Contents

ϵ-explicit analysis of DPG for convection-dominated diffusion

Tan Bui-Thanh Jesse Chan Leszek Demkowicz Norbert Heuer

1 Introduction

1.1 Singular perturbation problems and robustness

The finite element/Galerkin method has been widely utilized in engineering to solve partial differential equations related to simulation of solid mechanics problems. For many problems in engineering, the bilinear forms resulting from these partial differential equations are symmetric and coercive (positive-definite), and it is well known that the finite element method produces near-optimal results for such problems. However, standard Bubnov-Galerkin methods tend to perform poorly for the class of partial-differential equations known as singularly perturbed problems. Singularly perturbed problems are problems where a given parameter may be either very small or very large in the context of physical problems. An additional complication of singularly perturbed problems is that very often, in the limiting case of the parameter blowing up or decreasing to zero, the partial differential equation itself will change types (i.e from parabolic to hyperbolic). A canonical example is the convection-diffusion equation with Dirichlet boundary conditions.

1.1.1 Convection-diffusion

In 1D, the convection-diffusion equation is

$$u' - \epsilon u'' = f$$

The equation represents the change in the concentration of some quantity u in some medium from both convective and diffusive effects. The parameter ϵ represents the diffusivity, and in the limit of an inviscid medium as $\epsilon \to 0$, the partial differential equation changes types, going from a parabolic to a hyperbolic equation. For Dirichlet boundary conditions $u(0) = u_0$ and $u(1) = u_1$, the solution can develop sharp boundary layers of order ϵ near the outflow.

The finite element method has notoriously performed poorly for both this example, and for many other singularly perturbed problems. This poor performance is captured by the error bound for the finite element method applied to convection-diffusion equations - for the standard Bubnov-Galerkin method with $u \in H^1(0,1)$, we have the bound given in [14]

$$\|u - u_h\|_\epsilon \leq C \inf_{w_h} \|u - w_h\|_{H^1(0,1)},$$

for C independent of ϵ, and for $\|u\|_\epsilon \coloneqq \|u\|_{L^2} + \epsilon\|u'\|_{L^2}$. An alternative formulation of the above bound is

$$\|u - u_h\|_{H^1(0,1)} \leq C(\epsilon) \inf_{w_h} \|u - w_h\|_{H^1(0,1)},$$

where $C(\epsilon)$ grows as $\epsilon \to 0$.

The dependence on ϵ in the bound of our finite element error by best approximation error is referred to as a *loss of robustness*. As our singular perturbation parameter changes, our finite

element error is bounded more and more loosely by the best approximation error, allowing for more and more error in the finite element solution for as ϵ shrinks.

Intuitively, a non-robustness method will perform increasingly worse as ϵ decreases, independently of the best approximation error. For example, for a coarse mesh and small values of ϵ, the Galerkin approximation of the solution to the convection-diffusion equation with a boundary layer develops spurious oscillations of magnitude $O(\epsilon^{-1})$ everywhere in the domain, even where best approximation error is small.

1.1.2 Wave propagation

Another example of a singular perturbation problem which experiences loss of robustness is high frequency wave propagation problems, in which the singular perturbation parameter is the wavenumber, $k \to \infty$. The loss of robustness in this case manifests as "pollution" error, a phenomenon where the finite element solution degrades over many wavelengths for high wavenumbers (commonly manifesting as a phase error between the FE solution and the exact solution).

1.1.3 Stabilization terms

Traditionally, instability/loss of robustness has been dealt with using residual-based stabilization. Given some variational form, the problem is modified by adding to the bilinear form a factor of the residual, scaled by a stabilization constant τ. The most well-known example of this is the streamline-upwind Petrov-Galerkin method, which is a stabilized method for solving the convection-diffusion equation using piecewise linear continuous finite elements [2]. An important difference between residual-based stabilization techniques and other stabilizations is the idea of *consistency* - by adding stabilization terms based on the residual, the exact solution still satisfies the same variational problem (i.e Galerkin orthogonality still holds)[1].

The addition of residual-based stabilization terms can also be interpreted as a modification of the test functions - in other words, stabilization can be achieved by changing the test space for a given problem. This idea can be generalized under the idea of *optimal test functions*.

1.2 Optimal test functions

The idea of optimal test functions was introduced by Demkowicz and Gopalakrishnan in [7]. Conceptually, these optimal test functions are the natural result of the minimumization of a residual corresponding the operator form of a variational equation. The connection between stabilization and least squares/minimum residual methods has been observed previously; however, the concept of optimal test functions presents the idea from a different perspective.

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

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

[1] Contrast this to an artificial diffusion method, where a specific amount of additional viscosity is added based on the interplay between the convection and diffusion terms. The exact solution to the original equation no longer satisfies the new stabilized formulation.

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 determined by solving the auxiliary variational problem

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

This concept is not a new one; the idea of optimal testing is a classical one (see find citation). However, for standard conforming finite elements, test functions are continuous over the entire domain, and solving the above variational problem for optimal test functions is a global operation that must be performed once for every approximating function in the trial space, rendering the method impractical. The breakthrough came through the development of discontinous Galerkin (DG) methods; specifically, the use of discontinuous test spaces reduces the problem of determining optimal test functions to a local problem that can be solved in an element-by-element fashion.

In practice, this variational problem is difficult to solve exactly, and the solution is instead is approximately using the standard Bubnov-Galerkin method and an "enriched" subspace of V such that $\dim(V) > \dim(U)$. 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 [10].

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"

$$\|u\|_E = \sup \frac{b(u, v)}{\|v\|_V}.$$

The above norm can be determined via $\|u\|_E = \|v_u\|_V$, where v_u solves

$$(v_u, \delta v)_V = b(u, \delta v), \quad \forall v \in V$$

Due to the isometry of the Riesz map, this is equivalent to measuring the norm of the functional $b(u, v) = l_u(v) \in V'$ in the dual norm. An additional consequence of adopting such an energy norm is that, without knowing the exact solution, the actual energy error $\|u - u_h\|_E$ is computable through $\|e_h\|_V$, where

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

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 - the dual norm on V.

1.3 Variational form

The discontinuous Petrov-Galerkin (DPG) method is the pairing of the concept of locally computable optimal test functions with the so-called "ultra-weak formulation". For a given operator equation in strong form $Au = f$, the ultra-weak formulation is

$$b\left((\widehat{u}, u), v\right) - l(v) = \langle \widehat{u}, [v] \rangle - (u, A^* v) - (f, v) = 0$$

where A^* is the formal adjoint resulting from integration by parts, and $[v]$ is the jump of the test function v across element boundaries. Both the inner product and formal adjoint are understood in an element-wise fashion. Under the ultra-weak formulation, 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 DG method with locally supported optimal test functions, where both fluxes and traces are hybridized. The energy setting is

$$\begin{aligned} u &\in L^2(\Omega) \\ v &\in V = D(A^*) \\ \widehat{u} &\in T(D(A)) \end{aligned}$$

where $D(A^*)$ is the domain of the formal adjoint A^*, and $T(D(A))$ is the trace space of the domain of the operator A in its strong form. A more comprehensive discussion of the abstract setting of DPG, as well as a proof of well-posedness, is developed using the graph norm in [3].

1.4 Induced norms

Up to now, we have neglected discussion of the proper choice of inner product/norm on the space V. The energy norm in which DPG is optimal

$$\|u\|_E = \sup_{v \in V \setminus \{0\}} \frac{b(u,v)}{\|v\|_{V,E}}.$$

depends heavily on our choice of norm on the test space $\|v\|_{V,E}$.

An important relationship is the duality of the energy norm and test norm; under the assumption of injectivity of the adjoint of our variational operator B, for any choice of energy norm on U, the test norm $\|v\|_V$ inducing it can be recovered via duality

$$\|v\|_{V,U} = \sup_{u \in U \setminus \{0\}} \frac{b\,(u,v)}{\|u\|_U} \tag{1}$$

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

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

where we're able to switch the order of inf and sup due to the injectivity of B (see [4]).

As a consequence of (1), given two norms on U, $\|\cdot\|_{U_1}$ and $\|\cdot\|_{U,2}$, if $\|u\|_{U_1} \leq \|u\|_{U,2}$, we have the equivalent statement

$$\|v\|_{V,U_1} = \sup_{u \in U \setminus \{0\}} \frac{b\,(u,v)}{\|u\|_{U_1}} \geq \sup_{u \in U \setminus \{0\}} \frac{b\,(u,v)}{\|u\|_{U,2}} = \|v\|_{V,U,2}$$

In other words, showing an inequality between two norms on U is equivalent to showing the reverse inequality between the respective induced norms on V.

1.5 Optimal and quasi-optimal test norms

The duality between test norm and variational form holds for any particular variational form, though it is often difficult to determine analytically the form of $\|v\|_{V,U}$. Whether a particular choice of norm on V norm is "good" or not is problem-dependent; however, for the DPG ultra-weak formulation, there is an important canonical test norm that is induced by a specific norm on U.

Define the norm on U

$$\|u\| = C_1\|u\|_{L^2} + C_2\|\widehat{u}\|$$

where the norm on $\widehat{u}$ is taken to be the minimum energy extension norm

$$\|\widehat{u}\| = \inf_{w \in V,\, w|_{\Gamma_h} = \widehat{u}} \|w\|_V$$

Then, the induced test norm is

$$\begin{aligned}
\|v\| &= \sup_{u,\widehat{u}} \frac{\langle \widehat{u}, [v] \rangle - (u, A^* v)}{C_1 \|u\|_{L^2} + C_2 \|\widehat{u}\|} \\
&\leq \frac{1}{C_1} \sup_{u} \frac{(u, A^* v)}{\|u\|_{L^2}} + \frac{1}{C_2} \sup_{\widehat{u}} \frac{\langle \widehat{u}, [v] \rangle}{\|\widehat{u}\|} \\
&= \frac{1}{C_1} \|A^* v\|_{L^2} + \frac{1}{C_2} \sup_{\widehat{u}} \frac{\langle \widehat{u}, [v] \rangle}{\|\widehat{u}\|}
\end{aligned}$$

where we have used the Cauchy-Schwarz inequality, and eliminated the supremum by noting that the Cauchy-Schwarz inequality is tight. We define the optimal/ideal norm on U as $\|u\|_{\text{opt}}$ as such a norm with $C_1 = C_2 = 1$,

$$\|u\|_{\text{opt}} = \|u\|_{L^2} + \|\widehat{u}\|.$$

which induces the so-called "optimal test norm"

$$\|v\|_{\text{opt}} = \|A^* v\|_{L^2} + \left(\sup_{\widehat{u}} \frac{\langle \widehat{u}, [v] \rangle}{\|\widehat{u}\|} \right)$$

This is, in some sense, the canonical or ideal test norm; given any well-posed variational problem, DPG under the optimal test norm will return the best finite element approximation in a balanced norm on both field and flux variables.

Unfortunately, the norm $\|v\|_{\text{opt}}$ is non-localizable, meaning that it cannot be decomposed into the sum of independent contributions over individual elements. Consequently, the optimal test norm must be induced by a global inner product due to the presence of the jump terms. Solving for optimal test functions under this optimal norm (through inversion of Riesz map corresponding to this inner product) will result in expensive global problems, making this an impractical norm to work with. 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 norm

$$\|v\|_{\text{qopt}} \coloneqq \|A^* v\|_{L^2} + \alpha \|v\|_{L^2}$$

where $\alpha \in \mathbb{R}$ is chosen to complete the seminorm defined by $\|A^* v\|_{L^2}$. This optimal test norm has been shown to induce a well-posed DPG method for any well-posed ultra-weak formulation in the context of Friedrichs' systems of equations [3].

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 [15]. Similar results have also been obtained with the quasi-optimal test norm in the context of the elasticity [1] and linear Stokes equation [13].

2 Model problem and robustness

Our aim now is to translate the above abstract language to a concrete model problem in computational fluid dynamics. We consider the model convection-diffusion problem on domain $\Omega \in \mathbb{R}^d$ with boundary Γ

$$\nabla \cdot (\beta u) - \epsilon \Delta v = f \tag{2}$$

in first order form

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

with the corresponding variational form (prior to the application of boundary conditions)

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

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

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

where n is the outward normal vector to an element edge, and $\text{sgn}(n)$ is 1 if n aligns with a given normal orientation n_e on each edge, and -1 otherwise (see Examples in [3] for a more detailed explanation).

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. Similarly, we define the interior skeleton $\Gamma_h^0 = \Gamma_h \setminus \Gamma$. The divergence and gradient operators are likewise understood to act element-wise.

The functional setting is now well understood as well (see [6] 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 "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 $H(\text{div}, \Omega_h)$.

Finally, we split the boundary Γ into three portions

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

We adopt a new inflow boundary condition, given by Hesthaven et al in [11], where we set

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

on Γ_-, and use the standard wall boundary condition $u = 0$ on Γ_+. For our model problem, as for many problems of interest in computational fluid dynamics, we expect ∇u to be small near the inflow, and that the solutions to (2) using $\beta_n u - \sigma_n = f_n = u_0$ on Γ_- will converge to that using $u = u_0$ on Γ_- for sufficiently small ϵ. Under these new boundary conditions, our variational problem $b(u, v) = l(v)$ is now defined by

$$b\left((u, \sigma, \widehat{u}, \widehat{\sigma}_n), (\tau, v)\right) = (u, \nabla \cdot \tau - \beta \cdot \nabla v) + \left(\sigma, \epsilon^{-1}\tau + \nabla v\right) - \langle [\tau_n], \widehat{u}\rangle_{\Gamma_- \cup \Gamma_h^0} + \langle \widehat{f_n}, [v]\rangle_{\Gamma_+ \cup \Gamma_h^0}$$
$$l\left((\tau, v)\right) = (f, v) - \langle u_0, v\rangle_{\Gamma_-}$$

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 (taking into account the respective lifts/boundary conditions on $\widehat{u}$ and $\widehat{f_n}$ on Γ_+ and Γ_-, respectively)

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

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

$$\left\| \left(u, \sigma, \widehat{u}, \widehat{f_n}\right) \right\|_U = C_1 \|u\|_{L^2(\Omega_h)} + C_2 \|\sigma\|_{L^2(\Omega_h)} + C_3 \|\widehat{u}\| + C_4 \|\widehat{f_n}\|.$$

2.2 Norms on V

As $\tau \in H(\text{div}, \Omega_h)$ 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 $H^1(K) \times H(\text{div}, K)$ norm over a single element

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

The squared norm over the entire triangulation Ω_h is defined to be the squared sum of contributions from each element

$$\| (\tau, v) \|^2_{H^1(\Omega_h)\times H(\text{div},\Omega_h)} = \sum_{K\in\Omega_h} \| (\tau, v) \|^2_{H^1(K)\times H(\text{div},K)}$$

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

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

$$\|[\tau \cdot n]\| = \|[\tau \cdot n]\|_{\Gamma_h \setminus \Gamma_+} \coloneqq \sup_{w\in H^1(\Omega),\, w|_{\Gamma_+}=0} \frac{\langle [\tau \cdot n], w\rangle}{\|w\|_{H^1(\Omega)}}$$

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

2.3 Approximability of the quasi-optimal test norm

An ideal choice for the test norm would be the quasi-optimal norm; however, the form of the optimal test norm for convection-diffusion problems turns out to be quite problematic in the small diffusion limit.

For convection-diffusion, the quasi-optimal test norm is

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

for some choice of constants C_1 and C_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 [12]. Such systems also develop strong boundary layers for small diffusion, and transforming the problem to the reference element reveals that the inversion of the inner product associated with the optimal test norm will induce strong boundary layers of width ϵ/h^2 in the optimal test functions. In comparison, the quasi-optimal norm has yielded excellent results for the Helmholtz equation and other wave propagation problems. The difference between the two problems lies in the fact that, for wave propagation problems, the mesh size tends to be on the order of the wavenumber k/singular perturbation parameter. Transforming the variational problem using the quasi-optimal test norm for wave propagation yields smooth optimal test functions that are much easier to approximate over the reference element. Typically, the wavenumbers k of physical interest are $O(100) - O(1000)$ with respect to a unit domain. The corresponding finite element problems will typically be solved on meshes containing approximately $O(k^d)$ elements in $\mathbb{R}^d$, well within the range of a computationally tractable simulation. However, for convection diffusion problems, the relevant range of ϵ for physical problems can be as small as $1e-7$. Solving on partially under-resolved meshes is thus unavoidable, and the approximability of the optimal test functions must be addressed in order to take advantage of the properties of DPG.

In the application of DPG in [5, 7, 8, 15], the approximation of optimal test functions is done using the enriched test space $V = \cup_K P^{p+\Delta p}(K)$, where p is the polynomial order of the trial space on a given element K. In other words, optimal test functions are approximated element-by-element using polynomials whose order is Δp more than the local order of approximation. Under this scheme, the approximation of optimal test functions is tied to the effectiveness of the p-method. Unfortunately, for problems with boundary layers - including the approximation of optimal test functions under the quasi-optimal test norm - the p-method performs very poorly.

Resolving such boundary layers present in quasi-optimal test functions has been investigated numerically using specially designed (Shishkin) subgrid meshes by Niemi, Collier, and Calo in [12]. However, even with Shishkin meshes, the approximation of optimal test functions under the quasi-optimal norm is difficult and far more expensive and complex to implement than approximation of test functions using a simple p-enriched space for V. We approach this from a different perspective, aiming instead to design a test norm that does not induce boundary layers, yet still maintains what we will define as *robustness* - approximation properties that do not change over a range of ϵ.

2.4 Alternatives to the quasi-optimal test norm

We expect the quasi-optimal test norm to deliver results that are close to the optimal test norm; namely, that the DPG finite element solution minimizes error a norm equivalent to

$$\left\|\left(u,\sigma,\widehat{u},\widehat{f_n}\right)\right\|_U = \|u\|_{L^2(\Omega_h)} + \|\sigma\|_{L^2(\Omega_h)} + \|\widehat{u}\| + \|\widehat{f_n}\|.$$

This is an example of a *robust* error bound - under the optimal test norm, we will return the best approximation error in the same norm *independently* of ϵ. The precise definition of robustness for singularly perturbed problems can be given as follows - for a problem with singular perturbation parameter $\epsilon \to 0$ and exact solution u, a *robust* method will return a solution u_h such that

$$\|u - u_h\|_U \lesssim \inf_{w_h \in U} \|u - w_h\|_U$$

where $\|\cdot\|_U$ is some norm and $\lesssim$ is a bound by a constant, both of which are independent of ϵ.

We expect a similar bound to hold under the quasi-optimal norm; however, it is not clear what to expect for other choices of test norms on V. In particular, it is not clear whether it is possible to achieve a completely robust bound under alternative norms, especially ones that generate easily approximated test functions.

The estimates and proofs in this paper are largely based on previous work done in [9] and [6]. Our primary contributions in this paper are stronger stability estimates for a new inflow boundary condition and a slightly different test norm.

3 Strategy for estimating the DPG energy norm

The main goal of this paper is to derive a test norm that *does not induce boundary layers in the optimal test functions*, but still induces an energy norm performs well for $\epsilon \ll 1$. The avoidance of boundary layers in the test norm will be described in more detail in the section detailing numerical experiments. Towards the goal of a well-performing energy norm, we will seek to derive ϵ-explicit bounds on the finite element error by the best approximation error, and to characterize the dependence of such bounds upon the test norm. We seek to derive an estimate

$$\left\|\left(u,\sigma,\widehat{u},\widehat{f_n}\right)\right\|_{U,1} \lesssim \left\|\left(u,\sigma,\widehat{u},\widehat{f_n}\right)\right\|_{U,2}$$

where both $\|\cdot\|_{U,1}$ and $\|\cdot\|_{U,2}$ are characterized using

$$\left\|\left(u,\sigma,\widehat{u},\widehat{f_n}\right)\right\|_{U,i} := C_1^i\|u\| + C_2^i\|\sigma\| + C_3^i\|\widehat{u}\| + C_4^i\|\widehat{f_n}\|, \quad i = 1,2$$

To derive the above ϵ-explicit bounds on the finite element error in norm $U,1$ by the best approximation error in norm $U,2$, we need to show that the DPG energy norm (in which our finite element solution is optimal), is bounded from above and below

$$\left\|\left(u,\sigma,\widehat{u},\widehat{f_n}\right)\right\|_{U,1} \lesssim \left\|\left(u,\sigma,\widehat{u},\widehat{f_n}\right)\right\|_{E} \lesssim \left\|\left(u,\sigma,\widehat{u},\widehat{f_n}\right)\right\|_{U,2}.$$

Our goal is to explicitly derive the constants $C_1^1, \dots C_4^1$ and $C_1^2, \dots C_4^2$ that define the norms $\|\cdot\|_{U,1}$ and $\|\cdot\|_{U,2}$ respectively, taking into account any dependence of such C_j^i on ϵ. Our strategy for doing so is broken down into three parts.

3.1 Equate relations between norms on U and norms on V

We begin with a given norm on U

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

By (1), the duality of test norm and energy norm, a norm of this form induces the test norm

$$\begin{aligned}
\| (v, \tau) \|_{V,U} &= \sup_{(u,\sigma,\widehat{u},\widehat{f_n}) \in U} \frac{b\left(\left(u, \sigma, \widehat{u}, \widehat{f_n} \right), (\tau, v) \right)}{C_1 \|u\| + C_2 \|\sigma\| + C_3 \epsilon \|\widehat{u}\| + C_4 \sqrt{\epsilon} \|\widehat{u}\|} \\
&= \sup_{(u,\sigma,\widehat{u},\widehat{f_n}) \in U} \frac{(u, \nabla \cdot \tau - \beta \cdot \nabla v) + \left(\sigma, \epsilon^{-1} \tau + \nabla v \right) - \langle [\tau_n], \widehat{u} \rangle_{\Gamma_- \cup \Gamma_h^0} + \langle \widehat{f_n}, [v] \rangle_{\Gamma_+ \cup \Gamma_h^0}}{C_1 \|u\| + C_2 \|\sigma\| + C_3 \epsilon \|\widehat{u}\| + C_4 \sqrt{\epsilon} \|\widehat{u}\|} \\
&\leq \frac{1}{C_1} \|g\| + \frac{1}{C_2} \|f\| + \frac{1}{C_3} \sup_{\widehat{u} \neq 0, \widehat{u}|_{\Gamma_+} = 0} \frac{\langle [\tau \cdot n], \widehat{u} \rangle}{\|\widehat{u}\|} + \frac{1}{C_4} \sup_{\widehat{f_n} \neq 0, \widehat{f_n}|_{\Gamma_-} = 0} \frac{\langle \widehat{f_n}, [v] \rangle}{\|\widehat{f}\|}
\end{aligned}$$

where f and g are defined as

$$\begin{aligned}
g &\coloneqq \nabla \cdot \tau - \beta \cdot \nabla v \\
f &\coloneqq \epsilon^{-1} \tau + \nabla v
\end{aligned}$$

which, by definition of the boundary norms on $[\tau \cdot n]$ and $[v]$, gives the characterization of the induced test norm

$$\| (v, \tau) \|_{V,U} \leq \frac{1}{C_1} \|g\| + \frac{1}{C_2} \|f\| + \frac{1}{C_3} \|[\tau \cdot n]\| + \frac{1}{C_4} \|[v]\|$$

is this an equality? Is the inequality attainable, and thus attained in the supremum? We can now use this relation to compare two different norms on U by comparing their induced norms on V - recall that showing a robust inequality between two norms on U is similarly equivalent to showing the robust reverse inequality in the induced norms on V.

3.2 Decomposition into analyzable components

After reducing the problem of comparing norms on U to comparing norms on V, we break the analysis of $(\tau, v) \in V$ into the analysis of three subproblems. Define the decomposition

$$(v, \tau) = (v_0, \tau_0) + (v_1, \tau_1) + (v_2, \tau_2),$$

where (v_1, τ_1) satisfies

$$\begin{aligned}
\epsilon^{-1} \tau_1 + \nabla v_1 &= 0 \\
\nabla \cdot \tau_1 - \beta \cdot \nabla v_1 &= \nabla \cdot \tau - \beta \cdot \nabla v = g,
\end{aligned}$$

Likewise, we define (v_2, τ_2) as satisfying

$$\begin{aligned}
\epsilon^{-1} \tau_2 + \nabla v_2 &= \epsilon^{-1} \tau + \nabla v = f \\
\nabla \cdot \tau_2 - \beta \cdot \nabla v_2 &= 0
\end{aligned}$$

We will additionally require both (τ_1, v_1) and (τ_2, v_2) to come from $H(\mathrm{div}, \Omega) \times H^1(\Omega)$, and to satisfy boundary conditions

$$\tau_i \cdot n = 0, \quad x \in \Gamma_- \tag{3}$$

$$v_i = 0, \quad x \in \Gamma_+ \tag{4}$$

for $i = 1, 2$. The selection of $H(\mathrm{div}, \Omega) \times H^1(\Omega)$ conforming test functions satisfying the specific above boundary conditions removes the contribution of the jump terms over the skeleton Γ_h in the bilinear form, allowing us to bound components of the induced test norm separately from each other.

Finally, by construction, (v_0, τ_0) must satisfy

$$\epsilon^{-1}\tau_0 + \nabla v_0 = 0$$
$$\nabla \cdot \tau_0 - \beta \cdot \nabla v_0 = 0$$

with jumps

$$[v_0] = [v], \quad x \in \Gamma_h^0$$
$$[\tau_0 \cdot n] = [\tau \cdot n], \quad x \in \Gamma_h^0.$$

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

$$v_0 = v, \quad x \in \Gamma_+ \tag{5}$$
$$\tau_0 \cdot n = \tau \cdot n, \quad x \in \Gamma_- \cup \Gamma_0. \tag{6}$$

We have now decomposed any arbitrary test function (τ, v) into a single discontinuous portion and two conforming portions. Showing the robust bound of $\| (v, \tau) \|_V$ from above and below (by norms $\|\cdot\|_{V,1}$ and $\|\cdot\|_{V,2}$ on V, respectively) reduces to showing the robust bound of each component

$$\| (\tau, v) \|_{V,1} \lesssim \| (v_0, \tau_0) \|_V, \| (v_1, \tau_1) \|_V, \| (v_2, \tau_2) \|_V \lesssim \| (\tau, v) \|_{V,2}$$

Bounding $\| (\tau_0, v_0) \|$ requires the use of techniques and decompositions first developed in [6] and adapted to convection-diffusion in [9]. However, due to the $H(\mathrm{div}, \Omega) \times H^1(\Omega)$ conforming nature of (τ_1, v_1) and (τ_2, v_2), the bound from above of test functions $\| (v_1, \tau_1) \|_V, \| (v_2, \tau_2) \|_V \lesssim \|g\|$ is reduced to proving classical error estimates for the adjoint equations

$$\epsilon^{-1}\tau + \nabla v = f \tag{7}$$
$$\nabla \cdot \tau - \beta \cdot \nabla v = g, \tag{8}$$

for any choice of data $f, g \in L^2(\Omega_h)$ and boundary conditions $\tau \cdot n|_{\Gamma_-} = 0$ and $v|_{\Gamma_+} = 0$

3.3 Deriving adjoint estimates

The final step to estimating the induced norm on U by a given localizable test norm on V is to derive adjoint stability estimates on τ and v in terms of localizable normed quantities. These normed quantities will create "building-blocks" through which we can construct test norms, where each building block will correspond roughly to estimates for $\|u\|_{L^2}$, $\|\sigma\|_{L^2}$, $\|\widehat{u}\|$, or $\|\widehat{f}_n\|$.

We introduce first the bounds derived; the proofs will be given later. For this analysis, it will be necessary to assume certain technical 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 hold

$$\nabla \times \beta = 0, \quad 0 < C \leq |\beta|^2 + \frac{1}{2}\nabla \cdot \beta, \quad C = O(1) \tag{9}$$
$$\nabla\beta + \nabla\beta^T - \nabla \cdot \beta I = O(1) \tag{10}$$
$$\nabla \cdot \beta = 0 \tag{11}$$

We also require ϵ sufficiently small in Lemma 4.

Under proper assumptions on β, we have the robust bounds

- **Lemma 3**: For β satisfying (9) and (10), $f = 0$ and $v \in H^1(\Omega_h)$, satisfying boundary conditions (5) and (6)

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

Similarly, from $\nabla \cdot \tau - \beta \cdot \nabla v = g$, we get $\|\nabla \cdot \tau\| \lesssim \|g\|$ as well.

- **Lemma 4**: For β satisfying (9), and $v \in H^1(\Omega_h)$ satisfying boundary conditions (5) and (6), and for sufficiently small ϵ

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

- **Lemma 5**: For β satisfying (9), (11), and $f = g = 0$,

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

Together, these estimates will be sufficient to construct a complete norm on V and demonstrate bounds on induced test norms for different scalings of norms on U.

4 Relations between energy and U norms

Let the norm

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

be defined as the energy norm induced by DPG under the specific test norm

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

The primary result of this paper is that, if β satisfies 9,10, and 11, the DPG method provides robust control over the L^2 error in u and σ

$$\left\|\left(u, \sigma, \widehat{u}, \widehat{f}_n\right) - \left(u_h, \sigma_h, \widehat{u}_h, \widehat{f}_{n,h}\right)\right\|_{U,1} \lesssim \inf_{w \in U_h} \left\|\left(u, \sigma, \widehat{u}, \widehat{f}_n\right) - w_h\right\|_E$$

where $\|\cdot\|_{U,1}$ is defined as

$$\left\|\left(u, \sigma, \widehat{u}, \widehat{f}_n\right)\right\|_{U,1} \coloneqq \|u\| + \|\sigma\| + \epsilon\|\widehat{u}\| + \sqrt{\epsilon}\|\widehat{f}_n\|.$$

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

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

While the second bound is robust, we have derived such a bound only by explicitly accounting for a factor of $\epsilon^{-1/2}$ in the bounding norm $\|\cdot\|_{U,2}$. This implies that for very small ϵ, the L^2 error may be much smaller than the energy error (in this sense, DPG may lose some efficiency for small ϵ). However, numerical experiments demonstrate much better results, and suggest that, at least for our example problems, $\|u - u_h\|_E \sim \|u - u_h\|_U$, implying that the induced energy norm may be equivalent to a norm on the group variable $\left(u, \sigma, \widehat{u}, \widehat{f}_n\right)$ that does not grow with ϵ.

4.1 Bound from below

We begin by using the strategy outlined in sections 3.1, 3.2, and the estimates provided in 3.3 to prove the robust bound of the L^2 error in the field variables by the energy error.

As a consequence of the duality of norms (1), we know that the norm $\|u\|_{U,1}$ is induced by a specific 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* (9) *and* (10)*, 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 (1), for f and g defined in (7) and (8),

$$\begin{aligned} f &= \epsilon^{-1}\tau + \nabla v \\ g &= \nabla \cdot \tau - \beta \cdot \nabla v, \end{aligned}$$

we can characterize the test norm for $\left\|\left(u, \sigma, \widehat{u}, \widehat{f_n}\right)\right\|_U = C_1\|u\| + C_2\|\sigma\| + C_3\|\widehat{u}\| + C_4\|\widehat{u}\|$ as

$$\begin{aligned} \|\,(v,\tau)\,\|_V &\lesssim \frac{b\left(\left(u, \sigma, \widehat{u}, \widehat{f_n}\right), (\tau, v)\right)}{C_1\|u\| + C_2\|\sigma\| + C_3\|\widehat{u}\| + C_4\|\widehat{u}\|} \\ &\simeq \frac{1}{C_1}\|g\| + \frac{1}{C_2}\|f\| + \frac{1}{C_3}\sup_{\widehat{u}\neq 0,\, \widehat{u}|_{\Gamma_+}=0} \frac{\langle [\tau\cdot n], \widehat{u}\rangle}{\|\widehat{u}\|} + \frac{1}{C_4}\sup_{\widehat{f_n}\neq 0,\, \widehat{f_n}\big|_{\Gamma_-}=0} \frac{\langle \widehat{f_n}, [v]\rangle}{\|\widehat{f}\|}, \end{aligned}$$

which, by definition of the boundary norms, is

$$\|\,(v,\tau)\,\|_V \lesssim \frac{1}{C_1}\|g\| + \frac{1}{C_2}\|f\| + \frac{1}{C_3}\|[\tau\cdot n]\| + \frac{1}{C_4}\|[v]\|$$

We will bound $\|\,(v,\tau)\,\|_V$ for all (v,τ) satisfying (12) by decomposing $(v,\tau) = (v_0,\tau_0) + (v_1,\tau_1) + (v_2,\tau_2)$ as described in section 3.2. By the triangle inequality, robustly bounding $\|\,(v,\tau)\,\|_V$ from above reduces to robustly bounding each component

$$\|\,(v_0,\tau_0)\,\|_V, \|\,(v_1,\tau_1)\,\|_V, \|\,(v_2,\tau_2)\,\|_V \lesssim \frac{1}{C_1}\|g\| + \frac{1}{C_2}\|f\| + \frac{1}{C_3}\|[\tau\cdot n]\| + \frac{1}{C_4}\|[v]\|$$

Roughly speaking, 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.

- **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 [6] 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 now have control over as well.

- **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

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

  We have $\epsilon^{-1}\tau = f - \nabla v$, so $\epsilon^{-1}\|\tau\| \lesssim \|f\| + \|\nabla v\|$. Lemma 4 implies $\|\nabla v\|^2 \lesssim \|f\|^2$, so for $\epsilon \leq 1$, we control $\epsilon^{-1/2}\|\tau\| \leq \epsilon^{-1}\|\tau\| \lesssim \|f\|$. The remaining terms can be bounded by noting that, with $g = 0$, $\|\nabla \cdot \tau_2\| = \|\beta \cdot \nabla v_2\| \lesssim \|\nabla v_2\| \lesssim \|f\|$.

With these characterizations, we can conclude that $C_1 = C_2 = 1$, $C_3 = \epsilon$, and $C_4 = \epsilon^{1/2}$. □

4.2 Bound from above

We have shown the robust bound of the norm $\|\cdot\|_{U,1}$ on U by the energy norm; for a full equivalence statement, we require a bound on the energy norm by the norm $\|\cdot\|_{U,2}$ on U. By the duality of the energy and test norm, this is equivalent to bounding the test norm from below by the test norm induced by $\|\cdot\|_{U,2}$.

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

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

where f and g are again the loads of the adjoint problem defined in (7), (8). By the triangle inequality, we have the bounds

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

We estimate the supremum by following [9]; 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

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

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

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

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 $C_2 = C_3 = C_4 = \frac{1}{\sqrt{\epsilon}}$. □

5 Proof of lemmas/stability of the adjoint problem

We reduce the adjoint problem to the scalar second order equation

$$-\epsilon \Delta v - \beta \cdot \nabla v = g - \epsilon \nabla \cdot f \tag{12}$$

with boundary conditions

$$-\epsilon \nabla v \cdot n = f \cdot n, \quad x \in \Gamma_- \tag{13}$$

$$v = 0, \quad x \in \Gamma_+ \tag{14}$$

and treat the cases $f = 0$, $g = 0$ separately. The above boundary conditions are the reduced form of boundary conditions (5) and (6) corresponding to $\tau \cdot n|_{\Gamma_-} = 0$ and $v|_{\Gamma_+} = 0$. 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 appearance of the normal trace $f \cdot n$ necessitates the use of a smooth dense subset of L^2. Formally speaking, we define the adjoint equation (12) for $f \in C^\infty(\Omega)$. In doing so, we are allowed to speak of the normal trace $f \cdot n$ on the boundary. We derive inequalities that are independent of $f \cdot n$ and $\nabla \cdot f$ (quantities that are ill-defined for $L^2(\Omega)$ functions), and note that, since smooth functions are dense in $L^2(\Omega)$, we can take any L^2 function to be the limit of $C^\infty(\Omega)$ functions. Since the inequality holds for all smooth functions and is well-defined for $L^2(\Omega)$ functions, the inequality holds for $L^2(\Omega)$ functions in the limit as well.

Lemma 3. *Assume v satisfies* (12)*, with boundary conditions* (5)*, and* (6)*, and β satisfies* (9) *and* (10)*. 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 (12) 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* (9) *holds. Then, for v satisfying equation* (12) *with boundary conditions* (5), (6) *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 [9], we substitute w into the the left hand side of equation (12), 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 the final lines of the proof of Lemma 1 in [9]; an application of Young's inequality (with δ) to the right hand side and bounds on $\|v\|, \|\nabla v\|$ by $\|w\|, \|\nabla w\|$ complete the estimate. □

Lemma 5. *Let β satisfy* (9) *and* (11)*, and let v satisfy* (12) *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 [6] and [9] in decomposing τ into $z \in H(\text{curl}, \Omega)$ 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 (9), 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. □

5.1 Comparison of boundary conditions

It is worth addressing the effect of boundary conditions on stability. Specifically, a test norm that provides stability for one set of boundary conditions may perform poorly for another set. Take, for example, the test norm defined in Section 4 and the convection-diffusion problem with Dirichlet boundary conditions.

The bilinear form for the case of Dirichlet boundary conditions is

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

The robust bounds given in Section 4 hold in arbitrary dimension; however, we can show that for the case of Dirichlet boundary conditions, the same results will not hold, even in 1D.

In 1D, the problem reduces to

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

with bilinear form

$$\begin{aligned} b\left(\left(u, \sigma, \widehat{u}, \widehat{f}\right), (\tau, v)\right) = \sum_{k=1}^{N} &\left(\widehat{u}(x)\tau(x)|_{x_{k-1}}^{x_k} + \widehat{f}(x)v(x)\Big|_{x_{k-1}}^{x_k} \right) \\ &+ \left(\sigma, \epsilon^{-1}\tau\right)_{\Omega_h} + (u, \tau')_{\Omega_h} - \left(u - \epsilon\sigma, v'\right)_{\Omega_h} \end{aligned}$$

where $\widehat{u}(x), \widehat{f}(x) \in \mathbb{R}$ are now point values defined on mesh points x_k for $k = 1, \ldots, N$, where there are N elements. We can assume without loss of generality that Ω_h is the unit interval $(0, 1)$.

Consider now the 1D analogue of the estimate given by Lemma 3, which is necessary for robust control over the error in $\|u\|_{L^2}$. In 1D, $\|\beta \cdot \nabla v\| \lesssim \|g\|$ reduces to the inequality

$$\|v'\| \lesssim \|g\|, \quad g \in L^2\Omega_h$$

The adjoint problem corresponding to Lemma 3 in section 3.3 is likewise reduced in 1D to the scalar equation

$$\epsilon v'' + v' = -g \tag{15}$$

with $v \in H_0^1((0, 1))$. Multiplying this equation by v' and integrating over Ω_h gives

$$\int_0^1 \epsilon v'' v' + \|v'\|_{L^2}^2 = -\int_0^1 g v'$$

Integrating by parts and Young's inequality give

$$-\frac{\epsilon}{2} v'^2 \Big|_0^1 + \|v'\|_{L^2}^2 \leq \frac{1}{2}\|g\| + \frac{1}{2}\|v'\|$$

implying that

$$\frac{\epsilon}{2} v'(0)^2 + \|v'\|_{L^2}^2 \lesssim \|g\|$$

Let us restrict ourselves to the cases where v is sufficiently smooth for $v'(0)$ to be well defined. Taking $g = 1$ (corresponding to a constant approximation) we can solve (15) exactly to see that

$$v(x) = \frac{e^{-\frac{x}{\epsilon}}}{e^{\frac{1}{\epsilon}} - 1} \left(e^{\frac{1}{\epsilon}} \left(e^{\frac{x}{\epsilon}} - 1 \right) + \left(e^{\frac{1}{\epsilon}} - 1 \right) e^{\frac{x}{\epsilon}} x \right)$$

Plotting the solution shows that v develops strong boundary layers of width ϵ near the inflow. Consequently, $\frac{\epsilon}{2} v'(0)^2 \approx \epsilon^{-1}$, indicating already a lack of robustness in the bound of $\|v'\|$ by $\|g\|$ for constant data[2]. The robust bound from below of g constant corresponds to the robust control of the approximation of field variable u by a constant. More detailed 1D error bounds for Dirichlet boundary conditions are provided in [8], and also indicate a lack of robustness when using the standard broken Sobolev norm as the test norm on τ and v.

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 $\nabla v \approx \epsilon \|\nabla v\|^2 \approx O(\epsilon^{-1})$. An interesting phenomenon observed is that, for sufficiently small ϵ, these issues manifest themselves in numerical experiments as additional refinements near the inflow[3]. The presence of the new boundary condition relaxes the outflow boundary condition for the adjoint/dual problem, resulting in stronger stability estimates for the adjoint, and a better robustness result for the primal problem.

[2]Unlike the case of Dirichlet boundary conditions, the inflow condition on $\widehat{f} = u(0) - \epsilon u'(0)$ induces an adjoint boundary condition $\tau(0) = 0$, or equivalently $v'(0) = 0$, removing the non-robust term from the estimate.

[3]Demkowicz and Heuer proved in [9] that for Dirichlet boundary conditions, robustness as $\epsilon \to 0$ is achieved by the test norm

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

where $\|\cdot\|_{w+\epsilon}$ is a weighted L^2 norm, where the weight $w \in (0, 1)$ is required to vanish on Γ_- and satisfy $\nabla w = O(1)$. The need for the weight is necessary to account for the loss of robustness at the inflow.

6 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. We can see this by recovering the strong form of the variational problem defining test functions. Assume first that $\tau = 0$. Then, assuming $\nabla \cdot \beta = 0$ for illustrative purposes, we have

$$\begin{aligned}((0, v), (\delta\tau, \delta v))_V &= (v, \delta v) + \epsilon (\nabla v, \nabla \delta v) + (\beta \cdot \nabla v, \beta \cdot \nabla \delta v) \\ &= (v - \epsilon\Delta v - \nabla \cdot ((\beta \otimes \beta) \nabla v), \delta v)_{L^2} + \langle \epsilon \nabla v \cdot n, \delta v \rangle + \langle n \cdot (\beta \otimes \beta) \nabla v, \delta v \rangle\end{aligned}$$

after integration by parts, recovering the strong form of the operator L inducing such a variational problem

$$Lv \coloneqq v - \epsilon\Delta v - \nabla \cdot ((\beta \otimes \beta) \nabla v)$$

The streamline direction β induces an anisotropic diffusion, while the $\epsilon\|\nabla v\|_{L^2}$ term induces a small isotropic diffusion contribution everywhere. Any vector in the cross-stream direction is in the null space of the anisotropic diffusion tensor, such that in the cross-stream directions, the optimal test function is governed only by the cross-stream part of the operator L

$$L_{\text{cross}} \coloneqq v - \epsilon\Delta v$$

and may develop boundary layers of width ϵ in those directions. To avoid boundary layers in the optimal test functions, we follow [9] 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 in 2D, such that $h_1 = h_2 = h$, and $|K| = h^2$. Our test norm for an element K is now

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

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 necessary. This leads to a test norm which produces easily approximable optimal test functions, but still provides *asymptotically* the strongest test norm and tightest robustness results in the areas of highest error.

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 experience with roundoff effects in numerical experiments [9]; 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 conditioning of 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 [7]), 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 perfect 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.

6.1 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

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

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 $x = 0$ via the formula

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

6.1.1 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.

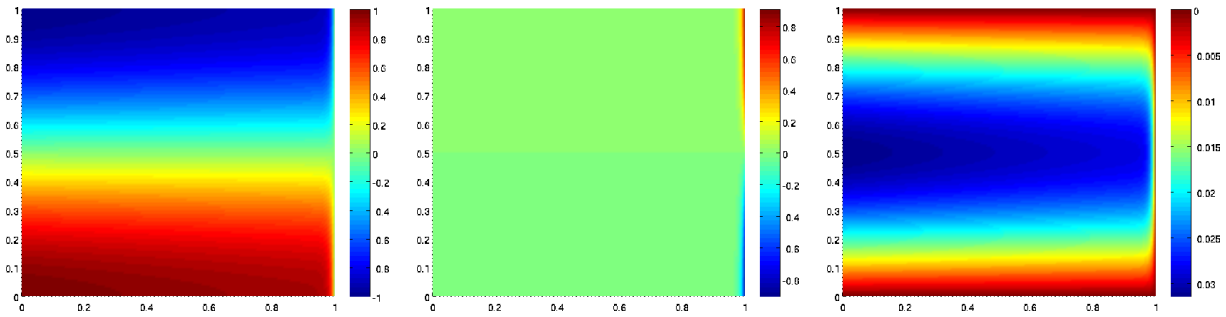

Figure 1: Solution for u, σ_x, and σ_y for $\epsilon = .01$, $C_1 = 1$, $C_n = 0$, $n \neq 1$

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.

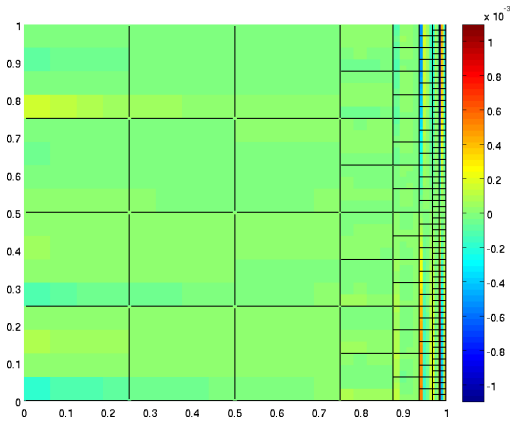

Figure 2: Adapted mesh and pointwise error for $\epsilon = .01$

We are especially interested in the ratio of energy error and total L^2 error in both σ and u, which we denote as $\|u - u_h\|_{L^2}$. 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, which plots the ratio of L^2 to energy error, seems to imply that (at least for this model problem) $C = O(1)$. Additionally, while we do not have a robust lower bound ($\|u-u_h\|_{L^2}/\|u-u_h\|_E$ can approach 0 as $\epsilon \to 0$), our numerical results appear to indicate the existence of an ϵ-independent lower bound.

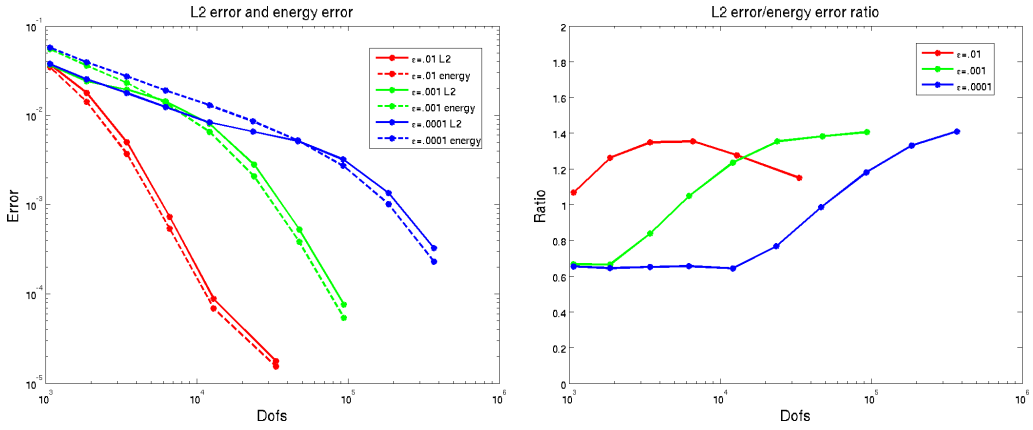

Figure 3: L^2 and energy errors, and their ratio for $\epsilon = .01$, $\epsilon = .001$, $\epsilon = .0001$

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 (providing the full robustness implied by our adjoint energy estimates), 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 whose size is of order ϵ through mesh refinement.

We examined how small ϵ needed to be in order to encounter roundoff effects as well. In [9], 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.

Without anisotropic refinements, it still becomes computationally difficult to fully resolve the solution for ϵ smaller than $1e-5$. Regardless, for all ranges of ϵ, DPG does not lose robustness, as indicted by the rates and ratio between L^2 and energy error in Figure 4 remaining bounded from both above and below. 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,

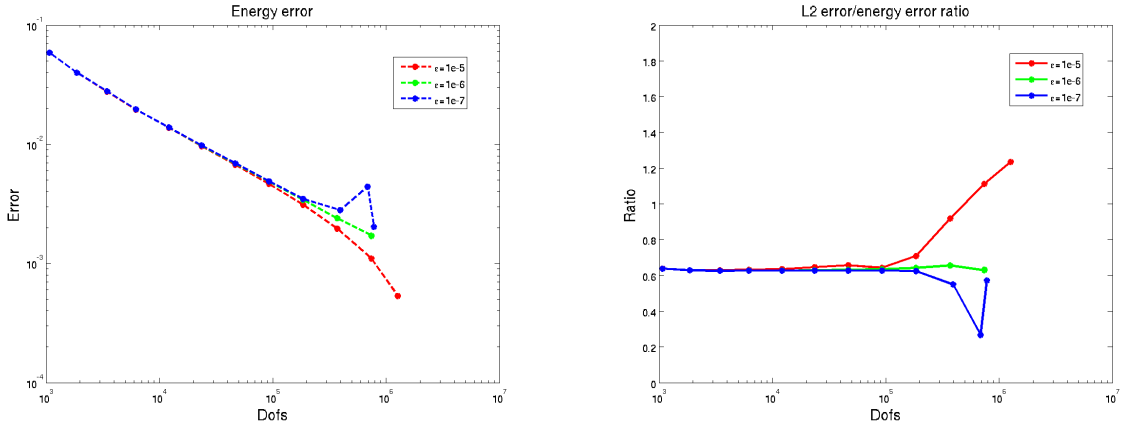

Figure 4: Energy error and L^2/energy error ratio for $\epsilon = 1e-5$, $\epsilon = 1e-6$, $\epsilon = 1e-7$. Non-monotonic behavior of the energy error indicates conditioning issues and roundoff effects.

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, possibly due to roundoff effects. For $\epsilon = 1e-6$, the ratio simply remains constant, but for $\epsilon = 1e-7$, we observe definite roundoff effects, as the energy error increases at the 11th refinement. Since DPG is optimal in the energy norm for a fixed test norm[4], 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 and conclude roundoff effects are at play.

6.1.2 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 in the inflow condition. For large $\epsilon = .01$, the L^2 error stagnates around $1e-3$, or about 7% relative error.

6.1.3 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 shows the solution u and overlaid trace variable $\widehat{u}$, which both demonstrate the regularizing effect of viscosity on the discontinuous boundary condition at $x = 0$. However, we do not have

[4]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 $\|v\|_1 \leq \|v\|_2$

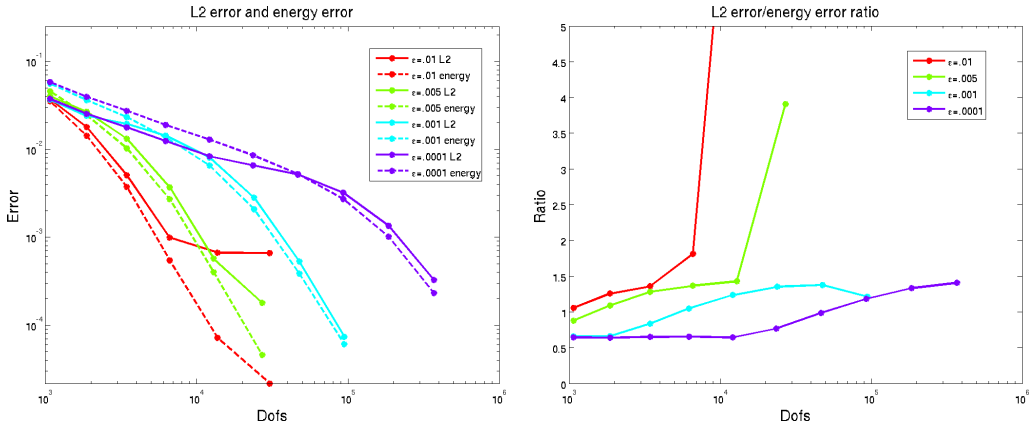

Figure 5: L^2 and energy errors and their ratio when neglecting σ_n at the inflow.

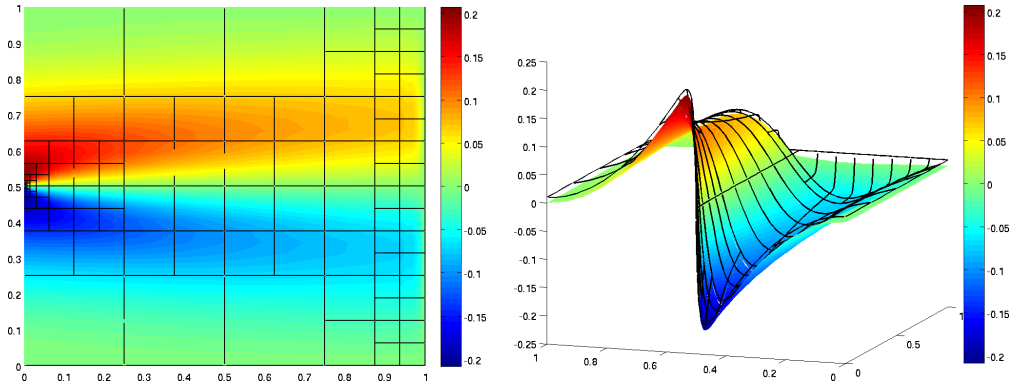

Figure 6: Solution variables u and $\widehat{u}$ with discontinuous inflow data u_0 for $\epsilon = .01$.

a closed-form solution with which to compare results for a stricly discontinuous u_0. In order to analyze convergence, we approximate u_0 with 20 terms of a Fourier series, giving a near-discontinuity for u_0.

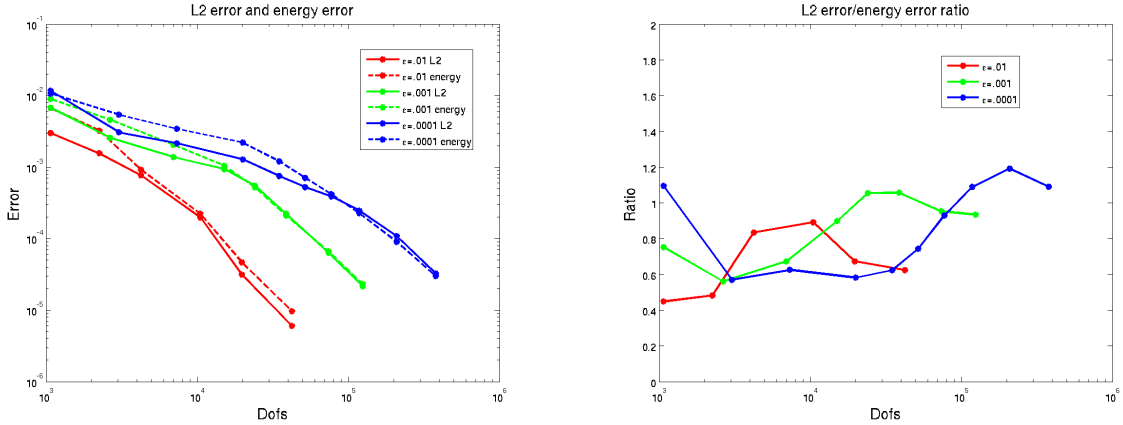

Figure 7: L^2 and energy errors, and their ratio for $\epsilon = .01$, $\epsilon = .001$, $\epsilon = .0001$, with discontinuous u_0 approximated by a Fourier expansion.

The ratios of L^2 to energy error are now less predictable than for the previous example, in part due to the difficulty in approximating highly 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. Even 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.

7 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] A.N. Brooks and T.J.R. Hughes. Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. *Comp. Meth. Appl. Mech. Engr*, 32:199–259, 1982.

[3] T. Bui-Thanh, L. Demkowicz, and O. Ghattas. A unified Discontinuous Petrov-Galerkin method and its analysis for Friedrichs' systems. Technical Report 11-34, ICES, 2011.

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

[5] 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.

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

[7] 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.

[8] L. Demkowicz, J. Gopalakrishnan, and A. Niemi. A Class of discontinuous Petrov-Galerkin methods. iii. Adaptivity. Technical Report 10-01, ICES, 2010.

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

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

[11] 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.

[12] 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.

[13] 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.

[14] H.G. Roos, M. Stynes, and L. Tobiska. *Robust numerical methods for singularly perturbed differential equations: convection-diffusion-reaction and flow problems.* Springer series in computational mathematics. Springer, 2008.

[15] 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.