Computers and Mathematics with Applications 95 (2021) 172–185

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

# An $L^p$-DPG method for the convection–diffusion problem

Jiaqi Li *, Leszek Demkowicz

*Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin, 201 E 24th St, Austin, TX 78712, USA*

ARTICLE INFO

*Article history:*
Available online 29 August 2020

*Keywords:*
Discontinuous Petrov–Galerkin methods
Residual minimization
Banach spaces
Convection-dominated diffusion

ABSTRACT

Following Muga and van der Zee (Muga and van der Zee, 2015), we generalize the standard Discontinuous Petrov–Galerkin (DPG) method, based on Hilbert spaces, to Banach spaces. Numerical experiments using model 1D convection-dominated diffusion problem are performed and compared with Hilbert setting. It is shown that Banach-based method gives solutions less susceptible to Gibbs phenomenon. h-adaptivity is implemented with the help of the error representation function as error indicator.

Published by Elsevier Ltd.

## 1. Introduction

The Discontinuous Petrov–Galerkin (DPG) Method can be applied to any well-posed variational problem [1], and it is best combined with broken test spaces [2] for most efficiency. The DPG method can be interpreted as a minimum-residual method with the residual measured in a dual norm. Consider the abstract problem

$$\begin{cases} \text{Find } u \in \mathcal{U} : \\ Bu = l \quad \text{in } \mathcal{V}' \end{cases} \tag{1.1}$$

where $\mathcal{U}, \mathcal{V}$ are trial and test spaces (Banach spaces in general), $B : \mathcal{U} \to \mathcal{V}'$ is a bounded linear operator dictated by the problem and the variational formulation we choose. For a well-posed variational problem, $B$ is bounded below as well.

Given a discrete trial space $\mathcal{U}_h \subset \mathcal{U}$, the ideal DPG method (test space is not discretized yet) solves the minimum residual problem

$$\begin{cases} \text{Find } u_h \in \mathcal{U}_h : \\ \|Bu_h - l\|_{\mathcal{V}'} \quad \text{is minimized.} \end{cases} \tag{1.2}$$

In Hilbert space setting, both $\mathcal{U}$ and $\mathcal{V}$ are Hilbert spaces, and we can make use of Riesz operator $R_{\mathcal{V}} : \mathcal{V} \to \mathcal{V}'$, defined by

$$\langle R_{\mathcal{V}} v, \delta v \rangle_{\mathcal{V}',\mathcal{V}} = (v, \delta v)_{\mathcal{V}} \quad \forall \delta v \in \mathcal{V}.$$

By Riesz representation theorem,

$$\|Bu_h - l\|_{\mathcal{V}'} = \|R_{\mathcal{V}}^{-1}(Bu_h - l)\|_{\mathcal{V}}.$$

Consequently, the minimum residual problem (1.2) can be reformulated as

$$\begin{cases} \text{Find } u_h \in \mathcal{U}_h : \\ \|R_{\mathcal{V}}^{-1}(Bu_h - l)\|_{\mathcal{V}}^2 \quad \text{is minimized.} \end{cases} \tag{1.3}$$

* Corresponding author.
*E-mail address:* jiaqi@oden.utexas.edu (J. Li).

https://doi.org/10.1016/j.camwa.2020.08.013
0898-1221/Published by Elsevier Ltd.

The optimality condition translates into

$$\left(R_{\mathcal{V}}^{-1}(Bu_h - l), R_{\mathcal{V}}^{-1}Bw_h\right)_{\mathcal{V}} = 0 \quad \forall w_h \in \mathcal{U}_h. \tag{1.4}$$

Let $\psi \in \mathcal{V}$ be the Riesz representation of the residual, the so called error representation function, i.e.

$$\psi = R_{\mathcal{V}}^{-1}(l - Bu_h). \tag{1.5}$$

The optimality condition (1.4) now reads as

$$\left(\psi, R_{\mathcal{V}}^{-1}Bw_h\right)_{\mathcal{V}} = \langle Bw_h, \psi\rangle_{\mathcal{V}',\mathcal{V}} = 0 \quad \forall w_h \in \mathcal{U}_h. \tag{1.6}$$

Finally, we can reformulate the minimum residual problem (1.3) as a mixed problem

$$\begin{cases} \text{Find } \psi \in \mathcal{V}, u_h \in \mathcal{U}_h : \\ (\psi, v)_{\mathcal{V}} + \langle Bu_h, v\rangle = l(v) \qquad \forall v \in \mathcal{V} \\ \langle Bw_h, \psi\rangle = 0 \qquad\qquad\quad \forall w_h \in \mathcal{U}_h \end{cases} \tag{1.7}$$

where we have omitted subscripts of the duality pairing for simplicity.

The goal of this paper is to replace Hilbert spaces with Banach spaces, focusing on Sobolev spaces $W^{1,p}(\Omega)$, $p \geq 2$ for test spaces. Using the Banach analog of Riesz operator, we will introduce error representation function and derive a mixed problem similar to (1.7). The key difference is that "Riesz operator" for Banach space is no longer linear; hence the need to solve a nonlinear system of equations. Newton's method with line search is applied. The theory is illustrated with numerical experiments for a 1D convection–diffusion model problem using both classical and ultraweak variational formulations. The experiments include h-adaptivity driven by the error representation function $\psi$.

*Related work.* Guermond was the first to approximate first-order PDEs in $L^p$. In [3], He generalizes least-squares methods to residual minimization in $L^p$, and applies the method to solve transport and convection–diffusion equation. Houston, Muga, Roggendorf, and van der Zee [4] prove the inf–sup condition for the convection–diffusion–reaction problem in a non-Hilbert Sobolev space setting. Recently, Houston, Roggendorf, and van der Zee [5] develop a nonlinear Petrov–Galerkin method for the convection–diffusion–reaction equation in the $W_0^{1,p'}(\Omega) - W_0^{1,p}(\Omega)$ setting, where $1/p + 1/p' = 1$. They show that as $p' \to 1$, the Gibbs phenomenon can be eliminated entirely on certain meshes for certain problems. In the development of the theory for DPG in Banach spaces, we follow closely the work of Muga and van der Zee [6].

## 2. Theory: From Hilbert to Banach

### 2.1. Convection–diffusion problem and variational formulations

To stay focused, we will consider a model convection–diffusion problem. Given a domain $\Omega \subset \mathbb{R}^N$, we want to solve

$$-\nabla \cdot (\epsilon \nabla u - \beta u) = f \quad \text{in } \Omega \tag{2.8}$$

where $\epsilon$ is the diffusion coefficient, $\beta$ denotes an incompressible advection field, and $f$ is a source term. We assume a flux boundary condition on the inflow boundary,

$$-\epsilon \frac{\partial u}{\partial n} + \beta_n u = \beta_n u_0 \quad \text{on } \Gamma_{\text{in}} \tag{2.9}$$

where $\beta_n = \beta \cdot n$ and

$$\Gamma_{\text{in}} = \{x \in \Gamma : \beta_n < 0\}.$$

On the remaining, outflow part $\Gamma_{\text{out}}$ of the boundary, homogeneous Dirichlet boundary condition $u = 0$ is imposed. As usual, $n$ denotes the outward normal unit vector on $\Gamma = \partial\Omega$.

*Classical variational formulation.* The standard variational formulation in Hilbert setting [7] is

$$\begin{cases} \text{Find } u \in \mathcal{U} : \\ \displaystyle\int_\Omega \epsilon \nabla u \cdot \nabla v - u\beta \cdot \nabla v = \int_\Omega fv - \int_{\Gamma_{\text{in}}} \beta_n u_0 v \quad \forall v \in \mathcal{V} \end{cases} \tag{2.10}$$

where

$$\mathcal{U} = \mathcal{V} = H^1_{\Gamma_{\text{out}}}(\Omega) := \{u \in H^1(\Omega) : u = 0 \text{ on } \Gamma_{\text{out}}\}.$$

Our goal is to replace the Hilbert trial and test spaces by Banach spaces:

$$\begin{aligned} \mathcal{U} &= W^{1,p'}_{\Gamma_{\text{out}}}(\Omega) := \{u \in W^{1,p'}(\Omega) : u = 0 \text{ on } \Gamma_{\text{out}}\} \\ \mathcal{V} &= W^{1,p}_{\Gamma_{\text{out}}}(\Omega) \end{aligned} \tag{2.11}$$

where $p \geq 2$, $1/p + 1/p' = 1$. For the proof of well-posedness of problem (2.10) and (2.11), we refer the readers to the work of Houston et al. [4]

*Ultraweak variational formulation.* By introducing total flux $\sigma = \epsilon\nabla u - \beta u$, we can rewrite the convection–diffusion problem as a first-order system:

$$\begin{cases} \epsilon^{-1}\sigma - \nabla u + \epsilon^{-1}\beta u = 0 & \text{in } \Omega \\ -\nabla\cdot\sigma = f & \text{in } \Omega \\ -\sigma\cdot n = \beta_n u_0 & \text{on } \Gamma_{\text{in}} \\ u = 0 & \text{on } \Gamma_{\text{out}}. \end{cases} \tag{2.12}$$

Eq. (2.8) can be split into a system of first order equations in more than one way. In particular, Broersen and Stevenson [8] proposed to 'split $\epsilon$ in half' by introducing a 'scaled' flux $\sigma := \epsilon^{1/2}\nabla u$ as the new variable. The split is indeed beneficial for problems with small values of $\epsilon$. In the presented work, we choose for the extra variable the *total flux*; the choice being dictated by the inflow boundary condition (BC) (2.9). Since, at present, we are concerned only with moderately small values of $\epsilon$, the Broersen–Stevenson split benefiting conditioning for small $\epsilon$, is a secondary issue.

We can now multiply the first equation with a vector-valued test function $\tau$, the second equation with a scalar-valued function $v$, and integrate over $\Omega$:

$$\begin{cases} (\epsilon^{-1}\sigma,\tau) - (\nabla u,\tau) + (\epsilon^{-1}\beta u,\tau) = 0 \\ -(\nabla\cdot\sigma, v) = (f, v) \end{cases}$$

where $(\cdot,\cdot)$ denotes standard $L^2(\Omega)$ inner product,

$$(u, v) = \int_\Omega uv\, dx.$$

For vector-valued functions, we take their inner product and then integrate over $\Omega$. Integrating by parts, and making use of boundary conditions, we finally obtain the ultraweak formulation [7]

$$\begin{cases} \text{Find } \sigma \in (L^2(\Omega))^N, u \in L^2(\Omega): \\ \left(\sigma, \epsilon^{-1}\tau\right) + \left(u, \nabla\cdot\tau + \epsilon^{-1}\beta\cdot\tau\right) = 0 & \forall \tau \in H_{\Gamma_{\text{in}}}(\text{div}, \Omega) \\ (\sigma, \nabla v) = (f, v) - \displaystyle\int_{\Gamma_{\text{in}}} \beta_n u_0 v & \forall v \in H^1_{\Gamma_{\text{out}}}(\Omega) \end{cases} \tag{2.13}$$

where

$$H_{\Gamma_{\text{in}}}(\text{div}, \Omega) := \{\tau \in H(\text{div}, \Omega) : \tau\cdot n = 0 \text{ on } \Gamma_{\text{in}}\}.$$

Consequently,

$$\begin{aligned} \mathcal{U} &= (L^2(\Omega))^N \times L^2(\Omega) \\ \mathcal{V} &= H_{\Gamma_{\text{in}}}(\text{div}, \Omega) \times H^1_{\Gamma_{\text{out}}}(\Omega). \end{aligned} \tag{2.14}$$

In the Banach setting, we have

$$\begin{aligned} \mathcal{U} &= (L^{p'}(\Omega))^N \times L^{p'}(\Omega) \\ \mathcal{V} &= W^p_{\Gamma_{\text{in}}}(\text{div}, \Omega) \times W^{1,p}_{\Gamma_{\text{out}}}(\Omega) \end{aligned} \tag{2.15}$$

where $p \geq 2$, $1/p + 1/p' = 1$, and

$$W^p_{\Gamma_{\text{in}}}(\text{div}, \Omega) := \{\tau \in (L^p(\Omega))^N : \text{div } \tau \in L^p(\Omega), \tau\cdot n = 0 \text{ on } \Gamma_{\text{in}}\}.$$

This motivates our study of the analog of Riesz operator for $W^{1,p}(\Omega)$ and $W^p(\text{div}, \Omega) \times W^{1,p}(\Omega)$, which is presented next.

### 2.2. Representation operator

The DPG method minimizes the error in dual norm. The introduction of Riesz operator is the key to computation of the dual norm. In Banach spaces, we need a similar representation operator that achieves this goal. Consider the test space with norm

$$\begin{aligned} \|v\|^p_{\mathcal{V}} &:= \|v\|^p_{L^p(\Omega)} + \|\nabla v\|^p_{(L^p(\Omega))^N} \\ &= \sum_{|\alpha|\leq 1} \|D^\alpha v\|^p_{L^p(\Omega)} \end{aligned} \tag{2.16}$$

for the classical variational formulation and

$$\|(\tau, v)\|^p_{\mathcal{V}} := \|\tau\|^p_{(L^p(\Omega))^N} + \|\text{div } \tau\|^p_{L^p(\Omega)} + \|v\|^p_{L^p(\Omega)} + \|\nabla v\|^p_{(L^p(\Omega))^N} \tag{2.17}$$

for the ultraweak formulation. We shall restrict our attention to $p \in [2, \infty)$.

Let $l \in \mathcal{V}'$. When $\mathcal{V}$ is Hilbert, the Riesz representation of $l$, $R_{\mathcal{V}}^{-1} l$, minimizes the total "potential energy":

$$R_{\mathcal{V}}^{-1} l = \arg\min_{v \in \mathcal{V}} \frac{1}{2}\|v\|_{\mathcal{V}}^2 - l(v). \tag{2.18}$$

Analogously, the "Banach version" of the Riesz representation of $l$, can be defined by considering the $p$-analog of the energy,

$$R_{\mathcal{V}}^{-1} l := \arg\min_{v \in \mathcal{V}} \frac{1}{p}\|v\|_{\mathcal{V}}^p - l(v). \tag{2.19}$$

Equivalently,

$$\langle R_{\mathcal{V}}(v), \delta v\rangle := \langle \partial J(v), \delta v\rangle = l(\delta v) \quad \forall \delta v \in \mathcal{V} \tag{2.20}$$

where $J : \mathcal{V} \to \mathbb{R}$ is defined by

$$J(v) := \frac{1}{p}\|v\|_{\mathcal{V}}^p. \tag{2.21}$$

The "Banach version" of Riesz map $R_{\mathcal{V}} : \mathcal{V} \to \mathcal{V}'$ is given by the Gâteaux derivative of $J$, $R_{\mathcal{V}} = \partial J$. We can compute the Gâteaux derivative explicitly,

$$\begin{cases} \langle R_{\mathcal{V}}(v), \delta v\rangle = \displaystyle\sum_{|\alpha|\le 1} \int_\Omega |D^\alpha v|^{p-2} D^\alpha v D^\alpha \delta v \quad \text{for the classical variational formulation} \\ \langle R_{\mathcal{V}}((\tau, v)), (\delta\tau, \delta v)\rangle = \displaystyle\sum_{i=1}^N \int_\Omega |\tau_i|^{p-2}\tau_i \delta\tau_i + \int_\Omega |\text{div } \tau|^{p-2}\text{div } \tau \text{ div } \delta\tau + \sum_{|\alpha|\le 1}\int_\Omega |D^\alpha v|^{p-2} D^\alpha v D^\alpha \delta v \\ \qquad \text{for the ultraweak variational formulation.} \end{cases} \tag{2.22}$$

**Theorem 1** (*Representation Theorem for the Dual Space*). *$R_{\mathcal{V}} : \mathcal{V} \to \mathcal{V}'$ is well defined, and it is bijective. Moreover, we have*

$$\|R_{\mathcal{V}}(v)\|_{\mathcal{V}'} = \|v\|_{\mathcal{V}}^{p-1}.$$

The proof is relegated to Appendix A. We point out that the presented representation theory is embodied within classical duality-mapping theory involving duality maps with weight functions. Our definition coincides with Definition 2.1 in [5] where the weight function is chosen to be $\varphi(t) = t^{p-1}$.

### 2.3. Minimization of residual in banach spaces

We first invoke the best approximation theorem in Banach spaces (see Theorem 2.A in [6], or Section 10.2 in [9]).

**Theorem 2** (*Best Approximation*). *Let $Y$ be a Banach space, and $y \in Y$.*

- *Suppose $M \subset Y$ is a finite-dimensional subspace, then there exists a best approximation $y_0 \in M$ such that*

$$\|y - y_0\| = \inf_{z \in M} \|y - z\|.$$

- *Suppose $M \subset Y$ is a convex subset of $Y$, and $Y$ is strictly convex. If there exists a best approximation $y_0$ to $y$, then $y_0$ must be unique.*

This theorem is classical. We include a proof for both completeness and readers' convenience.

**Proof.** Suppose first $M \subset Y$ is a finite-dimensional subspace. Then

$$d := \inf_{z \in M} \|y - z\| \le \|y - 0\| = \|y\|.$$

By definition of the infimum, there exists a sequence $\{z_j\}$ in $M$ such that $\|y - z_j\| \to d$. Note that

$$\|z_j\| = \|y - y + z_j\| \le \|y\| + \|y - z_j\|.$$

Therefore, $\{z_j\}$ is a bounded sequence in $M$. By Bolzano–Weierstrass theorem, there exists a subsequence, denoted with the same symbol, such that $z_j \to y_0$ for some $y_0 \in M$. Continuity of norm implies

$$\|y - y_0\| = d = \inf_{z \in M} \|y - z\|.$$

Now let $M \subset Y$ be a convex subset of $Y$, and $Y$ is strictly convex. Let $y_0, y_1$ be distinct best approximations to $y$, such that $\|y - y_0\| = \|y - y_1\| = d$, and $y_0 \neq y_1$. Consider now the convex combination $u = \alpha y_0 + (1-\alpha)y_1$, for any $\alpha \in (0,1)$. Due to convexity of $M$, we know that $u$ is also in $M$. The strict convexity of $Y$ leads to the inequality

$$\begin{aligned}\|y-u\| &= \|\alpha(y-y_0)+(1-\alpha)(y-y_1)\| \\ &< \alpha\|y-y_0\| + (1-\alpha)\|y-y_1\| \\ &= \alpha d + (1-\alpha)d \\ &= d\end{aligned}$$

which implies $u$ is a better approximation to $y$ in $M$, a contradiction. □

Consider the minimum residual problem (1.2) where $\mathcal{V} = W^{1,p}(\Omega)$. Now $Y = \mathcal{V}'$ is Banach, and $M = B\mathcal{U}_h \subset Y$ is a finite-dimensional subspace. Therefore there exists a solution to problem (1.2). Moreover, as shown in (2.22), the norm in $\mathcal{V}$ is Gâteaux differentiable, and we have a classical result that $\mathcal{V}$ is reflexive. The reflexivity of space $\mathcal{V}$ combined with Gâteaux differentiability of norm imply the strict convexity of dual space $\mathcal{V}'$ (see [10], Corollary 5.4.18 and Proposition 5.4.7). Therefore the best approximation to $l$ in $B\mathcal{U}_h$ is unique. The uniqueness of solution $u_h$ to problem (1.2) follows from the injectivity of $B$. Next we derive a necessary condition for $Bu_h$ to be the best approximation to $l$.

Introduce error representation function

$$\psi = R_{\mathcal{V}}^{-1}(l - Bu_h). \tag{2.23}$$

For an arbitrary $w_h \in \mathcal{U}_h$, let

$$\varphi = R_{\mathcal{V}}^{-1}(l - Bw_h) \tag{2.24}$$

be the representation function of the residual. For our problem, $B : \mathcal{U} \to \mathcal{V}'$ is induced by a bilinear form $b$ on $\mathcal{U} \times \mathcal{V}$

$$\langle Bu, v\rangle = b(u,v). \tag{2.25}$$

We seek to minimize $\|\varphi\|_{\mathcal{V}}^{p-1} = \|Bw_h - l\|_{\mathcal{V}'}$, or equivalently, minimize $J(\varphi) = \frac{1}{p}\|\varphi\|_{\mathcal{V}}^p$ under the constraint

$$\langle R_{\mathcal{V}}(\varphi), v\rangle + b(w_h, v) = l(v) \quad \forall v \in \mathcal{V}. \tag{2.26}$$

It can be shown that $J(\varphi)$ is strictly convex in $\varphi$.

Consider the functional $I : \mathcal{V} \times \mathcal{U}_h \to \mathbb{R}$, defined by

$$I(\varphi, w_h) = J(\varphi) + b(w_h, \varphi) - l(\varphi). \tag{2.27}$$

We shall study the sup–inf and inf–sup problems:

$$\sup_{w_h \in \mathcal{U}_h} \inf_{\varphi \in \mathcal{V}} I(\varphi, w_h) \quad \text{and} \quad \inf_{\varphi \in \mathcal{V}} \sup_{w_h \in \mathcal{U}_h} I(\varphi, w_h).$$

Let us look at the sup-inf problem first. For a given $w_h \in \mathcal{U}_h$, $I(\varphi, w_h)$ is the sum of a strictly convex functional and a linear functional in $\varphi$; hence it is strictly convex in $\varphi$. Then $\inf_{\varphi\in\mathcal{V}} I(\varphi, w_h)$ is achieved for the unique $\varphi^*(w_h)$ that satisfies (2.26), i.e., vanishing of Gâteaux derivative. Take the test function $v = \varphi^*$ in (2.26), we get

$$b(w_h, \varphi^*) - l(\varphi^*) = -\langle R_{\mathcal{V}}(\varphi^*), \varphi^*\rangle. \tag{2.28}$$

From the expression for $R_{\mathcal{V}}(\varphi)$, (2.22), we have

$$\langle R_{\mathcal{V}}(\varphi^*), \varphi^*\rangle = \|\varphi^*\|_{\mathcal{V}}^p = pJ(\varphi^*). \tag{2.29}$$

Thus

$$\sup_{w_h \in \mathcal{U}_h} \inf_{\varphi \in \mathcal{V}} I(\varphi, w_h) = \sup_{w_h \in \mathcal{U}_h} (1-p)J(\varphi^*(w_h)) \tag{2.30}$$

where $\varphi^*$ is related to $w_h$ by (2.26). Since we assume $p \geq 2$, we have $1 - p < 0$, and minimization of $J(\varphi)$ is equivalent to maximization of $(1-p)J(\varphi)$. Therefore, our residual minimization problem can be recast as the sup-inf problem

$$\sup_{w_h \in \mathcal{U}_h} \inf_{\varphi \in \mathcal{V}} I(\varphi, w_h).$$

Next we examine the inf-sup problem. For any given $\varphi \in \mathcal{V}$, $I(\varphi, w_h)$ is affine in $w_h$. Its supremum is $+\infty$ unless $b(\delta u_h, \varphi) = 0 \quad \forall \delta u_h \in \mathcal{U}_h$. Thus

$$\inf_{\varphi \in \mathcal{V}} \sup_{w_h \in \mathcal{U}_h} I(\varphi, w_h) = \inf_{\varphi \in (B\mathcal{U}_h)^\perp} J(\varphi) - l(\varphi) \tag{2.31}$$

where

$$(B\mathcal{U}_h)^\perp := \{v \in \mathcal{V} \mid \langle B\delta u_h, v\rangle = 0 \quad \forall \delta u_h \in \mathcal{U}_h\}$$

is the common definition of orthogonal complement. The inf-sup problem is now turned into a standard convex minimization problem, and we invoke a classical result from convex analysis (see Proposition 1.2 in Chapter 2, [11]).

**Lemma 1** (*Existence and Uniqueness of Solutions*). *Let $V$ be a reflexive Banach space, and $\mathcal{C}$ a non-empty closed convex subset of $V$. Assume $F : \mathcal{C} \to \mathbb{R}$ is convex, lower semi-continuous and proper. Moreover, the function $F$ is coercive over $\mathcal{C}$, i.e.*

$$\lim F(u) = +\infty \quad \text{for } u \in \mathcal{C}, \|u\| \to \infty.$$

*Then the problem $\inf_{u\in\mathcal{C}} F(u)$ has at least one solution. It has a unique solution if the function $F$ is strictly convex over $\mathcal{C}$.*

For our problem (2.31), $V = \mathcal{V}$ is reflexive Banach space, and $\mathcal{C} = (B\mathcal{U}_h)^\perp \subset \mathcal{V}$ is non-empty, closed, and convex. $F(\varphi) = J(\varphi) - l(\varphi)$ is strictly convex, continuous (thus lower semi-continuous), and proper. By "proper" we mean it nowhere takes the value $-\infty$ and is not identically equal to $+\infty$. $F$ is also coercive, since $J(\varphi) = \frac{1}{p}\|\varphi\|^p$ while $l(\varphi)$ grows only linearly in $\|\varphi\|$. Therefore problem (2.31) has a unique solution. To characterize the unique solution, we invoke another standard result from convex analysis (Proposition 2.1 in Chapter 2, [11]).

**Lemma 2** (*Characterization of Solutions*). *We assume that the function $F$ satisfies the condition in Lemma 1, and is Gâteaux differentiable with continuous derivative $F'$. Then if $u \in \mathcal{C}$, $u$ is a solution of $\inf_{u\in\mathcal{C}} F(u)$ if and only if*

$$\langle F'(u), v - u\rangle \geq 0 \quad \forall v \in \mathcal{C}.$$

Denote the unique solution of (2.31) by $\psi^*$. Then the lemma requires

$$\langle F'(\psi^*), \varphi - \psi^*\rangle \geq 0 \quad \forall \varphi \in (B\mathcal{U}_h)^\perp. \tag{2.32}$$

Let $v = \varphi - \psi^*$. Then $v$ can take any value in $(B\mathcal{U}_h)^\perp$. The optimality condition (2.32) is true if and only if

$$\langle F'(\psi^*), v\rangle = 0 \quad \forall v \in (B\mathcal{U}_h)^\perp. \tag{2.33}$$

This is equivalent to $F'(\psi^*) \in ((B\mathcal{U}_h)^\perp)^\perp = B\mathcal{U}_h$. The last equality is true since $B\mathcal{U}_h$ is finite dimensional, thus closed. Therefore, there exists $u_h^* \in U_h$, such that

$$F'(\psi^*) + Bu_h^* = 0 \quad \text{in } \mathcal{V}'. \tag{2.34}$$

Plugging in the expression for $F$, and applying the functional on test function $v \in \mathcal{V}$, we get

$$\langle R_{\mathcal{V}}(\psi^*), v\rangle + b(u_h^*, v) = l(v) \quad \forall v \in \mathcal{V}. \tag{2.35}$$

Moreover, $\psi^*$ need to satisfy the orthogonality condition

$$b(\delta u_h, \psi^*) = 0 \quad \forall \delta u_h \in \mathcal{U}_h \tag{2.36}$$

which is equivalent to $\psi^* \in (B\mathcal{U}_h)^\perp$. In summary, the unique solution $\psi^*$ to the inf-sup problem (2.31) can be obtained by solving the mixed system

$$\begin{cases} \text{Find } \psi^* \in \mathcal{V}, u_h^* \in \mathcal{U}_h : & & \\ \langle R_{\mathcal{V}}(\psi^*), v\rangle + b(u_h^*, v) & = l(v) & \forall v \in \mathcal{V} \\ b(\delta u_h, \psi^*) & = 0 & \forall \delta u_h \in \mathcal{U}_h. \end{cases} \tag{2.37}$$

On the other hand, the mixed problem admits a unique solution in $\psi^*$, and $Bu_h^* = l - R_{\mathcal{V}}(\psi^*)$ is also uniquely determined. Since B is an injection, $u_h^*$ is also unique.

Finally we prove the equivalence of the sup–inf and inf–sup problem. It suffices to prove the existence of a saddle point (see Theorem 7.16-1, [12]). Let $(\psi^*, u_h^*)$ solves the mixed problem (2.37). We claim that it is a saddle point of $I(\varphi, w_h)$, i.e.,

$$\sup_{w_h\in\mathcal{U}_h} I(\psi^*, w_h) = I(\psi^*, u_h^*) = \inf_{\varphi\in\mathcal{V}} I(\varphi, u_h^*). \tag{2.38}$$

By direct calculation,

$$I(\psi^*, w_h) - I(\psi^*, u_h^*) = b(w_h - u_h^*, \psi^*) = 0.$$

Thus the first equality in (2.38) holds. On the other hand,

$$I(\psi^*, u_h^*) = \inf_{\varphi\in\mathcal{V}} I(\varphi, u_h^*)$$

because $\psi^*$ makes the Gâteaux derivative of a convex functional vanish. We have proved that $(\psi^*, u_h^*)$ is indeed a saddle point of $I$. Therefore,

$$\sup_{w_h\in\mathcal{U}_h} \inf_{\varphi\in\mathcal{V}} I(\varphi, w_h) = I(\psi^*, u_h^*) = \inf_{\varphi\in\mathcal{V}} \sup_{w_h\in\mathcal{U}_h} I(\varphi, w_h). \tag{2.39}$$

We have managed to show that the minimum residual problem (1.2) is equivalent to the convex optimization problem (2.31), or the mixed problem (2.37). Moreover, the unique solution to the mixed system (2.37), $(\psi^*, u_h^*)$, coincides with the solution $u_h$ to the residual minimization problem (1.2) and its error representation function $\psi$ defined in (2.23). We will utilize this fact in our numerical computations.

## 3. Numerical algorithms and results

### 3.1. Discretization with broken test spaces

In our analysis above, we have only considered discrete trial space $\mathcal{U}_h$; the test space $\mathcal{V}$ is not yet discretized. We will use the standard technique in DPG to discretize $\mathcal{V}$: broken test spaces [1,2]. To stay focused, we consider the convection–diffusion problem (2.8). Suppose we discretize the domain $\Omega$ with a mesh $\Omega_h$. The collection of element boundaries $\partial K$ for all $K \in \Omega_h$, is denoted by $\partial\Omega_h$. The broken Sobolev space is defined by

$$H^1(\Omega_h) = \left\{u \in L^2(\Omega) \mid u|_K \in H^1(K), K \in \Omega_h\right\}. \tag{3.40}$$

Testing (2.8) with discontinuous test functions $v \in H^1(\Omega_h)$ and integrating by parts elementwise, we obtain the corresponding DPG formulation

$$\begin{cases} \text{Find } u \in H^1_{\Gamma_{\text{out}}}(\Omega),\ t \in H^{-1/2}(\partial\Omega_h), t = \beta_n u_0 \text{ on } \Gamma_{\text{in}}: \\ (\epsilon\nabla u, \nabla v)_h - (u, \beta\cdot\nabla v)_h + \langle t, v\rangle = (f, v)_\Omega \quad \forall v \in H^1(\Omega_h) \end{cases} \tag{3.41}$$

where $(\cdot,\cdot)_h$ denotes $L^2$ inner product for broken spaces, i.e.

$$(f, g)_h = \sum_{K\in\Omega_h} \int_K fg\, dx \tag{3.42}$$

$(\cdot,\cdot)_\Omega$ represents $L^2(\Omega)$ inner product, and $\langle\cdot,\cdot\rangle$ stands for the duality pairing of the extra unknown-flux $t$ with broken test functions. Flux $t$ can be identified with the normal trace of a $\sigma \in H(\text{div}, \Omega)$ to element boundaries.

$$\langle t, v\rangle = \sum_{K\in\Omega_h} \langle \sigma|_K \cdot n_K, v|_K\rangle_{\partial K}. \tag{3.43}$$

When we replace Hilbert spaces with Banach spaces (2.11), $\mathcal{V} = W^{1,p}(\Omega_h)$, and the space for flux becomes the trace of $W^{p'}(\text{div}, \Omega)$ to element boundaries, see Appendix B. In 1D, the flux is just numbers at vertex nodes. By introducing broken test functions, we are effectively replacing our original bilinear form with a new one, $b : \mathcal{U}\times\mathcal{V} \to \mathbb{R}$,

$$b((u, t), v) = (\epsilon\nabla u, \nabla v)_h - (u, \beta\cdot\nabla v)_h + \langle t, v\rangle. \tag{3.44}$$

The trial space $\mathcal{U} = \mathcal{U}_0 \times \hat{\mathcal{U}}$, where $\mathcal{U}_0$ denotes the original trial space when we have continuous test functions, and $\hat{\mathcal{U}}$ denotes the flux space. For ultraweak formulation, we can get

$$b((\sigma, u, \hat{\sigma}_n, \hat{u}), (\tau, v)) = \left(\sigma, \epsilon^{-1}\tau + \nabla v\right)_h + \left(u, \nabla\cdot\tau + \epsilon^{-1}\beta\cdot\tau\right)_h - \langle\tau\cdot n, \hat{u}\rangle - \langle\hat{\sigma}_n, v\rangle \tag{3.45}$$

where

$$\begin{aligned} &\tau \in W^p(\text{div}, \Omega_h), \quad && v \in W^{1,p}(\Omega_h), \\ &\sigma \in (L^{p'}(\Omega))^N, && u \in L^{p'}(\Omega), \\ &\hat{\sigma}_n \in W^{-\frac{1}{p'},p'}(\Gamma_h), && \hat{u} \in W^{1-\frac{1}{p'},p'}(\Gamma_h) \end{aligned} \tag{3.46}$$

(See Appendix B for details). This will give us the operator $B : \mathcal{U} \to \mathcal{V}'$ through (2.25). However, now the test space $\mathcal{V}$ is broken, so we need to study how the residual minimization problem is affected.

It suffices to consider $\mathcal{V} = \mathcal{V}_1 \times \mathcal{V}_2$, because the result easily generalizes to finite product $\mathcal{V} = W^{1,p}(\Omega_h) = \prod_{K\in\Omega_h} W^{1,p}(K)$. Suppose $\mathcal{V}_1 = W^{1,p}(\Omega_1)$, $\mathcal{V}_2 = W^{1,p}(\Omega_2)$, $\|(v_1, v_2)\|_{\mathcal{V}} := (\|v_1\|^p_{\mathcal{V}_1} + \|v_2\|^p_{\mathcal{V}_2})^{1/p}$. We define the representation operator on $\mathcal{V}$, $R_{\mathcal{V}} : \mathcal{V} \to \mathcal{V}'$, by

$$\langle R_{\mathcal{V}}(v_1, v_2), (\delta v_1, \delta v_2)\rangle = \langle R_{\mathcal{V}_1}(v_1), \delta v_1\rangle + \langle R_{\mathcal{V}_2}(v_2), \delta v_2\rangle \tag{3.47}$$

where $R_{\mathcal{V}_1}, R_{\mathcal{V}_2}$ is given by (2.22), with $\Omega$ replaced by $\Omega_1$, $\Omega_2$, respectively. Then we claim the following result.

**Theorem 3** (*Representation Theorem for Product Space*). *$R_{\mathcal{V}}$ defined in (3.47) is well defined and bijective. Moreover,*

$$\|R_{\mathcal{V}}(v_1, v_2)\|_{\mathcal{V}'} = \|(v_1, v_2)\|^{p-1}_{\mathcal{V}}.$$

**Proof.** By Theorem 1, $R_{\mathcal{V}_1}$, $R_{\mathcal{V}_2}$ are both bijective, $\|R_{\mathcal{V}_1}(v_1)\|_{\mathcal{V}_1'} = \|v_1\|^{p-1}$, and $\|R_{\mathcal{V}_2}(v_2)\|_{\mathcal{V}_2'} = \|v_2\|^{p-1}$. Note the isomorphism $i : \mathcal{V}' \to \mathcal{V}_1' \times \mathcal{V}_2'$, given by $i(l) = (l_1, l_2)$, where $l_1 \in \mathcal{V}_1'$, $l_2 \in \mathcal{V}_2'$ are defined as

$$l_1(v_1) := l((v_1, 0)), \quad l_2(v_2) := l((0, v_2)).$$

Now $i(R_{\mathcal{V}}(v_1, v_2)) = (R_{\mathcal{V}_1}(v_1), R_{\mathcal{V}_2}(v_2))$. The bijectivity of $R_{\mathcal{V}}(v_1, v_2)$ follows from that of $R_{\mathcal{V}_1}(v_1), R_{\mathcal{V}_2}(v_2)$. Moreover,

$$\begin{aligned}
|\langle R_{\mathcal{V}}(v_1, v_2), (\delta v_1, \delta v_2)\rangle| &\le \|R_{\mathcal{V}_1}(v_1)\|_{\mathcal{V}_1'} \|\delta v_1\|_{\mathcal{V}_1} + \|R_{\mathcal{V}_2}(v_2)\|_{\mathcal{V}_2'} \|\delta v_2\|_{\mathcal{V}_2} \\
&= \|v_1\|_{\mathcal{V}_1}^{p-1} \|\delta v_1\|_{\mathcal{V}_1} + \|v_2\|_{\mathcal{V}_2}^{p-1} \|\delta v_2\|_{\mathcal{V}_2} \\
&\le \left(\|v_1\|_{\mathcal{V}_1}^p + \|v_2\|_{\mathcal{V}_2}^p\right)^{(p-1)/p} \left(\|\delta v_1\|_{\mathcal{V}_1}^p + \|\delta v_2\|_{\mathcal{V}_2}^p\right)^{1/p} \\
&= \|(v_1, v_2)\|_{\mathcal{V}}^{p-1} \|(\delta v_1, \delta v_2)\|_{\mathcal{V}}.
\end{aligned}$$

The equality is achieved when $\delta v_1 = v_1, \delta v_2 = v_2$. Therefore,

$$\|R_{\mathcal{V}}(v_1, v_2)\|_{\mathcal{V}'} = \|(v_1, v_2)\|_{\mathcal{V}}^{p-1}. \quad \square$$

With the theorem proven, we are now in a position to analyze the minimum residual problem (1.2) where $\mathcal{V} = \mathcal{V}_1 \times \mathcal{V}_2 = W^{1,p}(\Omega_1) \times W^{1,p}(\Omega_2)$. Since $\mathcal{V}$ is the product of two reflexive Banach spaces, it is a reflexive Banach space. Our argument in Section 2.3 still holds. The minimum residual problem is equivalent to the convex optimization problem (2.31), or mixed system (2.37). Finally we discretize $\mathcal{V} = W^{1,p}(\Omega_h)$ with piecewise polynomials, which are globally discontinuous.

### 3.2. Newton's method with line search

After discretization, we seek to solve the convex optimization problem subject to linear constraints (see (2.31)):

$$\begin{aligned}
&\text{minimize} \quad f(\psi_h) \\
&\text{subject to} \quad b(\delta u_h, \psi_h) = 0 \quad \forall \delta u_h \in \mathcal{U}_h
\end{aligned} \tag{3.48}$$

where $f(\psi_h) = J(\psi_h) - l(\psi_h)$, and the domain is $\mathcal{V}_h \subset \mathcal{V}$.

Following standard practice in numerical optimization, we use Newton's method to solve the problem (see Section 10.2 in [13]). Define the stiffness matrix $\mathbf{B}_{ij} := b(e_j, g_i)$, where $e_j$ is the $j$th basis function for $\mathcal{U}_h$, and $g_i$ is the $i$th basis function for $\mathcal{V}_h$. Then the linear constraint can be written as

$$\mathbf{B}\boldsymbol{\psi}_h = \mathbf{0} \tag{3.49}$$

where $\boldsymbol{\psi}_h$ is the coefficient vector of $\psi_h$ under the basis $\{g_1, g_2, \ldots, g_n\}$. For the Newton iteration, We can always start with a feasible $\psi_h$. In practice, we start with $\psi_h = 0$. The Newton step $\Delta\boldsymbol{\psi}_{nt}$ at feasible $\psi_h$ is characterized by

$$\begin{bmatrix} \nabla^2 f(\psi_h) & \mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{\psi}_{nt} \\ \boldsymbol{u_h} \end{bmatrix} = \begin{bmatrix} -\nabla f(\psi_h) \\ \mathbf{0} \end{bmatrix}. \tag{3.50}$$

This is similar to what we obtain for DPG in Hilbert space [1]. With the broken test spaces, the Newton step of representation function $\Delta\boldsymbol{\psi}_{nt}$ can be condensed out element-wise. We assemble and solve the linear system for $u_h$; then we compute $\Delta\boldsymbol{\psi}_{nt}$ locally. After obtaining $\Delta\boldsymbol{\psi}_{nt}$, we do a backtracking line search to ensure the Armijo sufficient decrease condition (see Section 9.2 in [13]):

$$f(\psi_h + t\Delta\psi_{nt}) \le f(\psi_h) + \alpha t \nabla f(\psi_h)^T \Delta\psi_{nt} \tag{3.51}$$

where $\alpha$ is some constant in $(0, 1)$. In our computations, we choose $\alpha = 10^{-4}$.

The Newton decrement is defined as

$$\lambda(\psi_h) = (\Delta\boldsymbol{\psi}_{nt}^T \nabla^2 f(\psi_h) \Delta\boldsymbol{\psi}_{nt})^{1/2} \tag{3.52}$$

and serves as an error indicator for Newton's method. We stop the Newton iteration when $\lambda$ is small enough. The tolerance is set to $10^{-5}$ in our numerical experiments.

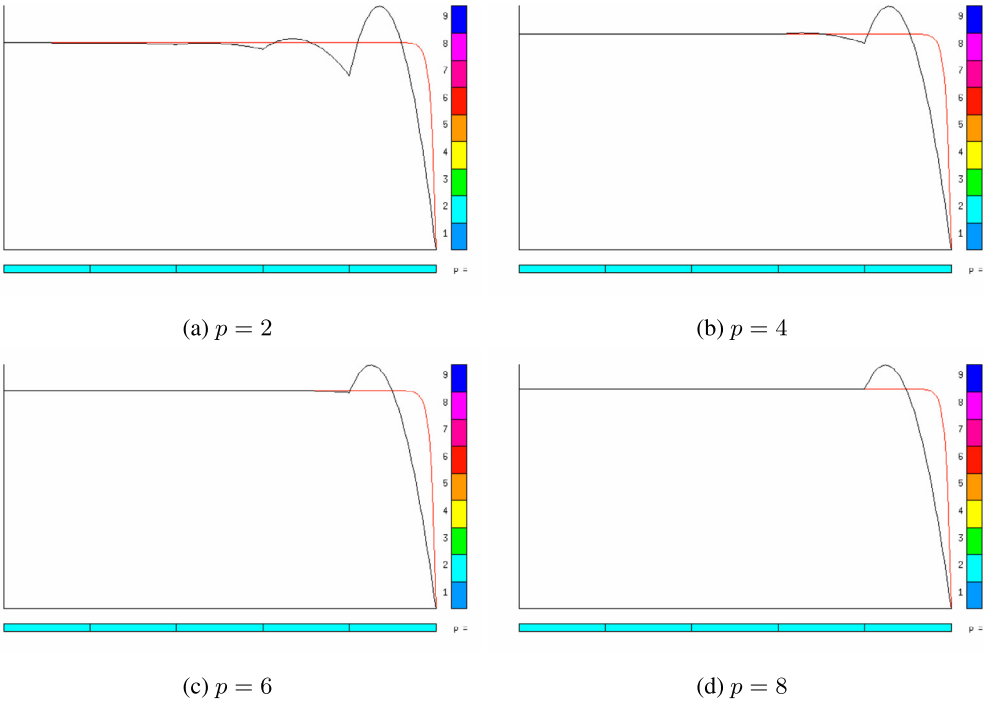
### 3.3. Numerical results for 1D problem

As an illustration, we solve the 1D convection-dominated diffusion problem:

$$\begin{cases} -\epsilon u'' + u' &= 0 \quad \text{in } (0, 1) \\ -\epsilon u' + u &= 1 \quad \text{at } x = 0 \\ u &= 0 \quad \text{at } x = 1. \end{cases} \tag{3.53}$$

We set $\epsilon = 10^{-2}$, and use a uniform mesh consisting of 5 elements. The choice of polynomial degree and our terminology follow the logic of the 1D polynomial exact sequence,

$$\begin{array}{ccc} H^1(0, 1) & \xrightarrow{\partial} & L^2(0, 1) \\ \cup & & \cup \\ \mathcal{P}^r(0, 1) & \xrightarrow{\partial} & \mathcal{P}^{r-1}(0, 1). \end{array}$$



(a) $p = 2$ (b) $p = 4$

(c) $p = 6$ (d) $p = 8$

**Fig. 1.** Solution of classical variational formulation for different $p$. The black line represents numerical solution, and the red line stands for the exact one.

By the order of elements, displayed in the following figures, we mean always the order $r$ of $H^1$-conforming element. The order of the corresponding $L^2$-conforming element is then $r-1$. The choice implies that the best $H^1$- and $L^2$-approximation errors, converge with the same rate.

We begin with the classical variational formulation. Second-order elements are used for the trial space $\mathcal{U}_h$, and fourth order elements for the test space $\mathcal{V}_h$. Fig. 1 shows the solution for different choices of $p$. For $p = 2$, our method coincides with the DPG method in Hilbert spaces. In this case, the numerical solution exhibits a clear overshoot near the boundary layer at $x = 1$. This is known as Gibbs phenomenon, which occurs when we approximate discontinuous/boundary-layer problems with continuous functions. As we increase $p$ (in the $W^{1,p}$ norm for the test space), oscillation is localized to the last element. It can also be seen that the solution does not change much if we further increase $p$ over 4.

For the ultraweak formulation, we use quadratic elements for the trial space as well. This means that $\mathcal{U}_h \subset L^{p'}(\Omega)$ is discretized with piecewise linear polynomials, in accordance with the exact sequence logic. For test space $\mathcal{V}_h$, we continue using 4th order elements. Fig. 2 illustrates behavior of the solution as we increase $p$. The oscillation is again localized. We do not go for higher $p$ like $p = 8$ for two reasons. First, the solution does not change much when we increase $p$ over 4. Second, while increasing $p$, we need more integration points for element integration and the condition number of Gram matrix grows fast; actually the ill-conditioning becomes so bad that the iteration for $p = 8$ does not converge.

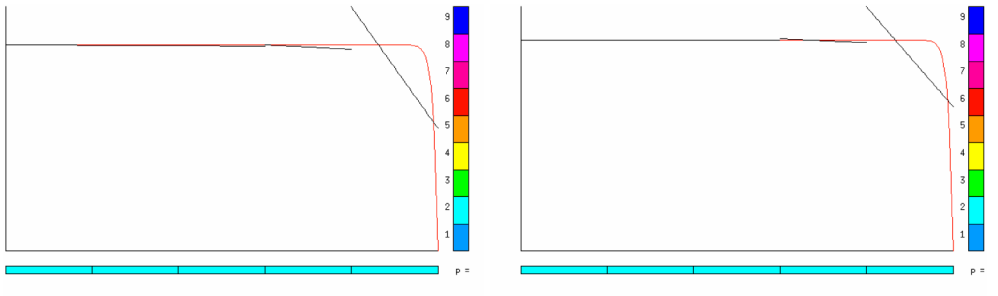
We implement $h$-adaptivity based on $\psi$. The following greedy algorithm is used:

- Solve the problem on the current mesh.
- Compute element residual $\|\psi\|^p$, and mark all elements that have residual larger than $\frac{1}{4}\|\psi\|^p_{\max}$.
- Refine each marked element; continue.

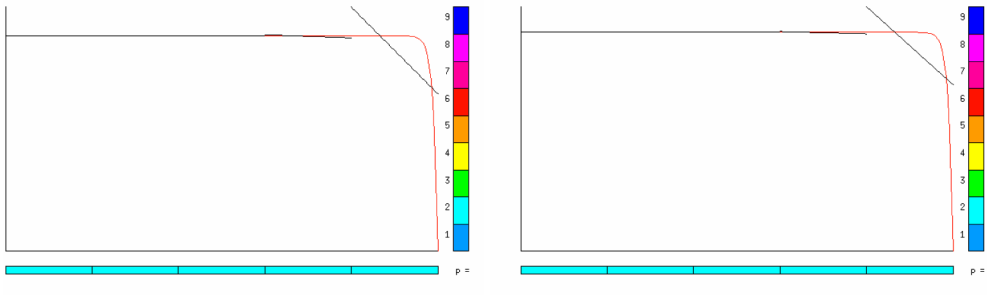
Solution of the problem using $h$-adaptivity is shown in Fig. 3. $\epsilon$ and polynomial orders are unchanged, and $p$ is set to 4. We start with a mesh consisting of 5 uniformly-spaced elements. The element containing the boundary layer is refined in each iteration step. This verifies that the largest error comes from elements near the boundary layer, as we observe in Figs. 1 and 2. After 4 refinements, the numerical solution and the exact one is almost indistinguishable.

## 4. Conclusion

We have presented a generalization of the DPG method to the case of reflexive Banach test spaces $W^{1,p}(\Omega)$ with $p \geq 2$. Our generalization is based on minimum residual method. The relation between the residual minimization problem and the mixed problem is explored, with the introduction of a Lagrangian $l(\varphi, w_h)$.

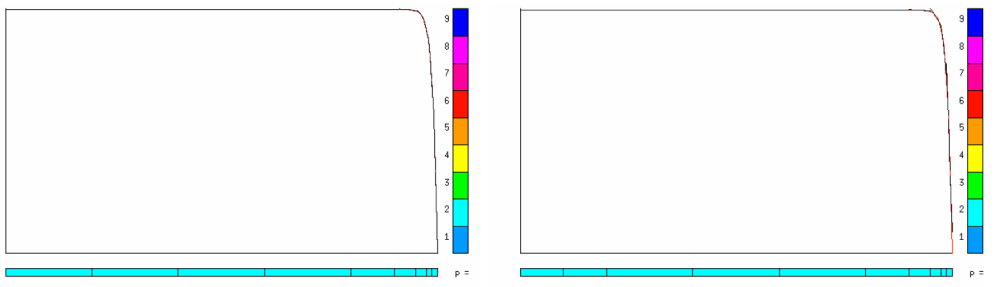


(a) $p = 2$ (b) $p = 3$



(c) $p = 4$ (d) $p = 5$

**Fig. 2.** Solution of ultraweak formulation for different $p$. The black line represents numerical solution, and the red line stands for the exact one.



(a) Classical variational formulation (b) Ultraweak formulation

**Fig. 3.** $h$-adaptivity result after 4 refinements. Starting mesh consists of 5 elements and is uniformly spaced. $p = 4$. The black line represents numerical solution, and the red line stands for the exact one.

With broken test spaces and Newton's method, we get a linear system similar to what we have for DPG in Hilbert space. Error representation function can be condensed out elementwise, and we solve for $u_h$ first; then $\psi$ is computed in each element. We have performed numerical experiments by solving 1D convection-dominated diffusion. It is demonstrated that the DPG method in Banach space localizes the oscillations. $h$-adaptivity based on the error representation function works well. Future work is to apply our method to 2D and 3D problems. In the meanwhile, new theory regarding traces of Banach spaces may need to be developed.

## CRediT authorship contribution statement

**Jiaqi Li:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization. **Leszek Demkowicz:** Conceptualization, Supervision, Writing - review & editing.

## Acknowledgment

J. Li and L. Demkowicz were partially supported with NSF, USA grant No. 1819101.

## Appendix A. Proof of representation theorem for Banach space

In the appendix, we present an elementary proof of Theorem 1. First consider $\mathcal{V} = W^{1,p}(\Omega)$. From the expression of $R_{\mathcal{V}}$, (2.22), we have

$$\begin{aligned}
|\sum_{|\alpha|\leq 1}\int_\Omega |D^\alpha v|^{p-2}D^\alpha v D^\alpha \delta v| &\leq \sum_{|\alpha|\leq 1}\int_\Omega |D^\alpha v|^{p-1}|D^\alpha \delta v| \\
&\leq \sum_{|\alpha|\leq 1}\left(\int_\Omega |D^\alpha v|^p\right)^{\frac{p-1}{p}}\left(\int_\Omega |D^\alpha \delta v|^p\right)^{1/p} \\
&= \sum_{|\alpha|\leq 1}\|D^\alpha v\|^{p-1}_{L^p(\Omega)}\|D^\alpha \delta v\|_{L^p(\Omega)} \\
&\leq \left(\sum_{|\alpha|\leq 1}\|D^\alpha v\|^p_{L^p(\Omega)}\right)^{\frac{p-1}{p}}\left(\sum_{|\alpha|\leq 1}\|D^\alpha \delta v\|^p_{L^p(\Omega)}\right)^{1/p} \\
&= \|v\|^{p-1}_{\mathcal{V}}\|\delta v\|_{\mathcal{V}}.
\end{aligned} \tag{A.54}$$

Equality is achieved for $\delta v = v$. Therefore $R_{\mathcal{V}}(v) \in \mathcal{V}'$, and

$$\|R_{\mathcal{V}}(v)\|_{\mathcal{V}'} = \|v\|^{p-1}_{\mathcal{V}}. \tag{A.55}$$

It remains for us to prove $R_{\mathcal{V}} : \mathcal{V} \to \mathcal{V}'$ is one to one and onto. Note that $R_{\mathcal{V}}$ is the Gâteaux derivative of $J : \mathcal{V} \to \mathbb{R}$, and $J$ is strictly convex. Thus $R_{\mathcal{V}}$ is strictly monotonic, i.e.

$$\langle R_{\mathcal{V}}(u) - R_{\mathcal{V}}(v), u - v\rangle > 0 \quad \text{for } u \neq v. \tag{A.56}$$

For $u \neq v$, we have $R_{\mathcal{V}}(u) - R_{\mathcal{V}}(v) \neq 0$, hence the injectivity of $R_{\mathcal{V}}$. Finally, to prove that $R_{\mathcal{V}}$ is surjective, consider the variational problem: given $l \in \mathcal{V}'$,

$$\begin{cases} \text{Find } v \in \mathcal{V} : \\ \langle R_{\mathcal{V}}(v), \delta v\rangle = l(\delta v) \quad \forall \delta v \in \mathcal{V}. \end{cases} \tag{A.57}$$

Due to the strict convexity of $J$, the variational problem (A.57) is equivalent to the minimization problem

$$v = \arg\min_{w\in\mathcal{V}} J(w) - l(w) \tag{A.58}$$

which has a unique solution by Lemma 1. The arguments for $\mathcal{V} = W^p(\text{div}, \Omega) \times W^{1,p}(\Omega)$ are identical.

## Appendix B. Traces for $L^p$ Banach spaces

Much of the DPG theory in $L^2$ Hilbert setting extends seamlessly to the $L^p$ Banach setting for $p \in (1, \infty)$. In this Appendix, we sketch the necessary theory for traces in multi-space dimensions, functional setting for the formulations with broken test spaces, and prove the well-posedness of the "broken" variational formulations. The critical result reported here is the Banach version of the duality lemma in [2].

We begin by recalling the Trace Theorem for spaces $W^{1,p}(\Omega)$ due to Gagliardo [14], see also [15].

**Theorem 4.** *Let $\Omega \subset \mathbb{R}^N$ be a Lipschitz domain, $N = 1, 2, 3, \ldots$, $\Gamma = \partial\Omega$ be its boundary, and let $p \in [1, \infty)$. There exists a unique continuous and surjective trace operator*

$$W^{1,p}(\Omega) \ni V \to v := \gamma V \in W^{1-\frac{1}{p},p}(\Gamma).$$

We shall stick with the reflexive spaces only, i.e. $p \in (1, \infty)$. We can use the surjectivity of the trace operator to replace the intrinsic norm on $W^{1-\frac{1}{p},p}(\Gamma)$ with the minimum extension norm,

$$\|v\|_{W^{1-\frac{1}{p},p}(\Gamma)} := \inf_{\substack{V\in W^{1,p}(\Omega)\\ \gamma V = v}} \|V\|_{W^{1,p}(\Omega)}. \tag{B.59}$$

We define now,

$$\begin{aligned} W^p(\text{div}, \Omega) &:= \{\sigma \in (L^p(\Omega))^N \ : \ \text{div}\,\sigma \in L^p(\Omega)\} \\ W^{-\frac{1}{p},p}(\Gamma) &:= (W^{1-\frac{1}{p'},p'}(\Gamma))' \end{aligned} \tag{B.60}$$

where $p'$ is the conjugate index to $p$: $1/p + 1/p' = 1$, and where space $W^{-\frac{1}{p},p}(\Gamma)$ is equipped with the standard dual norm. The integration by parts identity allows us to introduce the *normal trace operator*. Let $\sigma \in W^p(\mathrm{div}, \Omega)$. Define a functional $\sigma_n \in (W^{1-\frac{1}{p'},p'}(\Gamma))'$ by

$$\langle \sigma_n, v \rangle = \int_\Omega \mathrm{div}\,\sigma\, V + \int_\Omega \sigma \cdot \nabla V \tag{B.61}$$

where $v \in W^{1-\frac{1}{p'},p'}(\Gamma)$, $V \in W^{1,p'}(\Omega)$ is any function such that $\gamma V = v$. Then $\sigma_n$ is well-defined. Indeed, one have to show only that

$$\int_\Omega \mathrm{div}\,\sigma\, V + \int_\Omega \sigma \cdot \nabla V = 0 \tag{B.62}$$

for all $V \in W^{1,p'}(\Omega)$, $\gamma V = 0$. This is a consequence of the well-known fact that

$$\{V \in W^{1,p'}(\Omega) \,:\, \gamma V = 0\} = W_0^{1,p'}(\Omega) := \overline{C_0^\infty(\Omega)}^{W^{1,p'}(\Omega)}.$$

For any $V \in C_0^\infty(\Omega)$, (B.62) holds simply by definition of the div operator; moreover, the left hand side of (B.62) is a continuous function of $V$ in $W^{1,p'}(\Omega)$ norm. A density argument proves (B.62) for any $V \in W_0^{1,p'}(\Omega)$. Hölder inequality implies that $\sigma_n$ is continuous and

$$\|\sigma_n\|_{(W^{1-\frac{1}{p'},p'}(\Gamma))'} \leq \|\sigma\|_{W^p(\mathrm{div},\Omega)}.$$

The operator

$$\gamma_n \,:\, W^p(\mathrm{div}, \Omega) \ni \sigma \to \gamma_n \sigma := \sigma_n \in W^{-\frac{1}{p},p}(\Gamma) \tag{B.63}$$

defines thus a continuous trace operator. We shall show momentarily that the operator is surjective.

**Lemma 3.** *Let $p \in (1, \infty)$ and $p'$ denote the conjugate exponent. Let $u, v \in \mathbb{R}$. Then*

$$u = |v|^{p'-2}v \quad \Leftrightarrow \quad v = |u|^{p-2}u\,.$$

**Proof.** The proof relies on simple algebra. Let $u = |v|^{p'-2}v$. Then

$$\begin{aligned} |u|^{p-2} &= |v|^{(p'-2)(p-2)}|v|^{p-2} = |v|^{(p'-1)(p-2)} \\ &= |v|^{p'p-p-2p'+2} = |v|^{2-p'} \qquad (p'p = p' + p)\,. \end{aligned}$$

Consequently,

$$v = u|v|^{2-p'} = |u|^{p-2}u\,. \quad \square$$

Let now $\sigma_n \in W^{-\frac{1}{p},p}(\Gamma)$. Consider the "Banach version" of Riesz representation of $\sigma_n$,

$$V = \underset{U \in W^{1,p'}(\Omega)}{\arg\min} \frac{1}{p'}\|U\|^{p'}_{W^{1,p'}(\Omega)} - \langle \sigma_n, \gamma U \rangle\,.$$

$V$ satisfies the following Neumann boundary-value problem,

$$\begin{cases} |V|^{p'-2}V - \displaystyle\sum_j \frac{\partial}{\partial x_j}\left(\left|\frac{\partial V}{\partial x_j}\right|^{p'-2}\frac{\partial V}{\partial x_j}\right) &= 0 \quad \text{in } \Omega \\ \displaystyle\sum_j \left|\frac{\partial V}{\partial x_j}\right|^{p'-2}\frac{\partial V}{\partial x_j} n_j &= \sigma_n \quad \text{on } \Gamma\,. \end{cases} \tag{B.64}$$

Define now $\sigma_j = |\frac{\partial V}{\partial x_j}|^{p'-2}\frac{\partial V}{\partial x_j}$. The equation above implies that $\mathrm{div}\,\sigma = |V|^{p'-2}V$. Lemma 3 implies that

$$V = |\mathrm{div}\,\sigma|^{p-2}\,\mathrm{div}\,\sigma \quad \text{and} \quad \frac{\partial V}{\partial x_j} = |\sigma_j|^{p-2}\sigma_j\,.$$

This implies that $\sigma$ satisfies the Dirichlet boundary-value problem,

$$\begin{cases} |\sigma_j|^{p-2}\sigma_j - \dfrac{\partial}{\partial x_j}\left(|\mathrm{div}\,\sigma|^{p-2}\,\mathrm{div}\,\sigma\right) &= 0 \quad \text{in } \Omega \\ \sigma \cdot n &= \sigma_n \quad \text{on } \Gamma\,. \end{cases} \tag{B.65}$$

In other words, $\sigma \in W^p(\mathrm{div}, \Omega)$ is a minimum-energy extension of $\sigma_n$. A direct computation shows that the minimum energy extension norm coincides with the dual norm of $\sigma_n$. In exactly the same way, we show that the dual norm to the minimum energy extension norm for space $W^{-\frac{1}{p},p}(\Gamma)$ coincides with the minimum energy extension norm for $W^{1-\frac{1}{p'},p'}(\Gamma)$.

**Remark.** Our formal proof showing that the minimization problem for $V$, equivalent with Neumann boundary-value problem (B.64), is equivalent to Dirichlet boundary-value problem (B.65) for $\sigma$, can be made fully precise by using the duality theory [11]. Indeed, (B.65) corresponds to a maximization problem dual to the minimization problem for $V$. This explains why we term the result as the *duality lemma*.

**Theorem 5.** *Normal trace operator (B.63) defines is a continuous surjection with a norm equal one. The boundary spaces $W^{1-\frac{1}{p'},p'}(\Gamma)$ and $W^{-\frac{1}{p},p}(\Gamma)$ equipped with the minimum energy extension norms form a duality pairing.*

Existence of trace operators opens now up the analysis for broken variational formulations in the same way as in [2]. We start by introducing the broken test spaces.

$$\begin{aligned} W^{1,p}(\Omega_h) &:= \textstyle\prod_{K\in\Omega_h} W^{1,p}(K) \\ W^p(\mathrm{div}, \Omega_h) &:= \textstyle\prod_{K\in\Omega_h} W^p(\mathrm{div}, K). \end{aligned} \tag{B.66}$$

Let $\Gamma_h$ denote now the mesh skeleton. We introduce the trace spaces defined on the skeleton in the usual way,

$$\begin{aligned} W^{-\frac{1}{p'},p'}(\Gamma_h) &:= \{\sigma_n = \{\sigma_{K,n}\} \in \textstyle\prod_{K\in\Omega_h} W^{-\frac{1}{p'},p'}(\partial K) \,:\, \exists \sigma \in W^{p'}(\mathrm{div}, \Omega) \,:\, \gamma_{n,\partial K}\sigma|_K = \sigma_{K,n}\} \\ W^{1-\frac{1}{p'},p'}(\Gamma_h) &:= \{u = \{u_K\} \in \textstyle\prod_{K\in\Omega_h} W^{1-\frac{1}{p'},p'}(\partial K) \,:\, \exists U \in W^{1,p'}(\Omega) \,:\, \gamma_{\partial K} U|_K = u_K\}. \end{aligned} \tag{B.67}$$

By construction, the duality pairings on the mesh skeleton are well-defined,

$$\begin{aligned} \langle \hat{\sigma}_n, v\rangle_{\Gamma_h} &= \textstyle\sum_{K\in\Omega_h} \langle \sigma_{K,n}, \gamma_{\partial K} v_K\rangle_{\partial K} \quad \hat{\sigma}_n \in W^{-\frac{1}{p'},p'}(\Gamma_h),\ v \in W^{1,p}(\Omega_h) \\ \langle \hat{u}, \tau\rangle_{\Gamma_h} &= \textstyle\sum_{K\in\Omega_h} \langle u_K, \gamma_{n,\partial K}\tau_K\rangle_{\partial K} \quad \hat{u} \in W^{1-\frac{1}{p'},p'}(\Gamma_h),\ \tau \in W^p(\mathrm{div}, \Omega_h). \end{aligned}$$

Also, by the standard density arguments, we have,

$$\begin{aligned} v \in W^{1,p}(\Omega_h), \quad \langle \hat{\sigma}_n, v\rangle_{\Gamma_h} = 0 \quad \forall\, \hat{\sigma}_n \in W^{-\frac{1}{p'},p'}(\Gamma_h) \quad &\Leftrightarrow \quad v \in W^{1,p}(\Omega) \\ \tau \in W^p(\mathrm{div}, \Omega_h), \quad \langle \hat{u}, \tau\rangle_{\Gamma_h} = 0 \quad \forall\, \hat{u} \in W^{1-\frac{1}{p'},p'}(\Gamma_h) \quad &\Leftrightarrow \quad \tau \in W^p(\mathrm{div}, \Omega). \end{aligned}$$

*Well-posedness of broken variational formulations.* Consider now the Banach version of classical variational formulation (2.10),

$$\begin{cases} \text{Find } u \in W^{1,p'}(\Omega),\ u = 0 \text{ on } \Gamma_{\mathrm{out}} \\ (\epsilon\nabla u - \beta u, \nabla v) = (f, v) - \displaystyle\int_{\Gamma_{\mathrm{in}}} \beta_n u_0\, v \quad \forall v \in W^{1,p}(\Omega) \,:\, v = 0 \text{ on } \Gamma_{\mathrm{out}}. \end{cases} \tag{B.68}$$

The "broken" version of the formulation reads as follows.

$$\begin{cases} \text{Find } u \in W^{1,p'}(\Omega),\ \hat{\sigma}_n \in W^{-\frac{1}{p'},p'}(\Gamma_h) \,: \\ u = 0 \text{ on } \Gamma_{\mathrm{out}}, \quad \hat{\sigma}_n = -\beta_n u_0 \text{ on } \Gamma_{\mathrm{in}} \\ (\epsilon\nabla u - \beta u, \nabla_h v) - \langle \hat{\sigma}_n, v\rangle_{\Gamma_h} = (f, v) \quad \forall v \in W^{1,p}(\Omega_h). \end{cases} \tag{B.69}$$

As usual, $\nabla_h$ denotes the gradient computed element-wise.

Similarly, consider the Banach version of the ultraweak formulation (2.13),

$$\begin{cases} \text{Find } \sigma \in (L^{p'}(\Omega))^N,\ u \in L^{p'}(\Omega) \,: \\ (\sigma, \epsilon^{-1}\tau) + (u, \mathrm{div}\,\tau + \epsilon^{-1}\beta\cdot\tau) = 0 & \forall\, \tau \in W^p(\mathrm{div}, \Omega) \,:\, \tau_n = 0 \text{ on } \Gamma_{\mathrm{in}} \\ (\sigma, \nabla v) = (f, v) - \displaystyle\int_{\Gamma_{\mathrm{in}}} \beta_n u_0 v & \forall\, v \in W^{1,p}(\Omega) \,:\, v = 0 \text{ on } \Gamma_{\mathrm{out}} \end{cases} \tag{B.70}$$

with the corresponding "broken" version,

$$\begin{cases} \text{Find } \sigma \in (L^{p'}(\Omega))^N, u \in L^{p'}(\Omega), \hat{\sigma}_n \in W^{-\frac{1}{p'},p'}(\Gamma_h), \hat{u} \in W^{1-\frac{1}{p'},p'}(\Gamma_h) \,: \\ \hat{\sigma}_n = -\beta_n u_0 \text{ on } \Gamma_{\mathrm{in}},\ \hat{u} = 0 \text{ on } \Gamma_{\mathrm{out}} \\ (\sigma, \epsilon^{-1}\tau) + (u, \mathrm{div}_h\,\tau + \epsilon^{-1}\beta\cdot\tau) - \langle \hat{u}, \tau\rangle_{\Gamma_h} = 0 & \forall\, \tau \in W^p(\mathrm{div}, \Omega_h) \\ (\sigma, \nabla_h v) - \langle \hat{\sigma}_n, v\rangle_{\Gamma_h} = (f, v) & \forall\, v \in W^{1,p}(\Omega_h). \end{cases} \tag{B.71}$$

**Theorem 6.** *Assume variational problems (B.68) and (B.70) are well-posed. Then the broken counterparts (B.69) and (B.71) are well-posed as well, with inf-sup constants of the same order as those for the original formulations.*

**Proof.** Proof is identical to the reasoning in [2]. □

## References


[1] L. Demkowicz, J. Gopalakrishnan, Discontinuous Petrov–Galerkin (DPG) Method, ICES Report, 15-20, 2015.
[2] C. Carstensen, L. Demkowicz, J. Gopalakrishnan, Breaking spaces and forms for the DPG method and applications including Maxwell equations, Comput. Math. Appl. 72 (2016) 494–522.
[3] J.L. Guermond, A finite element technique for solving first-order PDEs in $L^p$, SIAM J. Numer. Anal. 42 (2004) 714–737.
[4] P. Houston, I. Muga, S. Roggendorf, K.G. van der Zee, The convection-diffusion-reaction equation in non-Hilbert Sobolev spaces: A direct proof of the inf-sup condition and stability of Galerkin's method, Comput. Methods Appl. Math. 19 (2019) 503–522.
[5] P. Houston, S. Roggendorf, K.G. van der Zee, Eliminating Gibbs phenomena: A non-linear Petrov–Galerkin method for the convection–diffusion–reaction equation, Comput. Math. Appl. 80 (2020) 851–873.
[6] I. Muga, K.G. van der Zee, Discretization of linear problems in Banach spaces: Residual minimization, nonlinear Petrov–Galerkin, and monotone mixed methods, 2015, arXiv e-prints, p. arXiv:1511.04400.
[7] L. Demkowicz, Various Variational Formulations and Closed Range Theorem, ICES Report, 15-03, 2015.
[8] D. Broersen, R.P. Stevenson, A robust Petrov-Galerkin discretisation of convection–diffusion equations, Comput. Math. Appl. 68 (2014) 1605–1618.
[9] I. Stakgold, M.J. Holst, Green'S Functions and Boundary Value Problems, Vol. 99, John Wiley & Sons, 2011.
[10] R.E. Megginson, An Introduction to Banach Space Theory, first ed., in: Graduate Texts in Mathematics, Springer, 1998.
[11] I. Ekeland, R. Temam, Convex Analysis and Variational Problems, Vol. 28, SIAM, 1999.
[12] P.G. Ciarlet, Linear and Nonlinear Functional Analysis with Applications, Vol. 130, SIAM, 2013.
[13] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.
[14] E. Gagliatrdo, Caratterizzazioni delle tracce sulla frontiera relative ad alcune classi d funzioni in $n$-variabili, Rend. Semin. Mat. Univ. Padova 27 (1957) 284–305.
[15] G. Geymonat, Trace theorem for Sobolev spaces on Lipschitz domains. Necessary conditions, Ann. Math. Blaise Pascal 14 (2007) 187–197.