partial_{\mathrm{out}}K} = \psi, \qquad \vec{\beta}\cdot\vec{\nabla}\, v = 0.$$

Then

$$\|v\|_K^2 \le \frac{h_{\beta,K}}{|\vec{\beta}|}\, \|\psi\|^2_{\beta,\partial_{\mathrm{out}}K}\,.$$

Proof. Map ψ from $\partial_{\mathrm{out}}K$ to D using the flow isomorphism to obtain a function $\hat{\psi}$ on D. Then, evaluating the norm $\|v\|_K^2$ as an iterated integral,

$$\begin{aligned}\|v\|_K^2 = &\sum_j \int_{f_{\mathrm{out}}(\eta_2,\ldots,\eta_N)}^{f_{\mathrm{in}}(\eta_2,\ldots,\eta_N)} \int_{D_j} \\ &\qquad |\hat{\psi}(\eta_2,\ldots,\eta_N)|^2\, d\eta_2 \ldots d\eta_N\; d\eta_1 \quad \text{(by Fubini)} \\ = &\sum_j \int_{f_{\mathrm{out}}(\eta_2,\ldots,\eta_N)}^{f_{\mathrm{in}}(\eta_2,\ldots,\eta_N)} \frac{|\vec{\beta}\cdot\vec{n}_{\mathcal{F}_{\mathrm{out}}(D_j)}|}{|\vec{\beta}|} \|\psi\|^2_{\mathcal{F}_{\mathrm{out}}(D_j)}\; d\eta_1 \text{ (cf. (61))} \\ \le &\frac{h_{\beta,K}}{|\vec{\beta}|}\, \|\psi\|^2_{\beta,\partial_{\mathrm{out}}K}\,,\end{aligned}$$

as the maximum value of $|f_{\mathrm{in}} - f_{\mathrm{out}}|$ is $h_{\beta,K}$. □

Proof of Theorem 2.2. To prove the bound on $\mathcal{Q}_p\phi$ in (21), we use Lemma A.2 and the fact that orthogonal projectors have unit norm, so that

$$\begin{aligned}\|\mathcal{Q}_p\phi\|^2_{\frac{1}{\beta},\partial_{\mathrm{out}}K} &= \sum_{F_{\mathrm{out}}\subseteq\partial_{\mathrm{out}}K} \frac{1}{|\vec{\beta}\cdot\vec{n}_{F_{\mathrm{out}}}|}\, \|\mathcal{Q}_p\phi\|^2_{F_{\mathrm{out}}} \\ &= \sum_{F_{\mathrm{out}}\subseteq\partial_{\mathrm{out}}K} \frac{1}{|\vec{\beta}\cdot\vec{n}_{F_{\mathrm{out}}}|}\, \|\mathit{\Pi}_M\mathcal{Q}\phi\|^2_{F_{\mathrm{out}}} \\ &\le \sum_{F_{\mathrm{out}}\subseteq\partial_{\mathrm{out}}K} \frac{1}{|\vec{\beta}\cdot\vec{n}_{F_{\mathrm{out}}}|}\, \|\mathcal{Q}\phi\|^2_{F_{\mathrm{out}}} = \|\mathcal{Q}\phi\|^2_{\frac{1}{\beta},\partial_{\mathrm{out}}K},\end{aligned}$$

and complete using (56).

To prove the bound on $\mathcal{U}_p\phi$, let v in $V_p(K)$ be a function as in Lemma A.3 satisfying

$$v|_{\partial_{\text{out}}K} = 0, \quad \vec{\beta} \cdot \vec{\nabla}\, v = \mathcal{U}_p\phi.$$

Substituting this v in the definition (17) of $\mathcal{U}_p$, we obtain

$$\begin{aligned} \|\mathcal{U}_p\phi\|_K^2 = \langle \phi, v\rangle_{\partial_{\text{in}}K} &\le \|\phi\|_{\frac{1}{\beta},\partial_{\text{in}}K}\, \|v\|_{\beta,\partial_{\text{in}}K} \\ &\le \|\phi\|_{\frac{1}{\beta},\partial_{\text{in}}K} \frac{h_{\beta,K}^{1/2}}{|\vec{\beta}|^{1/2}} \|\mathcal{U}_p\phi\|_K \end{aligned}$$

by Lemma A.3, so the proof of (21) is finished.

To prove (22), first consider $\mathcal{Q}_p^f z$. In its definition, namely in (18), we can set v to be the function given by Lemma A.4 with $\psi = \mathcal{Q}_p^f z / |\vec{\beta} \cdot \vec{n}|$. Then

$$\begin{aligned} \|\mathcal{Q}_p^f z\|^2_{\frac{1}{\beta},\partial_{\text{out}}K} = (z,v) &\le \|z\|_K \frac{h_{\beta,K}^{1/2}}{|\vec{\beta}|^{1/2}} \left\| \frac{1}{|\vec{\beta}\cdot\vec{n}|} \mathcal{Q}_p^f z \right\|_{\beta,\partial_{\text{out}}K} \\ &= \|z\|_K \frac{h_{\beta,K}^{1/2}}{|\vec{\beta}|^{1/2}} \|\mathcal{Q}_p^f z\|_{\frac{1}{\beta},\partial_{\text{out}}K}. \end{aligned}$$

Next, consider $\mathcal{U}_p^f z$. In this case, we set v to be the function in $V_p(K)$ with $\vec{\beta}\cdot\vec{\nabla}\, v = -\mathcal{U}_p^f z$ and $v|_{\partial_{\text{out}}K=0}$, given by Lemma A.3. Substituting this v into (18), and applying (60) of Lemma A.3, we have

$$\|\mathcal{U}_p^f z\|_K^2 = (z,v) \le \|z\|_K \|v\|_K \le \|z\|_K \frac{h_{\beta,K}}{|\vec{\beta}|} \|\mathcal{U}_p^f z\|_K.$$

This proves the last estimate of the theorem and finishes the proof. □

References

[1] I. Babuška, *Error-bounds for finite element method*, Numer. Math., 16 (1970/1971), pp. 322–333.

[2] T. Belytschko, N. Moës, S. Usui, and C. Parimi, *Arbitrary discontinuities in finite elements*, International Journal for Numerical Methods in Engineering, 50 (2001), pp. 993–1013.

[3] K. S. Bey and J. T. Oden, *hp-version discontinuous Galerkin methods for hyperbolic conservation laws*, Comput. Methods Appl. Mech. Engrg., 133 (1996), pp. 259–286.

[4] C. L. Bottasso, S. Micheletti, and R. Sacco, *The discontinuous Petrov-Galerkin method for elliptic problems*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 3391–3409.

[5] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, no. 15 in Springer Series in Computational Mathematics, Springer-Verlag, New York, 1991.

[6] P. Causin and R. Sacco, *A discontinuous Petrov-Galerkin method with Lagrangian multipliers for second order elliptic problems*, SIAM J. Numer. Anal., 43 (2005), pp. 280–302 (electronic).

[7] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland Publishing Company, Amsterdam, 1978.

[8] B. Cockburn, B. Dong, and J. Guzmán, *Optimal convergence of the original DG method for the transport-reaction equation on special meshes*, SIAM J. Numer. Anal., 46 (2008), pp. 1250–1265.

[9] B. Cockburn, J. Gopalakrishnan, and R. Lazarov, *Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems*, SIAM Journal on Numerical Analysis, 47 (2009), pp. 1319–1365.

[10] H. Egger and J. Schöberl, *A mixed-hybrid-discontinuous Galerkin finite element method for convection-diffusion problems*, IMA J. Numer. Anal., Preprint (to appear) (2009).

[11] R. S. Falk, *Analysis of finite element methods for linear hyperbolic problems*, in Discontinuous Galerkin Methods: Theory, Computation, and Applications, vol. 11 of Lecture Notes in Computational Science and Engineering, New York, 2000, Springer–Verlag, pp. 103–112.

[12] P. Houston, C. Schwab, and E. Süli, *Stabilized hp-finite element methods for first-order hyperbolic problems*, SIAM J. Numer. Anal., 37 (2000), pp. 1618–1643 (electronic).

[13] T. J. R. Hughes and A. Brooks, *A multidimensional upwind scheme with no crosswind diffusion*, in Finite element methods for convection dominated flows (Papers, Winter Ann. Meeting Amer. Soc. Mech. Engrs., New York, 1979), vol. 34 of AMD, Amer. Soc. Mech. Engrs. (ASME), New York, 1979, pp. 19–35.

[14] C. Johnson and J. Pitkäranta, *An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation*, Math. Comp., 46 (1986), pp. 1–26.

[15] P. Lasaint and P.-A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical aspects of finite elements in partial differential equations, C. de Boor, ed., Academic Press, New York, 1974, pp. 89–123. Proceedings of a symposium conducted by Math. Res. Center, Univ. of Wisconsin–Madison, in April, 1974.

[16] N. Nguyen, J. Peraire, and B. Cockburn, *An implicit high-order hybridizable discontinuous galerkin method for linear convection–diffusion equations*, Journal of Computational Physics, 228 (2009), pp. 3232–3254.

[17] T. E. Peterson, *A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation*, SIAM J. Numer. Anal., 28 (1991), pp. 133–140.

[18] W. H. Reed and T. R. Hill, *Triangular mesh methods for the neutron transport equation*, Tech. Rep. LA-UR-73-479, Los Alamos Scientific Laboratory, 1973.

[19] G. R. Richter, *An optimal-order error estimate for the discontinuous Galerkin method*, Math. Comp., 50 (1988), pp. 75–88.

[20] C. Schwab, *p- and hp-finite element methods*, Numerical Mathematics and Scientific Computation, The Clarendon Press, Oxford University, New York, 1998. Theory and applications in solid and fluid mechanics.

[21] G. H. Zhou and R. Rannacher, *Mesh orientation and anisotropic refinement in the streamline diffusion method*, in Finite element methods (Jyväskylä, 1993), vol. 164 of Lecture Notes in Pure and Appl. Math., Dekker, New York, 1994, pp. 491–500.

Institute of Computational Engineering and Sciences, 1 University Station, C0200, The University of Texas at Austin, TX 78712
E-mail address: `leszek@ices.utexas.edu`

University of Florida, Department of Mathematics, Gainesville, FL 32611–8105.
E-mail address: `jayg@math.ufl.edu`