Contents

Robust DPG stability II

Tan Bui Jesse Chan Leszek Demkowicz Norbert Heuer

1 Introduction

1.1 Singular perturbation problems and robustness

Standard Bubnov-Galerkin methods tend to perform poorly for the class of partial-differential equations known as singularly perturbed problems, where a given parameter ϵ may approach either 0 or ∞ in the context of physical problems. This poor performance is captured by the error bound

$$\|u - u_h\|_E \leq C(\epsilon) \inf_{w_h} \|u - w_h\|_U$$

The growth of C with ϵ is referred to as a loss of robustness, where the bound on the error in the finite element solution by the best approximation error increases as $\epsilon \to 0$. Intuitively, as our singular perturbation parameter changes, our finite element error is bounded more and more loosely to the best approximation error, allowing for degradation of the solution

A well-known example is the growth of spurious oscillations in the finite element solution of the 1D convection-diffusion problem $u' - \epsilon u'' = f$ (with Dirichlet boundary conditions) as $\epsilon \to 0$. Another class of examples are wave propagation problems, in which the singular perturbation parameter is the wavenumber, $k \to \infty$. The lack of robustness in this case manifests as "pollution" error, a phenomenon where the finite element solution degrades over many wavelengths (commonly manifesting as a phase error between the FE solution and the exact solution).

add bit about residual-based stabilization terms and test functions

1.2 Optimal test functions

The idea of optimal test functions was introduced by Demkowicz and Gopalakrishnan in [4]. Conceptually, these optimal test functions are the result of a minimum residual method applied to the operator form of a variational equation. Given Hilbert spaces U and V and a variational problem $b(u, v) = l(v)$, $\forall u \in U, v \in V$, we can identify $B : U \to V'$ and $l \in V'$

$$\left.\begin{array}{c} b(u_h, v) = \langle Bu_h, v\rangle \\ l(v) = \langle l, v\rangle \end{array}\right\} \Longleftrightarrow Bu_h = l$$

We seek the minimization of the residual in the dual norm

$$\min_{u_h \in U_h} J(u_h) = \frac{1}{2}\|Bu_h - l\|^2_{V'} = \frac{1}{2}\sup_{v \in V} \|b(u_h, v) - l(v)\|^2_V$$

Recall R_V, the Riesz operator $\langle R_V v, \delta v\rangle = (v, \delta v)_V$ identifying elements of Hilbert space with elements of the dual. As R_V and its inverse are isometries, $\|f\|_{V'} = \|R_V^{-1} f\|_V$, and

$$\min_{u_h \in U_h} J(u_h) = \frac{1}{2}\|Bu_h - l\|^2_{V'} = \frac{1}{2}\|R_V^{-1}(Bu_h - l)\|^2_V$$

First order optimality conditions require the Gateux derivative to be zero in all directions $u_h \in U_h$.

$$\begin{aligned}\left(R_V^{-1}(Bu_h - l), R_V^{-1}B\delta u_h\right)_V = 0, \quad & \forall \delta u_h \in U_h \\ \rightarrow \langle Bu_h - l, v_h\rangle = 0, \quad & v_h = R_V^{-1}B\delta u_h\end{aligned}$$

which returns a standard variational equation $b(u_h, v_h) - l(v_h) = 0$, for the specific choice of test functions $v_h = R_V^{-1}B\delta u_h$. We identify the *trial-to-test* operator $T = R_V^{-1}B$, which maps a trial function u_h to its corresponding *optimal* test function $v_h = Tu_h$. These test functions can be solved for by solving the auxiliary variational problem

$$(v_h, \delta v)_V = b(u_h, \delta v)$$

Finally, while the above variational problem is a global operation for standard continuous test spaces, the use of discontinuous test spaces reduces it to a local problem that can be solved in an element-by-element fashion. In practice, this variational problem can rarely be solved exactly, but instead is solved approximately using the standard Bubnov-Galerkin method and an "enriched" subspace of V. We assume the corresponding error in approximation of the optimal test functions is negligible for the scope of this paper. Further work concerning the effect of approximation error in the computation of optimal test functions can be found in [6].

Under the standard conditions for well-posedness of the continuous variational problem, the discrete DPG method delivers the best approximation error in the "energy norm" $\|u\|_E = \sup \frac{b(u,v)}{\|v\|_V}$. Additionally, the actual energy error $\|u - u_h\|_E$ is computable through $\|e_h\|_V$, where

$$(e_h, \delta v)_V = b(u - u_h, v) = b(u_h, v) - l(v)$$

This is simply a consequence of the least-squares nature of DPG; the energy error is simply the measure of the residual in the proper norm.

1.3 Variational form

The DPG (discontinuous Petrov-Galerkin) method is the combination of the concept of computable optimal test functions with the so-called "ultra-weak formulation", where the regularity requirement on each solution variable is relaxed through integration by parts. Moreover, to maintain conformity while seeking an L^2 setting on interior "field" variables, boundary terms are identified as additional new unknowns. The result is a hybridized DG method with locally supported optimal test functions. For a given operator equation $Au = f$, the ultra-weak formulation is

$$b\left((\widehat{u}, u), v\right) - l(v) = \langle \widehat{u}, v\rangle - (u, A^*v) - (f, v)$$

where A^* is the formal adjoint resulting from integration by parts.

1.4 Optimal and quasi-optimal test norms

Up to now, we have neglected discussion of the proper choice of inner product/norm on the space V. This choice is important, as the choice of norm on V determines the norm on U in which DPG is optimal. Whether a particular choice of norm on V norm is "good" or not is problem-dependent; however, there is an important canonical choice of test norm for any given variational form.

Under the assumption of injectivity of the adjoint of our variational operator B, we can define the *optimal test norm*

$$\|v\|_{\text{opt}} = \sup_{u \in U} \frac{b(u, v)}{\|u\|_U}$$

If the solution is in U, then choosing $\|v\| = \|v\|_{\text{opt}}$ implies $\|u\|_E = \|u\|_U$. The proof is simple; we have by definition that $b(u, v) \leq \|u\|_U \|v\|_V$. The reverse $b(u, v) \leq \|u\|_U \|v\|_V$ is proved by

$$\inf_{u \in U} \frac{\|u\|_E}{\|u\|_U} = \inf_{u \in U} \sup_{v \in V} \frac{b(u, v)}{\|v\|_V \|u\|_U} = \inf_{v \in V} \sup_{u \in U} \frac{b(u, v)}{\|v\|_V \|u\|_U} = \inf_{v \in V} \frac{\|v\|_V}{\|v\|_V} = 1$$

where we're able to switch the order of inf and sup due to the injectivity of B (see [2]). This is the ideal test norm; given any particular norm $\|u\|_U$, DPG under the optimal test norm will return the best finite element approximation in that norm.

The norm $\|v\|_{\text{opt}}$ is non-localizable, however - the inversion of the inner product corresponding to this norm is global, and solving for optimal test functions under this optimal norm will result in global problems as well, due to the coupling given by the boundary terms in our bilinear form. In practice, an approximation is made by removing such boundary terms and replacing them with scaled L^2 norms of the field variables, producing so-called *quasi-optimal* test norms. The DPG method is currently being analyzed for both acoustic and elastic waves in context of the Helmholtz equation and equations of linear elasticity. Numerically, DPG appears to provide a "pollution-free" method without phase error for these problems under the "quasi-optimal" test norm with a specific scaling of the regularizing L^2 terms [10]. Similar results have also been obtained with the quasi-optimal test norm in the context of the elasticity [1] and linearized Stokes equations [9].

2 Model problem and robustness

We consider the model convection-diffusion problem on domain Ω with boundary Γ

$$\nabla \cdot (\beta u) - \epsilon \Delta v = f \tag{1}$$

in first order form

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

with the corresponding variational form

$$b\left((u, \sigma, \widehat{u}, \widehat{\sigma}_n), (\tau, v)\right) = (u, \nabla \cdot \tau - \beta \cdot \nabla v) + \left(\sigma, \epsilon^{-1}\tau + \nabla v\right) - \langle [\tau_n], \widehat{u} \rangle + \langle \widehat{f}_n, [v] \rangle$$

where $[\tau_n]$ and $[v]$ denote the jumps in τ_n and v, defined as

$$\begin{aligned} [\tau_n] &= \tau^+ \cdot n^+ + \tau^- \cdot n^- \\ [v] &= v^+ \text{sgn}(n^+) + v^- \text{sgn}(n^-) \end{aligned}$$

Additionally, the inner products are taken element-wise over the finite element mesh Ω_h, and the duality pairings are taken on Γ_h, the mesh "skeleton", or union of edges of elements K in Ω_h. The divergence and gradient operators are likewise understood to act element-wise.

Define the interior skeleton $\Gamma_h^0 = \Gamma_h \setminus \Gamma$. The functional setting is now well understood as well (see [3] for details) - $u, \sigma \in L^2(\Omega)$, $v \in H^1(\Omega_h)$, $\tau \in H(\text{div}, \Omega_h)$, where $H^1(\Omega_h)$ and $H(\text{div}, \Omega_h)$ are element-wise "broken" Sobolev spaces. By duality, $\widehat{u}$ lives in $H^{1/2}(\Gamma_h)$, the trace space of $H^1(\Omega_h)$, while $\widehat{f}_n$ comes from $H^{-1/2}(\Gamma_h)$, the normal trace space of $H(\text{div}, \Omega_h)$.

We split the boundary Γ into three portions

$$\begin{aligned} \Gamma_- &\coloneqq \{x \in \Gamma; \beta_n(x) < 0\} \quad \text{(inflow)} \\ \Gamma_+ &\coloneqq \{x \in \Gamma; \beta_n(x) > 0\} \quad \text{(outflow)} \\ \Gamma_0 &\coloneqq \{x \in \Gamma; \beta_n(x) = 0\} \end{aligned}$$

We adopt a new inflow boundary condition, given by Hesthaven et al in [7], where we set

$$\beta_n u - \sigma_n = f_n = u_0$$

on Γ_-, and continue with the wall boundary condition $u = 0$ on Γ_+. For our model problem, as for most problems of interest in CFD, we expect ∇u to be small near the inflow, and that the solutions

to (1) using $u = u_0$ and $\beta_n u - \sigma_n = f_n = u_0$ should converge to each other for sufficiently small ϵ. Under these new boundary conditions, our variational problem $b(u, v) = l(v)$ is now defined by

$$b\left((u, \sigma, \widehat{u}, \widehat{\sigma}_n), (\tau, v)\right) = (u, \nabla \cdot \tau - \beta \cdot \nabla v) + \left(\sigma, \epsilon^{-1}\tau + \nabla v\right) - \langle [\tau_n], \widehat{u}\rangle_{\Gamma_- \cup \Gamma_h^0} + \langle \widehat{f_n}, [v]\rangle_{\Gamma_+ \cup \Gamma_h^0}$$
$$l\left((\tau, v)\right) = (f, v) + \langle [\tau_n], \widehat{u}\rangle_{\Gamma_+} - \langle \widehat{f_n}, [v]\rangle_{\Gamma_-}$$

2.1 Norms on U

Given the boundary conditions on $\widehat{u}$ and $\widehat{f_n}$, we can now also define the trace norms as (canonical) minimum energy extension norms

$$\|\widehat{u}\| = \inf_{w \in H^1(\Omega),\, w|_{\Gamma_+} = 0,\, w|_{\Gamma_h \setminus \Gamma_+} = \widehat{u}} \|w\|_{H^1(\Omega)}$$
$$\|\widehat{f_n}\| = \inf_{q \in H(\mathrm{div}, \Omega),\, q \cdot n|_{\Gamma_-} = 0,\, q \cdot n|_{\Gamma_h} = \widehat{f_n}} \|q\|_{H(\mathrm{div}, \Omega)}$$

Additionally, throughout the rest of this work, we denote norms on u and σ as

$$\|u\| = \|u\|_{L^2}$$
$$\|\sigma\| = \|\sigma\|_{L^2}$$

We will construct norms on U as combinations of norms on each variable

$$\left\| \left(u, \sigma, \widehat{u}, \widehat{f_n}\right) \right\|_U = C_1\|u\| + C_2\|\sigma\| + C_3\|\widehat{u}\| + C_4\|\widehat{f_n}\|.$$

2.2 Norms on V

As $\tau \in H(\mathrm{div}, \Omega_h)$ and $v \in H^1(\Omega_h)$, we will construct norms on v and τ using some combination of terms which are equivalent to the canonical broken $H^1(\Omega_h) \times H(\mathrm{div}, \Omega_h)$ norm

$$\|\,(\tau, v)\,\|_{H^1(\Omega_h) \times H(\mathrm{div}, \Omega_h)} = \|v\|^2_{L^2(\Omega_h)} + \|\nabla v\|^2_{L^2(\Omega_h)} + \|\tau\|^2_{L^2(\Omega_h)} + \|\nabla \cdot \tau\|^2_{L^2(\Omega_h)}$$

The exact norms that we will place on V will be determined later.

The boundary norms on Γ_h for v and τ are defined by duality from the bilinear form

$$\|[\tau \cdot n]\| = \|[\tau \cdot n]\|_{\Gamma_h \setminus \Gamma_+} \coloneqq \sup_{w \in H^1(\Omega),\, w|_{\Gamma_+} = 0} \frac{\langle [\tau \cdot n], w\rangle}{\|w\|_{H^1(\Omega)}}$$
$$\|[v]\| = \|[v]\|_{\Gamma_h^0 \cup \Gamma_+} \coloneqq \sup_{\eta \in H(\mathrm{div}; \Omega),\, \eta \cdot n|_{\Gamma_- \cup \Gamma_0} = 0} \frac{\langle v, \eta \cdot n\rangle}{\|\eta\|_{H(\mathrm{div}, \Omega)}}$$

2.3 Approximability of the quasi-optimal test norm

For convection-diffusion, the quasi-optimal norm is

$$\|\,(\tau, v)\,\|^2_V = \|\nabla \cdot \tau - \beta \cdot \nabla v\|^2_{L^2} + \|\epsilon^{-1}\tau + \nabla v\|^2_{L^2} + C_1\|v\|_{L^2} + C_2\|\tau\|_{L^2}$$

for some choice of constants C_1 and C_2. We note that the optimal test norm will automatically guarantee the robust bound $\|\,(\tau, v)\,\|_V \leq \|u\|_{L^2}$. However, use of this norm for the convection-diffusion problem is difficult - the variational problem for optimal test functions using the quasi-optimal test norm is equivalent to a reaction-diffusion system [8]. Transforming the problem to the reference element reveals that will induce strong boundary layers of width ϵ/h^2 (in comparison, in Helmholtz and wave propagation problems, the mesh size tends to be on the order of the wavelength/singular perturbation parameter, resulting in smooth optimal test functions that are

much easier to approximate over the reference element). As the relevant range of ϵ for physical problems is $1e-7$, solving on partially under-resolved meshes is unavoidable.

Resolving such boundary layers for an underresolved mesh has been investigated numerically using specially designed (Shishkin) subgrid meshes by Niemi, Collier, and Calo in [8]. However, even with Shishkin meshes, the approximation of optimal test functions under the quasi-optimal norm is difficult and far more expensive than approximation test functions using a simple p-enriched space for V. We approach this from a different perspective, looking instead for robust test norms which do not induce boundary layers in the approximation of optimal test functions, and do so by formulating a general approach to analyzing the behavior of DPG under a given test norm.

The estimates and proofs in this paper are largely based off of techniques and prior work done in [5] and [3]; for further details, we encourage readers to consult these two publications as well. The primary contributions in this paper are stronger stability estimates for a new inflow boundary condition and a slightly different test norm.

2.4 Adjoint stability

introduce the three lemmas An important relationship is the duality of the energy norm and test norm; for any choice of energy norm, the test norm $\|v\|_V$ inducing it can be recovered via duality

$$\|v\|_V = \sup_{u \in U \setminus \{0\}} \frac{b(u,v)}{\|u\|_E} \tag{2}$$

The proof for this is identical to proving $\|u\|_E = \|u\|_U$ for the choice of the optimal test norm.

As a consequence of (8), assume an energy norm $\|u\|_{E,1}$ is induced by a given test norm $\|v\|_{V,1}$. To bound $\|\cdot\|_E$ robustly from above or below by a given norm $\|u\|_{U,2}$ on U now only requires the robust bound in the opposite direction of $\|v\|_{V,1}$ by $\|v\|_{V,2}$.

We will use the notation $\lesssim$ to denote an ϵ-independent bound.

2.4.1 Induced adjoint boundary conditions

Demkowicz and Heuer proved in [5] that for Dirichlet boundary conditions, robustness as $\epsilon \to 0$ is achieved by the test norm

$$\|(\tau, v)\|^2_{V,w} = \|v\| + \epsilon\|\nabla v\| + \|\beta \cdot \nabla v\|_{w+\epsilon} + \|\nabla \cdot \tau\|_{w+\epsilon} + \frac{1}{\epsilon}\|\tau\|_{w+\epsilon}$$

where $\|\cdot\|_{w+\epsilon}$ is a weighted L^2 norm, where the weight $w \in (0,1)$ is required to vanish on Γ_-.

Explain how energy norm convergence in $\|u\| + \epsilon\|u'\|$ returns bad results, and the need for an $O(1)$ term in the test norm.

The need for the weight is intuitively explained by the induced adjoint problem. The bilinear form for the Dirichlet case is

$$b((u, \sigma, \widehat{u}, \widehat{\sigma}_n), (\tau, v)) = (u, \nabla \cdot \tau - \beta \cdot \nabla v) + \left(\sigma, \epsilon^{-1}\tau + \nabla v\right) + \langle \widehat{f}_n, [v] \rangle$$

We note that

$$\begin{aligned} b((u, \sigma, \widehat{u}, \widehat{\sigma}_n), (\tau, v)) &= (u, \nabla \cdot \tau - \beta \cdot \nabla v) + \left(\sigma, \epsilon^{-1}\tau + \nabla v\right) - \langle [\tau_n], \widehat{u} \rangle_{\Gamma_- \cup \Gamma_h^0} + \langle \widehat{f}_n, [v] \rangle_{\Gamma_+ \cup \Gamma_h^0} \\ &= \|u\|^2_{L^2} \end{aligned}$$

for the specific continuous test functions τ and v satisfying

$$\begin{aligned} \nabla \cdot \tau - \beta \cdot \nabla v &= u \\ \frac{1}{\epsilon}\tau + \nabla v &= 0 \end{aligned}$$

with boundary conditions

$$\tau_n = 0, \quad x \in \Gamma_- \cup \Gamma_0 \tag{3}$$
$$v = 0, \quad x \in \Gamma_+ \tag{4}$$

Choosing a continuous conforming test function v removes the contribution of the term $\langle \widehat{f_n}, [v] \rangle$ on the mesh skeleton, while the boundary conditions remove the contribution of boundary terms from the bilinear form. Then, for this specific choice of (τ, v),

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

If we can show $\|(\tau, v)\|_V \lesssim \|u\|_{L^2}$ for any choice of $u \in L^2$, dividing through by $\|u\|_{L^2}$ gives

$$\|u\|_{L^2} \lesssim \|u\|_E$$

which results in in robust control of the L^2 error by the error in the energy norm, in which DPG is optimal.

However, we are unable to prove this in the case of Dirichlet boundary conditions. Intuitively, the adjoint problem is similar to the primal problem with the direction of inflow reversed, such that the inflow becomes the outflow, and outflow inflow. The need for the weight arises due to the presence of the Dirichlet boundary condition on v near the inflow; this induces strong boundary layers at the inflow such that $\nabla v \approx O(\epsilon^{-1})$. It may be of interest to note that, for sufficiently small ϵ, these issues manifest themselves in numerical experiments as additional refinements near the inflow. The presence of the new boundary condition relaxes the outflow boundary condition for the adjoint/dual problem, resulting in stronger stability for the primal problem.

We improve upon the result of Demkowicz and Heuer not by modifying the test norm, but by adopting a new inflow boundary condition, which changes the stability properties of the corresponding adjoint equation. Under these new boundary conditions, we can prove robust control over $\|u\|_{L^2}$ using the unweighted version of the test norm in [5]

$$\|(\tau, v)\|_V^2 = \|v\|^2 + \epsilon\|\nabla v\|^2 + \|\beta \cdot \nabla v\|^2 + \|\nabla \cdot \tau\|^2 + \frac{1}{\epsilon}\|\tau\|^2$$

This test norm is implied by certain norm "building blocks"; stability analysis of the adjoint equation allows us to bound from above norms and seminorms on v and τ by $\|f\|_{L^2}$ and $\|g\|_{L^2}$ in ways that do not grow as $\epsilon \to 0$, and we assemble the test norm as combinations of such norms and seminorms on v and τ.

2.5 Equivalence of energy and U norms

For this analysis, it will be necessary to assume certain conditions on β. For each proof, we require $\beta \in C^2(\bar{\Omega})$ and $\beta, \nabla \cdot \beta = O(1)$. Additionally, we will assume some or all of the following assumptions

$$\nabla \times \beta = 0, \quad 0 < C \leq |\beta|^2 + \frac{1}{2}\nabla \cdot \beta, \quad C = O(1) \tag{5}$$
$$\nabla\beta + \nabla\beta^T - \nabla \cdot \beta I = O(1) \tag{6}$$
$$\nabla \cdot \beta = 0 \tag{7}$$

The primary result of this paper is that, if conditions 5,6, and 7 are all satisfied, the DPG method provides robust control over the L^2 error in u and σ

$$\left\| \left(u, \sigma, \widehat{u}, \widehat{f_n}\right) - \left(u_h, \sigma_h, \widehat{u}_h, \widehat{f}_{n,h}\right) \right\|_{U_1} \lesssim \inf_{w \in U_h} \left\| \left(u, \sigma, \widehat{u}, \widehat{f_n}\right) - w_h \right\|_E$$

using the test norm

$$\| (\tau, v) \|_V^2 = \|v\|^2 + \epsilon\|\nabla v\|^2 + \|\beta \cdot \nabla v\|^2 + \|\nabla \cdot \tau\|^2 + \frac{1}{\epsilon}\|\tau\|^2$$

where $\|\cdot\|_{U_1}$ is defined as

$$\left\|\left(u, \sigma, \widehat{u}, \widehat{f}_n\right)\right\|_{U_1}^2 := \|u\|^2 + \|\sigma\|^2 + \epsilon^2\|\widehat{u}\|^2 + \epsilon\|\widehat{f}_n\|^2.$$

A second, weaker result is an upper bound on the energy norm

$$\left\|\left(u, \sigma, \widehat{u}, \widehat{f}_n\right) - \left(u_h, \sigma_h, \widehat{u}_h, \widehat{f}_{n,h}\right)\right\|_E \lesssim \inf_{w \in U_h} \left\|\left(u, \sigma, \widehat{u}, \widehat{f}_n\right) - w_h\right\|_{U_2}$$

where $\|\cdot\|_{U_2}$ is defined as

$$\left\|\left(u, \sigma, \widehat{u}, \widehat{f}_n\right)\right\|_{U_2}^2 := \|u\|^2 + \frac{1}{\sqrt{\epsilon}}\left(\|\sigma\|^2 + \|\widehat{u}\|^2 + \|\widehat{f}_n\|^2\right).$$

The upper bound on the energy norm appears to depend on $\epsilon^{-1/2}$, implying that for very small ϵ, the L^2 error may be much smaller than the energy error. However, numerical experiments demonstrate much better results, and suggest that, at least for our examples, $\|u - u_h\|_E \sim \|u - u_h\|_U$, implying that the induced energy norm is equivalent to the canonical norms we have chosen on the group variable $\left(u, \sigma, \widehat{u}, \widehat{f}_n\right)$.

The proof proceeds in two steps; the bound from below is proven by providing stability estimates for the adjoint problem, while the bound from above of the energy norm is proven separately.

2.5.1 Bound from below

introduce the three lemmas With these three lemmas, we can prove the robust bound of both field variables and fluxes by the energy error. In doing so, we take advantage of the fact that, for any choice of energy norm, the test norm $\|v\|_V$ inducing it can be recovered via duality

$$\|v\|_V = \sup_{u \in U \setminus \{0\}} \frac{b(u, v)}{\|u\|_E} \tag{8}$$

The proof for this is identical to proving $\|u\|_E = \|u\|_U$ for the choice of the optimal test norm.

As a consequence of (8), assume an energy norm $\|u\|_{E,1}$ is induced by a given test norm $\|v\|_{V,1}$. To bound $\|\cdot\|_E$ robustly from above or below by a given norm $\|u\|_{U,2}$ on U now only requires the robust bound in the opposite direction of $\|v\|_{V,1}$ by $\|v\|_{V,2}$.

Lemma 1. *(Bound from below) If β satisfies (5) and (6), then for $\left(u, \sigma, \widehat{u}, \widehat{f}_n\right) \in U$,*

$$\|u\| + \|\sigma\| + \epsilon\|\widehat{u}\| + \sqrt{\epsilon}\|\widehat{u}\| \lesssim \left\|\left(u, \sigma, \widehat{u}, \widehat{f}_n\right)\right\|_E$$

Proof. By (8), proving the above result is equivalent to showing

$$\begin{aligned}\| (v, \tau) \|_V &\lesssim \frac{b\left(\left(u, \sigma, \widehat{u}, \widehat{f}_n\right), (\tau, v)\right)}{\|u\| + \|\sigma\| + \epsilon\|\widehat{u}\| + \sqrt{\epsilon}\|\widehat{u}\|} \\ &\simeq \|g\| + \|f\| + \epsilon^{-1} \sup_{\widehat{u} \neq 0, \widehat{u}|_{\Gamma_+} = 0} \frac{\langle [\tau \cdot n], \widehat{u} \rangle}{\|\widehat{u}\|} + \epsilon^{-1/2} \sup_{\widehat{f}_n \neq 0, \widehat{f}_n|_{\Gamma_-} = 0} \frac{\langle \widehat{f}_n, [v] \rangle}{\|\widehat{f}\|},\end{aligned}$$

which, by definition of the boundary norms, is

$$\| (v, \tau) \|_V \lesssim \|g\| + \|f\| + \epsilon^{-1}\|[\tau \cdot n]\| + \epsilon^{-1/2}\|[v]\|$$

We will bound $\| (v,\tau) \|_V$ for all (v,τ) satisfying (9) by first decomposing $(v,\tau) = (v_0,\tau_0)+(v_1,\tau_1)+(v_2,\tau_2)$, where (v_1,τ_1) satisfies our adjoint equation (9) with boundary conditions (10) and (11) and

$$\begin{aligned} f &= 0 \\ g &= \nabla\cdot\tau - \beta\cdot\nabla v, \end{aligned}$$

and (v_2,τ_2) satisfies (9) with boundary conditions (10) and (11) and

$$\begin{aligned} g &= 0 \\ f &= \epsilon^{-1}\tau + \nabla v. \end{aligned}$$

We note that, if v_2 satisfies $n\cdot\nabla v_2 = n\cdot f$, by definition of f and v_2,

$$\tau_2\cdot n = \epsilon n\cdot f - \epsilon n\cdot\nabla v_2 = 0, \quad x\in\Gamma_-\cup\Gamma_0$$

By construction, (v_0,τ_0) must satisfy (9) with $f=g=0$, with jumps

$$\begin{aligned} [v_0] &= [v], \quad x\in\Gamma_h^0 \\ [\tau_0\cdot n] &= [\tau\cdot n], \quad x\in\Gamma_h^0. \end{aligned}$$

By $(v_0,\tau_0) = (v,\tau)-(v_1,\tau_1)-(v_2,\tau_2)$, (v_0,τ_0) must satisfy boundary conditions

$$\begin{aligned} v_0 &= v, \quad x\in\Gamma_+ \\ \tau_0\cdot n &= \tau\cdot n, \quad x\in\Gamma_-\cup\Gamma_0. \end{aligned}$$

Robustly bounding $\| (v,\tau) \|_V$ from above reduces to robustly bounding each component

$$\| (v_0,\tau_0) \|_V, \| (v_1,\tau_1) \|_V, \| (v_2,\tau_2) \|_V \lesssim \|g\| + \|f\| + \epsilon^{-1}\|[\tau\cdot n]\| + \epsilon^{-1/2}\|[v]\|$$

- **Bound on** $\| (v_0,\tau_0) \|_V$

  Lemma 5 gives control over $\epsilon\|\nabla v_0\| + \frac{1}{\epsilon}\|\tau_0\|$ through

  $$\|\nabla v_0\| = \frac{1}{\epsilon}\|\tau_0\| \lesssim \frac{1}{\epsilon}\|[\tau_0\cdot n]\|_{\Gamma_h\setminus\Gamma_+} + \frac{1}{\sqrt{\epsilon}}\|[v_0]\|_{\Gamma_h^0\cup\Gamma_+} = \frac{1}{\epsilon}\|[\tau\cdot n]\|_{\Gamma_h\setminus\Gamma_+} + \frac{1}{\sqrt{\epsilon}}\|[v]\|_{\Gamma_h^0\cup\Gamma_+}$$

  Lemma 4.2 of [3] gives us the Poincare inequality for discontinuous functions

  $$\|v_0\| \lesssim \|\nabla v_0\| + \|[v]\|$$

  Since $g=0$, $\|\nabla\cdot\tau_0\| = \|\beta\cdot\nabla v_0\| \lesssim \|\nabla v_0\|$, which we have bounded previously.

- **Bound on** $\| (v_1,\tau_1) \|_V$

  With $f=0$, Lemma 4 provides the bound

  $$\|\beta\cdot\nabla v_1\| \lesssim \|g\|$$

  Noting that $\nabla\cdot\tau_1 = g+\beta\cdot\nabla v_1$ gives $\|\nabla\cdot\tau_1\| \lesssim \|g\|$ as well. Lemma 4 gives

  $$\epsilon\|\nabla v_1\|^2 + \|v_1\|^2 \lesssim \|g\|^2$$

  and noting that $\epsilon^{-1/2}\tau_1 = \epsilon^{1/2}v_1$ gives $\epsilon\|\nabla v_1\|^2 = \epsilon^{-1}\|\tau_1\|^2 \lesssim \|g\|^2$ as well.

- **Bound on** $\| (v_2,\tau_2) \|_V$

  Lemma 4 provides

  $$\epsilon\|\nabla v\|^2 + \|v\|^2 \lesssim \epsilon\|f\|^2 \le \|f\|^2$$

  We have $\epsilon^{-1}\tau = f - \nabla v$, so $\epsilon^{-1}\|\tau\| \lesssim \|f\| + \|\nabla v\|$. Lemma 1 implies $\|\nabla v\|^2 \lesssim \|f\|^2$, so for $\epsilon\le 1$, we control $\epsilon^{-1/2}\|\tau\| \le \epsilon^{-1}\|\tau\| \lesssim \|f\|$. The remaining terms can be bounded by noting that, with $g=0$, $\|\nabla\cdot\tau_2\| = \|\beta\cdot\nabla v_2\| \lesssim \|\nabla v_2\| \lesssim \|f\|$.

□

2.5.2 Bound from above

We have shown the robust bound of a norm on U by energy error; for a full equivalence statement, we require a bound on the energy norm by some norm on U. By the duality of the energy and test norm, this is equivalent to bounding the test norm from below, using the test norm induced by the norm on U bounding the energy norm.

Lemma 2. *(Bound from above) For* $\left(u, \sigma, \widehat{u}, \widehat{f_n}\right) \in U$,

$$\left\|\left(u, \sigma, \widehat{u}, \widehat{f_n}\right)\right\|_E \lesssim \|u\| + \frac{1}{\sqrt{\epsilon}}\left(\|\sigma\| + \|\widehat{u}\| + \|\widehat{u}\|\right)$$

Proof. For any norm on U of the form

$$\left\|\left(u, \sigma, \widehat{u}, \widehat{f_n}\right)\right\|_{U,*} = \|u\| + C_2\|\sigma\| + C_3\|\widehat{u}\| + C_4\|\widehat{f_n}\|,$$

the induced test norm is

$$\begin{aligned}
\|\,(\tau, v)\,\|_{V,*} &= \sup_{\left(u,\sigma,\widehat{u},\widehat{f_n}\right)\in U\setminus\{0\}} \frac{b\left(\left(u, \sigma, \widehat{u}, \widehat{f_n}\right), (\tau, v)\right)}{\left\|\left(u, \sigma, \widehat{u}, \widehat{f_n}\right)\right\|_E} \\
&= \sup_{\left(u,\sigma,\widehat{u},\widehat{f_n}\right)\in U\setminus\{0\}} \frac{(u, \nabla\cdot\tau - \beta\cdot\nabla v) + \left(\sigma, \epsilon^{-1}\tau + \nabla v\right) - \langle[\tau_n], \widehat{u}\rangle + \langle\widehat{f_n}, [v]\rangle}{\|u\| + C_2\|\sigma\| + C_3\|\widehat{u}\| + C_4\|\widehat{f_n}\|} \\
&\lesssim \|g\| + \frac{1}{C_2}\|f\| + \sup_{\widehat{u},\widehat{f_n}\neq 0} \frac{\langle[\tau_n], \widehat{u}\rangle + \langle\widehat{f_n}, [v]\rangle}{C_3\|\widehat{u}\| + C_4\|\widehat{f_n}\|}
\end{aligned}$$

where f and g are again the loads of the adjoint problem (9). By the triangle inequality,

$$\begin{aligned}
\epsilon\|f\|^2 \leq \epsilon^{-1}\|\tau\|^2 + \epsilon\|\nabla v\|^2 &\lesssim \|\,(\tau, v)\,\|_V^2 \\
\|g\| \leq \|\nabla\cdot\tau\| + \|\beta\cdot\nabla v\| &\lesssim \|\,(\tau, v)\,\|_V
\end{aligned}$$

We estimate the supremum by following [5]; we begin by choosing $\eta \in H(\mathrm{div}; \Omega)$, $w \in H^1(\Omega)$, such that $(\eta - \beta w)\cdot n|_{\Gamma_+} = 0$ and $w|_{\Gamma_-\cup\Gamma_0} = 0$, and integrating the boundary pairing by parts to get

$$\begin{aligned}
\langle[\tau\cdot n], w\rangle + \langle[v], (\eta - \beta w)\cdot n\rangle &\lesssim \|\,(\tau, v)\,\|_V \frac{1}{\sqrt{\epsilon}}\left(\|\eta\| + \|w\|\right) \\
&\lesssim \|\,(\tau, v)\,\|_V \frac{1}{\sqrt{\epsilon}}\left(\|\eta - \beta w\| + \|w\|\right)
\end{aligned}$$

since $\|\eta\| = \|\eta - \beta w + \beta w\| \leq \|\eta - \beta w\| + \|\beta w\| \lesssim \|\eta - \beta w\| + \|w\|$. Dividing through and taking the supremum gives

$$\sup_{w,\eta\neq 0} \frac{\langle[\tau\cdot n], w\rangle + \langle[v], (\eta - \beta w)\cdot n\rangle}{\left(\|\eta - \beta w\| + \|w\|\right)} \lesssim \|\,(\tau, v)\,\|_V \frac{1}{\sqrt{\epsilon}}$$

To finish the proof, define $\rho \in H^{1/2}(\Gamma_h)$ and $\phi \in H^{-1/2}(\Gamma_h)$ such that $\rho = w|_{\Gamma_h}$ and $\phi = (\eta - \beta w)\cdot n|_{\Gamma_h}$, and note that

$$\sup_{\rho,\phi\neq 0} \frac{\langle[\tau\cdot n], \rho\rangle + \langle[v], \phi\rangle}{\|\rho\| + \|\phi\|} = \sup_{w,\eta\neq 0} \frac{\langle[\tau\cdot n], w\rangle + \langle[v], (\eta - \beta w)\cdot n\rangle}{\|w\| + \|\eta - \beta w\|}$$

Together, the bounds on $\|g\|$ and $\|f\|$ imply $\left\|\left(u, \sigma, \widehat{u}, \widehat{f_n}\right)\right\|_E \lesssim \left\|\left(u, \sigma, \widehat{u}, \widehat{f_n}\right)\right\|_{U,*}$ for $C_2 = C_3 = C_4 = \frac{1}{\sqrt{\epsilon}}$. □

3 Proof of lemmas/stability of the adjoint problem

We reduce the adjoint problem to the scalar second order equation

$$-\epsilon\Delta v - \beta\cdot\nabla v = g - \epsilon\nabla\cdot f \tag{9}$$

with boundary conditions

$$-\epsilon\nabla v\cdot n = f\cdot n, \quad x\in\Gamma_- \tag{10}$$
$$v = 0, \quad x\in\Gamma_+ \tag{11}$$

and treat the cases $f=0$, $g=0$ separately. motivate boundary conditions here

Additionally, the $\nabla\cdot$ operator is understood now in the weak sense, as the dual operator of $-\nabla : H_0^1(\Omega)\to L^2(\Omega)$, such that $\nabla\cdot f\in\left(H_0^1(\Omega)\right)'$. The normal trace $f\cdot n$ is understood as a limit; smooth functions are dense in $H^1(\Omega)$, and we define $f\cdot n$ as the limit of $f_i\cdot n$, where $f_i\in C^\infty(\Omega)$, $f_i\to f$. clearly define what we mean here with the density argument

Lemma 4 concerns the robust control of u by the energy error; Lemmas 3 and 5 concern the robust control of field variable σ and flux/trace variables, respectively.

Lemma 3. *Assume v satisfies* (9)*, with boundary conditions* (10)*, and* (11)*, and β satisfies* (5) *and* (6)*. If $\nabla\cdot f=0$ and ϵ is sufficiently small,*

$$\|\beta\cdot\nabla v\|\lesssim\|g\|$$

Proof. Define $v_\beta=\beta\cdot\nabla v$. Multiplying the adjoint equation (9) by v_β and integrating over Ω gives

$$\|v_\beta\|^2 = -\int_\Omega g v_\beta - \epsilon\int_\Omega \Delta v v_\beta$$

Note that

$$-\int_\Omega \beta\cdot\nabla v\Delta v = -\int_\Omega \beta\cdot\nabla v\nabla\cdot\nabla v$$

Integrating this by parts, we get

$$-\int_\Omega \beta\cdot\nabla v\nabla\cdot\nabla v = \int_\Omega \nabla(\beta\cdot\nabla v)\cdot\nabla v - \int_\Gamma n\cdot\nabla v\beta\cdot\nabla v$$

Note that

$$\nabla(\beta\cdot\nabla v) = \nabla\beta\cdot\nabla v + \beta\cdot\nabla\nabla v$$

where $\nabla\beta$ and $\nabla\nabla v$ are understood to be tensors. Then,

$$\int_\Omega \nabla(\beta\cdot\nabla v)\cdot\nabla v = \int_\Omega (\nabla\beta\cdot\nabla v)\cdot\nabla v + \int_\Omega \beta\cdot\nabla\nabla v\cdot\nabla v$$

Noting that $\nabla v\cdot\nabla\nabla v = \nabla\frac{1}{2}(\nabla v\cdot\nabla v)$, if we integrate by parts again, we get

$$\begin{aligned}-\int_\Omega \Delta v v_\beta &= -\int_\Gamma n\cdot\nabla v\beta\cdot\nabla v + \frac{1}{2}\int_\Gamma \beta_n(\nabla v\cdot\nabla v) - \frac{1}{2}\int_\Omega \nabla\cdot\beta(\nabla v\cdot\nabla v) + \int_\Omega(\nabla\beta\cdot\nabla v)\cdot\nabla v\\ &= -\int_\Gamma n\cdot\nabla v\beta\cdot\nabla v + \frac{1}{2}\int_\Gamma \beta_n(\nabla v\cdot\nabla v) + \int_\Omega \nabla v\left(\nabla\beta - \frac{1}{2}\nabla\cdot\beta I\right)\nabla v\end{aligned}$$

Finally, substituting this into our adjoint equation multiplied by v_β, we get

$$\|v_\beta\|^2 = -\int_\Omega g\beta\cdot\nabla v + \epsilon\int_\Gamma\left(-n\cdot\nabla v\beta + \frac{1}{2}\beta_n\nabla v\right)\cdot\nabla v + \epsilon\int_\Omega \nabla v\left(\nabla\beta - \frac{1}{2}\nabla\cdot\beta I\right)\nabla v$$

The last term can be bounded by our assumption on $\|\nabla\beta - \frac{1}{2}\nabla\cdot\beta I\|^2 \leq C$.

$$\epsilon\int_\Omega \nabla v\left(\nabla\beta - \frac{1}{2}\nabla\cdot\beta I\right)\nabla v \leq C\frac{\epsilon}{2}\|\nabla v\|^2$$

For the boundary terms, on Γ_-, $\nabla v\cdot n = 0$, reducing the integrand over the boundary to $\beta_n|\nabla v|^2 \leq 0$. On Γ_+, $v = 0$ implies $\nabla v\cdot\tau = 0$, where τ is a tangential direction. An orthogonal decomposition yields $\nabla v = (\nabla v\cdot n)n$, reducing the above to

$$\epsilon\int_\Gamma -\frac{1}{2}|\beta_n|(\nabla v\cdot n)^2 \leq 0$$

leaving us with the estimate

$$\|v_\beta\|^2 \leq -\int_\Omega g\beta\cdot\nabla v + C\frac{\epsilon}{2}\|\nabla v\|^2$$

Since $C = O(1)$, an application of Young's inequality and Lemma 4 complete the estimate. □

Lemma 4. *Assume* (5) *holds. Then, for v satisfying equation* (9) *with boundary conditions* (10), (11) *and sufficiently small ϵ,*

$$\epsilon\|\nabla v\|^2 + \|v\|^2 \lesssim \|g\|^2 + \epsilon\|f\|^2$$

Proof. Since $\nabla\times\beta = 0$, and Ω is simply connected, there exists a scalar potential ψ, $\nabla\psi = \beta$ such that $e^\psi = O(1)$. Take the transformed function $w = e^\psi v$; following (2.26) in [5], we substitute w into the the left hand side of equation (9), arriving at the relation

$$-\epsilon\Delta w - (1-2\epsilon)\beta\cdot\nabla w + \left((1-\epsilon)|\beta|^2 + \epsilon\nabla\cdot\beta\right)w = e^\psi(g - \epsilon\nabla\cdot f)$$

Multiplying by w and integrating over Ω gives

$$-\epsilon\int_\Omega \Delta w w - (1-2\epsilon)\int_\Omega \beta\cdot\nabla w w + \int_\Omega\left((1-\epsilon)|\beta|^2 + \epsilon\nabla\cdot\beta\right)w^2 = \int_\Omega e^\psi(g-\epsilon\nabla\cdot f)w$$

Integrating by parts gives

$$-\epsilon\int_\Omega \Delta w w - (1-2\epsilon)\int_\Omega \beta\cdot\nabla w w = \epsilon\left(\int_\Omega|\nabla w|^2 - \int_\Gamma w\nabla w\cdot n\right) + \frac{(1-2\epsilon)}{2}\left(\int_\Omega\nabla\cdot\beta w^2 - \int_\Gamma\beta_n w^2\right)$$

Noting that $w = 0$ on Γ_+ reduces the boundary integrals over Γ to just the inflow Γ_-. Furthermore, we have the relation $\nabla w = e^\psi(\nabla v + \beta v)$. Applying the above and boundary conditions on Γ_-, the first boundary integral becomes

$$\int_{\Gamma_-} w\nabla w\cdot n = \int_{\Gamma_-} we^\psi(\nabla v + \beta v)\cdot n = \int_{\Gamma_-} we^\psi(f\cdot n + \beta_n v)$$

Noting $\int_{\Gamma_-}\beta_n w^2 \leq 0$ through $\beta_n < 0$ on the inflow gives

$$\epsilon\int_\Omega|\nabla w|^2 + \int_\Omega\left((1-\epsilon)|\beta|^2 + \frac{1}{2}\nabla\cdot\beta\right)w^2 - \epsilon\int_{\Gamma_-} we^\psi f\cdot n \leq \int_\Omega e^\psi(g-\epsilon\nabla\cdot f)w$$

assuming ϵ is sufficiently small. Our assumptions on β implies $\left((1-\epsilon)|\beta|^2 + \frac{1}{2}\nabla\cdot\beta\right) \lesssim 1$ and $e^\psi = O(1)$. We can then bound from below

$$\epsilon\|\nabla w\|^2 + \|w\|^2 - \epsilon\int_{\Gamma_-} we^\psi f\cdot n \lesssim \epsilon\int_\Omega|\nabla w|^2 + \int_\Omega\left((1-\epsilon)|\beta|^2 + \frac{1}{2}\nabla\cdot\beta\right)w^2 - \epsilon\int_{\Gamma_-} we^\psi f\cdot n$$

Interpreting $\nabla \cdot f$ as a functional, the right hand gives

$$\int_\Omega e^\psi (g - \epsilon \nabla \cdot f) w = \int_\Omega e^\psi g + \int_\Omega \epsilon f \cdot \nabla (e^\psi w) - \int_\Gamma \epsilon f \cdot n e^\psi w$$

The boundary integral on Γ reduces to Γ_-, which is then nullified by the same term on the left hand side, leaving us with

$$\epsilon \|\nabla w\|^2 + \|w\|^2 \lesssim \int_\Omega e^\psi g + \int_\Omega \epsilon f \cdot \nabla (e^\psi w) = \int_\Omega e^\psi g + \int_\Omega \epsilon f \cdot (\beta w + \nabla w)$$

From here, the proof is identical to [5]; an application of Peter-Paul Young's inequality to the right hand side and bounds on $\|v\|, \|\nabla v\|$ by $\|w\|, \|\nabla w\|$ complete the estimate. □

Lemma 5. *Let β satisfy* (5) *and* (7)*, and let v satisfy* (9) *with $f = g = 0$, and arbitrary boundary conditions. Then, there holds*

$$\|\nabla v\| = \frac{1}{\epsilon}\|\tau\| \lesssim \frac{1}{\epsilon}\|[\tau \cdot n]\|_{\Gamma_h \setminus \Gamma_+} + \frac{1}{\sqrt{\epsilon}}\|[v]\|_{\Gamma_h^0 \cup \Gamma_+}$$

Proof. We follow [3] and [5] in decomposing τ into $z \in H(\mathrm{curl}, \Omega)$ and $\psi \in H^1(\Omega)$ such that

$$\tau = (\epsilon \nabla \psi - \beta \psi) + \nabla \times z$$

Specifically, ψ is chosen as the solution to

$$\begin{aligned} -\epsilon \Delta \psi + \nabla \cdot (\beta \psi) &= -\nabla \cdot \tau \\ \epsilon \nabla \psi \cdot n - \beta_n \psi - \tau \cdot n = 0, &\quad x \in \Gamma_- \\ \psi = 0, &\quad x \in \Gamma_+ \end{aligned}$$

Since $\nabla \cdot \beta = 0$, we satisfy (5), and can use Lemma 4 (with $f = \frac{1}{\epsilon}\tau$) to bound

$$\epsilon \|\nabla \psi\|^2 + \|\psi\|^2 = \frac{1}{\epsilon}\|\tau\|^2$$

By the above bound and triangle inequality,

$$\|\nabla \times z\| \leq \epsilon \|\nabla \psi\| + \|\beta \psi\| + \|\tau\| \lesssim \frac{1}{\sqrt{\epsilon}}\|\tau\|$$

On the other hand, using the decomposition and boundary conditions directly, we can integrate by parts to arrive at

$$\begin{aligned} \|\tau\|^2 &= (\tau, \epsilon \nabla \psi - \beta \psi + \nabla \times z) = (\tau, \epsilon \nabla \psi) - (\tau, \beta \psi) + (\tau, \nabla \times z) \\ &= (\tau, \epsilon \nabla \psi) + \epsilon (\nabla v, \beta \psi) - \epsilon (\nabla v, \nabla \times z) \\ &= \epsilon \langle [\tau \cdot n], \psi \rangle - \epsilon \langle n \cdot \nabla \times z, [v] \rangle = \epsilon \langle [\tau \cdot n], \psi \rangle_{\Gamma_h \setminus \Gamma_+} - \epsilon \langle n \cdot \nabla \times z, [v] \rangle_{\Gamma_h \cup \Gamma_+} \\ &\lesssim \epsilon \|[\tau \cdot n]\| \|\psi\| + \epsilon \|[v]\| \|n \cdot \nabla \times z\| \end{aligned}$$

by definition of the boundary norms on $[\tau \cdot n]$ and $[v]$.

Applying the above estimates for $\|\psi\|$ and $\|n \cdot \nabla \times z\|$ and noting that $\|\nabla v\| = \frac{1}{\epsilon}\|\tau\|$ completes the proof. □

4 Numerical experiments

In our numerical experiments, we will be interested in the use of the unweighted test norm

$$\| (\tau, v) \|_V^2 = \|v\|^2 + \epsilon\|\nabla v\|^2 + \|\beta \cdot \nabla v\|^2 + \|\nabla \cdot \tau\|^2 + \frac{1}{\epsilon}\|\tau\|^2$$

However, the presence of both $\|v\|$ and $\epsilon\|\nabla v\|$ terms, and similarly $\|\nabla \cdot \tau\|$ and $\frac{1}{\epsilon}\|\tau\|$ terms, induces boundary layers in the optimal test functions, though only in the crosswind direction. To avoid this effect, we follow [5] in scaling the L^2 contributions of v and τ such that, when transformed to the reference element, both v and ∇v terms are $O(1)$. In this paper, we consider only isotropic refinements on quadrilateral elements, such that $h_1 = h_2 = h$, and $|K| = h^2$. Our test norm for an element K is now

$$\| (\tau, v) \|_{V,K}^2 = \min\left\{\frac{\epsilon}{|K|}, 1\right\}\|v\| + \epsilon\|\nabla v\| + \|\beta \cdot \nabla v\| + \|\nabla \cdot \tau\| + \min\left\{\frac{1}{\epsilon}, \frac{1}{|K|}\right\}\|\tau\|$$

This modified test norm avoids boundary layers, but for adaptive meshes, provides additional stability in areas of heavy refinement, where best approximation error tends to be large and stronger robustness is most needed.

In each numerical experiment, we vary $\epsilon = .01, .001, .0001$ in order to demonstrate robustness over a range of ϵ. This is intended to mirror the numerical experiments in [5]; for "worst-case" linear solvers, such as LU decomposition without pivoting, the effect of roundoff error becomes evident in the solving of optimal test functions for $\epsilon \leq O(1e-5)$. The roundoff itself comes from the Gram matrix under certain test norms; for example, if the weighted $\big(H(\text{div};\Omega), H^1(\Omega)\big)$ norm is used for the test norm $\| (\tau, v) \|_V$ (as was done in [4]), for an element of size h, $\|v\|_{L^2}^2 = O(h)$, while $\|\nabla v\|_{L^2}^2 = O(h^{-1})$. As $h \to 0$, the seminorm portion of the test norm dominates the Gram matrix, leading to a near-singular and ill-conditioned system.

The effect of roundoff error is often characterized by an increase in the energy error, which (assuming approximability of test functions) is proven to decrease for any series of refined meshes. These roundoff effects are dependent primarily on the mesh, appearing when trying to fully resolve boundary layers for very small ϵ.

4.1 Erickson model problem

For the choice of $\Omega = (0,1)^2$ and $\beta = (1,0)^T$, the convection diffusion equation reduces to

$$\frac{\partial u}{\partial x} - \epsilon\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = 0$$

which has an exact solution by separation of variables, allowing us to analyze convergence of DPG for a wide range of ϵ. For boundary conditions, we impose $u = 0$ on Γ_+ and $\beta_n u - \sigma_n$ on Γ_-, which reduces to

$$\begin{aligned} u - \sigma_x &= u_0 - \sigma_{x,0}, \quad x = 0 \\ \sigma_y &= 0, \quad y = 0, 1 \end{aligned}$$

In this case, our exact solution is the series

$$u(x,y) = C_0 + \sum_{n=1}^{\infty} C_n \frac{\exp(r_2(x-1) - \exp(r_1(x-1)))}{r_1\exp(-r2) - r_2\exp(-r1)}\cos(n\pi y)$$

where

$$\begin{aligned} r_{1,2} &= \frac{1 \pm \sqrt{1 + 4\epsilon\lambda_n}}{2\epsilon} \\ \lambda_n &= n^2\pi^2\epsilon \end{aligned}$$

The constants C_n depend on a given inflow condition u_0 at $x = 0$

$$C_n = \int_0^1 u_0(y) \cos(n\pi y)$$

4.1.1 Solution with $C_1 = 1, C_{n\neq 1} = 0$

We begin with the solution taken to be the first non-constant mode of the above series. We set the inflow boundary condition to be exactly the value of $u - \sigma_x$ corresponding to the exact solution.

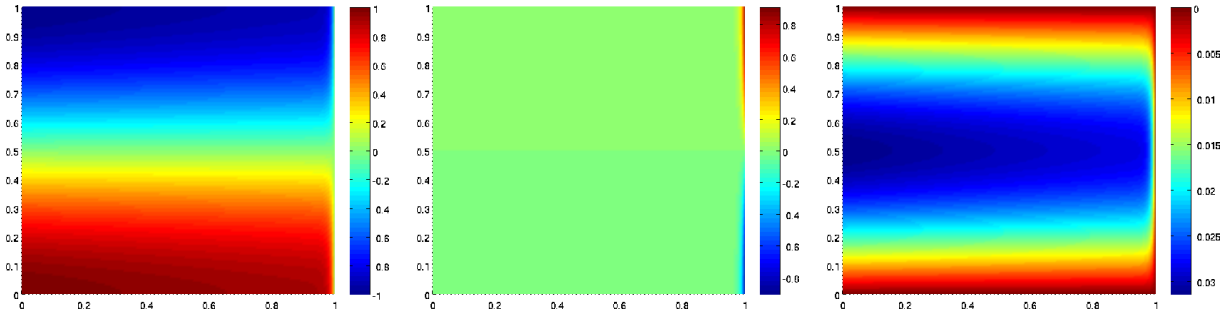

Figure 1: Solution for u, σ_x, and σ_y for $\epsilon = .01$, $C_1 = 1$, $C_n = 0$, $n \neq 1$

In each case, we begin with a square 4 by 4 mesh of quadrilateral elements with order $p = 3$. We choose $\Delta p = 5$, though we note that the behavior of DPG is nearly identical for any $\Delta p > 3$. h-refinements are executed using a greedy refinement algorithm, where element energy error e_K^2 is computed for all elements K, and elements such that $e_K^2 \leq \alpha \max_K e_K^2$ are refined. We make the arbitrary choice of taking $\alpha = .2$ for each of these experiments.

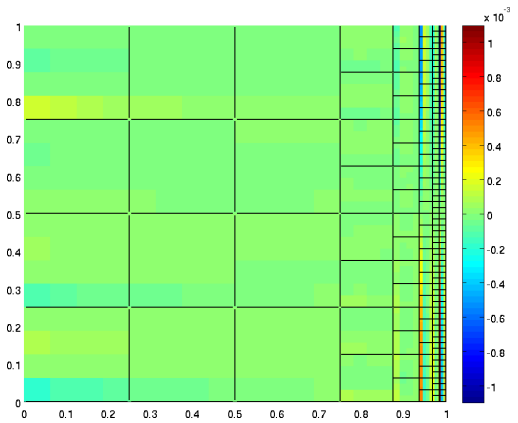

Figure 2: Adapted mesh and pointwise error for $\epsilon = .01$

We are especially interested in the ratio of energy error and total L^2 error in both σ and u. Stability estimates imply that, using the above test norm, $\|u - u_h\|_{L^2}/\|u - u_h\|_E \leq C$ independent of ϵ. Figure 3 seems to imply that, at least for this model problem, $C = O(1)$. Additionally, while we do not have a robust lower bound (i.e. $\|u - u_h\|_{L^2}/\|u - u_h\|_E$ can approach 0 as $\epsilon \to 0$), our numerical results present a robust lower bound.

The effect of a mesh dependent test norm can be seen in the ratios of L^2 to energy error; as the mesh is refined, the constants in front of the L^2 terms for v and τ converge to stationary values, and the ratio of L^2 to energy error transitions from a smaller to a larger value. The transition point happens later for smaller ϵ, which we expect, since the transition of the ratio corresponds to the introduction of elements of order ϵ through mesh refinement.

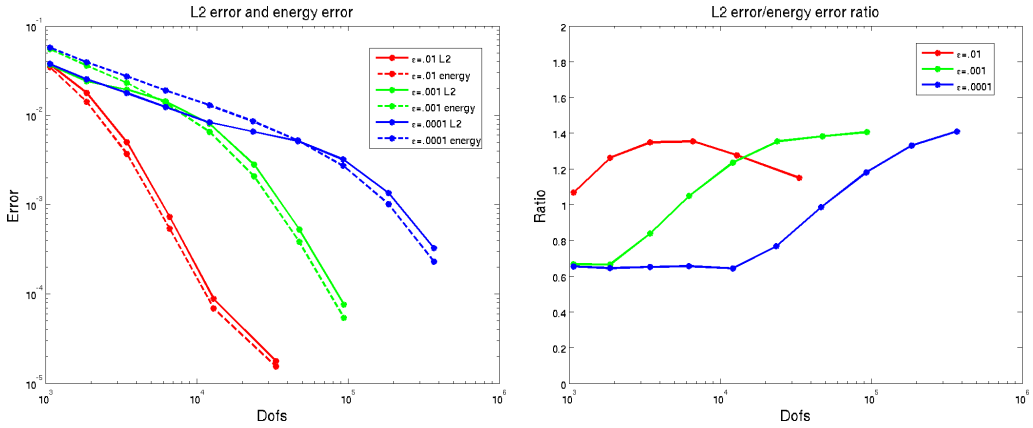

Figure 3: L^2 and energy errors, and their ratio for $\epsilon = .01$, $\epsilon = .001$, $\epsilon = .0001$

We examined smaller ϵ as well. In [5], the smallest resolvable ϵ using only double precision arithmetic was $1e-4$. The solution of optimal test functions is now done using both pivoting and equilibration, improving conditioning. Roundoff effects still appear, but at smaller values of ϵ.

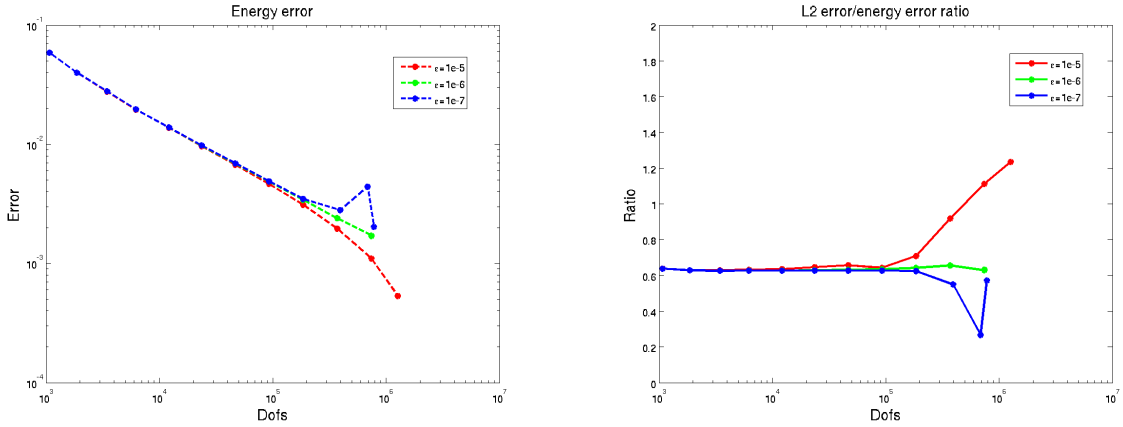

Figure 4: Energy error and L^2/energy error ratio for $\epsilon = 1e-5$, $\epsilon = 1e-6$, $\epsilon = 1e-7$. Non-monotonic behavior of the energy error indicates conditioning issues and roundoff effects.

Without anisotropic refinements, it still becomes computationally difficult to fully resolve the solution for smaller ϵ. Regardless, for all ranges of ϵ, the robustness of DPG remains constant, as indicted by the rates and ratio between L^2 and energy error in Figure 4. For $\epsilon = 1e-5$, we observe that the ratio of L^2 error increases, corresponding to the scaling of the test norm with mesh size (the transition in test norm occurs after 8 refinements, which, for an initial 4×4 mesh, implies a minimum element size of about $1.5e-05$. At this point, rescaled test norm allows us to take advantage of the full magnitude of the L^2 term for $\|v\|$ and $\|\tau\|$ implied by our adjoint estimates). By analogy, for smaller $\epsilon = 1e-6, 1e-7$, the transition period should begin near the 10th and 11th refinement iterations; however, we do not observe such behavior. For $\epsilon = 1e-6$, the ratio simply remains constant, but for $\epsilon = 1e-7$, we observe roundoff effects, as the energy error increases at the 11th refinement. Since DPG is optimal in the energy norm for a fixed test norm[1], we expect monotonic decrease of the energy error with mesh refinement. Non-monotonic behavior indicates either approximation or roundoff error, and as there was no qualitative difference between using $\Delta p = 5$ and $\Delta p = 6$ for these experiments, we expect that the approximation error is negligible

[1]While the test norm changes with the mesh, it increases monotonically. A strictly stronger test norm implies $\frac{b(u,v)}{\|v\|_1} \geq \frac{b(u,v)}{\|v\|_2}$ for any $\|v\|_1 \leq \|v\|_2$

and conclude roundoff effects are at play.

4.1.2 Neglecting σ_n

In practice, we will not have prior knowledge of σ_n at the inflow, and will have to set $\beta_n u - \sigma_n = u_0$, ignoring the viscous contribution to the boundary condition. The hope is that for small ϵ, this omission will be negligible. Figure 5 indicates that, between $\epsilon = .005$ and $\epsilon = .001$, the omission of σ_n in the boundary condition becomes negligible, and both our error rates and ratios of L^2 to energy error become identical to the case where σ_n is explicitly accounted for, well within the range of ϵ of physical interest. For large $\epsilon = .01$, the L^2 error stagnates around $1e-3$, or about 7% relative error.

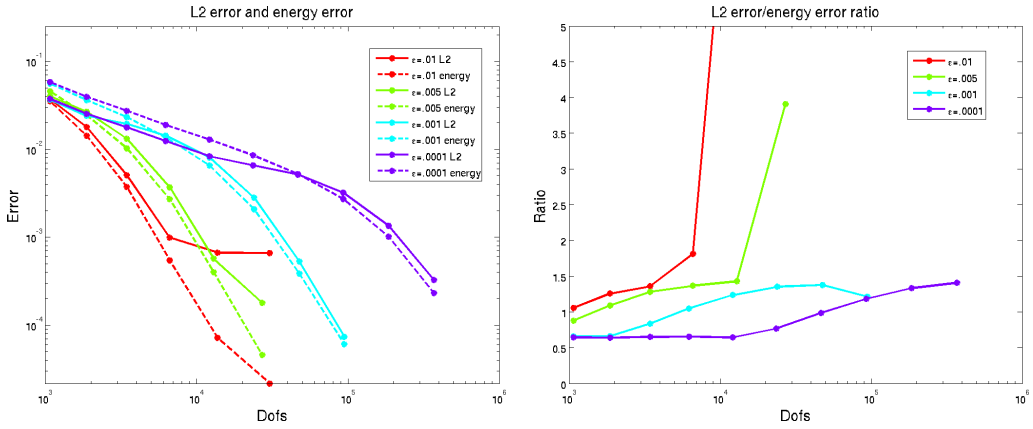

Figure 5: L^2 and energy errors and their ratio when neglecting σ_n at the inflow.

4.1.3 Discontinuous inflow data

We note also that an additional advantage of selecting this new boundary condition is a relaxation of regularity requirements; as $\widehat{f_n} \in H^{-1/2}(\Omega)$, strictly discontinuous inflow boundary conditions are no longer "variational crimes". We consider the discontinuous inflow condition

$$u_0(y) = \begin{cases} (y-1)^2, & y > .5 \\ -y^2, & y \leq .5 \end{cases}$$

as an example of a more difficult test case.

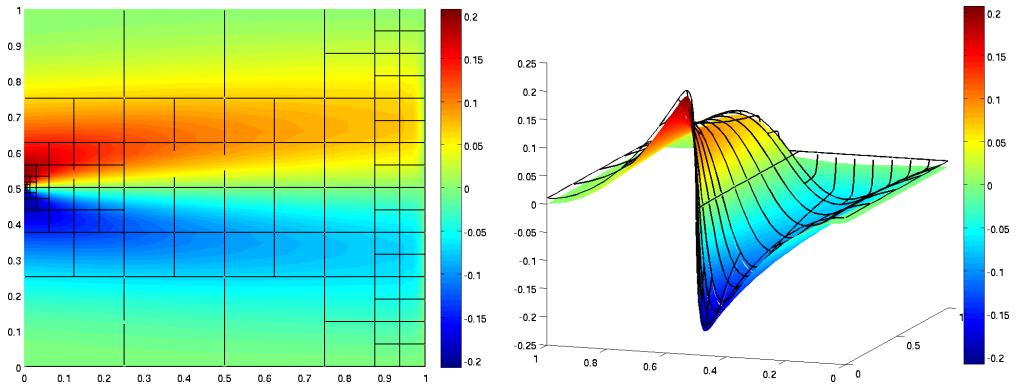

Figure 6: Solution variables u and $\widehat{u}$ with discontinuous inflow data u_0 for $\epsilon = .01$.

Figure 6 shows the solution u and overlaid trace variable $\widehat{u}$, which both demonstrate the regularizing effect of viscosity on the discontinuous boundary condition at $x = 0$. In order to compare with

an exact solution, we approximate u_0 with 20 terms of a Fourier series, giving a near-discontinuity for u_0.

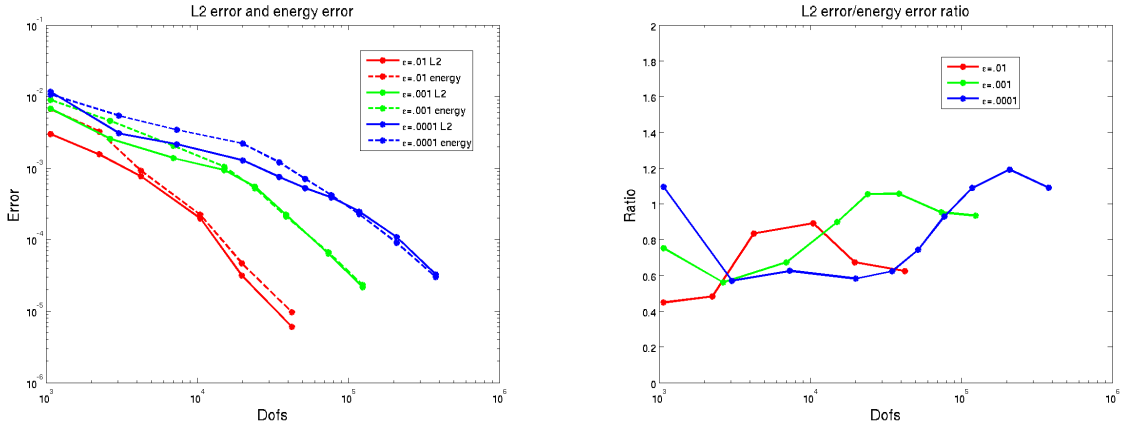

Figure 7: L^2 and energy errors, and their ratio for $\epsilon = .01$, $\epsilon = .001$, $\epsilon = .0001$, with discontinuous u_0 approximated by a Fourier expansion.

The ratios of L^2 to energy error are now less predictable than for the previous example, due to the "moving target" nature of the approximation of oscillatory boundary conditions. The numerical experiments were originally performed by applying boundary conditions via interpolation; the result was that the highly oscillatory inflow boundary condition was not sampled enough to be properly resolved, causing the solution to converge to a solution different than the exact solution. The experiments were repeated using the penalty method to enforce inflow conditions; however, we note that the proper way to do so is to use an L^2 projection at the boundary. When using the penalty method, however, the ratios still remain bounded and close to 1 for ϵ varying over two orders of magnitude, as predicted by theory.

5 Conclusions

None yet. Ideas:

- Explanation of choice of boundary condition and why it impacts stability
- Interpretation of weight as allowing diffusion to work on the boundary.
- Note results have worked for Burgers, and expand on how they would apply to Navier-Stokes

References

[1] Antti H. Niemi, Jamie A. Bramwell, and Leszek F. Demkowicz. Discontinuous Petrov-Galerkin Method with Optimal Test Functions for Thin-body Problems in Solid Mechanics. Technical Report 17, ICES, 2010.

[2] L. Demkowicz. Babuska <=> Brezzi? Technical Report 06-08, ICES, 2006.

[3] L. Demkowicz and J. Gopalakrishnan. An analysis of the DPG method for the Poisson equation. Technical Report 10-37, ICES, 2010.

[4] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. ii. Optimal test functions. *Num. Meth. for Partial Diff. Eq*, 27:70–105, 2011.

[5] L. Demkowicz and N. Heuer. Robust DPG method for convection-dominated diffusion problems. Technical Report 11-33, ICES, 2011.

[6] J. Gopalakrishnan and W. Qiu. An analysis of the practical DPG method. Technical report, IMA, 2011. submitted.

[7] J.S. Hesthaven. A stable penalty method for the compressible navier-stokes equations. iii. multi dimensional domain decomposition schemes. *SIAM J. SCI. COMPUT*, 17:579–612, 1996.

[8] Antti H. Niemi, Nathan O. Collier, and Victor M. Calo. Discontinuous Petrov-Galerkin method based on the optimal test space norm for one-dimensional transport problems. *Procedia CS*, 4:1862–1869, 2011.

[9] Nathan V. Roberts, Denis Ridzal, Pavel B. Bochev, and Leszek D. Demkowicz. A Toolbox for a Class of Discontinuous Petrov-Galerkin Methods Using Trilinos. Technical Report SAND2011-6678, Sandia National Laboratories, 2011.

[10] J. Zitelli, I. Muga, L. Demkowicz, J. Gopalakrishnan, D. Pardo, and V. Calo. A Class of Discontinuous Petrov-Galerkin Methods. Part IV: Wave Propagation Problems. Technical Report 17, ICES, 2010.