Research Article

Jiaqi Li* and Leszek Demkowicz

An L^p-DPG Method with Application to 2D Convection-Diffusion Problems

https://doi.org/10.1515/cmam-2021-0158
Received August 25, 2021; revised December 27, 2021; accepted March 7, 2022

Abstract: This article summarizes the L^p-DPG method presented in [18], where only 1D convection-diffusion problems are solved. We apply the same computational techniques to 2D convection-diffusion problems and report additional numerical results herein. Furthermore, we propose an L^p-DPG method with variable p and illustrate it with numerical experiments.

Keywords: Discontinuous Petrov–Galerkin Methods, Residual Minimization, Banach Spaces, Convection-Dominated Diffusion, Fortin Operators, $p(\cdot)$-Laplacian

MSC 2010: 65N30

1 Introduction

The Discontinuous Petrov–Galerkin (DPG) method has been proposed as a novel approach to designing finite element methods [4, 8, 9], and it offers many attractive features: guaranteed stability provided the problem is well posed, built-in a posteriori error estimator, as well as the ability to control the norm in which the convergence occurs. The DPG method admits the interpretation of a minimum residual method, where the residual is measured in a dual space to the space of test functions. Consider the following abstract problem:

$$\begin{cases} \text{find } u \in \mathcal{U} : \\ Bu = l \quad \text{in } \mathcal{V}', \end{cases}$$

where $\mathcal{U}, \mathcal{V}$ are trial and test spaces (Banach spaces in general), $B\colon \mathcal{U} \to \mathcal{V}'$ 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 (by *ideal*, we mean the test space is not yet discretized) solves the following minimum residual problem:

$$\begin{cases} \text{find } u_h \in \mathcal{U}_h : \\ \|Bu_h - l\|_{\mathcal{V}'} \text{ is minimized.} \end{cases} \tag{1.1}$$

Originally, the DPG method has dealt with Hilbert test and trial spaces only. Following the pioneering work by van der Zee et al. [15, 16, 19], we have investigated the DPG method in Banach spaces [18], focusing on Sobolev spaces $W^{1,p}(\Omega)$ and $W^p(\text{div}, \Omega)$ ($p \geq 2$) as test spaces, in particular. The trial spaces are chosen accordingly so as to ensure that the bilinear form $\langle Bu, v\rangle$ is bounded. In Banach spaces, the minimum residual problem (1.1) is shown to be equivalent to a convex minimization problem with linear constraints [19]. We solve the latter minimization problem using Newton's method, which will be detailed later. Our expe-

***Corresponding author: Jiaqi Li,** Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin, 201 E 24th St, Austin, TX 78712, USA, e-mail: jiaqi@utexas.edu. https://orcid.org/0000-0002-2415-8262
Leszek Demkowicz, Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin, 201 E 24th St, Austin, TX 78712, USA, e-mail: leszek@oden.utexas.edu. https://orcid.org/0000-0001-7839-8037

rience indicates that, when the residual is small, we may have trouble with Newton's method because the Hessian can become ill-conditioned. We propose an effective solution to the ill-conditioning problem based on a formulation with a variable exponent p.

This article first summarizes the results of our previous work [18]; for convenience, the details of discretization and Newton's method are described in a separate section. Then we proceed by solving various 2D convection-diffusion problems and present the numerical results. Finally, we conclude with a section on the L^p-DPG method with variable p.

2 Theory: L^p-DPG Method for the Convection-Diffusion Problem

We summarize the main theoretical result of [18] in this section. Consider problem (1.1), and assume that B is both bounded and bounded-below. Moreover, $\mathcal{U}_h$ is finite dimensional.

Theorem 1 (Existence and Uniqueness of the Solution). *When $\mathcal{V}'$ is strictly convex, there exists a unique solution u_h to problem* (1.1). *In particular, the dual space to $\mathcal{V} = W^{1,p}(\Omega)$ is strictly convex for $p \geq 2$.*

From now on, we shall concern ourselves with test spaces like $W^{1,p}(\Omega)$. Under such circumstances, it is proven in [18, 19] that the residual minimization problem (1.1) is equivalent to the convex optimization problem

$$\psi = \underset{\varphi \in (B\mathcal{U}_h)^\perp}{\arg\min} \frac{1}{p}\|\varphi\|_{\mathcal{V}}^p - l(\varphi), \tag{2.1}$$

where

$$(B\mathcal{U}_h)^\perp := \{v \in \mathcal{V} : \langle B\delta u_h, v\rangle = 0 \text{ for all } \delta u_h \in \mathcal{U}_h\}.$$

Through the classical optimization theory, one can show that problem (2.1) admits a unique solution, which is characterized by the following mixed system:

$$\begin{cases} \text{find } \psi \in \mathcal{V},\ u_h \in \mathcal{U}_h : \\ \langle R_{\mathcal{V}}(\psi), v\rangle + \langle Bu_h, v\rangle = l(v) \quad \text{for all } v \in \mathcal{V}, \\ \qquad\qquad \langle B\delta u_h, \psi\rangle = 0 \qquad \text{for all } \delta u_h \in \mathcal{U}_h, \end{cases} \tag{2.2}$$

where $R_{\mathcal{V}} : \mathcal{V} \to \mathcal{V}'$ is the Gâteaux derivative of the functional $J(\varphi) := \frac{1}{p}\|\varphi\|_{\mathcal{V}}^p$. Note that $R_{\mathcal{V}}$ is nonlinear for $p > 2$. When $p = 2$, $R_{\mathcal{V}}$ reduces to the familiar Riesz operator. We refer the readers to [18, 19] for details involving properties and formulae for $R_{\mathcal{V}}$.

The meaning of "equivalence" between the residual minimization problem (1.1) and convex optimization problem (2.1) is clarified by the following theorem.

Theorem 2 (Characterization of the Solution). *The unique solution u_h of problem* (1.1) *and the unique solution ψ of problem* (2.1) *satisfy the mixed system* (2.2). *Conversely, any solution (ψ, u_h) to the mixed system* (2.2) *consists of the minimizers of problem* (1.1) *and* (2.1).

In summary, our L^p-DPG method is motivated by the minimum residual problem (1.1). However, in practice, we solve the constrained convex optimization problem (2.1) instead, for which the techniques from the convex optimization can be applied.

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

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

where ϵ is the diffusion coefficient, β denotes an incompressible advection field, and f is a source term. We assume a non-homogeneous Dirichlet boundary condition

$$u = u_0 \quad \text{on } \Gamma = \partial\Omega.$$

Classical Variational Formulation. The standard variational formulation [7] is as follows:

$$\begin{cases} \text{find } u \in \tilde{u}_0 + \mathcal{U} : \\ \displaystyle\int_\Omega \epsilon \nabla u \cdot \nabla v - u\beta \cdot \nabla v = \int_\Omega fv \quad \text{for all } v \in \mathcal{V}, \end{cases}$$

where $\tilde{u}_0$ is a finite energy lift of u_0 into $W^{1,p'}(\Omega)$, i.e. $\tilde{u}_0 \in W^{1,p'}(\Omega)$, $\tilde{u}_0|_{\partial\Omega} = u_0$, and

$$\mathcal{U} = W_0^{1,p'}(\Omega) := \{u \in W^{1,p'}(\Omega) : u = 0 \text{ on } \partial\Omega\},$$
$$\mathcal{V} = W_0^{1,p}(\Omega),$$

where $p \geq 2$, $\frac{1}{p} + \frac{1}{p'} = 1$. The well-posedness of the convection-diffusion-reaction equation in the $W_0^{1,p'}(\Omega)$-$W_0^{1,p}(\Omega)$ setting is proven in [15], provided that Ω is bounded Lipschitz, $p \leq 4$ in 2D (or $p \leq 3$ in 3D), and Friedrich's positivity condition is satisfied. Although here we are not considering a reaction term, numerical stability is still observed.

Ultraweak Variational Formulation. To derive the ultraweak formulation, we introduce the total flux

$$\sigma = \epsilon\nabla u - \beta u$$

and rewrite the convection-diffusion problem as a first-order system. Then we multiply the system by test functions and integrate by parts. The final result is as follows:

$$\begin{cases} \text{find } \sigma \in (L^{p'}(\Omega))^N, u \in L^{p'}(\Omega) : \\ (\sigma, \epsilon^{-1}\tau) + (u, \operatorname{div}\tau + \epsilon^{-1}\beta \cdot \tau) = \langle u_0, \tau \cdot n\rangle \quad \text{for all } \tau \in W^p(\operatorname{div}, \Omega), \\ (\sigma, \nabla v) = (f, v) \qquad\qquad \text{for all } v \in W_0^{1,p}(\Omega), \end{cases} \tag{2.3}$$

where $p \geq 2$, $\frac{1}{p} + \frac{1}{p'} = 1$, and

$$W^p(\operatorname{div}, \Omega) := \{\tau \in (L^p(\Omega))^N : \operatorname{div}\tau \in L^p(\Omega)\}.$$

As usual, we use the notation

$$(u, v) = (u, v)_\Omega := \int_\Omega uv, \quad \langle u, v\rangle = \langle u, v\rangle_{\partial\Omega} := \int_{\partial\Omega} uv.$$

For details on the derivation and the definition of involved Sobolev spaces, we refer readers to [18].

3 Discretization and Linearization

3.1 Discretizing $\mathcal{V}$ with Broken Test Spaces

The test space $\mathcal{V}$ is discretized using the broken space technology [4]. Given a mesh Ω_h, we consider broken test spaces $W^p(\operatorname{div}, \Omega_h)$, $W^{1,p}(\Omega_h)$, defined as

$$W^p(\operatorname{div}, \Omega_h) := \{\sigma \in (L^p(\Omega))^N : \sigma|_K \in W^p(\operatorname{div}, K),\ K \in \Omega_h\} = \prod_{K\in\Omega_h} W^p(\operatorname{div}, K),$$
$$W^{1,p}(\Omega_h) := \{w \in L^p(\Omega) : w|_K \in W^{1,p}(K),\ K \in \Omega_h\} = \prod_{K\in\Omega_h} W^{1,p}(K).$$

As new test functions are no longer conforming, we must introduce interface fluxes as additional unknowns. The ultraweak formulation with broken test spaces is given as follows:

$$\begin{cases} \text{find } \sigma \in (L^{p'}(\Omega))^N,\ u \in L^{p'}(\Omega),\ \hat{\sigma}_n \in W^{-\frac{1}{p'},p'}(\Gamma_h),\ \hat{u} \in W^{1-\frac{1}{p'},p'}(\Gamma_h) : \\ \qquad\qquad \hat{u} = u_0 \qquad \text{on } \Gamma, \\ (\sigma, \epsilon^{-1}\tau) + (u, \operatorname{div}_h\tau + \epsilon^{-1}\beta \cdot \tau) - \langle \hat{u}, \tau \cdot n\rangle_{\Gamma_h} = 0 \qquad \text{for all } \tau \in W^p(\operatorname{div}, \Omega_h), \\ \qquad\qquad (\sigma, \nabla_h v) - \langle \hat{\sigma}_n, v\rangle_{\Gamma_h} = (f, v) \quad \text{for all } v \in W^{1,p}(\Omega_h), \end{cases}$$

where

$$\langle \hat{\sigma}_n, v\rangle_{\Gamma_h} := \sum_{K\in\Omega_h} \langle \hat{\sigma}_n, v_K\rangle_{\partial K}, \quad \langle \hat{u}, \tau\cdot n\rangle_{\Gamma_h} := \sum_{K\in\Omega_h} \langle \hat{u}, \tau_K\cdot n\rangle_{\partial K}$$

are the duality pairings on the mesh skeleton. For details on the trace spaces and well-posedness of the "broken" formulation, the readers are referred to [18, Appendix B]. We emphasize that broken spaces are easier to discretize than globally conforming ones; moreover, they lead to block diagonal Gram matrix, which can be inverted element-wise.

The last step of discretization is to replace $W^p(\mathrm{div}, \Omega_h)$, $W^{1,p}(\Omega_h)$ by piecewise polynomial spaces. If the trial space $\mathcal{U}$ is discretized with polynomials of degree r,[1] then we discretize the test space $\mathcal{V}$ with piecewise polynomials of degree $r + \Delta r$ on the same mesh with $\Delta r \geq 1$. As shown in [5, 16], $\Delta r = 1$ should suffice for the convection-diffusion problems, and this is the value we adopt in the reported numerical experiments.

3.2 Newton's Method for the Minimization Problem

Remark. In the following discussion, the solution u represents a group variable. For the ultraweak formulation with broken test spaces, $\mathsf{u} = (\sigma, u, \hat{\sigma}_n, \hat{u})$. Thus, in particular, the orthogonality condition $b(\delta\mathsf{u}_h, \varphi_h) = 0$ stands for four orthogonal conditions obtained by testing with the four components of $\delta\mathsf{u}_h$. Similarly, $\varphi_h = (\tau, v)$ represents also a group variable. In what follows, we drop the special font for u.

Let $\mathcal{V}_r$ denote the fully discrete test space. We seek to solve the discretized version of (2.1), a convex minimization problem with linear constraints,

$$\min_{\varphi_h\in\mathcal{V}_r} f(\varphi_h) \quad \text{subject to} \quad b(\delta u_h, \varphi_h) = 0 \quad \text{for all } \delta u_h \in \mathcal{U}_h,$$

where $f(\varphi_h) = \frac{1}{p}\|\varphi_h\|_{\mathcal{V}}^p - l(\varphi_h)$.

Following standard practice in numerical optimization, we use Newton's method to solve this problem (cf. [1, Section 10.2]). 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 $\mathcal{V}_r$. Then the linear constraint can be written as $\mathbf{B}^T\boldsymbol{\varphi}_h = \mathbf{0}$, where $\boldsymbol{\varphi}_h$ is the coefficient vector of φ_h under the basis $\{g_1, g_2, \ldots, g_n\}$, $n = \dim\mathcal{V}_r$. For the Newton iteration, we can always start with a feasible φ_h (by feasible, we mean it satisfies the constraint). In practice, we start with $\varphi_h = 0$. The Newton step $\Delta\boldsymbol{\varphi}_{\mathrm{nt}}$ at feasible φ_h is characterized by

$$\begin{bmatrix} \nabla^2\tilde{f}(\boldsymbol{\varphi}_h) & \mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{\varphi}_{\mathrm{nt}} \\ \boldsymbol{u_h} \end{bmatrix} = \begin{bmatrix} -\nabla\tilde{f}(\boldsymbol{\varphi}_h) \\ \mathbf{0} \end{bmatrix}.$$

Note that $\tilde{f}\colon \mathbb{R}^n \to \mathbb{R}$ is the discretized version of $f\colon \mathcal{V} \to \mathbb{R}$. It is defined as

$$\tilde{f}(\boldsymbol{\varphi}_h) := f\left(\sum_{i=1}^n \varphi_h^{(i)} g_i\right).$$

With broken test spaces, the Newton step $\Delta\boldsymbol{\varphi}_{\mathrm{nt}}$ can be condensed out element-wise. We assemble and solve the linear system for $\boldsymbol{u}_h$; then we compute $\Delta\boldsymbol{\varphi}_{\mathrm{nt}}$ locally. After obtaining $\Delta\boldsymbol{\varphi}_{\mathrm{nt}}$, we do a backtracking line search to ensure the Armijo sufficient decrease condition (see [1, Section 9.2]),

$$f(\varphi_h + t\Delta\varphi_{\mathrm{nt}}) \leq f(\varphi_h) + \alpha t\,\nabla f(\varphi_h)^T\Delta\varphi_{\mathrm{nt}},$$

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

The Newton decrement is defined as

$$\lambda(\varphi_h) = (\Delta\boldsymbol{\varphi}_{\mathrm{nt}}^T\nabla^2\tilde{f}(\boldsymbol{\varphi}_h)\Delta\boldsymbol{\varphi}_{\mathrm{nt}})^{\frac{1}{2}}$$

1 In the exact sequence logic. This amounts to order r for a $W^{1,p'}$-conforming element and order $r-1$ for an $L^{p'}$-conforming element.

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

For $p > 2$, say $p = 4$, we combine the Newton iteration with a continuation strategy. We start with $p = 2$ and solve for the minimizer ψ_h. Then we use this ψ_h as the initial point for $p = 3$. Next the minimizer is again used to initialize Newton's method for $p = 4$.

Computing the Hessian. Note that we need to invert $\nabla^2 \tilde{f}(\boldsymbol{\varphi}_h)$ in each Newton step. We provide the formula for the Hessian because of its great importance and influence on the numerical behavior of the algorithm. As an example, consider the ultraweak formulation (2.3) and mathematician's test norm,

$$\|(\tau, v)\|_V^p := \|\tau\|^p + \|\operatorname{div}\tau\|^p + \|v\|^p + \|\nabla v\|^p,$$

where $\|\cdot\|$ denotes standard $L^p(\Omega)$-norm. The Hessian of f in the functional form is

$$\langle \nabla^2 f(\tau, v); (\delta\tau, \delta v), (\Delta\tau, \Delta v)\rangle = (p-1)\Bigg[\sum_{i=1}^{N}\int_\Omega |\tau_i|^{p-2}\Delta\tau_i\delta\tau_i + \int_\Omega |\operatorname{div}\tau|^{p-2}\operatorname{div}\Delta\tau\operatorname{div}\delta\tau + \sum_{|\alpha|\le 1}\int_\Omega |D^\alpha v|^{p-2}D^\alpha\Delta v D^\alpha\delta v\Bigg].$$

This looks like a weighted "inner product", with $|\tau_i|^{p-2}$, $|\operatorname{div}\tau|^{p-2}$, $|D^\alpha v|^{p-2}$ being the weight. It is evident that, when the error representation function is small (τ_i, v and their derivatives have small absolute values), the Hessian is nearly singular. In particular, when the solution is exact, $\psi = (\tau, v) \equiv 0$, the Hessian is singular. A new method using variable p is proposed to circumvent this issue, which is the topic of Section 5.

4 Numerical Results

4.1 Eriksson–Johnson Problem

We consider the Eriksson–Johnson model problem [13]

$$\begin{cases} \dfrac{\partial u}{\partial x} - \epsilon\left(\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2}\right) = 0 & \text{in } (0,1)\times(0,1), \\ u = 0 & \text{if } x = 1,\ y = 0, 1, \\ u = \sin(\pi y) & \text{if } x = 0. \end{cases}$$

This is a 2D convection-diffusion problem with advection field $\beta = (1, 0)$ and a zero source term. The solution is driven by the inflow boundary condition. We can derive the exact solution using separation of variables,

$$u(x, y) = \frac{\exp(s_1(x-1)) - \exp(s_2(x-1))}{\exp(-s_1) - \exp(-s_2)}\sin(\pi y), \quad \text{where } s_1 = \frac{1+\sqrt{1+4\pi^2\epsilon^2}}{2\epsilon},\ s_2 = \frac{1-\sqrt{1+4\pi^2\epsilon^2}}{2\epsilon}.$$

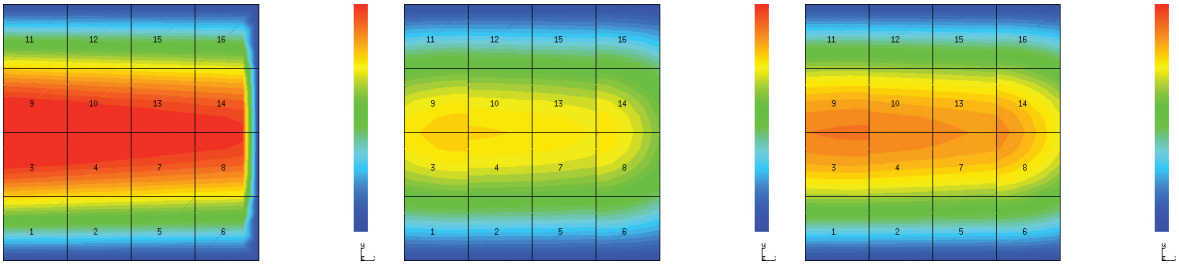
In our numerical experiments, we set $\epsilon = 0.01$, and we use the ultraweak formulation (which defines the operator B).

Choice of Test Norm. In our residual-minimization framework (2.1), the test norm enters the algorithm directly through the expression of a cost function. In DPG, the choices of a test norm can sometimes pose a challenge (see [10]). However, in this paper, we do not concern ourselves with small ϵ, and it suffices to work with mathematician's test norm and adjoint graph norm, which will be introduced now.

Mathematician's Test Norm. The mathematician's test norm is defined as

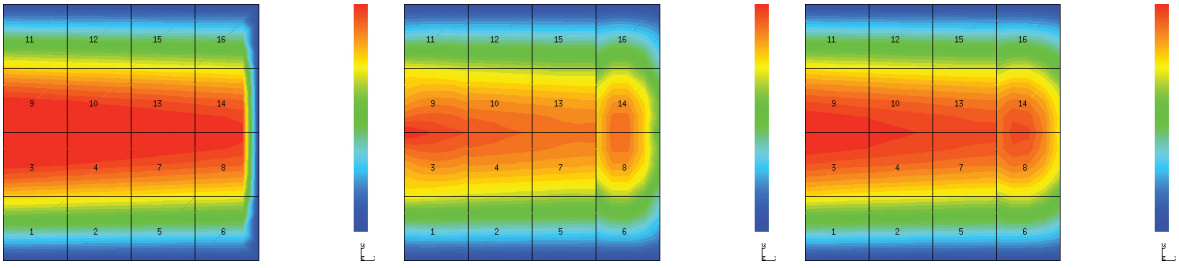
$$\|(\tau, v)\|_M^p := \|\tau\|^p + \|\operatorname{div}\tau\|^p + \|v\|^p + \|\nabla v\|^p,$$

where $\|\cdot\|$ denotes standard $L^p(\Omega)$-norm.



(a) Exact solution (b) $p = 2$ (c) $p = 4$

Figure 1: Solution of Eriksson–Johnson problem using mathematician's test norm. The field u is plotted using the same color scheme across three subfigures.



(a) Exact solution (b) $p = 2$ (c) $p = 4$

Figure 2: Solution of Eriksson–Johnson problem using adjoint graph norm. The field u is plotted using the same color scheme across three subfigures.

Adjoint Graph Norm. The definition of adjoint graph norm makes use of the adjoint operator A^*. In the convection-diffusion problem we consider, the primal operator A corresponding to the first-order system is defined as $A(\sigma, u) := (\epsilon^{-1}\sigma - \nabla u + \epsilon^{-1}\beta u, -\operatorname{div}\sigma)$, and A^* is its formal adjoint,

$$A^*(\tau, v) = (\epsilon^{-1}\tau + \nabla v,\ \operatorname{div}\tau + \epsilon^{-1}\beta \cdot \tau).$$

The adjoint graph norm is defined to be

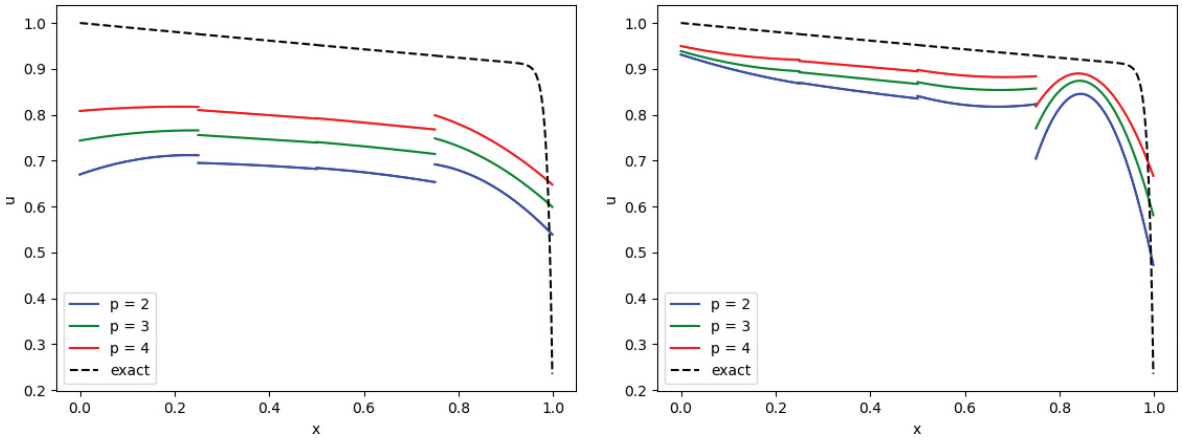
$$\|(\tau, v)\|_{AG}^p := \|(\tau, v)\|^p + \|A^*(\tau, v)\|^p,$$

where $\|(\tau, v)\|^p := \|\tau\|^p + \|v\|^p$.

We divide the domain into 4×4 square elements. The polynomial order (in the exact sequence logic) is set to be $(3, 3)$. Figure 1 presents the numerical solution obtained with $p = 2$ and $p = 4$ alongside the exact solution. As the red color means greater values of u, we can see that the Banach solution ($p = 4$) is closer to the exact solution (redder) than the Hilbert one ($p = 2$). We plot the same figure for adjoint graph norm in Figure 2, and the same trend can be observed.

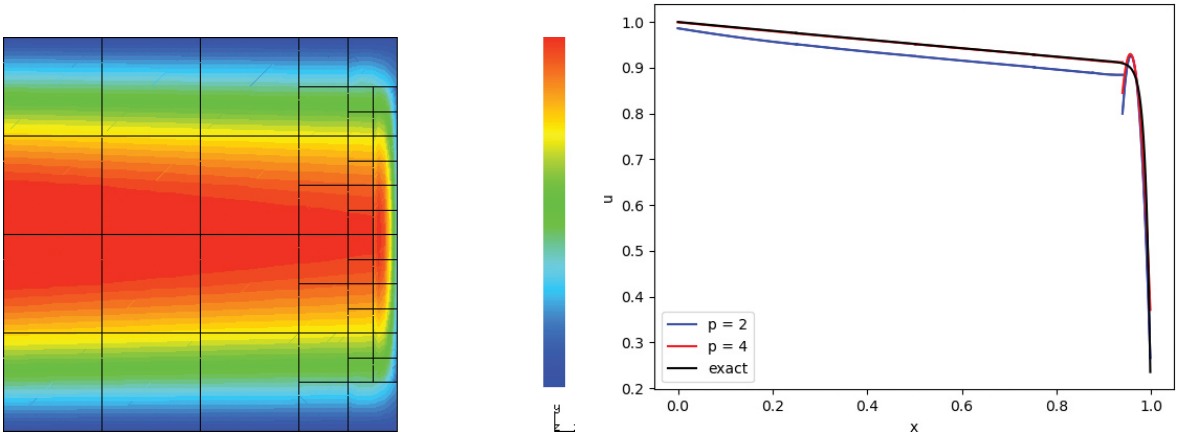
In order to better compare the solutions, we draw a profile of u along the line $y = 0.5$. We also plot the solution for $p = 3$ to make the trend more visible. In Figure 3, we can see that both the Hilbert and Banach solutions underestimate u; however, as we increase p from two to four, the solution is closer and closer to the exact one. At the same time, we also observe that use of adjoint graph norm produces a better solution than mathematician's test norm.

Adaptivity. The DPG method has one key advantage when it comes to adaptivity: $\|\psi\|$ as a built-in a posteriori error estimator [3]. In practice, we use $\eta = \frac{1}{p}\|\psi\|_V^p$ as the error estimator. We use the greedy strategy, marking for h-refinement those elements where $\eta >$ factor $* \eta_{\max}$. In our numerical experiments, we choose factor $= 0.25$. The initial mesh is chosen to be the same 4×4 mesh as before. Moreover, we work with the adjoint graph norm only as it has already been demonstrated to perform better than mathematician's test norm.



(a) Mathematician's test norm (b) Adjoint graph norm

Figure 3: Profile of u along $y = 0.5$. Dashed black line represents exact solution; solid blue line denotes Hilbert solution; solid green and red line stands for Banach solution with $p = 3$ and $p = 4$, respectively.



(a) Solution after 2 refinements, $p = 4$ (b) Profile along $y = 0.5$

Figure 4: Solution of Eriksson–Johnson problem using adjoint graph norm. (a) numerical solution of u after 2 refinements obtained for $p = 4$. (b) the profile of u along $y = 0.5$, for both $p = 2$ and $p = 4$. Black line represents exact solution, blue line denotes Hilbert solution, and red line stands for Banach solution.

Figure 4 (a) shows the solution u alongside the mesh after two refinements, where the same color scheme as in Figures 1 and 2 is used. As expected, mesh refinement occurs where the boundary layer resides. Figure 4 (b) displays the profile of u along $y = 0.5$. It can be seen that numerical solution obtained with $p = 4$ almost coincides with the exact solution. As the refinements proceed, the difference between the Hilbert and Banach versions becomes less significant.

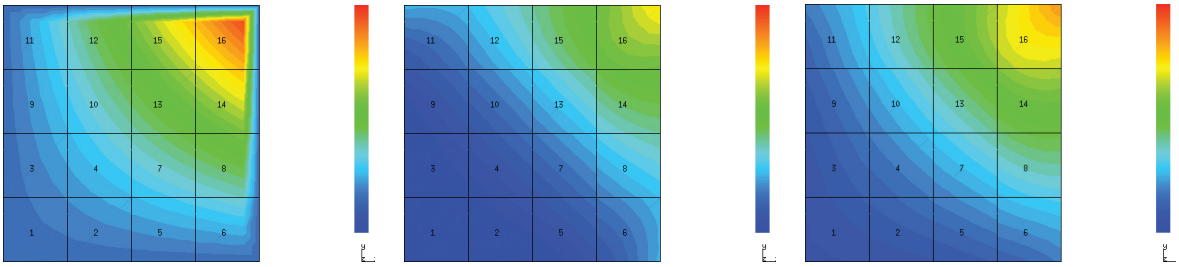
4.2 Egger–Schöberl Problem

We also study the Egger–Schöberl problem [12]

$$\begin{cases} -\epsilon\Delta u + \beta \cdot \nabla u = f & \text{in } \Omega := (0,1)\times(0,1), \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

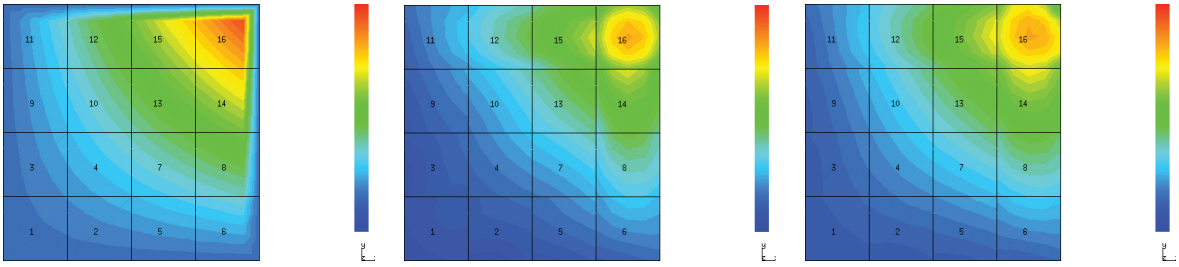
where f is chosen such that the exact solution is given by

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



(a) Exact solution (b) $p = 2$ (c) $p = 4$

Figure 5: Solution of Egger–Schöberl problem using mathematician's test norm. The field u is plotted using the same color scheme across three subfigures.



(a) Exact solution (b) $p = 2$ (c) $p = 4$

Figure 6: Solution of Egger–Schöberl problem using adjoint graph norm. The field u is plotted using the same color scheme across three subfigures.

In our numerical experiments, we set β to be $(1, 1)$ and ϵ to be 0.01. The same 4×4 mesh and polynomial order of $(3, 3)$ as for the Eriksson–Johnson are used.

Figure 5 and Figure 6 show the exact solution and numerical solution obtained with mathematician's test norm and adjoint graph norm, respectively. Figure 7 displays the profile of u along the line $y = 0.5$. We have the same findings as for the Eriksson–Johnson problem: increasing p improves the solution; the solution obtained with adjoint graph norm is overall better than that obtained with mathematician's test norm.

Adaptivity. Figure 8 (a) depicts the solution after three refinements, for $p = 4$ and adjoint graph norm, and Figure 8 (b) draws the profile of u along $y = 0.5$. The refinement occurs both near the top and the right side, in accordance with the location of the boundary layer. After three refinements, again, we observe that the numerical solution agrees reasonably well with the exact one.

4.3 A Posteriori Error Analysis

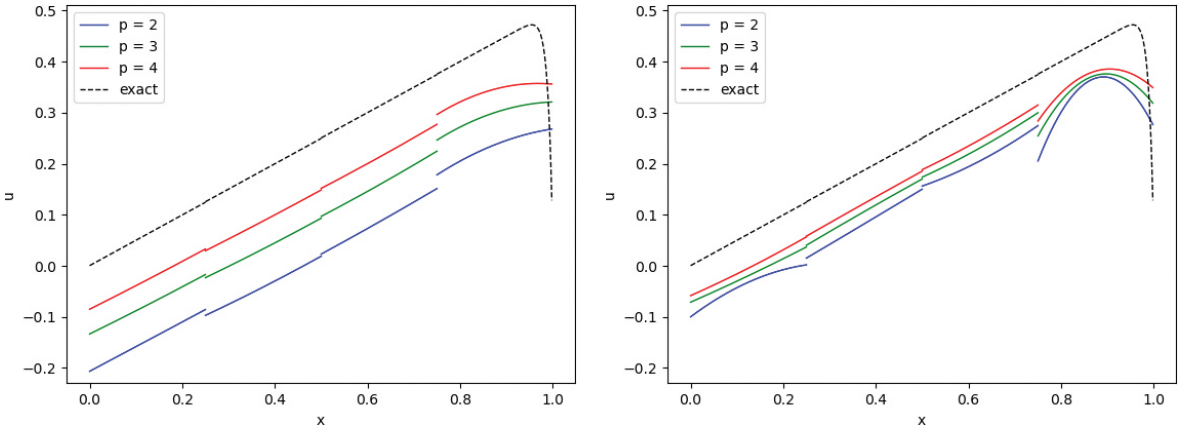
In Which Norm Should We Measure the Error? As proposed in [22], we can introduce the optimal test norm when $\mathcal{V}$ is reflexive and B is bijective. For our *unbroken* ultraweak formulation (2.3), $\|v\|_{\text{opt}} = \|A^* v\|_{L^p}$, where v denotes the group test variable. From (1.1), the minimum residual formulation of DPG, we know that DPG is a projection in the energy norm, i.e.

$$\|u - u_h\|_E := \|B(u - u_h)\|_{\mathcal{V}'} = \min_{w_h \in \mathcal{U}_h} \|u - w_h\|_E.$$

When we work with the optimal test norm, the energy norm reduces to the trial norm

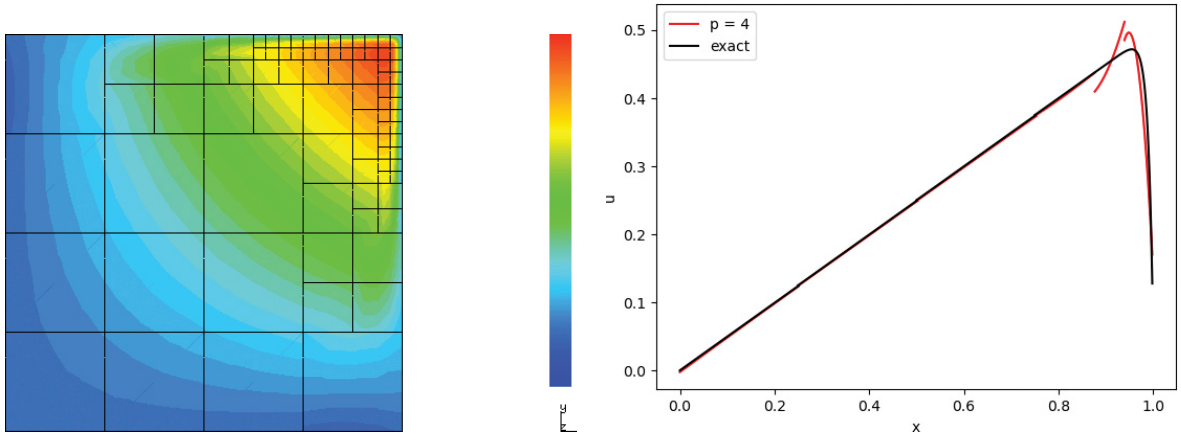
$$\|w\|_E = \sup_{v \in \mathcal{V}} \frac{\langle Bw, v\rangle}{\|v\|_{\text{opt}}} = \sup_{v \in \mathcal{V}} \frac{(w, A^* v)}{\|A^* v\|_{L^p}} = \|w\|_{L^{p'}},$$

where p' is the conjugate exponent to p. Thus the solution u_h would be the best approximation of u in $\mathcal{U}_h$ measured in $L^{p'}$ norm, provided we use *optimal test norm* and *ideal* DPG. In practice, we use the adjoint



(a) Mathematician's test norm (b) Adjoint graph norm

Figure 7: Profile of u along $y = 0.5$. Dashed black line represents exact solution; solid blue line denotes Hilbert solution; solid green and red line stands for Banach solution with $p = 3$ and $p = 4$, respectively.



(a) Solution after three refinements, $p = 4$ (b) Profile along $y = 0.5$

Figure 8: Solution of Egger–Schöberl problem using adjoint graph norm. Black line represents exact solution, and red line stands for Banach solution.

graph norm, also known as "quasi-optimal test norm", which is equivalent to the optimal test norm. Let γ_A be the boundedness below constant for A and A^*, i.e., $\|v\| \leq \gamma_A^{-1}\|A^*v\|$. We have

$$\|A^*v\|^p \leq \|v\|^p + \|A^*v\|^p \leq (\gamma_A^{-p} + 1)\|A^*v\|^p,$$

or equivalently,

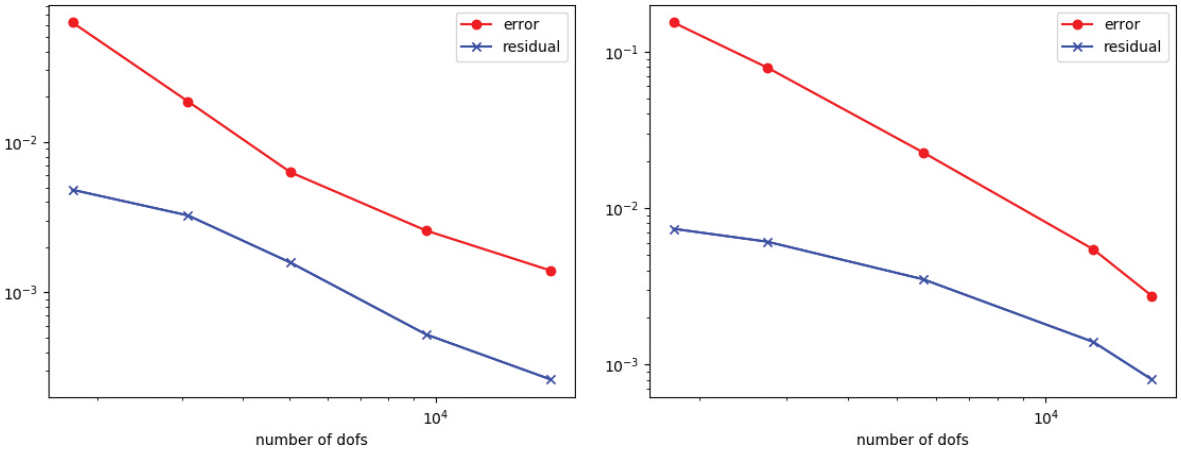
$$\|A^*v\| \leq \|v\|_{AG} \leq (\gamma_A^{-p} + 1)^{\frac{1}{p}}\|A^*v\|.$$

In this case, the energy norm satisfies

$$(\gamma_A^{-p} + 1)^{-\frac{1}{p}}\|w\|_{L^{p'}} \leq \|w\|_E = \sup_{v\in\mathcal{V}} \frac{(w, A^*v)}{\|v\|_{AG}} \leq \|w\|_{L^{p'}}.$$

Hence the equivalence between energy norm and $L^{p'}$ norm. Moreover, we adopt broken test spaces and have to discretize the test space to solve the problem (*practical* DPG instead of *ideal*). Still we expect to see near best approximation in $L^{p'}$ norm. Therefore, we measure the error in our solution with $L^{p'}$ norm.

How Do We Compute the Residual? In our program, the error representation function ψ is calculated element-wise after the solution of u_h. For simplicity, we use u_h to denote both the field and trace variables. Further,



(a) Eriksson–Johnson problem (b) Egger–Schöberl problem

Figure 9: Error and residual in adaptive solution, where $p = 4$ and adjoint graph norm is used. Red line indicates error, and blue line stands for residual. This is a log-log plot, where x-axis represents number of dofs.

$\psi \in \mathcal{V}_r$ satisfies the equation

$$\langle R_{\mathcal{V}_r}(\psi), v_r\rangle = l(v_r) - \langle Bu_h, v_r\rangle \quad \text{for all } v_r \in \mathcal{V}_r,$$

where $\mathcal{V}_r$ is the (discretized) enriched test space. As proven in [18, Theorem 3],

$$\|R_{\mathcal{V}_r}(\psi)\|_{\mathcal{V}_r'} = \|\psi\|_{\mathcal{V}_r}^{p-1} = \Big(\sum_K \|\psi_K\|_{\mathcal{V}(K)}^p\Big)^{\frac{p-1}{p}},$$

where K is the element index and $\mathcal{V}(K)$ is Sobolev space over the element. Specifically,

$$\mathcal{V}(K) = W^p(\mathrm{div}, K) \times W^{1,p}(K)$$

for the convection-diffusion problem with ultraweak formulation. This provides a formula for the residual $\|l - Bu_h\|_{\mathcal{V}_r'} = \|R_{\mathcal{V}_r}(\psi)\|_{\mathcal{V}_r'}$.

What Is the Relation between Error and Residual? This question in Hilbert space is answered in [3]. In general Banach space setting, Muga and van der Zee have proven the following a posteriori error estimate (for details, see [19, Theorem 4.7]):

$$\|u - u_h\|_{\mathcal{U}} \le \frac{1}{\gamma_B}\,\mathrm{osc}(l) + \frac{C_\Pi}{\gamma_B}\|l - Bu_h\|_{\mathcal{V}_r'},$$

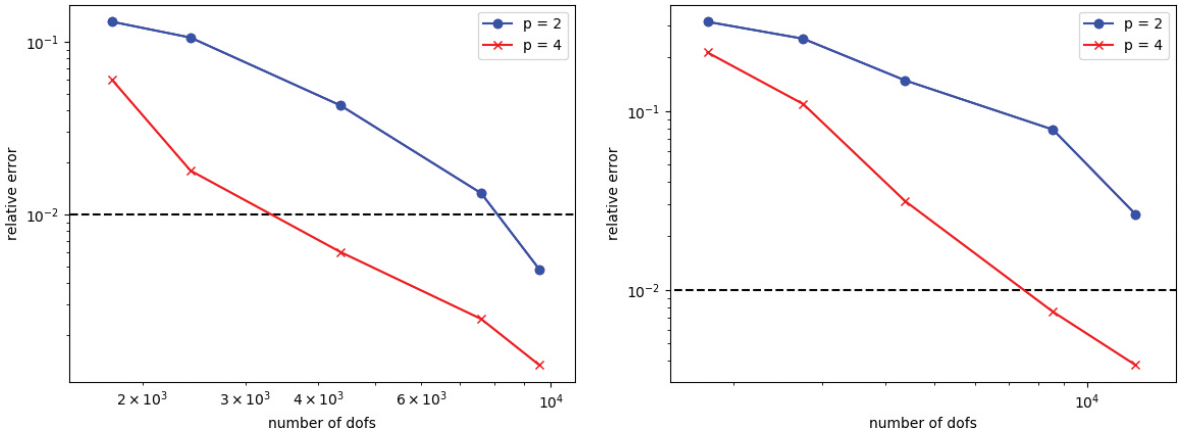
where γ_B is the boundedness-below constant for B, C_Π is the continuity constant for a Fortin operator $\Pi\colon \mathcal{V} \to \mathcal{V}_r$, and

$$\mathrm{osc}(l) := \sup_{v\in\mathcal{V}} \frac{\langle l, v - \Pi v\rangle}{\|v\|_{\mathcal{V}}}$$

is the data oscillation term. In essence, this theorem tells us that the residual $\|l - Bu_h\|_{\mathcal{V}_r'}$ is a good estimator of the error $\|u - u_h\|_{\mathcal{U}}$.

In Figure 9, we plot $\|u - u_h\|_{L^{p'}}$ and $\|l - Bu_h\|_{\mathcal{V}_r'}$ against number of degrees of freedom, for an h-adaptive solution of our model problems using adjoint graph norm and $p = 4$. It can be seen that, as we refine the mesh, both error and residual decrease monotonically, and they follow approximately the same trend.

Figure 10 illustrates the behavior of relative error as we refine the mesh, for both Hilbert and Banach solutions. Although we are not presenting the refined meshes here, we observe that Banach and Hilbert adaptive approaches lead to very similar meshes. It is evident that, for the same number of dofs, the Banach solution comes with a significantly smaller relative error. For the Egger–Schöberl problem, it is even impossible to reach the 1 % tolerance in relative error when $p = 2$, for the number of dofs we have calculated with (which can be finished on a laptop in several minutes). The improvement in accuracy has its price, naturally, in that



(a) Eriksson–Johnson problem (b) Egger–Schöberl problem

Figure 10: Relative error versus number of dofs. Blue line represents Hilbert solution ($p = 2$), while red line stands for Banach solution ($p = 4$); dashed line corresponds to 1 % relative error.

we have to deal with a nonlinear problem in the Banach setting ($p = 4$). In our problem setting, solution of the nonlinear system requires less than ten steps of Newton iteration, i.e. ten times the cost of the Hilbert version. However, the number of dofs necessary for the relative error to reach a particular tolerance is much less in the Banach setting. As the time required to solve the linear system scales quadratically or even cubically as the number of dofs grows, we contend that the additional effort involved with the Banach version pays off. Another advantage of the Banach version is the elimination of Gibbs phenomena, which is detailed in the paper by Houston, Roggendorf and van der Zee [16].

Fortin Operators. Our a posteriori error analysis is based on the existence of a Fortin operator. Construction of Fortin operators in [11, 14] generalizes immediately to the L^p spaces for $p \geq 2$. Recall the overall strategy.

Step 1: Establish L^2-continuity of the operators on the master element $\hat{K}$.

Step 2: Use (standard) scaling arguments to obtain the continuity of the operators on a physical element K (with h-independent continuity constant).

Step 3: Use the commutativity of operators to conclude continuity in the energy norms.

It is now sufficient simply to notice that the L^2-continuity on the master element implies immediately the continuity in the L^p-norm. Consider, e.g., the $H(\mathrm{div})$ Fortin operator Π^{div}. First, the L^p spaces on a bounded domain form a scale (see, e.g., [20, Proposition 3.9.3]). In other words, for $p \geq 2$,

$$\|\sigma\|_{H(\mathrm{div},\hat{K})} \leq C_1 \|\sigma\|_{W^p(\mathrm{div},\hat{K})}, \quad \sigma \in W^p(\mathrm{div}, \hat{K}),$$

with $C_1 > 0$. Additionally, by the finite-dimensionality argument, there exists a constant $C_2 > 0$ such that

$$\|\Pi^{\mathrm{div}}\sigma\|_{L^p(\hat{K})} \leq C_2 \|\Pi^{\mathrm{div}}\sigma\|_{L^2(\hat{K})}.$$

Consequently,

$$\|\Pi^{\mathrm{div}}\sigma\|_{L^p(\hat{K})} \leq C_2 \|\Pi^{\mathrm{div}}\sigma\|_{L^2(\hat{K})} \leq C_2 C \|\sigma\|_{H(\mathrm{div},\hat{K})} \leq C_2 C C_1 \|\sigma\|_{W^p(\mathrm{div},\hat{K})},$$

where C is the L^2-continuity constant. Steps 2 and 3 remain unchanged.

5 L^p-DPG Method with Variable p

As discussed in Section 3.2, when we try to solve problems for simple manufactured exact solutions like linear or quadratic function, we encounter trouble with a singular Hessian. This motivates us to propose the L^p-DPG method with a *variable* exponent p, in the spirit of the $p(\cdot)$-Laplacian problem [2]. We assume that the exponent p can vary element-wise. It is unnecessary to compute with $p > 2$ when the solution is simple and

can be captured by a Hilbert method. We use $p > 2$ where the residual is large and stay with $p = 2$ elsewhere. This section describes our modification of the L^p-DPG method with variable p.

In the convex optimization formulation of L^p-DPG (2.1), we can multiply the cost function by constant p,

$$\psi = \underset{\varphi \in (B\mathcal{U}_h)^\perp}{\arg\min} \|\varphi\|_{\mathcal{V}}^p - pl(\varphi).$$

This has no effect on the minimizer ψ, but the Lagrange multiplier u_h as the solution to the mixed system (2.2) will be affected. If we further multiply the constraint by p, then the solution ψ, u_h of the modified problem will coincide with the original one. The constrained optimization problem now reads

$$\underset{\varphi}{\text{minimize}} \ \|\varphi\|_{\mathcal{V}}^p - pl(\varphi) \quad \text{subject to} \quad pb(\delta u_h, \varphi) = 0 \quad \text{for all } \delta u_h \in \mathcal{U}_h,$$

where b is the bilinear form dictated by the problem we consider and the formulation we choose. The relation between b and B is $\langle Bu, v\rangle = b(u, v)$.

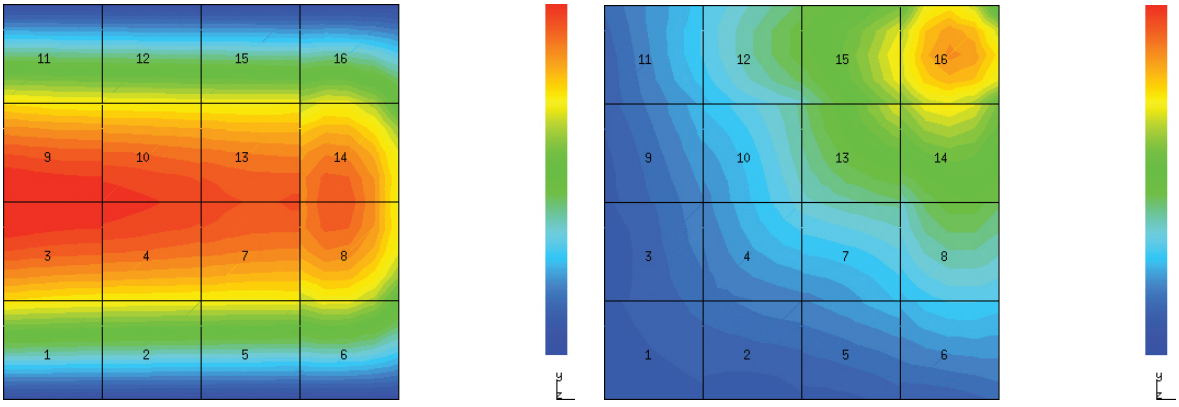
Such reformulation of the constrained optimization problem allows for an easy generalization of the L^p-DPG method to the L^p-DPG with a variable exponent. The latter approach, in particular, has the advantage of reduced condition number and better robustness. After discretization of the test space using broken space technology [4], the L^p-DPG method with variable p is defined by

$$\underset{\varphi}{\text{minimize}} \ \sum_K \|\varphi_K\|_{\mathcal{V}(K)}^{p_K} - p_K l_K(\varphi_K) \quad \text{subject to} \quad \sum_K p_K b_K(\delta u_h, \varphi_K) = 0 \quad \text{for all } \delta u_h \in \mathcal{U}_h,$$

where p_K is the constant exponent for element K, and l_K, b_K are restrictions of the linear and bilinear form on element K. Problems of this type are known as $p(\cdot)$-Laplacian [2]. Rather than venturing into theoretical analysis of the newly proposed method, which can be a future endeavor, we report results of some numerical experiments with the variable exponent.

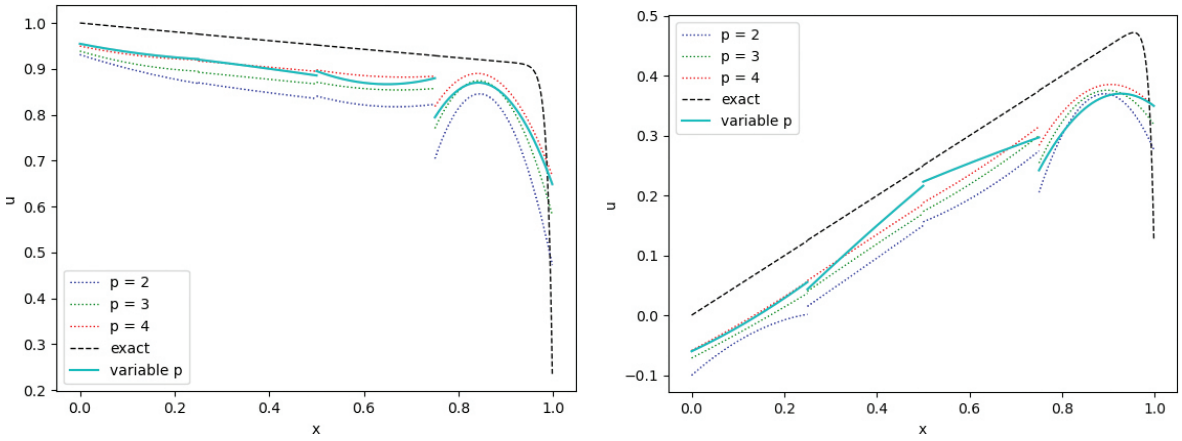
How Do We Determine p_K? This is the foremost question when we are concerned with the variable exponent. With the mesh given, we first set $p_K = 2$ in all elements; in this way, we recover the Hilbert solution. Next we evaluate the residual in each element, and wherever the residual is small (less than 1 % of the maximum value, for results to be reported), we retain the exponent; elsewhere we raise p_K. As a remark, this strategy for determining the variable p matches naturally the continuation in p used by the nonlinear solver – we proceed with local steps of Δp instead of a global step of $\Delta p = 1$.

Figure 11 displays numerical solution of both Eriksson–Johnson problem and Egger–Schöberl problem, using adjoint graph norm and variable exponent. The local exponent, as determined by our rule, is $p_K = 2$ in the four central elements, and $p_K = 4$ elsewhere. Figure 12 shows the profile of u along $y = 0.5$. We observe that the solution obtained with variable p lies approximately between the solutions for $p = 2$ and $p = 4$, as one would expect.



(a) Eriksson–Johnson problem (b) Egger–Schöberl problem

Figure 11: Solution obtained with adjoint graph norm and variable exponent. $p_K = 4$ for all elements adjacent to the boundary; $p_K = 2$ in the four central elements.



(a) Eriksson–Johnson problem (b) Egger–Schöberl problem

Figure 12: Profile of u along $y = 0.5$. Solid cyan line represents variable exponent solution. Dotted blue line denotes Hilbert solution; dotted green and red line stands for Banach solution ($p = 3, 4$). Exact solution is denoted by dashed black line.

6 Conclusion

In this paper, we apply the L^p-DPG method to 2D convection-diffusion problems. More specifically, Eriksson–Johnson problem and Egger–Schöberl problem are studied, with either mathematician's test norm or adjoint graph norm employed. The Banach solution ($p = 4$) is compared with the Hilbert one ($p = 2$), and the former is demonstrated to be generally better. We present an h-adaptivity result with L^p-DPG method, where the refinement occurs at the right place, i.e., near the boundary layer.

We comment shortly on connections between the reported results and the work of Sarah Roggendorf et al. [15–17, 21]. It has been shown in [16] that, for general unstructured meshes, convergence in $L^{p'}$ norm does not eliminate the Gibbs oscillations[2] as $p' \to 1$. However, this does not seem to be the case for standard structured rectangular meshes designed to capture the boundary layers. This is also in agreement with the practice of the computational fluid dynamics community, where hybrid (prismatic-tetrahedral) grids are employed to solve Navier–Stokes equation. In [6], Chen and Kallinderis suggest that the structured prisms permit the use of sufficient grid clustering near the body in the normal direction, while unstructured tetrahedra can cover remaining complicated topologies. Our numerical experience corroborates these observations. If we start with a uniform mesh and proceed with h-refinements driven by the method, the L^p version of the DPG method delivers significantly better results than the Hilbert version. The oscillations are smaller and more localized to the elements near the boundary. The overall global stability seems also to be better for the higher p; the global shift between the exact and the underresolved numerical solutions on coarse meshes is consistently smaller.

To solve the ill-conditioning problem associated with small residuals, we propose an L^p-DPG method with a variable exponent p. The numerical solution looks reasonable and approximately lies between the solutions for $p = 2$ and $p = 4$. This method has the potential of reducing condition number and speeding up the algorithm, while retaining the benefits of Banach solution.

Funding: J. Li and L. Demkowicz were partially supported with NSF grant No. 1819101.

2 The oscillations do not disappear even for certain (crisscross) structured meshes!

References

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University, Cambridge, 2004.

[2] D. Breit, L. Diening and S. Schwarzacher, Finite element approximation of the $p(\cdot)$-Laplacian, *SIAM J. Numer. Anal.* **53** (2015), no. 1, 551–572.

[3] C. Carstensen, L. Demkowicz and J. Gopalakrishnan, A posteriori error control for DPG methods, *SIAM J. Numer. Anal.* **52** (2014), no. 3, 1335–1353.

[4] C. Carstensen, L. Demkowicz and J. Gopalakrishnan, Breaking spaces and forms for the DPG method and applications including Maxwell equations, *Comput. Math. Appl.* **72** (2016), no. 3, 494–522.

[5] J. Chan, J. A. Evans and W. Qiu, A dual Petrov–Galerkin finite element method for the convection-diffusion equation, *Comput. Math. Appl.* **68** (2014), no. 11, 1513–1529.

[6] A. J. Chen and Y. Kallinderis, Adaptive hybrid (prismatic–tetrahedral) grids for incompressible flows, *Internat. J. Numer. Methods Fluids* **26** (1998), no. 9, 1085–1105.

[7] L. Demkowicz, Various variational formulations and closed range theorem, ICES Report 15-03, The University of Texas at Austin, 2015.

[8] L. Demkowicz and J. Gopalakrishnan, A class of discontinuous Petrov–Galerkin methods. II. Optimal test functions, *Numer. Methods Partial Differential Equations* **27** (2011), no. 1, 70–105.

[9] L. Demkowicz and J. Gopalakrishnan, *Encyclopedia of Computational Mechanics*, 2nd ed., Wiley, New York, 2018.

[10] L. Demkowicz and N. Heuer, Robust DPG method for convection-dominated diffusion problems, *SIAM J. Numer. Anal.* **51** (2013), no. 5, 2514–2537.

[11] L. Demkowicz and P. Zanotti, Construction of DPG Fortin operators revisited, *Comput. Math. Appl.* **80** (2020), no. 11, 2261–2271.

[12] H. Egger and J. Schöberl, A hybrid mixed discontinuous Galerkin finite-element method for convection-diffusion problems, *IMA J. Numer. Anal.* **30** (2010), no. 4, 1206–1234.

[13] K. Eriksson and C. Johnson, Adaptive streamline diffusion finite element methods for stationary convection-diffusion problems, *Math. Comp.* **60** (1993), no. 201, 167–188.

[14] J. Gopalakrishnan and W. Qiu, An analysis of the practical DPG method, *Math. Comp.* **83** (2014), no. 286, 537–552.

[15] P. Houston, I. Muga, S. Roggendorf and 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), no. 3, 503–522.

[16] P. Houston, S. Roggendorf and 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), no. 5, 851–873.

[17] P. Houston, S. Roggendorf and K. G. van der Zee, Gibbs phenomena for L^q-best approximation in finite element spaces, *ESAIM Math. Model. Numer. Anal.* **56** (2022), no. 1, 177–211.

[18] J. Li and L. Demkowicz, An Lp-DPG method for the convection-diffusion problem, *Comput. Math. Appl.* **95** (2021), 172–185.

[19] I. Muga and K. G. van der Zee, Discretization of linear problems in Banach spaces: residual minimization, nonlinear Petrov–Galerkin, and monotone mixed methods, *SIAM J. Numer. Anal.* **58** (2020), no. 6, 3406–3426.

[20] J. T. Oden and L. F. Demkowicz, *Applied Functional Analysis*, 3rd ed., CRC Press, Boca Raton, 2018.

[21] S. Roggendorf, *Eliminating the Gibbs phenomenon: The non-linear Petrov–Galerkin method for the convection-diffusion-reaction equation*, PhD thesis, University of Nottingham, 2019.

[22] J. Zitelli, I. Muga, L. Demkowicz, J. Gopalakrishnan, D. Pardo and V. M. Calo, A class of discontinuous Petrov–Galerkin methods. Part IV: The optimal test norm and time-harmonic wave propagation in 1D, *J. Comput. Phys.* **230** (2011), no. 7, 2406–2432.