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

1 Intoduction

The discontinuous Petrov-Galerkin (DPG) method with optimal test functions has been under active development for convection-diffusion type systems [4, 5, 7, 1, 8, 2, 9]. In this paper, we develop a theory for the locally conservative formulation of DPG for convection-diffusion type equations 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[10] 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.

The discontinuous Petrov-Galerkin finite element method has been described as least squares finite elements with a twist. The key difference is that least square methods seek to minimize the residual of the solution in the L^2 norm, while DPG seeks the minimization in a dual norm realized through the inverse Riesz map. Exact mass conservation has been an issue that has plagued least squares finite elements for a long time. Several approaches have been used to try to adress this. Chang and Nelson[3] developed the *restricted LSFEM*[3] by augmenting the least squares equations with a Lagrange multiplier explicitly enforcing mass conservation element-wise. Our conservative formulation of DPG takes a similar approach and both methods share similar negative of transforming a minimization method to a saddle-point problem.

1.2 DPG is a Minimum Residual Method

Roberts *et al.* presents a brief history and derivation of DPG with optimal test functions in [11]. We follow his derivation of the standard DPG method as a minumum 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 $B : U \to V'$. We seek to minimize the residual for the discrete space $U_h \subset U$:

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

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 $T \coloneqq R_V^{-1} B : U_h \to V$ 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, (5) still requires the inversion of the infinite-dimensional Riesz map, but approximating V by a finite dimensional V_h which is of a higher polynomial degree than U_h (hence "enriched space") works well in practice.

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.

2 Analysis

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 *skeleton* Γ_h and *interior skeleton* Γ_h^0,

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

We define broken Sobolev spaces element-wise:

$$H^1(\Omega_h) \quad \coloneqq \prod_K H^1(K),$$

$$\boldsymbol{H}(\mathrm{div}, \Omega_h) \quad \coloneqq \prod_K \boldsymbol{H}(\mathrm{div}, K).$$

We also need the trace spaces:

$$H^{\frac{1}{2}}(\Gamma_h) := \left\{\hat{v} = \{\hat{v}_K\} \in \prod_K H^{1/2}(\partial K) \,:\, \exists v \in H^1(\Omega) : v|_{\partial K} = \hat{v}_K\right\},$$

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

which are developed more precisely in [11].

2.1 Element Conservative Convection-Diffusion

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

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

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:

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

If we then multiply the top equation by some scalar test function v and the bottom equation by some vector-valued test function τ, we can integrate by parts over each element K:

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

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 considerations of DPG in the literature [] 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. 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

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

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

$$\tag{9}$$

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

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 a group variable. The element conservative DPG scheme is derived from the Lagrangian:

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

where $(1_K, \mathbf{0})$ is the test function in which $v = 1$ on element K and 0 elsewhere and $\boldsymbol{\tau} = \mathbf{0}$ everywhere.

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

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

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

Since the trial-to-test operator selectively produces test functions in $H^1(\Omega_h)$ and $\boldsymbol{H}(\mathrm{div}, \Omega_h)$, denote the result $v_{\delta\boldsymbol{u}_h}$ when $T(\delta u_h) \in H^1(\Omega_h)$ and $\boldsymbol{\tau}_{\delta\boldsymbol{u}_h}$ when $T(\delta u_h) \in \boldsymbol{H}(\mathrm{div}, \Omega_h)$. Then, putting (11) into more concrete terms for convection-diffusion, we get:

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

2.2 Local Problem

The optimal test functions are determined by solving local problems determined by the choice of test norm. We consider several options for the test norm. The graph norm [6] 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.* [2] derived a more robust alternative norm for convection diffusion (dubbed the robust norm). We recently developed a modification of the robust norm that appears to produce better results in the presence of singularities.

2.2.1 Robust Norm

For illustrative purposes we choose the robust test norm developed by Chan *et al.* [2]. With this choice, we get our local problem:

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

$$\begin{aligned} (\nabla \cdot \boldsymbol{\tau}_{\delta\boldsymbol{u}_h}, \nabla \cdot \delta\boldsymbol{\tau})_K + \min\left\{\frac{1}{\epsilon}, \frac{1}{|K|}\right\}(\boldsymbol{\tau}_{\delta\boldsymbol{u}_h}, \delta\boldsymbol{\tau})_K + \epsilon(\nabla v_{\delta\boldsymbol{u}_h}, \nabla\delta v)_K + (\boldsymbol{\beta}\cdot\nabla v_{\delta\boldsymbol{u}_h}, \boldsymbol{\beta}\cdot\nabla\delta v)_K & \\ + \alpha \min\left\{\sqrt{\frac{\epsilon}{|K|}}, 1\right\}(v_{\delta\boldsymbol{u}_h}, \delta v)_K = b(\delta\boldsymbol{u}_h, (\delta v, \delta\boldsymbol{\tau})) \qquad \forall \delta v \in H^{-1}(K),\, \delta\boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div}, K)\,, & \end{aligned} \tag{13}$$

where, typically, $\alpha = 1$.

We can decompose the discrete flux space $\hat{F}_h$ into element conserving fluxes and some algebraic complement:

$$\hat{F}_h = \hat{F}_h^c \oplus \hat{F}_h^{nc}\,, \quad \hat{F}_h^c := \{\hat{t} : \langle \hat{t}, (1_K, \mathbf{0}) \rangle = 0 \quad \forall K\}\,. \tag{14}$$

This may be accomplished by employing for fluxes in $\hat{F}_h^c$ traces of curls, $\nabla \times \phi$.

In practice, we solve (12) as a fully coupled system of equations, but for didactic purposes, we can examine it one piece at a time. If we restrict $(12)_1$ to $\delta\hat{t} \in \hat{F}_h^{nc}$, the Lagrange multiplier term drops out and we left with a system that we can solve for u_h, $\boldsymbol{\sigma}_h$, $\hat{u}_h$, and $\hat{t}_h^c \in \hat{F}_h^c$.

Restricting ourselves to $\delta\hat{t} \in \hat{F}_h^c$ in $(12)_1$, we obtain a system of equations that can be solved for u_h, $\boldsymbol{\sigma}_h$, $\hat{u}$, and $\hat{t} \in \hat{F}_h^c$. The remaining fluxes, $\hat{t} \in \hat{F}_h^{nc}$ are determined from $(12)_2$. Finally, testing with $\hat{t} \in \hat{F}_h^{nc}$ in $(12)_1$, we obtain a system of equations that can be solved for Lagrange multipliers λ_K.

Additionally, we can pass in local problem (13) with $\alpha \to 0$. The fact that the test functions will be determined then up to a constant does not matter, for $\delta\hat{t} \in \hat{F}_h^c$, equation $(12)_1$ is orthogonal to constants.

Mathematically, we are dealing with equivalence classes of functions, but in order to obtain a single function that we can deal with numerically, we replace the alpha term with a zero mean scaling condition to obtain the new test norm,

$$(\nabla \cdot \boldsymbol{\tau}_{\delta \boldsymbol{u}_h}, \nabla \cdot \delta\boldsymbol{\tau})_K + \min\left\{\frac{1}{\epsilon}, \frac{1}{|K|}\right\}(\boldsymbol{\tau}_{\delta \boldsymbol{u}_h}, \delta\boldsymbol{\tau})_K \\ + \epsilon(\nabla v_{\delta \boldsymbol{u}_h}, \nabla \delta v)_K + (\boldsymbol{\beta} \cdot \nabla v_{\delta \boldsymbol{u}_h}, \boldsymbol{\beta} \cdot \nabla \delta v)_K + \frac{1}{|K|^2}\int_K v_{\delta \boldsymbol{u}_h} \delta v \,, \tag{15}$$

where the $\frac{1}{|K|^2}$ coefficient is needed to make the zero mean term scale correctly with h. Numerically this helps a good deal with the condition numbers of the local solve.

It is convenient to be able to take $\alpha \to 0$ as we will see in some later numerical experiments.

2.3 Proof of Convergence

We cannot directly use Brezzi's theorem, but we can reproduce his argument. First of all, variational formulation (7) is equivalent to:

Find $\boldsymbol{u} \in \boldsymbol{U}$ such that

$$b(\boldsymbol{u}, T(\delta \boldsymbol{u})) = (f, T(\delta \boldsymbol{u})) \quad \forall \boldsymbol{v} \in \boldsymbol{V} \,. \tag{16}$$

This is because the trial-to-test operator $T = R_V^{-1} B$ is an isomorphism.

Restricting ourselves to $\delta \hat{t}_h \in \hat{F}_h^c$, and subtracting (16) from (12), we have the Galerkin orthogonality condition:

$$\frac{1}{\epsilon}(\boldsymbol{\sigma}_h - \boldsymbol{\sigma}, \boldsymbol{\tau}_{\delta \boldsymbol{u}_h}) + (u_h - u, \nabla \cdot \boldsymbol{\tau}_{\delta \boldsymbol{u}_h}) - \langle \hat{u}_h - \hat{u}, \boldsymbol{\tau}_{\delta \boldsymbol{u}_h} \cdot \boldsymbol{n} \rangle \\ - (\boldsymbol{\beta}(u_h - u) - (\boldsymbol{\sigma}_h - \boldsymbol{\sigma}), \nabla v_{\delta \boldsymbol{u}_h}) + \langle \hat{t}_h - \hat{t}, v_{\delta \boldsymbol{u}_h} \rangle \tag{17}$$
$$\forall (\delta u_h, \delta \boldsymbol{\sigma}_h, \delta \hat{u}_h, \delta \hat{t}_h) \in U_h \times \boldsymbol{S}_h \times \hat{U}_h \times \hat{F}_h^c \,.$$

We begin with an equivalent of Brezzi's *discrete inf-sup in kernel condition.*

Lemma 1. *The following condition holds:*

$$\sup_{\substack{(\delta u_h, \delta \boldsymbol{\sigma}_h, \delta \hat{u}, \delta \hat{t}) \in \\ U_h \times \boldsymbol{S}_h \times \hat{U}_h \times \hat{F}_h^c}} \frac{\left| b((u_h, \boldsymbol{\sigma}_h, \hat{u}_h, \hat{t}_h), (v_{\delta \boldsymbol{u}_h}, \boldsymbol{\tau}_{\delta \boldsymbol{u}_h})) \right|}{\|(v_{\delta \boldsymbol{u}_h}, \boldsymbol{\tau}_{\delta \boldsymbol{u}_h})\|} \geq \alpha \left\|(u_h, \boldsymbol{\sigma}_h, \hat{u}_h, \hat{t}_h)\right\|$$
$$\forall (u_h, \boldsymbol{\sigma}_h, \hat{u}_h, \hat{t}_h) \in U_h \times \boldsymbol{S}_h \times \hat{U}_h \times \hat{F}_h^c \,. \tag{18}$$

Proof. By the very construction of the trial-to-test operator, the supremum on the left is equal to

$$\left\|T(u_h, \boldsymbol{\sigma}_h, \hat{u}_h, \hat{t}_h)\right\|_V = \left\|R_V^{-1} B(u_h, \boldsymbol{\sigma}_h, \hat{u}_h, \hat{t}_h)\right\|_V = \left\|B(u_h, \boldsymbol{\sigma}_h, \hat{u}_h, \hat{t}_h)\right\|_{V'} \tag{19}$$

The condition follows now from the fact that operator B is bounded below, i.e. bilinear form $b((u_h, \boldsymbol{\sigma}_h, \hat{u}_h, \hat{t}_h), (v, \boldsymbol{\tau}))$ satisfies the inf-sup condition. □

Assume additionally that flux decomposition (14) is orthogonal. We begin by decomposing $\hat{t}_h$,

$$\hat{t}_h = \hat{t}_h^c + \hat{t}_h^{nc} \,. \tag{20}$$

Next, we use the triangle inequality,

$$\left\|(u - u_h, \boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \hat{u} - \hat{u}_h, \hat{t} - \hat{t}_h)\right\| = \left\|(u - u_h, \boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \hat{u} - \hat{u}_h, \hat{t} - (\hat{t}_h^c + \hat{t}_h^{nc}))\right\| \\ \leq \left\|(u - w_h, \boldsymbol{\sigma} - \boldsymbol{s}_h, \hat{u} - \hat{w}_h, \hat{t} - (\hat{r}_h^c + \hat{t}_h^{nc}))\right\| \\ + \left\|(w_h - u_h, \boldsymbol{s}_h - \boldsymbol{\sigma}_h, \hat{w}_h - \hat{u}_h, \hat{r}_h^c - \hat{t}_h^c)\right\| \tag{21}$$

where $w_h \in U_h$, $\boldsymbol{s}_h \in \boldsymbol{S}_h$, $\hat{w}_h \in \hat{U}_h$, and $\hat{r}_h^c \in \hat{F}_h$ are arbitrary. The inf-sup in kernel condition now implies:

$$\begin{aligned}
&\left\|(w_h - u_h, \boldsymbol{s}_h - \boldsymbol{\sigma}_h, \hat{w}_h - \hat{u}_h, \hat{r}_h^c - \hat{t}_h^c)\right\| \qquad (22)\\
&\quad \leq \frac{1}{\alpha} \sup_{\substack{(\delta u_h, \delta \boldsymbol{\sigma}_h, \delta \hat{u}, \delta \hat{t}) \in \\ U_h \times \boldsymbol{S}_h \times \hat{U}_h \times \hat{F}_h^c}} \frac{\left|b((w_h - u_h, \boldsymbol{s}_h - \boldsymbol{\sigma}_h, \hat{w}_h - \hat{u}_h, \hat{r}_h^c - \hat{t}_h^c), (v_{\delta \boldsymbol{u}_h}, \boldsymbol{\tau}_{\delta \boldsymbol{u}_h}))\right|}{\|(v_{\delta \boldsymbol{u}_h}, \boldsymbol{\tau}_{\delta \boldsymbol{u}_h})\|}\\
&\quad \leq \frac{1}{\alpha} \sup_{\substack{(\delta u_h, \delta \boldsymbol{\sigma}_h, \delta \hat{u}, \delta \hat{t}) \in \\ U_h \times \boldsymbol{S}_h \times \hat{U}_h \times \hat{F}_h^c}} \frac{\left|b((w_h - u_h, \boldsymbol{s}_h - \boldsymbol{\sigma}_h, \hat{w}_h - \hat{u}_h, \hat{r}_h^c + \hat{t}_h^{nc} - \hat{t}), (v_{\delta \boldsymbol{u}_h}, \boldsymbol{\tau}_{\delta \boldsymbol{u}_h}))\right|}{\|(v_{\delta \boldsymbol{u}_h}, \boldsymbol{\tau}_{\delta \boldsymbol{u}_h})\|}\\
&\quad \leq \frac{1}{\alpha} \left\|(w_h - u_h, \boldsymbol{s}_h - \boldsymbol{\sigma}_h, \hat{w}_h - \hat{u}_h, \hat{t} - (\hat{r}_h^c + \hat{t}_h^{nc}))\right\| \qquad (23)
\end{aligned}$$

where the second step follows from the Galerkin orthogonality condition (17), and the last step is a consequence of the minimum energy extension norm for fluxes. Cobining with (21), we get the final stability result:

$$\left\|(u - u_h, \boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \hat{u} - \hat{u}_h, \hat{t} - \hat{t}_h)\right\| \leq \left(1 + \frac{1}{\alpha}\right) \inf_{\substack{(w_h, \boldsymbol{s}_h, \hat{w}, \hat{r}_h^c) \in \\ U_h \times \boldsymbol{S}_h \times \hat{U}_h \times \hat{F}_h^c}} \left\|(u - w_h, \boldsymbol{\sigma} - \boldsymbol{s}_h, \hat{u} - \hat{w}, \hat{t} - (\hat{r}_h^c + \hat{t}_h^{nc}))\right\| . \qquad (24)$$

3 Numerical Experiments

In 3.1 we define each numerical experiment, and in 3.2 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 ($\hat{t}$), and fifth order test functions (v and $\boldsymbol{\tau}$).

Describe how we check local conservation.

3.1 Description of Problems

Unless otherwise noted, the problem domain is $\Omega = [0,1]^2$ and $f = 0$. Also note that except for the discontinuous source problem, blue corresponds to $u = 0$ and red to $u = 1$ with a linear scaling in between.

3.1.1 Erickson-Johnson Model Problem

The Erickson-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

$$u(x,y) = C_0 + \sum_{n=1}^{\infty} C_n \frac{\exp(s_n(x-1)) - \exp(r_n(x-1))}{r_n \exp(-s_n) - s_n \exp(-s_n)} \cos(n\pi y) . \qquad (25)$$

Taking $\epsilon = 10^{-2}$, $C_1 = 1$, and $C_{n\neq 1} = 0$, the exact solution looks like Figure 1.

3.1.2 Skewed Convection-Diffusion Problem

This is a standard convection-diffusion problem with a skewed convection vector relative to the mesh. We take $\boldsymbol{\beta} = (2,1)$, and on the left and bottom inflow boundaries we impose $\hat{t} = 1 - x - y$ while the right and top boundaries have $\hat{u} = 0$. We solve for $\epsilon = 10^{-4}$.

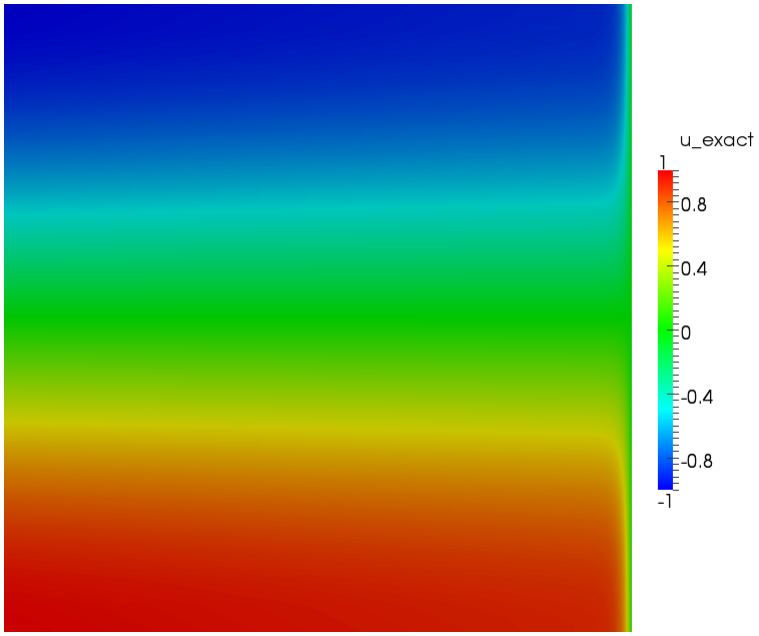


Figure 1: Erickson-Johnson exact solution

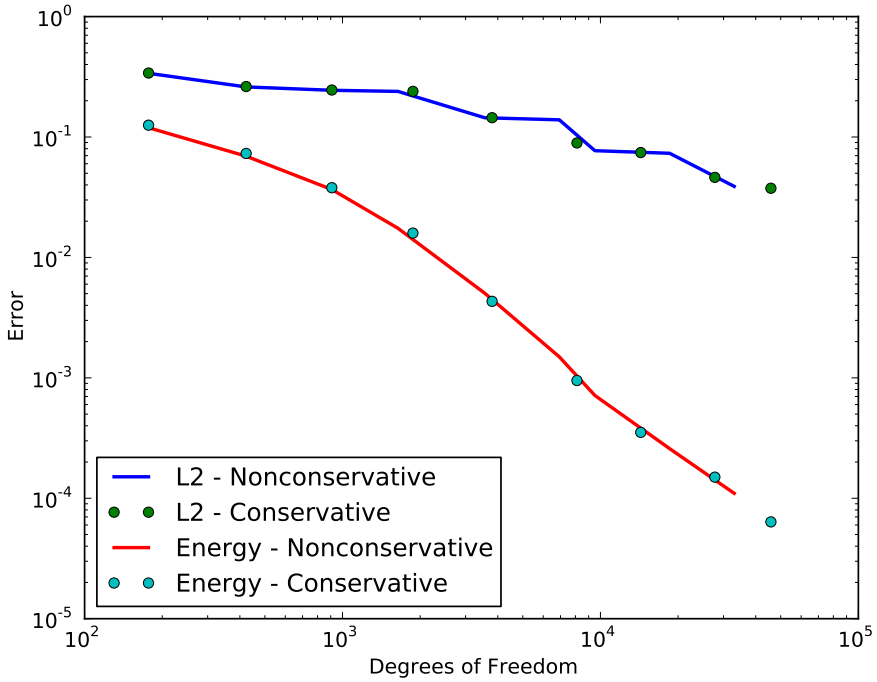


Figure 2: Error in Erickson-Johnson solutions

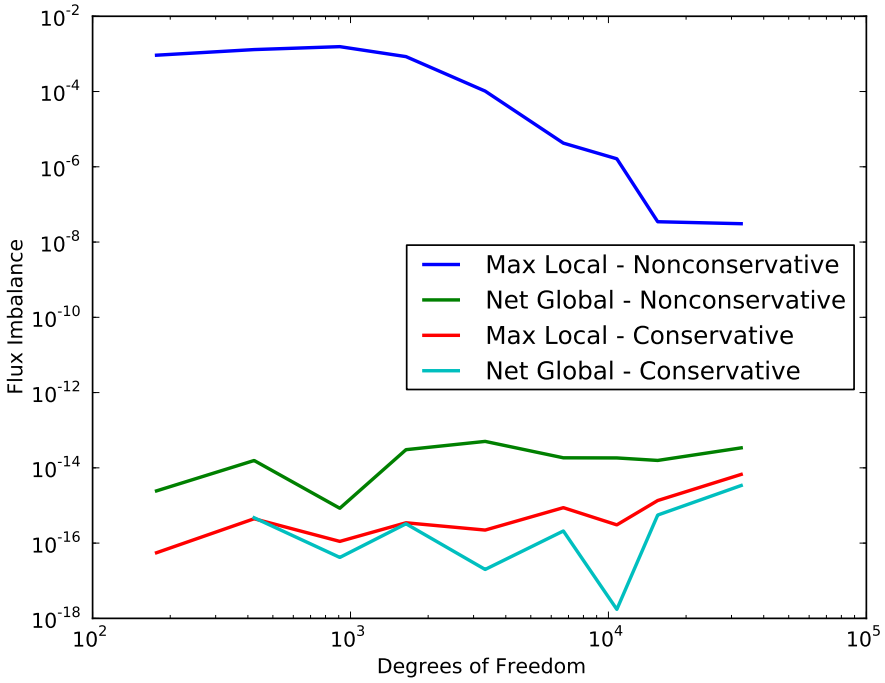


Figure 3: Flux imbalance in Erickson-Johnson solutions

3.1.3 Double Glazing Problem

This nominal problem definition takes the same unit square domain with

$$\boldsymbol{\beta} = \begin{pmatrix} 2(2y-1)(1-(2x-1)^2) \\ -2(2x-1)(1-(2y-1)^2) \end{pmatrix}; \quad \hat{u} = \begin{cases} 1 & \text{on } \Gamma_{right} \\ 0 & \text{else} \end{cases},$$

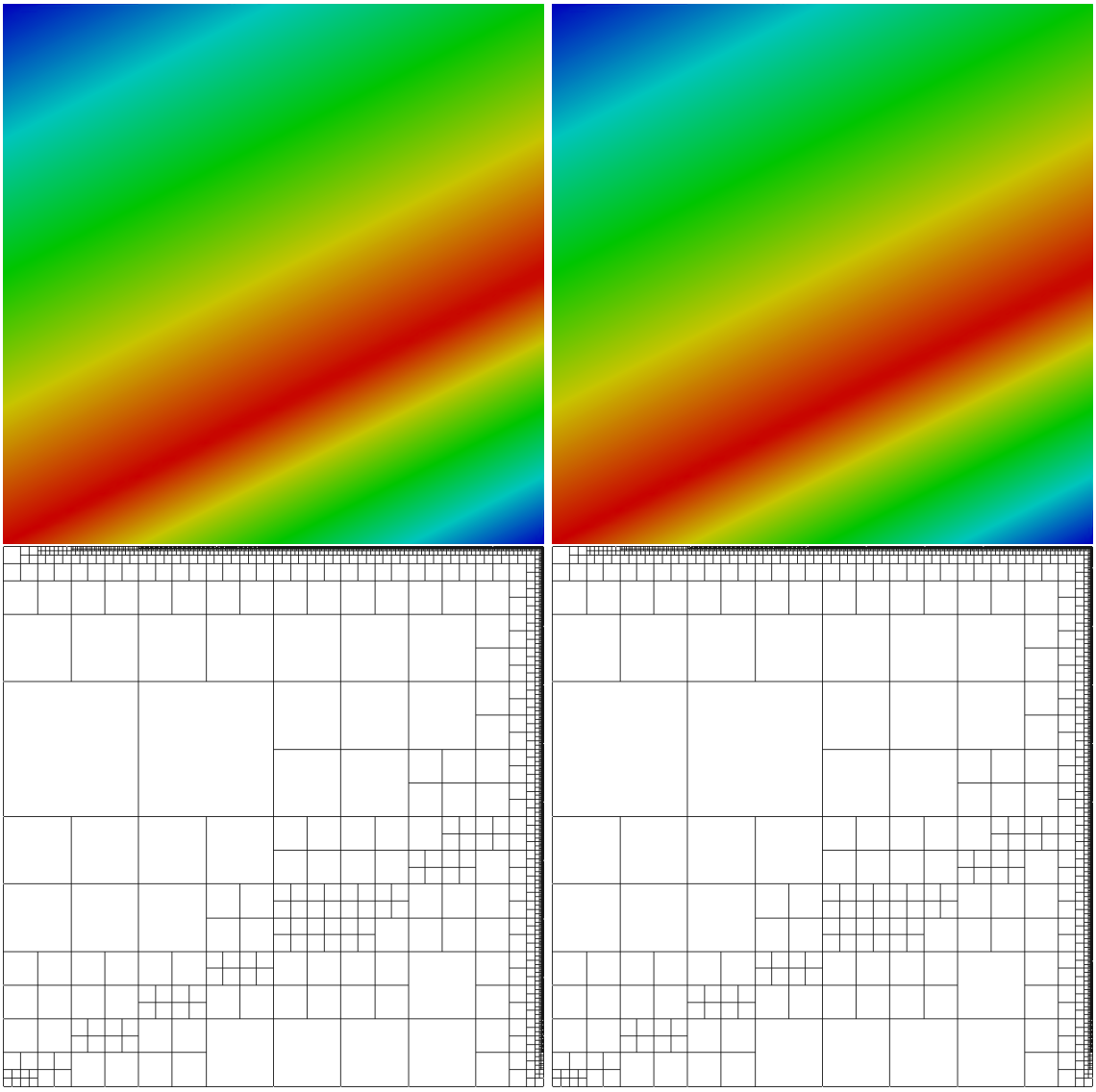
except that this definition for the boundary is not legal for $\hat{u} \in H^{\frac{1}{2}}(\Gamma_h)$ which is continuous. Therefore we use a ramp of width $\sqrt{\epsilon}$ on the right edge to go from 0 to 1. The results are for $\epsilon = 10^{-2}$.

3.1.4 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.

3.1.5 Wedge Problem

We devised this problem to examine some of the issues we were having blow up of the solution under the robust test norm. The domain is a rectangle $[-0.5,0.5] \times [-1,0.5]$ where the triangle connecting the bottom edges with the origin at $(0,0)$ is removed. The convection vector $\boldsymbol{\beta} = (1,0)^T$, $\epsilon = 10^{-1}$, and we apply boundary conditions $\hat{t} = 0$ on the left inflow edge, $\hat{u} = 0$ on the top edge, $\boldsymbol{\beta} \cdot \mathbf{n} \cdot \hat{u} - \hat{t} = \sigma_n = 0$ on the right outflow edge, $\hat{u} = 1$ on the leading edge of the wedge, and $\hat{t} = \boldsymbol{\beta} \cdot \mathbf{n} \cdot 1$ on the trailing edge of the wedge.



(a) Nonconservative (b) Conservative

Figure 4: Skewed convection-diffusion problem after 8 refinements

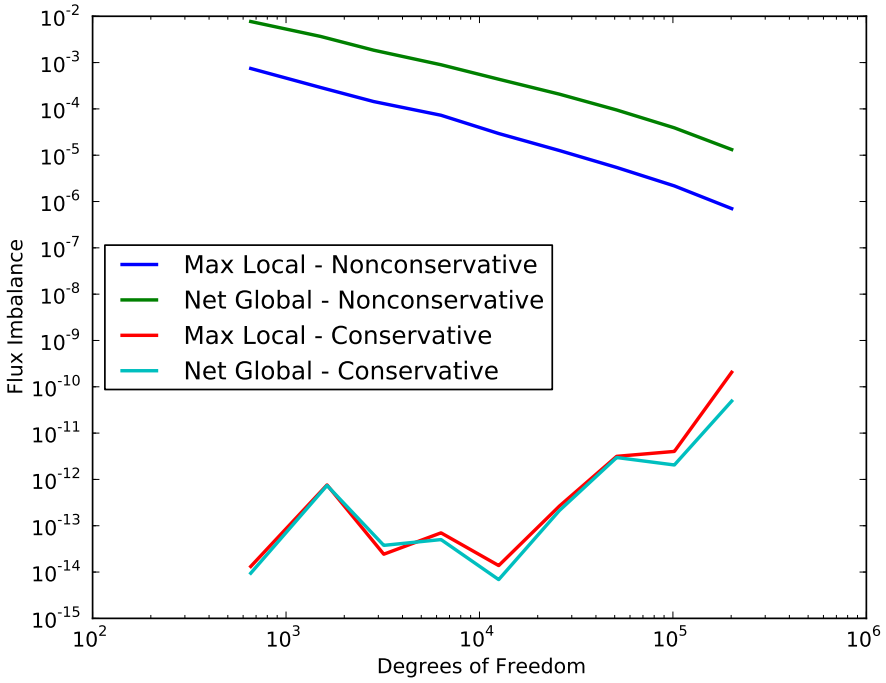


Figure 5: Flux imbalance in skewed convection-diffusion solutions

3.1.6 Inner Layer Problem

In this problem, $\boldsymbol{\beta} = (\frac{\sqrt{3}}{2}, \frac{1}{2})^T$ and we have a discontinuous flux inflow condition where $\hat{t} = 0$ if $y \leq 0.2$ and $\hat{t} = \boldsymbol{\beta} \cdot \mathbf{n}$ if $y > 0.2$. On the outflow, $\hat{u} = 0$. We use a very small diffusion scale of $\epsilon = 10^{-6}$ which creates a very thin inner layer.

3.1.7 Discontinuous Source Problem

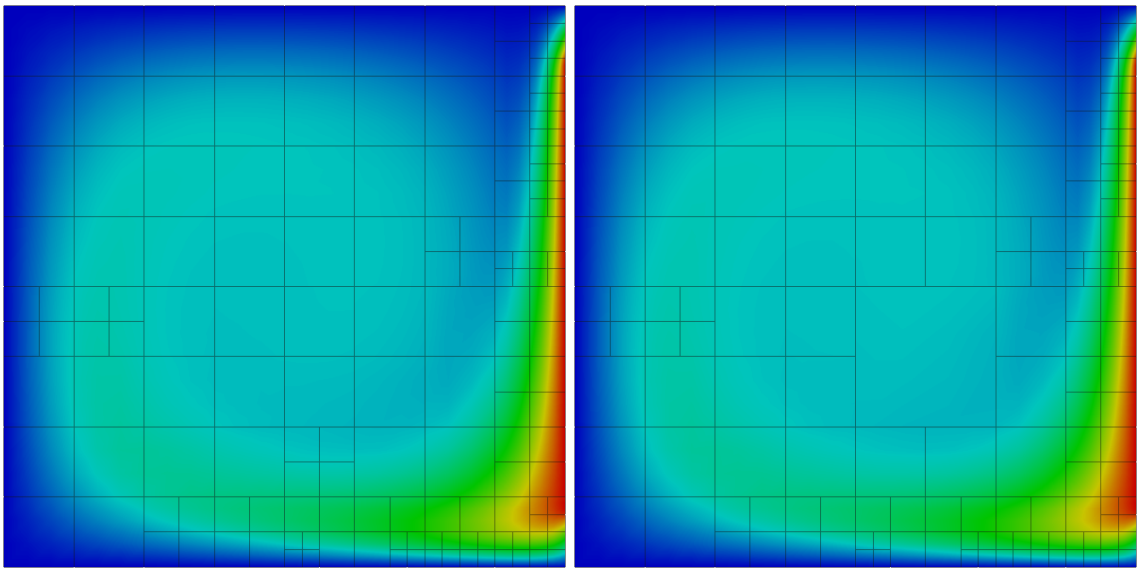
Here, $\boldsymbol{\beta} = (0.5, 1)^T/\sqrt{1.25}$, and we have a discontinuous source term such that $f = 1$ when $y \geq 2x$ and $f = -1$ when $y < 2x$. We apply boundary conditions of $\hat{t} = 0$ on the inflow and $\hat{u} = 0$ on the outflow. Contrary to the other problems discussed, the solution for this problem does not range from 0 to one. Rather, the colorbar in Figure 14 is scaled to $[-1.110, 0.889]$.

3.1.8 Hemker Problem

The Hemker problem is defined on a domain $\Omega = \{[-3, 9] \times [-3, 3]\} \setminus \{(x, y) : x^2 + y^2 < 1\}$, essentially a simplified model of flow past an infinite cylinder. We start with the initial mesh shown in Figure 16 in which we approximate the circle by 8 fourth order curvilinear polynomial segments. As we refine, the new elements are fitted to better approximate an exact circle.The boundary conditions are $\hat{t} = \boldsymbol{\beta} \cdot \mathbf{n} \cdot 1$ on the left inflow boundary, $\boldsymbol{\beta} \cdot \mathbf{n} \cdot \hat{u} - \hat{t} = \sigma_n = 0$ on the right outflow boundary, $\hat{t} = 0$ on the top and bottom edges, and $\hat{u} = 1$ on the cylinder.

3.2 Analysis of Results

The general trend we see looking at the results is that the solution quality of the conservative and conconservative formulations is nearly identical. This is ideal, as we wish to preserve the attractive features of DPG – namely that it minimizes error in the energy norm. In particular, when we examine the error convergence



(a) Nonconservative (b) Conservative

Figure 6: Double glazing problem after 5 refinements

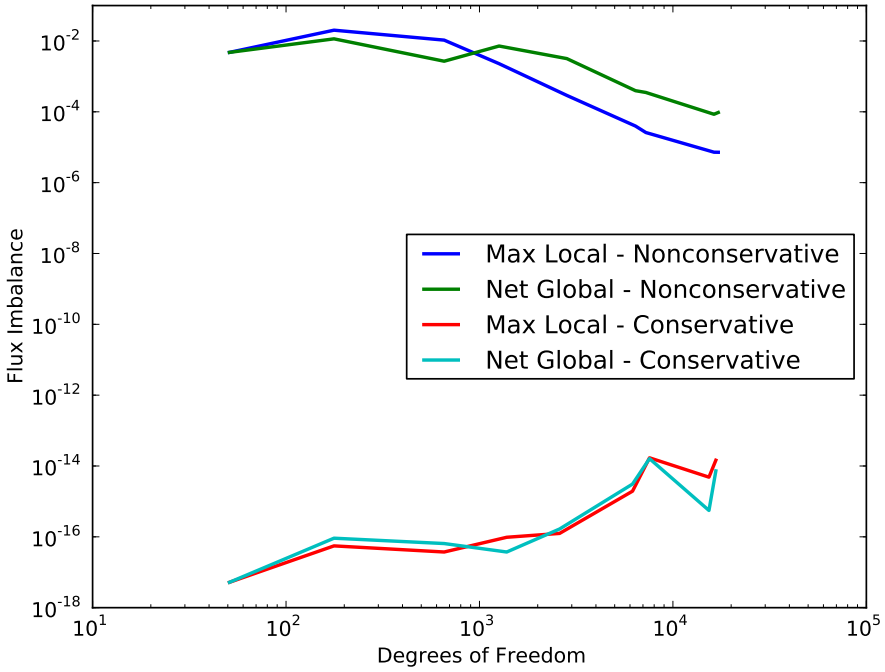
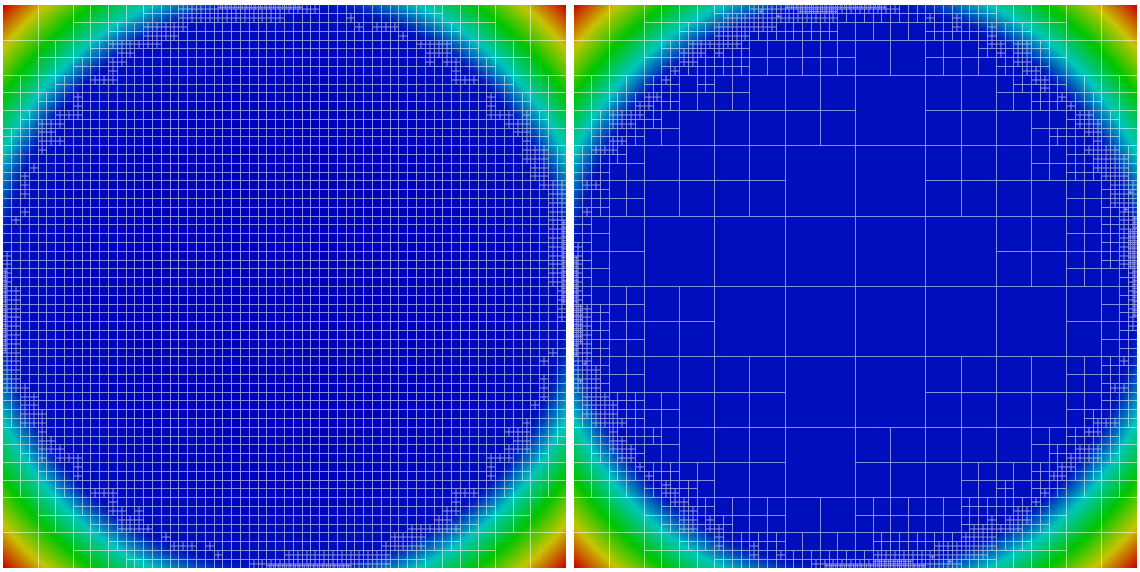


Figure 7: Flux imbalance in double glazing problem



(a) Nonconservative (b) Conservative

Figure 8: Vortex problem after 6 refinements

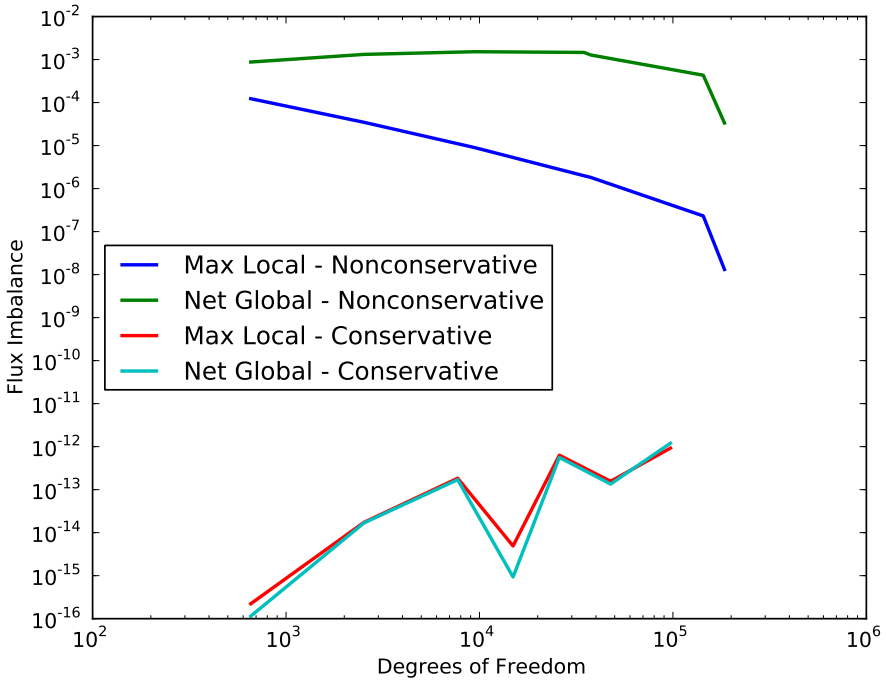
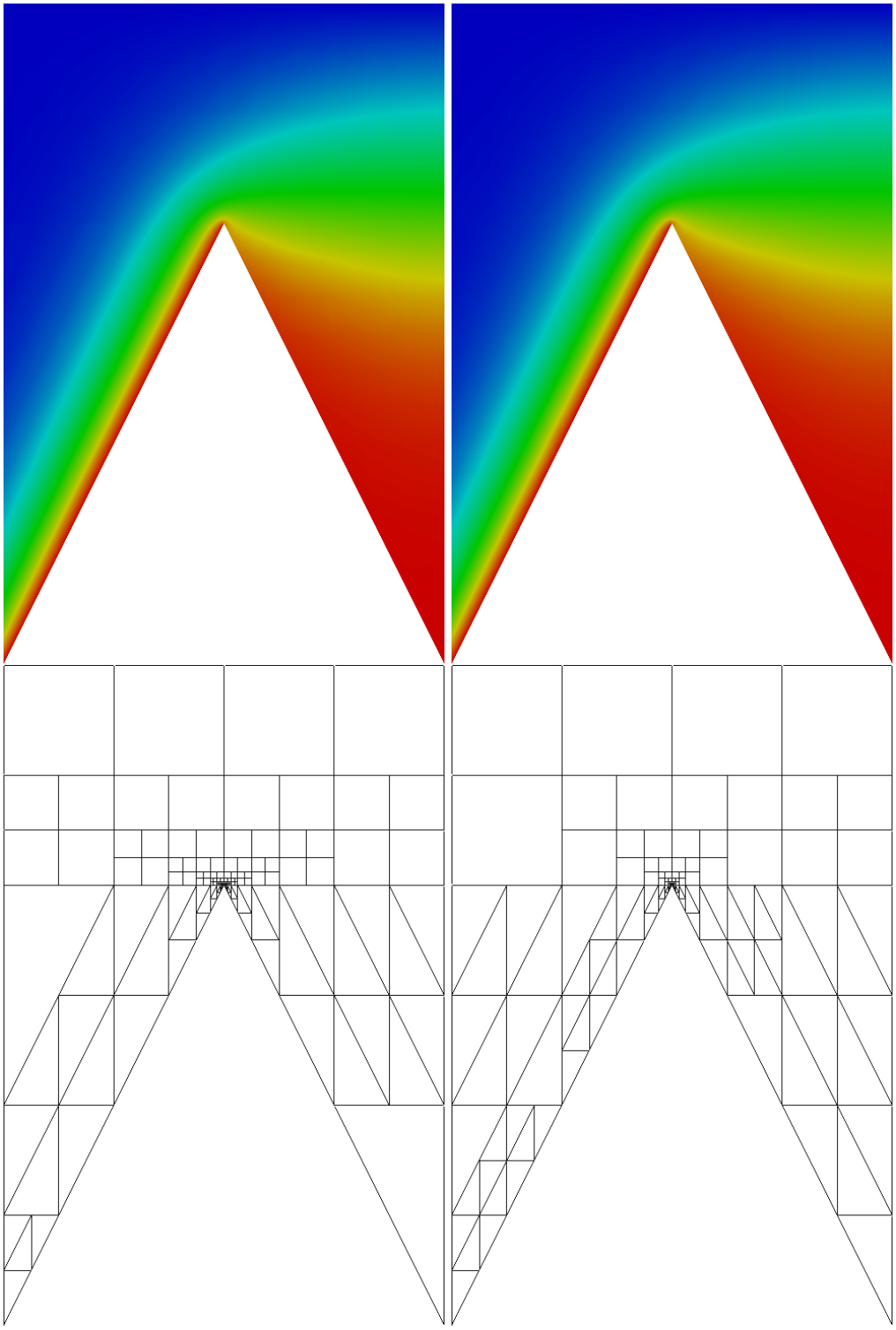


Figure 9: Flux imbalance in vortex solutions



(a) Nonconservative (b) Conservative

Figure 10: Wedge problem after 16 refinements

plots for the Erickson-Johnson problem, the energy levels for the two methods lie nearly right on top of each other. This seems to indicate at the least that enforcing local conservation doesn't hurt DPG too much.

Another measure of solution quality is to look at the magnitude of overshoots and undershoots in the solution. In most of the problems considered (barring the discontinuous source problem) the exact solution would range exactly between 0 and 1. The the resolved solutions, the divergence from these values would be negligible, but for the inner layer, in which $\epsilon = 10^{-6}$, we get overshoots and undershoots along the separation line. For the nonconservative formulation, the solution range for u was $[-0.392, 1.342]$, while the conservative solution range was $[0.389, 1.378]$. This difference is easily accounted for by the differences in refinement patterns between the two methods.

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 $\boldsymbol{\beta}$ [2] that break down for the vortex problem. Here we see the nonconservative method needlessly refine in the center of the domain where the solution is constant. The conservative scheme is more discerning about refinements and focuses refinements where solution features are changing. In general, though, both methods appear to follow very similar refinement patterns.

It shouldn't come as a surprise that the conservative and nonconservative solutions match each other so closely. The conservative formulation enforces local conservation more strictly, but if we examine the flux imbalance plots, the nonconservative 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.

References

[1] 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. submitted to J. Comp. Phys.

[2] J. Chan, N. Heuer, T Bui-Thanh, and L. Demkowicz. Robust DPG method for convection-dominated diffusion problems ii: a natural inflow condition. Technical Report 21, ICES, 2012.

[3] C. L. Chang and John 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.

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

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

[6] L. Demkowicz and J. Gopalakrishnan. An overview of the DPG method. Technical report, ICES, 2013.

[7] L. Demkowicz, J. Gopalakrishnan, and A. Niemi. A class of discontinuous Petrov-Galerkin methods. Part III: Adaptivity. Technical Report 1, ICES, 2010. submitted to App. Num. Math.

[8] L. Demkowicz and N. Heuer. Robust DPG method for convection-dominated diffusion problems. Technical Report 33, ICES, 2011.

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

[10] J. B. Perot. Discrete conservation properties of unstructured mesh schemes. *Annual Review of Fluid Mechanics*, 43:299–318, 2011.

[11] Nathan V. Roberts, Tan Bui-Thanh, and Leszek F. Demkowicz. The DPG method for the Stokes problem. Technical Report 12-22, ICES, 2012.

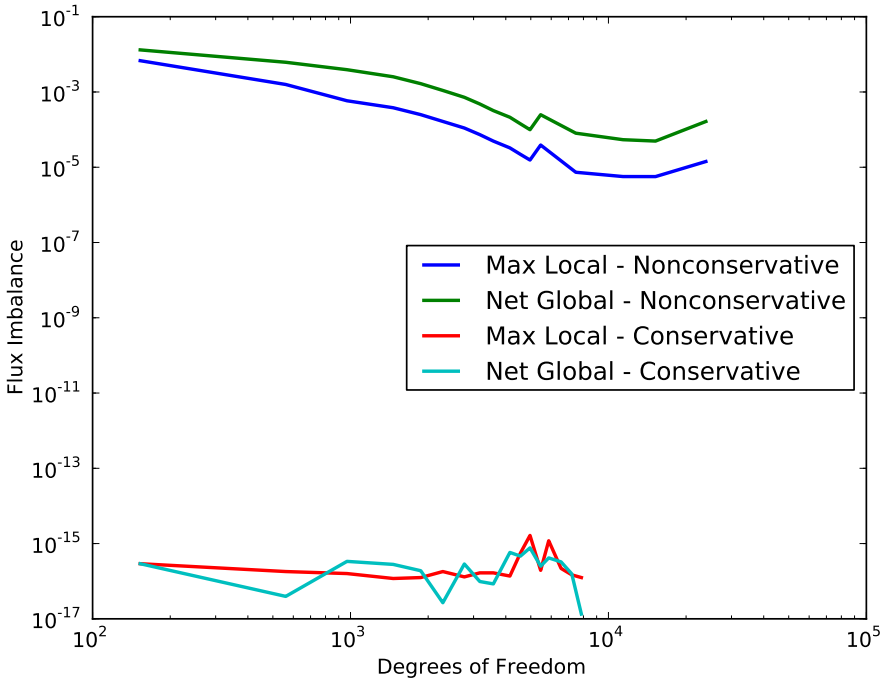
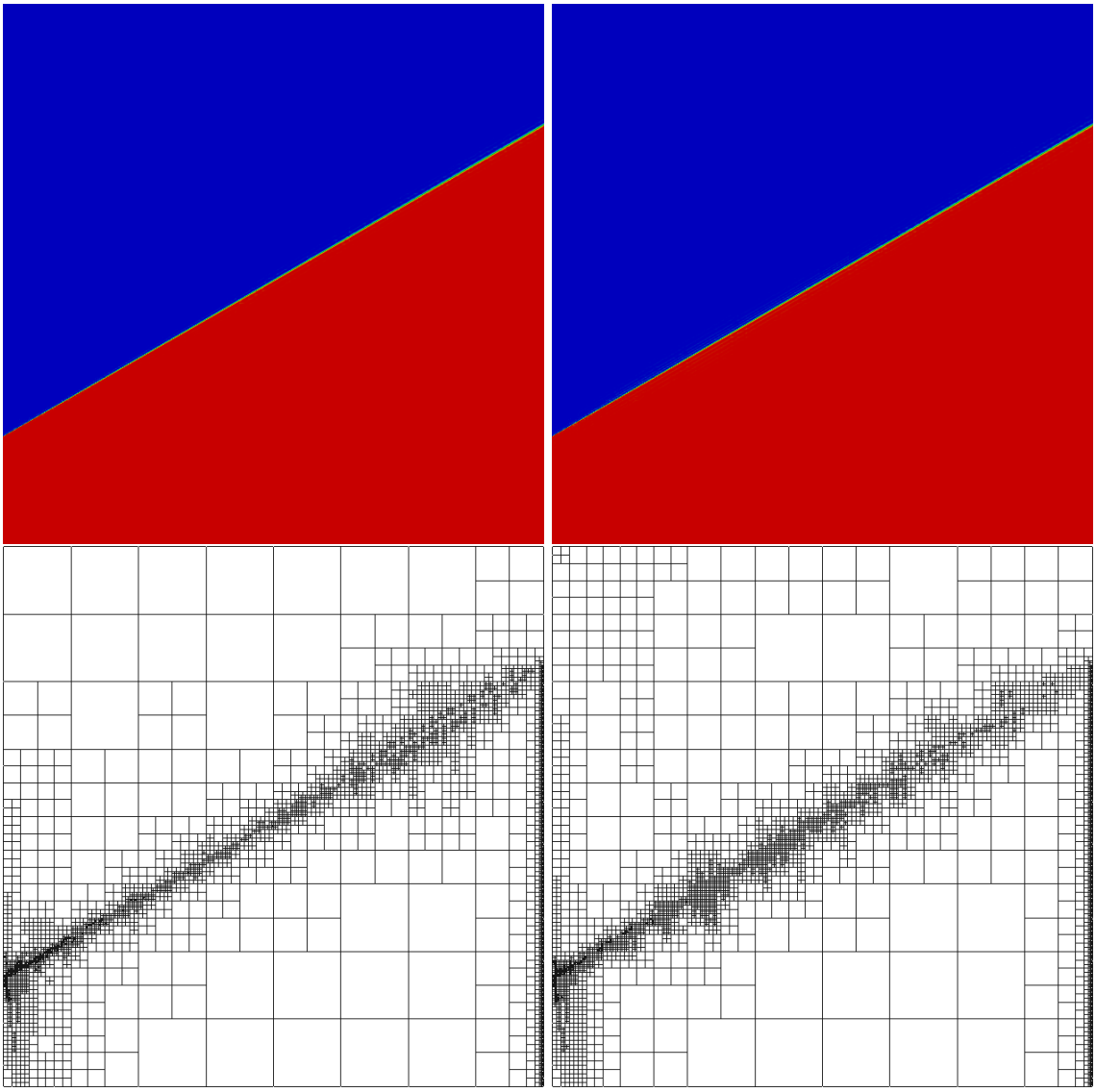


Figure 11: Flux imbalance in wedge solutions



(a) Nonconservative (b) Conservative

Figure 12: Inner layer problem after 8 refinements

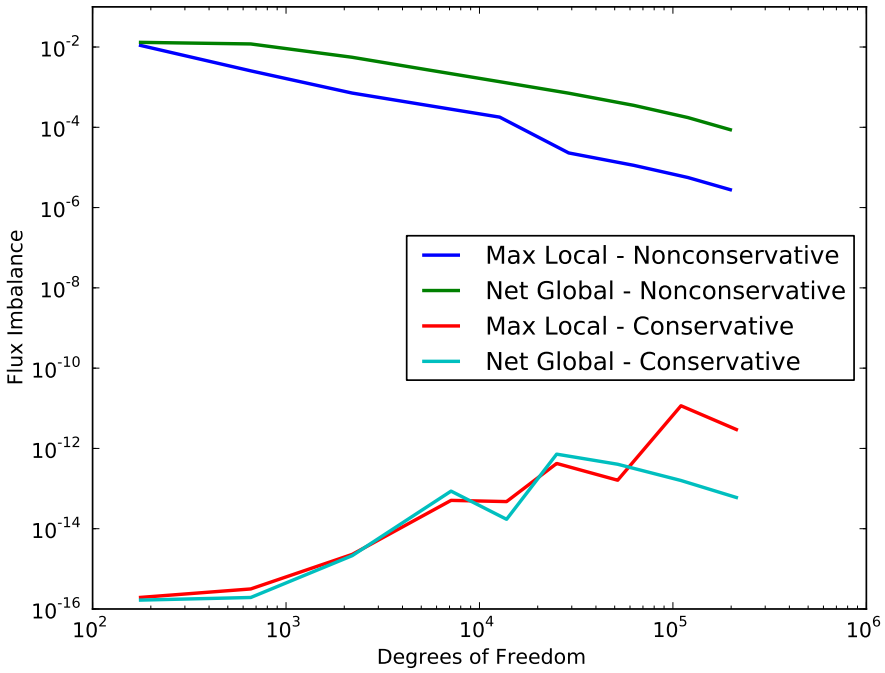
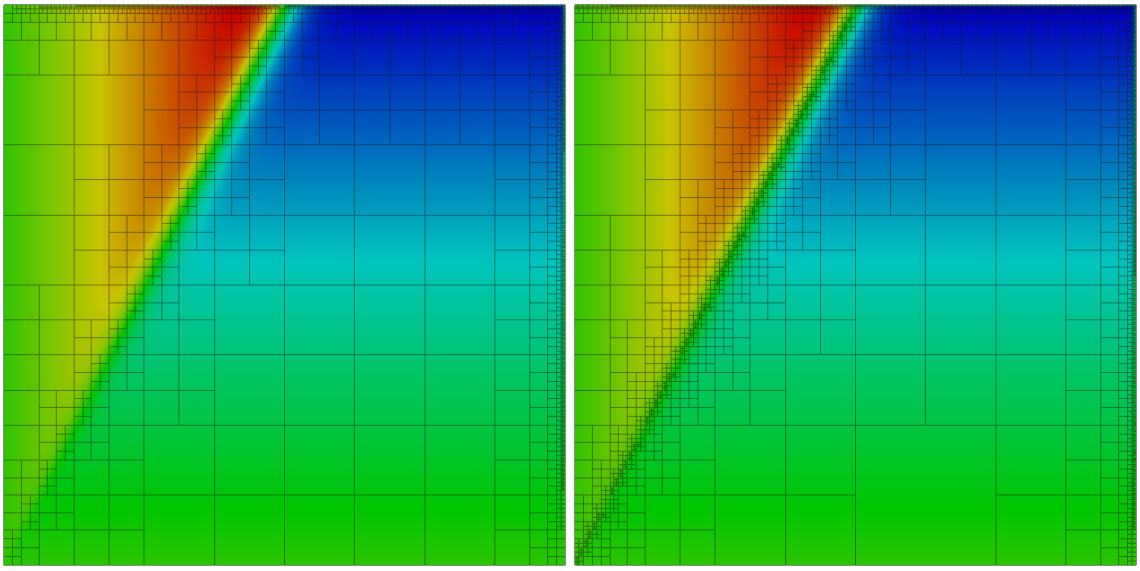


Figure 13: Flux imbalance in inner layer solutions



(a) Nonconservative (b) Conservative

Figure 14: Discontinuous source problem after 8 refinements

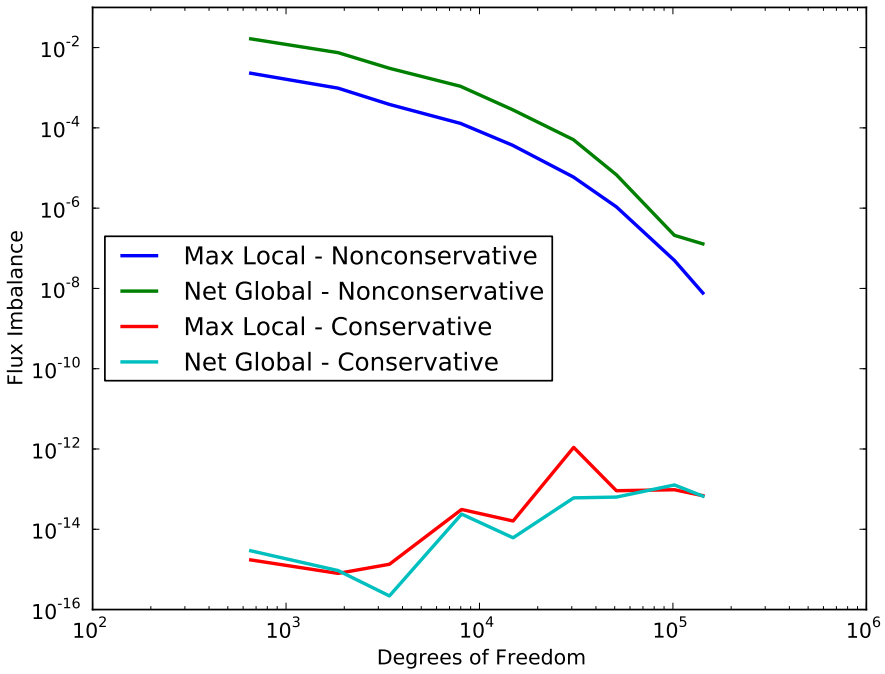


Figure 15: Flux imbalance in discontinuous source solutions

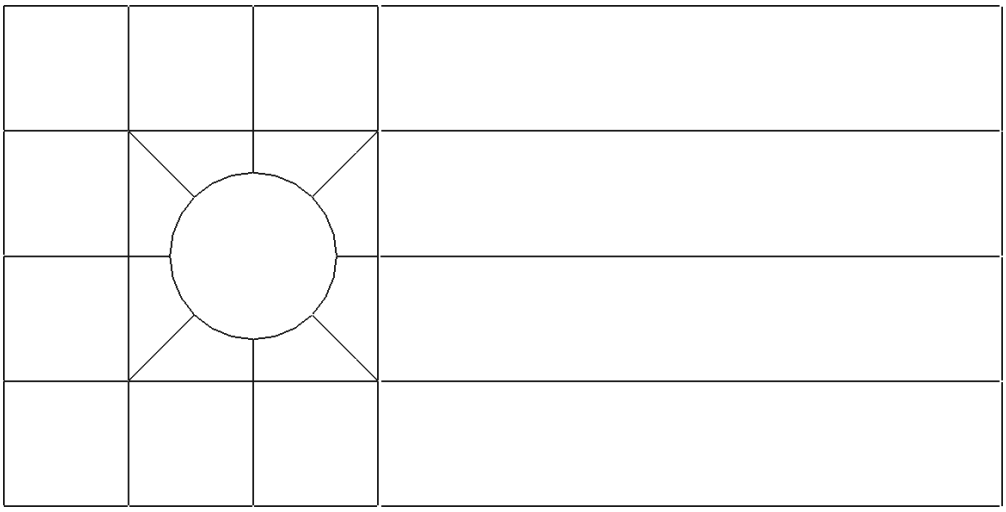
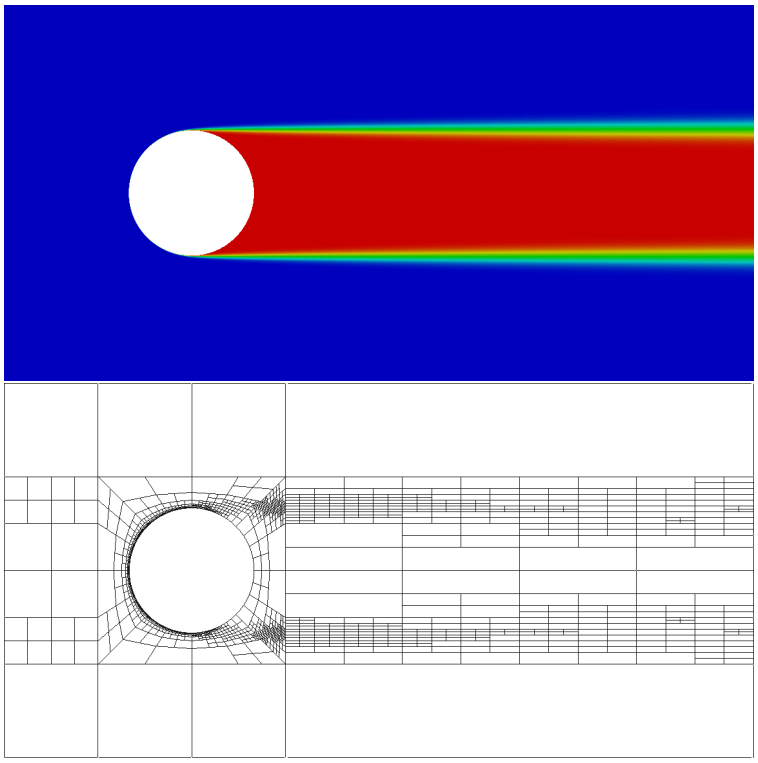
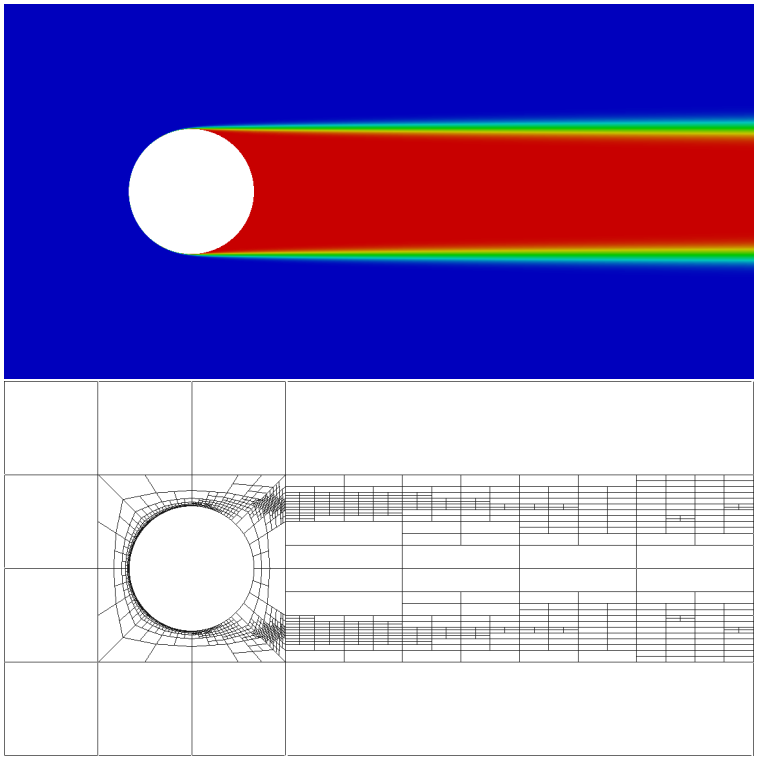


Figure 16: Initial mesh for the Hemker problem



(a) Nonconservative



(b) Conservative

Figure 17: Hemker problem after 8 refinements

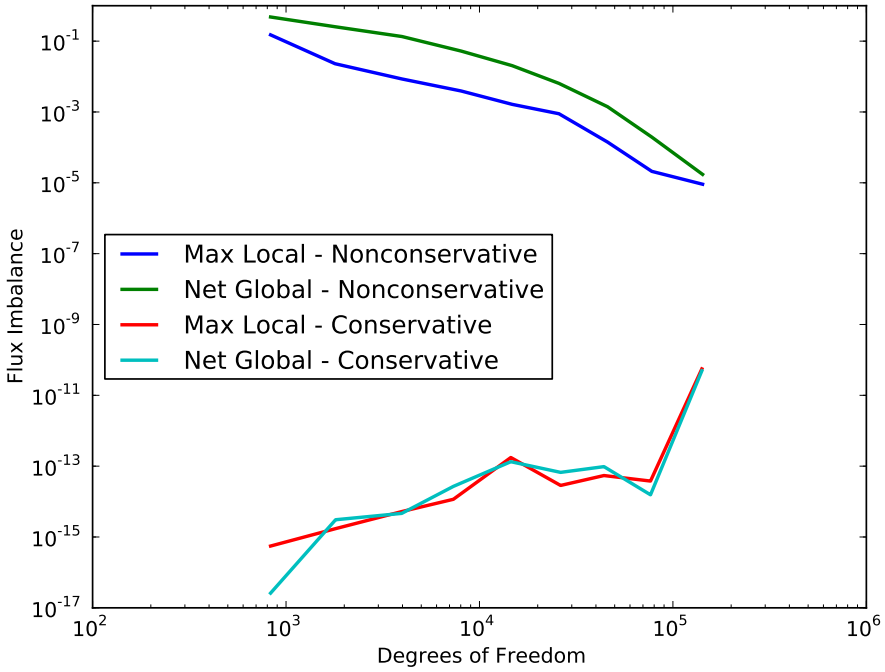


Figure 18: Flux imbalance in Hemker problem