MATHEMATICS OF COMPUTATION
Volume 82, Number 284, October 2013, Pages 1923–1952
S 0025-5718(2013)02697-X
Article electronically published on April 23, 2013

CONSTRUCTIVELY WELL-POSED APPROXIMATION METHODS WITH UNITY INF–SUP AND CONTINUITY CONSTANTS FOR PARTIAL DIFFERENTIAL EQUATIONS

TAN BUI-THANH, LESZEK DEMKOWICZ, AND OMAR GHATTAS

Abstract. Starting from the generalized Lax-Milgram theorem and from the fact that the approximation error is minimized when the continuity and inf–sup constants are unity, we develop a theory that provably delivers well-posed approximation methods with unity continuity and inf–sup constants for numerical solution of linear partial differential equations. We demonstrate our single-framework theory on scalar hyperbolic equations to constructively derive two different *hp* finite element methods. The first one coincides with a least squares discontinuous Galerkin method, and the other appears to be new. Both methods are proven to be trivially well-posed, with optimal *hp*-convergence rates. The numerical results show that our new discontinuous finite element method, namely a discontinuous Petrov-Galerkin method, is more accurate, has optimal convergence rate, and does not seem to have nonphysical diffusion compared to the upwind discontinuous Galerkin method.

1. Introduction

The work in this paper is inspired from the recent research on the discontinuous Petrov–Galerkin method (DPG) of Demkowicz and Gopalakrishnan [1, 2, 3] for the numerical solution of partial differential equations. The method starts by partitioning the domain of interest into nonoverlapping elements. Variational formulations are posed for each element separately and then summed up to form a global variational statement. Elemental solutions are connected by introducing hybrid variables (also known as fluxes or traces) that live on the skeleton of the mesh. This is therefore a mesh-dependent variational approach in which both bilinear and linear forms depend on the mesh under consideration.

In general, the trial and test spaces are not related to each other. In the standard Bubnov–Galerkin (also known as Galerkin) approach, the trial and test spaces are identical, while they differ in a Petrov–Galerkin scheme. Traditionally, one chooses either Galerkin or Petrov–Galerkin approaches, then proves the consistency and stability in both infinite and finite dimensional settings. The DPG method introduces a new paradigm in which one selects both trial and test spaces at the same time to satisfy well-posedness. In particular, one can select trial and test function spaces for which the continuity and inf–sup constants are unity. Given a finite dimensional trial subspace, the finite dimensional test space is constructed in

Received by the editor April 18, 2011 and, in revised form, October 31, 2011.
2010 *Mathematics Subject Classification.* Primary 65N30; Secondary 65N12, 65N15, 65N22.
Key words and phrases. Inf-sup condition, inf-sup constant, generalized Lax-Milgram theorem, discontinuous Galerkin method, discontinuous Petrov-Galerkin method, consistency, finite element method, stability, convergence, well-posedness hyperbolic partial differential equations.

such a way that the well-posedness of the finite dimensional setting is automatically inherited from the infinite dimensional counterpart.

The DPG method in [2] starts with a given norm in the trial space and then seeks a norm in the test space in order to achieve unity continuity and inf–sup constants. Another DPG method in [3] achieves the same goal but reverses the process, i.e., it looks for a norm in the trial space corresponding to a given norm in the test space. Clearly, this is one of the advantages of the DPG methodology, since it allows one to choose a norm of interest to work with, while rendering the error optimal, i.e., smallest in that norm. Furthermore, the DPG methods have been shown to be robust and provide optimal *hp*-convergence. We shall not discuss the advantages of the DPG methods any further here, and the readers are referred to the original DPG papers [1, 2, 3] for more details.

Here, we pursue a different option. In particular, we shall not prescribe norms in either the trial and test spaces. Instead, we permit the structure of the problem to determine its "natural" energy norms. Our goal is to develop a single framework that adapts to the problem at hand, while automatically generating accurate finite element methods with trivially-proven and guaranteed stability. Similar to other DPG methods, once one chooses the test or trial basis function, the other has to be solved for. Depending on how one applies our theory to problems under consideration, basis functions can be trivially obtained or solved for through an adjoint partial differential equation, as we shall show.

The remainder of the paper is organized as follows. In Section 2, we first develop a single framework for general variational problems that can be written in terms of bilinear and linear forms. Then, we develop a few analytical results that help us prove the optimality of the resulting finite dimensional approximations and their well-posedness. In Sections 3 and 4, we apply our single framework, but using two different points of view, to linear hyperbolic equations of advection-reaction type. We shall show that the framework constructively leads to an existing stabilized *hp*-discontinuous Galerkin (DG) method and a new *hp*-DPG method for advection-reaction equations. We discuss characteristics of each method and their *hp*-convergence in detail. The chief purpose of the paper is to introduce our theory and to demonstrate its usefulness to partial differential equations in deriving accurate and stable finite element methods. We further strengthen our findings by several one- and two-dimensional numerical examples in Section 5, and Section 6 concludes the paper.

2. Abstract theory development

In this section, we develop a theory for constructive approximations of linear partial differential equations. Our goal is to construct finite dimensional approximations that are guaranteed to be trivially well-posed with unity continuity and inf–sup constants. Here, trivial well-posedness means that the well-posedness of the finite dimensional problems is trivially inherited from their infinite dimensional counterparts. The starting point of our theory, to be shown, is not new since our work is inspired by the recent discontinuous Petrov–Galerkin (DPG) methodology of Demkowicz and Gopalakrishnan [1, 2, 3]. However, we shall point out the differences between our approach and the existing DPG methods.

Let U and V be Hilbert spaces over the real line (generalization of our theory to the complex field is straightforward). Consider the following variational problem:

$$\begin{cases} \text{Seek } u \in U \text{ such that} \\ a(u,v) = \ell(v), \quad \forall v \in V, \end{cases} \tag{1}$$

where $\ell(\cdot)$ is a linear form on V, $a(\cdot,\cdot)$ is a bilinear form satisfying the continuity condition with continuity constant M,

$$|a(u,v)| \leq M \|u\|_U \|v\|_V, \tag{2}$$

the inf–sup condition with the inf–sup constant γ,

$$\exists \gamma > 0 : \inf_{u \in U} \sup_{v \in V} \frac{a(u,v)}{\|u\|_U \|v\|_V} \geq \gamma, \tag{3a}$$

and the injectivity of the adjoint operator (to be defined),

$$(a(u,v) = 0, \quad \forall u \in U) \Rightarrow (v = 0). \tag{3b}$$

If (2) and (3) hold, then by the generalized Lax-Milgram theorem [4, 5] (also known as the Banach-Nečas-Babuška theorem [6]), (1) has a unique solution and the solution is stable in the following sense:

$$\|u\|_U \leq \frac{1}{\gamma} \|\ell\|_{V'},$$

where V' is the topological dual of V. Note that for convenience in writing, we have abused the notation $\sup_{v \in V}$ instead of $\sup_{v \in V, v \neq 0}$ (and similarly for inf).

Now let $U_h^n \subset U$ and $V_h^n \subset V$ be two finite dimensional trial and test spaces, and consider the following finite dimensional approximation problem:

$$\begin{cases} \text{Seek } u_h \in U_h^n \text{ such that} \\ a(u_h, v_h) = \ell(v_h), \quad \forall v_h \in V_h^n. \end{cases} \tag{4}$$

If $\dim U_h = \dim V_h = n$, and the following discrete inf–sup condition

$$\exists \gamma_h > 0 : \inf_{u_h \in U_h} \sup_{v_h \in V_h} \frac{a(u_h, v_h)}{\|u_h\|_U \|v_h\|_V} \geq \gamma_h \tag{5}$$

holds, then the finite dimensional problem (4) is well-posed by an application of the generalized Lax-Milgram theorem for finite dimensional problems (also known as the Babuška's Theorem [4, 7]). In general, however, the finite dimensional problem (4) does not inherit the well-posedness of the infinite dimensional counterpart (1) except for some special circumstances. One, therefore, has to prove the nontrivial discrete inf–sup condition [6].

In this paper, we constructively develop a class of finite dimensional approximations in which the discrete inf–sup condition (5) trivially follows from the continuous one (3a). Before doing so, let us first discuss the approximation error between the finite and infinite dimensional solutions. To begin, we recall the following projection result in which the norms of the projection and its complement are equal.

Lemma 2.1. *Let $U_h \subset U$ be a subspace of a Hilbert space U. Suppose $P : U \to U_h$, is a projection, i.e., $P^2 = P$, and P is not null or identity. Then*

$$\|P\| = \|I - P\|,$$

where the norm $\|P\|$ is induced from a norm in U, which is in turn induced from an inner product in U.

Proof. Many proofs of this result can be found in [8]. □

If we define P as $u_h = Pu$ through

$$a(u_h, v_h) = a(u, v_h), \quad \forall v_h \in V_h,$$

then, by the discrete inf–sup condition (5), it is trivial to see that P is a projection operator. In addition, we can bound the norm of the projection operator P as follows.

Lemma 2.2. *There holds* $\|P\| \leq \frac{M}{\gamma_h}$.

Proof. See [9] for a simple proof. □

Now comes the projection error result.

Theorem 2.3 (Babuška [7])**.** *Suppose that both the continuous problem* (1) *and discrete problem* (4) *are well-posed, then*

$$\|u - u_h\|_U \leq \frac{M}{\gamma_h} \inf_{w_h \in U_h} \|u - w_h\|_U .$$

Proof. A standard proof that uses Lemmas 2.1 and 2.2 can be found in [10, 9]. □

The following best approximation error result immediately follows from Theorem 2.3.

Corollary 1. *If* $M = \gamma_h$*, then*

$$\|u - u_h\|_U = \inf_{w_h \in U_h} \|u - w_h\|_U .$$

In particular, $M = \gamma_h = 1$ satisfies Corollary 1. That is, if the continuity constant and the discrete inf–sup constant are unity, then the error incurred from the discrete approximation (4) is the best, i.e., it is smallest. Up to this point, although not in the form presented here, the theory has already been discussed in the DPG methods [1, 2, 3]. As can be seen, there are two spaces to work with, namely the trial and test spaces, respectively. The first DPG method [2] starts with a given norm in the trial space U, and then seeks a norm in the test space V so that $M = \gamma = 1$. In the second DPG method [3], on the other hand, one defines a norm in U from a given norm in V such that $M = \gamma = 1$. Clearly, this is one of the advantages of the DPG methodology since it allows one to choose a norm of interest to work with while making the error optimal, i.e., smallest, in that norm.

In this paper, we explore a different option, namely, we shall not prescribe either norms in the spaces U and V. Instead, we let the problem determine its "natural" energy norms, thus which norms to be chosen to work with is out of the question in our new approach. Our single-framework method will be applied to linear hyperbolic equations of advection-reaction type, and we shall show that it constructively leads to an existing stabilized numerical method and a new one. Nevertheless, care must be taken since our idea may not be applicable for cases in which one prefers to work with particular norms.

The rest of this section exploits the goal of having $M = \gamma = 1$ to some extent. In particular, we characterize a few properties for problems having $M = \gamma = 1$, and then present sufficient conditions for the infinite dimensional problem to have $M = \gamma = 1$. Next, we study constructions of finite dimensional subspaces U_h^n and V_h^n such that the well-posedness of the resulting finite dimensional approximation

problems is trivially inherited from the infinite dimensional settings. In fact, as shall be shown, our method automatically delivers unity discrete continuity and inf–sup constants, i.e., $M_h = \gamma_h = 1$. It should be pointed out that the rest of our theory is general so that it can apply not only to our method in this paper, but also the other DPG methods.

We first note that the inf–sup condition (3a) is typically defined by taking first the supremum over the test space V and then the infimum over the trial space U. However, the following well-known result shows that as long as the well-posedness, and hence the continuity and inf–sup constants, is concerned, the distinction between the test and trial spaces are irrelevant. That is, there is no reason for us to favor one space over the other.

Lemma 2.4. *The problem* (1) *is well-posed if and only if*

$$\exists \gamma > 0 : \inf_{u \in U} \sup_{v \in V} \frac{a(u,v)}{\|u\|_U \|v\|_V} = \inf_{v \in V} \sup_{u \in U} \frac{a(u,v)}{\|v\|_V \|u\|_U} \geq \gamma.$$

Proof. See [9] for a proof. □

We first make the following simple observation.

Lemma 2.5. *The following are equivalent:*

(i) $M = \gamma = 1$.

(ii) $\forall u \in U$ *we have* $\|u\|_U = \sup_{v \in V} \frac{a(u,v)}{\|v\|_V}$.

(iii) $\forall v \in V$ *we have* $\|v\|_V = \sup_{u \in U} \frac{a(u,v)}{\|u\|_U}$.

Proof. By Lemma 2.4, it is sufficient to prove the equivalence of (i) and (ii).

(i) ⇒ (ii): From the continuity condition we have

$$\sup_{v \in V} \frac{a(u,v)}{\|v\|_V} \leq \|u\|_U ,$$

which, together with the inf–sup condition, implies (ii).

(ii) ⇒ (i): (ii) is equivalent to

$$\frac{a(u,v)}{\|v\|_V} \leq \sup_{v \in V} \frac{a(u,v)}{\|v\|_V} \leq \|u\|_U ,$$

$$\sup_{v \in V} \frac{a(u,v)}{\|v\|_V} \geq \|u\|_U ,$$

which implies (i). □

The following useful result will be used as guidelines to construct the "natural" norms in U and V spaces such that $M = \gamma = 1$.

Theorem 2.6. *Suppose the continuity condition holds with unity continuity constant, i.e.,*

$$a(u,v) \leq \|u\|_U \|v\|_V .$$

Then there holds $M = \gamma = 1$ *if either of the following conditions holds:*

(i) *For each* $u \in U \setminus \{0\}$, *there exists* $v_u \in V \setminus \{0\}$ *such that*

$$a(u, v_u) = \|u\|_U \|v_u\|_V .$$

(ii) *For each* $v \in V \setminus \{0\}$, *there exists* $u_v \in U \setminus \{0\}$ *such that*

$$a(u_v, v) = \|u_v\|_U \|v\|_V .$$

Proof. We shall show that (i) is a sufficient condition for $M = \gamma = 1$ to hold, and an analogous proof can be done for (ii). It is straightforward to have

$$\|u\|_U = \frac{a(u, v_u)}{\|v_u\|_V} \leq \sup_{v \in V} \frac{a(u, v)}{\|v\|_V} \leq \|u\|_U ,$$

and Lemma 2.5 concludes the proof. □

Remark 2.7. In general, the continuity and the inf–sup conditions are not related to each other, and it is typically more difficult to establish the latter. However, Theorem 2.6 shows that if the continuity constant is unity and the equality is attainable, then the continuity condition actually implies the inf–sup condition and the inf–sup constant is unity as well. This important observation is the key result in this paper and will be exploited throughout for the advection-reaction problem.

It should be pointed out that if both conditions in Theorem 2.6 hold, then the 3-tuple $(U, V, a(\cdot, \cdot))$ is known as a dual pair. That is, the bilinear form $a(\cdot, \cdot)$ puts U and V in duality.

To the end of the paper, we call the norms in U and V spaces *optimal norms* if both continuity and inf–sup constants are unity in these norms. Moreover, we also call the pair u and v_u (and hence for u_v and v) as the optimal trial and test functions, respectively. Here, optimality is in the sense of Corollary 1.

We are now in position to construct the approximation subspaces U_h^n and V_h^n such that the discrete continuity and inf–sup constants are unity.

Lemma 2.8. *Let the assumptions of Theorem* 2.6 *hold respectively for item (i) and (ii) below.*

(i) *Let* $U_h^n \subset U$ *be a subspace, and construct*

$$V_h^n = span\left\{v_{u_h} \in V : u_h \in U_h^n, a(u_h, v_{u_h}) = \|u_h\|_U \|v_{u_h}\|_V\right\}.$$

(ii) *Let* $V_h^n \subset V$ *be a subspace, and construct*

$$U_h^n = span\left\{u_{v_h} \in U : v_h \in V_h^n, a(u_{v_h}, v_h) = \|v_h\|_V \|u_{v_h}\|_U\right\}.$$

If the pair of test space V_h^n *and trial space* U_h^n *are constructed by either (i) or (ii), then there holds*

$$M_h = \gamma_h = 1,$$

and the discrete problem (4) *is well-posed if* $\dim U_h^n = \dim V_h^n$.

Proof. Again, it is sufficient to show item (i). $M_h = 1$ is a direct consequence of $M = 1$. By construction, we have, for each $u_h \in U_h^n$,

$$\|u_h\|_U = \frac{a(u_h, v_{u_h})}{\|v_{u_h}\|_V} = \sup_{v \in V} \frac{a(u_h, v)}{\|v\|_V} = \sup_{v_h \in V_h^n} \frac{a(u_h, v_h)}{\|v_h\|_V},$$

which implies $\gamma_h = 1$. The well-posedness of the discrete problem is now clear by the Babuška's theorem. □

It can be seen that Theorem 2.6 and Lemma 2.8 do not explicitly specify either the optimal test function v_u or the optimal trial function u_v. A general-purpose approach for choosing an optimal pair of functions is through the Riesz representation theorem as we now show. Moreover, if a basis of the trial space U_h^n is specified, we can determine the corresponding basis in the test space and vice versa so that the finite dimensional problem is well-posed with $M_h = \gamma_h = 1$.

Theorem 2.9. *Define the map* $T : U \ni u \mapsto Tu \in V'$ *as* $\langle Tu, v\rangle_{V'\times V} = a(u,v)$. *Denote* v_{Tu} *as the Riesz representation of* Tu *in* V. *Suppose* $a(\cdot,\cdot)$ *is continuous with unity constant and assumption (i) of Theorem* 2.6 *holds. Take* $U_h^n \subset U$ *and define*

$$V_h = span\left\{v_{Tu_h} \in V : u_h \in U_h^n\right\}.$$

Then, the following hold:

(i) $M_h = \gamma_h = 1$.

(ii) *Let* $U_h^n = span\left\{\varphi_i\right\}_{i=1}^n$, *where* $\varphi_i \in U, i = 1,\dots,n$. *Then* $\left\{v_{T\varphi_i}\right\}_{i=1}^n$ *is a basis of* V_h^n.

Proof.

(i) By Lemma 2.8, it is sufficient to show that $a(u_h, v_{Tu_h}) = \|u_h\|_U \|v_{Tu_h}\|_V$. But, by the proof of Theorem 2.6, the definition of norm in V', and the Riesz representation theorem, one readily has,

$$\|u_h\|_U = \sup_{v\in V} \frac{a(u_h,v)}{\|v\|_V} = \sup_{v\in V} \frac{\langle Tu_h, v\rangle_{V'\times V}}{\|v\|_V} = \|Tu_h\|_{V'} = \|v_{Tu_h}\|_V,$$

which yields,

$$a(u_h, v_{Tu_h}) = \langle Tu_h, v_{Tu_h}\rangle_{V'\times V} = \|v_{Tu_h}\|_V^2 = \|u\|_U \|v_{Tu_h}\|_V.$$

(ii) By virtue of the Riesz representation theorem, v_{Tu} is linear in Tu. Together with the linearity of T, we conclude that $V_h = span\left\{v_{T\varphi_i}\right\}_{i=1}^n$. It remains to prove that the set $\left\{v_{T\varphi_i}\right\}_{i=1}^n$ is independent. Assume, on the contrary, that there exists $i \in \{1,\dots,n\}$ such that $\alpha_i \neq 0$ and

$$\sum_{i=1}^n \alpha_i v_{T\varphi_i} = 0 \Rightarrow v_{T\left(\sum_{i=1}^n \varphi_i\right)} = 0,$$

which, by the injectivity of T, implies

$$\left\|\sum_{i=1}^n \alpha_i\varphi_i\right\|_U = \left\|T\left(\sum_{i=1}^n \alpha_i\varphi_i\right)\right\|_{V'} = 0,$$

which, by the independence of $\left\{\varphi_i\right\}_{i=1}^n$, in turn implies

$$\alpha_i = 0, \quad \forall i = 1,\dots,n,$$

a contradiction.

□

If we call the result in Theorem 2.9 as the primal approach, then using Lemma 2.4 we readily have the following "dual" analog.

Theorem 2.10. *Define the adjoint map* $T' : V \ni v \mapsto T'v \in U'$ *as* $\langle T'v, u\rangle_{U'\times U} = a(u,v)$. *Assume that* T' *is injective, i.e.,* (3b) *holds. Denote* $u_{T'v}$ *as the Riesz representation of* $T'v$ *in* U. *Suppose* $a(\cdot,\cdot)$ *is continuous with unity constant and assumption (ii) of Theorem* 2.6 *holds. Take* $V_h^n \subset V$ *and define*

$$U_h^n = span\left\{u_{T'v_h} \in U : v_h \in V_h^n\right\}.$$

Then, the following hold:

(i) $M_h = \gamma_h = 1$.

(ii) *Let* $V_h^n = span\left\{\phi_i\right\}_{i=1}^n$, *where* $\phi_i \in V, i = 1,\dots,n$. *Then* $\left\{u_{T'\phi_i}\right\}_{i=1}^n$ *is a basis of* U_h^n.

In the sequel, we shall not distinguish v_u and v_{Tu} as well as u_v and $u_{T'v}$ since we shall work exclusively with the Riesz representations.

We will use only the primal approach to investigate infinite dimensional settings and then deduce the properties of the resulting finite dimensional settings. That is, we start with a basis function in the trial space U and then derive the corresponding optimal basis function in the test space V. Of course, we can also take a dual approach, i.e, we start with a basis in the test space V and then seek the corresponding optimal basis in the trial space U. This will as well lead to well-posed approximation methods with $M_h = \gamma_h = 1$ by Theorem 2.10. However, the approximability is questionable since the resulting finite dimensional trial space is no longer designed to accurately approximate the exact solution.

3. The first well-posed approximation of linear scalar hyperbolic equations

In this section, we discuss in detail how to apply the primal method developed in Section 2 to a weak formulation of advection-reaction problems. The consistency and well-posedness of the infinite dimensional setting are analyzed at length. We shall use the Cauchy–Schwarz inequality to detect the natural norms so that the bilinear form under consideration is continuous with $M = 1$. The conditions for equalities to be achievable in the Cauchy–Schwarz inequality then allow us to find optimal pair of functions in U and V. In fact, the Riesz representation turns out to be a candidate for equalities to happen. Using this fact along with Theorem 2.9 we obtain a finite dimensional approximation method with guaranteed well-posedness and $M_h = \gamma_h = 1$.

3.1. Infinite dimensional setting.

Our problem of interest is the first order scalar linear hyperbolic equation of the form

$$\boldsymbol{\beta} \cdot \nabla u + \mu u = f, \quad \text{in } \Omega, \tag{6a}$$

$$u = g, \quad \text{on } \Gamma, \tag{6b}$$

where $\Gamma = \{\mathbf{x} \in \partial\Omega : \mathbf{n}(\mathbf{x}) \cdot \boldsymbol{\beta} < 0\}$ is the inflow boundary; $\mathbf{n}(\mathbf{x})$ denotes the outward normal vector at $\mathbf{x}$ on the boundary $\partial\Omega$. Assume $\boldsymbol{\beta} \in \left[W^{1,\infty}(\Omega)\right]^d$ with $d \in \{1,2,3\}$ denoting the dimension of the problem, $\mu \in L^\infty(\Omega)$, $f \in L^2(\Omega)$, and $g \in L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\Gamma)$ with

$$L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\Gamma) = \left\{ w : \|w\|^2_{L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\Gamma)} = \int_\Gamma |\boldsymbol{\beta}\cdot\mathbf{n}|\,|w|^2 \, d\Gamma < \infty \right\},$$

and $u \in H^1_{\boldsymbol{\beta}}$ with

$$H^1_{\boldsymbol{\beta}}(\Omega) = \left\{ u \in L^2(\Omega) : \boldsymbol{\beta}\cdot\nabla u \in L^2(\Omega) \right\}.$$

The following well-posedness of the transport equation (6) is proved in [11].

Lemma 3.1. *Assume that Ω is a Lipschitz domain. Let*

$$W = \left\{ u \in H^1_{\boldsymbol{\beta}}(\Omega) : u|_\Gamma = 0 \right\}$$

and define

$$W \ni u \mapsto Tu = \boldsymbol{\beta}\cdot\nabla u + \mu u \in L^2(\Omega).$$

Suppose that $\boldsymbol{\beta}$ is a filling field, i.e., there exists a characteristic line of the vector field $\boldsymbol{\beta}$ starting from Γ and arriving at (almost everywhere) $\mathbf{x} \in \Omega$ in finite time. Then T is a bijective map from W to $L^2(\Omega)$.

We partition the domain Ω into N^{el} nonoverlapping elements $K_j, j = 1, \ldots, N^{\mathrm{el}}$ such that $\Omega_h = \bigcup_{j=1}^{N^{\mathrm{el}}} K_j$ and $\overline{\Omega} = \overline{\Omega}_h$. Here, h is defined as $h = \max_{j\in\{1,\ldots,N^{\mathrm{el}}\}} \mathrm{diam}(K_j)$. In addition, we denote by $\mathcal{E}_h$, with cardinal number N^{ed}, the set of all unique faces in the mesh, namely, the mesh skeleton. In this paper, the term "faces" is used to indicate either edges of 2D elements or faces of 3D elements. Finally, we require $\boldsymbol{\beta}\cdot\mathbf{n}|_e \in L^\infty(e)$ for $e = 1, \ldots, N^{\mathrm{ed}}$, where $\mathbf{n}$ is the normal vector on face e. Multiplying (6a) by a test function v, integrating by parts, and introducing the single-valued flux $q \in L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\mathcal{E}_h)$ at the element interfaces, we have,

$$\sum_{j=1}^{N^{\mathrm{el}}} \int_{K_j} \left[-u\nabla\cdot(\boldsymbol{\beta}v) + \mu uv\right] d\mathbf{x} + \int_{\partial K_j} \mathbf{1}_{\partial K_j\setminus\Gamma} \boldsymbol{\beta}\cdot\mathbf{n} q v \, ds = \sum_{j=1}^{N^{\mathrm{el}}} \int_{K_j} fv \, d\mathbf{x} - \int_{\partial K_j\cap\Gamma} \boldsymbol{\beta}\cdot\mathbf{n} g v \, ds, \tag{7}$$

with $\mathbf{1}_{\partial K_j\setminus\Gamma}$ denoting the indicator function (also known as the characteristic function) of $\partial K_j \setminus \Gamma$. Clearly, for elements with characteristic faces, i.e., $\boldsymbol{\beta}\cdot\mathbf{n} = 0$ on ∂K_j, the boundary integrals corresponding to these faces simply drop out and q is allowed to be undefined on these boundaries. Next, integrating by parts one more time gives

$$\sum_{j=1}^{N^{\mathrm{el}}} \int_{K_j} (\boldsymbol{\beta}\cdot\nabla u + \mu u) v \, d\mathbf{x} + \int_{\partial K_j} \boldsymbol{\beta}\cdot\mathbf{n} \left(\mathbf{1}_{\partial K_j\setminus\Gamma} q - u\right) v \, ds = \sum_{j=1}^{N^{\mathrm{el}}} \int_{\partial K_j} fv \, d\mathbf{x} - \int_{\partial K_j\cap\Gamma} \boldsymbol{\beta}\cdot\mathbf{n} g v \, ds. \tag{8}$$

If we choose $v|_{K_j} \in L^2(K_j)$, then the trace $v|_{\partial K_j}$ is not defined. Therefore, we introduce a new hybrid variable r that lives in the space $\Pi_{j=1}^{N^{\mathrm{el}}} L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\partial K_j)$. Unlike q, which is single-valued on a face of the skeleton, r is allowed to have double values depending on the side of that face. With the introduction of r, (8) can be rewritten as

$$a(\mathbf{u},\mathbf{v}) = \sum_{j=1}^{N^{\mathrm{el}}} \int_{K_j} (\boldsymbol{\beta}\cdot\nabla u + \mu u) v \, d\mathbf{x} + \int_{\partial K_j} \boldsymbol{\beta}\cdot\mathbf{n} \left(\mathbf{1}_{\partial K_j\setminus\Gamma} q - u\right) r \, ds = \ell(\mathbf{v}) = \sum_{j=1}^{N^{\mathrm{el}}} \int_{K_j} fv \, d\mathbf{x} - \int_{\partial K_j\cap\Gamma} \boldsymbol{\beta}\cdot\mathbf{n} g r \, ds, \tag{9}$$

with $\mathbf{u} = (u, q)$, $\mathbf{v} = (v, r)$. We define the trial and test spaces as

$$U = \left\{\mathbf{u} : \mathbf{u}|_{K_j} \in H^1_{\boldsymbol{\beta}}(K_j) \times L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\partial K_j), j = 1, \ldots, N^{\mathrm{el}}\right\} = H^1_{\boldsymbol{\beta}}(\Omega_h) \times L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\mathcal{E}_h),$$

$$V = L^2(\Omega_h) \times \Pi_{j=1}^{N^{\mathrm{el}}} L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\mathcal{E}_h).$$

We first study the consistency of the weak formulation (9).

Proposition 1 (Consistency)**.** *If u is a solution of (6), then $\left(u, q = u|_{\mathcal{E}_h}\right)$ is a solution of (9). Conversely, if $(u,q) \in L^2(\Omega) \cap H^1_{\boldsymbol{\beta}}(\Omega_h) \times L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\mathcal{E}_h)$ is a solution of (9), then u is a solution of (6).*

Proof. If u is a solution of (6), then it is straightforward to see that $\left(u, q = u|_{\mathcal{E}_h}\right)$ is a solution of (9).

Conversely, let $(u,q) \in U$, with $u \in L^2(\Omega) \cap H^1_{\boldsymbol{\beta}\cdot\mathbf{n}}(\Omega_h)$, be a solution of (9) for all $(v,r) \in V$. First, taking $v = 0$ and $r \in L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(e)$ where $e \in \partial K_j$ is an internal or an outflow face, for some $j \in \left\{1, \ldots, N^{\mathrm{el}}\right\}$, yields $\boldsymbol{\beta}\cdot\mathbf{n}q = \boldsymbol{\beta}\cdot\mathbf{n}u$. This implies, for an internal face, $\boldsymbol{\beta}\cdot\mathbf{n}u^- = \boldsymbol{\beta}\cdot\mathbf{n}u^+$. Similarly, for an inflow face we obtain $\boldsymbol{\beta}\cdot\mathbf{n}u = \boldsymbol{\beta}\cdot\mathbf{n}g$.

Next, taking $v \in \mathcal{D}(K_j) = C_0^\infty(K_j)$ and $r = 0$, then (9) implies

$$\boldsymbol{\beta}\cdot\nabla u + \mu u = f, \quad \text{in } K_j, \quad j = 1, \ldots, N^{\mathrm{el}}, \tag{10}$$

in the distributional sense. Since $\boldsymbol{\beta}\cdot\nabla u + \mu u, f \in L^2(K_j)$, (10) also holds in $L^2(K_j)$. Finally, taking $v \in \mathcal{D}(\Omega)$, integrating by parts equation (10), and using $\boldsymbol{\beta}\cdot\mathbf{n}u^- = \boldsymbol{\beta}\cdot\mathbf{n}u^+$ and $u \in L^2(\Omega)$, we obtain the following identity,

$$-\int_\Omega u\boldsymbol{\beta}\cdot\nabla v \, d\mathbf{x} = \int_\Omega (f - \mu u + \nabla\cdot\boldsymbol{\beta}u)\, v \, d\mathbf{x},$$

which implies, in the distributional sense,

$$\nabla\cdot(\boldsymbol{\beta}u) = f - \mu u + \nabla\cdot\boldsymbol{\beta}u \in L^2(\Omega),$$

which in turn implies

$$\boldsymbol{\beta}\cdot\nabla u \in L^2(\Omega).$$

Hence, (10) is in fact valid globally in Ω. □

Guided by Theorem 2.6, we are seeking norms in spaces U and V such that the continuity constant is unity and the equality in the continuity condition is achievable. An approach to obtain this goal is to apply the Cauchy–Schwarz inequality, i.e.,

$$\begin{aligned} a(\mathbf{u},\mathbf{v}) &\le \sum_{j=1}^{N^{\mathrm{el}}} \|\boldsymbol{\beta}\cdot\nabla u + \mu u\|_{L^2(K_j)} \|v\|_{L^2(K_j)} \\ &\quad + \left\|\mathbf{1}_{\partial K_j\setminus\Gamma} q - u\right\|_{L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\partial K_j)} \|r\|_{L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\partial K_j)} \\ &\le \underbrace{\left\{\sum_{j=1}^{N^{\mathrm{el}}} \|\boldsymbol{\beta}\cdot\nabla u + \mu u\|^2_{L^2(K_j)} + \left\|\mathbf{1}_{\partial K_j\setminus\Gamma} q - u\right\|^2_{L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\partial K_j)}\right\}^{\frac{1}{2}}}_{\|\mathbf{u}\|_U} \\ &\quad \times \underbrace{\left\{\sum_{j=1}^{N^{\mathrm{el}}} \|v\|^2_{L^2(K_j)} + \|r\|^2_{L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\partial K_j)}\right\}^{\frac{1}{2}}}_{\|\mathbf{v}\|_V}. \end{aligned} \tag{11}$$

It is clearly that the functional $\|\cdot\|_V$ defined above is a norm. Before showing that the functional $\|\cdot\|_U$ is indeed a norm, let us check whether the equality is attainable.

Since the Cauchy–Schwarz inequalities are employed, one easily sees that, given $\mathbf{u} = (u, q) \in U$, if $\mathbf{v_u} = (v_\mathbf{u}, r_\mathbf{u})$ is defined by

$$v_\mathbf{u} = \boldsymbol{\beta} \cdot \nabla u + \mu u, \quad \text{in } K_j, \tag{12a}$$

$$r_\mathbf{u} = \operatorname{sgn}(\boldsymbol{\beta} \cdot \mathbf{n}) \left(\mathbf{1}_{\partial K_j \setminus \Gamma} q - u \right), \quad \text{on } \partial K_j, \tag{12b}$$

then the equality is attainable, i.e., $a(\mathbf{u}, \mathbf{v_u}) = \|\mathbf{u}\|_U \|\mathbf{v_u}\|_V$. Notice that $\mathbf{v_u}$ is in fact the Riesz representation of $T\mathbf{u}$ (see Theorem 2.9 for the definition of T). At this point, we immediately have the well-posedness of the weak formulation (9).

Theorem 3.2 (Well-posedness)**.** *Assume that $\boldsymbol{\beta}$ is a filling field. Then, the weak formulation* (9) *is well-posed with $M = \gamma = 1$ in the norms defined in* (11)*.*

Proof. $M = \gamma = 1$ is a direct consequence of Theorem 2.6. What remains to be proved is the adjoint injectivity condition (3b). Let $\mathbf{v} \in V$ and assume that

$$a(\mathbf{u}, \mathbf{v}) = 0, \quad \forall \mathbf{u} \in U. \tag{13}$$

Taking $u = 0$, then (13) implies

$$\sum_{j=1}^{N^{\mathrm{el}}} \int_{\partial K_j} \boldsymbol{\beta} \cdot \mathbf{n} \mathbf{1}_{\partial K_j \setminus \Gamma}\, qr\, ds = 0, \quad \forall q \in L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\mathcal{E}_h),$$

which in turn implies $r = 0$ on $\mathcal{E}_h \setminus \Gamma$ since the map

$$L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\partial K_j) \ni q \mapsto \boldsymbol{\beta} \cdot \mathbf{n} q \in L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\partial K_j)$$

is surjective due to $0 < c_1 \le \|\boldsymbol{\beta} \cdot \mathbf{n}\|_{L^\infty(\partial K_j)} \le c_2 < \infty$. Now, if $u \in W$, then (13) implies

$$\int_\Omega (\boldsymbol{\beta} \cdot \nabla u + \mu u)\, v\, d\mathbf{x} = 0, \quad \forall u \in W,$$

which, together with the surjectivity of $Tu = \boldsymbol{\beta} \cdot \nabla u + \mu u$ from Lemma 3.1, yields $v = 0$ in Ω. Finally, with $v = 0$ in Ω and $r = 0$ on $\mathcal{E}_h \setminus \Gamma$, (13) becomes

$$\sum_{j=1}^{N^{\mathrm{el}}} \int_{\partial K_j \cap \Gamma} \boldsymbol{\beta} \cdot \mathbf{n}\, ur\, ds = 0, \quad \forall u \in H^1_{\boldsymbol{\beta}}(\Omega_h).$$

Since the trace map $H^1_{\boldsymbol{\beta}}(K_j) \ni u \mapsto u|_{\partial K_j} \in L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\partial K_j)$ is surjective, the preceding equation shows that $r = 0$ on Γ. Hence $\mathbf{v} = 0$. □

Having proved the consistency and the well-posedness of the weak formulation (9), we proceed with finding optimal pairs of trial and test basis functions. For basis functions of the form $\boldsymbol{\phi} = (0, \phi) \in U$, where ϕ is a function in $L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\mathcal{E}_h)$, the corresponding basis functions in V for $j = 1, \ldots, N^{\mathrm{el}}$, are given by

$$v_{\boldsymbol{\phi}} = 0, \quad \text{in } K_j, \tag{14a}$$

$$r_{\boldsymbol{\phi}} = \mathbf{1}_{\partial K_j \setminus \Gamma} \phi \operatorname{sgn}(\boldsymbol{\beta} \cdot \mathbf{n}), \quad \text{on } \partial K_j. \tag{14b}$$

Similarly, for basis functions of the form $\boldsymbol{\varphi} = (\varphi, 0)$, where $\varphi \in H^1_{\boldsymbol{\beta}}(\Omega_h)$, the corresponding basis functions in V for $j = 1, \ldots, N^{\mathrm{el}}$ are given by

$$v_{\boldsymbol{\varphi}} = \boldsymbol{\beta} \cdot \nabla \varphi + \mu \varphi, \quad \text{in } K_j, \tag{15a}$$

$$r_{\boldsymbol{\varphi}} = -\varphi \operatorname{sgn}(\boldsymbol{\beta} \cdot \mathbf{n}), \quad \text{on } \partial K_j. \tag{15b}$$

Next, it is natural to substitute the test functions (14) or (15) into (9) to establish equations to solve for the unknowns $\mathbf{u} = (u, q)$. Let us proceed with the generic

test basis in (14) first. If ϕ is not a zero function on the interface $\partial K_i \cap \partial K_j$ of elements K_i and K_j and zero elsewhere, then testing with $\mathbf{v}_{\boldsymbol{\phi}}$ defined in (14) the weak formulation (9) becomes

$$\int_{\partial K_i \cap \partial K_j} |\boldsymbol{\beta} \cdot \mathbf{n}| \left(q - \{\{u\}\}\right) \phi \, ds = 0, \quad \forall \phi \in L^2_{\boldsymbol{\beta}\cdot\mathbf{n}} \left(\partial K_i \cap \partial K_j\right),$$

where $\{\{u\}\} = \frac{u^- + u^+}{2}$ denotes the average. The positive and negative signs indicate element interior and exterior, respectively. Since $(q - \{\{u\}\}) \in L^2_{\boldsymbol{\beta}\cdot\mathbf{n}} \left(\partial K_i \cap \partial K_j\right)$, the preceding equation implies

$$q = \{\{u\}\} \quad \text{ on } \partial K_i \cap \partial K_j.$$

Similarly, taking ϕ to be a nonzero function on $(\partial\Omega \setminus \Gamma) \cap \partial K_j$ and zero elsewhere, and then testing (9) with $\mathbf{v}_{\boldsymbol{\phi}}$ defined in (14), we obtain,

$$q = u \quad \text{ on } \partial\Omega \cap \partial K_j.$$

On the other hand, due to the indicator function $\mathbf{1}_{\partial K_j \setminus \Gamma}$, there is no test function $r_{\boldsymbol{\phi}}$ on the inflow boundary faces. In summary, testing with the test basis (14), the unknown q on the skeleton is found explicitly as

$$q = \begin{cases} \{\{u\}\} & \text{on } \partial K_i \cap \partial K_j \\ u & \text{on } (\partial\Omega \setminus \Gamma) \cap \partial K_j. \end{cases} \tag{16}$$

As a result, the trial space can be rewritten as

$$U = \left\{ u : u|_{K_j} \in H^1_{\boldsymbol{\beta}} \left(K_j\right) \right\} = H^1_{\boldsymbol{\beta}} \left(\Omega_h\right),$$

and the norm in U now becomes

$$\|u\|_U = \left\{ \sum_{j=1}^{N^{\text{el}}} \|\boldsymbol{\beta} \cdot \nabla u + \mu u\|^2_{L^2(K_j)} + \left\| \frac{1}{2} [\![u]\!] \right\|^2_{L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\partial K_j \setminus \Gamma)} + \|u\|^2_{L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\partial K_j \cap \Gamma)} \right\}^{\frac{1}{2}}, \tag{17}$$

which is essentially the norm in which the upwind DG method is stable; see [12], for example.

We are now in position to prove that (17) indeed defines a norm in U.

Proposition 2. *Assume that $\boldsymbol{\beta}$ is a filling field, then the functional*

$$\|\cdot\|_U : U \to [0, \infty)$$

defined in (17) *is a norm.*

Proof. The fact that the functional $\|\cdot\|_U$ satisfies the positive homogeneity and the triangle inequality is obvious. Therefore, it remains to prove that $\|u\|_U = 0$ implies $u = 0$. We first consider elements whose faces are subsets of the inflow boundary. In this case, $\|u\|_U = 0$ implies

$$\boldsymbol{\beta} \cdot \nabla u + \mu u = 0, \qquad \text{in } K_j, \tag{18a}$$

$$u = 0, \qquad \text{on } \partial K_j \cap \Gamma, \tag{18b}$$

$$[\![u]\!] = 0, \qquad \text{on } \partial K_j \setminus \Gamma. \tag{18c}$$

By Lemma 3.1, $u = 0$ in $\overline{K_j}$ is the unique solution of (18a) and (18b). Since $\boldsymbol{\beta}$ is a filling field, $\partial K_j \setminus \Gamma$ must be inflow boundaries of other adjacent elements. However, (18c) implies that $u = 0$ on these inflow boundaries. By induction, we conclude that $u = 0$ is zero on Ω_h, and this ends the proof. ∎

Next, substituting the optimal test basis function (15) into (9) we have

$$\sum_{j=1}^{N^{\mathrm{el}}} \int_{K_j} (\boldsymbol{\beta}\cdot\nabla u + \mu u)\,(\boldsymbol{\beta}\cdot\nabla\varphi + \mu\varphi)\; d\mathbf{x} - \int_{\partial K_j} |\boldsymbol{\beta}\cdot\mathbf{n}|\left(\mathbf{1}_{\partial K_j\setminus\Gamma} q - u\right)\varphi\, ds \tag{19}$$
$$= \sum_{j=1}^{N^{\mathrm{el}}} \int_{K_j} f\,(\boldsymbol{\beta}\cdot\nabla\varphi + \mu\varphi)\; d\mathbf{x} + \int_{\partial K_j\cap\Gamma} |\boldsymbol{\beta}\cdot\mathbf{n}|\, g\varphi\, ds.$$

We can simplify (19) further by using the explicit value of q in (16), renaming φ to v, and rewriting the bilinear and linear forms in (9) as

$$\begin{aligned} a(u,v) &= \sum_{j=1}^{N^{\mathrm{el}}} \int_{K_j} (\boldsymbol{\beta}\cdot\nabla u + \mu u)\,(\boldsymbol{\beta}\cdot\nabla v + \mu v)\; d\mathbf{x} \\ &\quad + \frac{1}{2}\int_{\partial K_j\setminus\partial\Omega} |\boldsymbol{\beta}\cdot\mathbf{n}|\,[\![u]\!]\, v\, ds + \int_{\partial K_j\cap\Gamma} |\boldsymbol{\beta}\cdot\mathbf{n}|\, uv\, ds \\ \ell(v) &= \sum_{j=1}^{N^{\mathrm{el}}} \int_{K_j} f\,(\boldsymbol{\beta}\cdot\nabla v + \mu v)\; d\mathbf{x} + \int_{\partial K_j\cap\Gamma} |\boldsymbol{\beta}\cdot\mathbf{n}|\, gv\, ds, \end{aligned} \tag{20}$$

where the jump operator $[\![u]\!] = u^- - u^+$ is employed. It is important to point out that the weak formulation (20) is completely equivalent to (9) except that the auxiliary hybrid variables q and r are now eliminated, and that the trial and test spaces are now the same, i.e.,

$$U = V = H^1_{\boldsymbol{\beta}}(\Omega_h).$$

3.2. **Finite dimensional setting and convergence analysis.** Now, we select the same finite dimensional subspace for both trial and test functions, i.e.,

$$V_h^n = U_h^n = span\,\{\varphi_i \in U, i = 1,\ldots,n\}.$$

The consistency of the weak formulation (9) on finite dimensional subspaces U_h^n and V_h^n is automatically inherited from Proposition 1. On the other hand, using Lemma 2.8 we immediately have the well-posedness of the finite dimensional approximate problem.

Theorem 3.3. *Let the bilinear form $a(\cdot,\cdot)$ and the linear form $\ell(\cdot)$ be defined as in* (20)*. The following problem,*

$$\begin{cases} \text{Seek } u_h \in U_h^n \text{ such that} \\ a(u_h, v_h) = \ell(v_h), \quad \forall v_h \in V_h^n, \end{cases}$$

is well-posed with $M_h = \gamma_h = 1$.

That is, our effort in Section 2 is now paid off by the trivial well-posedness of the finite dimensional problem, which is typically not using other methods [6].

In order to use polynomial approximation results [13, 14, 15, 16], we specify the finite dimensional subspaces U_h^n and V_h^n as

$$U_h^n = \left\{ u \in H^1_{\boldsymbol{\beta}}(\Omega_h) : u|_{K_j} \in \mathcal{P}^{p_j}(K_j), j = 1,\ldots,N^{\mathrm{el}} \right\} = V_h^n,$$

where $\mathcal{P}^p$ denotes the polynomial spaces of order at most p. For $d \in \{2,3\}$, $\mathcal{P}^p$ could be the usual polynomial spaces for triangular and tetrahedral meshes, and the tensor product polynomial spaces for quadrilateral and hexahedral meshes.

At this point, a closer look tells us that our simplified weak formulation (20) is one of the existing least squares DG (LSDG) methods [17]. However, it is important to emphasize that, for the DG methods in general, and the other LSDG methods in particular, one usually starts by introducing a numerical flux (typically borrowed from finite volume methods). One then defines bilinear and linear forms (adhoc typically) over finite dimensional polynomial subspaces, and proves the well-posedness of the resulting finite dimensional approximate problem with some prescribed norm. Here, starting from the requirement of having unity continuity and inf–sup constants, and choosing the weak formulation (9) to apply our theory developed in Section 2, we constructively (and accidentally) derive an LSDG method from a well-posed infinite dimensional setting. The distinct feature of our method is that the flux is introduced as a new unknown, and then found by fulfilling stability. It should be pointed out that our theory is general in the following sense. If it is applied to different weak formulations, one will constructively obtain different numerical methods, again, with trivial well-posedness for the finite dimensional approximation problem, as we shall show in the next sections.

Since our resulting approximation method coincides to an LSDG approach, the following *hp*-convergence result is immediate from [17].

Theorem 3.4. *Suppose that* $u|_{K_j} \in H^{s_j}(K_j), s_j > \frac{1}{2}, j = 1, \ldots, N^{el}$ *and the mesh is affine. Then, there exists a positive constant* C*, independent of* $h_j = diam(K_j)$*,* p_j*, and* u *such that*

$$\|u - u_h\|_U^2 \leq C \sum_{j=1}^{N^{el}} \frac{h_j^{2\sigma_j - 2}}{p_j^{2s_j - 2}} |u|^2_{H^{s_j}(K_j)},$$

with $p_j \geq 1$ *and* $0 \leq \sigma_j \leq \min(p_j + 1, s_j)$.

Furthermore, if the mesh is quadrilateral or hexahedral and the exact solution u *is elementwise analytic, then the following estimation holds:*

$$\|u - u_h\|_U^2 \leq C \sum_{j=1}^{N^{el}} \left(\frac{h_j}{2}\right)^{2\sigma_j - 2} p_j e^{-2b_j p_j} |meas(K_j)|,$$

where C *depends on* u*,* d*, and the shape-regularity,* b_j *depends on* d*, and* $meas(K_j)$ *denotes length, area, and volume of* K_j *for* $d = 1$*,* $d = 2$*, and* $d = 3$*, respectively.*

Proof. A proof can be found in [17] and the references therein. □

Remark 3.5. By Corollary 1, the error estimation is optimal in both h and p. In particular, if the exact solution is elementwise analytic, our approximation method delivers exponential convergence in p.

4. The second well-posed approximation of linear scalar hyperbolic equations

In the first application of our abstract theory for linear scalar hyperbolic equations in Section 3, we have chosen the weak formulation (9) obtained after integrating (6) by parts twice. In this section, we apply our theory to the weak formulation (7) obtained after integrating (6) by parts once. As will be shown, we obtain a different well-posed infinite dimensional setting, hence leading to different finite dimensional approximation with guaranteed well-posedness and $M_h = \gamma_h = 1$.

4.1. **Infinite dimensional setting.** Our starting point is the weak formulation (7) obtained by integrating (6) by parts once. This formulation can be written in a more compact form as

$$\sum_{j=1}^{N^{\mathrm{el}}} \int_{K_j} u\left[-\nabla\cdot(\boldsymbol{\beta}v)+\mu v\right]\,d\mathbf{x}+\frac{1}{2}\int_{\partial K_j}\boldsymbol{\beta}\cdot\mathbf{n}q\,[\![v]\!]\,ds = \sum_{j=1}^{N^{\mathrm{el}}}\int_{K_j} fv\,d\mathbf{x}-\int_{\partial K_j\cap\Gamma}\boldsymbol{\beta}\cdot\mathbf{n}gv\,ds, \tag{21}$$

where on the outflow boundary we conventionally define $v^+ = -v^-$ and on the inflow $v^+ = v^-$. We define the trial and test spaces as

$$U=\left\{\mathbf{u}:\mathbf{u}|_{K_j}\in L^2(K_j)\times L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\partial K_j)\right\}=L^2(\Omega_h)\times L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\mathcal{E}_h),$$
$$V=\left\{v:v|_{K_j}\in H^1_{\boldsymbol{\beta}}(K_j)\right\}=H^1_{\boldsymbol{\beta}}(\Omega_h).$$

Again, guided by Theorem 2.6, we are looking for natural norms in spaces U and V such that the continuity constant is unity and the equality in the continuity condition is achievable. Similar to Section 3, we apply the Cauchy–Schwarz inequality to obtain,

$$\begin{aligned} a(\mathbf{u},v) &\le \sum_{j=1}^{N^{\mathrm{el}}}\|-\nabla\cdot(\boldsymbol{\beta}v)+\mu v\|_{L^2(K_j)}\|u\|_{L^2(K_j)} \\ &\quad+\frac{1}{2}\|q\|_{L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\partial K_j)}\|[\![v]\!]\|_{L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\partial K_j)} \\ &\le \underbrace{\left\{\sum_{j=1}^{N^{\mathrm{el}}}\|-\nabla\cdot(\boldsymbol{\beta}v)+\mu v\|^2_{L^2(K_j)}+\left\|\frac{1}{\sqrt{2}}[\![v]\!]\right\|^2_{L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\partial K_j)}\right\}^{\frac{1}{2}}}_{\|\mathbf{v}\|_V} \\ &\quad\times\underbrace{\left\{\sum_{j=1}^{N^{\mathrm{el}}}\|u\|^2_{L^2(K_j)}+\left\|\frac{1}{\sqrt{2}}q\right\|^2_{L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\partial K_j)}\right\}^{\frac{1}{2}}}_{\|\mathbf{u}\|_U}. \end{aligned} \tag{22}$$

Equalities in (22) are achievable if we use the Riesz representations, i.e., given $\mathbf{u}=(u,q)\in U$,

$$u=-\nabla\cdot(\boldsymbol{\beta}v_{\mathbf{u}})+\mu v_{\mathbf{u}},\quad \text{in } K_j, \tag{23a}$$
$$q=\operatorname{sgn}(\boldsymbol{\beta}\cdot\mathbf{n})\,[\![v_{\mathbf{u}}]\!],\quad \text{on } \partial K_j. \tag{23b}$$

Once we know that the equality is obtainable under condition (23), the consistency and well-posedness of (21) with respect to the norms defined in (22) similarly follows Proposition 1 and Theorem 3.2.

We next use (23) to find optimal pairs of trial and (corresponding) test basis functions. For basis functions of the form $\boldsymbol{\phi}=(0,\phi)\in U$, where ϕ is a function in

$L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(\mathcal{E}_h)$, the corresponding basis functions in V are given by, for $j = 1, \ldots, N^{\text{el}}$,

$$-\nabla\cdot(\boldsymbol{\beta} v_\phi) + \mu v_\phi = 0, \quad \text{in } K_j, \tag{24a}$$

$$[\![v_\phi]\!] = \operatorname{sgn}(\boldsymbol{\beta}\cdot\mathbf{n})\,\phi, \quad \text{on } \partial K_j. \tag{24b}$$

Similarly, for basis functions of the form $\boldsymbol{\varphi} = (\varphi, 0) \in U$, where $\varphi \in L^2(\Omega_h)$, the corresponding basis functions in V are given by, for $j = 1, \ldots, N^{\text{el}}$,

$$-\nabla\cdot(\boldsymbol{\beta} v_\varphi) + \mu v_\varphi = \varphi, \quad \text{in } K_j, \tag{25a}$$

$$[\![v_\varphi]\!] = 0, \quad \text{on } \partial K_j. \tag{25b}$$

Once the test functions are found using (24) or (25), we can substitute them into (21) to establish equations to solve for the unknowns $\mathbf{u} = (u, q)$. Let us proceed with the generic test basis in (24) first. If ϕ is different from zero on $e \in \mathcal{E}_h$ and zero elsewhere on the skeleton, then testing (21) with $\mathbf{v} = (0, v_\phi)$ yields,

$$\int_e |\boldsymbol{\beta}\cdot\mathbf{n}|\ q\ \phi\ ds = \int_{\Omega_h} f\ v_\phi\ d\mathbf{x} - \int_\Gamma \boldsymbol{\beta}\cdot\mathbf{n}\ g\ v_\phi\ ds. \tag{26}$$

As can be seen in (26), for each $\phi \in L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(e), e \in \mathcal{E}_h$, $q \in L^2_{\boldsymbol{\beta}\cdot\mathbf{n}}(e)$ can be locally solved independently of u.

Now if $\varphi|_{Kj}$ is a nonzero function in K_j but zero elsewhere, then testing (21) with $\mathbf{v} = (0, v_\varphi)$ gives,

$$\int_{K_j} u\varphi\ d\mathbf{x} = \int_{\Omega_h} f\ v_\varphi\ d\mathbf{x} - \int_\Gamma \boldsymbol{\beta}\cdot\mathbf{n}\ g\ v_\varphi\ ds, \tag{27}$$

which shows that the unknown u can also be computed locally element-by-element and independently of q.

4.2. **Finite dimensional setting and convergence analysis.** Now, given a finite dimensional subspace

$$U_h^n = span\{\mathbf{u}_i, i = 1, \ldots, n\} \subset U,$$

where

$$\mathbf{u}_i = \begin{cases} \boldsymbol{\phi}_i = (0, \phi_i) & i = 1, \ldots, m, \\ \boldsymbol{\varphi}_i = (\varphi_i, 0) & i = m+1, \ldots, n, \end{cases}$$

where ϕ_i and φ_i are local functions previously used to obtain (26) and (27). The corresponding optimal finite dimensional test space is given by

$$V_h^n = span\{v_{\mathbf{u}_i}, i = 1, \ldots, n\},$$

with basis vectors $\mathbf{v}_{\mathbf{u}_i}$ computed as in (24) or (25).

The consistency of the weak formulation (26) and (27) on finite dimensional subspaces U_h^n and V_h^n is automatically inherited from the infinite dimensional setting. On the other hand, using Lemma 2.8 we immediately have the well-posedness of the finite dimensional approximate problem similar to Theorem 3.3.

Theorem 4.1. *Let the bilinear form $a(\cdot,\cdot)$ and the linear form $\ell(\cdot)$ be defined as in* (26) *and* (27)*. The following problem,*

$$\begin{cases} \text{Seek } \mathbf{u}_h^n = (u_h^n, q_h^n) \in U_h^n \text{ such that} \\ a(\mathbf{u}_h^n, \mathbf{v}_h^n) = \ell(\mathbf{v}_h^n), \quad \forall \mathbf{v}_h^n \in V_h^n, \end{cases}$$

is well-posed with $M_h = \gamma_h = 1$.

It is important to point out that the computation of each test basis function requires a solution of the adjoint hyperbolic equation (24) or (25). Since $\boldsymbol{\beta}$, and hence $-\boldsymbol{\beta}$, is assumed to be a filling field, it follows that starting from an element K_j, it is always possible to identify the neighbor elements, called outflow elements of K_j, whose inflow faces (with respect to the adjoint flow) are the outflow faces of K_j. We continue this procedure with the current outflow elements, and by induction, we can march to the inflow boundary in finite time since there is a finite number of elements in the mesh. In other words, the adjoint flow starting from an element K_j must arrive at the inflow boundary Γ by "mass conservation" in finite time for any triangulation. Once these elements are found, they form a "streamtube" starting from K_j to the inflow boundary. Note that a streamtube is allowed to have elements whose faces are both inflow and outflow with respect to the adjoint velocity field $-\boldsymbol{\beta}$. Therefore, we can either march the solution for an optimal test function element by element along a streamtube or solve for it on all elements simultaneously. Once the finite dimensional space V_h^n is constructed, the weak formulation (21) becomes decoupled weak formulations (26) and (27). This allows us to locally compute the trace q face-by-face, and the approximate solution u element-by-element, independently.

The solution procedure can be divided into offline and online stages as follows. Since the computation of the test basis functions does not depend on the boundary condition g and the forcing f, it is done once for the offline stage. Furthermore, solving for each optimal test function is an independent task, and hence can be done in parallel. As a result, the offline cost is in fact the same as the cost of solving for one optimal test function if the parallelization is fully exploited. After that, the test basis functions can be used for any g and f, while preserving the best approximation property. Therefore, in the online stage, our resulting finite dimensional approximation method provides fast and best (in the $\|\cdot\|_U$-norm) solution u for any admissible g and/or f. It is fast because we only need to locally solve for u element-by-element by inverting the elemental mass matrices as shown in (27). It is best in the $\|\cdot\|_U$-norm due to Corollary 1. We should point out that the forcing f is typically projected onto the subspace $span\,\{\varphi_i, i = m+1, \ldots, n\}$, and the boundary condition g onto $span\,\{\phi_i|_\Gamma, i = 1, \ldots, m\}$. Thus, the right side of (27) can be evaluated once in the offline stage for the basis functions in subspaces $span\,\{\varphi_i, i = m+1, \ldots, n\}$ and $span\,\{\phi_i|_\Gamma, i = 1, \ldots, m\}$. Clearly, if $f = 0$, the online stage is extremely fast and negligible in cost. This is particularly useful for real-time source or boundary condition inverse problems using optimization method in which one has to solve the forward equation (6) many times with different g and/or f. Moreover, our numerical results show that the offline cost may be much more than offset by the optimality and accuracy with no nonphysical diffusion of our method over the popular upwind DG approaches.

The above offline–online procedure can also be seen as a new model reduction (also known as reduced-order modeling) approach. Typical projection-based model reduction approaches [18, 19] begin with a finite, but large, dimensional approximation problem of dimension n, e.g., (4). When n is very large (i.e. for large-scale problems), the discrete problem (4) is intractable for real-time simulation, inverse problems, optimal control, uncertainty quantification, etc. [18, 20]. The idea behind model reduction is to seek a subspace

$$U_h^{n_r} = \{\xi_i, i = 1, \ldots, n_r\} \subset U_h^n, \quad n_r \ll n,$$

where the basis functions $\xi_i, i = 1, \ldots, n_r$ typically have global support. The discrete problem (4) is now solved for the pair $\{U_h^{n_r}, V_h^{n_r}\}$ instead of $\{U_h^n, V_h^n\}$, which is much cheaper since it has only $n_r \ll n$ unknowns. Clearly, one has to address the well-posedness of the reduced problem and its accuracy. Compared to the existing model reduction techniques, our reduction technique is more expensive in the offline stage. In the online stage, the cost of constructing an approximate solution is more or less similar, namely, $\mathcal{O}(n)$. However, the distinct feature of our method is that it is a "direct model reduction" technique. That is, it does not require to construct the reduced spaces $U_h^{n_r}$ and $V_h^{n_r}$. Moreover, the well-posedness of our direct model reduction method is trivial with $M_h = \gamma_h = 1$ by Lemma 2.8, and our "reduced" solution u_h^n is the best in the space U_h^n due to Corollary 1. In addition, we always have guaranteed optimal *hp*-convergence result for the reduced solution u_h^n, as we shall show momentarily.

Another important feature of our approximation is worth pointing out. That is, the test basis vectors computed from (24) or (25) are continuous across elements while the trial basis is completely discontinuous, implying discontinuous solution u in general. Thus, our approximation method can be considered as a new discontinuous finite element method which is in between continuous finite element methods [21] (in which both trial and test spaces are continuous), and the DG methods [22] or the DPG methods [1, 2, 3] (in which both trial and test spaces are discontinuous). It also has the flavor of the hybridized discontinuous Galerkin method [23] due to the present of the hybrid variable q. For ease in writing, we will also call our method a DPG method.

We next choose finite dimensional polynomial spaces and discuss the convergence of the resulting discontinuous finite element method. To begin, we specify

$$U_h = \left\{ \mathbf{u} = (u, q) \in U \;\middle|\; \begin{array}{ll} u|_{K_j} \in \mathcal{P}^{p_j}(K_j), & j = 1, \ldots, N^{\mathrm{el}}, \\ q|_e \in \mathcal{P}^{p_e}(e), & e = 1, \ldots, N^{\mathrm{ed}} \end{array} \right\}.$$

We first recall the following useful result on polynomial approximation theory [13, 14, 15, 16]. For $u \in H^s(D)$, where D could be an element in the mesh (i.e., $D = K_j$) or its faces (i.e., $D \subseteq \partial K_j$), there exists $\Pi u \in \mathcal{P}^p(D)$ such that

$$\|u - \Pi u\|_{L^2(D)}^2 \leq C \frac{h^{2\sigma}}{p^{2s}} |u|_{H^s(D)}^2, \quad s \geq 0, \tag{28}$$

where $\sigma = \min\{p+1, s\}$, and $h = \operatorname{diam}(D)$. Then, we have the following convergence result whose proof is almost trivial.

Theorem 4.2. *Suppose that* $u|_{K_j} \in H^{s_j}(K_j), s_j > \frac{1}{2}, j = 1, \ldots, N^{el}$, *and* $q|_e \in H^{s_e}(e), e = 1, \ldots, N^{ed}$. *Let the mesh be affine. Then, there exists a positive constant* C, *independent of* $h_j = \operatorname{diam}(K_j)$, $h_e = \operatorname{diam}(e)$, p_j, q, *and* u *such that*

$$\|\mathbf{u} - \mathbf{u}_h^n\|_U^2 \leq C \left\{ \sum_{j=1}^{N^{el}} \left[\frac{h_j^{2\sigma_j}}{p_j^{2s_j}} |u|_{H^{s_j}(K_j)}^2 + \sum_{e \cap \partial K_j, e \in \mathcal{E}_h} \frac{h_e^{2\sigma_e}}{p_e^{2s_e}} |q|_{H^{s_e}(e)}^2 \right] \right\},$$

with $p_j, p_e \geq 1$, $0 \leq \sigma_j \leq \min(p_j + 1, s_j)$, *and* $0 \leq \sigma_e \leq \min(p_e + 1, s_e)$.

Proof. The proof is obvious using the definition of the norm in (22), the best approximation error in Corollary 1, and the approximation error (28). □

5. Numerical results

In this section, we present several numerical results to support our findings. Since our first method coincides with a least squares DG described by Houston *et al.* [17], we refer the readers to this paper for extensive numerical results. We will therefore show only numerical results for our DPG method described in Section 4. Furthermore, since we are only interested in u_h^n, which is independent of q_h^n, we will ignore the computation of q_h^n.

5.1. **One-dimensional example.** We consider a one-dimensional example in which the optimal test functions can be computed analytically. In particular, we choose $\Omega = (0,1)$, $\beta = 1, \mu = 0$, $g = \arctan(-100)$, and the forcing function f to be

$$f(x) = \frac{100}{1 + 100^2 (x-1)^2}.$$

Under these assumptions, the exact solution can be found as

$$u(x) = \arctan[100(x-1)].$$

For one-dimensional problems, our mesh is simply given by $0 = x_0 < x_1 < \ldots < x_{N^{\text{el}}} = 1$, and $K_j = (x_{j-1}, x_j), j = 1, \ldots, N^{\text{el}}$. Let us denote $F_{K_j}(r)$, $r \in \mathcal{I} = [-1,1]$ as a diffeomorphic map from the master element $\mathcal{I}$ to K_j. From (25), the optimal test function is found by integrating from the outflow to the inflow boundaries, i.e.,

$$v_\varphi = \int_x^{x_j} \varphi \, dx, \quad j = 1, \ldots, N^{\text{el}},$$
$$[\![v_\varphi]\!] = 0, \text{ at } x_{j-1} \text{ and } x_j.$$

If supp $\varphi \subseteq K_j$, and $\varphi \circ F(r) = P_m(r)$ is the mth order Legendre polynomial, then we have,

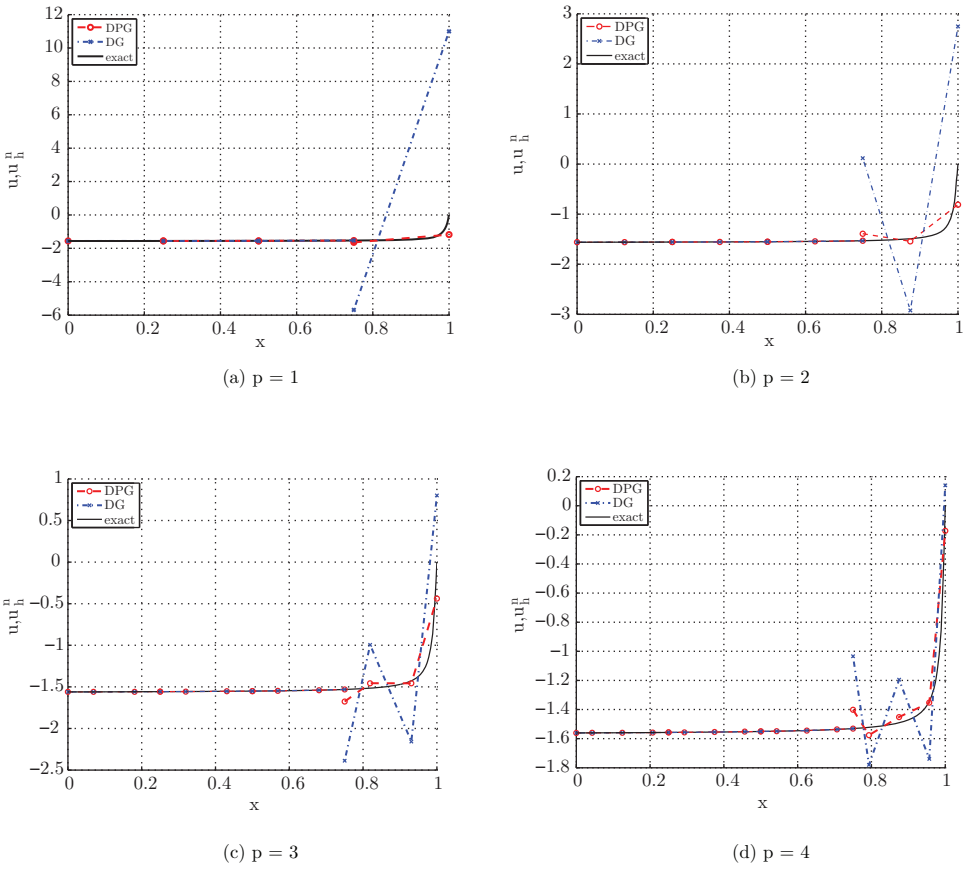
$$v_{P_m(r)} = J_{K_j} \frac{1-r^2}{m(m+1)} \frac{dP_m}{dr}, \quad m \neq 0, \tag{29}$$

where J_{K_j} is the Jacobian of the map F_{K_j}. From (25) and (29) we conclude that $v_{P_m(r)}$ has local support since $v_{P_m(-1)} = v_{P_m(1)} = 0$, in particular, supp $v_{P_m(r)} \subseteq K_j$, $\forall m \neq 0$. For $m = 0$, we obtain

$$v_{P_0(r)} = \begin{cases} 0 & \text{in } K_i, \quad i > j, \\ J_{K_j}(1-r) & \text{in } K_j, \\ 2J_{K_j} & \text{in } K_i, \quad i < j. \end{cases}$$

We have showed that most of the optimal test functions have locally compact supports. This makes the evaluation of the right side of (27) negligible. Moreover, the mass matrix on the left side of (27) becomes diagonal due to the orthogonality of the Legendre polynomials. This makes our method extremely efficient in the online stage since there is no need to invert any matrices to solve for u_h^n, no matter how fine the mesh is.

For the numerical results, we choose uniform order p for all elements. We will compare our DPG method with the original upwind DG [24, 25]. In Figure 1, we plot the DPG and DG solutions for various polynomial orders on a uniform mesh with four elements. In the region with small solution gradient, both methods are comparable. However, in the steep gradient region, the DPG solutions are less



(a) p = 1 (b) p = 2

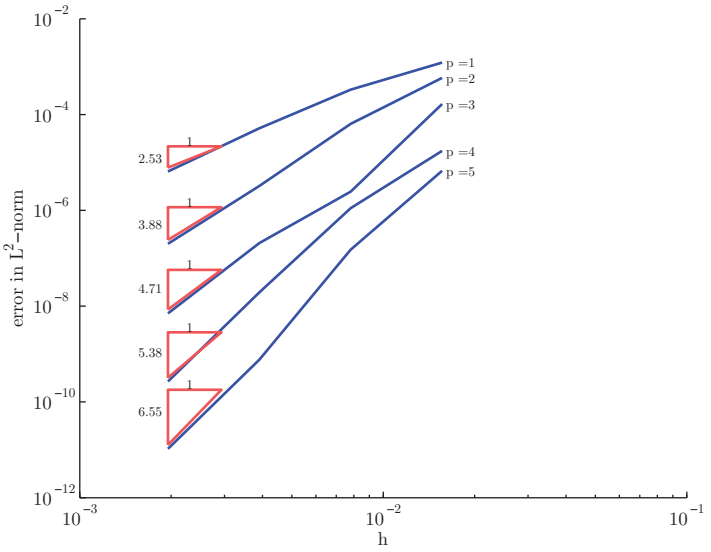
(c) p = 3 (d) p = 4

FIGURE 1. The DPG and DG solutions for various orders $p = 1, 2, 3, 4$ with $h = 0.25$. The mesh has four elements.

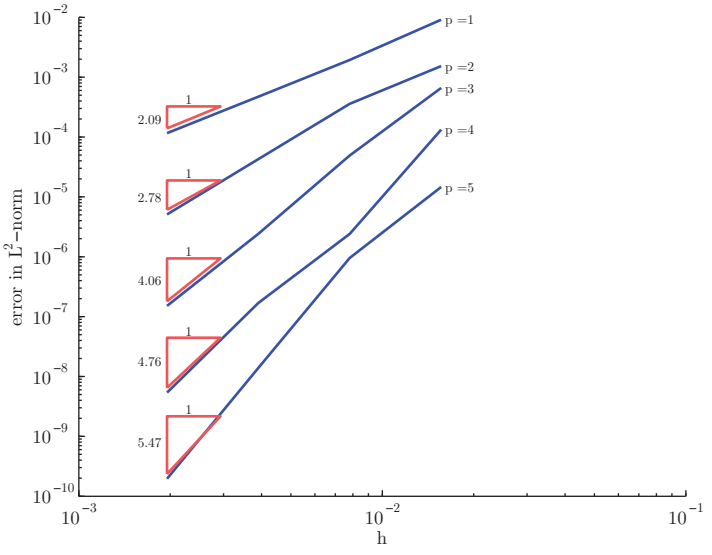
oscillatory than the DG ones. As can also be seen, the DPG solutions have smaller undershootings and overshootings. This is due to the fact that the DPG solutions minimize the error in the energy norm (22), a component of which is the error in the L^2 norm. Therefore, oscillations with big amplitude that have significant L^2 norm are not allowed in the DPG solutions. The DG method, however, does not have such a property.

Figure 2 plots the L^2 error for h refinements in the log-log scale. We approximate the convergence rate by least squares fitting the convergence curves with straight lines. As can be observed, not only are the DPG solutions more accurate by an order of magnitude, but also they have better convergence rates.

Figure 3 shows the L^2 error for p refinements in the linear-log scale. Both DPG and DG exhibits exponential convergence rate, which is consistent with Theorem 3.4. Again, for the same polynomial order and number of unknowns, the DPG method is more accurate than the DG method.

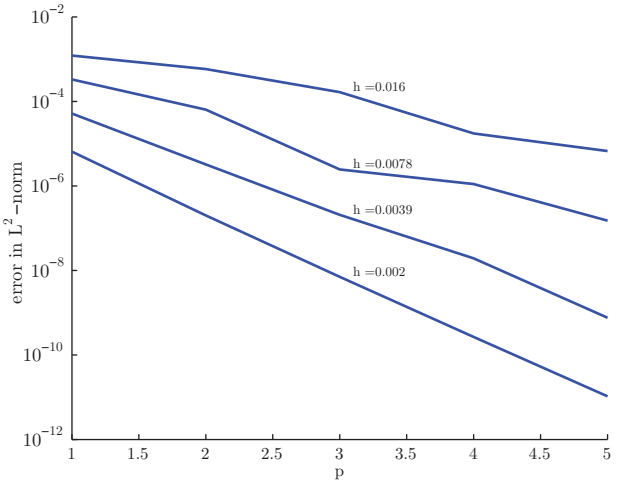


(a) DPG method

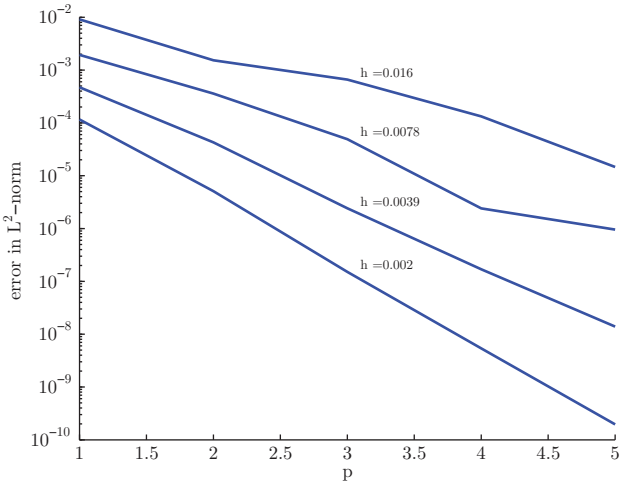


(b) DP method

FIGURE 2. h-convergence rate: 2(a) Log-log scale plot of the error of our DPG method in the L^2-norm, i.e., $\|u - u_h^n\|_{L^2(\Omega_h)}$; 2(b) Log-log scale plot of the error of the nodal DG method in the L^2-norm. The mesh is refined in h for different polynomial orders from $p = 1$ to $p = 5$. The convergence is shown for four different mesh sizes, $h = \{0.43, 0.22, 0.11, 0.054\}$.



(a) DPG method



(b) DG method

FIGURE 3. p convergence: 3(a) Linear-log scale plot of the error of our DPG method in the L^2-norm, i.e., $\|u - u_h^n\|_{L^2(\Omega_h)}$; 3(b) Linear-log scale plot of the error of the nodal DG method in the L^2-norm. The mesh is refined in p, and the convergence is shown for four different mesh sizes, $h \in \{0.016, 0.0078, 0.0039, 0.002\}$.

5.2. **Two-dimensional examples.** For the first example, we choose $\Omega = (-1,1)^2$, $\beta = (0.8, 0.6)$, $\mu = 0$, and the forcing

$$f = \frac{\pi}{10}(1+y)\left[1+y+1.5(1+x)\right]\cos\left(\pi(1+x)(1+y)^2/8\right),$$

so that the exact solution is

$$u = 1 + \sin\left(\pi(1+x)(1+y)^2/8\right).$$

The boundary condition is simply the restriction of the exact solution on the inflow boundaries

$$g = \begin{cases} 1 & y = 0, -1 \le x \le 1, \\ 1 & x = 0, -1 \le y \le 1. \end{cases}$$

For this problem, even though it may be still possible to compute the optimal test functions exactly, a more general and practical approach is to solve for them approximately. In particular, we choose to approximate the optimal test functions using the following finite dimensional test subspace

$$V_h^{\Delta p} = \left\{ v \in V : v|_{K_j} \in \mathcal{P}^{p_j + \Delta p_j}, j = 1, \dots, N^{\mathrm{el}} \right\} \subset V.$$

Clearly, $V_h^{\Delta p}$ asymptotically approaches V as $\Delta p_j \to \infty$, owing to the density of the space of polynomials. We now solve (25) for v_φ by marching along the streamtube starting from K_j to the inflow boundary Γ using the original DG method [24, 25], which seems to be natural for this problem. Figure 4 shows two streamtubes starting from two different elements (black) to the inflow boundary Γ. Clearly, the support of a streamtube, and hence the corresponding optimal test functions, can be significant if the mesh does not align with the streamlines. Alternatively, instead of carrying out the computation on the whole streamtube, one can make further approximation by creating a line of elements starting from K_j to Γ, and this issue will be addressed in our future work [26].

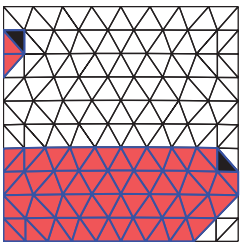


FIGURE 4. Two streamtubes starting from two different elements (black) to the inflow boundary Γ.

Once the optimal test functions v_φ are computed, the computation of the approximate solution u on each element via (27) is then followed by inverting the local mass matrix. Figure 5 presents the h-convergence of the DPG solution for various values of the enriched exponent $\Delta p_j = \{0, 1, 2, 3\}$ with $p_j = 4$, $j = 1, \dots, N^{\mathrm{el}}$. As can be observed, $\Delta p_j = 0$ is the least accurate choice, and the accuracy does not seem to increase for Δp_j greater than unity. For this reason, $\Delta p_j = 1$ will be used from here to the rest of this section.

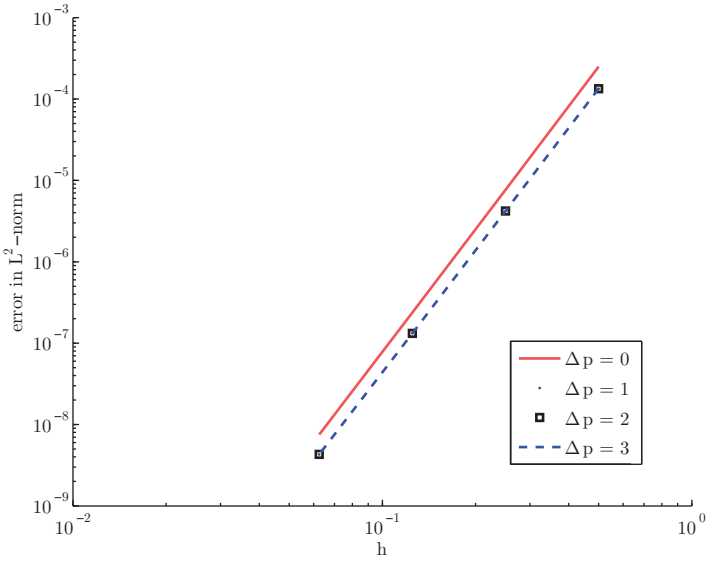


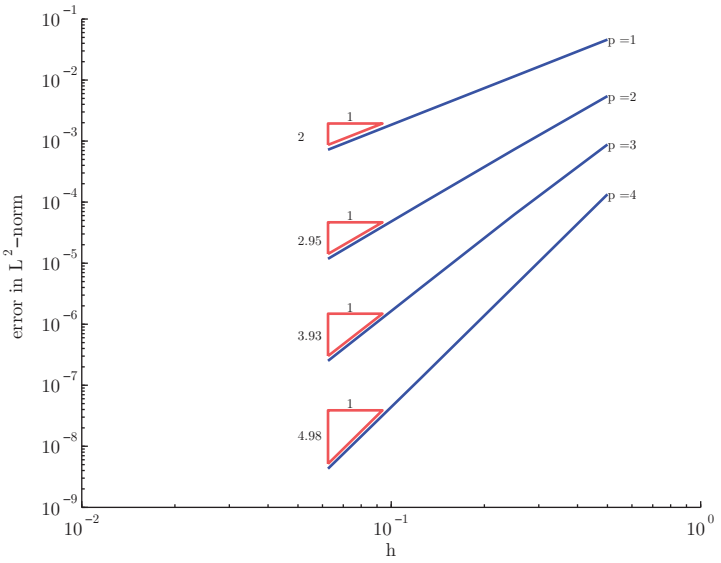
FIGURE 5. The DPG solutions for various values of the enriched exponent $\Delta p_j = \{0, 1, 2, 3\}$ with $p_j = 4$, $j = 1, \ldots, N^{\text{el}}$.

Figure 6 shows that the convergence rate of the DPG method is optimal in h and exponential in p.

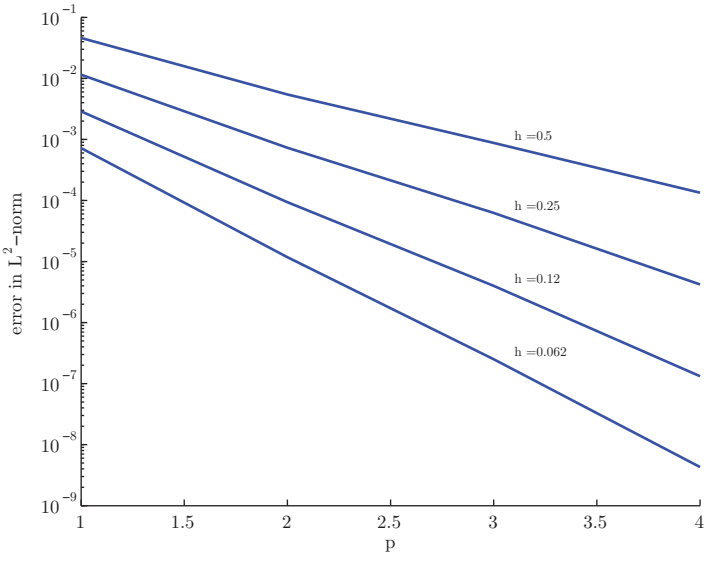
We next consider a more "challenging" problem in [2] on Peterson's meshes [27]. For this problem, we specify $\Omega = (0,1)^2$, $\boldsymbol{\beta} = (0,1)$, $\mu = 0$, $f = 0$, and the inflow boundary condition as

$$g = u(x,0) = \sin(6x), \quad x \in (0,1),$$

so that the exact solution is $u(x,y) = \sin(6x)$. In Figure 7 we compare the h-convergence rate of the DPG method with that of the upwind DG method [24, 25]. The DPG method delivers optimal convergence rates for all polynomial orders while the DG method has sharp sub-optimality. The sub-optimality of the DG method is well known [27], and the numerical result in Figure 7(b) interestingly indicates that it seems to happen only for polynomial orders below three. As can also be observed, the DPG is more accurate than the DG for all h- and p-cases, and this is a direct consequence of the best approximation property in Corollary 1.

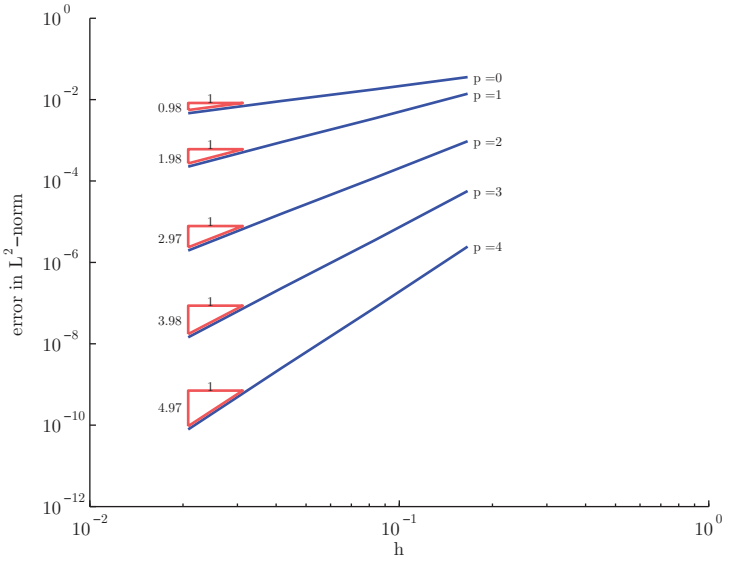


(a) DPG h-convergence

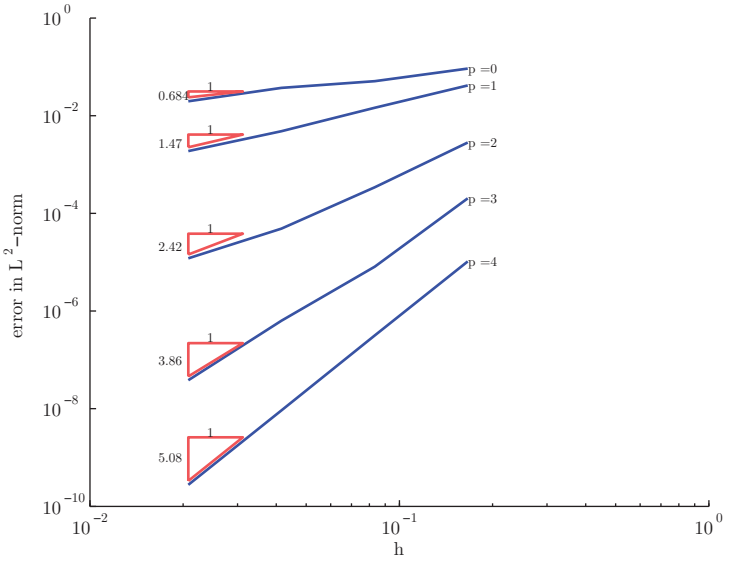


(b) DPG p-convergence

FIGURE 6. DPG h- and p-convergence rates: Log-log scale plot of the error of the DPG method in the L^2-norm, i.e., $\|u - u_h^n\|_{L^2(\Omega_h)}$; The mesh is refined both in h and p for $h = \{0.5, 0.25, 0.125, 0.0625\}$ and for $p = \{1, 2, 3, 4\}$.



(a) DPG h-convergence



(b) DG h convergence

FIGURE 7. h-convergence rate: 7(a) Log-log scale plot of the error of the DPG method in the L^2-norm, i.e., $\|u - u_h^n\|_{L^2(\Omega_h)}$; 7(b) Log-log scale plot of the error of the nodal DG method in the L^2-norm. The mesh is refined in h for different polynomial orders from $p = 0$ to $p = 4$. The convergence is shown for four different mesh sizes $h = \{0.1663, 0.0833, 0.0416, 0.0208\}$.

Since the DPG method can be considered as a least-squares method in the dual space V' [2], we would like to study its diffusion property numerically. To this end, the next example is extracted from [17], in particular, $\Omega = (0,2) \times (0,1)$, $\boldsymbol{\beta} = (1 + \sin(\pi y/2), 2)$, $\mu = 0$, $f = 0$, and the inflow boundary condition

$$g = \begin{cases} 1 & x = 0, 0 \le y \le 1, \\ \sin^6(\pi x) & 0 < x \le 1, y = 0, \\ 0 & 1 \le x \le 2, y = 0, \end{cases}$$

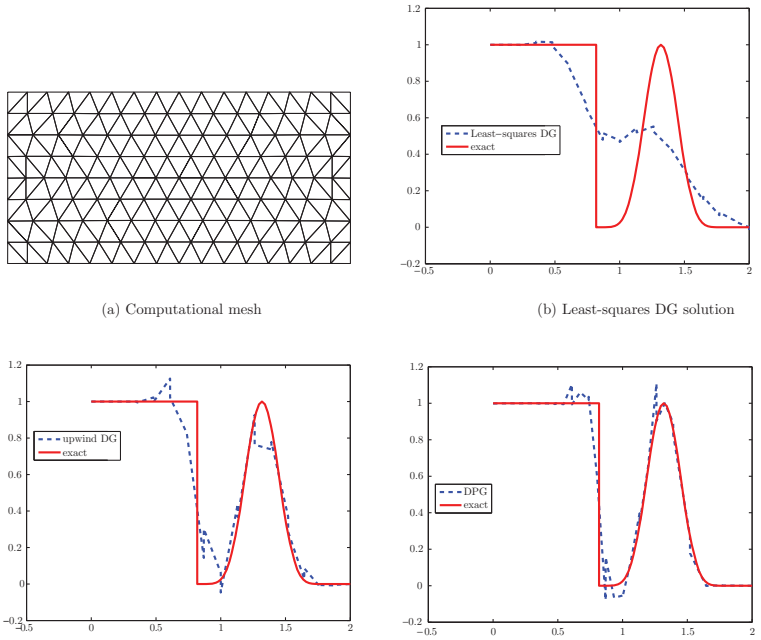
for which the exact solution can be found using the method of characteristics. Figure 8 compares the solutions using the least-squares DG method in Section 3, the upwind DG, and the DPG in Section 4. The computational mesh is shown in Figure 8(a), and the solution order $p = 1$ is chosen uniformly for all elements K_j, $j = 1, \ldots, N^{\text{el}}$. The results show that the least-squares DG is overly diffusive, the DG is less diffusive, and the DPG has the least diffusion. In addition, the DPG method is the most accurate one. It therefore indicates that while standard least-squares methods in primal spaces are very diffusive, those in dual spaces evidently do not seem to be the case. To further confirm this, we consider the following simple example [2] in which we specify $\Omega = (0,1)^2$, $\boldsymbol{\beta} = (0,1)$, $\mu = 0$, $f = 0$, and the inflow boundary condition as

$$g = u(x,0) = x^2, \quad x \in (0,1).$$

The $p = 0$ solutions using the upwind DG and the DPG methods are shown in Figure 9. Clearly, the DPG solution does not seem to have any noticeable cross-wind diffusion (i.e., the colors do not diffuse horizontally along the y-direction), whereas the upwind DG solution does. This reflects the fact that while most of numerical methods for hyperbolic equations introduce numerical diffusion either explicitly or implicitly (e.g., through the numerical fluxes as in many DG methods) to gain stability, and hence introduce nonphysical diffusion, the DPG stability comes directly from the functional setting through the Banach-Nečas-Babuška theorem on the infinite dimensional level.

We have solved for the optimal test functions in the subspace $V_h^{\Delta p} \subset V$ rather than V. Therefore, the inherited well-posedness with the optimal error estimate of the finite dimensional approximation problem does not generally hold, even though our numerical results show that it is in fact the case. Fortunately, a recent work [28] shows that the inheritance is still guaranteed under some suitable conditions for Laplace and linear elasticity equations. A similar result for our DPG method is a subject of our future work.

We emphasize again that our DPG method does not require any equation solves on the online stage if orthogonal polynomials or mass lumpings (e.g. collocation on Legendre-Gauss-Lobatto nodes) are used, and hence (partially) compensating for the cost of computing the optimal test functions. Work is in progress to minimize the test function supports in two- and three-dimensional problems [26]. We should mention that streamline meshes [29] may be an effective option in reducing the cost and complexity of the offline stage, since one only needs to solve for the optimal test basis functions along predefined streamtubes, accordingly. Streamline meshes also imply that the support of any optimal test basis function is confined inside a streamtube, allowing us to evaluate the right side of (27) only along streamtubes, and hence making the online stage even faster.



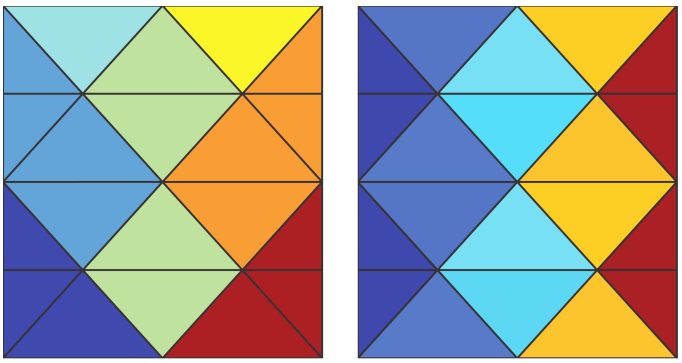
(a) Computational mesh

(b) Least-squares DG solution

(c) Upwind DG solution

(d) DPG solution

FIGURE 8. Solution at the outflow with $p_j = 1, j = 1, \ldots, N^{\text{el}}$ on $0 \leq 0 \leq 2, y = 1$: 8(a) The mesh; 8(b) The least-squares DPG method in Section 3; 8(c) Upwind DG method; 8(d) The DPG method. Solid lines are the exact solution, and numerical solutions are dashed lines.



(a) $p = 0$ DG solution

(b) $p = 0$ DPG solution

FIGURE 9. Numerical solutions with $p_j = 0, j = 1, \ldots, N^{\text{el}}$: 9(a) $p = 0$ DG solution; 9(b) $p = 0$ DPG solution.

6. Conclusions

We have developed a single-framework theory that adapts to the problems at hand, while automatically generating accurate finite element methods with trivial and guaranteed stability. The theory is devised for general variational problems that can be written in terms of bilinear and linear forms. We therefore expect that it can be applied to a wide range of partial differential equations. Nevertheless, due to space limitations, we present applications of the theory only to linear hyperbolic equations; additional applications are forthcoming. We have shown that our theory constructively leads to two different *hp* finite element methods, namely, an existing least squares discontinuous Galerkin method and a new discontinuous finite element method. We have analyzed the consistency, stability, and *hp*-convergence of these methods in detail. These analytical results are supported by numerical results which show that we have indeed obtained well-posed *hp* finite element methods with optimal convergence rates in the natural energy norms. Moreover, the numerical results show that our new discontinuous finite element method, namely the DPG method, is more accurate and does not seem to have nonphysical diffusion compared to the upwind discontinuous Galerkin method.

Acknowledgements

We thank the anonymous reviewers for their useful comments and suggestions that improved this paper significantly.

References

1. L. Demkowicz, J. Gopalakrishnan, A class of discontinuous Petrov–Galerkin methods. Part I: The transport equation, Computer Methods in Applied Mechanics and Engineering 199 (23-24) (2010) 1558–1572. MR2630162 (2011e:65263)
2. L. Demkowicz, J. Gopalakrishnan, A class of discontinuous Petrov–Galerkin methods. Part II: Optimal test functions, Numerical methods for Partial Differential Equations 27 (1) (2011) 70–105. MR2743600 (2011k:65155)
3. L. Demkowicz, J. Gopalakrishnan, A class of discontinuous Petrov–Galerkin methods. Part IV: The optimal test norm and time-harmonic wave propagation in 1D, Journal Computational Physics 230 (7) (2011) 2406–2432. MR2772923 (2012d:65294)
4. I. Babuška, A. Aziz, Survey lectures on the mathematical foundations of the finite element method, in: A. Aziz (Ed.), The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, Academic Press, New York, 1972, pp. 3–359. MR0421106 (54:9111)
5. J. T. Oden, L. F. Demkowicz, Applied functional analysis, CRC Press, 2010. MR2599487 (2011d:46001)
6. A. Ern, J.-L. Guermond, Theory and Practice of Finite Elements, Vol. 159 of Applied Mathematical Sciences, Spinger-Verlag, 2004. MR2050138 (2005d:65002)
7. I. Babuška, Error bounds for finite element method, Numerische Mathematik 16 (1971) 322–333. MR0288971 (44:6166)
8. D. B. Szyld, The many proofs of an identity on the norm of oblique projections, Numer. Algorithms 42 (2006) 309–323. MR2279449 (2007k:46040)
9. L. Demkowicz, "Babuška ↔ Brezzi?", Tech. Rep. 06-08, Institute for Computational Engineering and Sciences, the University of Texas at Austin (April 2006).
10. J. Xu, L. Zikatanov, Some observations on Babuška and Brezzi theories, Tech. Rep. AM222, Penn State University, http://www.math.psu.edu/ccma/reports.html (September 2000).
11. P. Azerad, Analyse des équations de Navier-Stokes en bassin peu profond et de l'équation de transport, Ph.D. thesis, Neuchatel (1996).
12. C. Johnson, J. Pitkäranta, An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation, Mathematics of Computation 46 (173) (1986) 1–26. MR815828 (88b:65109)

13. I. Babuška, M. Suri, The *hp*-version of the finite element method with quasiuniform meshes, Mathematical Modeling and Numerical Analysis 21 (1987) 199–238. MR896241 (88d:65154)
14. I. Babuška, M. Suri, The optimal convergence rate of the *p*-version of the finite element method, SIAM J. Numer. Anal. 24 (4) (1987) 750–776. MR899702 (88k:65102)
15. I. Babuška, M. Suri, The p and $h - p$ version of the finite element method, basic principles and properties, SIAM Review 36 (4) (1994) 578–632. MR1306924 (96d:65184)
16. C. Schwab, *p*– and *hp*–finite element methods: Theory and applications in solid and fluid mechanics, Oxford University Press, Oxford, 1998. MR1695813 (2000d:65003)
17. P. Houston, M. Jensen, E. Süli, *hp*-Discontinuous Galerkin finite element methods with least-squares stabilization, Journal of Scientific Computing 17 (1–4) (2002) 3–25. MR1910549
18. N. Nguyen, G. Rozza, D. Huynh, A. Patera, Reduced basis approximation and a posteriori error estimation for parametrized parabolic PDEs; Application to real-time Bayesian parameter estimation, in: L. Biegler, G. Biros, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Mallick, L. Tenorio, B. van Bloemen Waanders, K. Willcox (Eds.), Large Scale Inverse Problems and Quantification of Uncertainty, John Wiley & Sons, 2011.
19. T. Bui-Thanh, K. Willcox, O. Ghattas, Model reduction for large-scale systems with high-dimensional parametric input space, SIAM Journal on Scientific Computing 30 (2008) 3270–3288. MR2452388 (2009g:90084)
20. T. Bui-Thanh, Model-constrained optimization methods for reduction of parameterized large-scale systems, Ph.D. thesis, MIT (2007).
21. P. G. Ciarlet, The finite element method for elliptic problems, Vol. 40 of Classics in Applied Mathematics, SIAM, Philadelphia, PA, 2002, reprint of the 1978 original [North-Holland, Amsterdam]. MR0520174 (58:25001)
22. B. Cockburn, G. E. Karniadakis, C.-W. Shu, Discontinuous Galerkin Methods: Theory, Computation and Applications, Lecture Notes in Computational Science and Engineering, Vol. 11, Springer-Verlag, Berlin, Heidelberg, New York, 2000. MR1842160 (2002b:65004)
23. B. Cockburn, J. Gopalakrishnan, R. Lazarov, Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems, SIAM J. Numer. Anal. 47 (2009) 1319–1365. MR2485455 (2010b:65251)
24. W. H. Reed, T. R. Hill, Triangular mesh methods for the neutron transport equation, Tech. Rep. LA-UR-73-479, Los Alamos Scientific Laboratory (1973).
25. P. LeSaint, P. A. Raviart, On a finite element method for solving the neutron transport equation, in: C. de Boor (Ed.), Mathematical Aspects of Finite Element Methods in Partial Differential Equations, Academic Press, 1974, pp. 89–145. MR0658142 (58:31918)
26. T. Bui-Thanh, L. Demkowicz, O. Ghattas, A fast algorithm for inverse transport equation using a discontinuous Petrov-Galerkin method, In preparation.
27. T. E. Peterson, A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation, SIAM J. Numer. Anal. 28 (1) (1991) 133–140. MR1083327 (91m:65250)
28. J. Gopalakrishnan, W. Qiu, An analysis of the practical DPG method, to appear in Math. Comp.
29. M. Drela, Two-dimensional transonic aerodyanmic design and analysis using the Euler equations, Ph.D. thesis, MIT (1986).

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

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

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