ANALYSIS OF THE DPG METHOD FOR THE POISSON EQUATION

By

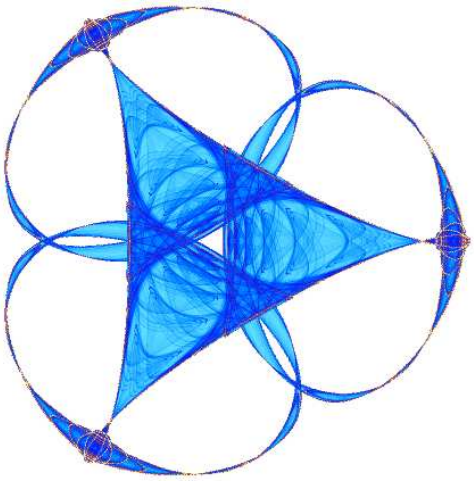
L. Demkowicz

and

J. Gopalakrishnan

IMA Preprint Series # 2340

(October 2010)



INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS

UNIVERSITY OF MINNESOTA

400 Lind Hall

207 Church Street S.E.

Minneapolis, Minnesota 55455–0436

Phone: 612-624-6066 Fax: 612-626-7370

URL: http://www.ima.umn.edu

ANALYSIS OF THE DPG METHOD FOR THE POISSON EQUATION

L. DEMKOWICZ AND J. GOPALAKRISHNAN

Abstract. We give an error analysis of the recently developed DPG method applied to solve the Poisson equation and a convection-diffusion problem. We prove that the method is quasioptimal. Error estimates in terms of both the mesh size h and the polynomial degree p (for various element shapes) can be derived from our results. Results of extensive numerical experiments are also presented.

1. Introduction

We present an analysis of a discontinuous Petrov-Galerkin (DPG) method for a simple model problem involving the Poisson equation. Although DPG methods for the Poisson equation have been in existence for long (e.g., [7]), we have in mind the new class of DPG methods constructed by following the framework we developed in [13, 14, 15, 23]. These papers explored how one can achieve the best possible stability (or close to it) by designing test spaces suitably.

Our first paper [13] gave an analysis of a DPG discretization of a simple transport problem in two dimensions. In this simple case we had the benefit of being able to tweak a hand-calculated test space that is "optimal" in terms of stability. With a view towards applying the DPG methodology to more complicated problems in an automatic fashion, we modified the DPG framework in [14], bringing it closer to the least-square Galerkin methods such as in [3]. What distinguishes our methodology from other least square methods is the possibility to *locally* compute a test space that is close to optimal. Our remaining papers [14, 15, 23] gave numerical evidence of the extraordinary stability of the resulting methods when applied to various problems. However, the theoretical analysis in these papers, unlike [13], was restricted to problems in one space dimension. The purpose of this paper is to give a few new techniques to analyze DPG methods for problems in higher space dimensions.

Although the subject here is the Poisson equation, we are not advocating that one should use the DPG method for the simple Poisson equation (for which many competitive methods exist). Indeed, the real potential of the DPG methodology is clearly evident only in high order simulation of more complex problems. What we aim for in this work is more limited, but clear-cut: We want theoretical techniques that permit a fairly complete analysis of the method for a simple elliptic multidimensional model problem. Since the DPG method is fundamentally different from other standard methods, few tools were available for its theoretical understanding. One of the techniques we introduce in this paper is a decomposition of a discontinuous function into a (conforming) solution of a mixed method and a (discontinuous) piecewise harmonic function. This leads to an

Key words and phrases. DPG method, discontinuous Galerkin, discontinuous Petrov-Galerkin, Helmholtz decomposition, adaptive, hp, finite element method, convection-diffusion.

This work was supported in part by the Department of Energy [National Nuclear Security Administration] under Award Number DE-FC52-08NA28615, the NSF under DMS-1014817, and the IMA.

inf-sup condition and facilitates the error analysis of the DPG method for the Poisson equation. Such techniques may find applicability beyond the Poisson example. Another unexpected example of the impact of the present theoretical study is that it indicates that for optimal convergence rates, while the numerical fluxes may be approximated using the same polynomial degrees as the interior variables, the DPG numerical traces need higher degree polynomials. Such insights can be valuable when applying the DPG methodology to more complex applications.

DPG methods fall into the general category of discontinuous Galerkin (DG) methods, so let us review DG methods for elliptic problems. The literature in this area is vast, so we will be brief and cite only works necessary to put this paper in broad perspective. One of the first DG methods, called the interior penalty (IP) method [2], used a penalization parameter. An inconvenient feature of this method is that for stability it required the penalization parameter to be "sufficiently" large (practically unknown) number. This was remedied by the LDG methods [6, 9] which also enjoyed the additional property that fluxes can be eliminated locally. The IP and LDG methods, and indeed all the "older" DG methods for the Poisson equation, have been reviewed thoroughly in [1]. They showed that almost all the DG methods in existence at the time of their writing could be recast into a system of two equations with *specific prescriptions* of the so-called "numerical trace" and the "numerical flux" on element interfaces (see (14) below for more on the terminology).

The further developments in this area yielding "newer" DG methods, not covered by [1], can be understood from various angles. To offer one perspective, many researchers considered the specific prescriptions of numerical traces and fluxes as adhoc and difficult to generalize for complex problems. Shouldn't a good method find the right numerical trace automatically? A partial answer was provided by the recently developed hybridized DG (HDG) method [8]. It lets the numerical trace be an unknown to be determined automatically by the method, and yet maintains the local elimination and flexible stabilization properties that endeared DG methods. Nonetheless, although the HDG method automatically finds the "right" numerical *trace*, its numerical *flux* must again be prescribed. In this perspective, the next natural question is whether there are stable DG methods which let *both* the numerical flux and the numerical trace to be unknowns (so that none of these needs to prescribed adhoc). The DPG method analyzed in this paper answers this question in the affirmative.

We begin by recalling a salient result for abstract DPG methods in Section 2. The development of the bilinear and linear forms that constitute the DPG method appears in Section 3. Next, in Section 4, we give an error analysis of the method. In Section 5 we point out how the new techniques of analysis can be extended to a more general second order elliptic problem. Finally, numerical experiments are presented in Section 6.

2. The abstract DPG method

In this section we summarize the DPG framework developed in [13, 14, 15, 23] and state an abstract result which we will use for error analysis.

Let U (the "trial" space) and V (the "test" space) be vector spaces over $\mathbb{R}$ and $b(\cdot,\cdot) : U \times V \mapsto \mathbb{R}$ be a bilinear form. Let U be a reflexive Banach space under the norm $\|\cdot\|_U$. We assume that V is a Hilbert space under an inner product $(\cdot,\cdot)_V$ with a corresponding

norm $\|\cdot\|_V$. We assume that

$$\|v\|_{\text{opt},V} = \sup_{u\in U} \frac{b(u,v)}{\|u\|_U} \tag{1}$$

is a norm on V. This is called the *optimal test space norm* for reasons explained in [23]. The norms $\|v\|_{\text{opt},V}$ and $\|v\|_V$ are not equal in general.

The variational problem we wish to approximate is as follows.

$$\begin{cases} \text{Find } u \in U \text{ such that} \\ b(u,v) = l(v), \quad \forall v \in V. \end{cases} \tag{2}$$

Here $l(\cdot)$ is a given real-valued continuous linear functional on V. The DPG approximation of $u \in U$ is denoted by u_h. It lies in U_h, a subspace of U. We define the *trial-to-test operator* $T : U \mapsto V$ by

$$(Tu, v)_V = b(u,v), \qquad \forall v \in V. \tag{3}$$

Let $V_h = T(U_h)$. The DPG approximation $u_h \in U_h$ satisfies

$$b(u_h, v_h) = l(v_h) \qquad \forall v_h \in V_h. \tag{4}$$

This is a Petrov-Galerkin type formulation as U_h and V_h are not generally identical. Nonetheless, the resulting stiffness matrix is symmetric and positive definite. This and other interesting properties are discussed in [14]. A basic convergence result for the abstract method is proved in [23, Theorem 2.1]. Let us restate it here in a form convenient for the current application.

Theorem 2.1. *Suppose u and u_h satisfy* (2) *and* (4)*, resp. Assume that*

$$\{w \in U : \ b(w,v) = 0, \ \forall v \in V\} = \{0\} \tag{5}$$

and that there are positive constants C_1, C_2 such that

$$C_1\|v\|_V \le \|v\|_{\text{opt},V} \le C_2\|v\|_V, \qquad \forall v \in V. \tag{6}$$

Then

$$\|u - u_h\|_U \le \frac{C_2}{C_1} \inf_{w_h \in U_h} \|u - w_h\|_U.$$

In the remainder, we wish to apply this theorem to the particular case of a DPG method for the Poisson equation. To this end, we will develop an *ultra-weak formulation* for the Poisson equation and verify the assumptions of Theorem 2.1.

3. Application to the Poisson equation

In this section, we derive a DPG method for the Poisson equation and set up a suitable functional framework. Let Ω be a bounded simply connected open subset of $\mathbb{R}^N$ with connected Lipschitz boundary, where $N = 2$ or 3. We assume that Ω_h is a disjoint partitioning of Ω into open "elements" K, i.e., $\cup\{\bar{K} : \ K \in \Omega_h\} = \bar{\Omega}$. At this point, we need not assume that elements are of any particular shape, but so as to apply trace theorems later, we will assume their boundaries ∂K are Lipschitz.

The boundary value problem targeted for approximation is

$$-\vec{\nabla}\cdot(\alpha\vec{\nabla}u) = f \qquad \text{on } \Omega \tag{7a}$$

$$u = 0 \qquad \text{on } \partial\Omega. \tag{7b}$$

Here $\alpha(\vec{x})$ is a given measurable coefficient function satisfying

$$0 < \alpha_0 \leq \alpha(\vec{x}) \leq \alpha_1, \qquad \forall \vec{x} \in \Omega. \tag{8}$$

The load f is in $L^2(\Omega)$ (although it will be clear later that this can be relaxed).

3.1. **The ultra-weak formulation.** We will now develop a variational formulation of (7) where u is only required to be in $L^2(\Omega)$ and the "flux" of the solution, namely $\vec{\sigma} = -\alpha\vec{\nabla}u$, is also only required to be in $L^2(\Omega)^N$. Hence the name "ultra"-weak formulation. (This name was used in the same spirit in [20] for a different method.)

To motivate the derivation of this ultra-weak formulation, let us temporarily assume that the solution and flux are smooth enough to allow integration by parts, i.e., we reformulate (7) as the first order system

$$\alpha^{-1}\vec{\sigma} + \vec{\nabla}u = 0, \tag{9a}$$

$$\vec{\nabla} \cdot \vec{\sigma} = 0, \tag{9b}$$

and integrate these equations by parts on one element K to get

$$(\alpha^{-1}\vec{\sigma}, \vec{\tau})_K - (u, \vec{\nabla} \cdot \vec{\tau})_K + \langle u, \vec{\tau} \cdot \vec{n} \rangle_{\partial K} = 0, \qquad \forall \vec{\tau} \in H(\mathrm{div}, K), \tag{10a}$$

$$-(\vec{\sigma}, \vec{\nabla}v)_K + \langle v, \vec{\sigma} \cdot \vec{n} \rangle_{\partial K} = (f, v)_K, \qquad \forall v \in H^1(K). \tag{10b}$$

Here $\vec{n}$ denotes the outward unit normal on ∂K. The outward unit normal on any other domain will also be generically denoted by $\vec{n}$ (the underlying domain will be clear from the context). The notations $(\cdot,\cdot)_D$ and $\langle \cdot, \cdot \rangle_{\partial D}$ denote the $L^2(D)$ and $L^2(\partial D)$ inner products, resp., on any domain D. Above and later, we use the standard notations for Sobolev spaces, such as $H^1(D)$ and $H(\mathrm{div}, D)$. Additionally, note that the completion of compactly supported smooth functions in the $H^1(D)$ and $H(\mathrm{div}, D)$-norms will be denoted by $H^1_0(D)$ and $H_0(\mathrm{div}, D)$, resp.

We now replace the terms $\langle u, \vec{\tau} \cdot \vec{n} \rangle_{\partial K}$ and $\langle v, \vec{\sigma} \cdot \vec{n} \rangle_{\partial K}$ by $\langle \hat{u}, \vec{\tau} \cdot \vec{n} \rangle_{1/2,\partial K}$ and $\langle v, \hat{\sigma}_n \rangle_{1/2,\partial K}$, resp., where $\hat{u}$ and $\hat{\sigma}_n$ are new unknowns, and $\langle \cdot, \ell \rangle_{1/2,\partial K}$ denotes the action of a functional ℓ in $H^{-1/2}(\partial K)$. This motivates the following ultra-weak formulation.

$$\begin{cases} \text{Find } (\vec{\sigma}, u, \hat{u}, \hat{\sigma}_n) \in U \text{ such that} \\ \qquad b(\ (\vec{\sigma}, u, \hat{u}, \hat{\sigma}_n),\ (\vec{\tau}, v)\ ) = l(\vec{\tau}, v), \qquad \forall (\vec{\tau}, v) \in V, \end{cases} \tag{11}$$

where the spaces are defined by

$$U = L^2(\Omega) \times L^2(\Omega) \times H^{1/2}_0(\partial\Omega_h) \times H^{-1/2}(\partial\Omega_h), \tag{12}$$

$$V = H(\mathrm{div}, \Omega_h) \times H^1(\Omega_h). \tag{13}$$

The notations here are defined by

$$H^{1/2}_0(\partial\Omega_h) = \{\eta : \exists\, w \in H^1_0(\Omega) \text{ such that } \eta|_{\partial K} = w|_{\partial K}\ \forall K \in \Omega_h\},$$

$$H^{-1/2}(\partial\Omega_h) = \{\eta \in \prod_K H^{-1/2}(\partial K) : \exists\, \vec{q} \in H(\mathrm{div}, \Omega) \text{ such that } \eta|_{\partial K} = \vec{q} \cdot \vec{n}|_{\partial K}\ \forall K \in \Omega_h\},$$

$$H(\mathrm{div}, \Omega_h) = \{\vec{\tau} :\ \vec{\tau}|_K \in H(\mathrm{div}, K),\ \forall K \in \Omega_h\}$$

$$H^1(\Omega_h) = \{v :\ v|_K \in H^1(K),\ \forall K \in \Omega_h\}.$$

Note that $w|_{\partial K}$, $\vec{q}\cdot\vec{n}|_{\partial K}$ etc. are abbreviated notations for appropriate trace operators. These traces are all well defined in the Sobolev spaces used above. The forms b and l in (11) are defined by

$$\begin{aligned} b(\ (\vec{\sigma},u,\hat{u},\hat{\sigma}_n),\ (\vec{\tau},v)\ ) &= (\alpha^{-1}\vec{\sigma},\vec{\tau})_{\Omega_h} - (u,\vec{\nabla}\cdot\vec{\tau})_{\Omega_h} + \langle\hat{u},\vec{\tau}\cdot\vec{n}\rangle_{\partial\Omega_h} \\ &\qquad - (\vec{\sigma},\vec{\nabla}v)_{\Omega_h} + \langle v,\hat{\sigma}_n\rangle_{\partial\Omega_h}, \\ l(\vec{\tau},v) &= (f,v)_{\Omega_h}. \end{aligned}$$

Note that the derivatives in the bilinear form are calculated element by element. Here and throughout, for concise notation that reflects the element by element calculations, we use

$$(r,s)_{\Omega_h} = \sum_{K\in\Omega_h}(r,s)_K, \qquad \langle w,\ell\rangle_{\partial\Omega_h} = \sum_{K\in\Omega_h}\langle w,\ell\rangle_{1/2,\partial K}.$$

We will also use $\|r\|_{\Omega_h}$ to denote the norm $(r,r)_{\Omega_h}^{1/2}$. The natural norms on the "broken" spaces $H^1(\Omega_h)$ and $H(\mathrm{div},\Omega_h)$ are defined by

$$\begin{aligned} \|v\|^2_{H^1(\Omega_h)} &= (v,v)_{\Omega_h} + (\vec{\nabla}v,\vec{\nabla}v)_{\Omega_h} \\ \|\vec{q}\|^2_{H(\mathrm{div},\Omega_h)} &= (\vec{q},\vec{q})_{\Omega_h} + (\vec{\nabla}\cdot\vec{q},\vec{\nabla}\cdot\vec{q})_{\Omega_h}. \end{aligned}$$

They determine the $\|\cdot\|_V$-norm for the space in (13).

The DPG solution of (11) consists of four components. While $\vec{\sigma}$ and u are simply called the flux and the solution components, resp., the customary names for the other two are

$$\text{numerical trace } (\hat{u}) \qquad \text{and} \qquad \text{numerical flux } (\hat{\sigma}_n). \tag{14}$$

As indicated above, these lie in $H_0^{1/2}(\partial\Omega_h)$ and $H^{-1/2}(\partial\Omega_h)$, resp. Note that these spaces consist of functions (or functionals, resp.) that can be interpreted as "single-valued" on element interfaces. They are normed by quotient norms, i.e.,

$$\|\hat{u}\|_{H_0^{1/2}(\partial\Omega_h)} = \inf\big\{\|w\|_{H^1(\Omega)} :\ \forall w\in H_0^1(\Omega) \text{ such that } \hat{u}|_{\partial K} = w|_{\partial K}\big\}, \tag{15a}$$

$$\|\hat{\sigma}_n\|_{H^{-1/2}(\partial\Omega_h)} = \inf\big\{\|\vec{q}\|_{H(\mathrm{div},\Omega)} :\ \forall\vec{q}\in H(\mathrm{div},\Omega) \text{ such that } \hat{\sigma}_n|_{\partial K} = \vec{q}\cdot\vec{n}|_{\partial K}\big\}. \tag{15b}$$

By standard arguments, we can conclude the existence of linear continuous liftings $E_{\mathrm{grad}} : H_0^{1/2}(\partial\Omega_h)\mapsto H_0^1(\Omega)$ and $E_{\mathrm{div}} : H^{-1/2}(\partial\Omega_h)\mapsto H(\mathrm{div},\Omega)$ such that

$$\|E_{\mathrm{grad}}\hat{u}\|_{H^1(\Omega)} = \|\hat{u}\|_{H_0^{1/2}(\partial\Omega_h)}, \qquad \|E_{\mathrm{div}}\hat{\sigma}_n\|_{H(\mathrm{div},\Omega)} = \|\hat{\sigma}_n\|_{H^{-1/2}(\partial\Omega_h)}. \tag{16}$$

This completes the description of the ultra-weak formulation (11) and its associated spaces and norms.

3.2. **The optimal test norm.** The optimal test norm was abstractly given in (1). For this specific application, it is easy to calculate it.

$$\begin{aligned} \|(\vec{\tau},v)\|_{\mathrm{opt},V} &= \sup_{(\vec{\sigma},u,\hat{u},\hat{\sigma}_n)\in U} \frac{b(\ (\vec{\sigma},u,\hat{u},\hat{\sigma}_n),\ (\vec{\tau},v)\ )}{\|\ (\vec{\sigma},u,\hat{u},\hat{\sigma}_n)\ \|_U} \\ &= \sup_{(\vec{\sigma},u,\hat{u},\hat{\sigma}_n)\in U} \frac{(\vec{\sigma},\alpha^{-1}\vec{\tau}-\vec{\nabla}v)_{\Omega_h} - (u,\vec{\nabla}\cdot\vec{\tau})_{\Omega_h} + \langle\hat{u},\vec{\tau}\cdot\vec{n}\rangle_{\partial\Omega_h} + \langle v,\hat{\sigma}_n\rangle_{\partial\Omega_h}}{\Big(\|\vec{\sigma}\|^2_{L^2(\Omega)} + \|u\|^2_{L^2(\Omega)} + \|\hat{u}\|^2_{H^{1/2}(\partial\Omega_h)} + \|\hat{\sigma}_n\|^2_{H^{-1/2}(\partial\Omega_h)}\Big)^{1/2}}. \end{aligned}$$

Elementary arguments then show that

$$\|(\vec{\tau},v)\|^2_{\mathrm{opt},V} = \|\alpha^{-1}\vec{\tau}-\vec{\nabla}v\|^2_{\Omega_h} + \|\vec{\nabla}\cdot\vec{\tau}\|^2_{\Omega_h} + \big\|[\vec{\tau}\cdot\vec{n}]\big\|^2_{\partial\Omega_h} + \big\|[v\vec{n}]\big\|^2_{\partial\Omega_h}, \tag{17}$$

where

$$\big\|[\vec{\tau}\cdot\vec{n}]\big\|_{\partial\Omega_h} \overset{\text{def}}{=} \sup_{\hat{u}\in H_0^{1/2}(\partial\Omega_h)} \frac{\langle \hat{u}, \vec{\tau}\cdot\vec{n}\rangle_{\partial\Omega_h}}{\|\hat{u}\|_{H_0^{1/2}(\partial\Omega_h)}} = \sup_{w\in H_0^1(\Omega)} \frac{\langle w, \vec{\tau}\cdot\vec{n}\rangle_{\partial\Omega_h}}{\|w\|_{H^1(\Omega)}}, \tag{18a}$$

$$\big\|[v\vec{n}]\big\|_{\partial\Omega_h} \overset{\text{def}}{=} \sup_{\hat{\sigma}_n\in H^{-1/2}(\partial\Omega_h)} \frac{\langle v, \hat{\sigma}_n\rangle_{\partial\Omega_h}}{\|\hat{\sigma}_n\|_{H^{-1/2}(\partial\Omega_h)}} = \sup_{\vec{q}\in H(\mathrm{div},\Omega)} \frac{\langle v, \vec{q}\cdot\vec{n}\rangle_{\partial\Omega_h}}{\|\vec{q}\|_{H(\mathrm{div},\Omega)}}. \tag{18b}$$

The last equalities in either case are a consequence of the definition of the spaces and their quotient norms, as defined in (15). E.g., to prove the last equality in (18b), observe that for every $\vec{q}$ in $H(\mathrm{div},\Omega)$, the trace $\hat{\sigma}_n|_{\partial K} = \vec{q}\cdot\vec{n}|_{\partial K}$ satisfies

$$\frac{\langle v, \hat{\sigma}_n\rangle_{\partial\Omega_h}}{\|\hat{\sigma}_n\|_{H^{-1/2}(\partial\Omega_h)}} \geq \frac{\langle v, \vec{q}\cdot\vec{n}\rangle_{\partial\Omega_h}}{\|\vec{q}\|_{H(\mathrm{div},\Omega)}}$$

due to (15b). Hence the first supremum in (18b) is greater than or equal to the second. The reverse inequality also holds because for every $\hat{\sigma}_n$ in $H^{-1/2}(\partial\Omega_h)$, there is a function $\vec{q}$ in $H(\mathrm{div},\Omega)$, namely $\vec{q} = E_{\mathrm{div}}\hat{\sigma}_n$ (see (16)) such that $\vec{q}\cdot\vec{n}|_{\partial K} = \hat{\sigma}_n$ and $\|\vec{q}\|_{H(\mathrm{div},\Omega)} = \|\hat{\sigma}_n\|_{H^{-1/2}(\partial\Omega_h)}$.

The norms in (18) measure the size of "jumps". Indeed, it is easy to see that when applied to functions without jumps they are zero, specifically,

$$\big\|[\vec{\rho}\cdot\vec{n}]\big\|_{\partial\Omega_h} = 0 \qquad \forall \vec{\rho}\in H(\mathrm{div},\Omega), \tag{19a}$$

$$\big\|[\phi\,\vec{n}]\big\|_{\partial\Omega_h} = 0 \qquad \forall \phi\in H_0^1(\Omega). \tag{19b}$$

E.g., to prove (19a), consider the numerator $\langle w, \vec{\rho}\cdot\vec{n}\rangle_{\partial\Omega_h}$ in (18a) for some w in $H_0^1(\Omega)$. We first integrate by parts locally, and next globally, to get

$$\begin{aligned}\langle w, \vec{\rho}\cdot\vec{n}\rangle_{\partial\Omega_h} &= (\vec{\nabla}w, \vec{\rho})_{\Omega_h} + (w, \vec{\nabla}\cdot\vec{\rho})_{\Omega_h}\\ &= (\vec{\nabla}w, \vec{\rho})_{\Omega} + (w, \vec{\nabla}\cdot\vec{\rho})_{\Omega}\\ &= \langle w, \vec{\rho}\cdot\vec{n}\rangle_{\partial\Omega}\end{aligned}$$

which vanishes due to the global boundary condition of w in $H_0^1(\Omega)$. Similarly, we can prove (19b).

As shown abstractly in [14, 23], if one is able to compute the trial-to-test operator with respect to the optimal test norm – i.e., use $\|\cdot\|_{\mathrm{opt},V}$ in place of $\|\cdot\|_V$ in (3) – then the DPG solution would coincide with the *best* approximation from U_h in U-norm. However, in most examples, including our current application, the optimal test norm is not easy to compute with. The optimal norm given in (17) is inconvenient for practical computations, due to the last two "jump" terms. These terms would make the trial-to-test computation in (3) non-local.

Therefore, a fundamental ingredient in our analysis (in the next section) is the proof of equivalence of the optimal norm in (17) with the simpler standard V-norm

$$\|\,(\vec{\tau},v)\,\|^2_V = \|\vec{\tau}\|^2_{H(\mathrm{div},\Omega_h)} + \|v\|^2_{H^1(\Omega_h)}. \tag{20}$$

This "broken" or "localizable" test space norm does not have any jump terms.

4. Error estimates

The DPG method uses the ultra-weak formulation developed in § 3.1 together with a (conforming) subspace U_h of the space U in (12). This section is devoted to proving the following results on bounds for the discretization error.

Theorem 4.1 (Quasioptimality)**.** *Suppose* $(\vec{\sigma}, u, \hat{u}, \hat{\sigma}_n) \in U$ *and* $(\vec{\sigma}_h, u_h, \hat{u}_h, \hat{\sigma}_{n,h}) \in U_h$ *are the exact and approximate solutions, resp. Let the discretization error* (Disc.Err.) *and the best approximation error* (Proj.Err.) *be denoted by*

$$\text{Disc.Err.} = \|\vec{\sigma} - \vec{\sigma}_h\|_{L^2(\Omega)} + \|u - u_h\|_{L^2(\Omega)} + \|\hat{u} - \hat{u}_h\|_{H_0^{1/2}(\partial\Omega_h)} + \|\hat{\sigma}_n - \hat{\sigma}_{n,h}\|_{H^{-1/2}(\partial\Omega_h)},$$

$$\text{Proj.Err.} = \inf_{(\vec{\rho}_h, w_h, \hat{z}_h, \hat{\eta}_{n,h}) \in U_h} \left(\|\vec{\sigma}_h - \vec{\rho}_h\|_{L^2(\Omega)} + \|u - w_h\|_{L^2(\Omega)} + \|\hat{u} - \hat{z}_h\|_{H_0^{1/2}(\partial\Omega_h)} + \|\hat{\sigma}_n - \hat{\eta}_{n,h}\|_{H^{-1/2}(\partial\Omega_h)} \right).$$

Then there is a $C(\alpha) > 0$ *independent of the subspace* U_h *and the partition* Ω_h *such that*

$$\text{Disc.Err.} \;\le\; C(\alpha)\ \text{Proj.Err.}$$

The value of $C(\alpha)$ *is an increasing function of* α_1 *and* $1/\alpha_0$.

The proof of Theorem 4.1 appears in § 4.4 below. Note that the discretization subspace U_h is unspecified in the theorem – the result holds for any U_h. In this sense, the theorem is comparable to Céa lemma (although the proof is much more involved). If we specify any particular finite element subspace U_h, with specific finite element shapes (simplices, quadrilaterals, etc.) whose best approximation properties are known, we can conclude rates of convergence.

As an example, let us consider the case of a finite element space built on a tetrahedral mesh. Let $P_p(D)$ denote the set of functions that are restrictions of (multivariate) polynomials of degree at most p on a domain D. Let

$$S_{h,p} = \{\vec{\rho} : \vec{\rho}|_K \in P_p(K)^N\}, \qquad W_{h,p} = \{v : v|_K \in P_p(K)\}.$$

Define the numerical trace and flux approximation spaces by

$$M_{h,p} = \{\eta : \exists\, w \in W_{h,p} \cap H_0^1(\Omega) \text{ such that } \eta|_{\partial K} = w|_{\partial K}\ \forall K \in \Omega_h\},$$
$$Q_{h,p} = \{\eta :\ \eta|_E \in P_p(E),\ \forall \text{ mesh faces } E\}.$$

We assume $p \ge 1$ so that $M_{h,p}$ is non-trivial.

Let us apply Theorem 4.1 with these as the trial spaces for each solution component. Then, to obtain rates of convergence, we only need to examine how the best approximation error converges in terms of h and p. It is well known that for $s > 0$,

$$\inf_{w_h \in W_{h,p}} \|u - w_h\|_\Omega \le C h^s p^{-s} |u|_{H^s(\Omega)}, \qquad (s \le p+1). \tag{21}$$

A similar best approximation estimate obviously holds for $\vec{\sigma}$ as well.

The only best approximation terms that need further explanation are the flux terms. Since the exact solution $\hat{u}$ is the trace of u and the exact flux $\hat{\sigma}_n$ is the trace of the interfacial normal components of $\vec{\sigma}$, we have

$$\inf_{\hat{z}_h \in M_{h,p}} \|\hat{u} - \hat{z}_h\|_{H_0^{1/2}(\partial\Omega_h)} \le \|u - \Pi_{\text{grad}} u\|_{H^1(\Omega)}$$
$$\inf_{\hat{\eta}_{n,h} \in Q_{h,p}} \|\hat{\sigma}_n - \hat{\eta}_{n,h}\|_{H^{-1/2}(\partial\Omega_h)} \le \|\vec{\sigma} - \Pi_{\text{div}} \vec{\sigma}\|_{H(\text{div},\Omega)},$$

where $\Pi_{\text{grad}} u \in H^1_0(\Omega)$ and $\Pi_{\text{div}} \vec{\sigma} \in H(\text{div}, \Omega)$ are suitable projections, such that their traces $\Pi_{\text{grad}} u|_E$ and $\Pi_{\text{div}} \vec{\sigma} \cdot \vec{n}|_E$ on any mesh edge E is in $P_p(E)$.

Such conforming projectors providing approximation estimates with constants independent of p are not easy to construct. However they are now available from recent results in [11, 12, 16, 17, 18]. In [12, Corollaries 1 and 2], and later in the corrected version of the results in [11, Theorem 5.3], projectors Π_{grad} and Π_{div} satisfying

$$\|u - \Pi_{\text{grad}} u\|_{H^1(\Omega)} \leq C \ln(p)^2 \, h^s p^{-s} |u|_{H^{s+1}(\Omega)}, \qquad (s \leq p), \tag{22a}$$

$$\|\vec{\sigma} - \Pi_{\text{div}} \vec{\sigma}\|_{L^2(\Omega)} \leq C \ln(p) \, h^s p^{-s} |\vec{\sigma}|_{H^{s+1}(\Omega)}, \qquad (s \leq p+1). \tag{22b}$$

were given under the assumption that $s > 1/2$ and the conjecture that a certain polynomial extension operator exists. The latter conjecture was recently proved in [16, 17, 18] and as a result, the estimates of (22) are finally proved.

Furthermore, let us now observe that Π_{div} can be chosen to be either a projector into $S_{h,p} \cap H(\text{div}, \Omega)$ or a projector into the Raviart-Thomas space.[1] This is because the normal traces of functions in both these spaces result in the same $Q_{h,p}$. Projectors Π_{div} into both these spaces have been analyzed in [11]. Let Π_p denote the L^2-orthogonal projection into $W_{h,p}$. The projector into the former space satisfies the commutativity property $\vec{\nabla} \cdot \Pi_{\text{div}} \vec{\sigma} = \Pi_{p-1} \vec{\nabla} \cdot \vec{\sigma}$, while the projector into the Raviart-Thomas space satisfies $\vec{\nabla} \cdot \Pi_{\text{div}} \vec{\sigma} = \Pi_p \vec{\nabla} \cdot \vec{\sigma}$. Hence, for the latter,

$$\|\vec{\nabla} \cdot (\vec{\sigma} - \Pi_{\text{div}} \vec{\sigma})\|_{L^2(\Omega)} \leq C h^s p^{-s} |\vec{\nabla} \cdot \vec{\sigma}|_{H^s(\Omega)}, \qquad (s \leq p+1). \tag{23}$$

Thus, although we could have used either projector for estimating best approximation error for the numerical flux, it is preferable to use the projector into the Raviart-Thomas space to get the best power of h.

Now, comparing the rates of convergence in (21), (22) and (23), we find that to obtain a full $O(h^{p+1})$ order of convergence, we must increase the polynomial degree of the numerical trace space to $p+1$. Combining these observations, we have the following corollary.

Corollary 4.1 (h and p convergence rates)**.** *Suppose Ω_h is a shape regular tetrahedral finite element mesh and let h denote the maximum of the diameters of its elements. Set*

$$U_h = S_{h,p} \times W_{h,p} \times M_{h,p+1} \times Q_{h,p}.$$

Then there is a constant C independent of h and p (but dependent on the shape regularity and α) such that

$$\text{Disc.Err.} \;\leq\; C \ln(p)^2 \, h^s p^{-s} (\|u\|_{H^{s+1}(\Omega)} + \|\vec{\sigma}\|_{H^{s+1}(\Omega)})$$

for all $1/2 < s \leq p+1$.

In the same way, one can derive convergence rates for other element shapes and spaces (triangles, hexahedra, etc.) from Theorem 4.1. In the remainder, we develop the results needed to prove Theorem 4.1. The proof proceeds by applying the abstract result of Theorem 2.1. Hence we must verify its assumptions. The injectivity assumption (5) is verified in § 4.1. Most of the work is in proving one side of the two sided inequality of the second assumption (6), which appears in § 4.3. The proof is completed in § 4.4.

[1] The Raviart-Thomas space [22] consists of functions in $H(\text{div}, \Omega)$, which when restricted to a mesh tetrahedron, takes the form $\vec{r}_p + \vec{x} s_p$ for some $\vec{r}_p \in P_p(K)^N$ and $s_p \in P_p(K)$.

4.1. **Uniqueness.** Let us verify the first assumption of Theorem 2.1, namely (5).

Lemma 4.1. *With U and V as set in* (12) *and* (13)*, suppose $(\vec{\sigma}, u, \hat{u}, \hat{\sigma}_n) \in U$ satisfies*

$$b(\ (\vec{\sigma}, u, \hat{u}, \hat{\sigma}_n),\ (\vec{\tau}, v)\ ) = 0 \tag{24}$$

for all $(\vec{\tau}, v) \in V$. Then

$$(\vec{\sigma}, u, \hat{u}, \hat{\sigma}_n) = 0.$$

Proof. Eq. (24) implies that on every mesh element K,

$$(\alpha^{-1}\vec{\sigma}, \vec{\tau})_K - (u, \vec{\nabla} \cdot \vec{\tau})_K + \langle \hat{u}, \vec{\tau} \cdot \vec{n} \rangle_{1/2,\partial K} = 0 \qquad \forall \vec{\tau} \in H(\mathrm{div}, K) \tag{25}$$

$$-(\vec{\sigma}, \vec{\nabla} v)_K + \langle v, \hat{\sigma}_n \rangle_{1/2,\partial K} = 0 \qquad \forall v \in H^1(K). \tag{26}$$

Choosing an infinitely smooth v compactly supported on K, we find that (26) implies

$$\vec{\nabla} \cdot \vec{\sigma} = 0 \tag{27}$$

distributionally. Similarly, (25) implies the distributional gradient of u satisfies

$$\vec{\nabla} u = -\alpha^{-1}\vec{\sigma}. \tag{28}$$

We also have that

$$u|_{\partial K} = \hat{u}|_{\partial K}, \qquad \hat{\sigma}_n|_{\partial K} = \vec{\sigma} \cdot \vec{n}|_{\partial K}. \tag{29}$$

This is obtained by integrating (25)–(26) by parts and using (27)–(28) to find that $\langle \hat{u} - u, \vec{\tau} \cdot \vec{n} \rangle_{1/2,\partial K} = \langle \hat{\sigma}_n - \vec{\sigma} \cdot \vec{n}, v \rangle_{1/2,\partial K} = 0$. Collecting these observations, let us note that (27) and (28) imply $\vec{\sigma} \in H(\mathrm{div}, K)$ and $u \in H^1(K)$ for every mesh element K. Furthermore, (29) then implies that $u \in H^1_0(\Omega)$ and $\vec{\sigma} \in H(\mathrm{div}, \Omega)$.

Due to this extra regularity of $\vec{\sigma}$ and u, we may set $\vec{\tau} = \vec{\sigma}$ and $v = u$ in (25)–(26). Summing these equations and canceling terms after integrating by parts, we find that

$$(\alpha^{-1}\vec{\sigma}, \vec{\sigma})_{\Omega_h} - \langle u, \vec{\sigma} \cdot \vec{n} \rangle_{\partial\Omega_h} + \langle \hat{u}, \vec{\sigma} \cdot \vec{n} \rangle_{\partial\Omega_h} + \langle u, \hat{\sigma}_n \rangle_{\partial\Omega_h} = 0. \tag{30}$$

The last two boundary terms vanish (cf. (19)). Furthermore, $\langle u, \vec{\sigma} \cdot \vec{n} \rangle_{\partial\Omega_h} = \langle u, \vec{\sigma} \cdot \vec{n} \rangle_{\partial\Omega} = 0$ as $u \in H^1_0(\Omega)$. Thus, (30) implies that $\vec{\sigma} = 0$. Consequently, by (28), u must be constant. Since $u \in H^1_0(\Omega)$, this implies that $u = 0$. Since both $\vec{\sigma}$ and u vanishes, by (29), the unknowns on the element boundary, $\hat{u}$ and $\hat{\sigma}_n$, also vanish. □

4.2. **A Poincaré inequality.** Next, we give a Poincaré-type inequality for discontinuous functions. Stronger results are available in [4] (under further conditions on mesh angles – see e.g., [4, Cor. 6.3]), but we only need the following simple lemma, which holds with no assumptions on the mesh. A simple proof is included for completeness.

Lemma 4.2. *There is a constant C_P independent of Ω_h such that for all v in $H^1(\Omega_h)$,*

$$\|v\|_{\Omega_h} \le C_P \left(\|\vec{\nabla} v\|_{\Omega_h} + \left\| [v\vec{n}] \right\|_{\partial\Omega_h} \right).$$

Proof. Let ϕ in $H^1_0(\Omega)$ solve the Dirichlet problem $-\Delta\phi = v$. Then, by the weak formulation for ϕ, we obviously have $\|\vec{\nabla}\phi\|^2_\Omega = (v,\phi)_{\Omega_h}$. Hence

$$\begin{aligned}\|v\|^2_\Omega &= (v,-\Delta\phi)_\Omega = (\vec{\nabla}v,\vec{\nabla}\phi)_{\Omega_h} + \langle v, \frac{\partial\phi}{\partial n}\rangle_{\partial\Omega_h}\\ &\le \|\vec{\nabla}v\|_{\Omega_h}\|\vec{\nabla}\phi\|_{\Omega_h} + \left(\frac{\langle v,\vec{\nabla}\phi\cdot\vec{n}\rangle_{\partial\Omega_h}}{\|\vec{\nabla}\phi\|_{H(\mathrm{div},\Omega)}}\right)\|\vec{\nabla}\phi\|_{H(\mathrm{div},\Omega)}\\ &\le \|\vec{\nabla}v\|_{\Omega_h}\left|(v,\phi)_{\Omega_h}\right|^{1/2} + \left(\sup_{\vec{q}\in H(\mathrm{div},\Omega)}\frac{\langle v,\vec{q}\cdot\vec{n}\rangle_{\partial\Omega_h}}{\|\vec{q}\|_{H(\mathrm{div},\Omega)}}\right)\left(\left|(v,\phi)_{\Omega_h}\right| + \|v\|^2_{\Omega_h}\right)^{1/2}.\end{aligned}$$

Applying the standard Poincaré inequality in $H^1_0(\Omega)$ for ϕ to bound $\|\phi\|_\Omega \le C\|\vec{\nabla}\phi\|_\Omega$, and performing obvious estimations using the arithmetic-geometric mean inequality, we obtain the result. □

4.3. **An inf-sup condition.** Next, we prove that the inf-sup condition

$$C_1\|(\vec{\tau},v)\|_V \le \sup_{(\vec{\sigma},u,\hat{u},\hat{\sigma}_n)\in U}\frac{b(\ (\vec{\sigma},u,\hat{u},\hat{\sigma}_n),\ (\vec{\tau},v)\ )}{\|\ (\vec{\sigma},u,\hat{u},\hat{\sigma}_n)\ \|_U} \tag{31}$$

holds with a constant C_1 independent of Ω_h. Note that the supremum on the right hand side is the same as the optimal test norm. Hence, this inf-sup condition is the same as the lower bound in the assumption (6) of Theorem 2.1.

The idea is to decompose v into two parts v_0 and v_1. The function v_1 is continuous across elements and solves a mixed problem with nonzero source in general (precisely described in Lemma 4.4). On the other hand, v_0 is harmonic (in the $\alpha = 1$ case) on each element but discontinuous across elements. That $\vec{\nabla}v_0$ can be controlled solely by the jump terms is the content of the next lemma. Let $\|\vec{r}\|_\alpha = (\alpha\,\vec{r},\vec{r})_{\Omega_h}$ and $\|\vec{r}\|_{1/\alpha} = (\alpha^{-1}\vec{r},\vec{r})_{\Omega_h}$. From now on, we will use C to denote a generic constant independent of Ω_h.

Lemma 4.3. *Let $\vec{\tau}_0$ in $H(\mathrm{div},\Omega_h)$ and v_0 in $H^1(\Omega_h)$ satisfy*

$$\alpha^{-1}\vec{\tau}_0 - \vec{\nabla}v_0 = 0, \qquad \text{on } K, \tag{32a}$$

$$\vec{\nabla}\cdot\vec{\tau}_0, = 0 \qquad \text{on } K, \tag{32b}$$

for every element K in Ω_h. Then

$$\|\vec{\tau}_0\|_{1/\alpha} = \|\vec{\nabla}v_0\|_\alpha \le C\left(\left\|[\vec{\tau}_0\cdot\vec{n}]\right\|_{\partial\Omega_h}\sqrt{1/\alpha_0} + \left\|[v_0\vec{n}]\right\|_{\partial\Omega_h}\sqrt{\alpha_1}\right) \tag{33}$$

Proof. Let us consider the three dimensional case first. We need the weighted Helmholtz decomposition

$$\vec{\tau}_0 = \alpha\vec{\nabla}\psi + \vec{\nabla}\times\vec{z} \tag{34}$$

for some ψ in $H^1_0(\Omega)$ and $\vec{z}$ in $H(\mathrm{curl},\Omega)$. To obtain this decomposition, we first find the $\psi\in H^1_0(\Omega)$ by solving

$$(\alpha\vec{\nabla}\psi,\vec{\nabla}\varphi)_\Omega = (\vec{\tau}_0,\vec{\nabla}\varphi)_{\Omega_h}, \quad \forall\varphi\in H^1_0(\Omega). \tag{35}$$

Then, since the distributional divergence $\vec{\nabla}\cdot(\vec{\tau}_0 - \alpha\vec{\nabla}\psi) = 0$, by a well-known exact sequence property implied by our topological assumptions on Ω, we can find a $\vec{z}$ in

$H(\mathrm{curl},\Omega)$ such that $\vec{\nabla}\times\vec{z}=\vec{\tau}_0-\alpha\vec{\nabla}\psi$, which gives (34). By construction, the two components $\alpha\vec{\nabla}\psi$ and $\vec{\nabla}\times\vec{z}$ are orthogonal in $(\alpha^{-1}\cdot,\cdot)$-inner product and

$$\|\vec{\nabla}\times\vec{z}\|_{1/\alpha}^2+\|\vec{\nabla}\psi\|_{\alpha}^2=\|\vec{\tau}_0\|_{1/\alpha}^2. \tag{36}$$

Using (35), let us estimate $\vec{\tau}_0$ as follows.

$$\begin{aligned}\|\vec{\tau}_0\|_{1/\alpha}^2&=(\alpha^{-1}\vec{\tau}_0,\vec{\tau}_0)=(\alpha^{-1}\vec{\tau}_0,\alpha\vec{\nabla}\psi+\vec{\nabla}\times\vec{z})_{\Omega_h}\\&=(\vec{\tau}_0,\vec{\nabla}\psi)_{\Omega_h}+(\vec{\nabla}v_0,\vec{\nabla}\times\vec{z})_{\Omega_h}\\&=-(\vec{\nabla}\cdot\vec{\tau}_0,\psi)_{\Omega_h}+\langle\psi,\vec{\tau}_0\cdot\vec{n}\rangle_{\partial\Omega_h}+\langle v_0,\vec{n}\cdot\vec{\nabla}\times\vec{z}\rangle_{\partial\Omega_h},\end{aligned}$$

One of the terms vanishes by (32b). Hence,

$$\begin{aligned}\|\vec{\tau}_0\|_{1/\alpha}^2&=\langle\psi,\vec{\tau}_0\cdot\vec{n}\rangle_{\partial\Omega_h}+\langle v_0,\vec{n}\cdot\vec{\nabla}\times\vec{z}\rangle_{\partial\Omega_h},\\&=\frac{\langle\psi,\vec{\tau}_0\cdot\vec{n}\rangle_{\partial\Omega_h}}{\|\psi\|_{H^1(\Omega)}}\|\psi\|_{H^1(\Omega)}+\frac{\langle v_0,\vec{n}\cdot\vec{\nabla}\times\vec{z}\rangle_{\partial\Omega_h}}{\|\vec{\nabla}\times\vec{z}\|_{H(\mathrm{div},\Omega)}}\|\vec{\nabla}\times\vec{z}\|_{L^2(\Omega)}\\&\le\left(\sup_{w\in H_0^1(\Omega)}\frac{\langle w,\vec{\tau}_0\cdot\vec{n}\rangle_{\partial\Omega_h}}{\|w\|_{H^1(\Omega)}}\right)\|\psi\|_{H^1(\Omega)}+\left(\sup_{\vec{q}\in H(\mathrm{div},\Omega)}\frac{\langle v_0,\vec{n}\cdot\vec{q}\rangle_{\partial\Omega_h}}{\|\vec{q}\|_{H(\mathrm{div},\Omega)}}\right)\|\vec{\nabla}\times\vec{z}\|_{L^2(\Omega)}\end{aligned}$$

It now follows from (36) and the standard Poincaré inequality $\|\psi\|_{\Omega}^2\le C\|\vec{\nabla}\psi\|_{\Omega}^2$ that

$$\|\vec{\tau}_0\|_{1/\alpha}^2\le\left\|[\vec{\tau}_0\cdot\vec{n}]\right\|_{\partial\Omega_h}\frac{\sqrt{1+C}}{\sqrt{\alpha_0}}\|\vec{\tau}_0\|_{1/\alpha}+\left\|[v_0\vec{n}]\right\|_{\partial\Omega_h}\sqrt{\alpha_1}\|\vec{\tau}_0\|_{1/\alpha}$$

which proves the lemma.

In the two dimensional case, the same argument works if the vector potential $\vec{z}$ is replaced by a scalar potential z and $\vec{\nabla}\times\vec{z}$ by the rotated gradient $\vec{\nabla}\wedge z=(-\partial_2 z,\partial_1 z)$. □

Lemma 4.4. *Let $\vec{G}\in L^2(\Omega)^N$ and $F\in L^2(\Omega)$. There is a $\vec{\tau}_1$ in $H(\mathrm{div},\Omega)$ and v_1 in $H_0^1(\Omega)$ satisfying*

$$\alpha^{-1}\vec{\tau}_1-\vec{\nabla}v_1=\vec{G}, \qquad \text{on } \Omega, \tag{37a}$$

$$\vec{\nabla}\cdot\vec{\tau}_1,=F \qquad \text{on } \Omega, \tag{37b}$$

and

$$\|\vec{\tau}_1\|_{\Omega}+\|\vec{\nabla}v_1\|_{\Omega}\le C\left((1+\alpha_1)\|\vec{G}\|_{\Omega}+\left(1+\frac{\alpha_1}{\alpha_0}\right)\|F\|_{\Omega}\right). \tag{38}$$

Proof. If $\vec{\tau}_1$ and v_1 satisfy (37), then they form the unique solution of the following well-known mixed weak formulation: Find τ_1 in $H(\mathrm{div},\Omega)$ and v in $L^2(\Omega)$ such that

$$(\alpha^{-1}\vec{\tau}_1,\vec{\rho})_{\Omega}+(v_1,\vec{\nabla}\cdot\vec{\rho})_{\Omega}=(\vec{G},\vec{\rho})_{\Omega} \qquad \forall\vec{\rho}\in H(\mathrm{div},\Omega) \tag{39a}$$

$$(\vec{\nabla}\cdot\vec{\tau}_1,w)_{\Omega}=(F,w)_{\Omega} \qquad \forall w\in L^2(\Omega). \tag{39b}$$

Uniqueness and stability of solutions for this formulation are well-known [5]. These follow by verifying the conditions of the Babuška-Brezzi theory, namely

$$\alpha_1^{-1}\|\vec{\rho}\|_{\Omega}^2\le(\alpha^{-1}\vec{\rho},\vec{\rho})_{\Omega}\le\alpha_0^{-1}\|\vec{\rho}\|_{H(\mathrm{div},\Omega)}^2, \qquad \forall\vec{\rho}\in H_0(\mathrm{div},\Omega),$$

$$\sup_{\vec{\rho}\in H_0(\mathrm{div},\Omega)}\frac{(v,\vec{\nabla}\cdot\vec{\rho})_{\Omega}}{\|\vec{\rho}\|_{H(\mathrm{div},\Omega)}}\ge C_{\Omega}\|v\|_{\Omega}, \qquad \forall v\in L^2(\Omega).$$

where C_Ω is a constant depending only on Ω. Hence by [5, Ch. II, Prop. 1.3], the solution of (37) satisfies

$$\|\vec{\tau}_1\|_{H(\mathrm{div},\Omega)} \leq \alpha_1 \|\vec{G}\|_\Omega + \frac{\alpha_1^{-1} + \alpha_0^{-1}}{C_\Omega \alpha_1^{-1}} \|F\|_\Omega \tag{40}$$

Clearly, (40) and (37a) also imply that

$$\|\vec{\nabla} v_1\|_\Omega \leq \|\vec{G}\|_\Omega + \alpha_0^{-1}\|\vec{\tau}_1\|_\Omega \leq (\alpha_1 + 1)\|\vec{G}\|_\Omega + C_\Omega(1 + \alpha_1/\alpha_0)\|F\|_\Omega.$$

Together, they prove the lemma. □

Theorem 4.2 (The inf-sup condition)**.** *The inequality* (31) *holds, i.e., with* $\|(\vec{\tau}, v)\|_{\mathrm{opt},V}$ *and* $\|(\vec{\tau}, v)\|_V$ *as in* (17) *and* (20)*, resp., the inequality*

$$\|(\vec{\tau}, v)\|_V \leq C_\alpha \, \|(\vec{\tau}, v)\|_{\mathrm{opt},V}\,, \tag{41}$$

holds for all $\vec{\tau} \in H(\mathrm{div}, \Omega_h)$ *and* $v \in H^1(\Omega_h)$*. Here* C_α *is independent of* Ω_h *and is an increasing function of* α_1 *and* $1/\alpha_0$*.*

Proof. Let $F = \vec{\nabla} \cdot \vec{\tau}$ and $\vec{G} = \alpha^{-1}\vec{\tau} - \vec{\nabla} v$, where as before, the derivatives are calculated element by element. Clearly $F \in L^2(\Omega)$ and $\vec{G} \in L^2(\Omega)^N$. Let $(\vec{\tau}_1, v_1)$ in $H(\mathrm{div}, \Omega) \times H_0^1(\Omega)$ be the solution of (37) with this F and $\vec{G}$. Then by Lemma 4.4,

$$\|\vec{\tau}_1\|_\Omega + \|\vec{\nabla} v_1\|_\Omega \leq C\left((1+\alpha_1)\|\vec{G}\|_\Omega + \left(1 + \frac{\alpha_1}{\alpha_0}\right)\|F\|_\Omega\right). \tag{42}$$

Let $\vec{\tau}_0 = \vec{\tau} - \vec{\tau}_1$ and $v_0 = v - v_1$. Then, the pair $(\vec{\tau}_0, v_0)$ obviously satisfies (32), so by Lemma 4.3,

$$\|\vec{\tau}_0\|_\Omega + \|\vec{\nabla} v_0\|_{\Omega_h} \leq C\left(\left(\frac{\sqrt{\alpha_1}}{\sqrt{\alpha_0}} + \frac{1}{\alpha_0}\right)\big\|[\vec{\tau}_0 \cdot \vec{n}]\big\|_{\partial\Omega_h} + \left(\alpha_1 + \frac{\sqrt{\alpha_1}}{\sqrt{\alpha_0}}\right)\big\|[v_0 \vec{n}]\big\|_{\partial\Omega_h}\right). \tag{43}$$

With these observations, we now proceed to prove the theorem.

Since $\vec{\tau} = \vec{\tau}_0 + \vec{\tau}_1$ and $v = v_0 + v_1$, by triangle inequality,

$$\|\vec{\tau}\|_\Omega \leq \|\vec{\tau}_0\|_\Omega + \|\vec{\tau}_1\|_\Omega \qquad \leq \kappa_\alpha\big(\|\vec{G}\|_\Omega + \|F\|_\Omega + \big\|[\vec{\tau}_0 \cdot \vec{n}]\big\|_{\partial\Omega_h} + \big\|[v_0\vec{n}]\big\|_{\partial\Omega_h}\big), \tag{44a}$$

$$\|\vec{\nabla} v\|_{\Omega_h} \leq \|\vec{\nabla} v_0\|_{\Omega_h} + \|\vec{\nabla} v_1\|_\Omega \leq \kappa_\alpha\big(\|\vec{G}\|_\Omega + \|F\|_\Omega + \big\|[\vec{\tau}_0 \cdot \vec{n}]\big\|_{\partial\Omega_h} + \big\|[v_0\vec{n}]\big\|_{\partial\Omega_h}\big). \tag{44b}$$

where κ_α is an increasing function of α_1 and $1/\alpha_0$. Now, observe that by (19), the terms $\big\|[\vec{\tau}_0 \cdot \vec{n}]\big\|_{\partial\Omega_h}$ and $\big\|[v_0\vec{n}]\big\|_{\partial\Omega_h}$ can be replaced by $\big\|[\vec{\tau} \cdot \vec{n}]\big\|_{\partial\Omega_h}$ and $\big\|[v\vec{n}]\big\|_{\partial\Omega_h}$, resp. Recalling from (17) that

$$\|(\vec{\tau}, v)\|_{\mathrm{opt},V}^2 = \|\vec{G}\|_{\Omega_h}^2 + \|F\|_{\Omega_h}^2 + \big\|[\vec{\tau} \cdot \vec{n}]\big\|_{\partial\Omega_h}^2 + \big\|[v\vec{n}]\big\|_{\partial\Omega_h}^2,$$

we conclude that (44) implies

$$\|\vec{\tau}\|_{\Omega_h} + \|\vec{\nabla} v\|_{\Omega_h} \leq C\kappa_\alpha \, \|(\vec{\tau}, v)\|_{\mathrm{opt},V}\,.$$

Therefore, to complete the proof of (41), we only need to bound the remaining terms that compose the norm $\|(\vec{\tau}, v)\|_V$. The needed bounds follow from

$$\|\vec{\nabla} \cdot \vec{\tau}\|_{\Omega_h} = \|F\|_{\Omega_h},$$

$$C\|v\|_{\Omega_h} \leq \|\vec{\nabla} v\|_{\Omega_h} + \big\|[v\vec{n}]\big\|_{\partial\Omega_h}.$$

The first statement above is obvious, while the second follows from Lemma 4.2. □

4.4. **Proof of Theorem 4.1.** We apply Theorem 2.1. Accordingly, we verify its assumptions. Assumption (5) is verified by Lemma 4.1, so we only need to verify (6). The lower bound of (6) is already verified by Theorem 4.2 with $C_1 = 1/C_\alpha$, so let us prove the upper bound $\|(\vec{\tau}, v)\|_{\text{opt},V} \leq C_2 \|(\vec{\tau}, v)\|_V$.

To this end, we consider each of the terms in the optimal norm in (17), beginning with the jump terms. Integrating by parts locally and applying Cauchy-Schwarz inequality,

$$\left\|[\vec{\tau}\cdot\vec{n}]\right\|_{\partial\Omega_h} = \sup_{w\in H_0^1(\Omega)} \frac{\langle w, \vec{\tau}\cdot\vec{n}\rangle_{\partial\Omega_h}}{\|w\|_{H^1(\Omega)}} = \sup_{w\in H_0^1(\Omega)} \frac{(\vec{\nabla} w, \vec{\tau})_{\Omega_h} + (w, \vec{\nabla}\cdot\vec{\tau})_{\Omega_h}}{\|w\|_{H^1(\Omega)}} \leq \|\vec{\tau}\|_{H(\text{div},\Omega_h)}.$$

A similar argument proves that

$$\left\|[v\vec{n}]\right\|_{\partial\Omega_h} \leq \|v\|_{H^1(\Omega_h)}.$$

The remaining terms are handled obviously:

$$\|\alpha^{-1}\vec{\tau} - \vec{\nabla} v\|_{\Omega_h} \leq \alpha_0^{-1}\|\vec{\tau}\|_{H(\text{div},\Omega_h)} + \|v\|_{H^1(\Omega_h)}$$
$$\|\vec{\nabla}\cdot\vec{\tau}\|_{\Omega_h} \leq \|\vec{\tau}\|_{H(\text{div},\Omega_h)}.$$

Combining these estimates for each of the terms in the optimal norm, the upper bound is proved. The result now follows from the abstract conclusion of Theorem 2.1. □

5. An extension

To show the potential of generalizing the above described technique of analysis to other problems, we now quickly describe the modifications needed to analyze the convection-diffusion problem. One of the major considerations in schemes for the convection-diffusion problem is "robustness" with respect to vanishing diffusion. While we have addressed this for the one-dimensional convection-diffusion problem in [14], the question of showing robustness of the DPG method remains open in higher dimensions. Nonetheless, below we will indicate how to show well-posedness and stability of the DPG scheme with diffusion-dependent constants. Even this is by no means obvious as we must prove an inf-sup condition with a *mesh independent* constant.

The boundary value problem under consideration now, in place of (7), is

$$-\vec{\nabla}\cdot(\alpha\vec{\nabla} u) - \vec{\beta}\cdot\vec{\nabla} u = f \qquad \text{on } \Omega \tag{45a}$$
$$u = 0 \qquad \text{on } \partial\Omega, \tag{45b}$$

where α is as before, and $\vec{\beta} : \Omega \mapsto \mathbb{R}^N$ represents the convection vector field satisfying $\vec{\nabla}\cdot\vec{\beta} = 0$. We can write it as a first order system similar to (9): Eq. (9a) remains the same, while (9b) is replaced by $\vec{\nabla}\cdot(\vec{\sigma} - \vec{\beta} u) = 0$. From this, a DPG ultra-weak formulation can be derived as before. It reads the same as (11), with the same spaces U and V, but with the modified forms

$$b(\ (\vec{\sigma}, u, \hat{u}, \hat{\sigma}_n),\ (\vec{\tau}, v)\ ) = (\alpha^{-1}\vec{\sigma}, \vec{\tau})_{\Omega_h} - (u, \vec{\nabla}\cdot\vec{\tau})_{\Omega_h} + \langle \hat{u}, \vec{\tau}\cdot\vec{n}\rangle_{\partial\Omega_h} - (\vec{\sigma} - \vec{\beta} u, \vec{\nabla} v)_{\Omega_h} + \langle v, \hat{\sigma}_n\rangle_{\partial\Omega_h},$$
$$l(\vec{\tau}, v) = (f, v)_{\Omega_h}.$$

Now $\hat{\sigma}_n$ represents an approximation of the total flux $(\vec{\sigma} - \vec{\beta} u)\cdot\vec{n}$ across element interfaces. Note that the stiffness matrix of the DPG scheme is symmetric and positive definite [14, § II] even though the original convection-diffusion operator is not.

With this setting we wish to apply the abstract result of Theorem 2.1. The first step is to prove uniqueness, i.e., verify the first assumption (5). This proceeds very much like the proof of Lemma 4.1, so we omit the details. To verify the second assumption (6), we need the optimal test norm, which is easy to calculate:

$$\|(\vec{\tau}, v)\|^2_{\mathrm{opt},V} = \|\alpha^{-1}\vec{\tau} - \vec{\nabla} v\|^2_{\Omega_h} + \|\vec{\nabla}\cdot(\vec{\tau} - \vec{\beta} v)\|^2_{\Omega_h} + \left\|[\vec{\tau}\cdot\vec{n}]\right\|^2_{\partial\Omega_h} + \left\|[v\vec{n}]\right\|^2_{\partial\Omega_h}.$$

The proof of the upper bound in the assumption (6) proceeds very similarly to the Poisson case. The only major difference in the entire analysis is the proof of the lower bound, i.e., the inf-sup condition.

The proof of the inf-sup condition will proceed as in Theorem 4.2 provided we have analogues of Lemmas 4.3 and 4.4. We develop these below. From now on we use $C_{\alpha,\beta}$ to denote a generic constant independent of Ω_h but dependent on α and $\vec{\beta}$. Its value may differ at different occurrences.

Lemma 5.1 (Modification of Lemma 4.3)**.** *Let $\vec{\tau}_0 \in H(\mathrm{div}, \Omega_h)$ and $v_0 \in H^1(\Omega_h)$ satisfy*

$$\alpha^{-1}\vec{\tau}_0 - \vec{\nabla} v_0 = 0, \qquad \text{on } K, \tag{46a}$$

$$\vec{\nabla}\cdot(\vec{\tau}_0 - \vec{\beta} v_0) = 0 \qquad \text{on } K, \tag{46b}$$

for every element K in Ω_h. Then

$$\|\vec{\tau}_0\|_\Omega + \|\vec{\nabla} v_0\|_{\Omega_h} \le C_{\alpha,\beta}\left(\left\|[\vec{\tau}_0\cdot\vec{n}]\right\|_{\partial\Omega_h} + \left\|[v_0\vec{n}]\right\|_{\partial\Omega_h}\right) \tag{47}$$

Proof. We consider only the three-dimensional case (as the two-dimensional one is simpler). Instead of the Helmholtz decomposition (appearing in the proof of Lemma 4.3), we now use the following decomposition:

$$\vec{\tau}_0 = (\alpha\vec{\nabla}\psi + \vec{\beta}\psi) + \vec{\nabla}\times\vec{z} \tag{48}$$

for a ψ in $H^1_0(\Omega)$ and $\vec{z}$ in $H(\mathrm{curl}, \Omega)$.

To see that such a decomposition exists, we first find ψ in $H^1_0(\Omega)$ satisfying

$$(\alpha\vec{\nabla}\psi, \vec{\nabla}\phi)_\Omega + (\vec{\beta}\psi, \vec{\nabla}\phi)_\Omega = (\vec{\tau}_0, \vec{\nabla}\phi)_\Omega, \qquad \forall\phi \in H^1_0(\Omega). \tag{49}$$

Let us check that there is a unique solution to this variational equation. By integration by parts, $(\vec{\beta}\phi, \vec{\nabla}\phi)_\Omega = -(\vec{\nabla}\cdot(\vec{\beta}\phi), \phi)_\Omega = -(\vec{\beta}\phi, \vec{\nabla}\phi)_\Omega$, hence $(\vec{\beta}\phi, \vec{\nabla}\phi)_\Omega$ must vanish. Hence

$$(\alpha\vec{\nabla}\phi, \vec{\nabla}\phi)_\Omega + (\vec{\beta}\phi, \vec{\nabla}\phi)_\Omega = \|\vec{\nabla}\phi\|^2_\alpha$$

so the bilinear form in (49) is coercive. Therefore, by the Lax-Milgram lemma, (49) is uniquely solvable and moreover,

$$\|\psi\|_{H^1(\Omega)} \le C_{\alpha,\beta}\|\vec{\tau}_0\|_\Omega. \tag{50}$$

To complete the proof of the existence of the decomposition (48), it suffices to note that (49) implies that the distributional divergence $\vec{\nabla}\cdot(\alpha\vec{\nabla}\psi - \vec{\beta}\psi) - \vec{\tau}_0) = 0$, hence there exists [19] a $\vec{z}$ in $H(\mathrm{curl}, \Omega)$ such that $\vec{\nabla}\times\vec{z} = (\alpha\vec{\nabla}\psi - \vec{\beta}\psi) - \vec{\tau}_0$, and

$$\|\vec{\nabla}\times z\|_\Omega \le \|\alpha\vec{\nabla}\psi - \vec{\beta}\psi\|_\Omega + \|\vec{\tau}_0\|_\Omega \le C_{\alpha,\beta}\|\vec{\tau}_0\|_\Omega, \tag{51}$$

where we have used (50). Thus the decomposition in (48) exists and is stable.

Next, let us use (48) to bound $\vec{\tau}_0$.

$$\begin{aligned}\|\vec{\tau}_0\|_{1/a}^2 &= (\alpha^{-1}\vec{\tau}_0, (\alpha\vec{\nabla}\psi + \vec{\beta}\psi) + \vec{\nabla}\times\vec{z})_\Omega && \text{by (48)}\\ &= (\vec{\tau}_0, \vec{\nabla}\psi)_{\Omega_h} + (\vec{\nabla}v_0, \vec{\beta}\psi)_{\Omega_h} + (\vec{\nabla}v_0, \vec{\nabla}\times\vec{z})_\Omega && \text{by (46a)}\\ &= -(\vec{\nabla}\cdot(\vec{\tau}_0 - \vec{\beta}v_0), \psi)_{\Omega_h} + \langle\psi, \vec{\tau}_0\cdot\vec{n}\rangle_{\partial\Omega_h} + \langle v_0, \vec{n}\cdot\vec{\nabla}\times\vec{z}\rangle_{\partial\Omega_h}\end{aligned}$$

by local integration by parts. The first term on the right hand side vanishes by (46b). The remaining two can be handled in exactly the same way as in the proof of Lemma 4.3. The only difference is that we must now use (50) and (51) in place of (36). □

Lemma 5.2 (Modification of Lemma 4.4)**.** *Let $\vec{G} \in L^2(\Omega)^N$ and $F \in L^2(\Omega)$. There is a $\vec{\tau}_1$ in $H(\mathrm{div}, \Omega)$ and v_1 in $H_0^1(\Omega)$ satisfying*

$$\alpha^{-1}\vec{\tau}_1 - \vec{\nabla}v_1 = \vec{G}, \qquad \text{on } \Omega, \tag{52a}$$

$$\vec{\nabla}\cdot(\vec{\tau}_1 - \vec{\beta}v_1) = F \qquad \text{on } \Omega, \tag{52b}$$

and

$$\|\vec{\tau}_1\|_\Omega + \|\vec{\nabla}v_1\|_\Omega \le C_{\alpha,\beta}\left(\|\vec{G}\|_\Omega + \|F\|_\Omega\right). \tag{53}$$

Proof. By the same argument as in proof of Lemma 5.1 (cf. (49)), there is a unique v_1 in $H_0^1(\Omega)$ satisfying

$$(\alpha\vec{\nabla}v_1, \vec{\nabla}w)_\Omega - (\vec{\beta}v_1, \vec{\nabla}w)_\Omega = -(\alpha\vec{G}, \vec{\nabla}w)_\Omega - (F, w)_\Omega \quad \forall w \in H_0^1(\Omega).$$

Setting $\vec{\tau}_1 = \alpha\vec{\nabla}v_1 + \alpha\vec{G}$ it is easy to see that $\vec{\tau}_1$ and v_1 satisfies (52). The same argument leading to (50) gives the bound

$$\|v_1\|_{H^1(\Omega)} \le C_{\alpha,\beta}(\|\vec{G}\|_\Omega + \|F\|_\Omega).$$

The norm $\|\vec{\tau}_1\|_\Omega$ is also bounded by the same because of (52a). □

With these modified lemmas, an analogue of Theorem 4.1 for the DPG approximation of the convection-diffusion problem can be easily proved along the lines of § 4.4.

6. Numerical experiments

We conducted numerical experiments using a code built with modules from an existing software package [10]. Partly to be able to fit into the input paramaters of the software, we modified the previously presented DPG method. (All modifications are listed below.) We now report the performance of this modified method under h and p refinements. Since our eventual goal is to apply the DPG method with fully automatic hp-adaptivity to more complex problems, we will also report results from an hp-adaptive algorithm (although a convergence theory for this adaptive algorithm is yet to be developed).

6.1. **Practical settings.**

For all our experiments, we will consider two-dimensional domains Ω subdivided into either fully geometrically conforming or 1-irregular quadrilateral meshes. Let Ω_h denote the collection of mesh elements as before, while $\mathcal{E}_h$ denote the collection of mesh edges. An element edge with a hanging node is considered as two separate edges. To each mesh element K is associated a polynomial degree $p_K \ge 1$ and to each

mesh edge E the degree p_E. The practical trial space U_h is set to $U_h = S_h \times W_h \times M_h \times Q_h$, where

$$W_h = \{v : v|_K \in \mathcal{Q}_{p_K,p_K}(K),\ \forall K \in \Omega_h\}, \tag{54a}$$
$$S_h = W_h \times W_h, \tag{54b}$$
$$M_h = \{\mu : \mu|_E \in P_{p_E}(E),\ \forall E \in \mathcal{E}_h \text{ and } \mu|_{\partial\Omega} = 0\}, \tag{54c}$$
$$Q_h = \{\eta : \eta|_E \in P_{p_E}(E),\ \forall E \in \mathcal{E}_h\}, \tag{54d}$$

where $\mathcal{Q}_{l,m}(K)$ is the space of bivariate polynomials which are of degree at most l in x and at most m in y.

At this point, we should note some discrepancies between the method we theoretically analyzed and the method we practically implement.

(1) The first "variational crime" we commit in our implementation is to let one of the component spaces in U_h be a nonconforming subspace of U-component. Clearly, the M_h in (54c) contains functions that are discontinuous globally on $\cup_K \partial K$ and hence $M_h \not\subset H_0^{1/2}(\partial\Omega_h)$. We chose the above defined M_h solely to fit within the available parameters of the software package [10].

(2) The polynomial degree used for M_h and Q_h are the same, although Corollary 4.1 suggests we should use one less degree in Q_h. Again, this is done only to fit within the input parameters of the software.

(3) Recall that the test space is determined by T, defined by (3). In implementations, in place of T, we use $\tilde{T} : U \mapsto \tilde{V}$ defined by

$$(\tilde{T}u, \tilde{v})_V = b(u, \tilde{v}), \qquad \forall\, \tilde{v} \in \tilde{V},$$

where $\tilde{V}$ is the finite dimensional subspace of V defined by

$$\tilde{V} = \{(\vec{\tau}, v) : \vec{\tau}|_K \in \mathcal{Q}^{\mathrm{div}}_{\tilde{p}_K} \text{ and } v|_K \in \mathcal{Q}_{\tilde{p}_K,\tilde{p}_K}(K)\} \qquad \text{and} \quad \tilde{p}_K = p_K + \delta p.$$

We set the *enrichment degree* δp to be the same for all mesh elements. Here, $\mathcal{Q}^{\mathrm{div}}_\ell(K) \stackrel{\mathrm{def}}{=} \mathcal{Q}_{\ell+1,\ell} \times \mathcal{Q}_{\ell,\ell+1}$ is a well known subspace of $H(\mathrm{div}, K)$, often called a Nédélec space of the first kind [21] or the Raviart-Thomas space [22] on squares.

In our earlier numerical experience [14, 15] on various equations, we found that $\delta p = 2$ is sufficient to obtain good results, so this will form our default choice. But below we will also report results with other choices.

The degree of polynomials approximating the numerical fluxes as well as numerical traces are, by default, set by the *maximum rule*. To describe this, first note that a while mesh edge can be shared by at most two mesh element in conforming meshes, on 1-irregular meshes, we consider edges split by a hanging node as shared by three adjacent mesh elements. For any edge E, let $\hat{p}_E$ denote the maximum of the degrees p_K for the two or the three adjacent elements K. When using the maximum rule we typically set p_E in (54) to $\hat{p}_E$, but to investigate the dependence of the errors with edge degrees, we will conduct experiments with

$$p_E = \hat{p}_E + \delta p_F \tag{55}$$

for some choices of integers δp_F (set equal for all edges E).

We will often report the *energy norm* of the error, in addition to the L^2-norm of the error. The energy norm [14] in the DPG framework is

$$\|u\|_E = \sup_{v \in V} \frac{|b(u,v)|}{\|v\|_V}.$$

It is easy to see [23] that whenever (6) holds, $C_1\|u\|_U \leq \|u\|_E \leq C_2\|u\|_U$. We will verify the practical manifestation of this equivalence by comparing the errors in the energy norm and the L^2-norm.

Finally, in all our adaptive schemes, we use (an approximation of) the energy norm of the error as the error indicator. To describe how we compute it, first define the *error representation* function $\tilde{e} \in \tilde{V}$ by $\tilde{e} = \tilde{T}(u - u_h)$. Since $(\tilde{T}(u-u_h), \tilde{v})_V = b(u-u_h, \tilde{v}) = l(\tilde{v}) - b(u_h, \tilde{v})$, the error representation function can be computed element by element by solving

$$(\tilde{e}, \tilde{v})_V = l(\tilde{v}) - b(u_h, \tilde{v}), \qquad \forall\ \tilde{v} \in \tilde{V}.$$

The energy norm of the error is then approximated by

$$\|u - u_h\|_E = \|T(u-u_h)\|_V \approx \|\tilde{T}(u-u_h)\|_V = \|\tilde{e}\|_V. \tag{56}$$

The contribution to $\|\tilde{e}\|_V^2$ from each element forms the *element error indicator*. Note that for the Poisson example, $\tilde{e}$ represents the error in *all* the four variables (u, $\vec{\sigma}$, $\hat{u}$ and $\hat{\sigma}_n$).

6.2. **Numerical Examples.** We consider the following two-dimensional examples. The first example is on the *unit square* i.e., we set $\Omega = (0,1) \times (0,1)$ and solve

$$\begin{cases} -\Delta u = f, & \text{on } \Omega, \\ u = 0, & \text{on } \partial\Omega, \end{cases}$$

with $f = 2\pi^2 \sin(\pi x)\sin(\pi y)$, so that the exact solution $u = \sin(\pi x)\sin(\pi y)$ is infinitely smooth. Initially, Ω_h is a uniform mesh of four congruent square elements with $p_K = 2$ for all $K \in \Omega_h$.

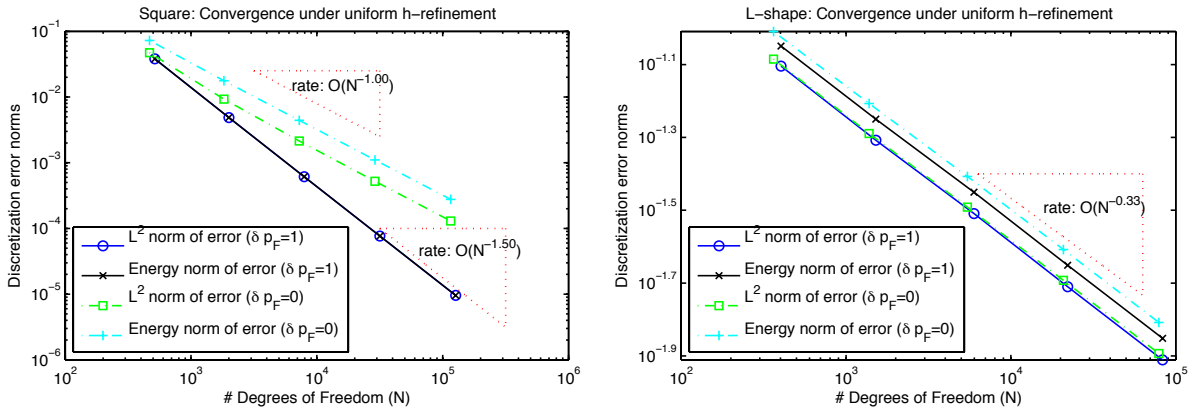
The second example involves an *L-shaped domain*, a classical test domain for adaptivity, namely $\Omega = (-1,1) \times (-1,1) \setminus [-1,0] \times [-1,0]$. We solve

$$\begin{cases} -\Delta u = 0, & \text{on } \Omega, \\ u = 0, & \text{on the edges of } \partial\Omega \text{ along } x \text{ and } y \text{ axes}, \\ \dfrac{\partial u}{\partial n} = g, & \text{on the remainder of the boundary } \partial\Omega. \end{cases}$$

The Neumann data g is set so that the exact solution is

$$u(r,\theta) = r^{2/3} \sin\big(\frac{2}{3}(\theta + \frac{\pi}{2})\big).$$

The derivatives of solution u are singular at the origin. It is well known that the solution u is in $H^{1+s}(\Omega)$ for all $s < 2/3$. The initial mesh Ω_h consists of three congruent squares with $p_K = 2$ for all three elements.



(a) The square case (b) The case of the L-shaped domain

FIGURE 1. h-convergence rates for the two examples

6.3. **Convergence rates in h and p.** The observed rates of convergence under uniform mesh refinement, holding the degree p fixed to 2 for all elements, is reported in Figure 1. The L^2 and energy norms of the errors versus degrees of freedom are plotted in these figures. When we report the "L^2-norm" of the error, we only include $\|u - u_h\|_{L^2(\Omega)}$ and $\|\vec{\sigma} - \vec{\sigma}_h\|_{L^2(\Omega)}$. We do not include the errors in numerical flux or trace.

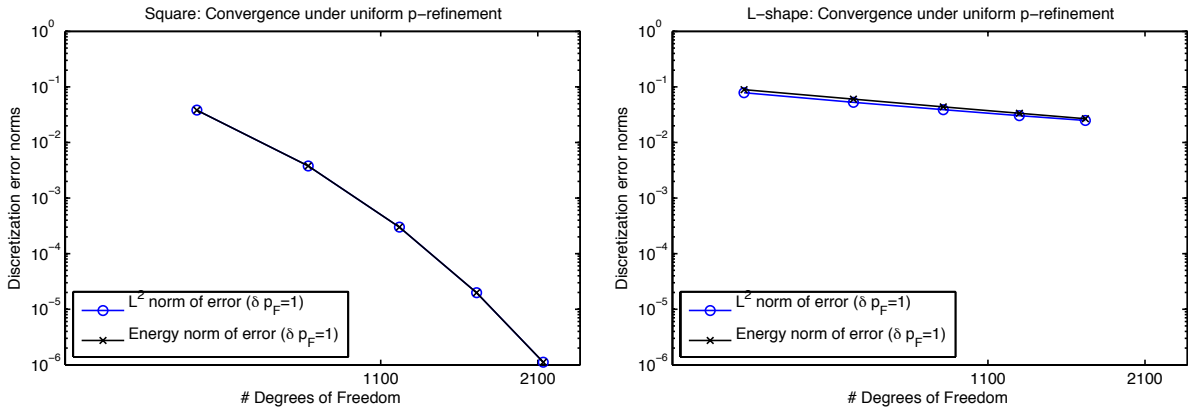
Under uniform h-refinement, the number of degrees of freedom N is $O(h^{-2})$. Thus, from Figure 1(a), we find that the observed h-convergence rates for the square domain are $O(N^{-1.5}) = O(h^3) = O(h^{p+1})$ when $\delta p_F = 1$, while it reduces to $O(N^{-1.0}) = O(h^2) = O(h^p)$ when $\delta p_F = 0$. This is indeed in accordance with Theorem 4.1 and the well-known best approximation rates for the square elements we used. In particular, it seems necessary, both theoretically and practically, to use higher order polynomials for numerical traces for optimal convergence rates.

Results from the L-shaped domain are in Figure 1(b). Here, we observe that the convergence rate under uniform h-refinement is $\approx O(N^{-1/3}) = O(h^{2/3})$. This is also in accordance with Theorem 4.1. The observed rates are the same for $\delta p_F = 0$ and $\delta p_F = 1$ due to the limited regularity of the solution.

To study p-convergence, we increased the polynomial degree uniformly on all elements, holding the mesh fixed (to the initial mesh). The results are in Figure 2. While the exponential convergence rate is evident (Figure 2(a)) for the infinitely smooth solution on the square, the p-convergence rate is limited (Figure 2(b)) for the less regular solution on the L-shaped domain.

6.4. **Results from the adaptive scheme.** We used the standard "greedy" strategy for h-adaptive refinements, i.e., all elements which contribute within 25% of the maximum element contribution to the square of total error in energy norm, are marked for the refinement. For hp-refinements, we have used Mark Ainsworth's flagging strategy: all elements adjacent to the origin (the singularity) are h-refined, while the remaining ones are p-refined. The error estimator, in both cases, is the previously mentioned energy error – see (56).

The results are in Figure 3. The first graph in Figure 3(a) compares the error reduction obtained by uniform h-refinement, adaptive h-refinement, and the above mentioned



(a) Results from the square domain (b) Results from the L-shaped domain

FIGURE 2. p-convergence rates for the two examples

adaptive hp-refinement. Clearly, the hp-adaptive strategy is superior. As expected, h-adaptivity restores the optimal rate of convergence for quadratic elements ($N^{-1.5}$) while the hp-adaptive strategy delivers exponential convergence despite the singularity of the solution.

In Figure 3(b), we examine the effectiveness of the error indicator. Comparing the energy error with the L^2-error, we find the two curves follow each other with a ratio between them close to unity. Note that the same behavior was also seen in Figures 1(a), 1(b), 2(b) and 2(a).

Next, we consider secondary effect of approximation of optimal test functions. The effect of approximating T by $\tilde{T}$ can be studied by varying the enrichment degree δp. Results from the h-adaptive algorithm applied to the L-shaped domain with varying δp are presented in Figure 3(c). The curves are practically identical for $\delta p = 2, 3$, and 4.

We also report the effect of varying the degrees of numerical fluxes and traces, as measured by δp_F, in the context of the same h-adaptive example. Figure 3(d) shows that when $\delta p_F = 0$, the energy error indicator seems to deviate from the optimal curve as the number of degrees of freedom increase.

Finally, as an illustration, Figure 4 shows the optimal hp mesh obtained after 15 refinements with the relative L^2-error less than 0.016%. This includes the error in u and its derivatives. The computed function u is also shown in the figure.

REFERENCES

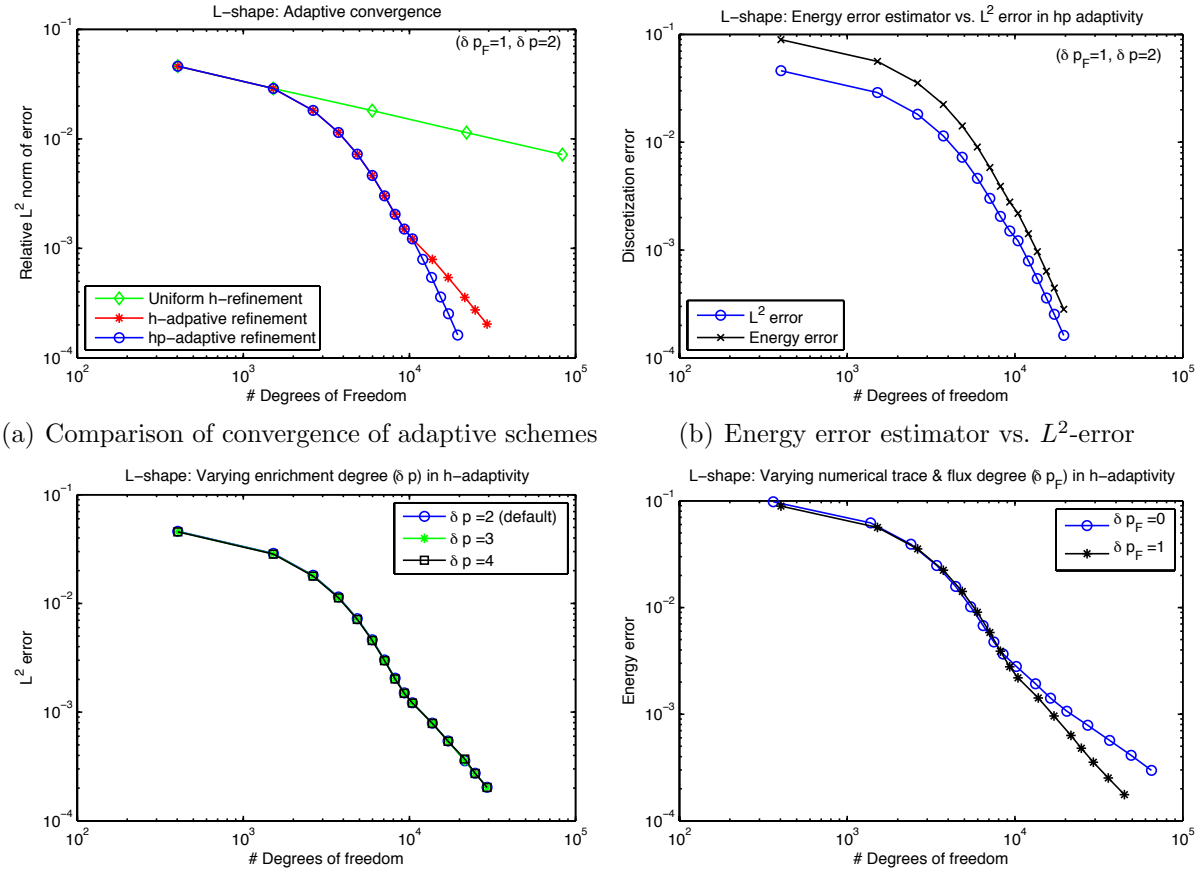
[1] D. N. ARNOLD, F. BREZZI, B. COCKBURN, AND L. D. MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2001/02), pp. 1749–1779 (electronic).

[2] G. A. BAKER, *Finite element methods for elliptic equations using nonconforming elements*, Math. Comp., 31 (1977), pp. 45–59.

[3] 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.

[4] S. C. BRENNER, *Poincaré-Friedrichs inequalities for piecewise H^1 functions*, SIAM J. Numer. Anal., 41 (2003), pp. 306–324 (electronic).

[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.



(a) Comparison of convergence of adaptive schemes

(b) Energy error estimator vs. L^2-error

(c) Effect of varying δp

(d) Effect of varying δp_F

FIGURE 3. Convergence curves from adaptive schemes

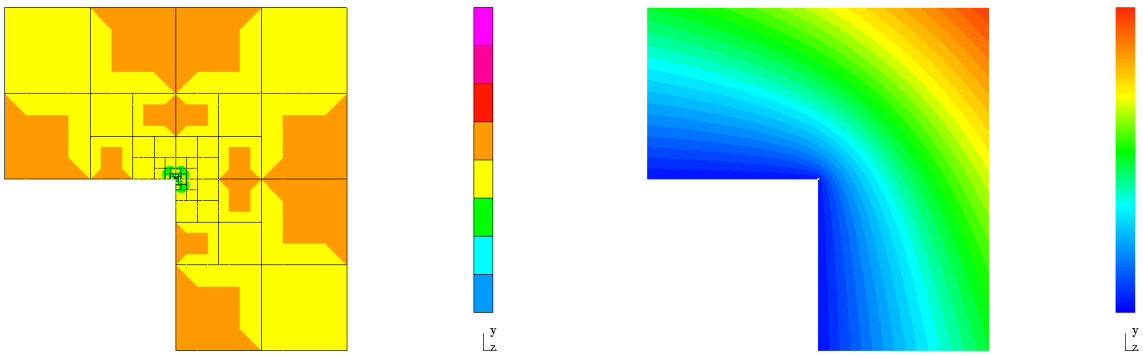


FIGURE 4. Left: The *hp* mesh found by the *hp*-adaptive algorithm after 15 refinements. (Color scale represents polynomial degrees.) Right: The corresponding solution u. (Color scale represent solution values.)

[6] P. Castillo, B. Cockburn, I. Perugia, and D. Schötzau, *An a priori error analysis of the local discontinuous Galerkin method for elliptic problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1676–1706 (electronic).
[7] 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).
[8] 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.
[9] 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).
[10] L. Demkowicz, *Computing with hp-adaptive finite elements. Vol. 1*, Chapman & Hall/CRC Applied Mathematics and Nonlinear Science Series, Chapman & Hall/CRC, Boca Raton, FL, 2007. One and two dimensional elliptic and Maxwell problems, With 1 CD-ROM (UNIX).
[11] L. F. Demkowicz, *Polynomial exact sequences and projection-based interpolation with application to maxwell equations*, in Mixed finite elements, compatibility conditions, and applications pp. x+235. Lectures given at the C.I.M.E. Summer School held in Cetraro, June 26–July 1, 2006, Edited by Boffi and Lucia Gastaldi.
[12] L. Demkowicz and A. Buffa, *H^1, H(curl) and H(div)-conforming projection-based interpolation in three dimensions. Quasi-optimal p-interpolation estimates*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 267–296.
[13] L. Demkowicz and J. Gopalakrishnan, *A class of discontinuous Petrov-Galerkin methods. Part I: The transport equation*, Computer Methods in Applied Mechanics and Engineering, 199 (2010), pp. 1558–1572.
[14] ——, *A class of discontinuous Petrov-Galerkin methods. Part II: Optimal test functions*, To appear in Numerical Methods for Partial Differential Equations, (2010).
[15] L. Demkowicz, J. Gopalakrishnan, and A. Niemi, *A class of discontinuous Petrov-Galerkin methods. Part III: Adaptivity*, Submitted, (2010).
[16] L. Demkowicz, J. Gopalakrishnan, and J. Schöberl, *Polynomial extension operators. Part I*, SIAM J. Numer. Anal., 46 (2008), pp. 3006–3031.
[17] ——, *Polynomial extension operators. Part II*, SIAM J. Numer. Anal., 47 (2009), pp. 3293–3324.
[18] ——, *Polynomial extension operators. Part III*, Submitted, (2009).
[19] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, no. 5 in Springer series in Computational Mathematics, Springer-Verlag, New York, 1986.
[20] T. Huttunen, P. Monk, and J. P. Kaipio, *Computational aspects of the ultra-weak variational formulation*, J. Comput. Phys., 182 (2002), pp. 27–46.
[21] J.-C. Nédélec, *Mixed Finite Elements in $\mathbb{R}^3$*, Numer. Math., 35 (1980), pp. 315–341.
[22] P.-A. Raviart and J. M. Thomas, *A mixed finite element method for 2nd order elliptic problems*, in Mathematical aspects of finite element methods (Proc. Conf., Consiglio Naz. delle Ricerche (C.N.R.), Rome, 1975), Springer, Berlin, 1977, pp. 292–315. Lecture Notes in Math., Vol. 606.
[23] J. Zitelli, I. Muga, L. Demkowicz, J. Gopalakrishnan, D. Pardo, and V. Calo, *A class of discontinuous Petrov-Galerkin methods. Part IV: Wave propagation*, Submitted, (2010).

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

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