Locally Conservative Discontinuous Petrov-Galerkin Finite Elements for Fluid Problems

Truman Ellis, Leszek Demkowicz, and Jesse Chan

Institute for Computational Engineering and Sciences,
The University of Texas at Austin,
Austin, TX 78712

Abstract

We develop a locally conservative formulation of the discontinuous Petrov-Galerkin finite element method (DPG) for convection-diffusion type problems using Lagrange multipliers to exactly enforce conservation over each element. We provide a proof of convergence as well as extensive numerical experiments showing that the method is indeed locally conservative. We also show that standard DPG, while not guaranteed to be conservative, is nearly conservative for many of the benchmarks considered. The new method preserves many of the attractive features of DPG, but turns the normally symmetric positive-definite DPG system into a saddle point problem.

1 Introduction

The discontinuous Petrov-Galerkin (DPG) method with optimal test functions has been under active development for convection-diffusion type systems [12, 13, 18, 7, 19, 8, 21, 6, 23]. In this paper, we develop a theory for a locally conservative formulation of DPG for convection-diffusion type equations (including Burgers' equation and Stokes flow) and supplement this with extensive numerical results.

1.1 Importance of Local Conservation

Locally conservative methods hold a special place for numerical analysts in the field of fluid dynamics. Perot[22] argues

> Accuracy, stability, and consistency are the mathematical concepts that are typically used to analyze numerical methods for partial differential equations (PDEs). These important tools quantify how well the mathematics of a PDE is represented, but they fail to say anything about how well the physics of the system is represented by a particular numerical method. In practice, physical fidelity of a numerical solution can be just as important (perhaps even more important to a physicist) as these more traditional mathematical concepts. A numerical solution that violates the underlying physics (destroying mass or entropy, for example) is in many respects just as flawed as an unstable solution.

There are also some mathematically attractive reasons to pursue local conservation. The Lax-Wendroff theorem guarantees that a conservative numerical solution to a system of hyperbolic conservation laws will converge to a weak solution.

The discontinuous Petrov-Galerkin finite element method has been described as least squares finite elements with a twist. The key difference is that while least squares methods seek to minimize the residual of the solution in the L^2 norm, DPG seeks the minimization in a dual norm realized through the inverse Riesz map. Exact mass conservation has been an issue that has long plagued least squares finite elements. Several approaches have been used to try to adress this. Bochev *et al.* [2] accomplish local conservation by using a pointwise divergence free velocity space in the Stokes formulation. Chang and Nelson[9] developed

the *restricted LSFEM*[9] by augmenting the least squares equations with Lagrange multipliers explicitly enforcing mass conservation element-wise. Our conservative formulation of DPG takes a similar approach and both methods share a similar negative of transforming a minimization method to a saddle point problem. In the interest of crediting Chang and Nelson's restricted LSFEM, we could call the following locally conservative DPG method the restricted DPG method, but we prefer to the term conservative DPG. It is worth mentioning that enforcing element conservation is possible within more standard DG schemes, e.g. see the first two discretization schemes presented in [1].

1.2 DPG is a Minimum Residual Method

Roberts *et al.* presents a brief history and derivation of DPG with optimal test functions in [24]. We follow his derivation of the standard DPG method as a minimum residual method. Let U be the trial Hilbert space and V the test Hilbert space for a well-posed variational problem $b(u,v) = l(v)$. In operator form this is $Bu = l$, where $B : U \to V'$ and $\langle Bu, v\rangle = b(u,v)$. We seek to minimize the residual for the discrete space $U_h \subset U$:

$$u_h = \arg\min_{u_h \in U_h} \frac{1}{2} \|Bu_h - l\|_{V'}^2 \,. \tag{1}$$

Recalling that the Riesz operator $R_V : V \to V'$ is an isometry defined by

$$\langle R_V v, \delta v\rangle = (v, \delta v)_V \,, \quad \forall \delta v \in V,$$

we can use the Riesz inverse to minimize in the V-norm rather than its dual:

$$\frac{1}{2} \|Bu_h - l\|_{V'}^2 = \frac{1}{2} \left\|R_V^{-1}(Bu_h - l)\right\|_V^2 = \frac{1}{2} \left(R_V^{-1}(Bu_h - l), R_V^{-1}(Bu_h - l)\right)_V \,. \tag{2}$$

The first order optimality condition for (2) requires the Gâteaux derivative to be zero in all directions $\delta u \in U_h$, i.e.,

$$\left(R_V^{-1}(Bu_h - l), R_V^{-1} B\delta u\right)_V = 0, \quad \forall \delta u \in U.$$

By definition of the Riesz operator, this is equivalent to

$$\left\langle Bu_h - l, R_V^{-1} B\delta u_h\right\rangle = 0 \quad \forall \delta u_h \in U_h \,. \tag{3}$$

Now, we can identify $v_{\delta u_h} \coloneqq R_V^{-1} B\delta u_h$ as the optimal test function for trial function δu_h. Define $T := R_V^{-1} B : U_h \to V$ as the trial-to-test operator. Now we can rewrite (3) as

$$b(u_h, v_{\delta u_h}) = l(v_{\delta u_h}). \tag{4}$$

The DPG method then is to solve (4) with optimal test functions $v_{\delta u_h} \in V$ that solve the auxiliary problem

$$(v_{\delta u_h}, \delta v)_V = \langle R_V v_{\delta u_h}, \delta v\rangle = \langle B\delta u_h, \delta v\rangle = b(\delta u_h, \delta v) \quad \forall \delta v \in V. \tag{5}$$

Using a continuous test basis would result in a global solve for every optimal test function. Therefore DPG uses a discontinuous test basis which makes each solve element-local and much more computationally tractable. Of course, even on the element level, (5) still requires the inversion of the infinite-dimensional Riesz map. In practice, the inversion is approximated using standard Galerkin and an "enriched test space". The effect of approximating the optimal test functions has been analyzed in [20]. In this contribution, we assume that error resulting from the approximation of test functions is of higher order and can be neglected.

No assumptions have been made so far on the definition of the inner product on V. In fact, proper choice of $(\cdot,\cdot)_V$ can make the difference between a solid DPG method and one that suffers from robustness issues (a robust technique remain stable as the singularly perturbed parameter approaches "extreme" values).

2 Element Conservative Convection-Diffusion

We now proceed to develop a locally conservative formulation of DPG for convection-diffusion type problems, but there are a few terms that we need to define first. If Ω is our problem domain, then we can partition it into finite elements K such that

$$\overline{\Omega} = \bigcup_K \bar{K}, \quad K \text{ open},$$

with corresponding external boundary Γ, *skeleton* Γ_h and *interior skeleton* Γ_h^0,

$$\Gamma_h := \bigcup_K \partial K \qquad \Gamma_h^0 := \Gamma_h - \Gamma.$$

We define broken Sobolev spaces element-wise:

$$\begin{aligned} H^1(\Omega_h) &:= \textstyle\prod_K H^1(K), \\ \boldsymbol{H}(\mathrm{div}, \Omega_h) &:= \textstyle\prod_K \boldsymbol{H}(\mathrm{div}, K). \end{aligned}$$

We also need the trace spaces:

$$\begin{aligned} H^{\frac{1}{2}}(\Gamma_h) &:= \left\{ \hat{v} = \{\hat{v}_K\} \in \textstyle\prod_K H^{1/2}(\partial K) \,:\, \exists v \in H^1(\Omega) : v|_{\partial K} = \hat{v}_K \right\}, \\ H^{-\frac{1}{2}}(\Gamma_h) &:= \left\{ \hat{\sigma}_n = \{\hat{\sigma}_{Kn}\} \in \textstyle\prod_K H^{-1/2}(\partial K) \,:\, \exists \boldsymbol{\sigma} \in \boldsymbol{H}(\mathrm{div}, \Omega) : \hat{\sigma}_{Kn} = (\boldsymbol{\sigma} \cdot \boldsymbol{n})|_{\partial K} \right\}, \end{aligned}$$

which are developed more precisely in [24].

2.1 Derivation

Now that we have briefly outlined the abstract DPG method, let us apply it to the convection-diffusion equation. The strong form of the steady convection-diffusion problem with homogeneous Dirichlet boundary conditions reads

$$\begin{cases} \nabla \cdot (\boldsymbol{\beta} u) - \epsilon \Delta u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \Gamma, \end{cases}$$

where u is the property of interest, $\boldsymbol{\beta}$ is the convection vector, and f is the source term. Nonhomogeneous Dirichlet and Neumann boundary conditions are straightforward but would add technicality to the following discussion. Let us write this as an equivalent system of first order equations:

$$\begin{aligned} \nabla \cdot (\boldsymbol{\beta} u - \boldsymbol{\sigma}) &= f \\ \frac{1}{\epsilon} \boldsymbol{\sigma} - \nabla u &= \boldsymbol{0}. \end{aligned}$$

If we then multiply the first equation by some scalar test function v and the bottom equation by some vector-valued test function $\boldsymbol{\tau}$, we can integrate by parts over each element K:

$$\begin{aligned} -(\boldsymbol{\beta} u - \boldsymbol{\sigma}, \nabla v)_K + \langle (\boldsymbol{\beta} u - \boldsymbol{\sigma}) \cdot \mathbf{n}, v \rangle_{\partial K} &= (f, v)_K \\ \frac{1}{\epsilon} (\boldsymbol{\sigma}, \boldsymbol{\tau})_K + (u, \nabla \cdot \boldsymbol{\tau})_K - \langle u, \tau_n \rangle_{\partial K} &= 0. \end{aligned} \tag{6}$$

The discontinuous Petrov-Galerkin method refers to the fact that we are using discontinuous optimal test functions that come from a space differing from the trial space. It does not specify our choice of trial space. Nevertheless, many versions of DPG in the literature (convection-diffusion [14], linear elasticity [3], linear acoustics [17], Stokes [24]) associate DPG with the so-called "ultra-weak formulation." We will follow the same derivation for the convection-diffusion equation, but we emphasize that other formulations are available (in particular, the primal DPG[15] method presents an alternative with continuous trial functions). Thus, we seek field variables $u \in L^2(K)$ and $\boldsymbol{\sigma} \in \mathbf{L^2}(K)$. Mathematically, this leaves their traces on element boundaries undefined, and in a manner similar to the hybridized discontinuous Galerkin method, we define

new unknowns for trace $\hat{u}$ and flux $\hat{t}$. Applying these definitions to (6) and adding the two equations together, we arrive at our desired variational problem.

Find $\boldsymbol{u} := (u, \boldsymbol{\sigma}, \hat{u}, \hat{t}) \in \boldsymbol{U} := L^2(\Omega_h) \times \boldsymbol{L}^2(\Omega_h) \times H^{1/2}(\Gamma_h) \times H^{-1/2}(\Gamma_h)$ such that

$$\underbrace{-(\boldsymbol{\beta} u - \boldsymbol{\sigma}, \nabla v)_K + \left\langle \hat{t}, v \right\rangle_{\partial K} + \frac{1}{\epsilon}(\boldsymbol{\sigma}, \boldsymbol{\tau})_K + (u, \nabla \cdot \boldsymbol{\tau})_K - \langle \hat{u}, \tau_n \rangle_{\partial K}}_{b(\mathbf{u},\mathbf{v})} = \underbrace{(f, v)_K}_{l(\mathbf{v})} \qquad \text{in } \Omega \tag{7}$$

$$\hat{u} = 0 \qquad \text{on } \Gamma \tag{8}$$

for all $\boldsymbol{v} := (v, \boldsymbol{\tau}) \in \boldsymbol{V} := H^1(\Omega_h) \times \boldsymbol{H}(\mathrm{div}, \Omega_h)$.

We note that for convection-diffusion problems we are particularly interested in designing a *robust* DPG method. Specifically, we are interested in designing methods whose behavior does not change as the diffusion parameter ϵ becomes very small. Naive Galerkin methods for convection-diffusion tend to suffer from a lack of robustness; specifically, the finite element error is bounded by a constant factor of the best approximation error, but the constant is often proportional to ϵ^{-1}. Our aim is to design a DPG method with this in mind. We follow the methodology introduced by Heuer and Demkowicz in [19]: the ultra-weak variational formulation for convection-diffusion can be refactored as

$$b\left(\left(u, \boldsymbol{\sigma}, \widehat{u}, \hat{t}\right), (v, \boldsymbol{\tau})\right) = \sum_{K \in \Omega_h} \left[\left\langle \hat{t}, v \right\rangle_{\partial K} + \langle \widehat{u}, \tau_n \rangle_{\delta K} + (u, \nabla \cdot \boldsymbol{\tau} - \boldsymbol{\beta} \cdot \nabla v)_{L^2(K)} + \left(\boldsymbol{\sigma}, \frac{1}{\epsilon}\boldsymbol{\tau} + \nabla v \right)_{L^2(K)} \right],$$

modulo application of boundary data. If we choose specific *conforming* test functions satisfying the adjoint equations

$$\begin{aligned} \nabla \cdot \boldsymbol{\tau} - \boldsymbol{\beta} \cdot \nabla v &= u, \\ \frac{1}{\epsilon}\boldsymbol{\tau} + \nabla v &= \boldsymbol{\sigma}, \end{aligned}$$

then evaluating $b\left(\left(u, \boldsymbol{\sigma}, \widehat{u}, \hat{t}\right), (v, \boldsymbol{\tau})\right)$ at these specific test functions returns back $\|u\|^2 + \|\boldsymbol{\sigma}\|^2$, the L^2 norm of our field variables. Multiplying and dividing through by the test norm $\|\boldsymbol{v}\|_V$, we have

$$\|u\|^2_{L^2(\Omega)} + \|\boldsymbol{\sigma}\|^2_{L^2(\Omega)} = b\left(\left(u, \boldsymbol{\sigma}, \widehat{u}, \hat{t}\right), (v, \boldsymbol{\tau})\right) = \frac{b\left(\left(u, \boldsymbol{\sigma}, \widehat{u}, \hat{t}\right), (v, \boldsymbol{\tau})\right)}{\|\boldsymbol{v}\|_V} \|\boldsymbol{v}\|_V \leq \left\|u, \boldsymbol{\sigma}, \widehat{u}, \hat{t}\right\|_E \|\boldsymbol{v}\|_V,$$

where

$$\left\|u, \boldsymbol{\sigma}, \widehat{u}, \hat{t}\right\|_E = \sup_{v \in V \setminus \{0\}} \frac{b\left(\left(u, \boldsymbol{\sigma}, \widehat{u}, \hat{t}\right), (v, \boldsymbol{\tau})\right)}{\|\boldsymbol{v}\|_V}$$

is the DPG energy norm. If we can robustly bound the test norm $\|\boldsymbol{v}\|_V \lesssim \left(\|u\|^2_{L^2(\Omega)} + \|\boldsymbol{\sigma}\|^2_{L^2(\Omega)}\right)^{1/2}$ (i.e. derive a bound from above with a constant independent of ϵ), then we can divide through to get

$$\left(\|u\|^2_{L^2(\Omega)} + \|\boldsymbol{\sigma}\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}} \lesssim \left\|(u, \boldsymbol{\sigma}, \widehat{u}, \hat{t})\right\|_E. \tag{9}$$

In other words, the energy norm in which DPG is optimal bounds the L^2 norm independently of epsilon. So, as we drive our energy error down to zero, we can expect that the L^2 error will also decrease uniformly in ϵ.

We note that the construction of the test norm $\|\boldsymbol{v}\|_V$ for a robust DPG method depends on two things: the test norm, as well as the adjoint equation. In [19], the standard problem with Dirichlet conditions enforced over the entire boundary was considered; in [8], boundary conditions were chosen for the forward problem such that the induced adjoint problem was regularized and contained no strong boundary layers, allowing for the construction of a stronger test norm on V. We adopt a slight modification of the test norm introduced in [8] for numerical experiments here, which is motivated and explained in more detail in Section 2.3.

Having reviewed and laid the foundation for DPG methods, we can now formulate our conservative DPG scheme. A numerical scheme for a conservation law is said to be *element conservative* if the flux integrals over the element boundary are balanced with the volume integral of the source,

$$\int_{\partial K} \hat{t} = \int_K f$$

In context of the DPG method, this translates in a simple condition that test function $(1_K, \mathbf{0})$ should belong to the element test space where 1_K denotes the element indicator function, i.e. $(1_K, \mathbf{0})$ is the test function in which $v = 1$ on element K and 0 elsewhere and $\boldsymbol{\tau} = \mathbf{0}$ everywhere. We note that this property is nowhere guaranteed for the space of optimal test functions. The idea proposed by Moro *et al.* [21] is based on replacing the (unconstrained) DPG minimum residual formulation with a constrained minimization where the constraint corresponds to enforcing the element conservation. Let $\boldsymbol{U}_h := U_h \times \boldsymbol{S}_h \times \hat{U}_h \times \hat{F}_h \subset L^2(\Omega_h) \times \boldsymbol{L}^2(\Omega_h) \times H^{\frac{1}{2}}(\Gamma_h) \times H^{-\frac{1}{2}}(\Gamma_h)$ be a finite-dimensional subspace, and let $\boldsymbol{u}_h := (u_h, \boldsymbol{\sigma}_h, \hat{u}_h, \hat{t}_h) \in \boldsymbol{U}_h$ be the group variable. The element conservative DPG scheme is derived from the Lagrangian:

$$L(\boldsymbol{u}_h, \lambda_K) = \frac{1}{2} \left\| R_V^{-1}(b(\boldsymbol{u}_h, \cdot) - (f, \cdot)) \right\|_{\boldsymbol{V}}^2 - \sum_K \lambda_K (b(\boldsymbol{u}_h, (1_K, \mathbf{0})) - l((1_K, \mathbf{0}))) \,. \tag{10}$$

Taking the Gâteaux derivatives as before, we arrive at the following system of equations:

$$\begin{cases} b(\boldsymbol{u}_h, T(\delta \boldsymbol{u}_h)) - \sum_K \lambda_K b(\delta \boldsymbol{u}_h, (1_K, \mathbf{0})) & = l(T(\delta \boldsymbol{u}_h)) \quad \forall \delta \boldsymbol{u}_h \in \boldsymbol{U}_h \\ b(\boldsymbol{u}_h, (1_K, \mathbf{0})) & = l((1_K, \mathbf{0})) \quad \forall K \,, \end{cases} \tag{11}$$

where $T := R_V^{-1} B : \boldsymbol{U}_h \to \boldsymbol{V}$ is the same trial-to-test operator as in the original formulation.

Denote $T(\delta \boldsymbol{u}_h) = (v_{\delta \boldsymbol{u}_h}, \boldsymbol{\tau}_{\delta \boldsymbol{u}_h}) \in H^1(\Omega_h) \times \boldsymbol{H}(\mathrm{div}, \Omega_h)$. Then, putting (11) into more concrete terms for convection-diffusion, we get:

$$\begin{cases} -(\boldsymbol{\beta} u - \boldsymbol{\sigma}, \nabla v_{\delta \boldsymbol{u}_h}) + \langle \hat{t}, v_{\delta \boldsymbol{u}_h} \rangle + \frac{1}{\epsilon}(\boldsymbol{\sigma}, \boldsymbol{\tau}_{\delta \boldsymbol{u}_h}) + (u, \nabla \cdot \boldsymbol{\tau}_{\delta \boldsymbol{u}_h}) - \langle \hat{u}, \boldsymbol{\tau}_{\delta \boldsymbol{u}_h} \cdot \boldsymbol{n} \rangle & \\ \qquad - \sum_K \lambda_K (\delta \hat{t}, (1_K, \mathbf{0})) & = (f, v_{\delta \boldsymbol{u}_h}) \quad \forall \delta \boldsymbol{u}_h \in \boldsymbol{U}_h \\ \langle \hat{t}, (1_K, \mathbf{0}) \rangle & = (f, 1_K) \quad \forall K \,. \end{cases} \tag{12}$$

2.2 Stability Analysis

In the following analysis we neglect the error due to the approximation of optimal test functions. See [20] for a defense of this assumption. We follow the classical Brezzi's theory [4, 11] for an abstract mixed problem:

$$\begin{cases} \boldsymbol{u} \in \boldsymbol{U}, p \in Q & \\ a(\boldsymbol{u}, \boldsymbol{w}) + c(p, \boldsymbol{w}) & = l(\boldsymbol{w}) \quad \forall \boldsymbol{w} \in \boldsymbol{U} \\ c(q, \boldsymbol{u}) & = g(q) \quad \forall q \in Q \end{cases} \tag{13}$$

where $\boldsymbol{U}, Q$ are Hilbert spaces, and a, c, l, g denote the appropriate bilinear and linear forms. Note that $a(\boldsymbol{u}, \boldsymbol{w}) = b(\boldsymbol{u}, T\boldsymbol{w}) = (T\boldsymbol{u}, T\boldsymbol{w})_V$ in the notation from the previous section.

Let the function $\boldsymbol{\psi}$ denote the $\boldsymbol{H}(\mathrm{div}, \Omega)$ extension of flux $\hat{t}$ that realizes the minimum in the definition of the quotient (minimum energy extension) norm. The choice of norm for the Lagrange multipliers λ_K is implied by the quotient norm used for $H^{-1/2}(\Gamma_h)$ and continuity bound for form $c(p, \boldsymbol{w})$ representing the

constraint:

$$
\begin{aligned}
|c(\textstyle\sum_K \lambda_K(1_K,\mathbf{0}),(u,\boldsymbol{\sigma},\hat{u},\hat{t}))| &= |\textstyle\sum_K \lambda_K \langle \hat{t}, 1_K\rangle_{\partial K}| \\
&= |\textstyle\sum_K \lambda_K \langle v_n, 1_K\rangle_{\partial K}| \\
&= |\textstyle\sum_K \lambda_K \int_K \mathrm{div}\boldsymbol{\psi}\; 1_K| \\
&\leq \textstyle\sum_K \lambda_K ||\mathrm{div}\boldsymbol{\psi}||_{L^2(K)} \mu(K)^{1/2} \\
&\leq (\textstyle\sum_K \mu(K)\lambda_K^2)^{1/2}\, (\textstyle\sum_K ||\mathrm{div}\boldsymbol{\psi}||^2_{L^2(K)})^{1/2} \\
&\leq \underbrace{\left(\sum_K \mu(K)\lambda_K^2\right)^{1/2}}_{=:||\boldsymbol{\lambda}||} ||\hat{t}||_{H^{-1/2}(\Gamma_h)} \\
&\leq \|\boldsymbol{\lambda}\|\, \|\boldsymbol{u}\|\,,
\end{aligned}
\tag{14}
$$

where $\mu(K)$ stands for the area (measure) of element K.

We proceed now with the discussion of the discrete inf-sup stability constants. We skip index h in the notation.

Inf Sup Condition relating spaces $\boldsymbol{U}$ and Q reads as follows:

$$
\sup_{\boldsymbol{w}\in\boldsymbol{U}} \frac{|c(p,\boldsymbol{w})|}{||\boldsymbol{w}||_{\boldsymbol{U}}} \geq \beta ||p||_Q\,. \tag{15}
$$

Let

$$
R\,:\, L^2(\Omega) \ni q \rightarrow \boldsymbol{\psi} \in \boldsymbol{H}(\mathrm{div},\Omega)\cap \boldsymbol{H}^1(\Omega) = \boldsymbol{H}^1(\Omega) \tag{16}
$$

be the continuous right inverse of the divergence operator constructed by Costabel and McIntosh in [10]. Let $\boldsymbol{\psi}_h$ denote the classical, lowest order Raviart-Thomas (RT) interpolant of the function

$$
\boldsymbol{\psi} = R(\sum_K \lambda_K 1_K)\,. \tag{17}
$$

Note that $\mathrm{div}\boldsymbol{\psi}_h = \mathrm{div}\boldsymbol{\psi} = \lambda_K$ in element K.

Classical h-interpolation interpolation error estimates for the lowest error Raviart-Thomas elements and continuity of operator R imply the stability estimate:

$$
\begin{aligned}
||\boldsymbol{\psi}_h|| &\leq ||\boldsymbol{\psi}_h - \boldsymbol{\psi}|| + ||\boldsymbol{\psi}|| \\
&\leq Ch||\boldsymbol{\psi}||_{H^1} + ||\boldsymbol{\psi}|| \\
&\leq C||\mathrm{div}\boldsymbol{\psi}|| = C(\textstyle\sum_K \mu(K)\lambda_K^2)^{1/2}\,.
\end{aligned}
\tag{18}
$$

Above, C is a generic, mesh independent constant incorporating constant from the interpolation error estimate and the continuity constant of R. Let $\hat{t}$ be the trace of $\boldsymbol{\psi}_h$. We have then,

$$
\sup_{\hat{t}\in H^{-1/2}(\Gamma_h)} \frac{|\sum_K \lambda_K \langle \hat{t}, 1_K\rangle_{\partial K}|}{||\hat{t}||_{H^{-1/2}(\Gamma_h)}} \geq \frac{|\sum_K \lambda_K \int_K \mathrm{div}\boldsymbol{\psi}_h\; 1_K|}{||\boldsymbol{\psi}_h||_{H(\mathrm{div},\Omega)}} \geq \frac{1}{C}(\sum_K \mu(K)\lambda_K^2)^{1/2}\,, \tag{19}
$$

where C is the constant from stability estimate (18).

Notice that we have considered traces of lowest order Raviart-Thomas elements for the discretization of flux $\hat{t}$. The inf-sup condition for the lowest order RT spaces implies automatically the analogous condition for elements of arbitrary order; increasing the dimension of space U only makes the discrete inf-sup constant bigger.

Inf Sup in Kernel Condition is satisfied automatically due to the use of optimal test functions. First of all, we characterize the "kernel" space:

$$\begin{aligned} \boldsymbol{U}_0 := & \quad \{\boldsymbol{w} \in \boldsymbol{U} \,:\, c(q,\boldsymbol{w}) = 0 \quad \forall q \in Q\} \\ = & \quad \{(u,\boldsymbol{\sigma},\hat{u},\hat{t}) \,:\, \langle \hat{t}, 1_K\rangle_{\partial K} = 0 \quad \forall K\}\,. \end{aligned} \tag{20}$$

In other words, the kernel space contains only the equilibrated fluxes. With $\boldsymbol{u} \in \boldsymbol{U}_0$, we have then:

$$\sup_{\boldsymbol{w}\in\boldsymbol{U}_0} \frac{|a(\boldsymbol{u},\boldsymbol{w})|}{||\boldsymbol{w}||_{\boldsymbol{U}}} \geq \frac{|b(\boldsymbol{u},T\boldsymbol{u})|}{||\boldsymbol{u}||} = \frac{|b(\boldsymbol{u},T\boldsymbol{u})|}{||T\boldsymbol{u}||}\frac{||T\boldsymbol{u}||}{||\boldsymbol{u}||} = \sup_{(v,\boldsymbol{\tau})} \frac{|b((u,\boldsymbol{\sigma},\hat{u},\hat{t}),(v,\boldsymbol{\tau}))|}{||(v,\boldsymbol{\tau})||}\frac{||T\boldsymbol{u}||}{||\boldsymbol{u}||} \geq \gamma^2 ||(u,\boldsymbol{\sigma},\hat{u},\hat{t})||\,, \tag{21}$$

where γ is the stability constant for the standard continuous DPG formulation. The first inequality follows as we plug in the definition for a and pick $\boldsymbol{w} = \boldsymbol{u}$. The second equality is trivial, while the next one follows by definition of the optimal test functions given through the trial-to-test operator T. The final inequality springs from the fact that $\sup_{\boldsymbol{v}} \frac{|b(\boldsymbol{u},\boldsymbol{v})|}{||\boldsymbol{v}||} \geq \gamma ||\boldsymbol{u}||$ and $||T\boldsymbol{u}||_V = ||R_V^{-1}B\boldsymbol{u}||_V = ||B\boldsymbol{u}||_{V'} \geq \gamma||\boldsymbol{u}||$.

With both discrete inf-sup constants in place, we have the standard result: the FE error is bounded by the best approximation error in the constrained space. Notice that the exact Lagrange multipliers are zero, so the best approximation error involves only the solution $(u,\boldsymbol{\sigma},\hat{u},\hat{t})$.

2.2.1 Robustness Analysis

Recall the line of analysis leading to the construction of robust test norms allowing us to bound the L^2 error of the field variables by the energy error, (9). With robust test norms, we have

$$\begin{aligned} \left(||u-u_h||^2 + ||\boldsymbol{\sigma}-\boldsymbol{\sigma}_h||^2\right)^{\frac{1}{2}} & \lesssim ||(u-u_h,\boldsymbol{\sigma}-\boldsymbol{\sigma}_h,\hat{u}-\hat{u}_h,\hat{t}-\hat{t}_h||_E \\ & = \inf_{(w_h,\boldsymbol{\varsigma}_h,\hat{w}_h,\hat{r}_h)} ||(u-w_h,\boldsymbol{\sigma}-\boldsymbol{\varsigma}_h,\hat{u}-\hat{w}_h,\hat{t}-\hat{r}_h||_E\,. \end{aligned} \tag{22}$$

The last equality follows from the fact that the DPG method delivers the best approximation error in the energy norm (minimizes the residual). This is no longer true for the conservative version. So, can we claim robustness in the sense of the inequality above for the conservative version as well?

One possible way to attack the problem is to switch to the energy norm in the Brezzi stability analysis. Dealing with the "inf-sup in kernel" condition is simple. Upon replacing the original norm of solution $\boldsymbol{u}$ with the energy norm, both constant γ and continuity constant become unity. In order to investigate the robustness of inf-sup constant β, we need to realize first what the energy norm of the flux $\hat{t}$ is. Given an element K, we solve for the optimal test functions corresponding to the flux $\hat{t}$,

$$\begin{cases} v_K \in H^1(K),\, \boldsymbol{\tau}_K \in \boldsymbol{H}(\mathrm{div},K) \\ ((v_K,\boldsymbol{\tau}_K),(\delta v,\delta\boldsymbol{\tau}))_V = \langle \hat{t},\delta v\rangle_{\partial K} \quad \forall \delta v \in H^1(K), \delta\boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div},K)\,. \end{cases} \tag{23}$$

The energy norm of $\hat{t}$ is then equal to

$$||\hat{t}||_E^2 = \sum_K ||(v_K,\boldsymbol{\tau}_K)||_V^2\,. \tag{24}$$

We need to establish sufficient conditions under which the inf-sup and continuity constants for the bilinear form representing the constraint are independent of viscosity ϵ.

Let us start with the inf-sup condition,

$$\sup_{\hat{t}} \frac{|\sum_K \lambda_K \langle \hat{t}, 1_K\rangle_{\partial K}|}{||\hat{t}||_E} \geq \beta \left(\sum_K \mu(K)\lambda_K^2\right)^{1/2}\,. \tag{25}$$

As in the previous analysis, we select for $\hat{t}$ the trace of Raviart-Thomas interpolant $\boldsymbol{\psi}_h$ of $\boldsymbol{\psi} = R(\sum_K \lambda_K 1_K)$ where R is the right-inverse of the divergence operator constructed by Costabel and McIntosh. The only

change compared with the previous analysis, is the evaluation of the norm of $\hat{t}_h$. For this, we need to solve the local problems:

$$((v_K, \boldsymbol{\tau}_K), (\delta v, \delta\boldsymbol{\tau}))_V = \langle \hat{t}, \delta v\rangle_{\partial K} = \int_K \mathrm{div}\boldsymbol{\psi}_h\, \delta v = \int_K \mathrm{div}\boldsymbol{\psi}\, \delta v$$
$$= \int_K \lambda_K \delta v = \lambda_K (1_K, \delta v)_K \quad \forall \delta v \in H^1(K)\, \forall \delta\boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div}, K)\,. \tag{26}$$

We need then an upper bound of the energy norm of $(v_h, \boldsymbol{\tau}_h)$:

$$\left(\sum_K ||(v_K, \boldsymbol{\tau}_K)||_V^2\right)^{1/2}.$$

Substituting $(v_K, \boldsymbol{\tau}_K)$ for $(\delta v, \delta\boldsymbol{\tau})$ in (26), we get,

$$||(v_K, \boldsymbol{\tau}_K)||_V^2 = \lambda_K (1_K, v_K)_K\,. \tag{27}$$

If we have a robust stability estimate:

$$|(1_K, v_K)_K| \le C\mu(K)^{1/2} ||(v_K, \boldsymbol{\tau}_K)||_V \tag{28}$$

(i.e. constant C is independent of ϵ), then

$$||(v_K, \boldsymbol{\tau}_K)||_V \le C\mu(K)^{1/2} |\lambda_K| \tag{29}$$

and, eventually as needed,

$$\sum_K ||(v_K, \boldsymbol{\tau}_K)||_V^2 \le C^2 \sum_K \mu(K)\lambda_K^2\,, \tag{30}$$

which leads to the robust estimate of inf-sup constant β. For example, it is sufficient if

$$||v||_{L^2(K)} \le ||(v_K, \boldsymbol{\tau}_K)||_V\,. \tag{31}$$

Notice that the stability analysis with the energy norm was, in a sense, easier than with the quotient norm. Only the divergence of the interpolant $\boldsymbol{\psi}_h$ enters (26) and it coincides with the divergence of $\boldsymbol{\psi}$.

We arrive at a similar situation in the continuity estimate of

$$\sum_K \lambda_K \langle \hat{t}, 1_K\rangle_{\partial K}\,.$$

Testing with $(1_K, \mathbf{0})$ in the local problem (23), we obtain,

$$((v, \boldsymbol{\tau}), (1_K, \mathbf{0}))_V = \langle \hat{t}, 1_K\rangle_{\partial K}\,. \tag{32}$$

If we have a robust estimate,

$$|((v, \boldsymbol{\tau}), (1_K, \mathbf{0}))_V| \le C\mu(K)^{1/2}\, ||(v, \boldsymbol{\tau})||_V\,, \tag{33}$$

then

$$\begin{aligned}
|\textstyle\sum_K \lambda_K \langle \hat{t}, 1_K\rangle| &= |\textstyle\sum_K \mu^{1/2}(K)\lambda_K\, \mu^{-1/2}(K)\langle \hat{t}, 1_K\rangle| \\
&\le \left(\textstyle\sum_K \mu(K)\lambda_K^2\right)^{1/2} \left(\textstyle\sum_K \frac{1}{\mu(K)} |\langle \hat{t}, 1_K\rangle|^2\right)^{1/2} && \text{(discrete Cauchy-Schwartz)} \\
&\le \left(\textstyle\sum_K \mu(K)\lambda_K^2\right)^{1/2} \left(\textstyle\sum_K \frac{1}{\mu(K)} |((v_k, \boldsymbol{\tau}_K), (1_K, 0))_V|^2\right)^{1/2} && (23) \\
&\le C \left(\textstyle\sum_K \mu(K)\lambda_K^2\right)^{1/2} \left(\textstyle\sum_K \|(v_k, \boldsymbol{\tau}_K)\|^2\right)^{1/2} && \text{(assumption (33))} \\
&= C \left(\textstyle\sum_K \mu(K)\lambda_K^2\right)^{1/2} \|\hat{t}\|_E && \text{(definition of energy norm)} \\
&\le C\|\boldsymbol{\lambda}\|\, \|\boldsymbol{u}\|_E
\end{aligned} \tag{34}$$

as needed.

For instance, condition (33) will be satisfied if the test inner product in (32) reduces to the L^2 term only,

$$((v,\boldsymbol{\tau}),(1_K,\mathbf{0}))_V = (v,1_K)_{L^2(K)}\,. \tag{35}$$

With the robust stability and continuity constants for the mixed problem, the energy error of solution $(u,\boldsymbol{\sigma},\hat{u},\hat{t})$ (and Lagrange multipliers λ_K as well) is bounded robustly by the *best approximation error* of $(u,\boldsymbol{\sigma},\hat{u},\hat{t})$ measured in the energy norm. We arrive thus at the same situation as in the standard DPG method.

2.3 Robust Test Norms

The optimal test functions are determined by solving local problems determined by the choice of test norm. There are several options to consider. The graph norm [16] is one of the most natural norms to consider as it is derived directly from the adjoint of the problem supplemented with (possibly scaled) L^2 field terms to upgrade it from a semi-norm. Chan *et al.* [8] derived a more robust alternative norm for convection diffusion (dubbed the robust norm). We recently developed a modification of the robust norm that produces better results in the presence of singularities; for more details and motivation, see [5].

$$\begin{aligned}\|(v,\boldsymbol{\tau})\|^2_{V,K} \coloneqq \min\left\{\frac{1}{\epsilon},\frac{1}{\mu(K)}\right\}\|\boldsymbol{\tau}\|^2_K\,, + \|\nabla\cdot\boldsymbol{\tau}-\boldsymbol{\beta}\cdot\nabla v\|^2_K \\ + \|\boldsymbol{\beta}\cdot\nabla v\|^2_K + \epsilon\,\|\nabla v\|^2_K + \|v\|^2_K\,,\end{aligned} \tag{36}$$

where $||\cdot||_K$ signifies the L^2 norm over element K.

2.3.1 Adaptation for a Locally Conservative Formulation

With this choice of test norm, our local problem now becomes:

Find $v_{\delta\boldsymbol{u}_h}\in H^1(K)$, $\boldsymbol{\tau}_{\delta\boldsymbol{u}_h}\in\boldsymbol{H}(\mathrm{div},K)$ such that:

$$\begin{aligned}\min\left\{\frac{1}{\epsilon},\frac{1}{\mu(K)}\right\}(\boldsymbol{\tau}_{\delta\boldsymbol{u}_h},\delta\boldsymbol{\tau})_K + (\nabla\cdot\boldsymbol{\tau}_{\delta\boldsymbol{u}_h}-\boldsymbol{\beta}\cdot\nabla v_{\delta\boldsymbol{u}_h},\nabla\cdot\delta\boldsymbol{\tau}-\boldsymbol{\beta}\cdot\nabla\delta v)_K + (\boldsymbol{\beta}\cdot\nabla v_{\delta\boldsymbol{u}_h},\boldsymbol{\beta}\cdot\nabla\delta v)_K \\ + \epsilon(\nabla v_{\delta\boldsymbol{u}_h},\nabla\delta v)_K + \alpha(v_{\delta\boldsymbol{u}_h},\delta v)_K = b(\delta\boldsymbol{u}_h,(\delta v,\delta\boldsymbol{\tau})) \qquad \forall\delta v\in H^1(K),\,\delta\boldsymbol{\tau}\in\boldsymbol{H}(\mathrm{div},K)\,,\end{aligned} \tag{37}$$

where the coefficient α scaling the L^2 term can be selected using different criteria, see [19, 8] for details.

With a locally conservative formulation, we can eliminate the L^2 term altogether by setting $\alpha=0$ in local problem (37). This is related to the fact that, with constants $(1_K,0)$ present in the test space, we need to determine the optimal test functions *only* in the L^2-orthogonal complement of the constants (the semi-norm is a norm in the quotient space). In practice, we replace the L^2-term with the product of integrals of $v_{\delta\boldsymbol{u}_h}$ and δv. Notice that the extra term vanishes in the orthogonal complement and, therefore, produces implicitly a specific constant component for the optimal test functions *without* affecting the test space.

$$\begin{aligned}\min\left\{\frac{1}{\epsilon},\frac{1}{\mu(K)}\right\}(\boldsymbol{\tau}_{\delta\boldsymbol{u}_h},\delta\boldsymbol{\tau})_K + (\nabla\cdot\boldsymbol{\tau}_{\delta\boldsymbol{u}_h}-\boldsymbol{\beta}\cdot\nabla v,\nabla\cdot\delta\boldsymbol{\tau}-\boldsymbol{\beta}\cdot\nabla v)_K \\ + (\boldsymbol{\beta}\cdot\nabla v_{\delta\boldsymbol{u}_h},\boldsymbol{\beta}\cdot\nabla\delta v)_K + \epsilon(\nabla v_{\delta\boldsymbol{u}_h},\nabla\delta v)_K + \frac{1}{\mu(K)}\int_K v_{\delta\boldsymbol{u}_h}\int_K\delta v\,,\end{aligned} \tag{38}$$

Notice that the coefficient $\frac{1}{\mu K}$, convenient for the analysis presented next, can be replaced with any other constant in practical computations. In all numerical experiments reported in this paper, we have used $\frac{1}{\mu(K)^2}$ which renders the term to be $O(1)$ in terms of element size h, and seems to have been the best scaling for the numerical inversion of the approximate Riesz map.

2.3.2 Verification of Robust Stability Estimate

In the robustness analysis in Section 2.2.1, we argued that if we have robust stability estimates:

$$(1_K, v_K) \leq C\mu(K)^{1/2}||(v, \boldsymbol{\tau})||_K \tag{28 revisited}$$

and

$$|((v, \boldsymbol{\tau}), (1_K, \mathbf{0}))_V| \leq C\mu(K)^{1/2}\,||(v, \boldsymbol{\tau})||_V\,. \tag{33 revisited}$$

then the conservative DPG method is robust.

We now proceed to show that the robust norms we are using satisfy this requirement. Consider the inner product from (37), with $\alpha = 1$. We wish to verify condition (28) with the norm derived from this inner product on the right hand side. By Cauchy-Schwarz

$$\int_K v \cdot 1 \leq \mu(K)^{1/2}\,\|v\|_{L^2(K)} \leq \mu(K)^{1/2}\,\|(v, \boldsymbol{\tau})\|_K\,, \tag{39}$$

where $\|(v, \boldsymbol{\tau})\|_K$ is the norm derived from the inner product. Condition (33) comes out the same since

$$|((v, \boldsymbol{\tau}), (1_K, \mathbf{0}))| \leq \sum_K |(1_K, v_K)| \leq \sum_K \mu(K)^{1/2}\,\|(v, \boldsymbol{\tau})\|_K$$

element-wise.

Now we need to perform the same analysis for the modified inner product in (38). In this case, condition (28) follows even more naturally as

$$\int_K v \cdot 1 \leq \mu(K)^{1/2}\frac{1}{\mu(K)^{1/2}}\left|\int_K v\right| \leq \mu(K)^{1/2}\,\|(v, \boldsymbol{\tau})\|_K\,, \tag{40}$$

where $\|(v, \boldsymbol{\tau})\|$ now refers to the norm generated by inner product (38). Condition (33) follows by the same reasoning.

3 Application to Other Fluid Model Problems

Extension of these ideas to other fluid flow problems is relatively trivial. For the following problems, we just use the graph norm for the local problems.

3.1 Inviscid Burgers' Equation

We include the inviscid Burgers' equation in our suite of tests because, being a nonlinear hyperbolic conservation law, it falls under the scope of the Lax-Wendroff theorem. The inviscid Burger's equation is

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = f\,.$$

Define the space-time gradient: $\nabla_{xt} = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial t}\right)^T$. We can now rewrite this as

$$\nabla_{xt} \cdot \begin{pmatrix} u^2/2 \\ u \end{pmatrix} = f\,.$$

Multiplying by a test function v, and integrating by parts:

$$-\left(\begin{pmatrix} u^2/2 \\ u \end{pmatrix}, \nabla_{xt} v\right) + \langle \hat{t}, v\rangle = (f, v)\,,$$

where $\hat{t}$ is the trace of $\begin{pmatrix} u^2/2 \\ u \end{pmatrix} \cdot \mathbf{n}_{xt}$ on element boundaries, and $\mathbf{n}_{xt}$ is the space-time normal vector. As in convection-diffusion, local conservation implies that $\int_{\partial K} \hat{t} = \int_K f$ for all elements, K.

In order to solve this nonlinear problem, we linearize and do a simple Newton iteration until the solution converges. The linearized equation is

$$-\left(\begin{pmatrix} u \\ 1 \end{pmatrix}\Delta u, \nabla_{xt} v\right) + \langle \hat{t}, v\rangle = (f, v) + \left(\begin{pmatrix} u^2/2 \\ u \end{pmatrix}, \nabla_{xt} v\right),$$

where u is the previous solution iteration and Δu is the update. The results follow in Section 4.1.4.

3.2 Stokes Flow

We start with the VGP (velocity, gradient pressure) Stokes formulation:

$$\begin{aligned} \mu\Delta \boldsymbol{u} + \nabla p &= \boldsymbol{f} \\ \nabla \cdot \boldsymbol{u} &= 0\,, \end{aligned}$$

where $\boldsymbol{u}$ is the velocity vector field. As a first order system of equations, this is

$$\begin{aligned} \frac{1}{\mu}\boldsymbol{\sigma} - \nabla \boldsymbol{u} &= 0 \\ \nabla \cdot \boldsymbol{\sigma} + \nabla p &= \boldsymbol{f} \\ \nabla \cdot \boldsymbol{u} &= 0\,, \end{aligned}$$

where $\boldsymbol{\sigma}$ is a tensor valued stress field. Multiplying by test functions $\boldsymbol{\tau}$ (tensor valued), $\boldsymbol{v}$ (vector valued), and q (scalar valued), and integrating by parts:

$$\begin{aligned} \left(\frac{1}{\mu}\boldsymbol{\sigma}, \boldsymbol{\tau}\right) + (\boldsymbol{u}, \nabla \cdot \boldsymbol{\tau}) - \langle \hat{\boldsymbol{u}}, \boldsymbol{\tau} \cdot \mathbf{n}\rangle &= 0 \\ -(\boldsymbol{\sigma}, \nabla \boldsymbol{v}) - (p, \nabla \cdot \boldsymbol{v}) + \left\langle \hat{\boldsymbol{t}}, \boldsymbol{v}\right\rangle &= (\boldsymbol{f}, \boldsymbol{v}) \\ -(\boldsymbol{u}, \nabla q) + \langle \hat{\boldsymbol{u}} \cdot \mathbf{n}, q\rangle &= 0\,, \end{aligned}$$

where $\hat{\boldsymbol{u}}$ is the trace of $\boldsymbol{u}$, and $\hat{\boldsymbol{t}}$ is the trace of $(\boldsymbol{\sigma} + p\mathbf{I}) \cdot \mathbf{n}$. The solve for p is only unique up to a constant, so we also impose a zero mean condition, $\int_\Omega p = 0$. Local conservation for Stokes flow means that over each element, $\int_K \hat{\boldsymbol{u}} \cdot \mathbf{n} = 0$. Results follow in Sections 4.1.5 and 4.1.6.

4 Numerical Experiments

In 4.1 we define each numerical experiment, and in 4.2 we discuss the solution properties in general. We solve with second order field variables and flux (u, $\boldsymbol{\sigma}$, and $\hat{t}$), third order traces ($\hat{u}$), and fifth order test functions (v and $\boldsymbol{\tau}$).

We measure flux imbalance by looping over each element in the mesh and integrating the flux over each side and summing them together. We then integrate the source term over the volume of the element. The two should match each other, and the remainder is the flux imbalance. We get the net global flux imbalance by summing these quantities and taking the absolute value. The max local flux imbalance is the maximum absolute value of these flux imbalances.

4.1 Description of Problems

Unless otherwise noted, the problem domain is $\Omega = [0, 1]^2$ and $f = 0$. Also note that unless otherwise noted, for all of the pseudo-color plots, blue corresponds to 0 and red to 1 with a linear scaling in between. Also, all convection-diffusion plots are of the field variable u. Inviscid Burgers' and Stokes results will be dealt with individually.

4.1.1 Eriksson-Johnson Model Problem

The Eriksson-Johnson problem is one of the few convection-diffusion problems with a known analytical solution. Take $\boldsymbol{\beta} = (1,0)^T$ and boundary conditions $\hat{t} = \boldsymbol{\beta} \cdot \mathbf{n} u_0$ when $\beta_n \leq 0$, where u_0 is the trace of the exact solution, and $\hat{u} = 0$ when $\beta_n > 0$. For $n = 1, 2, \cdots$, let $\lambda_n = n^2\pi^2\epsilon$, $r_n = \frac{1+\sqrt{1+4\epsilon\lambda_n}}{2\epsilon}$, and $s_n = \frac{1-\sqrt{1+4\epsilon\lambda_n}}{2\epsilon}$. The exact solution is

$$u(x,y) = C_0 + \sum_{n=1}^{\infty} C_n \frac{\exp(s_n(x-1)) - \exp(r_n(x-1))}{r_n \exp(-s_n) - s_n \exp(-s_n)} \cos(n\pi y)\,. \tag{41}$$

The exact solution for $\epsilon = 10^{-2}$, $C_1 = 1$, and $C_{n\neq 1} = 0$ is shown in Figure 1.

4.1.2 Vortex Problem

This problem models a mildly diffusive vortex convecting fluid in a circle. We deal with domain $\Omega = [-1,1]^2$, with $\epsilon = 10^{-4}$, and $\boldsymbol{\beta} = (-y,x)^T$. Note that $\boldsymbol{\beta} = \mathbf{0}$ at the domain center. We have an inflow boundary condition when $\boldsymbol{\beta} \cdot \mathbf{n} < 0$, in which case we set $\hat{t} = \boldsymbol{\beta} \cdot \mathbf{n} \cdot u_0$ where $u_0 = \frac{\sqrt{x^2+y^2}-1}{\sqrt{2}-1}$ which will vary from 0 at the center of boundary edges to 1 at corners. We don't enforce an outflow boundary.

4.1.3 Discontinuous Source Problem

Here, $\boldsymbol{\beta} = (0.5,1)^T/\sqrt{1.25}$, and we have a discontinuous source term such that $f = 1$ when $y \geq 2x$ and $f = -1$ when $y < 2x$. We apply boundary conditions of $\hat{t} = 0$ on the inflow and $\hat{u} = 0$ on the outflow. Contrary to the other problems discussed, the solution for this problem does not range from zero to one. Rather, the colorbar in Figure 6 is scaled to $[-1.110, 0.889]$.

4.1.4 Inviscid Burgers' Equation

This is a standard test problem for Burgers' equation. The domain is a unit square. We assign boundary conditions $\hat{t} = -(1-2x)$ on the bottom, $\hat{t} = -1/2$ on the left, while $\hat{t} = 1/2$ on the right. Since this is a hyperbolic equation, there is no need to set a boundary condition on the top.

4.1.5 Stokes Flow Around a Cylinder

This is a common problem used to stress-test local conservation properties of least squares finite element methods. Since DPG can be viewed as a generalized least squares methods[16], we might expect it to struggle with this problem as well. The problem domain is detailed in Figure 10 with inlet and outlet velocity profiles

$$\boldsymbol{u}_{in} = \boldsymbol{u}_{out} = \begin{pmatrix} (1-y)(1+y) \\ 0 \end{pmatrix},$$

and zero flow on the cylinder and at the top and bottom walls. We use $\mu =$ with both Stokes problems and set velocity boundary conditions on $\hat{\boldsymbol{u}}$.

Bochev *et al.* [2] run this test with both $r = 0.6$ and $r = 0.9$; we repeat the same experiments with standard and conservative DPG methods starting from the very coarse meshes shown in Figure 11 while adaptively refining toward a resolved solution. The extreme pressure gradient in the $r = 0.9$ case obviously makes local conservation more challenging.

We measure mass loss more directly in these two Stokes problems. Because fluid enters and leaves the domain only through the inlet and outlet boundaries, we should be able to integrate the mass flux over any cross-section of the mesh and get the same value. Unfortunately, it is not mathematically well-defined to take line integrals of our field variables which only live in L^2. We can however integrate the trace and flux variables over element boundaries. This carries the unfortunate limitation that we can only measure mass loss where there is a clear vertical mesh line. We therefore pick integration lines from the initial coarse mesh and measure the mass flux after each adaptive refinement step. The percent mass loss is thus

$$\%m_{loss} = \frac{\int_{\Gamma_{in}} \boldsymbol{u} \cdot \boldsymbol{n}_{in} d\ell - \int_S \boldsymbol{u} \cdot \boldsymbol{n}_S d\ell}{\int_{\Gamma_{in}} \boldsymbol{u} \cdot \boldsymbol{n}_{in} d\ell} \times 100,$$

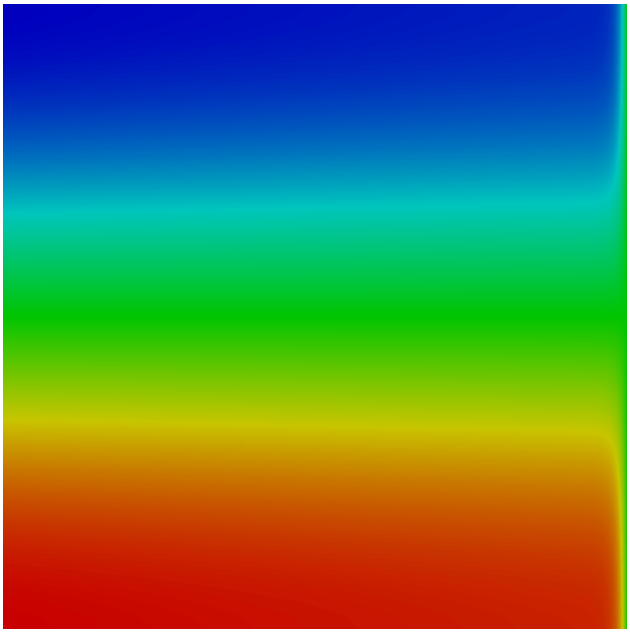


Figure 1: Erickson-Johnson exact solution

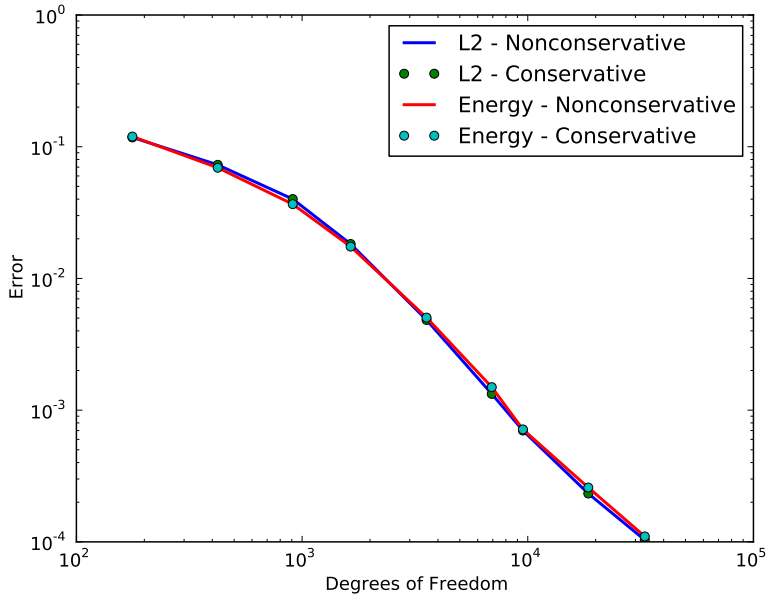


Figure 2: Error in Erickson-Johnson solutions

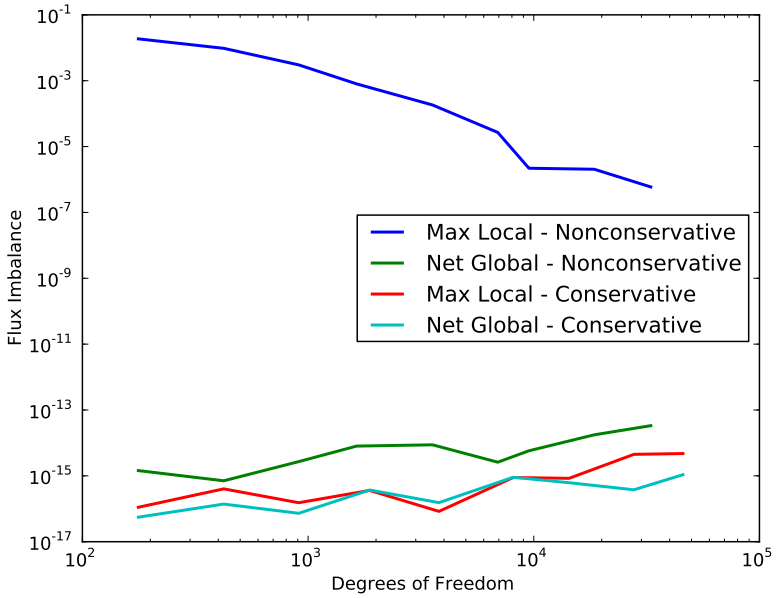


Figure 3: Flux imbalance in Erickson-Johnson solutions

where S is some vertical mesh line.

4.1.6 Stokes Flow Over a Backward Facing Step

Similarly, least squares methods have historically performed very poorly when calculating Stokes flow over a backward facing step shown in Figure 15. The stress singularity at the reentrant corner seems to destroy local conservation. We assign parabolic inlet and outlet velocity boundary conditions

$$\boldsymbol{u}_{in} = \begin{pmatrix} 8(y-0.5)(1-y) \\ 0 \end{pmatrix} \quad \text{and} \quad \boldsymbol{u}_{out} = \begin{pmatrix} y(1-y) \\ 0 \end{pmatrix}$$

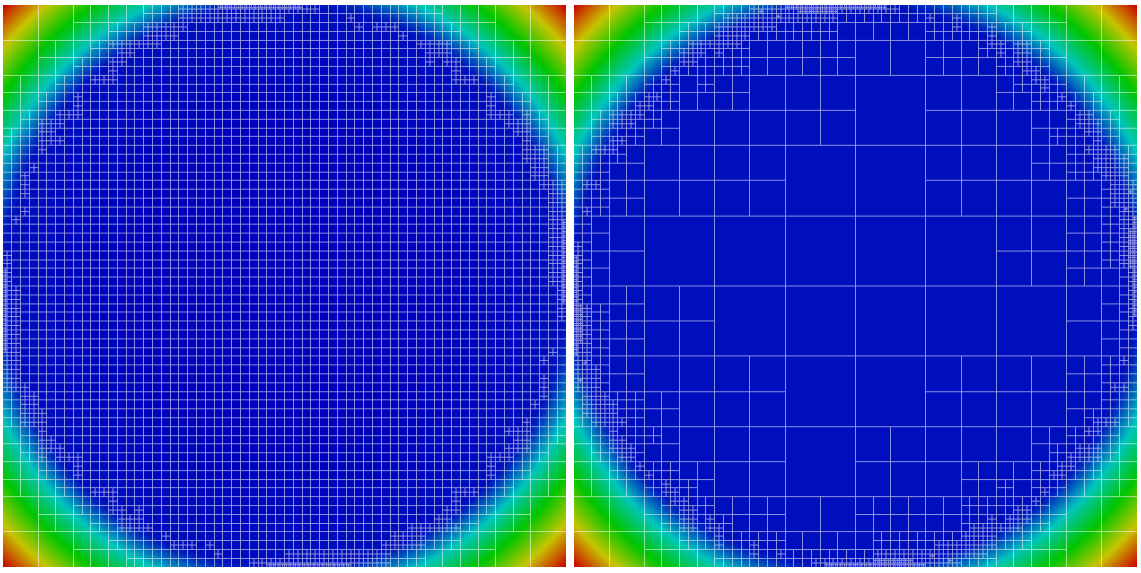
and zero velocity on all other boundaries. In this problem, we solve with fourth order field and flux variables, fifth order traces, and sixth order test functions.

4.2 Analysis of Results

4.2.1 Convection-Diffusion Results

The general trend we observe from the results is that the solution quality of the standard and conservative formulations is nearly identical once sufficiently resolved.

It is clear when comparing the refinement patterns that the two methods appear to calculate slightly different error representation functions (which determine which elements to adaptively refine). Standard DPG minimizes the error in the energy norm, but the Lagrange multipliers in the conservative formulation shift the solution slightly, so we should see somewhat higher error and different elements will get chosen for refinement. The choice of test norm also plays into this calculation of the error representation function. As discussed earlier, the conservative formulation allows us to throw away the L^2 term on v. The inclusion of this term required certain assumptions on $\boldsymbol{\beta}$ [8] that break down for the vortex problem, where $|\boldsymbol{\beta}| \to 0$ in the center of the domain. Here, we see the standard method needlessly refines in the center of the domain where the solution is constant. The conservative scheme is more discerning about refinements and focuses them where solution features are changing. In general, though, both methods appear to follow very similar refinement patterns.



(a) Nonconservative (b) Conservative

Figure 4: Vortex problem after 6 refinements

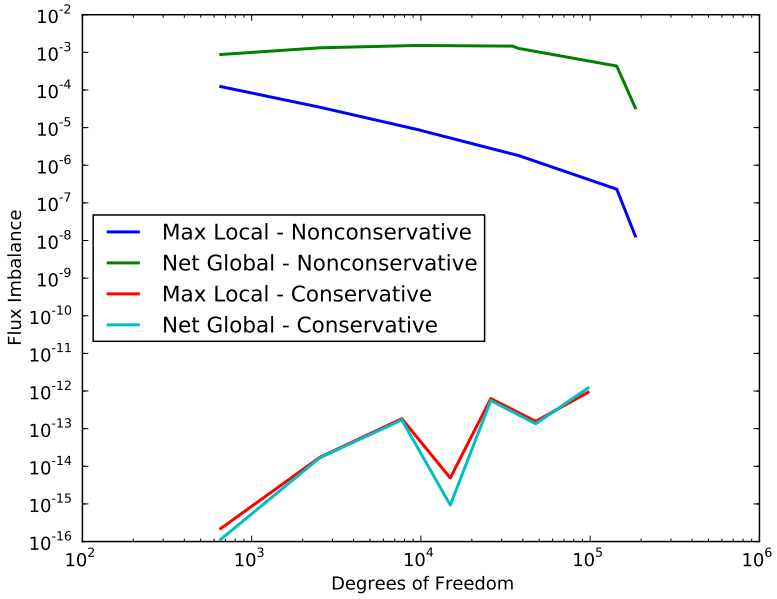
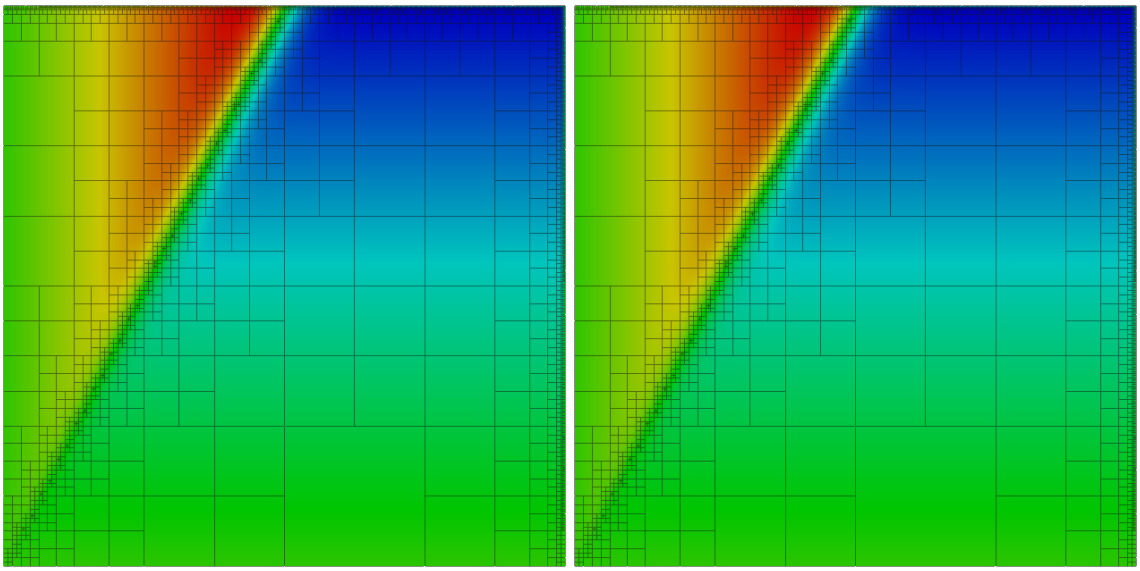


Figure 5: Flux imbalance in vortex solutions



(a) Nonconservative (b) Conservative

Figure 6: Discontinuous source problem after 8 refinements

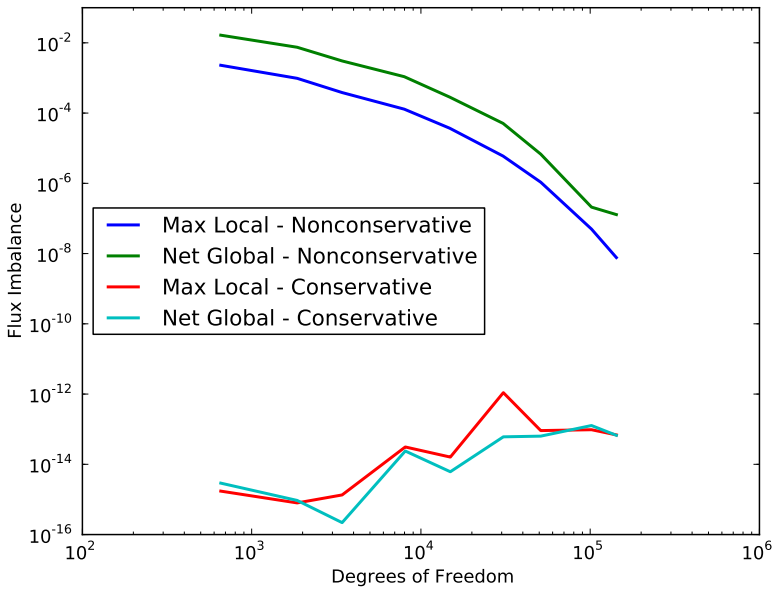
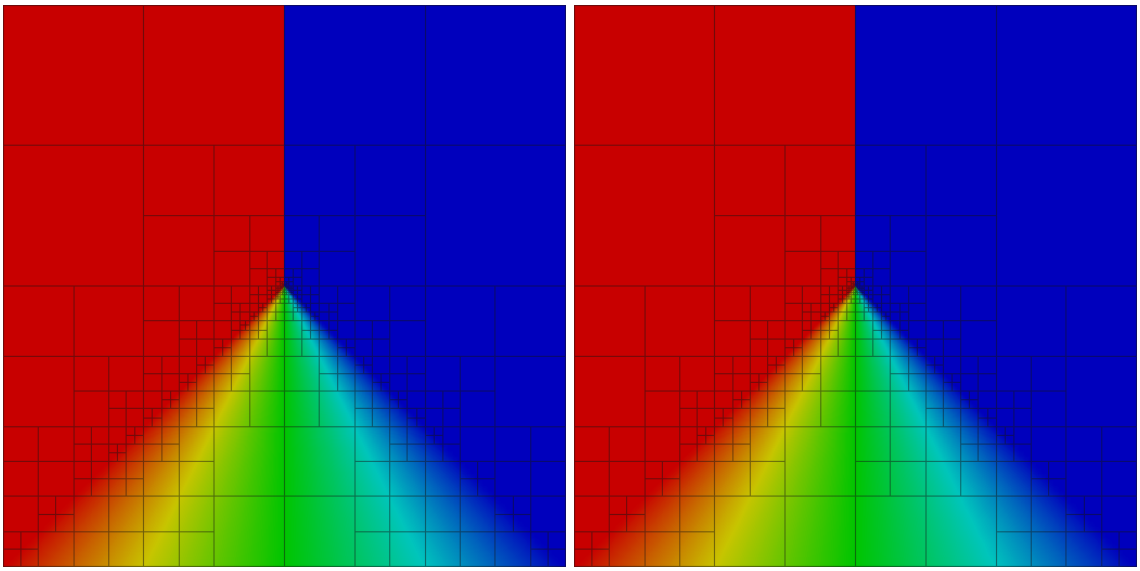


Figure 7: Flux imbalance in discontinuous source solutions



(a) Nonconservative (b) Conservative

Figure 8: Burgers' problem after 8 refinements

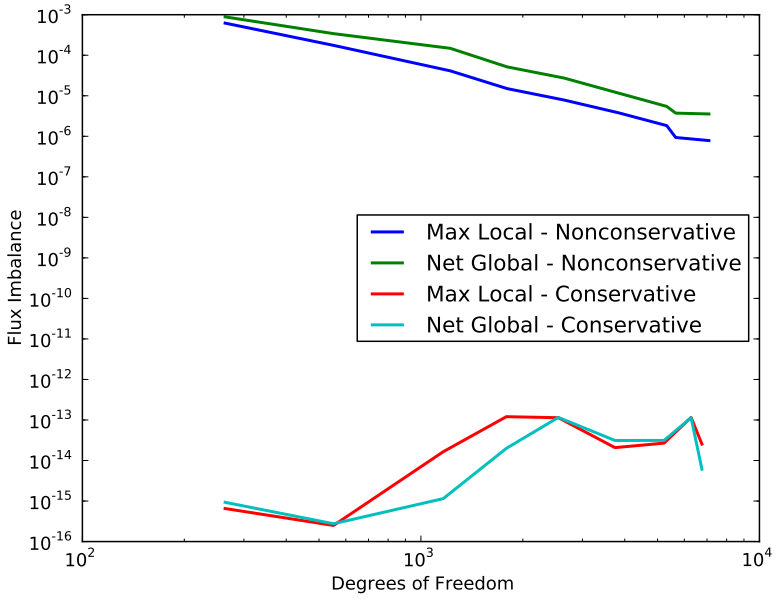


Figure 9: Flux imbalance in Burgers' solutions

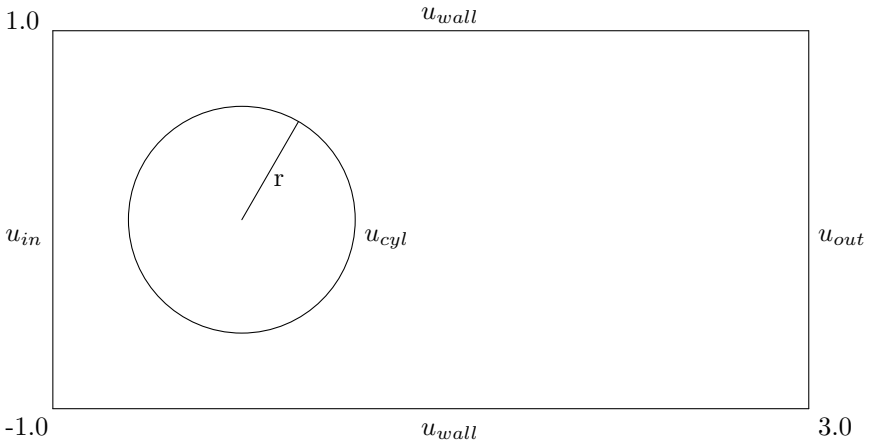
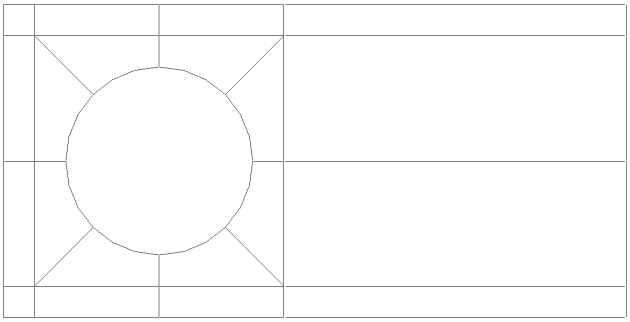
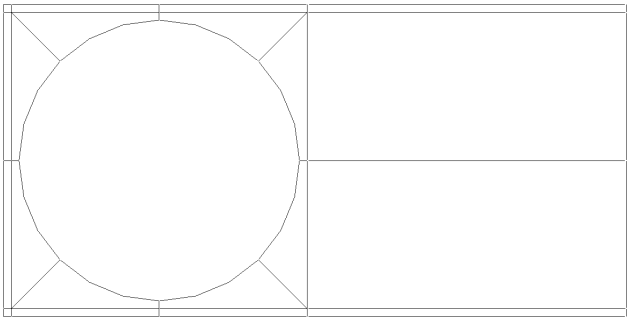


Figure 10: Stokes cylinder domain

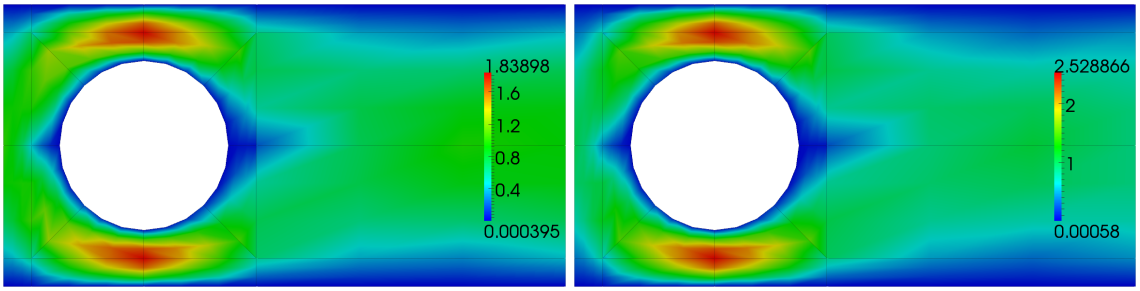


(a) Mesh for $r = 0.6$

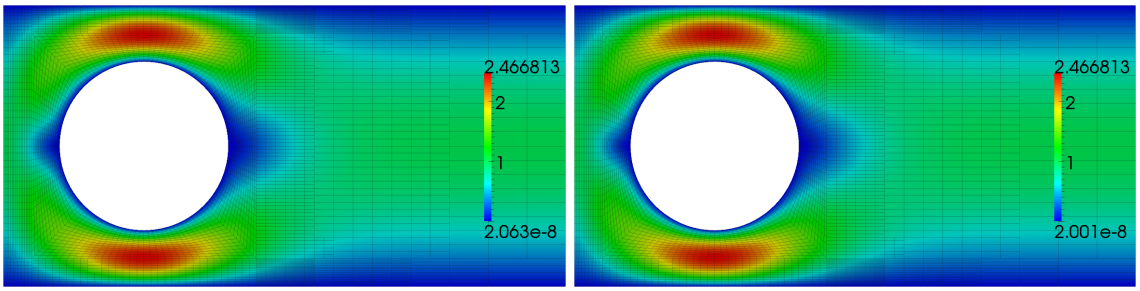


(b) Mesh for $r = 0.9$

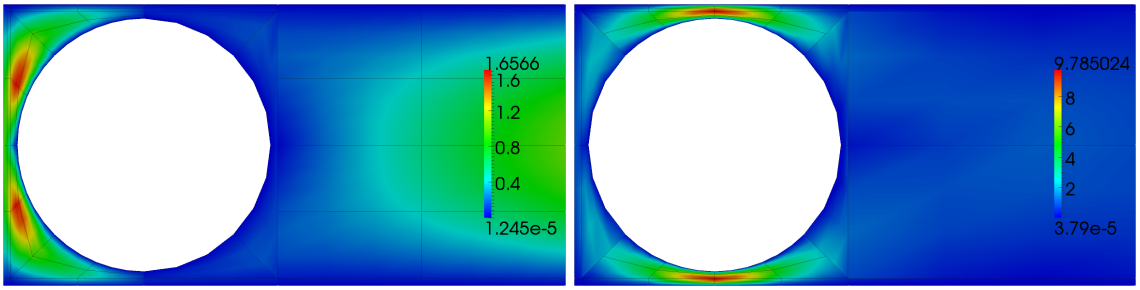
Figure 11: Initial mesh for Stokes flow over a cylinder



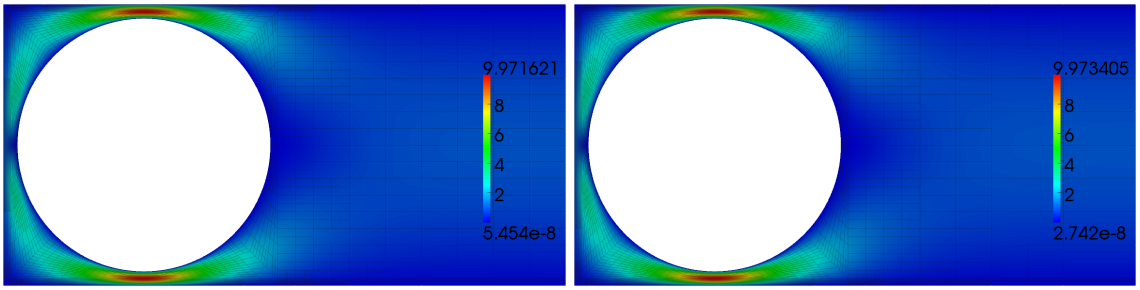
(a) Nonconservative on initial mesh with $r = 0.6$ (b) Conservative on initial mesh with $r = 0.6$



(c) Nonconservative after 6 refinements with $r = 0.6$ (d) Conservative after 6 refinements with $r = 0.6$

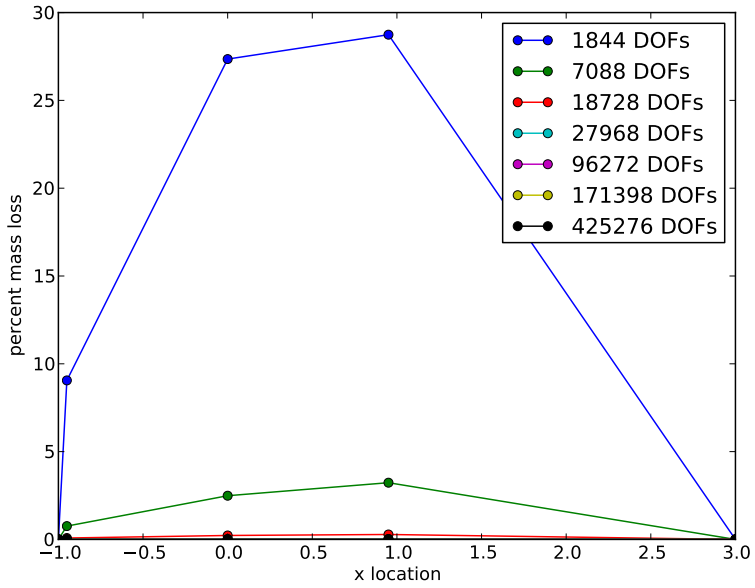


(e) Nonconservative after 1 refinement with $r = 0.9$ (f) Conservative after 1 refinement with $r = 0.9$

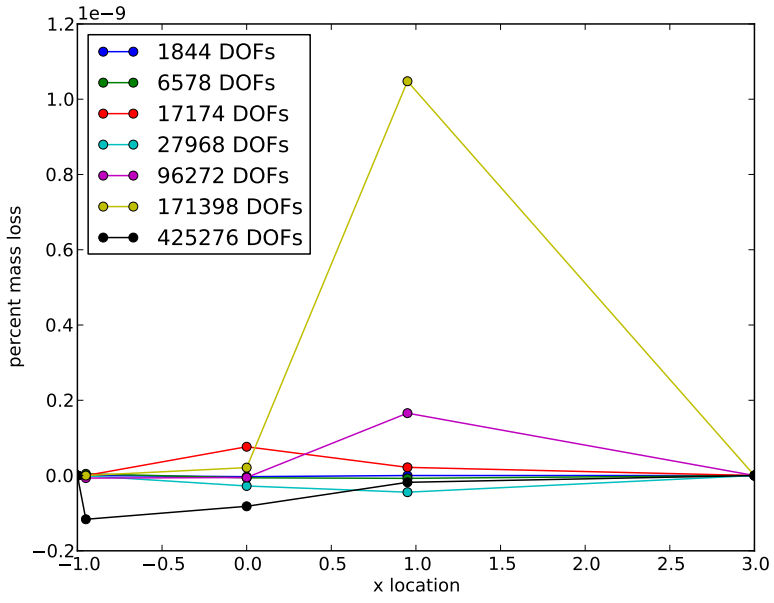


(g) Nonconservative after 6 refinements with $r = 0.9$ (h) Conservative after 6 refinements with $r = 0.9$

Figure 12: Stokes flow around a cylinder - velocity magnitude

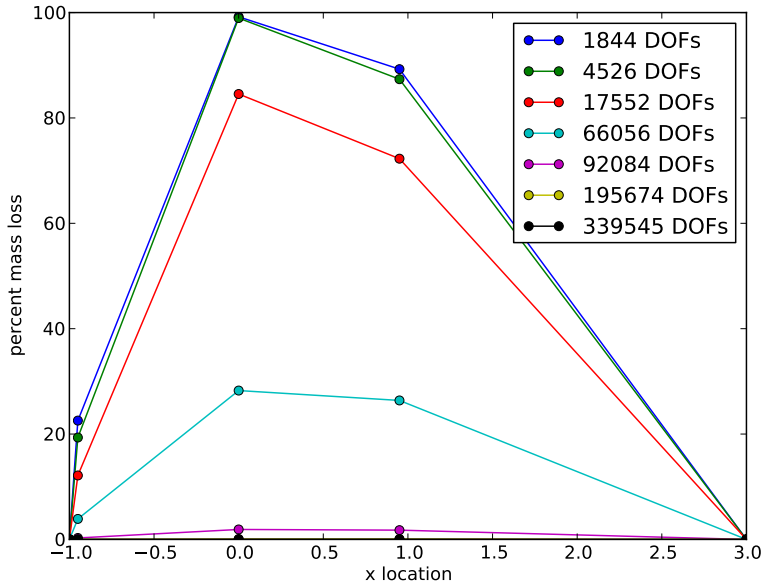


(a) Nonconservative

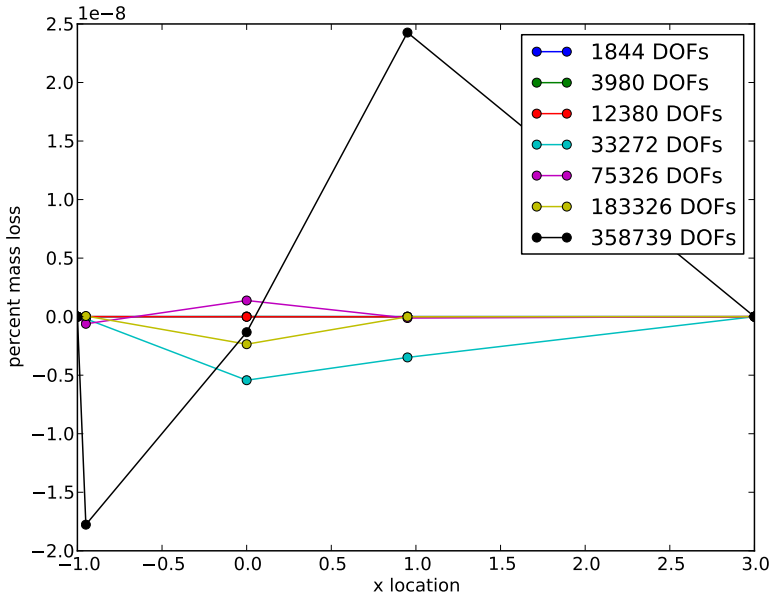


(b) Conservative

Figure 13: Mass loss in Stokes flow around a cylinder of radius 0.6



(a) Nonconservative



(b) Conservative

Figure 14: Mass loss in Stokes flow around a cylinder of radius 0.9

Figure 15: Stokes step domain

It should not come as a surprise that the standard and conservative solutions match each other so closely. The conservative formulation enforces local conservation more strictly, but if we examine the flux imbalance plots, the standard DPG formulation is nearly conservative on its own – and appears to become more conservative with refinement. The flux imbalance of the conservative methods appears to bounce around close to the machine epsilon (plus a few orders of magnitude). The level of enforcement appears to creep up with more degrees of freedom, indicating possible accruement of numerical error.

4.2.2 Burgers' Results

Standard and conservative DPG perform nearly identically for the inviscid Burgers' problem. It is obvious that the Lax-Wendroff condition of local conservation is a sufficient, but not necessary condition for numerical solutions to hyperbolic conservation laws. We see the same behavior with the flux imbalance plots that was so common with convection-diffusion.

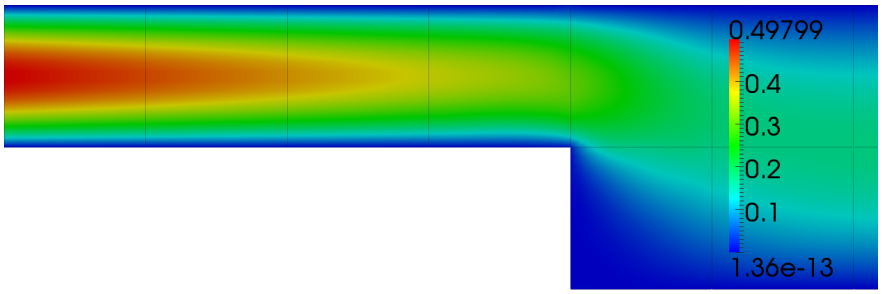
4.2.3 Stokes Results

The two Stokes problems are the first ones we encounter that stress the local conservation property of standard DPG. With a cylinder radius of 0.6, standard DPG loses nearly 30% of the mass post-cylinder, but quickly recovers most of that with further refinement. As we increase the cylinder radius to 0.9, the problem only exacerbates. Nearly 100% of the mass is lost in the constricted region on coarse meshes. It takes a much higher level of resolution to recover the mass loss. The stress singularity at the reentrant corner of the backward facing step causes issues for standard DPG on coarse meshes. It seems that the error in approximating the singularity outweighs the error of missed mass conservation. If we focus refinements at the singularity, the error eventually drops far enough for the method to become nearly conservative. The small amount of mass loss for the conservative method is clearly due to accumulation of floating point error.

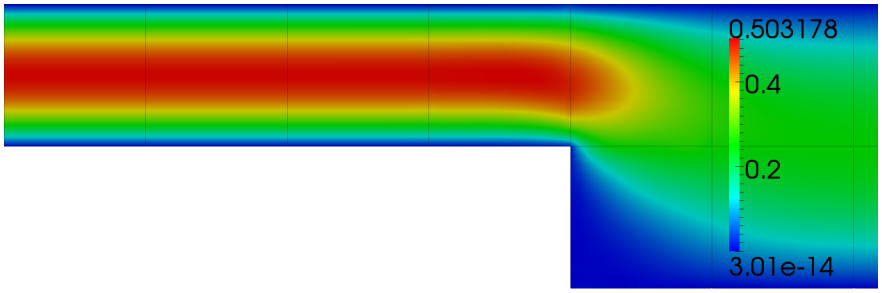
The most significant benefit of enforcing local conservation for these problems is that it allows us to recover the essential flow features with much coarser meshes. On the $r = 0.6$ cylinder problem, the peak velocity magnitude of the conservative solution is fairly close on the coarsest mesh, while the nonconservative solution severely underpredicts the peak. With the $r = 0.9$ cylinder, this problem is only worse. After just one adaptive refinement, the conservative solution nails the peak velocity. The nonconservative solution is completely useless at this point. We see the same thing with the backward facing step problem. The conservative solution preserves qualitative features even on the coarsest mesh, while standard DPG requires far higher resolution to achieve a similar solution.

5 Conclusions

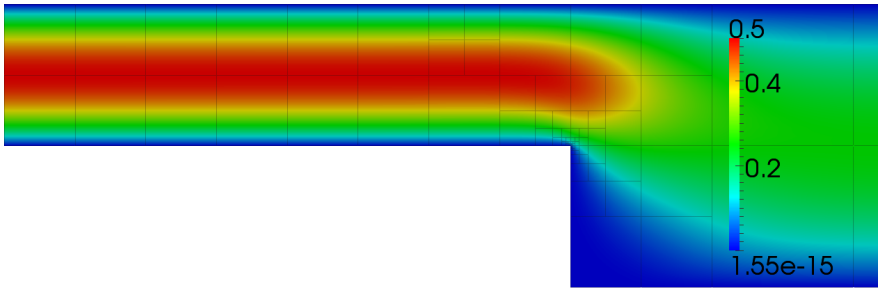
We have developed a locally conservative formulation of DPG by leveraging Lagrange multipliers. While the formulation is provably convergent and robust, it does increase the number of unknowns in the system and changes the structure from symmetric positive-definite to a saddle point problem. Numerical results indicate that the method delivers what it promises with local flux imbalances hovering close to machine precision, but they also indicate that for most of the problems considered, standard DPG is close to conservative and becomes more so with refinements. For the Stokes problems, standard DPG suffered similar mass loss as standard least squares on coarse meshes, but made up the lost mass with further resolution. For



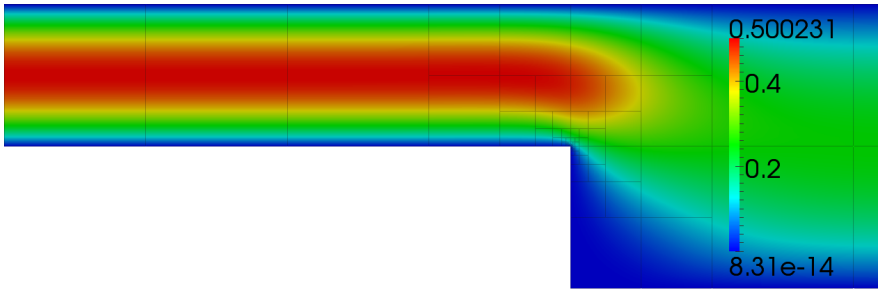
(a) Nonconservative on initial mesh



(b) Conservative on initial mesh

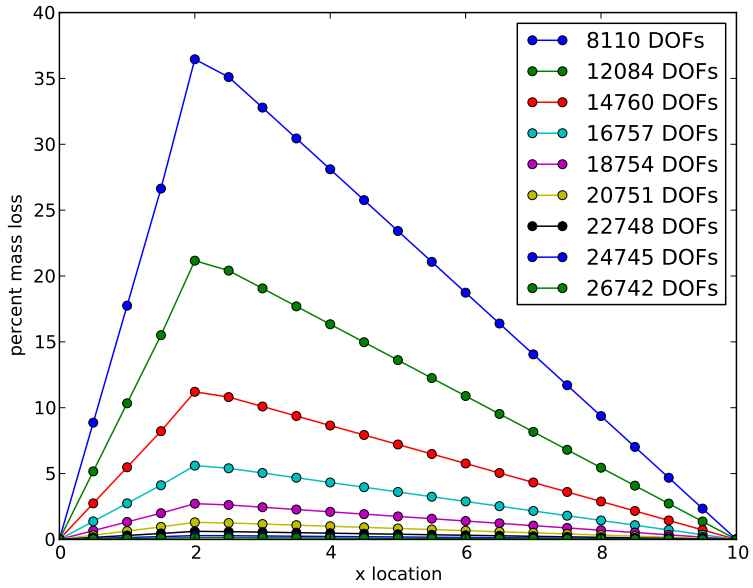


(c) Nonconservative after 8 refinement steps

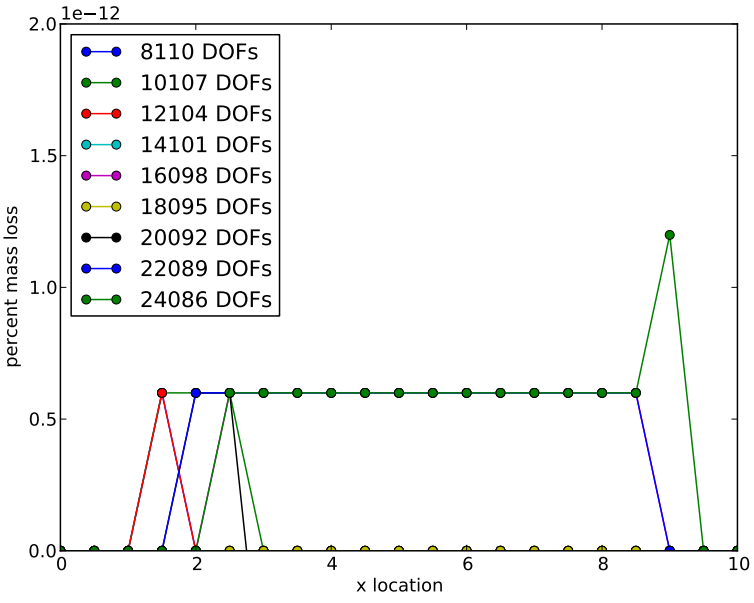


(d) Conservative after 8 refinement steps

Figure 16: Stokes backward facing step - velocity magnitude



(a) Nonconservative



(b) Conservative

Figure 17: Mass loss in Stokes backward facing step

problems where local conservation was stressed, conservative DPG was able to deliver reasonable solutions with much less resolution than standard DPG. Probably the most encouraging result of these experiments is that enforcing local conservation did not change the nature of the solutions too significantly. Standard DPG has a lot of attractive features, and we wish to preserve those.

References

[1] B. Ayuso and L.D. Marini. Discontinuous Galerkin methods for advection-diffusion-reaction problems. *SIAM J. Numer. Anal.*, 47(2):1391–1420, February 2009.

[2] P. Bochev, J. Lai, and L. Olson. A locally conservative, discontinuous least-squares finite element method for the Stokes equations. *Int. J. Numer. Methods Fluids*, 68:782–804, 2010.

[3] J. Bramwell, L. Demkowicz, J. Gopalakrishnan, and W. Qiu. A locking-free *hp* DPG method for linear elasticity with symmetric stresses. *Num. Math.*, 2012.

[4] F. Brezzi. On the existence, uniqueness, and approximation of saddle point problems arising from Lagrangian multipliers. *R.A.I.R.O., Anal. Numér.*, 2:129–151, 1974.

[5] J. Chan. *A DPG Method for Convection-Diffusion Problems*. PhD thesis, University of Texas at Austin, 2013.

[6] J. Chan, L. Demkowicz, and R. Moser. A DPG method for steady viscous compressible flow. *Comput. Fluids*, 98(0):69 – 90, 2014.

[7] J. Chan, L. Demkowicz, R. Moser, and N. Roberts. A class of discontinuous Petrov-Galerkin methods. Part V: Solution of 1D Burgers and Navier-Stokes equations. Technical Report 25, ICES, 2010.

[8] J. Chan, N. Heuer, T. Bui-Thanh, and L. Demkowicz. A robust DPG method for convection-dominated diffusion problems II: Adjoint boundary conditions and mesh-dependent test norms. *Comp. Math. Appl.*, 67(4):771 – 795, 2014. High-order Finite Element Approximation for Partial Differential Equations.

[9] C. Chang and J. Nelson. Least-squares finite element method for the Stokes problem with zero residual of mass conservation. *SIAM J. Num. Anal.*, 34:480–489, 1997.

[10] M. Costabel and A. McIntosh. On Bogovskiĭ and regularized Poincaré integral operators for de Rham complexes on Lipschitz domains. *Mathematische Zeitschrift*, 265(2):297–320, 2010.

[11] L. Demkowicz. Babuška ↔ Brezzi? Technical report, ICES, 2006.

[12] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. Part I: The transport equation. *Comput. Methods Appl. Mech. Engrg.*, 2009.

[13] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. Part II: Optimal test functions. *Numer. Meth. Part. D. E.*, 2010.

[14] L. Demkowicz and J. Gopalakrishnan. Analysis of the DPG method for the Poisson problem. *SIAM J. Num. Anal.*, 49(5):1788–1809, 2011.

[15] L. Demkowicz and J. Gopalakrishnan. A primal DPG method without a first order reformulation. *Comp. Math. Appl.*, 66:1058–1064, 2013.

[16] L. Demkowicz and J. Gopalakrishnan. *Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations (eds. X. Feng, O. Karakashian, Y. Xing)*, volume 157, chapter An Overview of the DPG Method, pages 149–180. IMA Volumes in Mathematics and its Applications, 2014.

[17] L. Demkowicz, J. Gopalakrishnan, I. Muga, and J. Zitelli. Wavenumber explicit analysis for a DPG method for the multidimensional Helmholtz equation. *Comput. Methods Appl. Mech. Engrg.*, 213-216:126–138, 2012.

[18] L. Demkowicz, J. Gopalakrishnan, and A. Niemi. A class of discontinuous Petrov-Galerkin methods. Part III: Adaptivity. *Appl. Numer. Math.*, 62(4):396–427, April 2012.

[19] L. Demkowicz and N. Heuer. Robust DPG method for convection-dominated diffusion problems. *SIAM J. Numer. Anal.*, 51(5):1514–2537, 2013.

[20] J. Gopalakrishnan and W. Qiu. An analysis of the practical DPG method. *Math. Comp.*, 2012.

[21] D. Moro, N.C. Nguyen, and J. Peraire. A hybridized discontinuous Petrov-Galerkin scheme for scalar conservation laws. *Int.J. Num. Meth. Eng.*, 2011.

[22] J.B. Perot. Discrete conservation properties of unstructured mesh schemes. *Annu. Rev. Fluid Mech.*, 43:299–318, 2011.

[23] N. Roberts. *A Discontinuous Petrov-Galerkin Methodology for Incompressible Flow Problems*. PhD thesis, University of Texas at Austin, 2013.

[24] N. Roberts, T. Bui-Thanh, and L. Demkowicz. The DPG method for the Stokes problem. *Comp. Math. Appl.*, 67(4):966 – 995, 2014. High-order Finite Element Approximation for Partial Differential Equations.