A CLASS OF DISCONTINUOUS PETROV-GALERKIN METHODS. PART II: OPTIMAL TEST FUNCTIONS

L. DEMKOWICZ AND J. GOPALAKRISHNAN

ABSTRACT. We lay out a program for constructing discontinuous Petrov-Galerkin (DPG) schemes having test function spaces that are automatically computable to guarantee stability. Given a trial space, a DPG discretization using its optimal test space counterpart inherits stability from the well-posedness of the undiscretized problem. Although the question of stable test space choice had attracted the attention of many previous authors, the novelty in our approach lies in the fact we identify a discontinuous Galerkin (DG) framework wherein test functions, arbitrarily close to the optimal ones, can be *locally* computed. The idea is presented abstractly and its feasibility illustrated through several theoretical and numerical examples.

1. INTRODUCTION

In this paper, we develop a class of Petrov-Galerkin methods that are self-adapting to a given boundary value problem, in the following sense: Given a variational formulation of a boundary value problem, and an approximation space to find numerical solutions, a space of test functions to construct a stable scheme is automatically computed. Traditionally, finite element methods are constructed by fixing test and trial spaces (usually polynomials) on each mesh element. From this perspective, the idea of adapting test function spaces on the fly, forms a fresh paradigm in the construction of Galerkin methods. Of course, its success is dependent on how easily one can compute appropriate test function spaces. We show that this can indeed be easily (in fact, locally) done, once we use variational formulations of the discontinuous Galerkin (DG) type. Our study proceeds by considering some specific examples of boundary value problems, and applying the paradigm to obtain new Petrov-Galerkin schemes. In this paper, we are particularly interested in the example of convection-dominated diffusion. While we are able to present a number of theoretical results for one dimensional examples, the main purpose of this paper is to illustrate the feasibility of the idea computationally in two dimensions.

By a Petrov-Galerkin method, we mean a generalization of the original Galerkin method (also known as the Bubnov-Galerkin method), in which one uses different trial and test spaces. For a detailed historical review on the Galerkin method, we refer to the introduction in the book of Mikhlin [29] who, in particular, refers to the original contribution of Petrov [33]. The idea of Petrov-Galerkin method was exploited early by Mitchell and Griffiths in the context of finite difference methods – see [30] – but it was fully realized in the famous Streamline Upwind Petrov Galerkin Method (SUPG) of Hughes et al., see e.g. [24, 25].

Demkowicz was supported in part by the Department of Energy [National Nuclear Security Administration] under Award Number [DE-FC52-08NA28615], and by a research contract with Boeing.

Gopalakrishnan was supported in part by the National Science Foundation under grant DMS-0713833.

One of the main features of our methods presented here is the use of discontinuous approximation spaces. DG methods, originating from the early papers for the purely advective case [26, 28, 34], have emerged as a powerful alternatives for advective problems and their perturbations. In particular, for our primary example of the convection dominated diffusion problem, DG methods have remained an active area of research [13, 15, 20, 21, 22, 23, 31]. While earlier research concentrated on tailoring numerical fluxes for upwinding, maintaining high approximation order, and addition of stabilization terms [4, 15, 22], there is a resurgence of interest [20, 31] focusing on adapting the recent hybridized (HDG) framework [14] to convective problems. A feature common to our methods and the HDG methods is the idea of letting certain inter-element numerical traces to be unknowns.

One of our points of departure from the standard DG methods is in the Petrov-Galerkin formalism. In Part I [17] of this series, we clarified the main design principle in Petrov-Galerkin schemes: Namely, while it is theoretically necessary to set trial spaces with good approximation properties, the test spaces can be chosen without regard to their approximation properties and solely to obtain stability. In this sequel to [17], we take this idea further and investigate how one may design test spaces to achieve stability in a natural "energy norm" (to be defined in the next section). Specifically, given a trial space, we are interested in automatically computing a basis for a test space that has an (almost) optimal stability constant in the energy norm. The concept of determining optimal test functions numerically is not new and resurfaces from time to time in literature, see [3, 11, 18, 19, 35, 36], just to mention a few. Its connection with the least squares Galerkin method is also previously known [8, Remark 2.4] (and we will explain this at length in Section 2). The novelty in our approach is a clear identification of possibility of locally computing such optimal test functions, made possible by using a DG framework.

We are not the first to consider Petrov-Galerkin methods in the DG framework. In fact, schemes christened "DPG methods" (discontinuous Petrov-Galerkin methods) already exist [5, 6, 11, 12]. In [5], a method of the mixed form for the Laplace's equation with test spaces enriched with bubbles are considered. An error analysis, higher order generalizations and hybridization aspects of such bubble-enriched methods are discussed in [11]. Methods along the same theme for advection-diffusion equations were considered in [6] and [12]. While the methods we present in this paper are also DPG methods (in that we also use discontinuous, non-equal test and trial spaces), the major difference in our aim from existing works is that we want to automatically compute almost optimal test functions for any given trial space. This leads to methods that are rather different from the previously proposed ones (e.g., we do not use bubbles in our test spaces).

The paper is organized as follows. The next section presents the idea in an abstract framework, and details the points to be noted in a practical realization, as well as the connection with least squares Galerkin methods. Section 3 discusses concretely the application of the idea to the transport equation, yielding two new DPG methods, which are then compared to the DPG method of [17]. Later sections deal with the convection dominated diffusion problem in one and two dimensions.

2. The concept of optimal test functions

We now explain what we mean by "optimal test functions" in the general context of an abstract variational boundary-value problem:

$$\text{Find } u \in U: \quad b(u,v) = l(v) \qquad \forall v \in V. \tag{2.1}$$

Here U and V are real Hilbert spaces (normed by $\|\cdot\|_U$ and $\|\cdot\|_V$, resp.), l is a continuous linear form defined on test space V, $b(\cdot,\cdot)$ denotes a bilinear form defined on $U \times V$ that is continuous, i.e.,

$$|b(u,v)| \le M\|u\|_U \, \|v\|_V, \tag{2.2}$$

and satisfies the inf-sup [1, 9] condition

$$\inf_{\|u\|_U=1} \sup_{\|v\|_V=1} b(u,v) \ge \gamma \tag{2.3}$$

with $\gamma > 0$. Above the infimum and supremum runs over all u and v in the unit balls of U and V, resp. (we will tacitly use such notations throughout). Additionally we assume that

$$\{v \in V: \ b(u,v) = 0 \quad \forall\, u \in U\} = \{0\}. \tag{2.4}$$

Under these conditions, it is well known [2] that problem (2.1) has a unique solution for any $\ell \in V'$ (primes are used to denote dual spaces).

Let us also recall the famous result of Babuška on approximation of (2.1) by the following Galerkin method obtained using subspaces $U_n \subseteq U$ and $V_n \subseteq V$, with $\dim U_n = \dim V_n$:

$$\begin{cases} \text{Find } u_n \in U_n \text{ satisfying} \\ \qquad b(u_n, v_n) = l(v_n) \quad \forall v \in V_n. \end{cases} \tag{2.5}$$

We assume that the subspaces satisfy the discrete analogue of (2.3) , namely

$$\inf_{\|u_n\|_{U_n}=1} \sup_{\|v_n\|_{V_n}=1} b(u_n, v_n) \ge \gamma_n \tag{2.6}$$

with $\gamma_n > 0$. Then the following result holds:

Theorem 2.1 (Babuška)**.** *Under the above assumptions, the exact and the discrete problems* (2.1) *and* (2.5) *are uniquely solvable. Furthermore,*

$$\|u - u_n\|_U \le \frac{M}{\gamma_n} \inf_{w_n \in U_n} \|u - w_n\|_U.$$

In the early paper [2], we find this result with the constant M/γ_n replaced by $1 + M/\gamma_n$. It is now well known that in the Hilbert space setting, one can remove the "1" in the constant (see [27], [37, Theorem 2], or [16, Theorem 4]).

The starting point of our analysis is the definition of an alternative norm, which we call the *energy norm* on the trial space U. It is defined by

$$\|u\|_E \stackrel{\text{def}}{=} \sup_{\|v\|_V=1} b(u,v). \tag{2.7}$$

From this definition the following result is immediate:

Proposition 2.1. *The energy norm $\|\cdot\|_E$ is an equivalent norm on U, specifically,*

$$\gamma\|u\|_U \le \|u\|_E \le M\|u\|_U, \qquad \forall u \in U,$$

if and only if (2.2) *and* (2.3) *hold.*

Next, consider the map from trial to test space $T : U \mapsto V$ defined as follows: For every $u \in U$, we define Tu in V as the unique solution of

$$(Tu, v)_V = b(u, v), \qquad \forall\, v \in V, \tag{2.8}$$

where $(\cdot, \cdot)_V$ denotes the inner product of V. By the Riesz representation theorem, T is a well defined map. The following proposition is now obvious from Hilbert space theory:

Proposition 2.2. *For any u in U, the supremum in (2.7) is attained by $v = Tu \in V$. The norm $\|u\|_E$ is generated by the inner product*

$$(u, u)_E \stackrel{\text{def}}{=} (Tu, Tu)_V.$$

Finally, let us consider a Petrov-Galerkin scheme of the form (2.5), with a finite dimensional trial subspace

$$U_n = \text{span}\{e_j : \ j = 1, \dots n\} \tag{2.9}$$

for some linearly independent set of functions e_j in U.

Definition 2.1. Every trial subspace U_n, as in (2.9), has its corresponding ***optimal test space***, defined by

$$V_n = \text{span}\{Te_j : \ j = 1, \dots n\}.$$

Test spaces defined above are "optimal" in the sense that it generates the best possible ratio of continuity constant to stability constant when U is endowed with the energy norm. Specifically, we have the following result:

Theorem 2.2. *Let V_n be the optimal test space corresponding to a finite dimensional trial space U_n. Then the error in the Petrov-Galerkin scheme (2.5) using $U_n \times V_n$ equals the best approximation error in the energy norm, i.e.*

$$\|u - u_n\|_E = \inf_{w_n \in U_n} \|u - w_n\|_E \tag{2.10}$$

Proof. We apply Theorem 2.1, but with $\|\cdot\|_E$ as the norm for U. Then, by (2.8), the continuity inequality

$$b(u, v) = (Tu, v)_V \le \|u\|_E \|v\|_V,$$

holds with unit constant for all $u \in U$, $v \in V$. The inf-sup condition (2.6) also holds with unit constant:

$$\begin{aligned} \sup_{\|v_n\|_{V_n}=1} b(u_n, v_n) &= \sup_{\|v_n\|_{V_n}=1} (Tu_n, v_n)_V \\ &\ge (Tu_n, \frac{Tu_n}{\|Tu_n\|_V})_V = \|u_n\|_E, \end{aligned}$$

for all $u_n \in U_n$ and $v_n \in V_n$, where we have used Proposition 2.2. Hence, by Theorem 2.1, the left hand side of (2.10) is bounded by the right hand side. The reverse inequality is obvious. □

A practical realization of this Petrov-Galerkin method with optimal test space involves approximating the operator T in (2.8) by some computable analogue. Application of this approximate operator is required to be inexpensive. With this in mind, our derivations of practical schemes proceed in the following general steps:

S1. Given a boundary value problem, as a first step, we develop mesh dependent variational formulations $b(\cdot,\cdot)$ with an underlying space V which allows *inter-element discontinuities*. It is for this reason our schemes are named "discontinuous" Petrov-Galerkin (DPG) schemes.

S2. The next step is to choose a trial subspace U_n. As is clear from Theorem 2.2, trial spaces must always be chosen with *good approximation* properties. Hence they are typically standard piecewise polynomial spaces, with degree determined by the local order of accuracy needed.

S3. The third step is to approximately compute optimal test functions. Since we allowed inter-element discontinuities in V in step S1, we are able to approximate T by a *local, element-by-element* computable approximation $T_n : U_n \mapsto \tilde{V}_n$ such that

$$(T_n u_n, \tilde{v}_n)_V = b(u_n, \tilde{v}_n), \quad \forall \tilde{v}_n \in \tilde{V}_n, \tag{2.11a}$$

and

$$T_n \text{ is injective on } U_n, \tag{2.11b}$$

where $\tilde{V}_n \subseteq V$ is a computationally convenient space of discontinuous functions, used to represent the approximate optimal test space. If $\{e_j\}$ forms a basis for U_n, then we set $V_n = \text{span}\{t_j\}$ where $t_j = T_n e_j$. Note that $\{t_j\}$ forms a basis for V_n due to (2.11b).

S4. The final step is to solve a *symmetric positive definite* matrix system. Indeed, regardless of any asymmetry of $b(\cdot,\cdot)$, we always arrive at a symmetric linear system, because the (i,j)th entry of the stiffness matrix of (2.5) is

$$\begin{aligned} b(e_j, t_i) &= (T_n e_j, t_i)_V && \text{by (2.11a)}, \\ &= (T_n e_j, T_n e_i)_V && \text{as } t_i = T_n e_i \\ &= (T_n e_i, T_n e_j)_V \\ &= b(e_i, t_j), \end{aligned}$$

thus coinciding with the (j,i)th entry. The positive definiteness is a consequence of (2.11b).

We do not have a universal prescription for selecting $\tilde{V}_n$. Its dimension must at least be $\dim(U_n)$, as otherwise (2.11b) would be violated. The motivation is that as $\tilde{V}_n$ gets richer, the discrete energy norm $\|T_n u_n\|_V$ may be expected to converge to $\|T u_n\|_V$, so the discrete method should increasingly inherit the stability properties of the exact problem.

It may seem like the ambiguity in the choice of $\tilde{V}_n$ makes the design of the method less automatic. But it is possible to use *hp*-adaptivity within each mesh element to make the computation of the right $\tilde{V}_n$ almost automatic. This is our eventual goal. However, before realizing this goal, we must study the local problems whose solutions form the optimal test space, which is one of the purposes of this paper. In the remaining sections, we will numerically some simple choices of $\tilde{V}_n$ for specific examples.

We close this section by explaining the connection with the ***least squares Galerkin method***. The equation defining the Petrov-Galerkin method with U_n and its optimal test space V_n, namely,

$$b(u_n, v_n) = \langle l, v_n \rangle_V, \quad \forall\, v_n \in V_n, \tag{2.12}$$

can be rewritten as

$$(Tu_n, Tw_n)_V = \langle l, Tw_n\rangle_V \quad \forall\, w_n \in U_n,$$

where $\langle\cdot,\cdot\rangle_V$ denotes the duality pairing in V. In other words, u_n solves $T^*\circ Tu_n = T^*R_V l$, where T^* is the V-adjoint of T, and R_V is the inverse of the Riesz map defined by $R_V : V' \mapsto V$ by $(R_V(\ell), v)_V = \langle \ell, v\rangle_V$. Thus our method is indeed of the least squares type. It is also related to the so-called "negative-norm least squares method" [7, 8, 10]. To see this, first note that (2.8) implies that

$$T = R_V \circ B$$

where $B : U \mapsto V'$ is the operator generated by the bilinear form , i.e., $\langle Bu, v\rangle_V = b(u, v)$ for all $u \in U, v \in V$. Then (2.12) can equivalently be rewritten as

$$\langle Bu_n, R_V \circ Bw_n\rangle_V = \langle l, R_V \circ Bw_n\rangle_V, \tag{2.13}$$

for all $w_n \in U_n$. A typical setting for negative-norm least squares techniques is the above equation with $V = H_0^1(\Omega)$. Here is where we differ from these techniques: It is not easy to obtain *local* and easily computable approximations to $T = R_V \circ B$ when V has global H^1-conformity. Techniques in [7, 8] approximate T by approximating R_V by a preconditioner. E.g., a standard multigrid preconditioner for the Laplace operator can serve as a good approximation for $R_V : H^{-1}(\Omega) \mapsto H_0^1(\Omega)$. However such operators are global, and require multilevel meshes and other such overhead. In contrast, in our approach, we use spaces V without inter-element continuity constraints, thus permitting simpler and local approximations to T.

3. First example: Pure convection

In this section, we will present two new DPG methods for the transport equation. They are derived by following the program of steps S1–S4 introduced abstractly in the previous section. Both are different from the DPG method introduced in Part I [17]. While maintaining the excellent approximation qualities of the first DPG method, one of the new methods is easier to implement. These methods will be presented in § 3.3, after we illustrate the concepts using simple one-dimensional examples in § 3.1 and § 3.2.

Consider the convection problem

$$\begin{cases} \boldsymbol{\beta}\cdot\nabla u &= f \qquad \text{in } \Omega \\ u &= u_0 \qquad \text{on } \Gamma_{in}. \end{cases} \tag{3.14}$$

Here $\Omega \subset \mathbb{R}^n$, $n = 1, 2$, and Γ_{in} denotes the inflow boundary,

$$\Gamma_{in} = \{\boldsymbol{x} \in \partial\Omega \,:\, \boldsymbol{\beta}\cdot\boldsymbol{n}(\boldsymbol{x}) < 0\}. \tag{3.15}$$

Given a partition of Ω into finite elements K, we multiply the convection equation with a test function v supported on K, and integrate by parts over the element K to obtain

$$-\int_K u\,\partial_\beta v + \int_{\partial K} \beta_n uv = \int_K fv$$

Here $\partial_\beta v = \boldsymbol{\beta}\cdot\nabla v$ and $\beta_n = \boldsymbol{\beta}\cdot\boldsymbol{n}$. Whenever the measures in the integration are obvious, we omit them, for simplifying notation (e.g., the first integral is to be read with

n-dimensional measure dx, while the second with $n-1$-dimensional measure ds). The flux,

$$q = |\beta_n| u \tag{3.16}$$

will be identified as an independent, new unknown. Due to a possible degeneration of β_n to zero, it is more suitable to work with the product $\beta_n u$ than u alone (see [17, § 2.3] for a detailed explanation). Let Γ_h denote the union of all interelement boundaries minus the inflow boundary Γ_{in}. Then, the above leads to the following variational formulation:

$$\begin{cases} \text{Find } u \in L^2(\Omega),\ q \in L^2(\Gamma_h) :\ \text{such that} \\ b(\,(u,q),\ v) = \ell(v), \quad \forall v \in H_\beta(K),\ \forall K \end{cases} \tag{3.17}$$

where

$$b(\,(u,q),\ v) = \sum_K \int_K -u\,\partial_\beta v + \int_{\partial K \setminus \Gamma_{in}} \mathrm{sgn}(\beta_n) q v, \tag{3.18a}$$

$$\ell(v) = \sum_K \int_K f v + \int_{\partial K \cap \Gamma_{in}} \beta_n u_0 v, \tag{3.18b}$$

$$H_\beta(K) = \{v \in L^2(K)\ :\ \partial_\beta v \in L^2(K)\} \tag{3.18c}$$

and $\mathrm{sgn}(x)$ denotes the sign of x. Note that q is single-valued on element interfaces, thus coupling the mesh elements. The development of a mesh dependent variational formulation, such as the above, is the first step (which we labeled S1 previously) according to the program outlined in Section 2. Recall that in the step S1, the test space was required to allow functions with inter-element discontinuities. In this example,

$$U = L^2(\Omega) \times L^2(\Gamma_h), \tag{3.19a}$$

$$V = \{v :\ v \in H_\beta(K),\ \forall \text{ elements } K\}, \tag{3.19b}$$

so neither the test nor the trial space has any inter-element continuity. Let us now proceed with the remaining steps in constructing the method discussed in Section 2, beginning with a simple one-dimensional (1D), one-element, scenario.

3.1. **A spectral discretization in 1D.** In the 1D case, we shall assume $\beta = 1$. Let $K = (x_1, x_2)$ be a finite element. The space $H_\beta(K)$ coincides with $H^1(K)$. We can endow it with a Hilbert structure using the inner product

$$(v,w)_V = \int_{x_1}^{x_2} v'\,w' \,+\, v(x_2)\,w(x_2), \tag{3.20}$$

where the primes denote differentiation. Then, the bilinear form

$$b(\,(u,q),\ v) = \int_{x_1}^{x_2} u\,v' + q\,v(x_2)$$

satisfies (2.2) with the natural L^2-norm on U and the above defined norm on V. Moreover, its is easy to see that the inf-sup condition (2.3) also holds.

According to step S2, we now select the trial space. Let $\mathcal{P}^p(K)$ denote polynomials of degree at most p on K. Set the trial space to

$$U_p = \mathcal{P}_p(K) \times \mathbb{R}.$$

In other words, a function in U_p is of the form (u_p, q) for some $u_p \in \mathcal{P}^p(K)$ (wherein u is approximated) *and* in addition one point value q (used for approximating the flux at x_2).

In this simple 1D case, it is possible to exactly calculate the optimal test functions. Since we can analytically compute the action of the exact T-operator defined in (2.8), there is no need to approximate T by any T_h. Equation (2.8), when written out to give the variational problem for the optimal test function corresponding to the flux at x_2, takes the following form: The test function $v_q \equiv T(0,1)$ is the unique function in $H^1(x_1,x_2)$ satisfying

$$\int_{x_1}^{x_2} v_q' \, \delta_v' \, + \, v_q(x_2) \, \delta_v(x_2) = \delta_v(x_2), \tag{3.21}$$

for all $\delta_v \in H^1(K)$. Its easily computed solution is

$$v_q \equiv 1, \tag{3.22}$$

i.e., the constant extension of the (outflow) unit flux at x_2.

The optimal test function $v_u \equiv T(u,0,0)$ corresponding to the interior polynomial trial function $u \in \mathcal{P}^p(K)$, is the unique function in $H^1(K)$ satisfying

$$\int_{x_1}^{x_2} v_u' \, \delta_v' + v_u(x_2) \, \delta_v(x_2) = - \int_{x_1}^{x_2} u \, \delta_v', \tag{3.23}$$

for all $\delta_v \in H^1(K)$. This can also be solved easily leading to the next optimal test function

$$v_u(x) = \int_x^{x_2} u(s) \, ds. \tag{3.24}$$

Notice that this test function is a polynomial of (one higher) degree $p+1$.

Thus, combining (3.22) and (3.24), we find that the optimal test space corresponding to our chosen trial space U_p is

$$V_p = \text{span}\{v_u, v_q : \; u \in \mathcal{P}^p(K), \; q \in \mathbb{R}\}. \tag{3.25}$$

We can now apply Theorem 2.2 to this method with the pair U_p, V_p and obtain p-optimal error estimates.

To conclude this discussion of our simplest example, we make the following remarks.

Remark 3.1. Different choices of inner products for V leads to different optimal test functions. Choosing, e.g., a more standard inner product,

$$(v, \delta_v)_V = \int_{x_1}^{x_2} (v' \delta_v' + v \delta_v),$$

in place of (3.20), it is easy to see that we obtain non-polynomial optimal test functions (even for polynomial trial functions).

Remark 3.2. To compare with the spectral 1D analogue of our first DPG method in Part I, [17, Section 2], observe that the test space in (3.25) equals $\mathcal{P}^{p+1}(K)$, which is the same as the first DPG method considered in 1D.

3.2. **A multielement 1D discretization.** To obtain the multielement version of the method in § 3.1, we consider the domain $\Omega = (x_0, x_n)$ split into elements (x_i, x_{i+1}). As in § 3.1, we start by setting the inner product on V to

$$(u,w)_V = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} v' \, w' \, + \, \alpha_i \, v^{\text{up}}(x_i) \, w^{\text{up}}(x_i),$$

where $v^{\rm up}(x_i)$ denotes the limit of $v(x)$ as x approaches x_i from the left (upwind), and α_i's are positive scaling factors to be determined. We will also use the downwind limit (from the right) at x_i, denoted by $v^{\rm dn}(x_i)$.

Next, let us set the trial space by

$$U_h = \{(w_h, q_1, \ldots, q_n) : w_h|_{(x_i,x_{i+1})} \in \mathcal{P}^p(x_i, x_{i+1}), \\ \text{and } q_i \in I\!R\}. \tag{3.26}$$

The first component w_h is used to approximate u, while the remaining components are outward (rightward) fluxes, i.e., q_i are used to outward flux (to the right) at x_i. The bilinear form (3.18a) now has the form

$$b(\,(u,q),\ v) = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} -u\,v' + q_i\,v^{\rm up}(x_i) - q_{i-1}\,v^{\rm dn}(x_{i-1}),$$

with the understanding that $q_0 = 0$.

As in § 3.1, we can exactly calculate the optimal test functions in this case. The optimal test function corresponding to a trial function w in $\mathcal{P}^p(x_i, x_{i+1})$ is obtained as in (3.24) by

$$v^w(x) = \int_x^{x_{i+1}} w(s)\,ds. \tag{3.27}$$

The optimal test function corresponding to the flux q_i, unlike the one-element case of § 3.1, is now supported on two adjacent elements (except for the last flux q_n). The optimal test function v_i corresponding to the unit flux $q_i = 1$ at x_i, for all $i = 1, \ldots, n-1$, is obtained by solving the two equations

$$\int_{x_{i-1}}^{x_i} v_i'\,\delta_v' \;+\; \alpha_i\,v_i^{\rm up}(x_i)\,\delta_v^{\rm up}(x_i) = \delta_v^{\rm up}(x_i),$$

for all $\delta_v \in H^1(x_{i-1}, x_i)$, and

$$\int_{x_i}^{x_{i+1}} v_i'\,\delta_v' \;+\; \alpha_{i+1}v_i^{\rm up}(x_{i+1})\,\delta_v^{\rm up}(x_{i+1}) = -\delta_v^{\rm dn}(x_i),$$

for all $\delta_v \in H^1(x_i, x_{i+1})$, independently. The solution is

$$v_i(x) = \begin{cases} \dfrac{1}{\alpha_i} & \text{if } x \in (x_{i-1}, x_i) \\ x - \dfrac{1+\alpha_{i+1}x_{i+1}}{\alpha_{i+1}} & \text{if } x \in (x_i, x_{i+1}) \\ 0 & \text{elsewhere.} \end{cases} \tag{3.28}$$

Similarly, we see that the exactly optimal test function corresponding to the unit flux at x_n, namely v_n, is the indicator function of the last element scaled by $1/\alpha_n$.

Thus, the exactly optimal test space corresponding to the U_h in (3.26) equals

$$V_h = \text{span}\{v^w : \ \text{for all } w \in \mathcal{P}^p(K), \text{ and} \\ v_i : \ \text{for all } i = 1, \ldots, n\}.$$

Theorem 3.1. *For the 1D DPG method with the above defined $U_h \times V_h$, the following statements hold:*

(1) *The energy norm for this example is given by*

$$\|(u, q_1, \ldots, q_n)\|_E^2 = \sum_{i=1}^{n} \frac{|q_i - q_{i-1}|^2}{\alpha_i} + \int_{x_{i-1}}^{x_i} |u - q_{i-1}|^2. \tag{3.29}$$

(2) *For all $u \in L^2(x_0, x_n)$ and all $q = (q_1, \ldots, q_n) \in \mathbb{R}^n$, the inf-sup condition*

$$\|u\|^2 + \|q\|_h^2 \le \gamma \|(u, q_1, \ldots, q_n)\|_E^2 \tag{3.30}$$

holds where $\|u\|$ denotes the $L^2(x_0, x_n)$-norm, $\|q\|_h^2 = \sum_{i=1}^{n} |q_{i-1}|^2 (x_i - x_{i-1})$, and $\gamma = \max(3\kappa, 2)$, with $\kappa = \sum_{\ell=1}^{n} \sum_{j=1}^{\ell-1} \alpha_j (x_\ell - x_{\ell-1})$.

(3) *The test space can be characterized as*

$$V_h = \{v : \ v|_K \in \mathcal{P}^{p+1}(K) \text{ for all elements } K\}.$$

(4) *The solution $(u_h, q_{h,1}, \ldots, q_{h,n})$ of this DPG method is independent of $\{\alpha_i\}$.*
(5) *The error in the fluxes $q_{h,i}$ is zero, i.e., $q_{h,i} = q_i$.*
(6) *The solution u_h equals the L^2-projection of the exact solution u.*

Proof. See Appendix A. □

Remark 3.3. The 1D analogue of the method considered in [17] is equivalent to the method presented in this subsection. Indeed by Theorem 3.1, item 3, we find that the optimal test space is the same as the test space for the 1D analogue of the method considered in [17]. Hence their solutions coincide. Note however that while we solve a symmetric positive definite system for the current DPG method, the first DPG method of [17] obtains the solution by backsubstitution of a block triangular system.

3.3. **Discretization in 2D.** To derive the new DPG methods for the pure advection problem, we follow the program of steps S1–S4 in Section 2. First, we set a particular form of the inner product on the test space, namely

$$(v, \delta_v)_V = \int_K \partial_\beta v \, \partial_\beta \delta_v + \int_K v \, \delta_v, \tag{3.31}$$

The spaces U, V, and the bilinear form $b(\cdot, \cdot)$ are as before (see (3.19) and (3.18a)). Thus, we have performed step S1 in the derivation of the method.

Since $\boldsymbol{\beta}$ can vary from point to point even within mesh elements, the well-posedness of the transport problem is not clear in all situations. We will therefore assume that

$$\nabla \cdot \boldsymbol{\beta} = 0, \tag{3.32}$$

and that the inf-sup condition (2.3) and (2.4) hold for our form $b(\cdot, \cdot)$. It is easy to construct examples with variable advection where these assumptions break down, e.g., if λ_i denotes the barycentric coordinates of a triangle, then setting $v = \lambda_1 \lambda_2 \lambda_3$ and $\boldsymbol{\beta} = \mathbf{curl}(\lambda_1 \lambda_2 \lambda_3)$, we find that $\boldsymbol{\beta} \cdot \nabla v = 0$ and $v|_{\partial K} = 0$. This function v violates (2.4). Our assumptions rule out such advection fields. To handle cellular convection and other such important examples of advection with closed loops, we must at least start with a mesh that is sufficiently refined so that situations like the above do not occur.

The next step S2, is to set the space of trial functions. We set this as in the first DPG method [17] to facilitate comparison, namely

$$U_h = \{(w_h, \phi_h) : \ w_h|_K \in \mathcal{P}^p(K), \ \phi_h|_E \in \mathcal{P}^{p+1}(E), \\ \forall \text{ elements } K \text{ and } \forall \text{ edges } E \subseteq \Gamma_h\}.$$

The optimal test functions corresponding to this U_h solve the following variational problem:

$$\begin{cases} \text{Find } v \in H_\beta(K) \text{ satisfying} \\ \displaystyle\int_K \partial_\beta v \, \partial_\beta \delta_v + \int_K v \, \delta_v \ = -\int_K u \, \partial_\beta \delta_v \ + \int_{\partial K} \mathrm{sgn}(\beta_n) q \, \delta_v \\ \text{for all } \delta_v \in H_\beta(K), \end{cases} \tag{3.33}$$

for every (u, q) in U_h.

The third step S3 involves approximating (3.33) by replacing $H_\beta(K)$ by a computable finite dimensional subspace $\tilde{V}_h$. We set

$$\tilde{V}_h = \{v : \ v|_K \in \mathcal{P}^{p+2}(K) \text{ for all mesh elements } K\}.$$

The approximately optimal test functions $v_h \in \tilde{V}_h$ are computed by solving (3.33) for all δ_v in $\tilde{V}_h$ (instead of $H_\beta(K)$), i.e., by solving the discrete problem:

$$\begin{cases} \text{Find } v \in \mathcal{P}^{p+2}(K) \text{ satisfying} \\ \displaystyle\int_K \partial_\beta v \, \partial_\beta \tilde{\delta}_v + v \tilde{\delta}_v \ = -\int_K u \partial_\beta \tilde{\delta}_v + \int_{\partial K} \mathrm{sgn}(\beta_n) q \, \tilde{\delta}_v, \\ \text{for all } \tilde{\delta}_v \in \mathcal{P}^{p+2}(K), \end{cases} \tag{3.34}$$

for each member of any local basis of U_h. The span of such v's forms the test space V_h. The DPG method, discretizing the convection problem (3.17) using these spaces is

$$\begin{cases} \text{Find } (u_h, q_h) \in U_h \text{ satisfying} \\ \qquad\qquad b(\,(u_h, q_h), v_h) = \ell(v_h), \quad \forall \, v_h \in V_h, \end{cases}$$

with b and ℓ as in (3.18), and will be called the *"DPG-A" method* or the DPG method with approximately optimal test functions. In contrast, the method of [17] will be called the *"DPG-1" method*.

The DPG-1 method was presented in [17] for the case of *constant* advection, i.e., assuming that $\boldsymbol{\beta}$ is constant. In this case, one may wonder if it is possible to analytically compute exactly optimal test functions. This is indeed possible and gives rise to what we denote by the *"DPG-X" method*. To make the calculations convenient, the DPG-X method is derived using a different inner product in V, i.e., instead of (3.35), we use the (equivalent) inner product

$$(v, \delta_v)_V = \int_K \partial_\beta v \ \partial_\beta \delta_v + \int_{\partial_{\mathrm{out}} K} v \ \delta_v, \tag{3.35}$$

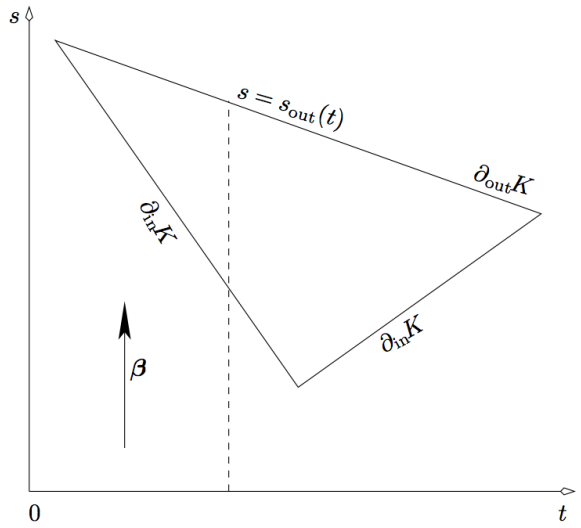


FIGURE 1. Streamline crossing an element.

where $\partial_{\text{out}}K$ denotes the outflow part of boundary ∂K. The optimal test function problem, modified from (3.33), now reads as follows:

$$\begin{cases} \text{Find } v \in H_\beta(K) \text{ satisfying} \\ \displaystyle\int_K \partial_\beta v\, \partial_\beta \delta_v + \int_{\partial_{\text{out}}K} v\, \delta_v = \quad -\int_K u\, \partial_\beta \delta_v + \int_{\partial_{\text{out}}K} q\, \delta_v - \int_{\partial_{\text{in}}K} q\, \delta_v, \\ \text{for all } \delta_v \in H_\beta(K). \end{cases} \tag{3.36}$$

To describe the method, it now suffices to describe the exact solution of this problem.

To this end, rewrite (3.36) as a classical boundary value problem (obtained by integrating the weak form by parts and using (3.32)):

$$\partial_\beta \partial_\beta v = -\partial_\beta u \qquad \text{in } K \tag{3.37a}$$

$$\beta_n \partial_\beta v = -\beta_n u - q \qquad \text{on } \partial_{\text{in}}K \tag{3.37b}$$

$$\beta_n \partial_\beta v + v = -\beta_n u + q \qquad \text{on } \partial_{\text{out}}K. \tag{3.37c}$$

To describe the solutions v for any (u, q) in U_h, first consider the case $q|_{\partial_{\text{out}}K} = q|_{\partial_{\text{in}}K} = 0$. Then we obtain the optimal test function for u by solving (3.37), namely by integrating u along streamlines with a zero initial condition on the outflow boundary. This function is a polynomial of degree $p + 1$ in the streamline coordinate (say s). In the remaining coordinate (say t), it is of degree p, if there is only one outflow edge. Irrespective of the number of outflow edges however, these functions are included in the test space of the DPG-1 method [17]. Next, consider the optimal test function for $q|_{\partial_{\text{out}}K}$ by setting $q|_{\partial_{\text{in}}K}$ and u to 0. Now the solutions v are the values of $q|_{\partial_{\text{out}}K}$ extended as constant along streamlines. The span of these functions together with the optimal test functions for u form the test space of the original DPG-1 method [17].

The sole difference between the test function spaces of DPG-X and DPG-1 methods lies in the optimal test functions for $q|_{\partial_{\text{in}}K}$. While these were set to zero in the DPG-1 method, the DPG-X method sets them by solving (3.36). The exact solution is a function linear in the streamline coordinate s. Indeed, rewriting the problem (3.36) for $v(s, t)$ in

terms of s,

$$\begin{aligned} \frac{\partial^2 v}{\partial s^2} &= 0 && \text{in } K \\ B_{\text{in}} \frac{\partial v}{\partial s} &= -q|_{\partial_{\text{in}} K} && \text{on } \partial_{\text{in}} K \\ B_{\text{out}} \frac{\partial v}{\partial s} + v &= 0 && \text{on } \partial_{\text{out}} K. \end{aligned}$$

where $B_{\text{in}} = \boldsymbol{\beta} \cdot \boldsymbol{n}_{\text{in}} |\boldsymbol{\beta}|$ and $B_{\text{out}} = \boldsymbol{\beta} \cdot \boldsymbol{n}_{\text{out}} |\boldsymbol{\beta}|$. Solving, we obtain the explicit formula

$$v = \frac{q}{B_{\text{in}}}(-s + s_{\text{out}} + B_{\text{out}}). \tag{3.38}$$

where s_{out} is the value of the s-coordinate on the outflow boundary (see Fig. 1). If the inflow boundary contains only one edge, then the optimal test function is already in the span of test functions corresponding to the outflow flux and solution u. If, however, the inflow boundary contains two edges, the corresponding test space is enriched with piecewise polynomial test functions, the same way as for the case when the outflow boundary contains two edges. The resulting test space is then different from the one used in our first paper [17]. Unless the streamline is parallel to one of the edges, the test space *always* involves piecewise polynomials generated by boundary fluxes.

3.4. **Numerical comparison.** In the previous subsection, we discussed three methods, DPG-1, DPG-A, and DPG-X, the latter two being newly proposed methods. We will now compare the performance of these three schemes.

First, let us reiterate the following features of these methods for comparison:

- DPG-1 gives rise to a block triangular system, and it can be solved by backsubstitution, marching from the inflow to the outflow edges. The fluxes q_h can be solved for independently of u_h, and u_h can be found by a local postprocessing using q_h (see [17] for details).
- DPG-A and DPG-X gives symmetric positive definite systems. This permits the use of well-developed iterative techniques for solving such systems. (For these methods, we cannot use backsubstitution like in DPG-1.)
- DPG-A is the easiest to implement as its spaces within an element are the most standard – unlike the other methods, it does not have piecewise polynomial test functions, so there is no need for using composite Gauss quadrature.

We now present the numerical results when the method is applied to two well known examples. The first example is due to Peterson [32], who constructed it to show that the h convergence rate of the standard DG method is suboptimal by $h^{1/2}$ (where h denotes the mesh size). Peterson constructed a specific sequence of quasiuniform meshes of obtained by manipulating an $n \times n$ partition of the unit square and discretized the problem of finding u satisfying $\partial u/\partial y = 0, u(x, 0) = \sin 6x, x \in (0, 1)$ (see [32] for further details on this experiment, and also see [17]). We present the $L^2(\Omega)$-norm of the error in u obtained by the three DPG methods and the standard DG method in Fig. 2.

The second example is from [22] which is designed to test any advantages DG methods may have over conforming methods like SUPG when the solution is discontinuous. They

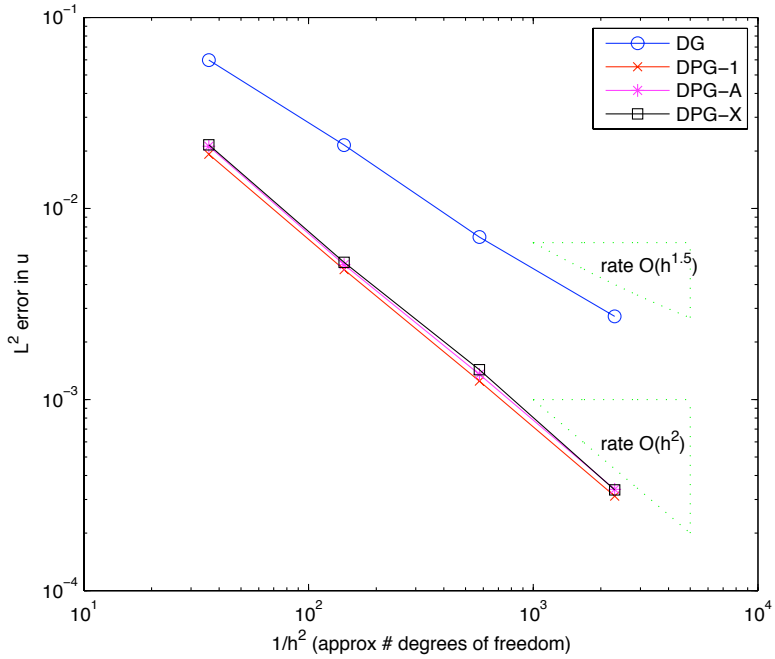


FIGURE 2. Loglog plot of $\|u - u_h\|_{L^2(\Omega)}$ for the Peterson example

set up an advection problem on $\Omega = (-1,1)^2$ with $\boldsymbol{\beta} = (1, 9/10)$ so that the exact solution

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

is discontinuos along the line $y = 9x/10$. While they [22] demonstrated the advantage of DG methods over streamline diffusion type methods, our purpose here is to compare the DPG methods with the DG methods. Fig. 3 presents our results for $p = 1, 2, 3, 4, 5$.

We draw the following conclusions from Figures 2 and 3:

- From Fig. 2, we see that all the DPG methods, including the new DPG-A (easier to implement than the other DPG methods) converged at the *optimal* rate for the Peterson example, while the standard DG method converged suboptimally.
- For the example of [22], the convergence rate for all the DPG methods is again optimal. The methods DPG-1 and DPG-X methods gave almost the same results. DPG-A lagged behind in accuracy, but only slightly. All three methods outperforms the standard DG method.

It is perhaps surprising that the method DPG-A which approximates the optimal piecewise polynomial test functions (discontinuous within element) by polynomials of one higher degree, performs so well. Note however that unlike the DPG-1 methods, we have no theoretical estimates for DPG-A. Moreover, from Figure 3, there appears to be a change in the slope of the convergence of the DPG-A method as one proceeds to the point of the highest degree of freedom. We believe that this is due to local ill-conditioning,

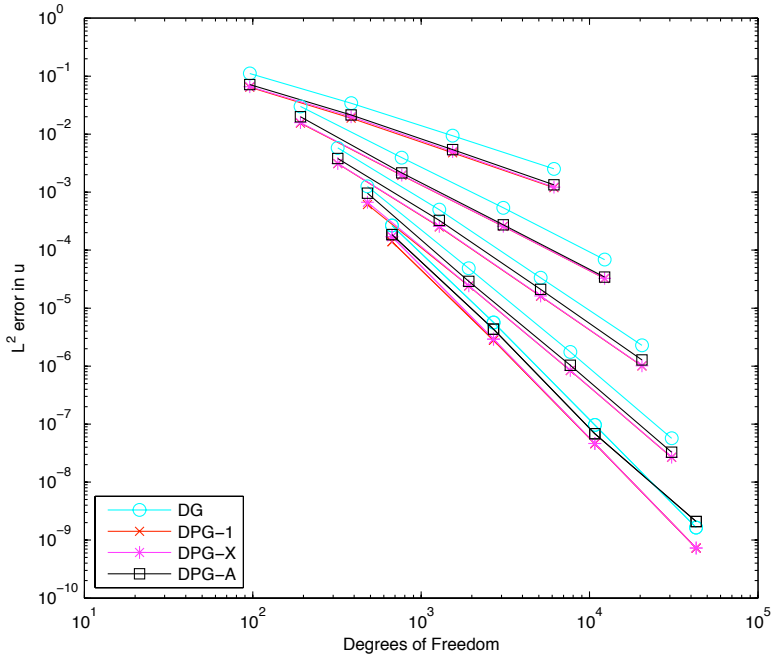


FIGURE 3. Performance of DPG-1, DPG-A, and DPG-X applied to the example of [22] (h convergence rates for $p = 1, 2, 3, 4, 5$)

although more investigations are needed before we can definitively make this conclusion and suggest remedies.

4. Convection-Dominated Diffusion in 1D

This section is devoted to a study of the application of the previous ideas to the convection-diffusion problem. Again, we apply the steps S1–S4, but now to the convection-diffusion problem. Accordingly, the first step is the derivation of a weak formulation for the convection-diffusion problem involving a Hilbert space V that allows discontinuous functions. This often gives rise to nonstandard bilinear forms. A major question that arises then, in the case of singularly perturbed problems like convection-dominated diffusion, is the inf-sup condition for such bilinear forms. We learned in Section 2 that the Petrov-Galerkin method with optimal test functions delivers the best approximation error in what we have named the *energy norm*. However, the inf-sup condition is needed to translate the energy norm estimates to error estimates in more standard norms (the original norm on U). In particular, if the diffusion is an arbitrarily small ϵ, we would like to know how the inf-sup constant changes with ϵ.

To study such issues, we will first calculate the energy norm in the simplest setting of a single-element one-dimensional spectral approximation and study its equivalence with L^2-type norms. Next, we consider an arbitrary *hp* mesh and illustrate the optimal test functions corresponding to fluxes and solution components. Finally, we finish the section with numerical examples illustrating the method.

4.1. **Spectral Method (One Element Case).** Consider the 1D model problem,

$$\begin{cases} u(0) = u_0, \quad u(1) = 0 \\ \dfrac{1}{\epsilon}\sigma - u' \quad\quad = 0 \\ -\sigma' + u' \quad\quad = f \end{cases} \tag{4.39}$$

and the corresponding variational formulation: Find $\sigma \in L^2(0,1), \hat{\sigma}(0) \in \mathbb{R}, \hat{\sigma}(1) \in \mathbb{R}$, and $u \in L^2(0,1)$ such that

$$\begin{cases} \dfrac{1}{\epsilon}\displaystyle\int_0^1 \sigma\tau + \int_0^1 u\tau' & = -u_0\,\tau(0), \\ \displaystyle\int_0^1 \sigma v' \;\; - \int_0^1 uv' \;\; + \hat{\sigma}(0)v(0) - \hat{\sigma}(1)v(1) & = \displaystyle\int_0^1 fv + u_0 v(0) \end{cases}$$

for all τ and v in $H^1(0,1)$. Given $\sigma, \hat{\sigma}(0), \hat{\sigma}(1), u$, we define the corresponding optimal test functions by formulating the following variational problem,

$$(\,(\tau, v), (\delta_\tau, \delta_v)\,)_V = b(\,(\sigma, \hat{\sigma}(0), \hat{\sigma}(1), u),\, (\delta_\tau, \delta_v)\,)$$

where the inner product corresponds to the norm for test functions defined as follows:

$$\begin{aligned} \|(\tau, v)\|^2 &= (1-\alpha)\|\tau\|^2 + \alpha\|v\|^2 \\ \|\tau\|^2 &= \int_0^1 |\tau'|^2 + |\tau(0)|^2 \\ \|v\|^2 &= \int_0^1 |v'|^2 + |v(1)|^2. \end{aligned} \tag{4.40}$$

Here $\alpha \in (0,1)$ is a scaling parameter. The choice of the particular norm for the test function is somehow arbitrary and it may be used to design different versions of the method.

In this example, we are again able to calculate the exactly optimal test functions. Testing individually with δ_τ and δ_v, we obtain two variational problems for the test functions:

$$(1-\alpha)\left(\int_0^1 \tau'\delta_\tau' + \tau(0)\delta_\tau(0)\right) = \frac{1}{\epsilon}\int_0^1 \sigma\delta_\tau + \int_0^1 u\delta_\tau'$$

for all δ_τ in $H^1(0,1)$ and

$$\begin{aligned} \alpha\left(\int_0^1 v'\delta_v' + v(1)\delta_v(1)\right) = \int_0^1 \sigma\delta_v' - \int_0^1 u\delta_v' \\ + \hat{\sigma}(0)\delta_v(0) - \hat{\sigma}(1)\delta_v(1) \end{aligned}$$

for all $\delta_v \in H^1(0,1)$. Rewriting these weak formulations in their corresponding classical formulations, we obtain

$$\begin{cases} -(1-\alpha)\tau'' = \dfrac{1}{\epsilon}\sigma - u' \\ (1-\alpha)\tau'(1) = u(1) \\ (1-\alpha)(-\tau'(0) + \tau(0)) = -u(0) \end{cases} \tag{4.41}$$

and

$$\begin{cases} -\alpha v'' = -\sigma' + u' \\ \alpha(v'(1) + v(1)) = \sigma(1) - u(1) - \hat{\sigma}(1) \\ -\alpha v'(0) = -\sigma(0) + u(0) + \hat{\sigma}(0), \end{cases} \tag{4.42}$$

respectively.

Solving these equations, we can obtain the optimal test functions and thus an explicit expression for the energy norm. Integrating the above formulae, we are led to

$$\begin{aligned} (1-\alpha)\tau' &= \frac{1}{\epsilon}\int_x^1 \sigma + u \\ (1-\alpha)\tau(0) &= \frac{1}{\epsilon}\int_0^1 \sigma \end{aligned} \tag{4.43}$$

and

$$\begin{aligned} \alpha v' &= \sigma - u - \hat{\sigma}(0) \\ \alpha v(1) &= \hat{\sigma}(0) - \hat{\sigma}(1). \end{aligned} \tag{4.44}$$

and the final formula for the energy norm (squared)

$$\begin{aligned} \|(\sigma, \hat{\sigma}(0), \hat{\sigma}(1), u)\|_E^2 = (1-\alpha) &\left(\left\| \frac{1}{\epsilon}\int_x^1 \sigma + u \right\|^2 + |\frac{1}{\epsilon}\int_0^1 \sigma|^2 \right) \\ &+ \alpha \left(\|\sigma - u - \hat{\sigma}(0)\|^2 + |\hat{\sigma}(0) - \hat{\sigma}(1)|^2 \right). \end{aligned}$$

By selecting different coefficient α we can scale the two residuals' contributions to the final norm. We select $\alpha = \frac{1}{2}$ to obtain

$$\begin{aligned} \|(\sigma, \hat{\sigma}(0), \hat{\sigma}(1), u)\|_E^2 = &\left\| \frac{1}{\epsilon}\int_x^1 \sigma + u \right\|^2 + \left| \frac{1}{\epsilon}\int_0^1 \sigma \right|^2 \\ &+ \|\sigma - u - \hat{\sigma}(0)\|^2 + |\hat{\sigma}(0) - \hat{\sigma}(1)|^2. \end{aligned} \tag{4.45}$$

This norm represents the natural stability of the variational formulation This energy norm can be related to standard L^2 norm:

Theorem 4.1 (Inf-sup condition)**.** *There exists a (unit order) constant $C > 0$ independent of ϵ such that*

$$\max\left\{ \|\sigma\|,\ \epsilon^{\frac{1}{2}}\|u\|,\ \epsilon^{\frac{1}{2}}\left\| \frac{1}{\epsilon}\int_x^1 \sigma(s)\, ds \right\|,\ \epsilon^{\frac{1}{2}}|\hat{\sigma}(0)|,\ \epsilon^{\frac{1}{2}}|\hat{\sigma}(1)| \right\} \leq\ C\, \|(\sigma, \hat{\sigma}(0), \hat{\sigma}(1), u)\|_E.$$

Proof. See Appendix B. □

4.2. **The composite DPG method.** We now extend the analysis to the multi-element case using an arbitrary partition,

$$0 = x_0 < x_1 < \ldots < x_{k-1} < x_k < \ldots < x_N = 1$$

The unknowns include $\sigma_k, u_k \in L^2(x_{k-1}, x_k)$ and fluxes $\hat{\sigma}(x_k), \hat{u}(x_k)$, $k = 0, \ldots, N$. Fluxes $\hat{u}(0) = u_0$, $\hat{u}(1) = 0$ are known from the boundary conditions. For each element $K = (x_{k-1}, x_k)$, we have test functions $(\tau, v) = (\tau_k, v_k)$, $\tau_k, v_k \in H^1(x_{k-1}, x_k)$.

For each $k = 1, \ldots, N$, we satisfy the following variational equations,

$$\begin{cases} \dfrac{1}{\epsilon}\displaystyle\int_{x_{k-1}}^{x_k} \sigma_k \tau + \int_{x_{k-1}}^{x_k} u_k \tau' + \hat{u}(x_{k-1})\tau(x_{k-1}) - \hat{u}(x_k)\tau(x_k) = 0 \\ \displaystyle\int_{x_{k-1}}^{x_k} \sigma_k v' - \int_{x_{k-1}}^{x_k} u_k v' - \hat{u}(x_{k-1})v(x_{k-1}) + \hat{u}(x_k)v(x_k) \\ \qquad\qquad + \ \hat{\sigma}(x_{k-1})v(x_{k-1}) - \ \hat{\sigma}(x_k)v(x_k) = \displaystyle\int_{x_{k-1}}^{x_k} fv \end{cases}$$

for every $\tau, v \in H^1(x_{k-1}, x_k)$. Again, for $k = 1$, $\hat{u}(0) = u_0$ is known and is moved to the right-hand side. Similarly, $\hat{u}(1) = 0$ in the last equation for $k = N$.

We choose to work with the following norm for the test functions,

$$\begin{aligned} \|(\boldsymbol{\tau}, \boldsymbol{v})\| &= \left(\sum_{k=1}^{N} \|\tau_k\|^2 + \|v_k\|^2 \right)^{\frac{1}{2}} \\ \|\tau_k\|^2 &= \int_{x_{k-1}}^{x_k} |\tau_k'|^2 + |\tau_k(x_{k-1})|^2 \\ \|v_k\|^2 &= \int_{x_{k-1}}^{x_k} |v_k'|^2 + |v_k(x_k)|^2 \end{aligned} \tag{4.46}$$

The choice is not unique. Different norms lead to different optimal test functions. The particular choice of norms for the H^1-spaces enables determination of the optimal test functions in closed form.

The local variational problems for determining optimal test functions are as follows:

$$\begin{aligned} \int_{x_{k-1}}^{x_k} \tau\delta_\tau + \tau(x_{k-1})\delta_\tau(x_{k-1}) = \frac{1}{\epsilon}\int_{x_{k-1}}^{x_k} \sigma_k \delta_\tau + \int_{x_{k-1}}^{x_k} u_k \delta_\tau' \\ + \hat{u}(x_{k-1})\delta_\tau(x_{k-1}) - \hat{u}(x_k)\delta_\tau(x_k), \end{aligned} \tag{4.47a}$$

for all $\delta_\tau \in H^1(x_{k-1}, x_k)$

$$\begin{aligned} \int_{x_{k-1}}^{x_k} v'\delta_v' + v(x_k)\delta_v(x_k) = \int_{x_{k-1}}^{x_k} \sigma_k v' - \int_{x_{k-1}}^{x_k} u_k v' \\ + \hat{\sigma}(x_{k-1})v(x_{k-1}) - \hat{\sigma}(x_k)v(x_k) \\ - \hat{u}(x_{k-1})v(x_{k-1}) + \hat{u}(x_k)v(x_k), \end{aligned} \tag{4.47b}$$

for all $\delta_v \in H^1(x_{k-1}, x_k)$. For each flux unknown $\hat{\sigma}(x_k)$, we have an optimal test function which spans across neighboring elements (x_{k-1}, x_k) and (x_k, x_{k+1}). For the first flux $\hat{\sigma}(0)$, the corresponding test function spans over the first element and, similarly, for the last flux $\hat{\sigma}(1)$, the corresponding test function spans over the last element only. Variational problem (4.47) leads to the following differential equations and boundary conditions for the optimal test functions.

$$\begin{cases} -\tau'' = \dfrac{1}{\epsilon}\sigma_k - u_k' \\ \tau'(x_k) = u_k(x_k) - \hat{u}(x_k) \\ -\tau'(x_{k-1}) + \tau(x_{k-1}) = -u_k(x_{k-1}) + \hat{u}(x_{k-1}), \end{cases}$$

and

$$\begin{cases} -v'' = -\sigma_k' + u_k' \\ v'(x_k) + v(x_k) = \sigma(x_k) - \hat{\sigma}(x_k) - u_k(x_k) + \hat{u}(x_k) \\ -v'(x_{k-1}) = -\sigma_k(x_{k-1}) + \hat{\sigma}(x_{k-1}) \\ \qquad\qquad + u_k(x_{k-1}) - \hat{u}(x_{k-1}). \end{cases}$$

This leads to the formulas

$$\tau' = \frac{1}{\epsilon}\int_x^{x_k} \sigma_k(s)\, ds + u_k(x) - \hat{u}(x_k) \tag{4.48a}$$

$$\tau(x_{k-1}) = \frac{1}{\epsilon}\int_{x_{k-1}}^{x_k} \sigma_k(s)\, ds + \hat{u}(x_{k-1}) - \hat{u}(x_k) \tag{4.48b}$$

$$\begin{aligned} \tau(x) = &\int_{x_{k-1}}^{x} (s - x_{k-1})\sigma_k(s)\, ds \\ &+ (x - x_{k-1})\int_x^{x_k} \sigma_k(s) \\ &+ \int_{x_{k-1}}^{x} u_k(s) + \frac{1}{\epsilon}\int_{x_{k-1}}^{x_k} \sigma_k(s)\, ds \\ &+ \hat{u}(x_{k-1}) - \hat{u}(x_k)(x - x_{k-1} + 1), \end{aligned} \tag{4.48c}$$

and

$$v'(x) = \sigma_k(x) - u_k(x) - \hat{\sigma}(x_{k-1}) + \hat{u}(x_{k-1}) \tag{4.49a}$$

$$v(x_k) = \hat{\sigma}(x_{k-1}) - \hat{\sigma}(x_k) - \hat{u}(x_{k-1}) + \hat{u}(x_k) \tag{4.49b}$$

$$\begin{aligned} v(x) = &\int_{x_{k-1}}^{x} \sigma_k(s)\, ds - \int_{x_{k-1}}^{x} u_k(s)\, ds \\ &+ \hat{\sigma}(x_{k-1})(1 - x + x_{k-1}) \\ &+ \ \hat{u}(x_{k-1})(x - x_{k-1} - 1) - \hat{\sigma}(x_k) + \hat{u}(x_k). \end{aligned} \tag{4.49c}$$

Formulas above allow us to construct optimal test functions for trial functions corresponding to L^2-variables $\sigma_k(x)$ and $u_k(x)$ as well as fluxes $\hat{\sigma}(x_k)$ and $\hat{u}(x_k)$. Notice that except for the test function corresponding to flux $\hat{\sigma}(x_k)$, all test functions are vector-valued, i.e. they have *both* τ and v components.

For instance, the test function corresponding to flux $\hat{u}(x_k) = 1$ is given by the formulas

$$\tau(x) = \begin{cases} x_{k-1} - x - 1 & \text{if } x \in (x_{k-1}, x_k) \\ 1 & \text{if } x \in (x_k, x_{k+1}) \end{cases}$$

$$v(x) = \begin{cases} 1 & \text{if } x \in (x_{k-1}, x_k) \\ x - x_{k-1} - 1 & \text{if } x \in (x_k, x_{k+1}). \end{cases}$$

The test function, illustrated in Fig. 4 shows a clear upwinding effect. For the h-method, i.e. with element size converging to zero, both τ and v converge to step functions. For $\sigma_k, u_k \in \mathcal{P}^p(x_{k-1}, x_k)$, formulas (4.48) and (4.49) imply that $\tau = \tau_k \in \mathcal{P}^{p+2}(x_{k-1}, x_k)$ and $v = v_k \in \mathcal{P}^{p+1}(x_{k-1}, x_k)$. Contrary to the pure convection problem, this does not allow for the construction of a simple Petrov-Galerkin method with the test functions being

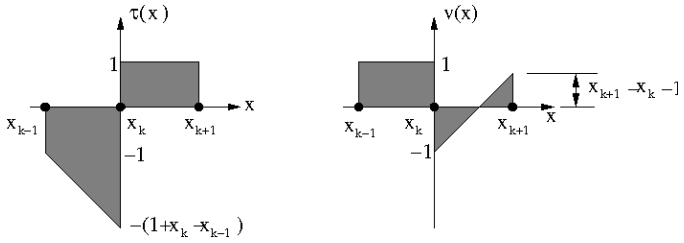


FIGURE 4. Optimal test function corresponding to flux $\hat{u}(x_k) = 1$.

polynomials of higher order. For elements of uniform order p, we would have a total of $2N(p+1)$ unknowns for σ_k, u_k, plus $2+2(N-1) = 2N$ fluxes, a total of $N(2p+4)$ unknowns. At the same time, the number of test d.o.f. would be $N(p+3+p+2) = N(2p+5)$. The numbers *do not* match each other.

Formulas (4.48) and (4.49) lead also to the formula for the energy norm,

$$\begin{aligned}
&\|(\boldsymbol{\sigma}, \boldsymbol{u}, \hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{u}})\|^2 \\
&= \sum_{k=1}^{N} \Bigg[\quad \left\| \frac{1}{\epsilon} \int_x^{x_k} \sigma_k(s)\, ds + u_k(x) - \hat{u}(x_k) \right\|^2 \\
&\qquad\qquad + \|\sigma_k(x) - u_k(x) - \hat{\sigma}(x_{k-1}) + \hat{u}(x_{k-1})\|^2 \Bigg] \\
&+ \sum_{k=1}^{N} \Bigg[\quad \left| \frac{1}{\epsilon} \int_{x_{k-1}}^{x_k} \sigma_k(s)\, ds + \hat{u}(x_{k-1}) - \hat{u}(x_k) \right|^2 \\
&\qquad\qquad + |\hat{\sigma}(x_{k-1}) - \hat{\sigma}(x_k) - \hat{u}(x_{k-1}) + \hat{u}(x_k)|^2 \Bigg]
\end{aligned}$$

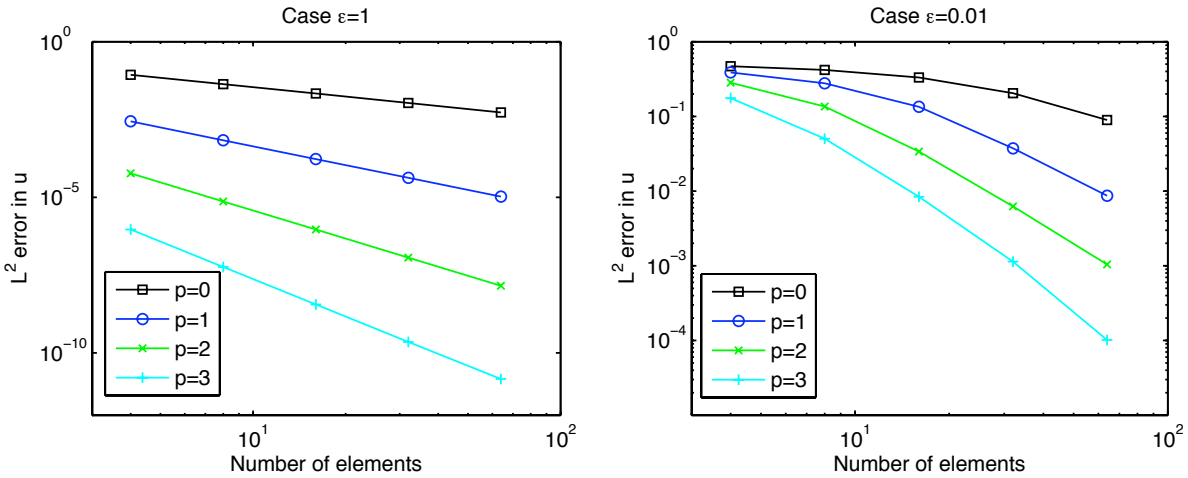
with $\hat{u}(0) = \hat{u}(1) = 0$, and $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)$, $\boldsymbol{u} = (u_1, \ldots, u_N)$, $\hat{\boldsymbol{\sigma}} = (\hat{\sigma}(0), \hat{\sigma}(x_1), \ldots, \hat{\sigma}(1))$, $\hat{\boldsymbol{u}} = (\hat{u}(x_1), \ldots, \hat{u}(x_{N-1}))$. The generalization of the one element analysis of §4.1 to this multielement case seems feasible, but we shall not attempt it in this paper.

4.3. **Numerical Experiments.**

Implementation details. We used equal order polynomials for σ and u. In our numerical implementation we chose to work with standard H^1-norms for the test space,

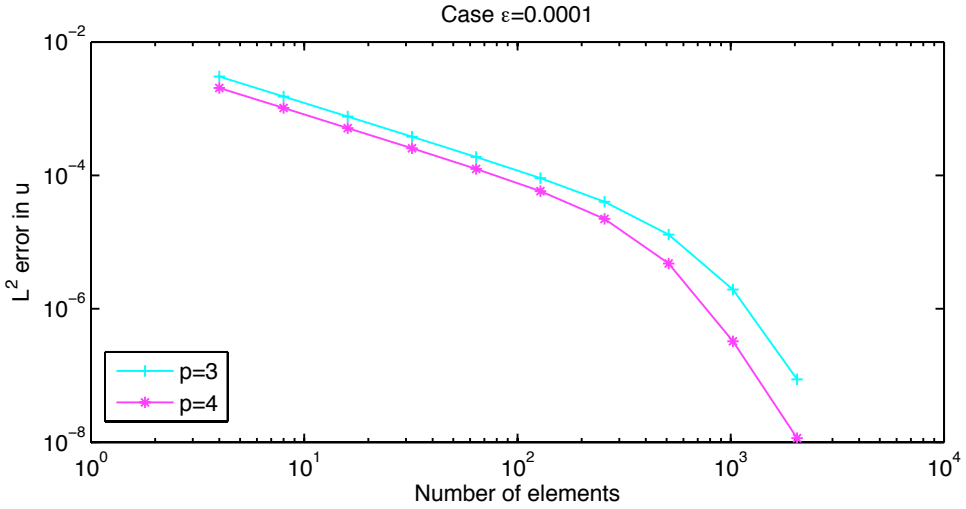
$$\begin{aligned}
\|(\boldsymbol{\tau}, \boldsymbol{v})\| &= \left(\sum_{k=1}^{N} \|\tau_k\|^2 + \|v_k\|^2 \right)^{\frac{1}{2}} \\
\|\tau_k\|^2 &= \int_{x_{k-1}}^{x_k} \left\{ |\tau_k'|^2 + |\tau|^2 \right\} \\
\|v_k\|^2 &= \int_{x_{k-1}}^{x_k} \left\{ |v_k'|^2 + |v_k|^2 \right\}.
\end{aligned} \tag{4.50}$$

Variational equations (4.47) for the optimal test functions have been solved approximately using polynomials of three degrees higher.



(a) Case $\epsilon = 1$, $p = 0, 1, 2, 3$.

(b) Case $\epsilon = 0.01$, $p = 0, 1, 2, 3$.



(c) Case $\epsilon = 10^{-4}$, $p = 3, 4$.

FIGURE 5. h-convergence of errors in u in L^2-norm for various p

All numerical experiments will be reported for the case with data $f(x) = 0$, $u_0 = 1$. The corresponding solution develops a boundary layer at $x = 1$,

$$\sigma(x) = -\frac{1}{1 - e^{-\frac{1}{\epsilon}}} e^{\frac{x-1}{\epsilon}},$$

$$u(x) = \frac{1}{1 - e^{-\frac{1}{\epsilon}}} \left(1 - e^{\frac{x-1}{\epsilon}}\right).$$

Convergence rates. We begin by reporting h-convergence rates for different orders of approximation for the L^2-norm of the solution and different values of diffusion constant ϵ. Fig. 5(a) presents h convergence for the solution in the diffusion regime $\epsilon = 1$. Fig. 5(b) presents h convergence for the solution in the convection-dominated regime, for $\epsilon = 0.01$. Finally, Fig. 5(c) presents h convergence for the solution in the convection-dominated regime, for $\epsilon = 10^{-4}$, $p = 3, 4$ and a greater number of elements. In all cases, the solution remains very stable in the preasymptotic regime.

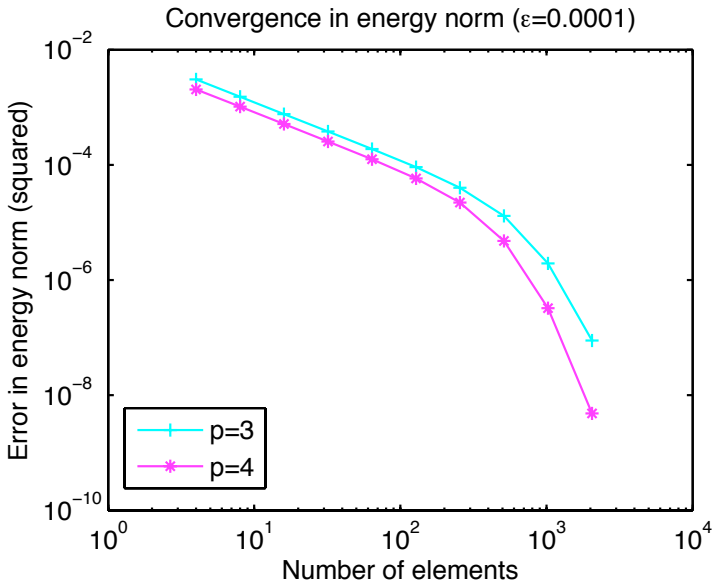


FIGURE 6. Case: $\epsilon = 10^{-4}, p = 3, 4$. h-convergence in the approximate energy norm.

Convergence in the energy norm. Fig. 6 presents convergence for $\epsilon = 10^{-4}$ and the energy norm. The norm is computed approximately using the same approximate test space (polynomials of order $p + 3$) as for the determination of the (approximate) optimal test functions. Consistently with the presented theory, the energy error decreases monotonically.

Dependence on accuracy of the optimal test functions. Fig. 7 presents convergence for $\epsilon = 10^{-4}$ and $p = 4$ with optimal test functions determined using polynomials of order $p + 1, p + 2, p + 3$ and $p + 4$. Whereas there is a significant difference between $p + 1$ and $p + 2$ cases, further increase in order produces visually indistinguishable results.

Examples of hp-adapted meshes. Finally, we present a few snapshots of interactively produced *hp*-adaptive meshes and the corresponding solutions for the case $\epsilon = 0.01$. Fig. 8(a) presents behavior of the approximate solution on a "bad" mesh, i.e. with refinements made in a wrong place, away from the boundary layer. Consistently with the analysis presented for the spectral case, the stress is approximated well, but the approximate velocity is clearly off by a constant.

Fig. 8(b) presents behavior of the approximate solution on a "good" mesh, obtained from the "bad" mesh by refining elements in the boundary layer. The exact and approximate solutions overlap each other and appear indistinguishably.

5. Convection-dominated diffusion in 2D

In this section, we numerically study the 2D convection-dominated diffusion. We discuss implementation details and present numerical experiments for a 2D model problem.

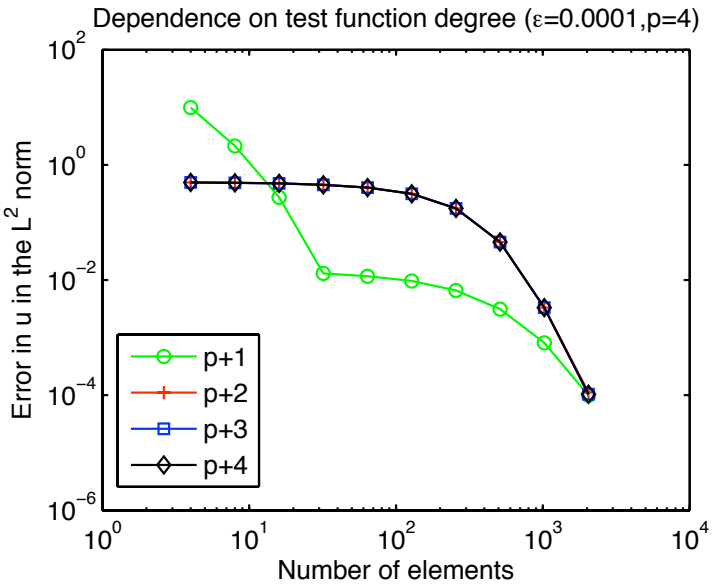


FIGURE 7. Case: $\epsilon = 10^{-4}, p = 4$. h-convergence in the L^2-norm for approximately optimal test functions determined with order $p+1, p+2, p+3, p+4$.

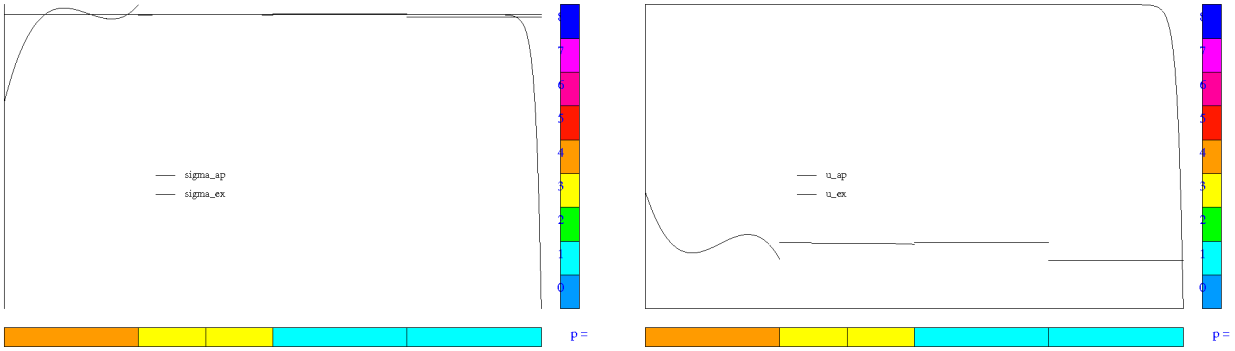
We consider the following model problem.

$$\begin{cases} u = u_0 & \text{on } \partial\Omega \\ \dfrac{1}{\epsilon}\sigma_1 - \dfrac{\partial u}{\partial x_1} = 0 & \text{in } \Omega \\ \dfrac{1}{\epsilon}\sigma_2 - \dfrac{\partial u}{\partial x_2} = 0 & \text{in } \Omega \\ -\text{div}\boldsymbol{\sigma} + \boldsymbol{\beta}\cdot\nabla u = f & \text{in } \Omega. \end{cases} \tag{5.51}$$

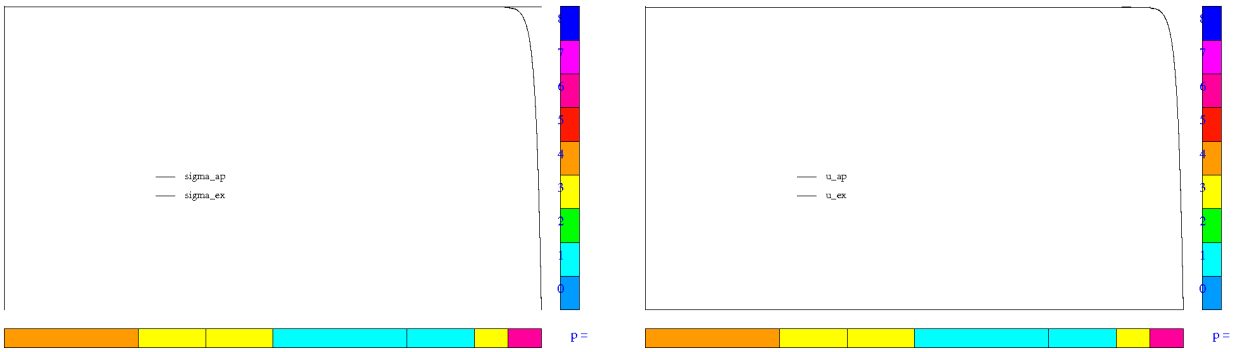
We assume that Ω has been partitioned into a FE mesh with elements K. In presented numerical examples, we will restrict outlived to 1-irregular triangular meshes only. Upon multiplying with test functions τ_1, τ_2 and v, integrating over an element K, and integrating by parts, we arrive at the following variational problem.

For each element K in the mesh, find functions $\boldsymbol{\sigma}_K$ in $\boldsymbol{L}^2(K)$, u_K in $L^2(K)$, and fluxes $\hat{u}_e \in U(e)$, $\hat{\sigma}_e \in L^2(e)$, satisfying

$$\begin{cases} \dfrac{1}{\epsilon}\displaystyle\int_K \boldsymbol{\sigma}_K\boldsymbol{\tau} + \int_K u_K \text{div}\boldsymbol{\tau} - \sum_{e\in\partial K\setminus\partial\Omega}\int_e \hat{u}_e\tau_n & = \ell_1(\tau) \\ \displaystyle\int_K \boldsymbol{\sigma}_K\nabla v - \int_K u_k\boldsymbol{\beta}\nabla v & \\ \quad + \displaystyle\sum_{e\in\partial K\setminus\partial\Omega}\int_e \beta_n\hat{u}_e v - \sum_{e\in\partial K}\int_e \hat{\sigma}_e \text{sgn}(\boldsymbol{n}_K) v & = \ell_2(v), \end{cases}$$



(a) Case $\epsilon = 10^{-2}$ and a "bad" hp mesh.



(b) Case $\epsilon = 10^{-2}$ and a "good" hp mesh.

FIGURE 8. Snapshots of mesh (bottom bar shows elements), the exact and approximate flux σ (left) and velocity u (right). Colors used in the bottom bar show polynomial degrees used in each element (legend in the side bar).

for all $\boldsymbol{\tau}$ in $\boldsymbol{H}(\mathrm{div}, K)$ and for all $v \in H^1(K)$, where

$$\ell_1(\tau) = \sum_{e \in \partial K \cap \partial\Omega} \int_e u_0 \tau_n$$

$$\ell_2(v) = \int_K f v - \sum_{e \in \partial K \cap \partial\Omega} \int_e \beta_n u_0 v$$

The unknowns include $\boldsymbol{\sigma}_K, u_K$, for each element K, fluxes $\hat{\sigma}_e$, for each edge e, and fluxes $\hat{u}$, for each *internal* edge e. We assume that each edge e has been assigned a particular normal $\boldsymbol{n}_e$. We define then

$$\mathrm{sgn}(\boldsymbol{n}_K) = \begin{cases} 1 & \text{if } \boldsymbol{n}_K = \boldsymbol{n}_e \\ -1 & \text{if } \boldsymbol{n}_K = -\boldsymbol{n}_e \end{cases} \tag{5.52}$$

Remark 5.1. We assume $f \in L^2(\Omega)$, and $u_0 \in H^{\frac{1}{2}}(\partial\Omega)$. The choice of energy space for fluxes u_e is far from trivial. Normal component τ_n lives in $H^{-\frac{1}{2}}(\partial K)$, which indicates (at least formally) that regularity $u_e \in L^2(e)$ may be insufficient. For the sake of this

paper, we shall assume that the problem is well posed, i.e. it has a unique solution in an appropriate functional setting. A systematic analysis of the well-posedness is postponed for a future work.

The variational problem for optimal test functions $\boldsymbol{\tau}, v$ is formulated as follows: Find $\boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div}, K)$ and $v \in H^1(K)$ satisfying

$$\int_K (\mathrm{div}\boldsymbol{\tau}\, \mathrm{div}\boldsymbol{\delta\tau} + \boldsymbol{\tau}\boldsymbol{\delta\tau}) = \frac{1}{\epsilon}\int_K \boldsymbol{\sigma}_K \boldsymbol{\delta\tau} \int_K u_K\, \mathrm{div}\boldsymbol{\delta\tau} - \sum_{e\in\partial K\setminus\partial\Omega}\int_e \hat{u}_e \delta_{\tau n}$$

for all $\boldsymbol{\delta\tau} \in \boldsymbol{H}(\mathrm{div}, K)$, and

$$\begin{aligned}\int_K (\nabla v \nabla\delta_v + v\delta_v) = &\int_K \boldsymbol{\sigma}_K \nabla\delta_v - \sum_{e\in\partial K}\int_e \hat{\sigma}_e \mathrm{sgn}(\boldsymbol{n}_K)\delta_v \\ &- \int_K u_k \boldsymbol{\beta}\nabla\delta v + \sum_{e\in\partial K-\partial\Omega}\int_e \beta_n \hat{u}_e \delta_v\end{aligned}$$

for all $\delta_v \in H^1(K)$. We have used equal polynomial order discretization for $\boldsymbol{\sigma}_K, u_K$ and one order higher approximation for fluxes. More precisely, if an edge e is shared by a number of elements K, the order for the fluxes is set to the maximum order of the adjacent elements. Also, if an edge has two small element neighbors on one side (we are using 1-irregular meshes only), the fluxes are approximated with *piecewise* polynomials rather then polynomials.

We use a very crude approximation for the optimal test functions. If $\boldsymbol{\sigma}_K, u_K \in \mathcal{P}^p(K)$, then $\boldsymbol{\tau}, v$ are approximated in $\mathcal{P}^{p+3}(K)$. It is worth mentioning that use of a lower order has resulted in a singular stiffness matrix (small pivots reported by a frontal solver). With piecewise polynomial fluxes, it would be more natural to divide the triangular element into four subelements and use $\boldsymbol{H}(\mathrm{div})$-conforming discretization for $\boldsymbol{\tau}$. This and a full *hp*-discretization of the $\boldsymbol{H}(\mathrm{div})$ problem are postponed for future studies.

5.1. **Verification of convergence rates.** To numerically verify the convergence rates, we have used an example presented by Egger and Schoeberl [20]. The convection-dominated diffusion problem is solved on a unit square with $\epsilon = 0.01, \boldsymbol{\beta} = (2, 1)$, homogeneous boundary conditions and source f corresponding to the exact solution,

$$u(x, y) = \left(x + \frac{e^{\frac{\beta_1 x}{\epsilon}} - 1}{1 - e^{\frac{\beta_1}{\epsilon}}}\right)\left(y + \frac{e^{\frac{\beta_2 y}{\epsilon}} - 1}{1 - e^{\frac{\beta_2}{\epsilon}}}\right) \tag{5.53}$$

The solution has a boundary layer along top and right edges. The problem was solved on a sequence of uniform triangular meshes (with positively sloped diagonals) with $4, 8, 16, 32$ elements on one side, and $p = 1, 2, 3, 4, 5$. Fig. 9 reports h-convergence rates for L^2-norm (for all three components of the solution, i.e. σ_1, σ_2, u) relative to L^2-norm. The source term f has been integrated using standard Gaussian integration.

Notice the difference in the h-convergence curves in the examples of Fig. 9 and Fig. 3. In the case of problems with boundary layers, it is typical to observe a speed-up of convergence rate as one resolves the layer, resulting in non-linear h-convergence curves as in Fig. 9 (cf. also the difference in the curves of Fig. 5(a) and 5(c)).

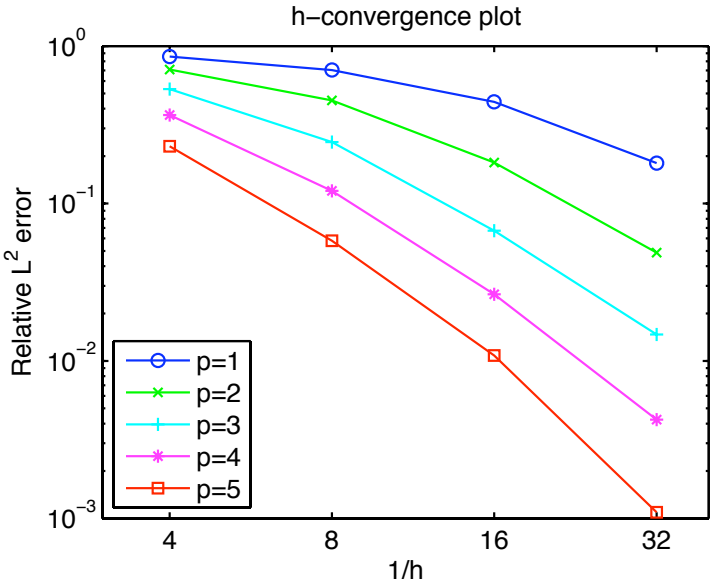


FIGURE 9. DPG method applied to the example of [20] under uniform h-refinement.

5.2. **Solution with a boundary layer.** We finish with an example illustrating the power of hp-adaptivity. We select the following data:

$$\begin{aligned} \Omega &= (0,1)\times(0,1) \\ f &= 0 \\ \epsilon &= 0.01 \\ \boldsymbol{\beta} &= (1,2) \\ u_0 &= \begin{cases} 1-x & \text{for } y=0 \\ 1-y & \text{for } x=0 \\ 0 & \text{for } x=1 \\ 0 & \text{for } y=1 \end{cases} \end{aligned} \tag{5.54}$$

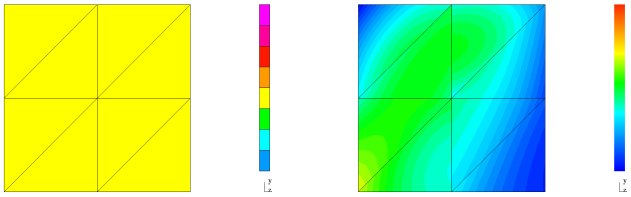
The solution develops a boundary layer along top and right edges. Fig. 10 present a sequence of hand-refined meshes and the corresponding solution (component u). The last solution is also shown in Fig. 11 with the mesh removed. Use of higher order elements allows for an accurate resolution of the boundary layer.

The main point of this illustration, however, is the fact that the solution remains very stable throughout the whole preasymptotic range.

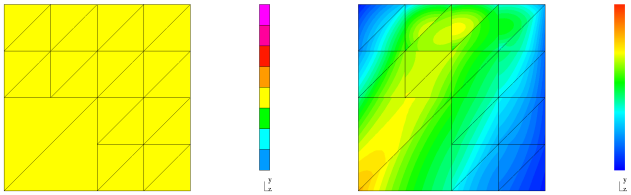
Finally, Fig. 12 presents evolution of the (approximate) energy norm (squared) vs. problem size (total number of d.o.f.). As predicted by the theory, the norm decreases monotonically.

6. Conclusions and Future Work

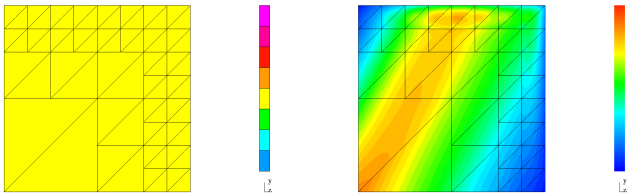
We have presented new *Discontinuous Petrov-Galerkin* (DPG) methods. Although we only discussed the advection and convection-diffusion problems, the ideas here can be



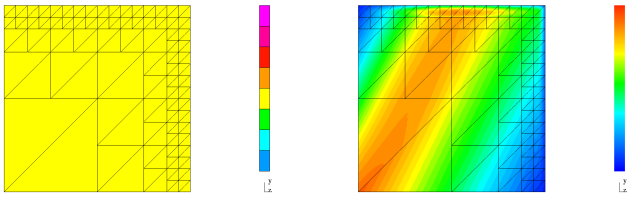
(a) First hp mesh and the corresponding solution u



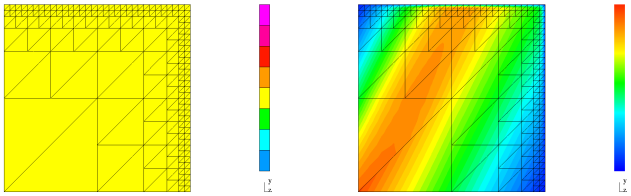
(b) Second hp mesh and the corresponding solution u



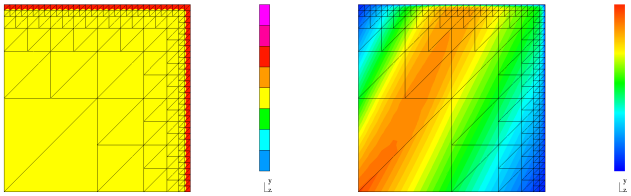
(c) Third hp mesh and the corresponding solution u



(d) Fourth hp mesh and the corresponding solution u



(e) Fifth hp mesh and the corresponding solution u



(f) Sixth hp mesh and the corresponding solution u

FIGURE 10. hp-refinements for a problem with a boundary layer

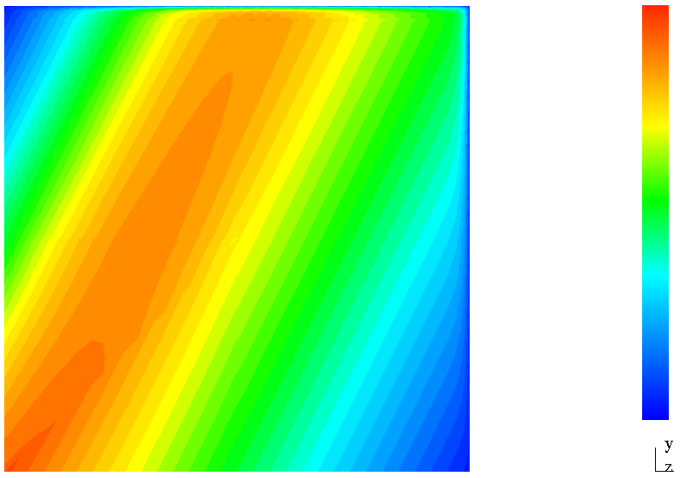


FIGURE 11. A plot of the finally computed solution u (on the sixth hp-mesh) for the problem with a boundary layer.

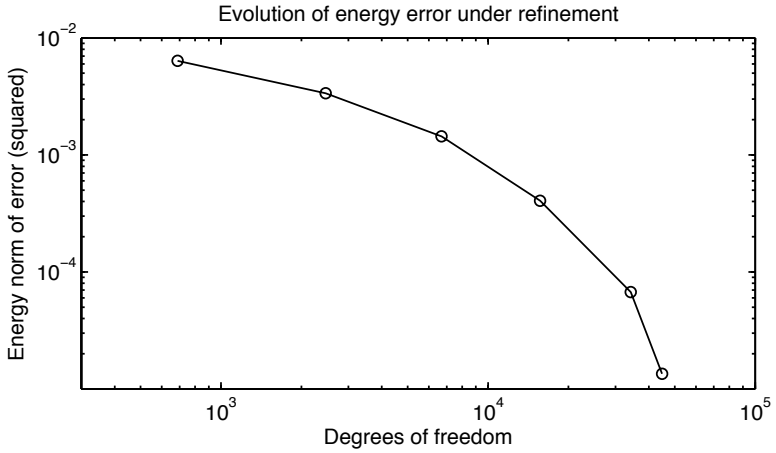


FIGURE 12. Decrease of error in the energy norm with hp-refinement for the problem with a boundary layer. The corresponding sequence of refined meshes is shown in Fig. 10.

used to develop new methods to solve general systems of PDE's in variational form. To summarize:

- The boundary value problem is formulated as a system of first order partial differential equations.
- A mesh dependent weak formulation is derived in the spirit of DG methods or "ultra-weak" formulations, i.e., all equations are treated in the sense of distributions with all derivatives passed to test functions.
- The problem is posed in L^2-energy setting with resulting fluxes treated as independent unknowns, in the spirit of hybridized methods. Variational equations are

formulated for each individual element in the mesh and these element problems are "connected" through fluxes shared by adjacent element.
- For each trial function, corresponding to either an interior or flux variable, we determine an optimal test function by solving auxiliary local variational problems. These variational problems for the test functions are derived by the choice of a test space norm. The optimal test functions then realize the supremum in the inf-sup condition.
- The norm on the test functions implies a special, "energy" norm for the solution space.
- With the *exact* optimal test functions, the method delivers *the best* approximation error in the energy norm (independent of the problem being solved).
- In practice, the local variational problems for the optimal test functions are solved *approximately* using an enriched space. In this paper, we have used simply polynomials of order $p + 3$.
- The resulting global stiffness matrix is *always* symmetric and positive definite. This enables use of iterative solvers.

To better understand the properties of the energy norm, we have theoretically analyzed one-dimensional convection and convection-dominated diffusion problems, specifically studying the the energy norm and characterizing it in terms of more standard norms. In particular, we have shown that the method delivers L^2 stability for the stress *uniformly* in diffusion constant ϵ. We have performed a number of both 1D and 2D numerical experiments for general *hp* meshes including 1-irregular grids. All our numerical results are in accordance with the theory we have so far. Rigorous error analyses in 2D for specific problems are postponed to future work.

The proposed method displays amazing stability properties on all meshes we experimented with. The proposed methodology is not restricted to standard element shapes (triangles, quadrilaterals in 2D, tetrahedra, hexahedra, prisms, and pyramids in 3D) and can be applied, in particular, to general polygons or polyhedra. The formulation enables the use of *hp*-adaptivity driven by the energy norm. Our current research focuses on two topics. We continue studying the convection-dominated diffusion with variable advection and the interdependence of different norms. Simultaneously, we have started a study on steady-state laminar compressible Navier-Stokes equations, the very problem that originated this research.

Appendix A. Proof of Theorem 3.1

Proof of Theorem 3.1. Proof of item 1: To calculate the energy norm, we can use Proposition 2.2. Given any u and q_i's, let the optimal test functions be denoted by t_u and t_q, i.e, $t_u(x)$ takes the value $\int_x^{x_i} u(s)$ if $x_{i-1} < x < x_i$, and $t_q(x)$ is given by $\sum_{i=1}^n q_i v_i$ where v_i is as in (3.28). Then by Proposition 2.2,

$$\|(u, q_1, \ldots, q_n)\|_E^2 = b(\,(u, q_1, \ldots, q_n), t_u + t_q).$$

Simple calculations show that

$$b(\,(u, q_1, \ldots, q_n), t_u) = \sum_{i=1}^n \int_{x_{i-1}}^{x_i} u^2 \,-\, q_{i-1} \int_{x_{i-1}}^{x_i} u,$$

and

$$b(\,(u,q_1,\ldots,q_n),t_q) = \sum_{i=1}^{n} \frac{(q_i-q_{i-1})^2}{\alpha_i} + (x_i-x_{i-1})q_{i-1}^2 - \int_{x_{i-1}}^{x_i} u\, q_{i-1},$$

and by adding, we obtain (3.29).

Proof of item 2: To prove the inf-sup condition, we first prove a discrete Poincaré inequality. Since

$$q_{\ell-1} = \sum_{j=1}^{\ell-1}(q_j - q_{j-1}),$$

by Cauchy-Schwarz inequality, we have

$$q_{\ell-1}^2 \le \Big(\sum_{j=1}^{\ell-1}\alpha_j\Big)\Big(\sum_{j=1}^{\ell-1}\frac{|q_j-q_{j-1}|^2}{\alpha_j}\Big).$$

Summing over ℓ and and increasing the last sum, we obtain the discrete Poincaré inequality

$$\|q\|_h^2 \;\le\; \kappa \sum_{j=1}^{n}\frac{|q_j-q_{j-1}|^2}{\alpha_j}. \tag{A.55}$$

Moreover,

$$\begin{aligned}\|u\|^2 &= \sum_{i=1}^{n}\int_{x_{i-1}}^{x_i} u^2 \le 2\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i} |u-q_{i-1}|^2 + |q_{i-1}|^2\\ &= 2\sum_{i=1}^{n} q_{i-1}^2(x_i-x_{i-1}) + \int_{x_{i-1}}^{x_i}|u-q_{i-1}|^2\\ &= 2\Big(\|q\|_h^2 + \sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}|u-q_{i-1}|^2\Big).\end{aligned}$$

Thus,

$$\|u\|^2 + \|q\|_h^2 \le 3\|q\|_h^2 + 2\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i}|u-q_{i-1}|^2,$$

from which (3.30) follows after combining with (A.55).

Proof of item 3: We proceed considering the last element (x_{n-1},x_n) first: On this element, the span of v_n and v^w for all $w\in\mathcal{P}^p(x_{n-1},x_n)$ equals $\mathcal{P}^{p+1}(x_{n-1},x_n)$. Using this and proceeding to the next element on the left, we can inductively prove the statement.

Proof of item 4: From item 3 it follows that the test space is independent of α_i. Since the bilinear form and the trial space are also independent of α_i, the solution is independent of α_i.

Proof of item 5: Applying Theorem 2.2, the error between the exact solution $(u,q_1,\ldots,q_n)$ and the discrete solution $(u_h,q_{h,1},\ldots,q_{h,n})$ satisfies

$$\begin{aligned}&\|(u,q_1,\ldots,q_n)-(u_h,q_{h,1},\ldots,q_{h,n})\|_E^2\\ &\qquad\le \|(u,q_1,\ldots,q_n)-(w_h,\phi_{h,1},\ldots,\phi_{h,n})\|_E^2\end{aligned}$$

for any $(w_h, \phi_{h,1}, \dots, \phi_{h,n}) \in U_h$. Since $\phi_{h,i} \in \mathbb{R}$ can be chosen to coincide with the exact fluxes $q_i \in \mathbb{R}$,

$$\begin{aligned} \|(u, q_1, \dots, q_n) - (u_h, q_{h,1}, \dots, q_{h,n})\|_E^2 \\ \leq \inf_{w_h} \|u - w_h\|^2. \end{aligned} \tag{A.56}$$

By item 1, this implies that

$$\sum_{i=1}^{n} \frac{\left|(q_i - q_{h,i}) - (q_{i-1} - q_{h,i-1})\right|^2}{\alpha_i} \leq \gamma \inf_{w_h} \|u - w_h\|^2.$$

Choosing all $\alpha_i = \varepsilon$ for an arbitrarily small ε, and multiplying the above inequality by ε, we find that the right hand side is $O(\varepsilon)$, while by item 4, the left hand side is independent of ε. Hence the left hand side must vanish, which by the Poincaré inequality (A.55) implies that $q_{h,i} = q_i$.

Proof of item 6: Returning to (A.56), we find that item 5 implies

$$\|u - u_h\| \leq \inf_{w_h} \|u - w_h\|$$

so u_h must coincide with the L^2 projection of u. □

Some of the statements in Theorem 3.1 can be proved more easily by direct arguments. Nonetheless, our purpose in the above proof is to illustrate the optimal test function techniques in perhaps the simplest possible example. Note that if α_i is chosen to be $x_i - x_{i-1}$, then the inf-sup constant in (3.30) can be chosen independent of the meshes $\{x_i\}$.

Appendix B. Proof of Theorem 4.1

Proof of Theorem 4.1. From the definition of the energy norm (4.45), it is clear that triangle inequality implies that the numbers

$$\left\| \sigma + \frac{1}{\epsilon} \int_x^1 \sigma - \hat{\sigma}(0) \right\|, \quad \left| \frac{1}{\epsilon} \int_0^1 \sigma \right|$$

are controlled by the energy norm times a constant independent of ϵ. Let

$$g(x) := \sigma(x) + \frac{1}{\epsilon} \int_x^1 \sigma - \hat{\sigma}(0), \quad A := \frac{1}{\epsilon} \int_0^1 \sigma.$$

We then solve for σ in terms of g and A as follows. From the definition of g,

$$\sigma(x) = e^{\frac{x-1}{\epsilon}} \hat{\sigma}(0) + g(x) - \frac{1}{\epsilon} \int_x^1 e^{\frac{x-s}{\epsilon}} g(s)\, ds. \tag{B.57}$$

Integrating both sides from 0 to 1, and dividing by ϵ, we obtain equation for $\hat{\sigma}(0)$,

$$A = (1 - e^{-\frac{1}{\epsilon}})\hat{\sigma}(0) + \frac{1}{\epsilon} \int_0^1 e^{-\frac{s}{\epsilon}} g(s)\, ds. \tag{B.58}$$

Solving for $\hat{\sigma}(0)$ and substituting into formula (B.57), we get the final formula for σ in terms of function g and constant A:

$$\begin{aligned}\sigma(x) = \frac{A}{(1-e^{-\frac{1}{\epsilon}})} e^{\frac{x-1}{\epsilon}} - \frac{e^{\frac{x-1}{\epsilon}}}{\epsilon(1-e^{-\frac{1}{\epsilon}})} \int_0^1 e^{-\frac{s}{\epsilon}} g(s)\, ds \\ + g(x) - \frac{1}{\epsilon}\int_x^1 e^{\frac{x-s}{\epsilon}} g(s)\, ds.\end{aligned} \tag{B.59}$$

With the help of this formula, we can now estimate the L^2-norm of σ in terms of the energy norm. Each of the four terms in (B.59) can be bounded with $\|g\|$ and A with constants *independent of* ϵ. The first three estimates are straightforward and we will leave them to the reader. The last one is tricky. For this, we first integrate by parts:

$$\begin{aligned}&\int_0^1 \left| \int_x^1 e^{\frac{x-s}{\epsilon}} g(s)\, ds \right|^2 dx = \int_0^1 e^{\frac{2x}{\epsilon}} \left(\int_x^1 e^{-\frac{s}{\epsilon}} g(s)\, ds \right)^2 dx \\ &= \left[\frac{\epsilon}{2} e^{\frac{2x}{\epsilon}} \left(\int_x^1 e^{-\frac{s}{\epsilon}} g(s)\, ds \right)^2 \right] \Bigg|_0^1 \\ &\quad + \frac{\epsilon}{2} \int_0^1 e^{\frac{2x}{\epsilon}}\, 2 \int_x^1 e^{-\frac{s}{\epsilon}} g(s)\, ds\, e^{-\frac{x}{\epsilon}} g(x)\, dx \\ &= -\frac{\epsilon}{2} \left(\int_0^1 e^{-\frac{s}{\epsilon}} g(s)\, ds \right)^2 + \epsilon \int_0^1 e^{\frac{x}{\epsilon}} g(x) \int_x^1 e^{-\frac{s}{\epsilon}} g(s)\, ds\, dx.\end{aligned}$$

Dropping the first negative term and using the Cauchy-Schwarz inequality, we obtain,

$$\begin{aligned}\int_0^1 \left| \int_x^1 e^{\frac{x-s}{\epsilon}} g(s)\, ds \right|^2 dx &\le \epsilon \int_0^1 g(x) \int_x^1 e^{\frac{x-s}{\epsilon}} g(s)\, ds\, dx \\ &\le \left[\int_0^1 g^2(x)\, dx \right]^{\frac{1}{2}} \left[\int_0^1 \left(\int_x^1 e^{\frac{x-s}{\epsilon}} g(s)\, ds \right)^2 dx \right]^{\frac{1}{2}}.\end{aligned}$$

Subdividing both sides by the last term on the right-hand side, we obtain,

$$\left[\int_0^1 \left(\int_x^1 e^{\frac{x-s}{\epsilon}} g(s)\, ds \right)^2 dx \right]^{\frac{1}{2}} \le \epsilon \|g\|$$

This leads to the final estimate of the last term,

$$\frac{1}{\epsilon^2} \int_0^1 | \int_x^1 e^{\frac{x-s}{\epsilon}} g(s)\, ds|^2\, dx \le \frac{1}{\epsilon^2} \epsilon^2 \|g\|^2 = \|g\|^2. \tag{B.60}$$

Concluding, we have

$$\|\sigma\| \le C(\|g\| + A) \tag{B.61}$$

where constant C is *independent* of ϵ, in other words,

$$\|\sigma\| \le C' \|(\sigma, \hat{\sigma}(0), \hat{\sigma}(1), u)\|_E$$

holds with a C' independent of ϵ.

To finish, by (B.58),

$$|\hat{\sigma}(0)| \leq \frac{1}{1-e^{-\frac{1}{\epsilon}}} A + \epsilon^{-\frac{1}{2}} \left[\frac{1-e^{-\frac{2}{\epsilon}}}{2}\right]^{\frac{1}{2}} \|g\|. \tag{B.62}$$

Estimate (B.62) and the energy norm definition (4.45) imply the final estimates of the theorem. □

Remark B.1. A straightforward use of the Cauchy-Schwarz inequality on the last term in (B.59) gives only a suboptimal estimate:

$$\begin{aligned}\int_0^1 \frac{1}{\epsilon^2} \left| \int_x^1 e^{\frac{x-s}{\epsilon}} g(s)\, ds \right|^2 dx &\leq \frac{1}{\epsilon^2} \int_0^1 \int_x^1 e^{\frac{2(x-s)}{\epsilon}}\, ds \int_x^1 |g|^2\, ds\, dx \\ &\leq \frac{1}{2\epsilon} \int_0^1 (1 - e^{\frac{2(x-1)}{\epsilon}})\, dx \int_0^1 |g|^2\, dx \\ &\leq \frac{1}{2\epsilon} \int_0^1 |g|^2\, dx.\end{aligned}$$

Remark B.2. One would like to establish an estimate analogous to (B.61) for the primal variable u as well. However, for this variable, we are only able to prove a suboptimal estimate. This is because equation (B.58) seems to yield only a weaker bound for constant $\hat{\sigma}(0)$, namely (B.62).

Remark B.3. Notice that the uniform estimate for $\|\sigma\|$ and (4.45) imply a uniform estimate for

$$\|u - \hat{\sigma}(0)\| \leq C \|(\sigma, \hat{\sigma}(0), \hat{\sigma}(1), u)\|_E$$

Any degeneracy that may occur in the L^2-stability for u must be global in nature due to the fact that we only have weak control of the constant $|\hat{\sigma}(0)|$.

References

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

[2] I. Babuška and A. K. Aziz, *Survey lectures on the mathematical foundations of the finite element method*, in The mathematical foundations of the finite element method with applications to partial differential equations (Proc. Sympos., Univ. Maryland, Baltimore, Md., 1972), Academic Press, New York, 1972, pp. 1–359. With the collaboration of G. Fix and R. B. Kellogg.

[3] I. Babuška, G. Caloz, and J. E. Osborn, *Special finite element methods for a class of second order elliptic problems with rough coefficients*, SIAM J. Numer. Anal., 31 (1994), pp. 945–981.

[4] 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.

[5] 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.

[6] ——, *A multiscale formulation of the discontinuous Petrov-Galerkin method for advective-diffusive problems*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 2819–2838.

[7] J. H. Bramble, R. D. Lazarov, and J. E. Pasciak, *A least-squares approach based on a discrete minus one inner product for first order systems*, Math. Comp., 66 (1997), pp. 935–955.

[8] J. H. Bramble and J. E. Pasciak, *A new approximation technique for div-curl systems*, Math. Comp., 73 (2004), pp. 1739–1762 (electronic).

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

[10] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations. I*, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.

[11] 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).

[12] P. Causin, R. Sacco, and C. L. Bottasso, *Flux-upwind stabilization of the discontinuous Petrov-Galerkin formulation with Lagrange multipliers for advection-diffusion problems*, M2AN Math. Model. Numer. Anal., 39 (2005), pp. 1087–1114.

[13] 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.

[14] 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.

[15] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463 (electronic).

[16] L. Demkowicz, *Babuska $\Longleftrightarrow$ Brezzi*, Tech. Rep. 06-08, ICES, The University of Texas at Austin, April 2006.

[17] L. Demkowicz and J. Gopalakrishnan, *A class of discontinuous Petrov-Galerkin methods. part i: The transport equation*, Submitted, (2009). Available online as ICES Report 09-12.

[18] L. Demkowicz and J. T. Oden, *An adaptive characteristic Petrov-Galerkin finite element method for convection-dominated linear and nonlinear parabolic problems in one space variable*, J. Comput. Phys., 67 (1986), pp. 188–213.

[19] ——, *An adaptive characteristic Petrov-Galerkin finite element method for convection-dominated linear and nonlinear parabolic problems in two space variables*, Comput. Methods Appl. Mech. Engrg., 55 (1986), pp. 63–87.

[20] 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).

[21] J. Gopalakrishnan and G. Kanschat, *A multilevel discontinuous Galerkin method*, Numer. Math., 95 (2003), pp. 527–550.

[22] 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).

[23] ——, *Discontinuous hp-finite element methods for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 39 (2002), pp. 2133–2163 (electronic).

[24] 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.

[25] C. Johnson, U. Nävert, and J. Pitkäranta, *Finite element methods for linear hyperbolic problems*, Comput. Methods Appl. Mech. Engrg., 45 (1984), pp. 285–312.

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

[27] T. Kato, *Estimation of iterated matrices, with application to the von Neumann condition*, Numer. Math., 2 (1960), pp. 22–29.

[28] 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.

[29] S. G. Mikhlin, *Variational methods in Mathematical Physics*, Pergamon Press, Oxford, 1964.

[30] A. R. Mitchell and D. F. Griffiths, *The finite difference method in partial differential equations*, John Wiley & Sons Ltd., Chichester, 1980. A Wiley-Interscience Publication.

[31] 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.

[32] 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.
[33] G. I. Petrov, *Application of the method of Galerkin to a problem involving the stationary flow of a viscous fluid*, Prikl. Matem. i Mekh., 4 (1940).
[34] 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.
[35] T. Russell and M. Celia, *An overview of research on Eulerian-Lagrangian localized adjoint methods (ELLAM)*, Advances in Water Resources, 25 (2002), pp. 1215–1231.
[36] W. Wendland, *On Galerkin methods*, in Proceedings of ICIAM 95, vol. 76 (Supplement 2), Muhlenstrasse 33-34, D-13187 Berlin, Germany, 1996, Akademie-Verlag GMBH, pp. 257–260.
[37] J. Xu and L. Zikatanov, *Some observations on Babuška and Brezzi theories*, Numer. Math., 94 (2003), pp. 195–202.

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