A DPG method for compressible flow problems

by

Jesse Chan, B.A.

DISSERTATION PROPOSAL

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2012

Table of Contents

Chapter 1

Introduction

1.1 Motivations

Over the last three decades, Computational Fluid Dynamics (CFD) simulations have become commonplace as a tool in the engineering and design of high-speed aircraft. Wind tunnel experiments are often complemented by computational simulations, and CFD technologies have proved very useful in both the reduction of aircraft development cycles and the simulation of experimentally difficult conditions. Great advances have been made in the field since its introduction, especially in areas of meshing, computer architecture, and solution strategies. Despite this, there still exist many computational limitations in existing CFD methods:

- **Higher order methods :** Higher order methods stand to offer large computational savings through a more efficient use of discrete degrees of freedom. However, there are very few working higher-order CFD codes in existence, and most higher order methods tend to degrade to first-order accuracy near shocks. The use of higher order codes to solve the steady state equations is even rarer, where convergence of discrete nonlinear steady equations is a tricky issue [54].

- **Automatic adaptivity :** The use of adaptive meshes is crucial to many CFD applications, where the solution can exhibit very localized sharp gradients and shocks. Good resolution for such problems under uniform meshes is computationally prohibitive and impractical for most physical regimes of interest. However, the construction of "good" meshes is a difficult

task, usually requiring a-priori knowledge of the form of the solution [2]. An alternative set of strategies are automatically adaptive schemes; such methods usually begin with a coarse mesh and refine based on the minimization of some error. However, this task is difficult, as the convergence of numerical methods for problems in CFD is notoriously sensitive to mesh quality. Additionally, the use of adaptivity becomes even more difficult in the context of higher order and *hp* methods [54].

Both of these issues are tied to the notion of *robustness*. We define robustness loosely as the degradation of the quality of numerical solutions with respect to a given problem parameter. In the context of CFD simulations, the parameter of interest is the Reynolds number (the nondimensional equivalent of the inverse of the viscosity) — for typical physical conditions of interest for the compressible Navier-Stokes equations, the Reynolds number is extremely high, on the order of 1e7, yielding solutions with two vastly different scales - inviscid phenomena at an $O(1)$ scale, and $O(1e-7)$ viscous phenomena.

The full Navier-Stokes equations are not well understood in a mathematical sense — in order to more clearly illustrate the issue of robustness for problems in CFD, we will study first the important model problem of convection-dominated diffusion.

1.1.1 Singular perturbation problems and robustness

Standard numerical methods tend to perform poorly across the board for the class of PDEs known as singular perturbation problems; these problems are often characterized by a parameter that may be either very small or very large. An additional complication of singular perturbation problems is that very often, in the limiting case of the parameter blowing up or decreasing to zero, the PDE itself will change types (e.g. from elliptic to hyperbolic). A canonical example of a singularly

perturbed problem is the convection-diffusion equation in domain $\Omega \in \mathbb{R}^3$,

$$\nabla \cdot (\beta u) - \epsilon \Delta u = f.$$

The equation models the steady-state distribution of the scalar quantity u, representing the concentration of a quantity in a given medium, taking into account both convective and diffusive effects. Vector $\beta \in \mathbb{R}^3$ specifies the direction and magnitude of convection, while the singular perturbation parameter ϵ represents the diffusivity of the medium. In the limit of an inviscid medium as $\epsilon \to 0$, the equation changes types, from elliptic to hyperbolic, and from second order to first order.

We will illustrate the issues associated with numerical methods for this equation using one dimensional examples. In 1D, the convection-diffusion equation is

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

For Dirichlet boundary conditions $u(0) = u_0$ and $u(1) = u_1$, the solution can develop sharp boundary layers of width ϵ near the outflow boundary $x = 1$.

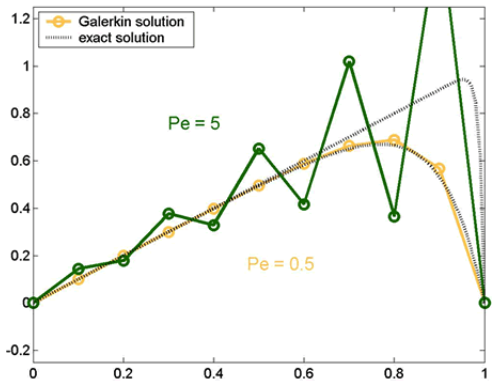


Figure 1.1: Oscillations in the 1D finite element solution of the convection-diffusion equation for small diffusion [33]. Standard finite volume and finite difference methods exhibit similar behavior.

We consider for now the Galerkin finite element method as applied to convection-dominated diffusion. Standard finite element methods (as well as standard finite volume and finite difference

methods) perform poorly for the case of small ϵ. The poor performance of the finite element method for this problem is reflected in the bound on the error in the finite element solution — under the standard Bubnov-Galerkin method with $u \in H^1(0,1)$, we have the bound given in [52]:

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

for $\|u\|_\epsilon^2 \coloneqq \|u\|_{L^2}^2 + \epsilon \|u'\|_{L^2}^2$, with C independent of ϵ. 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 of the constant C on ϵ is what we refer to as a *loss of robustness* — as the singular perturbation parameter ϵ decreases, our finite element error is bounded more and more loosely by the best approximation error. As a consequence, the finite element solution can diverge significantly from the best finite element approximation of the solution for very small values of ϵ. For example, Figure 1.1 shows an example of how, on a coarse mesh, and for small values of ϵ, the Galerkin approximation of the solution to the convection-diffusion equation with a boundary layer develops spurious oscillations everywhere in the domain, even where the best approximation error is small. These oscillations grow in magnitude as $\epsilon \to 0$, eventually polluting the entire solution.[1]

1.2 Goal

From the perspective of the full Navier-Stokes equations, this loss of robustness is doubly problematic. Not only will any nonlinear solution suffer from similar non-physical oscillations, but

[1] For nonlinear shock problems, the solution often exhibits sharp gradients or discontinuities, around which the solution would develop spurious Gibbs-type oscillations. These are a result of underresolution of the solution, and are separate from the oscillations resulting from a lack of robustness.

nonlinear solvers themselves may fail to yield a solution due to such instabilities. A nonlinear solution is almost always computed by solving a series of linear problems whose solutions will converge to the nonlinear solution under appropriate assumptions, and the presence of such oscillations in each linear problem can cause the solution convergence to slow significantly or even diverge.

Our aim is to develop a *stable, higher order adaptive scheme* for the steady compressible laminar Navier-Stokes equations in transonic/supersonic regimes that is *robust for very small viscosity*. In particular, we hope to present a method for which *hp*-adaptivity can be applied to problems in compressible flow. The goal of this dissertation will be to develop a mathematical theory demonstrating the robustness in ϵ of our method for singularly perturbed convection-diffusion problems, and to demonstrate its feasibility as a CFD solver by applying it to several benchmark problems.

1.3 Literature review

For the past half-century, problems in CFD have been solved using a multitude of methods, many of which are physically motivated, and thus applicable only to a small number of problems and geometries. We consider more general methods, whose framework is applicable to a larger set of problems; however, our specific focus will be on the problems of compressible aerodynamics involving small-scale viscous phenomena (i.e. boundary layers and, if present, shock waves). Broadly speaking, the most popular general methods include (in historical order) finite difference methods, finite volume methods, and finite element methods.

1.3.1 Finite difference and finite volume methods

For linear problems, finite difference (FD) methods approximate derivatives based on interpolation of pointwise values of a function. In the context of conservation laws, FD methods were popularized first by Lax, who introduced the concepts of the monotone scheme and numerical flux.

For the conservation laws governing compressible aerodynamics, FD methods approximate the conservation law, using some numerical flux to reconstruct approximations to the derivative at a point. Finite volume (FV) methods are similar to finite difference methods, but approximate the integral version of a conservation law as opposed to the differential form. FD and FV have roughly the same computational cost/complexity; however, the advantage of FV methods over FD is that FV methods can be used on a much larger class of problems and geometries than FD methods, which require uniform or smooth structured meshes.

Several ideas were introduced to deal with oscillations in the solution near a sharp gradient or shock: artificial diffusion, total variation diminishing (TVD) schemes, and slope limiters. However, each method had its drawback, either in terms of loss of accuracy, dimensional limitations, or problem-specific parameters to be tuned [53]. Harten, Enquist, Osher and Chakravarthy introduced the essentially non-oscillatory (ENO) scheme in 1987 [38], which was improved upon with the weighted essentially non-oscillatory (WENO) scheme in [44]. WENO remains a popular choice today for both finite volume and finite difference schemes. Most of these methods can be interpreted as adding some specific artificial diffusion to the given numerical scheme. We refer to such schemes as *modified equation* methods, as the exact solution no longer satisfies the discrete system due to the presence of additional artificial diffusion terms.

Historically, finite volumes and finite difference methods have been the numerical discretizations of choice for CFD applications; the simplicity of implementation of the finite difference method allows for quick turnaround time, and the finite volume method is appealing due to its locally conservative nature and flexibility. More recently, the finite element (FE) method has gained popularity as a discretization method for CFD applications for its stability properties and rigorous mathematical foundations. Early pioneers of the finite element method for CFD included Zienkiewicz, Oden,

Karniadakis, and Hughes [19].

1.3.2 Stabilized finite element methods

The finite element/Galerkin method has been widely utilized in engineering to solve partial differential equations governing the behavior of physical phenomena in engineering problems. The method relates the solution of a partial differential equation (PDE) to the solution of a corresponding variational problem. The finite element method itself provides several advantages — a framework for systematic mathematical analysis of the behavior of the method, weaker regularity constraints on the solution than implied by the strong form of the equations, and applicability to very general physical domains and geometries for arbitrary orders of approximation.

Historically, the Galerkin method has been very successfully applied to a broad range of problems in solid mechanics, for which the variational problems resulting from the PDE are often symmetric and coercive (positive-definite). It is well known that the finite element method produces optimal or near-optimal results for such problems, with the finite element solution matching or coming close to the best approximation of the solution in the finite element space. However, standard Galerkin methods tend to perform poorly for singular perturbation problems, developing instabilities when the singular perturbation parameter is very small.

Traditionally, instability/loss of robustness in finite element methods has been dealt with using residual-based stabilization techniques. Given some variational form, the problem is modified by adding to the bilinear form the strong form of the residual, weighted by a test function and scaled by a stabilization constant τ. The most well-known example of this technique is the streamline-upwind Petrov-Galerkin (SUPG) method, which is a stabilized FE method for solving the convection-diffusion equation using piecewise linear continuous finite elements [12]. SUPG stabilization not only

removes the spurious oscillations from the finite element solution of the convection-diffusion equation, but delivers the best finite element approximation in the H^1 norm.

1.3.2.1 SUPG

All Galerkin methods involve both trial (approximating) and test (weighting) functions. Standard Galerkin methods, where these trial and test functions are taken from the same space, are referred to as Bubnov-Galerkin methods. Petrov-Galerkin methods refer most often to methods where test and trial functions *differ*, leading to differing test and trial spaces.[2] The Streamline Upwind Petrov Galerkin (SUPG) method is a stabilization method for H^1-conforming finite elements, the idea of which was originally motivated by artificial diffusion techniques in finite differences. In particular, for the homogeneous 1D convection-diffusion equation, it is possible to recover, under a finite difference method, the exact solution at nodal points by adding an "exact" artificial diffusion based on the mesh size h and the magnitudes of the convection β and the viscosity ϵ. The idea of "exact" artificial viscosity was adapted to finite elements not through the direct modification of the equations, but through the *test* functions and weighting of the residual. [3]

We will introduce the SUPG method at the abstract level for illustrative purposes only. Further details and perspectives on the SUPG method can be found in [12], as well as in an upcoming book by Hughes. The convection-diffusion equation can be written as follows:

$$Lu = (L_{\text{adv}} + L_{\text{diff}})\, u = f,$$

where $L_{\text{adv}} u \coloneqq \nabla \cdot (\beta u)$ is the first order advective operator, and $L_{\text{diff}} u \coloneqq \epsilon \Delta u$ is the second-

[2] Hughes takes the more general definition of a Petrov-Galerkin method to be any Galerkin method other than a classical Bubnov-Galerkin method.

[3] Finite element and Galerkin methods are often referred to as "weighted residual" methods, since the starting point of both is to multiply the residual by a particular test, or weighting, function. Standard Bubnov-Galerkin methods simply choose these weighting functions to be the same as the the basis functions used to approximate the solution.

order diffusive operator. Let us assume u to be a linear combination of piecewise-linear basis functions $\phi_i, i = 0, \ldots, N$ (then, within each element, $L_{\text{diff}} u = 0$). If $b(u, v)$ and $l(v)$ are the bilinear form and load for the standard Galerkin method (resulting from multiplying by a test function v and integrating both convective and diffusion terms by parts), the SUPG method is then to solve $b_{\text{SUPG}}(u, v) = l_{\text{SUPG}}(v)$, where $b_{\text{SUPG}}(u, v)$ and $l_{\text{SUPG}}(v)$ are defined as

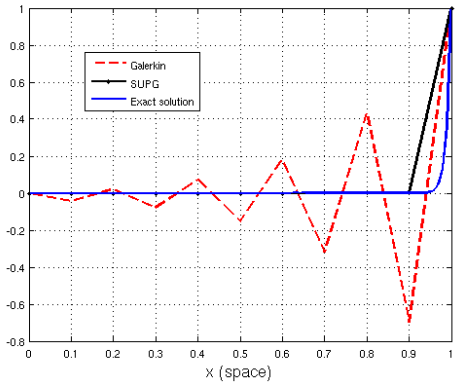
$$b_{\text{SUPG}}(u, v) = b(u, v) + \sum_K \int_K \tau \left(L_{\text{adv}} v\right) \left(Lu - f\right)$$
$$l_{\text{SUPG}}(v) = l(v) + \sum_K \int_K \tau \left(L_{\text{adv}} v\right) f$$

for where τ is the SUPG parameter. For uniform meshes in 1D, τ is chosen such that, for $f = 0$, the matrix system resulting from SUPG is exactly equal to the finite difference system under "exact" artificial diffusion. However, unlike exact artificial diffusion, for $f \neq 0$, the SUPG method still delivers optimal stabilization. In fact, the SUPG finite element solution in 1D is nothing less than the nodal interpolant and the best H_0^1 approximation of the exact solution, as seen in Figure 1.2.

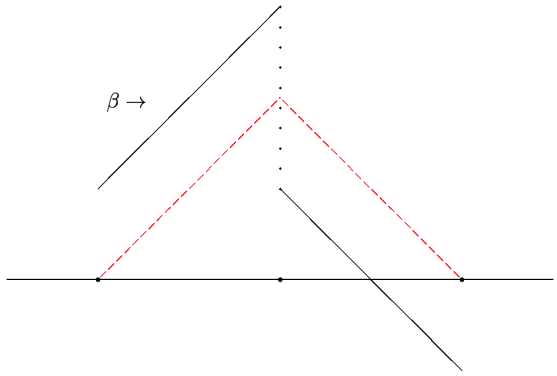
The idea of emphasizing the upwind portion of a test function is an old idea, introduced in 1977 by Zienkiewicz et al. in [39]. However, the precise amount of upwinding,[4] as well as the connection to residual-based stabilization methods, were novel to SUPG.

For appropriately chosen τ, the method can be generalized for higher order elements as well. In higher dimensions, and the SUPG solution is very close to, but no longer the H_0^1 best approximation for 2D and 3D problems [42]. Since its inception, SUPG is and has been the most popular stabilization method of choice for convection-diffusion type problems, in both academic and industry applications.

[4] Insufficient upwinding results in a method which still exhibits oscillations and instabilities, while excessive upwinding leads to an overly diffusive method.



(a) SUPG and standard FEM solutions



(b) SUPG test function

Figure 1.2: SUPG and standard Bubnov-Galerkin solutions to the 1D convection-dominated diffusion equation, and a modified SUPG test function (in black) corresponding to a linear basis "hat" function (in red). The upwind portion of the element is emphasized, while the downwind portion is decreased. The magnitude of the discontinuity between the upwind and downwind portion is controlled by the intrinsic timescale parameter τ.

An important feature of SUPG and other residual-based stabilization techniques that separates it from modified equation methods 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). Contrast this to the artificial diffusion methods in finite difference and finite volume methods, where a specific amount of additional viscosity is added based on the magnitude of the convection and diffusion parameters: unlike residual-based stabilization schemes, the exact solution to the original equation no longer satisfies the new stabilized formulation.

This addition of residual-based stabilization terms can be interpreted as a modification of the test functions as well. For SUPG, the formulation can equivalently be written as

$$b\left(u, \tilde{v}_i\right) = l\left(\tilde{v}_i\right), \quad \forall i = 1, \ldots, N-1$$

where the SUPG test function $\tilde{v}_i$ is defined elementwise as

$$\tilde{v}_i = \phi_i(x) + \tau L_{\text{adv}} \phi_i.$$

In other words, the test functions $\tilde{v}_i$ is a perturbation of the basis function ϕ_i by a scaled advective operator applied to ϕ_i. For a linear C^0 basis function (the "hat" function), this naturally leads to a bias in the upwind or streamline direction of the flow β, as seen in Figure 1.2.

An important connection can now be made — stabilization can be achieved by changing the test space for a given problem. We will discuss in Section 3.0.1 approaching the idea of stabilization through the construction of *optimal test functions* to achieve optimal approximation properties.

1.3.2.2 DG methods

Discontinuous Galerkin (DG) methods form a subclass of FEM; first introduced by Reed and Hill in [49], these methods were later analyzed by Cockburn and Shu [22] and have rapidly gained popularity for CFD problems. Advantages of DG methods include the local conservation property, easily modified local orders of approximation, ease of adaptivity in both h and p, and efficient parallelizability. Rather than having a continuous basis where the basis function support spans multiple element cells, DG opts instead for a discontinuous, piecewise polynomial basis, where, like FV schemes, a *numerical flux* facilitates communication between neighboring elements (unlike FV methods, however, there is no need for a reconstruction step).

The formal definition of the numerical flux (attributed to Peter Lax) on an element boundary is some function of the values on the edges of both the neighboring elements. An additional reason for the popularity of DG methods is that they can be interpreted as stabilized FE methods through appropriate choices of this numerical flux [11]. We will illustrate this with the steady convection

equation in 1D:

$$\frac{\partial\left(\beta(x)u\right)}{\partial x} = f, \quad u(0) = u_0.$$

The DG formulation is derived by multiplying by a test function v with support only on a single element $K = [x_K, x_{K+1}]$ and integrating by parts. The boundary term is left alone, such that the local formulation is

$$\beta uv|_{x_K}^{x_{K+1}} + \int_K -\beta u \frac{\partial v}{\partial x} = \int_K fv,$$

and the global formulation is recovered by summing up all element-wise local formulations. However, the boundary term in the local formulation is presently ill-defined, as both u and v are dual-valued over element boundaries. Consequently, we make the choice to define the values of u on the boundary (the *traces* of u) as

$$u(x_K) \coloneqq u(x_K^-), \quad u(x_{K+1}) \coloneqq u(x_{K+1}^-),$$

where $u(x_K^-)$ is the value of u at x_K as seen from the left, and $u(x_K^+)$ the value as seen from the right. Similarly, the traces of v are defined to be

$$v(x_K) \coloneqq v(x_K^+), \quad v(x_{K+1}) \coloneqq v(x_{K+1}^-),$$

For β positive, $v(x_K^+)$ is the *upwind* value of $v(x_K)$, and we refer to DG under this specific choice of traces as upwind DG. This specific choice of $v(x_K)$ as the upwind value is crucial; similarly to SUPG, the upwind DG emphasizes the test function in the direction of convection and changes the way the residual is measured. As it turns out, the performance of DG for convection-type problems is closely tied to this upwinding — choosing the value of $v(x_K)$ to be the downwind value $v(x_K^-)$ leads to an unstable method, while choosing $v(x_K)$ to be the average of the upwind and downwind

values leads to a DG method with suboptimal stability properties, similar to an H^1-conforming continuous Galerkin approximation[11].[5]

Another perspective on the use of the numerical flux in DG methods is that the selection of specific DG fluxes imparts *additional regularity* where needed. For example, for the pure convection problem, the solution has a distributional derivative in the streamline direction, but is only L^2 in the crosswind direction. As a consequence of the regularity of the solution, the boundary trace of the solution is defined only in the direction of convection. The upwind DG method addresses the above issue by choosing the numerical flux to be the upwind flux; in this case, the DG numerical flux can be viewed as imparting additional regularity to the discrete solution than is implied by the continuous setting [24, 28].

1.3.2.3 HDG

A more recent development in DG methods is the idea of *hybridized* DG (HDG), introduced by Gopalakrishnan and Lazarov [20]. The hybridized DG framework identifies degrees of freedom with support only on element edges, which can be interpreted as Lagrange multipliers enforcing weak continuity of the trial space. HDG methods treat numerical *traces* and numerical *fluxes* differently depending on the form of the boundary term resulting from integration by parts. The numerical trace (the result of integrating by parts the gradient) in HDG methods is chosen to be an unknown, while the numerical flux (the result of integrating by parts the divergence) is chosen to be an appropriate function of both function values on neighboring elements and the numerical trace.

By a careful choice of the numerical flux, the global HDG formulation can be reduced to a single equation involving only the numerical trace degrees of freedom, referred to as the *global*

[5] For second-order convection-diffusion problems with small diffusion, the additional regularity imparted by choice of the DG numerical flux is often insufficient, and SUPG-type stabilization is also applied.

problem. Once the global problem is solved, interior degrees of freedom can be recovered in parallel through so-called *local* problems [21].

HDG methods are an active topic of current research, since they address several criticisms of common DG methods (large number of globally coupled degrees of freedom, complicated/inefficient implementation procedures, suboptimal convergence of approximate fluxes). Note that HDG methods still fall under the category of stabilized methods — stabilization techniques are employed through the choice of the HDG numerical flux, which involves some stabilization parameter τ.

1.4 Scope

This PhD thesis proposal will proceed in four main parts. We will begin by introducing the abstract Discontinuous Petrov-Galerkin (DPG) method as a minimum residual method for linear problems and highlighting some important properties of the method. Our next step will be to formulate and prove the robustness of a DPG method (with respect to ϵ) for the model problem of convection-dominated diffusion. Finally, we will extend and apply the DPG method to singularly perturbed nonlinear problems in CFD, presenting preliminary results for the Burgers and compressible Navier-Stokes equations.

Chapter 2

Range of problems

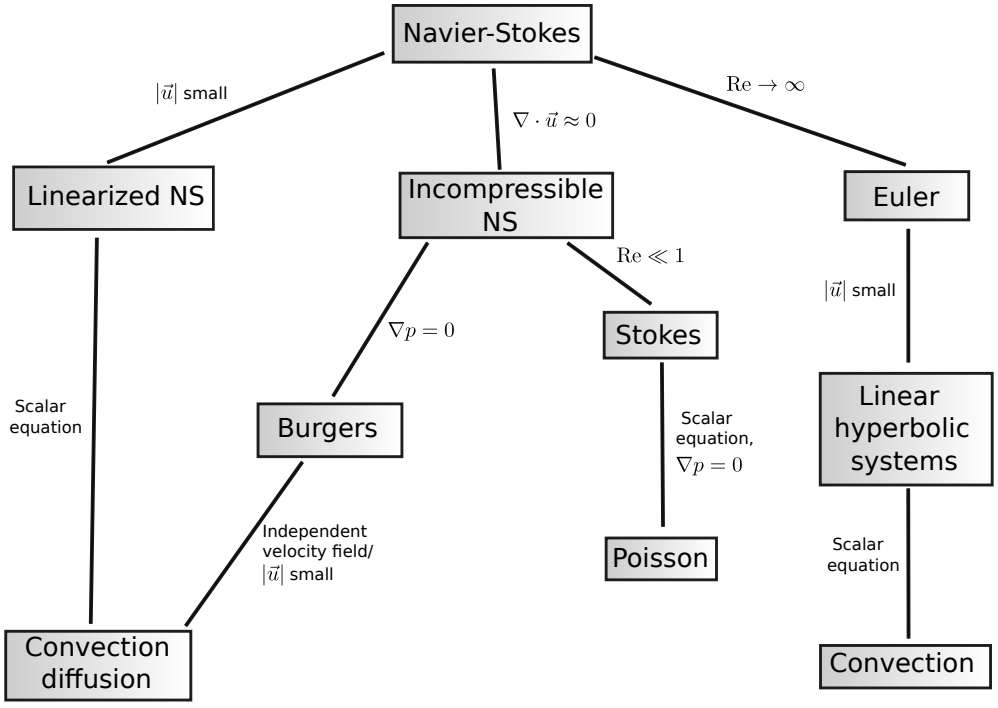


Figure 2.1: A diagram of common CFD problems and their simplifying assumptions.

2.1 The compressible Navier-Stokes equations

We consider the transient compressible Navier-Stokes equations. For simplicity, we present them in two spatial dimensions. Each equation of the Navier-Stokes system represents the conser-

vation of some physical quantity in the behavior of a fluid inside a general control volume.[1]

In 2D, the classical form of the Navier-Stokes equations involve the fluid density ρ, velocity in the x and y directions u and v (or u_1 and u_2), respectively, temperature T, energy per unit mass e, and stress and heat flux vectors $\boldsymbol{\sigma}_i$ and $\vec{q}$. The equations are as follows:

- **Mass conservation**

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \begin{bmatrix} \rho u \\ \rho v \end{bmatrix} = 0$$

- **Momentum conservation**

$$\frac{\partial \rho u_1}{\partial t} + \nabla \cdot \left(\begin{bmatrix} \rho u^2 + p \\ \rho uv \end{bmatrix} - \boldsymbol{\sigma}_1 \right) = 0$$
$$\frac{\partial \rho u_2}{\partial t} + \nabla \cdot \left(\begin{bmatrix} \rho uv \\ \rho v^2 + p \end{bmatrix} - \boldsymbol{\sigma}_2 \right) = 0$$

- **Energy conservation**

$$\frac{\partial \rho e}{\partial t} + \nabla \cdot \left(\begin{bmatrix} ((\rho e) + p)u \\ ((\rho e) + p)v \end{bmatrix} - \boldsymbol{\sigma}_1 \cdot \boldsymbol{u} - \boldsymbol{\sigma}_2 \cdot \boldsymbol{u} + \vec{q} \right) = 0$$

We assume our fluid satisfies standard stress laws for $\boldsymbol{\sigma}$ and $\boldsymbol{q}$ as well. For viscous stresses $\boldsymbol{\sigma}$, we assume a Newtonian fluid

$$\sigma_{ij} = \mu(u_{i,j} + u_{j,i}) + \lambda u_{k,k} \delta_{ij}.$$

The coefficients λ and μ are the viscosity and bulk viscosity, respectively. The bulk viscosity is often set implicitly through $2\mu + 3\lambda = 0$, known as Stokes' hypothesis. However, since the effect of

[1] The derivation of the compressible Navier-Stokes equations is a standard result of the Reynolds transport theorem, and can be found in many elementary fluid dynamics books. See [34] for one example.

bulk viscosity can become important for compressible flows, we treat both coefficients separately. In general, μ and λ are functions of temperature, obeying the power law

$$\mu = \left(\frac{T}{T_0}\right)^{\beta},$$

where T_0 is a reference temperature. We choose $\beta = 2/3$ in this case.

We assume our fluid satisfies Fourier's law, which relates the heat flux $\boldsymbol{q}$ to the gradient of the temperature through

$$\boldsymbol{q} = \kappa \nabla T,$$

where κ, the coefficient of heat conductivity, is generally a function of temperature.

Finally, we assume our fluid is a thermally and calorically perfect ideal gas. Let c_p and c_v be the specific heats at constant pressure and volume, respectively. Then,

$$\begin{aligned} p &= (\gamma - 1)\rho\iota \\ \iota &= e - \frac{1}{2}(u_1^2 + u_2^2) \\ \iota &= c_v T \end{aligned}$$

where e and ι are energy and internal energy per unit mass, respectively.

As mentioned before, the compressible Navier-Stokes equations are especially of interest in the simulation of high-speed air flows. In other contexts, however, the compressible Navier-Stokes equations may be simplified based on physical assumptions about the problem at hand. We briefly cover several simplifying assumptions common in CFD applications.

2.1.1 Incompressibility

Under appropriate assumptions on the behavior of density and temperature, the behavior of the compressible Navier-Stokes equations can be sufficiently represented by the incompressible

Navier-Stokes equations for some fluid flows. For example, the incompressible Navier-Stokes equations accurately model nearly incompressible mediums such as water, as well as low Mach number flows of compressible fluids. The study of the incompressible Navier-Stokes equations is an open area in mathematics, and is one of the most famous Millenium Problems posed by the Clay Mathematics Institute. The equations of incompressible flow pose a difficult problem computationally as well, in part due to the problem of the simulation of turbulent phenomena.

For highly viscous "creeping" flows, the incompressible Navier-Stokes equations reduce down to the Stokes equations. We remark that determining good finite element spaces for the Stokes problem is still an active area of research. [6] lists several choices of finite element discretizations suitable for the Stokes equation.

The scope of this dissertation will not deal with these two equations — the Stokes equations are treated in [50], and the incompressible Navier-Stokes are covered in an upcoming dissertation.

2.1.2 The linearized Navier-Stokes equations

The linearized Navier-Stokes equations are the result of small perturbation assumptions applied to the full Navier-Stokes equations. Under such assumptions, the flow in a domain consists only of slight variations (to a given background flow) that are small compared to the magnitude of the free stream velocity. Mathematically speaking, the linearized Navier-Stokes equations are the results of the linearization of the full equations with respect to a specific background flow. The linearized NS equations are frequently referred to as an "acoustic problem", as they are used to study sound propagation.

We are interested in the linearized Navier-Stokes equations mainly for mathematical purposes - as the solution to the full Navier-Stokes equations involves a series of solutions for linearized

Navier-Stokes, we wish to investigate the behavior of our numerical method with respect to this system.

2.2 The scalar convection-diffusion equation

Recall that the scalar convection-diffusion equation models mathematically the distribution of the concentration u of a substance in a medium due to both convective and diffusive effects. Scalar convection-diffusion has significant historical importance, as it is the prototypical model problem for solving the full Navier-Stokes equations — most stabilized methods consider first the scalar convection-diffusion equation as a test case before attempting a solution of the full Navier-Stokes equations. As discussed previously, an important feature of the convection-diffusion equation is that solutions can develop boundary layers whose thickness depends on the viscosity, a physical feature found in most applications of interest for compressible flow.

2.2.1 Burgers' equation

The Burgers' equation is physically derived from the incompressible Navier-Stokes equations under the assumption that $\nabla p \approx 0$, or that the pressure field is near constant. A feature of the Burgers' equation not present in convection-diffusion is that, due to the presence of the nonlinear term, it can develop shock discontinuities in its solutions in finite time. The Burgers' equation has also been used to study the phenomenon of turbulence; however, the Burgers equation does not exhibit the chaotic nature and sensitivity to initial conditions that characterizes turbulence as observed in the full and incompressible Navier-Stokes equations.

The Burgers' equation is also the simplest nonlinear extension of the linear convection-diffusion equation, and exact solutions can sometimes be found using the method of characteristics. In the scope of this dissertation proposal, Burgers shall be used as to test the extension of our

numerical method to nonlinear problems.

2.3 The inviscid case

The pure convection equation is a result of neglecting the viscous term in the convection-diffusion equation. Physically speaking, these assumptions correspond to the inviscid limit, as well as a particular class of boundary conditions (for example, a prescribed inflow condition may be incompatible with the wall boundary condition $u = 0$ in the inviscid limit). The Euler equations are likewise a result of neglecting the viscous terms in the Navier-Stokes equations. However, these problems can be ill-posed in the continuous setting. Take, for example, the vortex problem in Figure 2.2. A feature of the convection equation is that there is no crosswind diffusion - thus, materials do not mix across streamlines. However, for the vortex problem, this also implies that the solution on any closed streamline can take any arbitrary value, and is thus undefined.

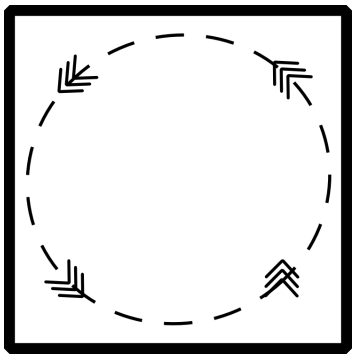


Figure 2.2: Setup for the vortex problem.

Formally speaking, the solution to the vortex problem is taken to be the solution to the convection-diffusion equation (with appropriate outflow boundary conditions) as the viscosity tends towards zero, in which case, the solution in the interior would be uniformly zero (this technique is referred to in mathematical literature as the "vanishing viscosity" method, and is used to define

unique solutions in the inviscid limit). This motivates the need for *artificial viscosity* methods with which to regularize inviscid solutions. The topic is expansive, and we direct the reader towards [4] for a more detailed discussion of past and present artificial viscosity methods.

The full Navier-Stokes models have proven difficult to solve due to the mathematical nature of the equations — due to the lack of robustness of most methods, solving the Navier-Stokes for high Reynolds numbers requires very fine meshes and is an incredibly expensive task. Additionally, the problem of turbulence for high Reynolds numbers further complicates the Navier-Stokes solutions for high speed compressible flow. Without turbulence models, turbulent effects can prevent convergence to a solution. However, common turbulence models, such as Reynolds Averaged Navier-Stokes (RANS), can lead to nonphysical solutions, such as the existence of a steady-state solution when there is none.

In comparison, the coupling of the inviscid Euler equations with boundary layer models has been successful in simulating many phenomena in compressible flow at a computational cost orders of magnitude below that of the full Navier-Stokes equations [5]. The method has been extended to a wide array of physical conditions, and is an active area of current research in both industry and academia.

Chapter 3

Discontinuous Petrov-Galerkin: a minimum residual method for linear problems

The Discontinuous Petrov-Galerkin (DPG) method of Demkowicz and Gopalakrishnan was first formulated in [24] as a scheme for the pure convection problem. The method demonstrated promise — in particular, the method was able to achieve optimal convergence rates for the Peterson problem where standard DG methods were suboptimal. A breakthrough came in [26], where the concept of locally-computable optimal test functions led to the development of the DPG method in its current form. Soon after, the DPG method was successfully applied to solve the convection-diffusion problem in the small-diffusion limit with $\epsilon = 1e-7$ [26, 28].

Historically, the name DPG was given by Bottasso, Micheletti, Sacco and Causin to their method for elliptic problems in [7]. The method was then extended to other problems, including convection-diffusion, in [8, 16, 17]. The key point in the method of Bottasso et al. was their method of hybridization of fluxes. Whereas HDG methods identify an additional flux unknown on the boundary, the numerical flux coupling neighboring elements together is still typically computed in part by using contributions from interior degrees of freedom on neighboring elements. In the DPG formulation, all numerical fluxes are declared to be independent unknowns, leaving the interior field degrees of freedom completely uncoupled from element to element. This formulation is often referred to as the *ultra-weak* variational formulation.

Since its inception, the DPG method has been applied to a wide range of problems with

great success, including elasticity [10], thin body shell problems [47], the cloaking problem in electromagnetics [30], both Helmholtz and elastic wave propagation problems [55], and recently, the linear Stokes equation [50]. DPG has also been shown to be a generalization of several successful finite element methods — comparisons between DPG and standard DG methods can be found in [24], and more recently, several existing DG methods have been shown to be derivable using the DPG framework [14, 15].

Detailed analysis of the energy setting and well-posedness of DPG has been done for the Poisson and convection-diffusion equations in [25]. The well-posedness of DPG has also recently been extended to the large class of Friedrichs systems in [13]. Significant efforts have also been made in demonstrating the robustness of the DPG method for singular perturbation problems, both in wave propagation [55, 27] and convection-diffusion problems [29, 18].

3.0.1 Optimal Petrov-Galerkin methods

Petrov-Galerkin methods, in which the test space differs from the trial space, have been explored for over 30 years, beginning with the approximate symmetrization method of Barrett and Morton [3]. The idea was continued with the SUPG method of Hughes, and the characteristic Petrov-Galerkin approach of Demkowicz and Oden [32], which introduced the idea of tailoring the test space to change the norm in which a finite element method would converge.

The idea of optimal test functions was introduced by Demkowicz and Gopalakrishnan in [26]. Conceptually, these optimal test functions are the natural result of the minimization of a residual corresponding to the operator form of a variational equation. The connection between stabilization and least squares/minimum residual methods has been observed previously [41]. However, the method in [26] distinguishes itself by measuring the residual of the natural *operator form of the*

equation, which is posed in the dual space, and measured with the dual norm, as we now discuss.

Throughout this work, we assume that the trial space U and test space V are real Hilbert spaces, and denote U' and V' as the respective topological dual spaces. Let $U_h \subset U$ and $V_h \subset V$ be finite dimensional subsets. We are interested in the following problem: given $l \in V'$, find $u_h \in U_h$ such that

$$b(u_h, v_h) = l(v_h), \quad \forall v_h \in V_h, \tag{3.1}$$

where $b(\cdot,\cdot) : U \times V \to \mathbb{R}$ is a continuous bilinear form. U is chosen to be some trial space of approximating functions, but V_h is as of yet unspecified. Throughout this work, we suppose the variational problem (3.1) to be well-posed.

We can identify a unique operator $B : U \to V'$ such that

$$\langle Bu, v\rangle_V \coloneqq b(u, v), \quad u \in U, v \in V$$

with $\langle\cdot,\cdot\rangle_V$ denoting the duality pairing between V' and V, to obtain the operator form of the continuous variational problem

$$Bu = l \quad \text{in } V'. \tag{3.2}$$

In other words, we can represent the continuous form of our variational equation (3.1) equivalently as the operator equation (3.2) with values in the dual space V'. This motivates us to consider the conditions under which the solution to (3.1) is the solution to the minimum residual problem in V'

$$u_h = \underset{u_h \in U_h}{\arg\min}\, J(u_h),$$

where $J(w)$ is defined for $w \in U$ as

$$J(w) = \frac{1}{2}\|Bw - l\|_{V'}^2 \coloneqq \frac{1}{2} \sup_{v \in V\setminus\{0\}} \frac{|b(w, v) - l(v)|^2}{\|v\|_V^2}.$$

For convenience in writing, we will abuse the notation $\sup_{v\in V}$ to denote $\sup_{v\in V\setminus\{0\}}$ for the remainder of this work.

Let us define $R_V : V \to V'$ as the Riesz map, which identifies elements of V with elements of V' by

$$\langle R_V v, \delta v\rangle_V \coloneqq (v, \delta v)_V, \quad \forall \delta v \in V.$$

Here, $(\cdot,\cdot)_V$ denotes the inner product in V. As R_V and its inverse, R_V^{-1}, are both isometries, e.g. $\|f\|_{V'} = \|R_V^{-1} f\|_V, \forall f \in V'$, we have

$$\min_{u_h\in U_h} J(u_h) = \frac{1}{2}\left\|Bu_h - l\right\|_{V'}^2 = \frac{1}{2}\left\|R_V^{-1}(Bu_h - l)\right\|_V^2. \tag{3.3}$$

The first order optimality condition for (3.3) requires the Gâteaux derivative to be zero in all directions $\delta u \in U_h$, i.e.,

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

We define, for a given $\delta u \in U$, the corresponding *optimal test function* $v_{\delta u}$

$$v_{\delta u} \coloneqq R_V^{-1} B\delta u \quad \text{in } V. \tag{3.4}$$

The optimality condition then becomes

$$\langle Bu_h - l, v_{\delta u}\rangle_V = 0, \quad \forall \delta u \in U$$

which is exactly the standard variational equation in (3.1) with $v_{\delta u}$ as the test functions. We can define the optimal test space $V_{\text{opt}} \coloneqq \{v_{\delta u} \text{ s.t. } \delta u \in U\}$. Thus, the solution of the variational problem (3.1) with test space $V_h = V_{\text{opt}}$ minimizes the residual in the dual norm $\|Bu_h - l\|_{V'}$. This is the key idea behind the concept of optimal test functions.

Since $U_h \subset U$ is spanned by a finite number of basis functions $\{\varphi_i\}_{i=1}^N$, (3.4) allows us to compute (for each basis function) a corresponding optimal test function v_{φ_i}. The collection $\{v_{\varphi_i}\}_{i=1}^N$ of optimal test functions then forms a basis for the optimal test space. In order to express optimal test functions defined in (3.4) in a more familiar form, we take $\delta u = \varphi$, a generic basis function in U_h, and rewrite (3.4) as

$$R_V v_\varphi = B\varphi, \quad \text{in } V',$$

which is, by definition, equivalent to

$$(v_\varphi, \delta v)_V = \langle R_V v_\varphi, \delta v\rangle_V = \langle B\varphi, \delta v\rangle_V = b(\varphi, \delta v), \quad \forall \delta v \in V.$$

As a result, optimal test functions can be determined by solving the auxiliary variational problem

$$(v_\varphi, \delta v)_V = b(\varphi, \delta v), \quad \forall \delta v \in V. \tag{3.5}$$

However, in general, for standard H^1 and $H(\text{div})$-conforming finite element methods, test functions are continuous over the entire domain, and hence solving variational problem (3.5) for each optimal test function requires a global operation over the entire mesh, rendering the method impractical. A breakthrough came through the development of discontinous Galerkin (DG) methods, for which basis functions are discontinuous over elements. In particular, the use of discontinuous test functions δv reduces the problem of determining global optimal test functions in (3.5) to local problems that can be solved in an element-by-element fashion.

We note that solving (3.5) on each element exactly is infeasible since it amounts to inverting the Riesz map R_V exactly. Instead, optimal test functions are approximated using the standard Bubnov-Galerkin method on an "enriched" subspace $\tilde{V} \subset V$ such that $\dim(\tilde{V}) > \dim(U_h)$ elementwise [24, 26]. In this document, we assume the error in approximating the optimal test functions

is negligible, and refer to the work in [36] for estimating the effects of approximation error on the performance of DPG.

It is now well known that the DPG method delivers the best approximation error in the "energy norm" — that is [14, 26, 55]

$$\|u - u_h\|_{U,E} = \inf_{w \in U_h} \|u - w\|_{U,E}\,, \tag{3.6}$$

where the energy norm $\|\cdot\|_{U,E}$ is defined for a function $\varphi \in U$ as

$$\|\varphi\|_{U,E} \coloneqq \sup_{v \in V} \frac{b(\varphi, v)}{\|v\|_V} = \sup_{\|v\|_V = 1} b(\varphi, v) = \sup_{\|v\|_V = 1} \langle B\varphi, v\rangle_V = \|B\varphi\|_{V'} = \|v_\varphi\|_V\,, \tag{3.7}$$

where the last equality holds due to the isometry of the Riesz map R_V (or directly from (3.5) by taking the supremum). An additional consequence of adopting such an energy norm is that, without knowing the exact solution, the energy error $\|u - u_h\|_{U,E}$ can be determined by computing $\|v_{u-u_h}\|_V$ from the following identity

$$(v_{u-u_h}, \delta v)_V = b(u - u_h, \delta v) = l\,(\delta v) - b(u_h, \delta v).$$

This is simply a consequence of the minimum-residual nature of DPG; the energy error is simply the norm of the residual in V'.

Practically speaking, this implies that the DPG method is not only discretely stable, but delivers the *best approximation in the energy norm* on any mesh. In particular, the stability and optimality of DPG apply naturally to higher order adaptive meshes, where discrete stability is often an issue.

3.0.2 Ultra-weak variational formulation

The naming of the discontinuous Petrov-Galerkin method refers to the fact that the method is a Petrov-Galerkin method, and that the test functions are specified to be discontinuous across

element boundaries. There is no specification of the regularity of the trial space, and we stress that the idea of DPG is not inherently tied to a single variational formulation [14].

In most of the DPG literature, however, the discontinuous Petrov-Galerkin method refers to the combination of the concept of locally computable optimal test functions in Section 3.0.1 with the so-called "ultra-weak formulation" [24, 26, 28, 55, 47, 46]. Unlike the previous two sections in which we studied the general equation (3.1) given by abstract bilinear and linear forms, we now consider a concrete instance of (3.1) resulting from an ultra-weak formulation for an abstract first-order system of PDEs $Au = f$. Additionally, from this section onwards, we will refer to DPG as the pairing of the ultra-weak variational formulation with the concept of locally computable optimal test functions.

We begin by partitioning the domain of interest Ω into N^{el} non-overlapping elements $K_j, j = 1, \ldots, N^{\text{el}}$ such that $\Omega_h = \cup_{j=1}^{N^{\text{el}}} K_j$ and $\overline{\Omega} = \overline{\Omega}_h$. Here, h is defined as $h = \max_{j \in \{1,\ldots,N^{\text{el}}\}} \text{diam}\,(K_j)$. We denote the mesh "skeleton" by $\Gamma_h = \cup_{j=1}^{N^{\text{el}}} \partial K_j$; the set of all faces/edges e, each of which comes with a normal vector n_e. The internal skeleton is then defined as $\Gamma_h^0 = \Gamma_h \setminus \partial\Omega$. If a face/edge $e \in \Gamma_h$ is the intersection of ∂K_i and ∂K_j, $i \neq j$, we define the following jumps:

$$[\![v]\!] = \text{sgn}\left(n^-\right) v^- + \text{sgn}\left(n^+\right) v^+, \quad [\![\tau \cdot n]\!] = n^- \cdot \tau^- + n^+ \cdot \tau^+,$$

where

$$\text{sgn}\left(n^\pm\right) = \begin{cases} 1 & \text{if } n^\pm = n_e \\ -1 & \text{if } n^\pm = -n_e \end{cases}.$$

For e belonging to the domain boundary $\partial\Omega$, we define

$$[\![v]\!] = v, \quad [\![\tau \cdot n]\!] = n_e \cdot \tau.$$

Note that we allow arbitrariness in assigning "-" and "+" quantities to the adjacent elements K_i and K_j.

The ultra-weak formulation for $Au = f$ on Ω_h, ignoring boundary conditions for now, reads

$$b\left((u, \widehat{u}), v\right) \coloneqq \langle \widehat{u}, [\![v]\!] \rangle_{\Gamma_h} - (u, A_h^* v)_{\Omega_h} = (f, v)_{\Omega_h}, \tag{3.8}$$

where we have denoted $\langle \cdot, \cdot \rangle_{\Gamma_h}$ as the duality pairing on Γ_h, $(\cdot, \cdot)_{\Omega_h}$ the L^2-inner product over Ω_h, and A_h^* the formal adjoint resulting from element-wise integration by parts. Occasionally, for simplicity in writing, we will ignore the subscripts in the duality pairing and L^2-inner product if they are Γ_h and Ω_h. Both the inner product and formal adjoint are understood to be taken element-wise. Using the ultra-weak formulation, the regularity requirement on solution variable u is relaxed, that is, u is now square integrable for the ultra-weak formulation (3.8) to be meaningful, instead of being (weakly) differentiable. The trade-off is that u does not admit a trace on Γ_h even though it did originally. Consequently, we need to introduce an additional new "trace" variable $\widehat{u}$ in (3.8), that is defined only on Γ_h.

The energy setting is now clear; namely,

$$u \in L^2(\Omega_h) \equiv L^2(\Omega), \quad v \in V = D(A_h^*), \quad \widehat{u} \in \gamma(D(A)),$$

where $D(A_h^*)$ denotes the broken graph space corresponding to A_h^*, and $\gamma(D(A))$ the trace space (assumed to exist) of the graph space of operator A. The first discussion of the well-posedness of DPG with the ultra-weak formulation can be found in [25], where the proof is presented for the Poisson and convection-diffusion equations. A more comprehensive discussion of the abstract setting for DPG with the ultra-weak formulation using the graph space, as well as a more general proof of well-posedness, can be consulted in [13].

3.0.3 Choices of test and trial norms

A clear property of the energy norm defined by (3.7) is that the trial norm $\|\cdot\|_{U,E}$ is induced by a given test norm. However, the reverse relationship holds as well; for any trial norm, the test

norm that induces such a norm is recoverable through duality. We have a result, proved in [14]: assuming, for simplicity, that the bilinear form $b(u,v)$ is definite, given any norm $\|\cdot\|_U$ on the trial space U, for $\varphi \in U$, we can represent $\|\varphi\|_U$ via

$$\|\varphi\|_U = \sup_{v\in V} \frac{b\,(w,v)}{\|v\|_{V,U}}.$$

where $\|v\|_{V,U}$ is defined through

$$\|v\|_{V,U} = \sup_{w\in U} \frac{b\,(w,v)}{\|w\|_U}.$$

In particular, given two arbitrary norms $\|\cdot\|_{U,1}$ and $\|\cdot\|_{U,2}$ in U such that $\|\cdot\|_{U,1} \leq c\,\|\cdot\|_{U,2}$ for some constant c, they generate two norms $\|\cdot\|_{V,U,1}$ and $\|\cdot\|_{V,U,2}$ in V defined by

$$\|v\|_{V,U,1} \coloneqq \sup_{w\in U} \frac{b\,(w,v)}{\|w\|_{U,1}}, \quad \text{and} \quad \|v\|_{V,U,2} \coloneqq \sup_{w\in U} \frac{b\,(w,v)}{\|w\|_{U,2}},$$

such that $\|\cdot\|_{V,U,1}$ and $\|\cdot\|_{V,U,2}$ induce $\|\cdot\|_{U,1}$ and $\|\cdot\|_{U,2}$ as energy norms in U, respectively. That is,

$$\|\varphi\|_{U,1} = \sup_{v\in V} \frac{b(\varphi,v)}{\|v\|_{V,U,1}}, \quad \text{and } \|\varphi\|_{U,2} = \sup_{v\in V} \frac{b(\varphi,v)}{\|v\|_{V,U,2}}.$$

A question that remains to be addressed is to establish the relationship between $\|\cdot\|_{V,U,1}$ and $\|\cdot\|_{V,U,2}$, given that $\|\cdot\|_{U,1} \leq c\,\|\cdot\|_{U,2}$. But this is straightforward since we have

$$\|v\|_{V,U,2} = \sup_{u\in U} \frac{b\,(w,v)}{\|w\|_{U,2}} \leq c \sup_{w\in U} \frac{b\,(w,v)}{\|w\|_{U,1}} = c\|v\|_{V,U,1}.$$

Consequently, a stronger energy norm in U will generate a weaker norm in V and vice versa. In other words, to show that an energy norm $\|\cdot\|_{U,1}$ is weaker than another energy norm $\|\cdot\|_{U,2}$ in U, one simply needs to show the reverse inequality on the corresponding norms in V, that is, $\|\cdot\|_{V,U,1}$ is stronger than $\|\cdot\|_{V,U,2}$.

From now on, unless otherwise stated, we will refer to $\|\cdot\|_{V,U}$ as the test norm that induces a given norm $\|\cdot\|_U$. Likewise, we will refer $\|\cdot\|_{U,V}$ as the trial norm induced by a given test norm

$\|\cdot\|_V$. In this work, for simplicity of exposition, we shall call a pair of norms in U and V that induce each other as an *energy norm pairing*.

From the discussion above of energy norm and test norm pairings, we know that specifying either a test norm or trial norm is sufficient to define an energy pairing. We now derive and discuss two important energy norm pairings, the first of which begins by specifying the canonical norm in U and inducing a test norm on V. The second pairing begins instead by specifying the canonical norm on V under the ultra-weak formulation (3.8) and inducing an energy norm on the trial space U.

We begin first with the canonical norm in U. Since $\widehat{u} \in \gamma\left(D\left(A\right)\right)$, the standard norm for $\widehat{u}$ is the so-called minimum energy extension norm defined as

$$\|\widehat{u}\| = \inf_{w \in D(A),\, w|_{\Gamma_h} = \widehat{u}} \|w\|_{D(A)}. \tag{3.9}$$

The canonical norm for the group variable $(u, \widehat{u})$ is then given by

$$\|\left(u, \widehat{u}\right)\|_U^2 = \|u\|_{L^2(\Omega)}^2 + \|\widehat{u}\|^2,$$

Applying the Cauchy-Schwarz inequality, we arrive at

$$b\left(\left(u, \widehat{u}\right), v\right) \leq \|(u, \widehat{u})\|_U \|v\|_{V,U},$$

where

$$\|v\|_{V,U}^2 = \|A_h^* v\|_{L^2(\Omega)}^2 + \left(\sup_{\widehat{u} \in \gamma(D(A))} \frac{\langle \widehat{u}, [\![v]\!] \rangle_{\Gamma_h}}{\|\widehat{u}\|}\right)^2.$$

On the other hand, since $v \in D\left(A_h^*\right)$, the canonical norm for v is the broken graph norm:

$$\|v\|_V^2 = \|A_h^* v\|_{L^2(\Omega)}^2 + \|v\|_{L^2(\Omega)}^2.$$

Using the Cauchy-Schwarz inequality again, we obtain

$$b\left(\left(u,\widehat{u}\right),v\right)\leq\left\|\left(u,\widehat{u}\right)\right\|_{U,V}\left\|v\right\|_{V},$$

where

$$\left\|\left(u,\widehat{u}\right)\right\|_{U,V}^{2}=\left\|u\right\|_{L^{2}(\Omega)}^{2}+\sup_{v\in D\left(A_{h}^{*}\right)}\frac{\langle\widehat{u},[\![v]\!]\rangle_{\Gamma_{h}}^{2}}{\left\|v\right\|_{V}^{2}},\tag{3.10}$$

Using the framework developed in [14], one can show that both pairs $\left(\left\|\left(u,\widehat{u}\right)\right\|_{U},\left\|v\right\|_{V,U}\right)$ and $\left(\left\|\left(u,\widehat{u}\right)\right\|_{U,V},\left\|v\right\|_{V}\right)$ are energy norm pairings in the sense discussed in Section 3.0.3. That is, the canonical norm $\left\|\left(u,\widehat{u}\right)\right\|_{U}$ in U induces (generates) the norm $\left\|v\right\|_{V,U}$ in V, while the canonical norm $\left\|v\right\|_{V}$ in V induces (generates) the energy norm $\left\|\left(u,\widehat{u}\right)\right\|_{U,V}$ in U. In the DPG literature [55], $\left\|v\right\|_{V,U}$ is known as the *optimal test norm*, while $\left\|v\right\|_{V}$ is known as the *quasi-optimal test norm*.

Trial norm		Test norm
$\boxed{\left\|u\right\|_{L^{2}(\Omega)}^{2}+\left\|\widehat{u}\right\|^{2}}$	$\Longrightarrow$	$\left\|A_{h}^{*}v\right\|_{L^{2}(\Omega)}^{2}+\left(\sup_{\widehat{u}}\frac{\langle\widehat{u},[\![v]\!]\rangle_{\Gamma_{h}}}{\left\|\widehat{u}\right\|}\right)^{2}$
$\left\|u\right\|_{L^{2}(\Omega)}^{2}+\sup_{v}\left(\frac{\langle\widehat{u},[\![v]\!]\rangle_{\Gamma_{h}}}{\left\|v\right\|_{V}}\right)^{2}$	$\Longleftarrow$	$\boxed{\left\|A_{h}^{*}v\right\|_{L^{2}(\Omega)}^{2}+\left\|v\right\|_{L^{2}(\Omega)}^{2}}$

Figure 3.1: A summary of the derivation of test/trial norm pairings; we begin with the boxed norm on either the trial or test space, and induce the norm on the other space through duality. The optimal *test* norm is naturally derived by beginning with the canonical norm on the trial space, while the quasi-optimal *trial* norm is derived from beginning with the canonical norm on the test space.

The canonical norm $\left\|\left(u,\widehat{u}\right)\right\|_{U}$ in U provides a good balance between the standard norms on the field u and the flux $\widehat{u}$ [55]. As a result, if the induced norm $\left\|v\right\|_{V,U}$ (namely, the optimal test norm) is used to compute optimal test functions in (3.5), the finite element error in the canonical norm is the best in the sense of (3.6).

Unfortunately, the optimal test norm is non-localizable due to the presence of the jump term $[\![v]\!]$.[1] Since the jump terms couple elements together, the evaluation of the jump terms re-

[1] A localizable norm can be written as $\left\|v\right\|_{V(\Omega_{h})}=\sum_{K\in\Omega_{h}}\left\|v\right\|_{V(K)}$, where $\left\|v\right\|_{V(K)}$ is a norm over K.

quires contributions from all the elements in the mesh. Consequently, solving for an optimal test function amounts to inverting the Riesz map over the entire mesh Ω_h, making the optimal test norm impractical.

On the other hand, the quasi-optimal test norm $\|v\|_V$, namely the canonical norm in V, *is* localizable, and hence practical. However, it's worth noting the difference between the induced energy norm $\|(u, \widehat{u})\|_{U,V}$ and the canonical norm in U; under the induced norm $\|(u, \widehat{u})\|_{U,V}$ there is no natural interpretation for the norm in which the error in the flux variable $\widehat{u}$ is measured.

Using a variant of the quasi-optimal test norm, numerical results show that the DPG method appears to provide a "pollution-free" method without phase error for the Helmholtz equation [55], and analysis of the pollution-free nature of DPG is currently under investigation. Similar results have also been obtained in the context of elasticity [47] and the linear Stokes equations [51]. On the theoretical side, the quasi-optimal test norm has been shown to yield a well-posed DPG methodology for the Poisson and convection-diffusion equations [25]. More recently, this theory has been generalized to show the well-posedness of DPG under the quasi-optimal test norm for the large class of Friedrichs' type PDEs [13].

Chapter 4

A robust DPG method for convection-diffusion

The majority of this chapter will focus on the convection-diffusion problem using the abstract theory that we have discussed in the previous chapter. In particular, we shall use the DPG method based on the ultra-weak formulation with optimal test functions to solve this model problem and analyze its behavior as $\epsilon \to 0$. Our goal is to show the robustness of the method with respect to ϵ, and demonstrate its usefulness as a numerical method for solving singular-perturbed problems.

4.1 DPG formulation for convection-diffusion

The convection-diffusion problem is given on a domain $\Omega \subset \mathbb{R}^d$ with boundary $\partial\Omega \equiv \Gamma$

$$\nabla \cdot (\beta u) - \epsilon \Delta u = f \in L^2(\Omega), \tag{4.1}$$

which can be cast into the first order form on the group variable (u, σ) as

$$A(u,\sigma) \coloneqq \begin{bmatrix} \nabla \cdot (\beta u - \sigma) \\ \frac{1}{\epsilon}\sigma - \nabla u \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix}. \tag{4.2}$$

Using the abstract ultra-weak formulation developed in Section 3.0.2 for the first order system of PDEs (4.2) we obtain

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

where (v, τ) is the group test function. It should be pointed out that the divergence and gradient operators are understood to act element-wise on test functions (v, τ) in the broken graph space

$D(A_h^*) \coloneqq H^1(\Omega_h) \times H(\mathrm{div}, \Omega_h)$, but globally as usual on conforming test functions, i.e. $(v, \tau) \in H^1(\Omega) \times H(\mathrm{div}, \Omega)$. It follows that the canonical test norm can be written as

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

where

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

In order to define the proper norm on the trial space, boundary conditions need to be specified. We begin by splitting the boundary Γ as follows

$$\Gamma_- \coloneqq \{x \in \Gamma; \beta_n(x) < 0\}, \quad \text{(inflow)}$$

$$\Gamma_+ \coloneqq \{x \in \Gamma; \beta_n(x) > 0\}, \quad \text{(outflow)}$$

$$\Gamma_0 \coloneqq \{x \in \Gamma; \beta_n(x) = 0\},$$

where $\beta_n \coloneqq \beta \cdot n$. Previous work in [29] adopted Dirichlet boundary conditions everywhere on Γ. We employ instead the inflow condition of Hesthaven *et al.* [40], where we set

$$\beta_n u - \sigma_n = u_0, \quad \text{on } \Gamma_-,$$

instead of $\beta_n u = u_0$. The former resembles the latter as ϵ approaches zero[1]; however, the latter induces a more "well-behaved" adjoint problem than the former, which, as we will discuss, affects the performance of DPG.

On the outflow boundary, we apply standard homogeneous Dirichlet boundary conditions

$$u = 0, \quad \text{on } \Gamma_+.$$

[1]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 (4.1) 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.

The material in this chapter is intended to act as an extension of the theory developed by Heuer and Demkowicz in [29]. The primary focus of this chapter is to analyze the DPG method and extend previous results under this new choice of inflow boundary conditions. The difference in the performance of DPG under both new and old boundary conditions is connected to the difference in the adjoint problems induced under each boundary condition. The secondary contribution of this work will be to analyze the performance of DPG under a new mesh-dependent test norm.

4.1.1 Norms on U

With the above boundary conditions at hand, the ultra-weak formulation (3.8) can be fitted in the abstract form (3.1) as

$$b\left(\left(u,\sigma,\widehat{u},\widehat{f}_n\right),(v,\tau)\right) = (u,\nabla\cdot\tau-\beta\cdot\nabla v)_{\Omega_h} + \left(\sigma,\epsilon^{-1}\tau+\nabla v\right)_{\Omega_h} \\ - \langle[\![\tau\cdot n]\!],\widehat{u}\rangle_{\Gamma_h\setminus\Gamma_+} + \left\langle\widehat{f}_n,[\![v]\!]\right\rangle_{\Gamma_h\setminus\Gamma_-} = (f,v) - \langle u_0,v\rangle_{\Gamma_-} = l\left((v,\tau)\right),$$

which, after using the setting in Section 3.0.2, suggests the following trial space (see [25, 13] for details):

$$u,\sigma\in L^2(\Omega), \quad \text{and} \quad \left(\widehat{u},\widehat{f}_n\right)\in\gamma\left(D\left(A\right)\right)\subset\gamma\left(H^1(\Omega)\times H(\mathrm{div},\Omega)\right) = H^{\frac{1}{2}}(\Gamma_h)\times H^{-\frac{1}{2}}(\Gamma_h).$$

The space for u and σ are simply scalar and vector L^2 spaces over Ω, while the space for $\left(\widehat{u},\widehat{f}_n\right)$ is the trace space of the graph space of the operator A subject to the boundary conditions.

The minimum energy extension norm (3.9) now reads

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

As a result, the canonical norm on U is given by

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

4.1.2 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 which are equivalent to the canonical $H^1(K) \times H(\mathrm{div}, K)$ norm over a single element

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

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

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

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

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

4.1.3 Approximability of the quasi-optimal test norm

An obvious choice for the test norm would be the quasi-optimal norm; it is the canonical test norm, and DPG has been shown to be well-posed and robust under such an optimal test norm for a large class of problems [29, 13, 50]. However, computations with the quasi-optimal test norm for convection-diffusion problems turn out to be quite problematic for small diffusion and coarse meshes.

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

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

Use of this norm for the convection-diffusion problem is difficult — since the problem (3.5) for optimal test functions is local, we can transform the problem over a single element K to the reference element $\widehat{K}$ and show that it is equivalent to a reaction-diffusion system, with diffusion parameter $\frac{\epsilon}{|K|}$, where $|K|$ is the element measure [46]. We refer to the inverse of this parameter $\frac{|K|}{\epsilon}$ as the element Peclet number Pe. For a coarse mesh and small diffusion parameter ϵ, we will have a large element Peclet number, and optimal test functions under the quasi-optimal test norm will develop strong boundary layers of width Pe, as seen in Figure 4.1.

In the application of DPG in [24, 26, 28, 55], the approximation of optimal test functions is done using polynomial enrichment. We search for the solution to (3.5) in the enriched test space $\tilde{V} \approx \prod_K P^{p+\Delta p}(K)$, where p is the polynomial order of the trial space on a given element K.[2] 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 error in approximation of test functions is tied to the effectiveness of the p-method. Unfortunately, for problems with boundary layers — including the approximation of test functions under the quasi-optimal test norm — the p-method performs very poorly. As a result of this poor approximation, the numerical solutions of the convection-dominated diffusion equation under DPG using the quasi-optimal test norm tend to be of poor quality, and do not exhibit all the proven properties of DPG (for example, the energy error may increase after mesh refinement, even though, by virtue of DPG delivering a best approximation, the energy error for a coarse mesh must be greater than or equal

[2] V is only *approximately* equal to the space $\prod_K P^{p+\Delta p}(K)$. In practice, V is constructed using locally H^1-conforming and Raviart-Thomas elements of appropriate order.

to the energy error for a finer mesh). We conclude that the error in approximation of optimal test functions using simple polynomial enrichment pollutes and ruins the performance of DPG under the quasi-optimal test norm.

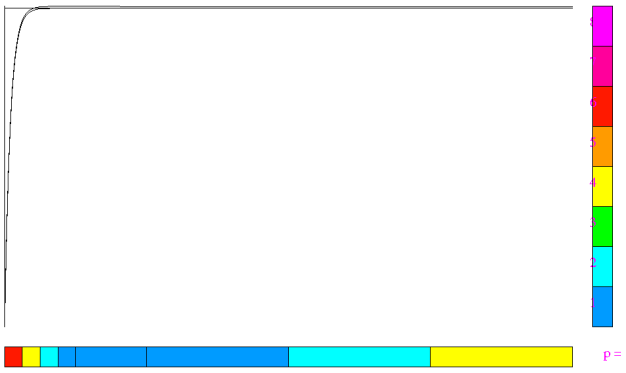


Figure 4.1: v and τ components of the 1D optimal test functions corresponding to the flux $\widehat{f}_n$ on the right-hand side of a unit element for $\epsilon = 0.01$. The solution has been obtained using automatic hp-adaptivity driven by the test norm with the error tolerance set at 1%.

The difficulty in using the quasi-optimal test norm for convection-diffusion is perplexing at first, considering that 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 — the singular perturbation parameter. Transforming the variational problem using the quasi-optimal test norm for wave propagation yields smooth optimal test functions that are approximated much more accurately using only polynomials over the reference element. Typically, the wavenumbers k of physical interest are $O(100)$ with respect to a unit domain. The corresponding finite element problems will typically be solved on meshes containing approximately $O\big(k^d\big)$ 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 under-resolved meshes is thus unavoidable, and the approximability of optimal test functions must be addressed in order to take advantage of the properties of DPG.

Resolving such boundary layers present in test functions under the quasi-optimal test norm has been investigated numerically using specially designed (Shishkin) subgrid meshes by Niemi, Collier, and Calo in [46]. However, even with Shishkin meshes, the approximation of optimal test functions under the quasi-optimal norm is far more expensive and complex to implement than approximation of test functions using a simple p-enriched space for V. We therefore aim instead to design a test norm that does not induce boundary layers, but still delivers good approximation results over a range of ϵ.

4.1.4 Analysis of a DPG test norm

We are interested in computing DPG optimal test functions for the convection-diffusion equation with very small values of ϵ; due to the difficulty of approximating optimal test functions, we conclude that the use of the quasi-optimal test norm is infeasible towards this goal.

However, if we naively choose a test norm that does not generate boundary layers, the performance of DPG may be adversely affected. For example, if $\| (v,\tau) \|_V^2 = \|v\|_{H^1(\Omega_h)}^2 + \|\tau\|_{H(\mathrm{div},\Omega_h)}^2$, the $H^1(\Omega_h) \times H(\mathrm{div},\Omega_h)$ norm, then the corresponding test functions will be smooth and free of boundary layers; however, the performance of DPG will provide approximations which worsen in quality as ϵ becomes very small [28, 29].

Our goal is to construct a test norm that compromises between performance of DPG and approximability of test functions. This test norm should not produce boundary layers in the optimal test functions, but still induce an energy norm that yields good approximation properties for small

ϵ. We note that, even under the quasi-optimal norm, the norms on the flux and trace variables will likely depend on ϵ. Thus, we aim to construct a test norm for which the DPG method will be robust in ϵ with respect to the *field variables*.

For now, we discuss the steps necessary to analyze the performance of DPG with respect to a non-canonical test norm. We require a priori that the test norm has separable τ and v components — in other words, that there are no terms in the test norm that couple τ and v together. Problem (3.5) then decouples, such that the components of the vector-valued test function (v, τ) can be solved for independently of each other. The decoupled variational problems are no longer systems but scalar equations in τ and v, for which it is easier to conclude whether or not there are boundary layers in the solutions (the avoidance of boundary layers in the test norm will be discussed in more detail in Section 4.2, which describes our numerical experiments). **This will ensure that the resulting DPG method does not suffer from approximation errors in the optimal test functions**.

We begin with the following test norm:

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

The use of this norm is problematic for practical computations; we will discuss the reasons why and present a modification of it in Section 4.1.4.3.

We can see how this norm will differ from the canonical $H^1(\Omega_h)\times H(\mathrm{div},\Omega_h)$ norm: the clearest difference is the fact that the gradient in the streamline direction is $O(1)$, while the full gradient is $O(\sqrt{\epsilon})$, so that, in our test norm, the streamline gradient of v will be emphasized over the full gradient of v for small ϵ.

The choice of this test norm is implied by the mathematics of the adjoint problem. Roughly speaking, necessary conditions for the performance of DPG to not degenerate as $\epsilon\to 0$ are derived

through analysis of specific test functions. For example, if u is the first L^2 component of the solution to the variational problem defined in Section 4.1, by choosing $(v,\tau) \in H^1(\Omega) \times H(\mathrm{div},\Omega)$ such that

$$\nabla \cdot \tau - \beta \cdot \nabla v = u$$
$$\frac{1}{\epsilon}\tau - \nabla v = 0,$$

we have

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

and we recover the L^2 norm of u from the bilinear form.

Let $\|a\| \lesssim \|b\|$ denote an ϵ-independent bound; specifically, that $\|a\| \leq C\|b\|$ for a constant C independent of ϵ. Consequently, if for any $u \in L^2(\Omega_h)$, $\|(v,\tau)\|_V \lesssim \|u\|_{L^2}$, then dividing through by $\|u\|_{L^2}$ gives the bound

$$\|u\|_{L^2} \lesssim \left\|\left(u,\sigma,\widehat{u},\widehat{f}_n\right)\right\|_E .$$

In other words, there is the guarantee that the L^2 error in u is at least robustly bounded from above by the energy error. Then, if the energy error (which DPG minimizes) approaches zero, the L^2 error in u will as well. The same exercise can be repeated for the stress σ, as well as the flux variables $\widehat{u}$, $\widehat{f}_n$.

This methodology gives constraints on the quantities found in the test norm; any quantity present in $\|(v,\tau)\|_V$ must be shown to be bounded from above independently of ϵ by the load of the adjoint problem. However, showing this simply amounts to showing *standard energy estimates* for H^1 and $H(\mathrm{div})$-conforming finite elements. A more detailed discussion on the reasoning behind the construction of test norms can be found in [29].

The second step will be to **show the equivalence of the energy norm to explicit norms on** U. Since we do not generally have a closed form expression for the DPG energy norm,

we seek to understand the behavior of DPG by finding a norm on U to which the DPG energy norm is equivalent. Since $\left(u, \sigma, \widehat{u}, \widehat{f_n}\right) \in U$ is a group variable from a tensor product space, we construct norms on U through the combination of norms on u, σ, $\widehat{u}$, and $\widehat{f_n}$. Specifically, we use the norm on U

$$\left\|\left(u, \sigma, \widehat{u}, \widehat{f_n}\right)\right\|_U^2 \coloneqq \|u\|^2 + \|\sigma\|^2 + \|\widehat{u}\|^2 + \left\|\widehat{f_n}\right\|^2. \tag{4.3}$$

For equivalence between norms, two constants are specified. However, since this norm on U is a norm on four separate variables, we can specify not just two but eight equivalence constants.[3] In order to simplify analysis, we phrase this equivalence statement in an alternative form.

Let $\|\cdot\|_E \coloneqq \|\cdot\|_{U,V}$, the energy norm induced by the test norm described above. We seek the bound of $\|\cdot\|_E$ from above and below:

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

where both $\|\cdot\|_{U,1}$ and $\|\cdot\|_{U,2}$ are defined as scaled combinations of the norms on $u, \sigma, \widehat{u}$, and $\widehat{f_n}$

$$\left\|\left(u, \sigma, \widehat{u}, \widehat{f_n}\right)\right\|_{U,i}^2 \coloneqq \left(C_u^i \|u\|\right)^2 + \left(C_\sigma^i \|\sigma\|\right)^2 + \left(C_{\widehat{u}}^i \|\widehat{u}\|\right)^2 + \left(C_{\widehat{f}_n}^i \|\widehat{f_n}\|\right)^2, \quad i = 1, 2 \tag{4.4}$$

Our goal is to explicitly derive the equivalence constants that define the norms $\|\cdot\|_{U,1}$ and $\|\cdot\|_{U,2}$ respectively, taking into account any dependency on ϵ. To do so, we need a relation between trial norms on U and test norms on V.

Recall from Section 3.0.3 that every test norm induces a corresponding trial norm, and vice versa. Let $\|\cdot\|_{U,1} \simeq \|\cdot\|_{U,2}$ mean that the norms $\|\cdot\|_{U,1}$ and $\|\cdot\|_{U,2}$ are equivalent, with equivalence constants independent of ϵ. By equivalence of finite dimensional norms and the discussion

[3] Sharper estimates are attainable if these constants are allowed to vary over the mesh Ω_h. See Section 4.1.4.4 for a discussion.

in Section 3.0.3 on the duality between test norms/energy norms, the norms (4.4) on U induce the equivalent test norms on $(v,\tau) \in H^1(\Omega_h) \times H(\mathrm{div}, \Omega_h)$

$$\begin{aligned}
\|(v,\tau)\|_{V,U,i} &\simeq \sup_{(u,\sigma,\widehat{u},\widehat{f}_n)\in U} \frac{b\left(\left(u,\sigma,\widehat{u},\widehat{f}_n\right),(v,\tau)\right)}{C_u^i\|u\| + C_\sigma^i\|\sigma\| + C_{\widehat{u}}^i\|\widehat{u}\| + C_{\widehat{f}_n}^i\|\widehat{f}_n\|} \\
&= \sup_{(u,\sigma,\widehat{u},\widehat{f}_n)\in U} \frac{(u, \nabla\cdot\tau - \beta\cdot\nabla v) + \left(\sigma, \epsilon^{-1}\tau + \nabla v\right) - \langle [\![\tau_n]\!], \widehat{u}\rangle_{\Gamma_-\cup\Gamma_h^0} + \langle \widehat{f}_n, [\![v]\!]\rangle_{\Gamma_+\cup\Gamma_h^0}}{C_u^i\|u\| + C_\sigma^i\|\sigma\| + C_{\widehat{u}}^i\|\widehat{u}\| + C_{\widehat{f}_n}^i\|\widehat{f}_n\|} \\
&\simeq \frac{1}{C_u^i}\|g\| + \frac{1}{C_\sigma^i}\|f\| + \frac{1}{C_{\widehat{u}}^i} \sup_{\widehat{u}\neq 0,\, \widehat{u}|_{\Gamma_+}=0} \frac{\langle [\![\tau\cdot n]\!], \widehat{u}\rangle}{\|\widehat{u}\|} + \frac{1}{C_{\widehat{f}_n}^i} \sup_{\widehat{f}_n\neq 0,\, \widehat{f}_n|_{\Gamma_-}=0} \frac{\langle \widehat{f}_n, [\![v]\!]\rangle}{\|\widehat{f}_n\|},
\end{aligned}$$

where f and g are defined element-wise over Ω_h as

$$g \coloneqq \nabla\cdot\tau - \beta\cdot\nabla v$$

$$f \coloneqq \epsilon^{-1}\tau + \nabla v.$$

By definition of the norms on the quantities defined on the skeleton Γ_h, this gives the characterization of the induced test norm

$$\|\,(v,\tau)\,\|_{V,U,i} \simeq \frac{1}{C_u^i}\|g\| + \frac{1}{C_\sigma^i}\|f\| + \frac{1}{C_{\widehat{u}}^i}\|\,[\![\tau\cdot n]\!]\,\| + \frac{1}{C_{\widehat{f}_n}^i}\|\,[\![v]\!]\,\|, \quad i = 1,2.$$

We can now use this relation to compare different norms on U by comparing their induced norms on V (recall that showing a robust inequality between two norms on U is equivalent to showing the robust *reverse* inequality in the induced norms on V). Namely, we can show the bound of $\|\cdot\|_{U,1} \lesssim \|\cdot\|_E$ by showing the bound $\|(v,\tau)\|_{V,U,1} \gtrsim \|(v,\tau)\|_V$, and likewise for $\|\cdot\|_E \lesssim \|\cdot\|_{U,2}$.

Since the techniques used to show such bounds are more involved, we break the procedure up into two steps:

1. Decompose test functions (v,τ) into three separate, more easily analyzable components (Section 4.1.4.1).

2. Derive adjoint estimates (Section 4.1.4.2).

4.1.4.1 Decomposition into analyzable components

Having reduced the problem of comparing norms on U to the comparison of norms on V, we break the analysis of $(v, \tau) \in V$ 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

$$\epsilon^{-1}\tau_1 + \nabla v_1 = 0,$$

$$\nabla \cdot \tau_1 - \beta \cdot \nabla v_1 = \nabla \cdot \tau - \beta \cdot \nabla v = g,$$

and (v_2, τ_2) satisfies

$$\epsilon^{-1}\tau_2 + \nabla v_2 = \epsilon^{-1}\tau + \nabla v = f,$$

$$\nabla \cdot \tau_2 - \beta \cdot \nabla v_2 = 0.$$

Both $(v_1, \tau_1), (v_2, \tau_2) \in H(\mathrm{div}; \Omega) \times H^1(\Omega)$ are understood to satisfy these relations in a conforming sense over the domain Ω; however, the divergence of τ and gradient of v on the right hand side are still understood to be taken in an element-wise fashion.

We will additionally require both $(v_1, \tau_1), (v_2, \tau_2)$ to satisfy the adjoint homogeneous boundary conditions

$$\tau_i \cdot n = 0, \quad \text{on } \Gamma_- \tag{4.5}$$

$$v_i = 0, \quad \text{on } \Gamma_+ \tag{4.6}$$

for $i = 1, 2$. The selection of $H(\mathrm{div}, \Omega) \times H^1(\Omega)$ conforming test functions satisfying the specific boundary conditions above removes the contribution of the jump terms over the skeleton Γ_h in

the bilinear form, allowing us to analyze field terms in the induced test norms separately from the boundary/jump terms.

Finally, by construction, $(v_0, \tau_0) \in H^1(\Omega_h) \times H(\mathrm{div}, \Omega_h)$ must satisfy

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

$$\nabla \cdot \tau_0 - \beta \cdot \nabla v_0 = 0$$

with jumps

$$[\![v_0]\!] = [\![v]\!], \quad \text{on } \Gamma_h^0$$

$$[\![\tau_0 \cdot n]\!] = [\![\tau \cdot n]\!], \quad \text{on } \Gamma_h^0.$$

and boundary conditions

$$v_0 = v, \quad \text{on } \Gamma_+$$

$$\tau_0 \cdot n = \tau \cdot n, \quad \text{on } \Gamma_- \cup \Gamma_0.$$

Notice that the evaluation the bilinear form $b\left(\left(u, \sigma, \widehat{u}, \widehat{f}_n\right), (v, \tau)\right)$ with each specific test functions returns only one part of the bilinear form. Furthermore, by choosing the proper loads $g = u$ and $f = \sigma$, we can recover from the bilinear form the norms of u and σ (as described in Section 4.1.4), as well as the norms on $\widehat{u}$, and $\widehat{f}_n$.[4]

We have now decomposed an arbitrary test function (τ, v) into a discontinuous contribution and two continuous contributions. Recall that our goal is to show the robust bound from above and below of the DPG energy norm by $\|\cdot\|_{U,1}$ and $\|\cdot\|_{U,2}$:

$$\left\|\left(u, \sigma, \widehat{u}, \widehat{f}_n\right)\right\|_{U,1} \lesssim \left\|\left(u, \sigma, \widehat{u}, \widehat{f}_n\right)\right\|_E \lesssim \left\|\left(u, \sigma, \widehat{u}, \widehat{f}_n\right)\right\|_{U,2}.$$

[4]To recover the norms on $\widehat{u}$, and $\widehat{f}_n$, the loads f, and g must be zero, and the jumps of the test function (v, τ) must be chosen specifically.

Under the duality of trial and test norms and the decomposition of test functions $(\tau, v) \in V$ into $(\tau_0, v_0), (\tau_1, v_1)$, and (τ_2, v_2), the above bound is equivalent to bounding each component

$$\| (v, \tau) \|_{V,U,1} \gtrsim \sum_{i=0}^{2} \| (v_i, \tau_i) \|_V \gtrsim \| (v, \tau) \|_{V,U,2}.$$

Bounding $\| (v_0, \tau_0) \|$ requires the use of techniques first developed in [25] and adapted to convection-diffusion in [25] and [29]. However, since $(\tau, v) \in H(\mathrm{div}, \Omega) \times H^1(\Omega)$, the bound from above of test functions $\| (v_1, \tau_1) \|_V$ and $\| (v_2, \tau_2) \|_V$ is reduced to proving classical error estimates for the adjoint equations

$$\begin{aligned} \epsilon^{-1}\tau_1 + \nabla v_1 &= 0 \\ \nabla \cdot \tau_1 - \beta \cdot \nabla v_1 &= g, \\ \tau_1 \cdot n|_{\Gamma_-} &= 0, \\ v_1|_{\Gamma_+} &= 0. \end{aligned}$$

and

$$\begin{aligned} \epsilon^{-1}\tau_2 + \nabla v_2 &= f \\ \nabla \cdot \tau_2 - \beta \cdot \nabla v_2 &= 0, \\ \tau_2 \cdot n|_{\Gamma_-} &= 0, \\ v_2|_{\Gamma_+} &= 0. \end{aligned}$$

More generally, we can analyze the adjoint equations

$$\epsilon^{-1}\tau + \nabla v = f \tag{4.7}$$

$$\nabla \cdot \tau - \beta \cdot \nabla v = g, \tag{4.8}$$

for arbitrary data $f, g \in L^2(\Omega)$ and boundary conditions $[\![\tau \cdot n]\!]_{\Gamma_-} = 0$ and $[\![v]\!]_{\Gamma_+} = 0$. In other words, we want to analyze the stability properties of the adjoint equations by deriving bounds of the form $\| (v_1, \tau_1) \|_V \lesssim \|g\|_{L^2}$ and $\| (v_2, \tau_2) \|_V \lesssim \|f\|_{L^2}$.

4.1.4.2 Adjoint estimates

The final step to estimating the induced norm on U by a selected localizable test norm on V is to derive adjoint stability estimates on τ and v in terms of localizable normed quantities. We will construct complete test norms on V through combinations of these normed quantities.

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 that some or all of the following assumptions hold:

$$\nabla \times \beta = 0, \quad 0 < C \leq |\beta|^2 + \frac{1}{2} \nabla \cdot \beta, \quad C = O(1), \tag{4.9}$$

$$\nabla \beta + \nabla \beta^T - \nabla \cdot \beta I = O(1), \tag{4.10}$$

$$\nabla \cdot \beta = 0. \tag{4.11}$$

Under proper assumptions on β, we have the robust bounds, which are proved in the Appendix.

- **Lemma 2**: *For β satisfying* (4.9) *and* (4.10)*, and $v_1 \in H^1(\Omega)$, satisfying equations* (4.7) *and* (4.8) *with $f = 0$, and with boundary conditions* (4.5) *and* (4.6),

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

  Similarly, from $\nabla \cdot \tau_1 - \beta \cdot \nabla v_1 = g$, we get $\|\nabla \cdot \tau_1\| \lesssim \|g\|$ as well.

- **Lemma 3**: *For β satisfying* (4.9)*, and $v \in H^1(\Omega)$ satisfying equations* (4.7) *and* (4.8) *and*

boundary conditions (4.5) *and* (4.6)*, and for sufficiently small* ϵ,

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

We can characterize both v_1 and v_2 in the above decompositions using this theorem by setting either $f = 0$ or $g = 0$.

- **Lemma 4**: *For* β *satisfying* (4.9)*,* (4.11)*, and solutions* $v_0 \in H^1(\Omega_h)$ *and* $\tau_0 \in H(\mathrm{div}, \Omega_h)$ *of equations* (4.7) *and* (4.8) *with* $f = g = 0$*,*

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

We are interested in showing the equivalence of the DPG energy norm with norms $\|\cdot\|_{U,1}$ and $\|\cdot\|_{U,2}$, respectively. We will show this by bounding $\|\cdot\|_V$ from below by $\|\cdot\|_{V,U,1}$ and from above by $\|\cdot\|_{V,U,2}$ (the induced test norms for $\|\cdot\|_{U,1}$ and $\|\cdot\|_{U,2}$, respectively).

4.1.4.3 A mesh-dependent test norm

Ideally, we would be interested in the use of the test norm

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

for practical computations. However, the presence of the term $\|v\|$ together with $\sqrt{\epsilon}\|\nabla v\|$ (and similarly $\|\nabla \cdot \tau\|$ and $\frac{1}{\sqrt{\epsilon}}\|\tau\|$ terms) induces boundary layers in the optimal test functions for under-resolved meshes. We can see this by recovering the strong form of the variational problem defining test functions. We first note that the variational problems for the v and τ components of optimal test functions decouple from each other under this test norm. Then, examining the variational problem for the v component only of an optimal test function, and assuming $\nabla \cdot \beta = 0$ for illustrative purposes,

we have

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

After integration by parts, we recover the strong form of the operator L inducing such a variational problem

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

where we neglect the resulting boundary terms from integration by parts for now.

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

$$L_{\beta^\perp} \coloneqq v-\epsilon\Delta v,$$

and can develop boundary layers in those directions. The presence of boundary layers has been verified through numerical computation as well; using an H^1-conforming finite element code with *hp*-adaptivity [23], the solution to the variational problem defining the optimal test function under the above test norm was computed. Figure 4.2 shows the result of such a computation for the v component of an optimal test function under the above test norm. To avoid boundary layers in the optimal test functions, we follow [29] in scaling the L^2 contributions of v by $C_v(K)$, such that, when transformed to the reference element, both $C_v(K)\|v\|^2$ and $\epsilon\|\nabla v\|^2$ are of the same magnitude. Similarly, we scale the L^2 contributions of τ by $C_\tau(K)$ such that $\frac{C_\tau(K)}{\epsilon}\|\tau\|^2$ and $\|\nabla\cdot\tau\|^2$ are of the same magnitude as well. For this work, we consider only isotropic refinements on quadrilateral elements in 2D.

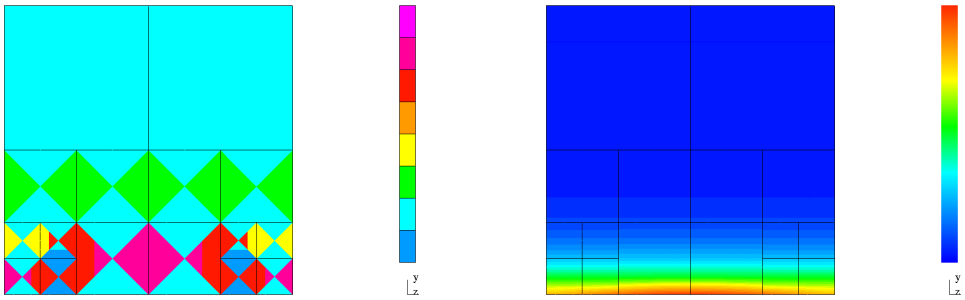


Figure 4.2: The v component of the optimal test function corresponding to flux $\widehat{u} = x(1-x)$ on the bottom side of a unit element for $\epsilon = 0.01$. The corresponding *hp*-mesh used to compute the solution is displayed to the left.

Our test norm, as defined over a single element K, is now

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

This modified test norm avoids boundary layers in the locally computed optimal test functions, but for adaptive meshes, provides additional stability in areas of heavy refinement, where the 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.

4.1.4.4 Equivalence of energy norm with $\|\cdot\|_U$

The main theoretical result of this work can now be given:

Lemma 1. *Under the mesh-dependent test norm*

$$\| (v,\tau) \|^2_{V,\Omega_h} = \|C_v v\|^2 + \epsilon\|\nabla v\|^2 + \|\beta\cdot\nabla v\|^2 + \|\nabla\cdot\tau\|^2 + \|C_\tau\tau\|^2,$$

where $C_v, C_\tau \in L^2(\Omega)$ are defined elementwise through

$$C_v|_K = \min\left\{\sqrt{\frac{\epsilon}{|K|}}, 1\right\}$$
$$C_\tau|_K = \min\left\{\frac{1}{\sqrt{\epsilon}}, \frac{1}{\sqrt{|K|}}\right\}.$$

If β satisfies (4.9)*,* (4.10)*, and* (4.11)*, the DPG energy norm $\|\cdot\|_E$ satisfies the following equivalence relations*

$$\|u\|_{L^2} + \|\sigma\|_{L^2} + \epsilon\,\|\widehat{u}\| + \sqrt{\epsilon}\left\|\widehat{f}_n\right\| \lesssim \left\|\left(u,\sigma,\widehat{u},\widehat{f}_n\right)\right\|_E$$
$$\left\|\left(u,\sigma,\widehat{u},\widehat{f}_n\right)\right\|_E \lesssim \|u\|_{L^2} + \left\|\frac{1}{\epsilon C_\tau}\sigma\right\|_{L^2} + \frac{1}{\sqrt{\epsilon}}\left(\|\widehat{u}\| + \left\|\widehat{f}_n\right\|\right).$$

Proof. We begin by proving the bound from below. As a consequence of the duality of norms discussed in Section 3.0.3, we know that the norm $\|u\|_{U,1}$ is induced by a specific test norm $\|v\|_{V,U,1}$. 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 $\|v\|_{V,U,1}$ by $\|v\|_{V,U,2}$.

For f and g defined in (4.7) and (4.8),

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

we can characterize the test norm for

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

through the equivalence relation

$$\|\,(v,\tau)\,\|_{V,U,1} \simeq \sup_{u,\sigma,\widehat{u},\widehat{f}_n} \frac{b\left(\left(u,\sigma,\widehat{u},\widehat{f}_n\right),(\tau,v)\right)}{\|u\| + \|\sigma\| + \epsilon\|\widehat{u}\| + \sqrt{\epsilon}\|\widehat{u}\|}$$
$$\simeq \|g\| + \|f\| + \frac{1}{\epsilon}\sup_{\widehat{u}\neq 0,\,\widehat{u}|_{\Gamma_+}=0} \frac{\langle[\![\tau\cdot n]\!],\widehat{u}\rangle}{\|\widehat{u}\|} + \frac{1}{\sqrt{\epsilon}}\sup_{\widehat{f}_n\neq 0,\,\widehat{f}_n|_{\Gamma_-}=0} \frac{\langle\widehat{f}_n,[\![v]\!]\rangle}{\|\widehat{f}\|},$$

which, by definition of the boundary norms, is

$$\| (v,\tau) \|_{V,U,1} \simeq \|g\| + \|f\| + \frac{1}{\epsilon}\| [\![\tau \cdot n]\!] \| + \frac{1}{\sqrt{\epsilon}}\| [\![v]\!] \|.$$

We wish to show the bound

$$\|(v,\tau)\|_{V,\Omega_h} \lesssim \|g\| + \|f\| + \frac{1}{\epsilon}\| [\![\tau \cdot n]\!] \| + \frac{1}{\sqrt{\epsilon}}\| [\![v]\!] \|.$$

By noting that both

$$\|C_v v_0\| \leq \|v_0\|,$$
$$\|C_\tau \tau_0\| \leq \frac{1}{\sqrt{\epsilon}} \|\tau_0\|,$$

we have that $\|(v,\tau)\|_{V,\Omega_h} \leq \|(v,\tau)\|_V$, so it suffices to prove the bound for the mesh-independent test norm

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

We will bound $\| (v,\tau) \|_V$ for all (v,τ) by decomposing $(v,\tau) = (v_0,\tau_0) + (v_1,\tau_1) + (v_2,\tau_2)$ as described in Section 4.1.4.1.

By the triangle inequality, robustly bounding $\| (v,\tau) \|_V$ from above reduces to robustly bounding each component

$$\| (v_0,\tau_0) \|_V, \| (v_1,\tau_1) \|_V, \| (v_2,\tau_2) \|_V \lesssim \|g\| + \|f\| + \frac{1}{\epsilon}\| [\![\tau \cdot n]\!] \| + \frac{1}{\sqrt{\epsilon}}\| [\![v]\!] \|.$$

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

  Lemma 4 gives control over $\sqrt{\epsilon}\|\nabla v_0\| + \frac{1}{\epsilon}\|\tau_0\|$ 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 [25] 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 2 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 3 gives

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

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

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

  Lemma 3 provides, for ϵ sufficiently small,

  $$\epsilon\|\nabla v_2\|^2 + \|v_2\|^2 \lesssim \epsilon\|f\|^2 \leq \|f\|^2.$$

  We have $\epsilon^{-1}\tau_2 = f - \nabla v_2$, so $\epsilon^{-1}\|\tau_2\| \lesssim \|f\| + \|\nabla v_2\|$. Lemma 3 implies $\|\nabla v_2\|^2 \lesssim \|f\|^2$, so for $\epsilon \leq 1$, we have $\epsilon^{-1/2}\|\tau_2\| \leq \epsilon^{-1}\|\tau_2\| \lesssim \|f\|$. The remaining terms can be bounded by noting that, with $g = 0$, $\|\nabla \cdot \tau_2\| = \|\beta \cdot \nabla v_2\| \lesssim \|\nabla v_2\| \lesssim \|f\|$.

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 from above 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 $\|\cdot\|_{U,2}$. For a norm on U of the form

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

the induced test norm is equivalent to

$$\begin{aligned}
\|\,(\tau,v)\,\|_{V,U,2} &\simeq \sup_{(u,\sigma,\widehat{u},\widehat{f}_n)\in U\setminus\{0\}} \frac{b\left(\left(u,\sigma,\widehat{u},\widehat{f}_n\right),(\tau,v)\right)}{\left\|\left(u,\sigma,\widehat{u},\widehat{f}_n\right)\right\|_E} \\
&\simeq \sup_{(u,\sigma,\widehat{u},\widehat{f}_n)\in U\setminus\{0\}} \frac{(u,\nabla\cdot\tau-\beta\cdot\nabla v)+\left(\sigma,\epsilon^{-1}\tau+\nabla v\right)-\langle[\![\tau_n]\!],\widehat{u}\rangle+\langle\widehat{f}_n,[\![v]\!]\rangle}{\|u\|+\left\|\left(\epsilon C_\tau\right)^{-1}\sigma\right\|+\frac{1}{\sqrt{\epsilon}}\left(\|\widehat{u}\|+\|\widehat{f}_n\|\right)} \\
&\simeq \|g\|+\|\epsilon C_\tau f\|+\sqrt{\epsilon}\left(\sup_{\widehat{u},\widehat{f}_n\neq 0}\frac{\langle[\![\tau_n]\!],\widehat{u}\rangle+\langle\widehat{f}_n,[\![v]\!]\rangle}{\|\widehat{u}\|+\|\widehat{f}_n\|}\right),
\end{aligned}$$

where f and g are

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

the loads of the adjoint problem defined in (4.7), (4.8).

Note that $\epsilon C_\tau \leq \sqrt{\epsilon}$. Then, by the triangle inequality, we have the bounds

$$\|\epsilon C_\tau f\| \leq C_\tau\|\tau\|+\epsilon C_\tau\|\nabla v\| \lesssim \|\,(\tau,v)\,\|_{V,\Omega_h}$$

$$\|g\| \leq \|\nabla\cdot\tau\|+\|\beta\cdot\nabla v\| \lesssim \|\,(\tau,v)\,\|_{V,\Omega_h}$$

We estimate the supremum on the jumps of (τ,v) by following [29]; 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

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

where we have used that $\epsilon < 1$, $\nabla \cdot \beta = O(1)$, and that $\|\beta \cdot \nabla w\| \lesssim \|\nabla w\|$.

Without loss of generality, assume the problem is scaled such that $\max_{K \in \Omega_h} |K| \leq 1$. Then, $\frac{1}{C_\tau^2} \leq \frac{1}{C_v^2} \leq \frac{1}{\epsilon}$, and an application of discrete Cauchy-Schwarz gives us

$$\begin{aligned}\langle [\![\tau \cdot n]\!], w\rangle + \langle [\![v]\!], (\eta - \beta w) \cdot n\rangle &\lesssim \| (\tau, v) \|_{V,\Omega_h} \frac{1}{\sqrt{\epsilon}} \left(\|\eta\|_{H(\mathrm{div},\Omega)} + \|w\|_{H^1(\Omega)} \right), \\ &\lesssim \| (\tau, v) \|_{V,\Omega_h} \frac{1}{\sqrt{\epsilon}} \left(\|\eta - \beta w\|_{H(\mathrm{div},\Omega)} + \|w\|_{H^1(\Omega)} \right),\end{aligned}$$

since $\|\eta\|_{H(\mathrm{div},\Omega)} = \|\eta - \beta w + \beta w\|_{H(\mathrm{div},\Omega)} \leq \|\eta - \beta w\|_{H(\mathrm{div},\Omega)} + \|\beta w\|_{H(\mathrm{div},\Omega)} \lesssim \|\eta - \beta w\|_{H(\mathrm{div},\Omega)} + \|w\|_{H^1(\Omega)}$. Dividing through and taking the supremum gives

$$\sup_{w,\eta \neq 0} \frac{\langle [\![\tau \cdot n]\!], w\rangle + \langle [\![v]\!], (\eta - \beta w) \cdot n\rangle}{\left(\|\eta - \beta w\|_{H(\mathrm{div},\Omega)} + \|w\|_{H^1(\Omega)} \right)} \lesssim \| (\tau, v) \|_{V,\Omega_h} \frac{1}{\sqrt{\epsilon}}.$$

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, from [25], by the definition of the trace norms on $[\![\tau \cdot n]\!]$ and $[\![v]\!]$

$$\sup_{\rho,\phi \neq 0} \frac{\langle [\![\tau \cdot n]\!], \rho\rangle + \langle [\![v]\!], \phi\rangle}{\|\rho\|_{H^{1/2}(\Gamma_h)} + \|\phi\|_{H^{-1/2}(\Gamma_h)}} = \sup_{w,\eta \neq 0} \frac{\langle [\![\tau \cdot n]\!], w\rangle + \langle [\![v]\!], (\eta - \beta w) \cdot n\rangle}{\|w\|_{H^1(\Omega)} + \|\eta - \beta w\|_{H(\mathrm{div},\Omega)}}.$$

Together, the bounds on the jump terms and the bounds on $\|g\|$ and $\|f\|$ imply $\left\| \left(u, \sigma, \widehat{u}, \widehat{f_n} \right) \right\|_E \lesssim \left\| \left(u, \sigma, \widehat{u}, \widehat{f_n} \right) \right\|_{U,2}$. □

4.1.4.5 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 4.1.4.4 and the convection-diffusion problem with Dirichlet boundary conditions.

The bilinear form for the case of Dirichlet boundary conditions is

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

Notice that the boundary terms in the final bilinear form are different; hence, the adjoint problems associated with Section 4.1.4.2 will now carry different boundary conditions as well. Likewise, the stability properties proven previously will not hold under a different set of boundary conditions.

As it turns out, the robust bounds given in Section 4.1.4.4 hold in $\mathbb{R}^d$ for arbitrary d; however, we can show that for the case of Dirichlet boundary conditions, the same results do not hold, even in 1D. Consider now the 1D analogue of the estimate given by Lemma 2. In 1D, $\|\beta \cdot \nabla v_1\| \lesssim \|g\|$ reduces to the inequality

$$\|\beta v_1'\| \lesssim \|g\|, \quad g \in L^2(\Omega_h)\,.$$

Without this inequality, we are unable to prove the robust bound on the L^2 error $\|u - u_h\|_{L^2} \lesssim \left\|(u, \sigma, \widehat{u}, \widehat{f_n}) - (u_h, \sigma_h, \widehat{u}_h, \widehat{f}_{n,h})\right\|_E$.

The adjoint problem corresponding to Lemma 2 in Section 4.1.4.2 is likewise reduced in 1D to the scalar equation

$$\epsilon v_1'' + \beta v_1' = -g \tag{4.12}$$

with $v_1 \in H_0^1((0,1))$. After multiplying this equation by $\beta v_1'$ and integrate by parts over Ω_h, we can apply Young's inequality to get

$$\frac{\epsilon}{2}\beta v_1'^2\Big|_0^1 + \|\beta v_1'\|_{L^2}^2 \leq \frac{1}{2}\|g\|^2 + \frac{1}{2}\|\beta v_1'\|^2,$$

implying that

$$\|\beta v_1'\|_{L^2}^2 \lesssim \|g\|^2 + \beta\epsilon v_1'(0)^2.$$

Let us restrict ourselves to the cases where v_1 is sufficiently smooth for $v'(0)$ to be well defined. Taking $g = 1$ (corresponding to a piecewise constant approximation) we can solve (4.12) exactly.

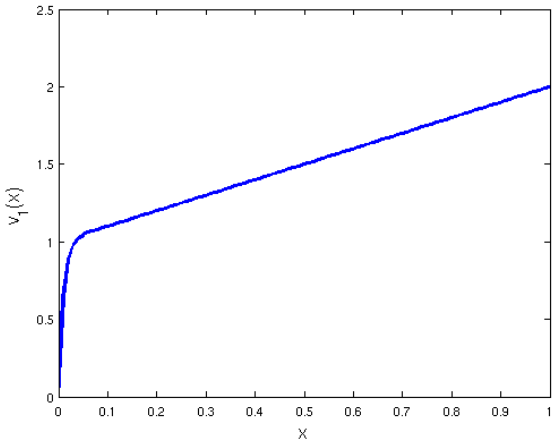


Figure 4.3: $v_1(x) = \frac{e^{-\frac{x}{\epsilon}}}{e^{\frac{1}{\epsilon}}-1}\left(e^{\frac{1}{\epsilon}}\left(e^{\frac{x}{\epsilon}}-1\right)+\left(e^{\frac{1}{\epsilon}}-1\right)e^{\frac{x}{\epsilon}}x\right)$, the solution to the adjoint equation for $f=0$ and constant β and load g for $\epsilon = .01$.

The solution v_1 is plotted in Figure 4.3, where we can see that $v_1(x)$ develops strong boundary layers of width ϵ near the inflow boundary $x=0$. Consequently, $\frac{\epsilon}{2}v_1'(0)^2 \approx \epsilon^{-1}$. Thus, we cannot conclude $\|\beta v'\| \lesssim \|g\|$ when g is a constant,[5] and as a consequence cannot conclude that the robust error bound $\|u-u_h\|_{L^2} \lesssim \|(u,\sigma,\widehat{u},\widehat{f}_n)-(u_h,\sigma_h,\widehat{u}_h,\widehat{f}_{n,h})\|_E$ holds for the solution u_h. More detailed 1D error bounds for Dirichlet boundary conditions are provided in [28], and indicate the same lack of robustness under the test norm used thus far.[6]

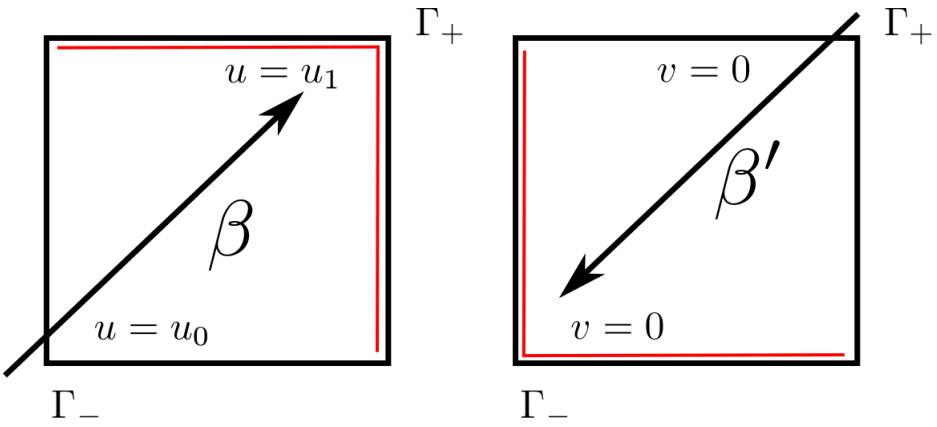
In higher dimensions, the adjoint problem is of the same form as the primal problem with the direction of convection reversed. However, the primal problem determines adjoint boundary conditions on Γ_- and Γ_+. Thus, wheareas for the primal problem, data is convected from the inflow

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

[6]Demkowicz and Heuer proved in [29] that for Dirichlet boundary conditions, robustness as $\epsilon \to 0$ is achieved by the test norm

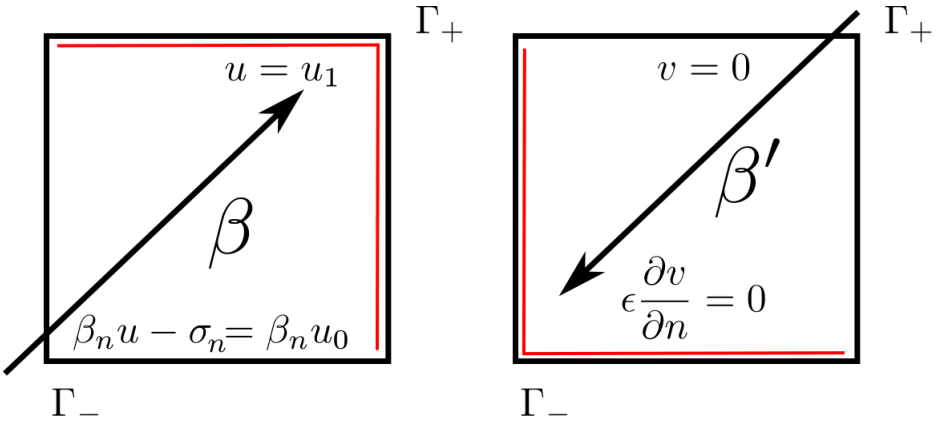
$$\|(\tau,v)\|^2_{V,w} = \|v\| + \epsilon\|\nabla v\| + \|\beta\cdot\nabla v\|_{w+\epsilon} + \|\nabla\cdot\tau\|_{w+\epsilon} + \frac{1}{\epsilon}\|\tau\|_{w+\epsilon}$$

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



(a) Primal problem, Dirichlet inflow BC

(b) Adjoint problem, Dirichlet inflow BC



(c) Primal problem, new inflow BC

(d) Adjoint problem, new inflow BC

Figure 4.4: Comparison of primal and adjoint problems under both the standard Dirichlet and the new inflow boundary condition. The outflow boundary for each problem is denoted in red. For the standard Dirichlet inflow condition, the solution to the adjoint problem can develop strong boundary layers at the outflow of the adjoint problem. Notice, under the new inflow conditions, the relaxation of a wall-stop boundary condition with a zero-stress condition at the outflow boundary of the adjoint problem.

to the outflow, in the adjoint problem, data is convected from the outflow to the inflow boundary instead.

We can intuitively explain the loss of robustness under our derived test norm by the presence of the Dirichlet boundary condition on v at the inflow boundary. Since the direction of convection

is reversed in the adjoint equation, we can interpret the adjoint as representing the convection of a concentration v from the outflow to the inflow boundary. In the presence of a Dirichlet boundary condition at the inflow, v can develop strong boundary layers at the inflow. As a consequence, the quantities $\|\beta \cdot \nabla v\|$ and $\sqrt{\epsilon}\|\nabla v\|$ are no longer robustly bounded by $\|f\|$ and $\|g\|$, and we can no longer derive robust bounds on the error $\|u - u_h\|_{L^2}$ by the error in the energy norm.

Recall our strategy for analysis was to decompose of (v, τ) into continuous and discontinuous portions. Mathematically speaking, the use of Dirichlet boundary conditions on the primal problem introduces strong boundary layers into the solution v of the adjoint equation — in other words, boundary layers are introduced into the continuous portions of our decomposition of (v, τ).[7] The new inflow boundary condition on the primal problem relaxes the wall boundary condition induced on the adjoint/dual problem with a boundary condition that does not generate boundary layers, resulting in stronger stability estimates for the adjoint, and a better result for the primal problem.

4.2 Numerical experiments: Eriksson-Johnson problem

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 [29]; 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 $H(\text{div}; \Omega) \times H^1(\Omega)$ norm is used for the test norm $\| (\tau, v) \|_V$ (as was done in [26]), for an element of size h, $\|v\|^2_{L^2} = O(h)$, while $\|\nabla v\|^2_{L^2} = O(h^{-1})$. As $h \to 0$, the seminorm

[7] The boundary conditions do not introduce boundary layers into the actual computed test functions. However, an interesting phenomenon observed is that, for small ϵ, a lack of robustness can manifest itself during numerical experiments as additional refinements near the inflow boundary, precisely where the continuous parts of the decomposition of (v, τ) develop boundary layers.

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 negligible error in the approximation 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 very thin boundary layers by introducing elements of size ϵ through adaptivity. The effects of roundoff error were successfully treated in [28] by dynamically rescaling the test norms based on element size, a practical remedy not covered yet by the present analysis.

To confirm our theoretical results, we adopt a modification of a problem first proposed by Eriksson and Johnson in [35]. For the choice of $\Omega = (0,1)^2$, $f = 0$, 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

$$u - \sigma_x = u_0 - \sigma_{x,0}, \quad x = 0,$$

$$\sigma_y = 0, \quad y = 0, 1,$$

$$u = 0, \quad x = 1.$$

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

$$r_{1,2} = \frac{1 \pm \sqrt{1 + 4\epsilon\lambda_n}}{2\epsilon},$$

$$\lambda_n = n^2\pi^2\epsilon.$$

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

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

All computations have been done using the adaptive DPG code Camellia, built on the Sandia toolbox Trilinos [51].

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

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

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 $\Delta p \leq 3$, and qualitatively the same for $\Delta p = 2$. 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.

We are especially interested in the ratio of energy error and total L^2 error in both σ and u, which we denote as $\|u - u_h\|_{L^2}$. The bounds on $\|\cdot\|_E$ presented in Section 4.1.4.4 imply that, using the above test norm, $\|u - u_h\|_{L^2}/\|u - u_h\|_E \leq C$ independent of ϵ. Figure 4.7, which plots the ratio of L^2 to energy error, seems to imply that (at least for this model problem) $C = O(1)$. Additionally,

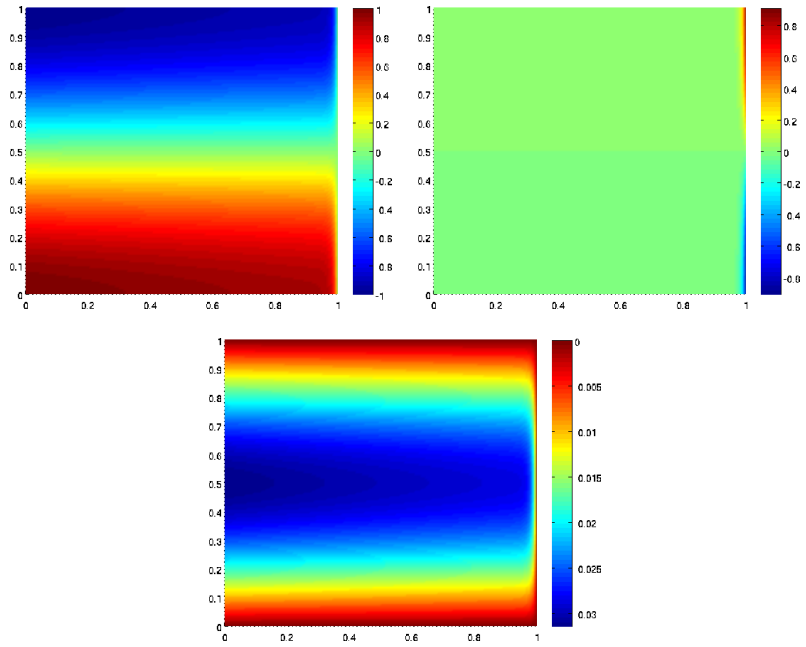


Figure 4.5: Solution for u, σ_x, and σ_y for $\epsilon = .01$, $C_1 = 1$, $C_n = 0$, $n \neq 1$

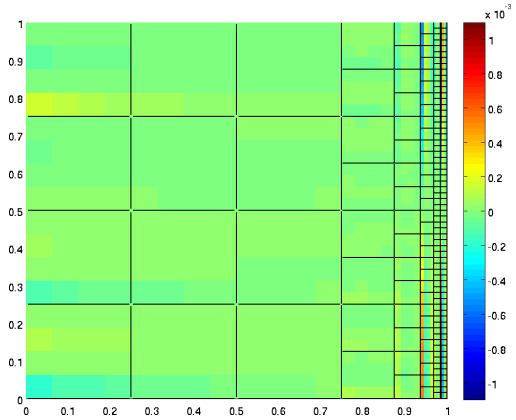


Figure 4.6: Adapted mesh and pointwise error for $\epsilon = .01$

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.

The effect of a mesh dependent scalings on the $\|v\|^2$ and $\|\tau\|^2$ terms in the test norm can

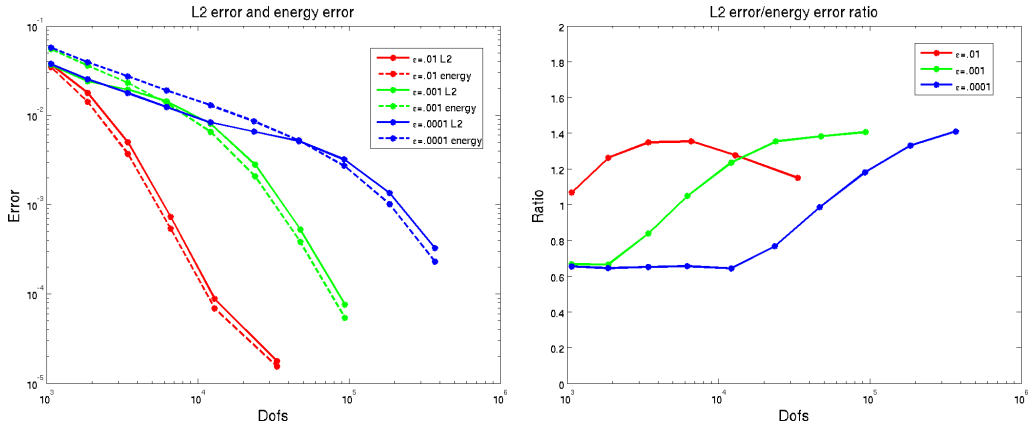


Figure 4.7: L^2 and energy errors, and their ratio for $\epsilon = .01$, $\epsilon = .001$, $\epsilon = .0001$

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 [29], 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 indicated by the rates and ratio between L^2 and energy error in Figure 4.8 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,

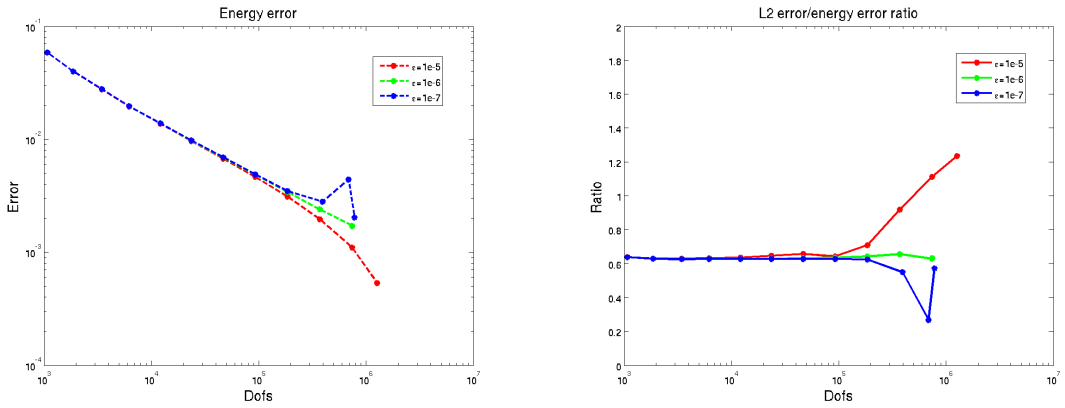


Figure 4.8: Energy error and L^2/energy error ratio for $\epsilon = 1e-5$, $\epsilon = 1e-6$, $\epsilon = 1e-7$. Non-monotonic behavior of the energy error indicates conditioning issues and roundoff effects.

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 mesh-independent test norm[8], we expect monotonic decrease of the energy error with mesh refinement. Non-monotonic behavior indicates either approximation or roundoff error, and as we observed 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 when these phenomena are observed.

It is worth noting that for $\epsilon \leq 1e-5$, we do not perform enough refinements to completely

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

resolve the boundary layer, so $|K| \geq \epsilon$ for all $K \in \Omega_h$. Thus, any roundoff effects observed are not due to the conditioning issues associated with the differing scales of the $\|v\|_{L^2(K)}$ and $\|\nabla v\|_{L^2(K)}$ terms discussed previously.

4.2.2 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 4.9 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.

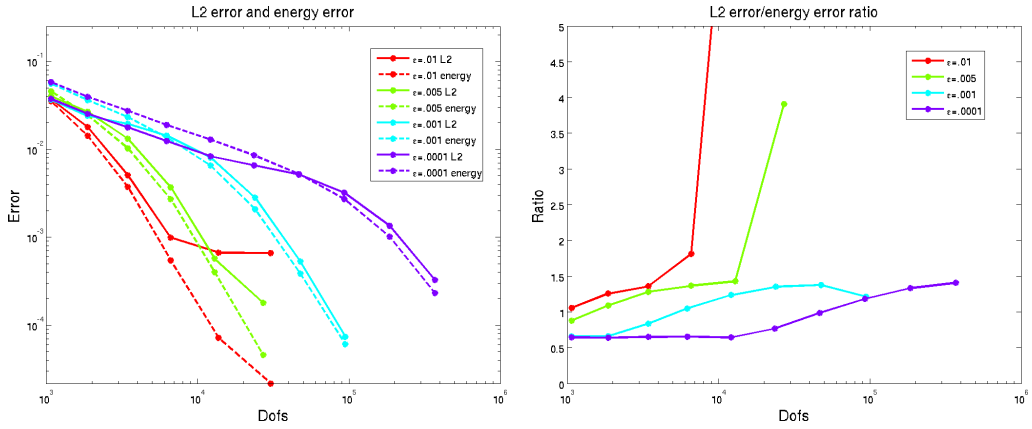


Figure 4.9: L^2 and energy errors and their ratio when neglecting σ_n at the inflow.

4.2.3 Discontinuous inflow data

We note also that an additional advantage of selecting this new boundary condition is a relaxation of regularity requirements: as $\widehat{f_n} \in H^{-1/2}(\Gamma_h)$, 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.

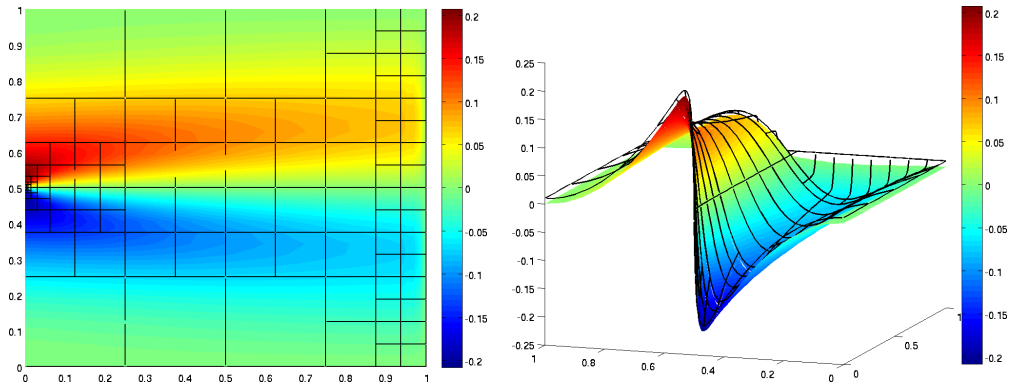


Figure 4.10: Solution variables u and $\widehat{u}$ with discontinuous inflow data u_0 for $\epsilon = .01$.

Figure 4.10 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 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.

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

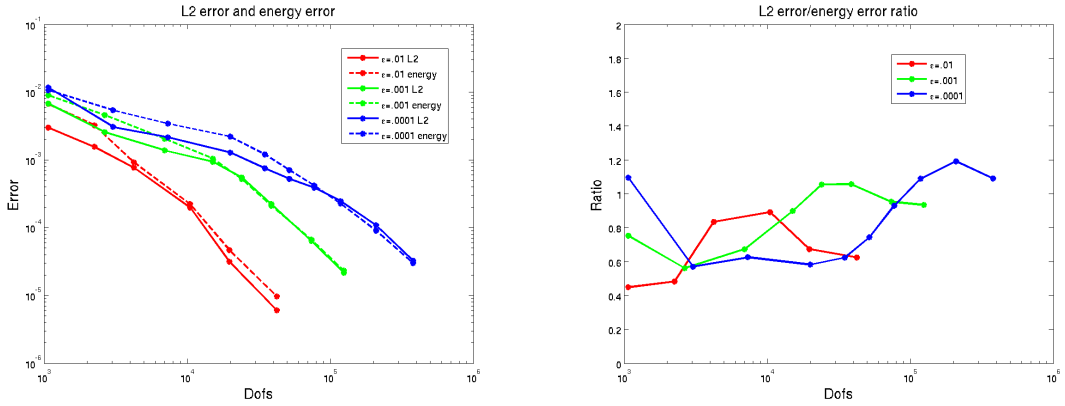


Figure 4.11: L^2 and energy errors, and their ratio for $\epsilon = .01$, $\epsilon = .001$, $\epsilon = .0001$, with discontinuous u_0 approximated by a Fourier expansion.

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.

Chapter 5

Proposed work

5.1 DPG for nonlinear problems

In this chapter, we extend DPG to the nonlinear setting and apply it to two problems in computational fluid dynamics. We take as our starting point a nonlinear variational formulation $b(u,v) = l(v)$, which is linear in v, but not in u. An appropriate linearization gives

$$b_u(\Delta u, v) = l(v) - b(u,v),$$

where $b_u(\Delta u, v)$ is the linearization of $b(u,v)$ with respect to u. Let $B(u)$ and $B_u\Delta u$ be the variational operators associated with $b(u,v)$ and $b_u(\Delta u, v)$, respectively. We define two additional measures:

$$\|\Delta u\|_E := = \|B_u\Delta u\|_{V'} = \left\|R_V^{-1}B_u\Delta u\right\|_V = \sup_{v\in V}\frac{b_u(\Delta u, v)}{\|v\|_V}$$
$$\|R(u)\|_E := = \|B(u)-l\|_E = \|B(u)-l\|_{V'} = \left\|R_V^{-1}B(u)-l\right\|_V = \sup_{v\in V}\frac{b(u,v)-l(v)}{\|v\|_V}$$

These two quantities are measures of the linearized update Δu and the nonlinear residual in the appropriate norm in the dual space V'. The first will be used to measure convergence of a nonlinear solution scheme to a stable discrete solution, while the second will be used to assess the convergence of the discrete solution to the continuous solution.

5.1.1 Nonlinear solution strategies

The solution of a nonlinear problem is most commonly found using an iterative method, where a series of solutions to linear problems is expected to converge to the nonlinear solution. We use two main methods to iterate to a nonlinear solution.

- **(Damped) Newton iteration :** Given the linearized system $b_u(\Delta u, v) = b(u, v) - l(v)$, we begin with some initial guess $u \coloneqq u_0$ and solve for Δu_0. The process is then repeated with $u \coloneqq u_{i+1} \coloneqq u_i + \alpha_i \Delta u_i$, where $\alpha_i \in (0, 1]$ is some damping parameter that may limit the size of the Newton step in order to optimize the rate of convergence or preserve physicality of the solution. The solution is considered converged when $\|\Delta u\|_E \leq tol$.

- **Pseudo-time stepping:** An alternative method for the solution of steady-state systems is to use a pseudo-timestepping method. The most common approach is to discretize the equations in time using a stable, implicit method — if $Lu = f$ is our nonlinear problem and L_u is the linearization of the nonlinear operator L with respect to u, then the pseudo-timestepping method solves at each discrete time t_i

$$\frac{\partial u}{\partial t} + Lu = f \rightarrow \frac{u(t_i) - u(t_{i-1})}{\Delta t} + L_{u(t_i)} \Delta u(t_i) = (f - Lu(t_i)).$$

  The solution at the next timestep is then set $u(t_{i+1}) \coloneqq u(t_i) + \Delta u(t_i)$. This procedure is then repeated for the next timestep t_{i+1} until the transient residual decreases such that $\|u(t_i) - u(t_{i-1})\|_{L^2(\Omega)} = \|\Delta u(t_i)\|_{L^2(\Omega)} \leq tol$.[1]

In practice, most compressible flow solvers opt for the pseudo-time step method over the direct Newton iteration due to the difficulty of convergence and sensitivity of the Newton iteration

[1] Strictly speaking, seeking the solution at each timestep involves the solution of a nonlinear problem, requiring a Newton-type iteration to solve for $u(t_i)$. However, for most applications, it is sufficient to approximate the nonlinear solution using a single Newton solve at each time step.

to initial guess[43]. Though the convergence of the pseudo-time step is slower, the addition of the zero-order transient terms "regularizes" the problem and makes it less difficult to solve.[2]

A second class of nonlinear solvers are optimization methods. Since DPG allows for the formulation of a discrete nonlinear residual, it is possible to formulate the nonlinear DPG problem as a minimization problem and use an optimization method to solve the discrete nonlinear problem. This approach has been successfully implemented by Peraire et al. in solving compressible gas dynamics problems on uniform grids using a modified version of the ultra-weak variational formulation [45]. An additional advantage of such an approach would be the more direct enforcement of physical constraints, which are treated in an ad-hoc manner in most compressible Navier-Stokes solvers.

5.1.2 DPG as a nonlinear minimum residual method

A recent theoretical development is the formulation of a DPG method that aims to minimize a nonlinear residual. Given two Hilbert spaces — a trial space U and test space V — our nonlinear variational formulation can be written as $b(u, v) = l(v)$, with the corresponding operator form of the formulation in V'

$$B(u) = l.$$

We can apply the steps used to derive the DPG method for linear problems to the nonlinear setting as well. Given a finite dimensional subspace $U_h \subset U$, we consider the discrete nonlinear residual

$$J(u_h) \coloneqq \frac{1}{2} \left\| R_V^{-1} \left(B(u_h) - l \right) \right\|_V^2 .$$

Our goal is to solve

$$u_h = \arg \min_{w_h \in U_h} J(w_h).$$

[2]The addition of a zero-order term "regularizes" an equation by adding to it a positive-definite L^2 projection operator. In the limit as $\Delta t \to 0$, the solution at t_i will simply return the L^2 projection of the solution at the previous timestep.

Similarly to the linear case, we can take the Gateáux derivative to arrive at a necessary condition for u_h to minimize $J(u_h)$

$$\langle J'(u_h), \delta u_h \rangle = \left(R_V^{-1} \left(B(u_h) - l \right), R_V^{-1} B'(u_h; \delta u_h) \right)_V, \quad \delta u_h \in U_h.$$

As the above is a nonlinear equation, we seek its solution through linearization. Differentiating a second time in u, we arrive at

$$\begin{aligned}
\langle J''(u_h), \Delta u_h \rangle =& \langle B'(u_h; \Delta u_h), B'(u_h; \delta u_h) \rangle_V \\
&+ \langle \left(B(u_h) - l \right), B''(u_h; \delta u_h, \Delta u_h) \rangle_V \\
=& \left(R_V^{-1} B'(u_h; \Delta u_h), R_V^{-1} B'(u_h; \delta u_h) \right)_V \\
&+ \left(R_V^{-1} \left(B(u_h) - l \right), R_V^{-1} B''(u_h; \delta u_h, \Delta u_h) \right)_V
\end{aligned}$$

where $B''(u_h; \delta u_h, \Delta u_h)$ denotes the Hessian of $B(u_h)$, evaluated using both δu_h and Δu_h.

Examining the above formulation, we note that DPG as applied to the linearized problem produces the term $\left(R_V^{-1} B'(u_h; \Delta u_h), R_V^{-1} B'(u_h; \delta u_h) \right)_V$. However, in approaching the nonlinear problem through the minimization of the discrete residual, we gain a second-order term involving the Hessian

$$\left(R_V^{-1} \left(B(u_h) - l \right), R_V^{-1} B''(u_h; \delta u_h, \Delta u_h) \right)_V.$$

The evaluation of this term can be done in a computationally efficient manner — if we define the image of the nonlinear residual under the Riesz inverse

$$v_{R(u)} = R_V^{-1} \left(B(u_h) - l \right),$$

then we can compute this additional term through

$$\left(v_{R(u)}, R_V^{-1} B''(u_h; \delta u_h, \Delta u_h) \right)_V = \left\langle v_{R(u)}, B''(u_h; \delta u_h, \Delta u_h) \right\rangle_V$$

which can be computed in the same fashion as a Bubnov-Galerkin stiffness matrix. This addition, though not positive definite, is symmetric due to the nature of second order derivatives.

We note that we have not implemented this Hessian-based nonlinear solver in the numerical experiments to follow, and instead plan to do so in the proposed work outlined in Section 5.4.

5.2 The viscous Burgers equation

We will illustrate the application of DPG to nonlinear problems using a viscous Burgers' equation on domain $\Omega = [0,1]^2 \in \mathbb{R}^2$

$$\frac{\partial\left(u^2/2\right)}{\partial x} + \frac{\partial u}{\partial y} - \epsilon\Delta u = f.$$

If we remove the viscous term, the above problem reduces to the form of the 1D transient Burgers equation, whose solution we can determine via the method of characteristics. For boundary conditions

$$u(x,y) = 1 - 2x, \quad x = 0, y = 0,$$

the solution forms a shock discontinuity starting at $(x,y) = (.5,.5)$, which then propagates upward in the y-direction. The addition of the viscous term smears this discontinuity, leading to a solution with a smooth shock of width ϵ.

We begin by writing the equation as a first order system. Defining $\beta(u) = (u/2, 1)$, the above Burgers equation can be written as

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

Analogously to the convection-diffusion problem, the DPG nonlinear variational formulation can

then be given

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

Linearizing the above then gives us

$$\begin{aligned}b_u\left(\left(\Delta u,\sigma,\widehat{u},\widehat{f}_n\right),(v,\tau)\right)&=\left(\Delta u,\nabla\cdot\tau-\begin{bmatrix}u\\1\end{bmatrix}\cdot\nabla v\right)+\left(\sigma,\frac{1}{\epsilon}\tau+\nabla v\right)+\langle\widehat{u},\tau\cdot n\rangle+\left\langle\widehat{f}_n,v\right\rangle\\&=(u,\nabla\cdot\tau-\beta(u)\cdot\nabla v)\end{aligned}$$

Notice that the nonlinear term is only dependent on u, and thus there is no need to linearize in the variables σ, $\widehat{u}$, and $\widehat{f}_n$.

Since the linearized Burgers' equation is of the form of a convection-diffusion problem with non-homogeneous load, we adopt the test norm described in Section 4.1.4 with convection vector $\beta=(u,1)$.

Recall that we did not linearize in the flux variables $\widehat{u}$ and $\widehat{f}_n$, so we can directly apply the nonlinear boundary conditions to our variational formulation. Additionally, since Burgers' equation does not have any physical constraints, we can employ a direct Newton iteration to solve the nonlinear equation. The adaptivity algorithm is identical to the greedy algorithm described previously, except that the linear solve is replaced by a nonlinear solve. The results of an adaptive simulation are shown in Figure 5.1 for $\epsilon=1e-4$.

5.3 The compressible Navier-Stokes equations

We briefly review the compressible Navier-Stokes equations, given in Section 2.1, and formulate DPG for the nonlinear system.

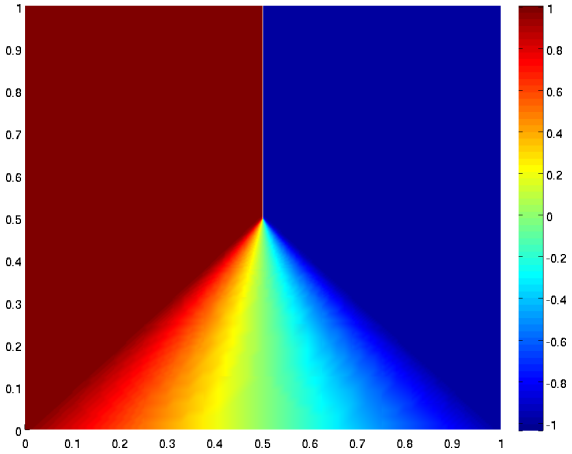


Figure 5.1: Shock solution for Burgers' equation with $\epsilon = 10^{-4}$.

- **Conservation equations**

$$\begin{aligned}
\nabla \cdot \begin{bmatrix} \rho u_1 \\ \rho u_2 \end{bmatrix} &= 0 \\
\nabla \cdot \begin{bmatrix} \rho u_1^2 + p \\ \rho u_1 u_2 \end{bmatrix} &= \nabla \cdot (\vec{\sigma_{i1}}) \\
\nabla \cdot \begin{bmatrix} \rho u_1 u_2 \\ \rho u_2^2 + p \end{bmatrix} &= \nabla \cdot (\vec{\sigma_{i2}}) \\
\nabla \cdot \begin{bmatrix} ((\rho e) + p) u_1 \\ ((\rho e) + p) u_2 \end{bmatrix} &= \nabla \cdot [\boldsymbol{\sigma} U + \vec{q}] \,,
\end{aligned}$$

  where $\boldsymbol{\sigma}$ is the stress tensor whose ijth term is σ_{ij}.

- **Newtonian fluid laws**

  We represent $\boldsymbol{\sigma}$ using the Newtonian fluid law

$$\sigma_{ij} = \mu(u_{i,j} + u_{j,i}) + \lambda u_{k,k} \delta_{ij}$$

  where μ is viscosity and λ is bulk viscosity. We can invert the stress tensor under isotropic and plane strain assumptions to get

$$\frac{1}{2}\left(\nabla U + \nabla^T U\right) = \frac{1}{2\mu}\sigma_{ij} - \frac{\lambda}{4\mu(\mu + \lambda)}\sigma_{kk}\delta_{ij}$$

We also have

$$\frac{1}{2}\left(\nabla U + \nabla^T U\right) = \nabla U - \boldsymbol{\omega}$$

where $\boldsymbol{\omega}$ is the antisymmetric part of the infinitesimal strain tensor:

$$\boldsymbol{\omega} = \frac{1}{2}\left(\nabla U - \nabla^T U\right).$$

Thus our final form is

$$\nabla U - \boldsymbol{\omega} = \frac{1}{2\mu}\boldsymbol{\sigma} - \frac{\lambda}{4\mu(\mu+\lambda)}\mathrm{tr}(\boldsymbol{\sigma})\boldsymbol{I}.$$

Notice that $\boldsymbol{\omega}$ is implicitly defined to be the symmetric part of ∇u by taking the symmetric part of the above equation.

We note that, though this is a standard approach in solid mechanics, it is nonstandard compared to the usual finite element and DG approaches to the viscous stresses. We adopt such an approach to better mirror our experiences with the convection-diffusion equation [29, 18].

- **Fourier's heat conduction law**

  We assume Fourier's law

  $$\vec{q} = \kappa \nabla T,$$

  We introduce here the Prandtl number here as well

  $$\mathrm{Pr} = \frac{\gamma c_v \mu}{\kappa}.$$

  In this case, we assume a constant Prandlt number, which implies that the heat conductivity κ is proportional to viscosity μ.

5.3.1 Nondimensionalization

To nondimensionalize our equations, we introduce nondimensional quantities for length, density, velocity, temperature, and viscosity.

$$\boldsymbol{x}^* = \frac{\boldsymbol{x}}{L}, \qquad \rho^* = \frac{\rho}{\rho_\infty}, \qquad u_1^* = \frac{u_1}{V_\infty}, \qquad u_2^* = \frac{u_2}{V_\infty}, \qquad T^* = \frac{T}{T_\infty}, \qquad \mu^* = \frac{\mu}{\mu_\infty}$$

Pressure, internal energy, and bulk viscosity are then nondimensionalized with respect to the above variables

$$p^* = \frac{p}{\rho_\infty V_\infty^2}, \qquad \iota^* = \frac{\iota}{V_\infty^2}, \qquad \lambda^* = \frac{\lambda}{\mu_\infty}$$

We introduce, for convenience, the Reynolds number

$$\mathrm{Re} = \frac{\rho_\infty \mathrm{V}_\infty \mathrm{L}}{\mu_\infty}$$

and the reference (free stream) Mach number

$$M_\infty = \frac{V_\infty}{\sqrt{\gamma(\gamma-1)c_v T_\infty}}$$

Note that

$$a = \sqrt{\frac{\gamma p_\infty}{\rho_\infty}} = \sqrt{\gamma p_\infty} = \sqrt{\gamma(\gamma-1)c_v T_\infty}$$

The equations take the same form as before after nondimensionalization, so long as we define new material constants

$$\tilde{\mu} = \frac{\mu^*}{\mathrm{Re}}, \qquad \tilde{\lambda} = \frac{\lambda^*}{\mathrm{Re}}, \qquad \tilde{c}_v = \frac{1}{\gamma(\gamma-1)M_\infty^2}, \qquad \tilde{\kappa} = \frac{\gamma \tilde{c}_v \tilde{\mu}}{\mathrm{Pr}}$$

From here on, we drop the * superscript and assume all variables refer to their nondimensionalized quantities.

To summarize, our system of equations in the classical variables is now

$$\begin{aligned}
\nabla \cdot \begin{bmatrix} \rho u \\ \rho v \end{bmatrix} &= 0 \\
\nabla \cdot \left(\begin{bmatrix} \rho u^2 + p \\ \rho uv \end{bmatrix} - \boldsymbol{\sigma}_1 \right) &= 0 \\
\nabla \cdot \left(\begin{bmatrix} \rho uv \\ \rho v^2 + p \end{bmatrix} - \boldsymbol{\sigma}_2 \right) &= 0 \\
\nabla \cdot \left(\begin{bmatrix} ((\rho e) + p)u \\ ((\rho e) + p)v \end{bmatrix} - \boldsymbol{\sigma}\mathbf{u} + \vec{q} \right) &= 0 \\
\frac{1}{2\mu}\boldsymbol{\sigma} - \frac{\lambda}{4\mu(\mu+\lambda)}\mathrm{tr}(\boldsymbol{\sigma})\boldsymbol{I} &= \nabla \mathbf{u} - \mathrm{Re}\,\boldsymbol{\omega} \\
\frac{1}{\kappa}\vec{q} &= \nabla T
\end{aligned}$$

We strongly enforce symmetry of the stress tensor $\boldsymbol{\sigma}$ by setting $\sigma_{21} = \sigma_{12}$. Additionally, we have scaled the antisymmetric tensor $\boldsymbol{\omega}$ by the Reynolds number to ensure that $\boldsymbol{\omega} = O(1)$ for all ranges of Re.

5.3.2 Linearization

5.3.2.1 Conservation laws

The Navier-Stokes conservation laws can be written as

$$\begin{aligned}
\nabla \cdot \begin{bmatrix} \rho u \\ \rho v \end{bmatrix} &= 0 \\
\nabla \cdot \left(\begin{bmatrix} \rho u^2 + p \\ \rho uv \end{bmatrix} - \boldsymbol{\sigma}_1 \right) &= 0 \\
\nabla \cdot \left(\begin{bmatrix} \rho uv \\ \rho v^2 + p \end{bmatrix} - \boldsymbol{\sigma}_2 \right) &= 0 \\
\nabla \cdot \left(\begin{bmatrix} ((\rho e) + p)u \\ ((\rho e) + p)v \end{bmatrix} - \boldsymbol{\sigma}_1 \cdot \boldsymbol{u} - \boldsymbol{\sigma}_2 \cdot \boldsymbol{u} + \vec{q} \right) &= 0
\end{aligned}$$

or generally,

$$\nabla \cdot (F_i(\boldsymbol{U}) - G_i(\boldsymbol{U}, \boldsymbol{\sigma}) = 0, \qquad i = 1, \ldots, 4$$

The variational form restricted to a single element gives

$$\langle \widehat{F}_i \cdot n, v \rangle - \int_K (F(\boldsymbol{U}) - G_i(\boldsymbol{U}, \boldsymbol{\sigma})) \cdot \nabla v_i = 0, \qquad i = 1, \ldots, 4$$

and the variational form over the entire domain is given by summing up the element-wise contributions.

The presence of terms such as $\boldsymbol{\sigma}_i \cdot \boldsymbol{u}$ means that we will need to linearize in the stress variables $\sigma_i j$ in addition to our Eulerian quantities. Since fluxes and traces are linear, we do not need to linearize them. Instead, fluxes $\widehat{F}_{i,n}$ and traces $\widehat{u}, \widehat{v}, \widehat{T}$ will represent normal traces and traces of the accumulated nonlinear solution. The linearized variational formulation is thus

$$\begin{aligned} \langle \widehat{F}_i \cdot n, v \rangle - \int_K & (F_{i,\boldsymbol{U}}(\boldsymbol{U}) \cdot \Delta \boldsymbol{U} - G_{i,\boldsymbol{U}}(\boldsymbol{U}, \boldsymbol{\sigma}) \cdot \Delta \boldsymbol{U} - G_{i,\boldsymbol{\sigma}}(\boldsymbol{U}, \boldsymbol{\sigma}) \cdot \Delta \boldsymbol{\sigma}) \cdot \nabla v_i \\ &= \int_K (F_i(\boldsymbol{U}) - G_i(\boldsymbol{U})) \cdot \nabla v_i \\ i &= 1, \ldots, 4 \end{aligned}$$

where $F^i_{j,\boldsymbol{U}}$, $G^i_{j,\boldsymbol{U}}$, and $G^i_{j,\boldsymbol{\sigma}}$ are the Eulerian and two viscous Jacobians (linearized w.r.t. the Eulerian/viscous variables), respectively.

5.3.2.2 Viscous equations

We have two equations left to linearize - the constitutive laws defining our viscous stresses and heat flux terms.

$$\begin{aligned} \frac{1}{2\mu}\boldsymbol{\sigma} - \frac{\lambda}{4\mu(\mu+\lambda)}\mathrm{tr}(\boldsymbol{\sigma})\boldsymbol{I} + \mathrm{Re}\boldsymbol{\omega} &= \nabla \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \\ \frac{1}{\kappa} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} &= \nabla T \end{aligned}$$

We treat the first tensor equation as two vector equations by considering each column:

$$\frac{1}{2\mu}\begin{bmatrix}\sigma_{11}\\ \sigma_{12}\end{bmatrix} - \frac{\lambda}{4\mu(\mu+\lambda)}\begin{bmatrix}\sigma_{11}+\sigma_{22}\\ 0\end{bmatrix} + \mathrm{Re}\begin{bmatrix}0\\ -\omega\end{bmatrix} - \nabla \mathrm{u}_1 = 0$$
$$\frac{1}{2\mu}\begin{bmatrix}\sigma_{12}\\ \sigma_{22}\end{bmatrix} - \frac{\lambda}{4\mu(\mu+\lambda)}\begin{bmatrix}0\\ \sigma_{11}+\sigma_{22}\end{bmatrix} + \mathrm{Re}\begin{bmatrix}\omega\\ 0\end{bmatrix} - \nabla \mathrm{u}_2 = 0$$

Since all equations are linear in variables q_1, q_2, w for all combinations of variables, we do not need to linearize any equations in q_1, q_2, w.

We do not linearize the viscosities μ and λ, but instead set them based on the power law and the solution at the previous timestep for simplicity.

5.3.3 Test norm

Recall the convection-diffusion problem

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

On domain Ω, with mesh Ω_h and mesh skeleton Γ_h, the DPG variational formulation is

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

with $v \in H^1$ and $\tau \in H(\mathrm{div}, \Omega_h)$. The test norm adopted for convection-diffusion in Section 4.1.4 and in [18] is defined elementwise on K as

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

This test norm both delivers robust control of the error in the L^2 variables u and σ and avoids boundary layers in the computation of local test functions.

Our test norm is extrapolated to the Navier-Stokes equations as follows: we denote the vector of H^1 test functions as $v = \{v_1, v_2, v_3, v_4\}$, and similarly for $W = \{\tau_1, \tau_2, \tau_3\}$. Similarly,

we group our Eulerian and stress variables into the vector variables U and Σ, respectively. If $R_{\text{Euler}}(U,\Sigma)$ and $R_{\text{visc}}(U,\Sigma)$ are Eulerian and viscous nonlinear residuals, our formulation for the linearized Navier-Stokes equations can be written as

$$\nabla \cdot (A_{\text{Euler}} U - A_{\text{visc}} \Sigma) = R_{\text{Euler}}(U,\Sigma)$$

$$E_{\text{visc}} \Sigma - \nabla U = R_{\text{visc}}(U,\Sigma)$$

with variational formulation

$$\left\langle \widehat{F}_n, V \right\rangle_{\Gamma_h} + \left(U, \nabla \cdot W - A_{\text{Euler}}^T \nabla V\right) + \left\langle \widehat{U}, W \cdot n \right\rangle_{\Gamma_h} + \left(\Sigma, E_{\text{visc}}^T W - A_{\text{visc}}^T \nabla V\right) = 0$$

Identifying vector-valued terms in the Navier-Stokes formulation with equivalent scalar terms in the convection-diffusion equation allows us to extrapolate our test norm to systems of equations

$$\begin{aligned} \| (V,W) \|_{V,K}^2 = & \min\left\{ \frac{\text{Re}}{|K|}, 1 \right\} \|v\|^2 + \frac{1}{\text{Re}} \|A_{\text{visc}}^T \nabla v\|^2 + \|A_{\text{Euler}}^T \nabla v\|^2 \\ & + \|\nabla \cdot W\|^2 + \min\left\{ \text{Re}, \frac{1}{|\text{K}|} \right\} \|E_{\text{visc}}^T \tau\|^2. \end{aligned}$$

An advantage of this extrapolation approach is that the incompletely parabolic nature of the Navier-Stokes equation is taken into account; there is no diffusive term present in the mass conservation equation, and the test norm reflects that by requesting only limited regularity of v_1.[3]

5.3.4 Boundary conditions

As a consequence of the ultra-weak variational formulation, our solution is linear in the flux and trace variables. Thus, the nonlinear boundary conditions can be applied directly to our fluxes $\widehat{f}_{i,n}, i = 1, \ldots, 4$, and traces $\widehat{u}_1$, $\widehat{u}_2$, and $\widehat{T}$.

[3]The situation is analogous to using the full $H^1(\Omega_h)$ norm for the pure convection equation — the optimal test norm $\|v\|_V = \|\beta \cdot \nabla v\| + \|v\|$ implies only streamline regularity, whereas taking $\|v\|_V = \|\nabla v\| + \|v\|$ implies stronger regularity on the test space V than the graph norm. Consequently, convergence is suboptimal for DPG applied to the convection problem under the $H^1(\Omega_h)$ test norm.

Additionally, inflow boundary conditions are applied not directly to the trace variables $\widehat{u}_1$, $\widehat{u}_2$, and $\widehat{T}$, but to the fluxes $\widehat{f}_{i,n}$. Extrapolating from the convection-diffusion problem, this allows us to use a stronger test norm without experiencing adverse effects for smaller diffusion/higher Reynolds numbers [29, 18].

5.3.5 Numerical experiments: Carter flat plate

For our solver, we use a standard pseudo-time solver and greedy refinement scheme. Though we have not yet implemented adaptive timestepping, we are able to take large enough uniform timesteps over a coarse mesh such that convergence occurs sufficiently quickly.

The numerical parameters used are as follows:

- DPG parameters: $p = 2$ and $\Delta p = 2$ uniformly across the mesh.
- Adaptivity parameters: Energy threshold for refinements is $\alpha = .2$ or $\alpha = .15$.
- Time-stepping parameters: $\Delta t = .1$, and tolerance for transient residual $\epsilon_t = 5e-7$.

Our problem of interest is the Carter flat plate problem. An infinitesmally thin flat plate disrupts a free stream flow and causes a shock to form at the tip of the plate.

- **Inflow boundary conditions:** free stream conditions are applied here to all four fluxes $\widehat{f}_{i,n}$.
- **Symmetry boundary conditions:** $u_n = q_n = \frac{\partial u_s}{\partial n} = 0$. Here, this implies $u_2 = q_2 = \sigma_{12} = 0$. We impose the stress condition by noting that, for the flat plate geometry, if $u_2 = 0$, then at the top and bottom, with $n = (0, 1)$, $\widehat{f}_{2,n} = \sigma_{12}$, and $\widehat{f}_{4,n} = q_2$ if σ_{12} and $u_2 = 0$.
- **Flat plate boundary conditions:** $u_1 = u_2 = 0$, and $T = T_w = \left[1 + (\gamma - 1)M_\infty^2/2\right] T_\infty = 2.8T_\infty$ (for Mach 3 flow). We impose these strongly on the trace variables $\widehat{u}_1, \widehat{u}_2, \widehat{T}$.

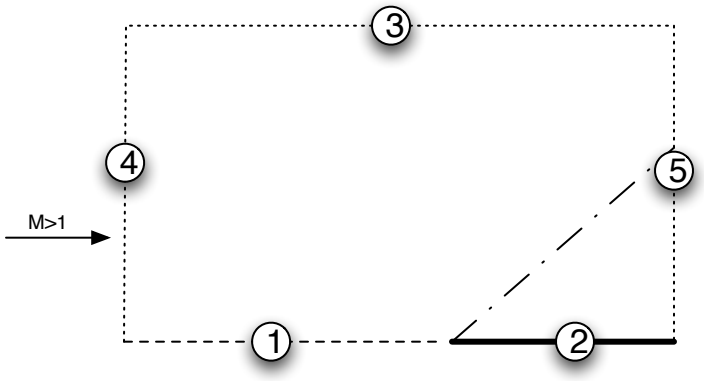


Figure 5.2: Carter flat plate problem.

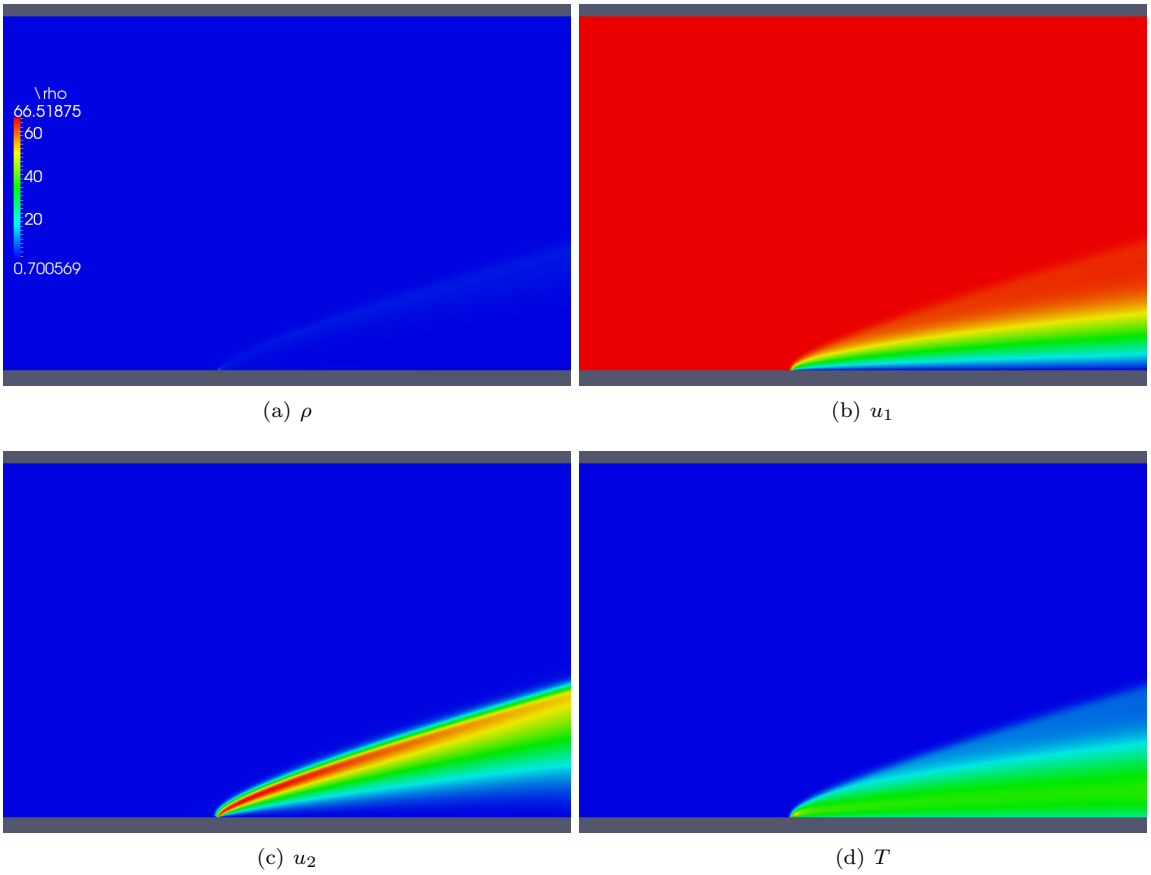
- **Outflow boundary conditions:** the exact boundary conditions to enforce here are not universally agreed on. Many enforce $\frac{\partial u_1}{\partial n} = \frac{\partial u_2}{\partial n} = 0$ and $\frac{\partial T}{\partial n} = 0$, while others enforce an outflow boundary condition only in regions where the flow is subsonic.[31]. We adopt a "no boundary condition" outflow condition, first introduced in [48]. A mathematical analysis and explanation of this boundary condition for standard H^1 elements is given in [37].

We initialize our solution to

$$\rho = 1, \qquad u_1 = 1, \qquad u_2 = 0, \qquad T = 1$$

which is consistent with what was done in Demkowicz and Oden in [31]. Stresses are set uniformly to zero.

We take the computational domain to be $\Omega = [0, 2] \times [0, 1]$. We begin with a mesh of 8 by 16 elements. Under Dirichlet wall boundary conditions for all 3 traces u_1, u_2, and T, the solution develops a singularity in the density ρ at the plate beginning, and both T and u_1 form a boundary layer along the leading edge of the plate. Due to the presence of the singularity, the solution for ρ is scaled such that the features of the solution away from the singularity are visible in Figure 5.4.



(a) ρ (b) u_1

(c) u_2 (d) T

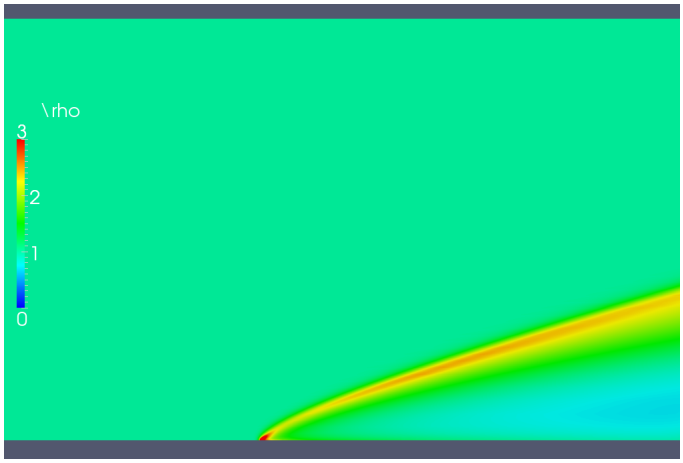
Figure 5.3: Solutions after seven refinements for $p = 2$ and $\mathrm{Re} = 1000$.

5.4 Area requirements

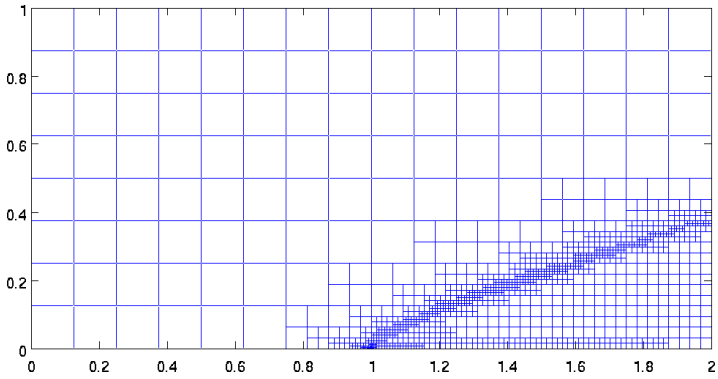
5.4.1 Area A: Applicable mathematics

- **Completed: Prove robustness of DPG method for the scalar convection-diffusion problem.**

  Analysis has been done in [18], extending the results of [29]. In particular, we have introduced a test norm under which the DPG method robustly bounds the L^2 error in the field variables u and the scaled stress σ. Numerical results confirm the theoretical bounds given.



(a) ρ



(b) Mesh

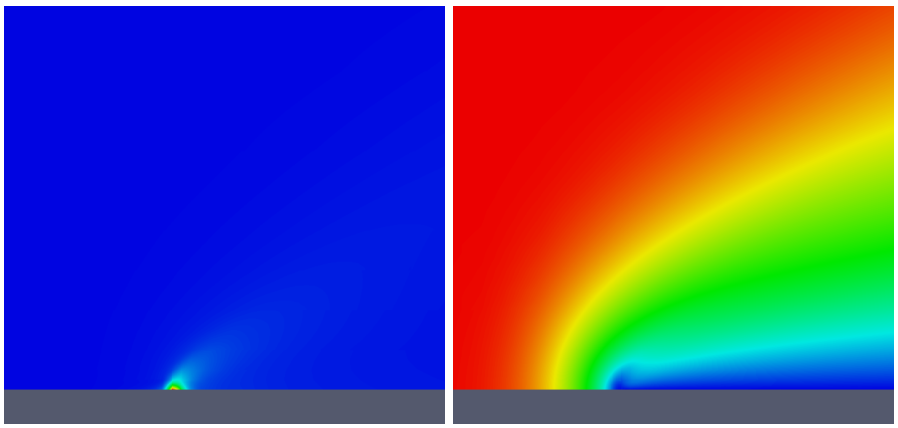
Figure 5.4: Rescaled solution for ρ in the range $[0, 3]$ and adaptive mesh after seven refinements.

- **Proposed: Attempt analysis of the linearized Navier-Stokes system.**

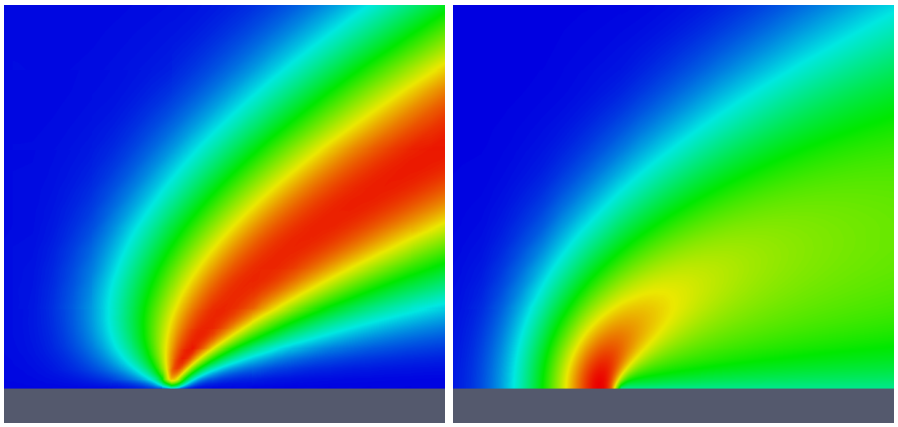
  We hope to analyze the linearized Navier-Stokes equations to determine an optimal extrapolation of the test norm for the scalar convection-diffusion problem to systems.

5.4.2 Area B: Numerical analysis and scientific computation

- **Completed: Collaborative work with Nathan Roberts on the higher order parallel adaptive DPG code Camellia.**



(a) ρ (b) u_1



(c) u_2 (d) T

Figure 5.5: Zoom of solutions at the beginning of the plate for $p = 2$ and $\mathrm{Re} = 1000$.

Most numerical experiments have been done using the higher-order adaptive codebase Camellia, built upon the Trilinos library and designed by Nathan Roberts. The library allows for rapid implementation of problems through a symbolic syntax similar to the Fenics project. Work has been done to support arbitrary-irregularity refinements in both h and p, and the framework for anisotropic refinements is in place as well. Similarly, the code is partially parallelized - the determination of the stiffness matrix is perfectly scalable, and we are currently

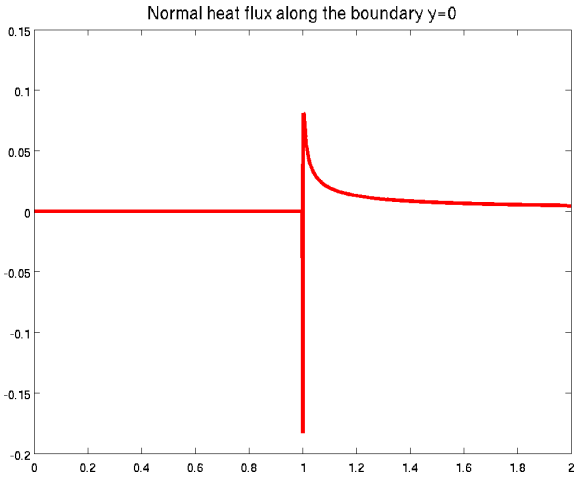


Figure 5.6: Heat flux at the bottom boundary.

exploring options for more scalable solvers.

- **Proposed: Distributed iterative static condensation.**

  A clear choice for a parallelized solver under the ultra-weak variational formulation is static condensation/the Schur-complement method. Given a block matrix structure of a stiffness matrix K, we can view the DPG system as

$$Ku = \begin{bmatrix} A & B \\ B^T & D \end{bmatrix} \begin{bmatrix} u_{\text{flux}} \\ u_{\text{field}} \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix} = l$$

  where D has a block-diagonal structure, and A and D are both square matrices with $\dim A < \dim D$. The system can be reduced to yield the condensed system

$$\left(A - BD^{-1}B^T\right) u_{\text{flux}} = f - BD^{-1}g$$

  where D^{-1} can be inverted block-wise. Once the globally coupled flux degrees of freedom are solved for, the field degrees of freedom can be reconstructed locally.

  Additionally, though the Schur complement $A - BD^{-1}B^T$ is generally a dense matrix, it does not have to be explicitly formed, and we can use a matrix-free method to solve the above

condensed system. The question remains on whether or not the condensed DPG stiffness matrix is well-conditioned. It has been shown that, unlike standard least-squares methods, DPG generates for the Poisson matrix a stiffness matrix with condition number $O(h^{-2})$ [36]. It is well known that, under standard finite element methods, if the condition number of the global stiffness matrix K is $O(h^{-2})$, the condition number of the Schur complement is $O(h^{-1})$. Additionally, through either diagonal preconditioning or matrix equilibration, the condition number of the Schur complement can often be made significantly smaller than $O(h^{-1})$, and the positive-definiteness of the resulting system allows the use of fast iterative solvers in solving the condensed system. We hope to implement such a distributed solver and determine the scalability of this solver under DPG.

- **Proposed: a Nonlinear Hessian-based DPG method.**

  We hope to implement the Hessian-based nonlinear DPG approach to the viscous Burgers and compressible Navier-Stokes problems and compare convergence and error rates for each problem after applying this nonlinear strategy.

- **Proposed: Anisotropic refinements and *hp*-schemes.**

  The effectiveness and necessity of anisotropic refinement schemes for problems in CFD has been demonstrated several times in the literature [9, 1]. As the error representation function has been shown to yield an effective and natural residual with which to drive refinement, we hope to generalize the use of the error representation function to yield anisotropic adaptive schemes. Additionally, we hope to explore the possibility of inferring an optimal choice between h and p refinement using the error representation function.

5.4.3 Area C: Mathematical modeling and applications

- **Completed: convection-dominated diffusion, Burgers, and a model problem for Navier-Stokes.**

  I have applied the DPG method to a several convection diffusion problems which mimic difficult problems in compressible flow simulations, including boundary layer and closed streamline problems. I have also solved the viscous Burgers' equation as an extension of DPG to nonlinear problems. I have also explored the regularization of the pure convection equation using the full convection-diffusion equation in the small-viscosity limit. A quadratic-order adaptive DPG method has also been applied to the Navier-Stokes equations to achieve a solution to the flat plate problem for Re = 1000.

- **Proposed: Higher Reynolds number, ramp problem, Gaussian bump, airfoil.**

  The effectiveness of DPG as a numerical method for compressible flow will be assessed with the application of DPG to common transonic and superonic benchmark problems. In particular, I will use DPG to solve the flat plate problem, the ramp problem, flow over a Gaussian bump, over a range of Mach numbers and laminar Reynolds numbers. Time permitting, I will attempt to also present solutions to flow over an airfoil.

- **Proposed: regularized Euler.**

  Time permitting, I will also explore the regularization of the Euler equations using the full Navier-Stokes equations in the inviscid limit.

Bibliography

[1] R. Almeida, R. Feijóo, A. Galeão, C. Padra, and R. Silva. Adaptive finite element computational fluid dynamics using an anisotropic error estimator. *Computer Methods in Applied Mechanics and Engineering*, 182(3–4):379 – 400, 2000.

[2] J. Anderson. *Modern Compressible Flow With Historical Perspective*. McGraw-Hill, 2003.

[3] J. Barrett and K. Morton. Optimal Petrov—Galerkin methods through approximate symmetrization. *IMA Journal of Numerical Analysis*, 1(4):439–468, 1981.

[4] G. Barter. *Shock Capturing with PDE-Based Artificial Viscosity for an Adaptive, Higher-Order Discontinuous Galerkin Finite Element Method*. PhD thesis in Aeronautics and Astronautics, Massachusetts Institute of Technology, 2008.

[5] M. Bieterman, R. Melvin, F. Johnson, J. Bussoletti, D. Young, W. Huffman, and C. Hilmes. Boundary layer coupling in a general configuration full potential code. Technical Report BCSTech-94-032, Boeing Computer Services, 1994.

[6] D. Boffi, F. Brezzi, and M Fortin. Finite elements for the Stokes problem. In *Mixed Finite Elements, Compatibility Conditions, and Applications*, volume 1939 of *Lecture Notes in Mathematics*, pages 45–100. Springer Berlin / Heidelberg.

[7] C.L. Bottasso, S. Micheletti, and R. Sacco. The discontinuous Petrov-Galerkin method for elliptic problems. *Comput. Methods Appl. Mech. Engrg.*, 191:3391–3409, 2002.

[8] C.L. Bottasso, S. Micheletti, and R. Sacco. A multiscale formulation of the discontinuous Petrov-Galerkin method for advective-diffusive problems. *Comput. Methods Appl. Mech. Engrg.*, 194:2819–2838, 2005.

[9] Y. Bourgault, M. Picasso, F. Alauzet, and A. Loseille. On the use of anisotropic a posteriori error estimators for the adaptive solution of 3D inviscid compressible flows. *International Journal for Numerical Methods in Fluids*, 59(1):47–74, 2009.

[10] J. Bramwell, W. Qiu, and L. Demkowicz. A locking-free hp DPG method for linear elasticity with symmetric stresses. Technical Report 2369, Institute for Mathematics and Its Applications, June 2011.

[11] F. Brezzi, B. Cockburn, L.D. Marini, and E. Süli. Stabilization mechanisms in discontinuous Galerkin finite element methods. *Computer Methods in Applied Mechanics and Engineering*, 195(25–28):3293 – 3310, 2006.

[12] A. Brooks and T. 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.

[13] T. Bui-Thanh, L. Demkowicz, and O. Ghattas. A unified discontinuous Petrov-Galerkin method and its analysis for Friedrichs' systems. *Submitted to SIAM J. Numer. Anal.*, 2011. Also ICES report 11-34, November 2011.

[14] T. Bui-Thanh, Leszek Demkowicz, and Omar Ghattas. Constructively well-posed approximation method with unity inf-sup and continuity constants for partial differential equations. *Mathematics of Computation*, 2011. To appear.

[15] T. Bui-Thanh, O. Ghattas, and L. Demkowicz. A relation between the Discontinuous Petrov-Galerkin method and the Discontinuous Galerkin method. Technical report, ICES, 2011.

[16] P. Causin and R. Sacco. A discontinuous Petrov-Galerkin method with Lagrangian multipliers for second order elliptic problems. *SIAM J. Numer. Anal.*, 43, 2005.

[17] P. Causin, R. Sacco, and C.L. Bottasso. Flux-upwind stabilization of the discontinuous Petrov-Galerkin formulation with Lagrange multipliers for advection-diffusion problems. *M2AN Math. Model. Numer. Anal.*, 39:1087–1114, 2005.

[18] J. Chan, N. Heuer, T. Bui Thanh, and L. Demkowicz. Robust DPG method for convection-diffusion problems II: natural inflow conditions. Technical Report 12-21, ICES, June 2012.

[19] T. Chung. *Computational Fluid Dynamics*. Cambridge University Press, 1st edition edition, 2002.

[20] B. Cockburn, J. Gopalakrishnan, and R. Lazarov. Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems. *SIAM J. Numer. Anal.*, 47(2):1319–1365, February 2009.

[21] B. Cockburn, J. Gopalakrishnan, and F. Sayas. A projection-based error analysis of HDG methods. *Math. Comp.*, 79:1351–1367, 2010.

[22] B. Cockburn and W. Shu. The Runge-Kutta Discontinuous Galerkin method for conservation laws: V. Multidimensional systems. *Journal of Comp. Phys.*, 141(2):199–224, 1998.

[23] L. Demkowicz. *Computing With hp-adaptive Finite Elements: One and two dimensional elliptic and Maxwell problems*. Chapman & Hall/CRC Applied Mathematics and Nonlinear Science Series. Chapman & Hall/CRC, 2006.

[24] 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.

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

[26] 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.

[27] L. Demkowicz, J. Gopalakrishnan, I. Muga, and J. Zitelli. Wavenumber explicit analysis for a DPG method for the multidimensional Helmholtz equation. Technical Report 24, ICES, November 2011.

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

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

[30] L. Demkowicz and J. Li. Numerical simulations of cloaking problems using a DPG method. Technical Report 31, ICES, November 2011.

[31] L. Demkowicz, J. Oden, and W. Rachowicz. A new finite element method for solving compressible Navier-Stokes equations based on an operator splitting method and hp-adaptivity. *Comput. Methods Appl. Mech. Eng.*, 84(3):275–326, December 1990.

[32] L. Demkowicz and J.T Oden. An adaptive characteristic Petrov-Galerkin finite element method

for convection-dominated linear and nonlinear parabolic problems in one space variable. *Journal of Computational Physics*, 67(1):188 – 213, 1986.

[33] J. Donea and A. Huerta. *Finite Element Methods for Flow Problems*. Wiley, 2003.

[34] G. Emanuel. *Analytical Fluid Dynamics*. CRC Press, Abingdon, 2001.

[35] K. Eriksson and C. Johnson. Adaptive streamline diffusion finite element methods for stationary convection-diffusion problems. *Mathematics of Computation*, 60(201):pp. 167–188, 1993.

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

[37] D. Griffiths. The 'no boundary condition' outflow boundary condition. *International Journal for Numerical Methods in Fluids*, 24(4):393–411, 1997.

[38] A. Harten, B. Engquist, S. Osher, and S. Chakravarthy. Uniformly high-order accurate non-oscillatory schemes. *SIAM Journal on Numerical Analysis*, 24(1):279–309, 1987.

[39] J. Heinrich, P. Huyakorn, O. Zienkiewicz, and A. Mitchell. An upwind finite element scheme for two-dimensional convective transport equation. *International Journal for Numerical Methods in Engineering*, 11(1):131–143, 1977.

[40] J. 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.

[41] T. Hughes, L. Franca, and G. Hulbert. A new finite element formulation for computational fluid dynamics: VIII. The Galerkin/least-squares method for advective-diffusive equations. *Comp. Meth. Appl. Mech. Engr*, 73:173–189, 1989.

[42] T. Hughes and G. Sangalli. Variational Multiscale Analysis: the Fine-scale Green's Function, Projection, Optimization, Localization, and Stabilized Methods. *SIAM J. Numer. Anal.*, 45(2):539–557, February 2007.

[43] B. Kirk, J. Peterson, R. Stogner, and G. Carey. `libMesh`: A C++ Library for Parallel Adaptive Mesh Refinement/Coarsening Simulations. *Engineering with Computers*, 22(3–4):237–254, 2006. http://dx.doi.org/10.1007/s00366-006-0049-3.

[44] X. Liu, S. Osher, and T. Chan. Weighted essentially nonoscillatory schemes. *Journal of Comp. Phys.*, 115:200–212, 1994.

[45] D. Moro-Lude na, J. Peraire, and N. Nguyen. A Hybridized Discontinuous Petrov-Galerkin scheme for compressible flows. Master's thesis, Massachusetts Institute of Technology, Dept. of Aeronautics and Astronautics, Boston, USA, 2011.

[46] A. Niemi, N. Collier, and V. Calo. Discontinuous Petrov-Galerkin method based on the optimal test space norm for one-dimensional transport problems. *Procedia CS*, 4:1862–1869, 2011.

[47] 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. *Computer Methods in Applied Mechanics and Engineering*, 200(9–12):1291 – 1300, 2011.

[48] T. Papanastasiou, N. Malamataris, and K. Ellwood. A new outflow boundary condition. *International Journal for Numerical Methods in Fluids*, 14(5):587–608, 1992.

[49] W. Reed and T. Hill. Triangular mesh methods for the neutron transport equation. Technical Report LA-UR-73-479, Los Alamos Scientific Laboratory, 1973.

[50] N. Roberts, T. Bui Thanh, and L. Demkowicz. The DPG method for the Stokes problem. Technical Report 12-22, ICES, June 2012.

[51] N. Roberts, D. Ridzal, P. Bochev, and L. Demkowicz. A Toolbox for a Class of Discontinuous Petrov-Galerkin Methods Using Trilinos. Technical Report SAND2011-6678, Sandia National Laboratories, 2011.

[52] H. 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.

[53] C. Shu. Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws. Lecture notes, Brown University.

[54] V. Venkatakrishnan, S. Allmaras, D. Kamenetskii, and F. Johnson. Higher order schemes for the compressible Navier-Stokes equations. In *16th AIAA Computational Fluid Dynamics Conference*, Orlando, FL, June 2003.

[55] J. Zitelli, I. Muga, L. Demkowicz, J. Gopalakrishnan, D. Pardo, and V.M. Calo. A class of discontinuous Petrov–Galerkin methods. Part IV: The optimal test norm and time-harmonic wave propagation in 1D. *Journal of Computational Physics*, 230(7):2406 – 2432, 2011.

Appendix

Appendix 1

Proof of lemmas/stability of the adjoint problem

We present now the proofs of the three lemmas used in Section 4.1.4.2 to show the equivalence of the DPG energy norm to norms on U. We reduce the adjoint problem to the scalar second order equation

$$-\epsilon\Delta v - \beta\cdot\nabla v = g - \epsilon\nabla\cdot f \tag{1.1}$$

with boundary conditions

$$-\epsilon\nabla v\cdot n = f\cdot n, \quad x\in\Gamma_- \tag{1.2}$$

$$v = 0, \quad x\in\Gamma_+ \tag{1.3}$$

and treat the cases $f=0$, $g=0$ separately. The above boundary conditions are the reduced form of boundary conditions (4.5) and (4.6) 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 normal trace of $f\cdot n$ is treated using a density argument — for $f\in C^\infty(\Omega)$, we derive inequalities that are independent of $f\cdot n$ and $\nabla\cdot f$. We extend these inequalities to $f\in L^2(\Omega)$ by taking f to be the limit of smooth functions.

Lemma 2. *Assume v satisfies* (1.1)*, with boundary conditions* (4.5) *and* (4.6)*, and β satisfies* (4.9)

and (4.10). *If* $\nabla \cdot f = 0$ *and* ϵ *is sufficiently small, then*

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

Proof. Define $v_\beta = \beta \cdot \nabla v$. Multiplying the adjoint equation (1.1) 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

$$-\int_\Omega \beta \cdot \nabla v \triangle v = -\int_\Omega \beta \cdot \nabla v \nabla \cdot \nabla v.$$

Integrating this by parts, we get

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

Since $\nabla(\beta \cdot \nabla v) = \nabla\beta \cdot \nabla v + \beta \cdot \nabla\nabla v$, where $\nabla\beta$ and $\nabla\nabla v$ are understood to be tensors,

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

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

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

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 any tangential direction. An orthogonal decomposition in the normal and tangential directions yields $\nabla v = (\nabla v \cdot n)n$, reducing the above to

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

Applying these inequalities to our expression for $\|v_\beta\|^2$ leaves 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 3 complete the estimate. □

Lemma 3. *Assume β satisfies* (4.9)*. Then, for v satisfying equation* (1.1) *with boundary conditions* (4.5) *and* (4.6) *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, $\nabla\psi = \beta$ by properties of the exact sequence. The potential is non-unique up to a constant, and we choose the constant such that $e^\psi = O(1)$. Take the transformed function $w = e^\psi v$; following (2.26) in [29], we substitute w into the the left hand side of equation (1.1), 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

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

Note that $w = 0$ on Γ_+ reduces the boundary integrals over Γ to just the inflow Γ_-. Furthermore, we have $\nabla w = e^{\psi}(\nabla v + \beta v)$. 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 imply $\left((1-\epsilon)|\beta|^2 + \frac{1}{2}\nabla \cdot \beta\right) \lesssim 1$ 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 [29]; 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 4. *Let β satisfy conditions* (4.9) *and* (4.11)*, and let $v \in H^1(\Omega_h)$, $\tau \in H(\mathrm{div}, \Omega_h)$ satisfy equations* (4.7) *and* (4.8) *with $f = g = 0$. Then*

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

Proof. We begin by choosing ψ as the unique solution to the following problem

$$-\epsilon\Delta\psi + \nabla\cdot(\beta\psi) = -\nabla\cdot\tau$$

$$\epsilon\nabla\psi\cdot n - \beta_n\psi - \tau\cdot n = 0, \quad x\in\Gamma_-$$

$$\psi = 0, \quad x\in\Gamma_+.$$

Since $\nabla\cdot\beta = 0$, we can conclude that the bilinear form is coercive and the problem is well posed [29]. The well-posedness of the above problem directly implies that $\nabla\cdot(\tau - (\epsilon\nabla\psi - \beta\psi)) = 0$ in a distributional sense, and thus there exists a $z\in H(\mathrm{curl},\Omega)$ such that

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

Since $\nabla\cdot\beta = 0$, we satisfy condition (4.9). Noting that the sign on β is opposite now of the sign on $\epsilon\Delta\psi$, the problem for ψ matches the adjoint problem for $f = \frac{1}{\epsilon}\tau$. Given the boundary conditions on ψ, we can use a trivial modification of the proof of Lemma 3 to bound

$$\epsilon\|\nabla\psi\|_{L^2}^2 + \|\psi\|_{L^2}^2 \lesssim \frac{1}{\epsilon}\|\tau\|_{L^2}^2.$$

By the above bound and the triangle inequality,

$$\|\nabla\times z\|_{L^2} \le \epsilon\|\nabla\psi\|_{L^2} + \|\beta\psi\|_{L^2} + \|\tau\|_{L^2} \lesssim \frac{1}{\sqrt{\epsilon}}\|\tau\|_{L^2}.$$

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

$$\begin{aligned}\|\tau\|_{L^2}^2 &= (\tau, \epsilon\nabla\psi - \beta\psi + \nabla\times z)_{\Omega_h} = (\tau,\epsilon\nabla\psi) - (\tau,\beta\psi) + (\tau,\nabla\times z)\\ &= (\tau,\epsilon\nabla\psi) + \epsilon(\nabla v,\beta\psi) - \epsilon(\nabla v,\nabla\times z)\\ &= \epsilon\langle[\tau\cdot n],\psi\rangle - \epsilon\langle n\cdot\nabla\times z,[\![v]\!]\rangle - \epsilon(\nabla\cdot\tau,\psi) + \epsilon(\nabla\cdot(\beta v),\psi).\end{aligned}$$

Note that $\nabla \cdot (\beta v) - \nabla \cdot \tau = 0$ removes the contribution of the pairings on the domain and leaves us with only boundary pairings. By definition of the boundary norms on $[\![\tau \cdot n]\!]$ and $[\![v]\!]$ and the fact that $\nabla \times z$ is trivially in $H(\mathrm{div}, \Omega)$,

$$\begin{aligned}\|\tau\|_{L^2}^2 &= \epsilon\langle [\tau \cdot n], \psi\rangle - \epsilon\langle n \cdot \nabla \times z, [\![v]\!]\rangle = \epsilon\langle [\tau \cdot n], \psi\rangle_{\Gamma_h \setminus \Gamma_+} - \epsilon\langle n \cdot \nabla \times z, [\![v]\!]\rangle_{\Gamma_h \setminus (\Gamma_- \cup \Gamma_0)} \\ &\lesssim \epsilon \|[\tau \cdot n]\| \|\psi\|_{H^1(\Omega)} + \epsilon \|[\![v]\!]\| \|\nabla \times z\|_{L^2}.\end{aligned}$$

Applying the bounds $\|\psi\|_{H^1(\Omega)} \leq \frac{1}{\epsilon}\|\tau\|_{L^2}$ and $\|\nabla \times z\|_{L^2} \lesssim \frac{1}{\sqrt{\epsilon}}\|\tau\|_{L^2}$, and noting that $\|\nabla v\| = \frac{1}{\epsilon}\|\tau\|$ completes the proof. □