The Dissertation Committee for Jesse L. Chan
certifies that this is the approved version of the following dissertation:

A DPG method for convection-diffusion problems

Committee:

Leszek Demkowicz, Supervisor

Robert Moser, Co-supervisor

Todd Arbogast

Omar Ghattas

Venkat Raman

A DPG method for convection-diffusion problems

by

Jesse L. Chan, B.A.; M.S.C.S.E.M.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2013

Acknowledgments

This dissertation represents five years of work and study, which seem to have simply flown by much quicker than I expected. First and foremost, I have to thank my advisor (and self-designated local parent figure during my time in Austin) Leszek Demkowicz. Your insight, experience, and most of all, your drive to think outside the box were irreplaceable in building up both my work and my understanding throughout these last 5 years. I have very much enjoyed all the fruitful discussions I have had with you, and have relished the fact that as I've mature mathematically, I've been able to contribute my own share into those discussions as well. My experience in Austin would not have been nearly as pleasant without your kindness as well; your hospitality and warmth are something all of your students appreciate greatly.

None of the work in this thesis would have been possible without the help and expertise of fellow graduate student Nate Roberts, who almost singlehandedly built and maintains the parallel *hp*-adaptive DPG codebase Camellia, with which I ran all numerical experiments in this thesis. Thanks for teaching me software design, for your help in debugging my code, for helpful discussions on each of our mathematical problems, and for good conversations and company in and outside of the ACES (now POB) building.

I'm grateful to the members of my committee – Professors Robert Moser, Todd Arbogast, Omar Ghattas, and Venkat Raman – for helpful discussions and feedback in evaluating my work, as well as for perspective on how my thesis work fits into the larger scheme of computational science and engineering.

To my co-authors on several papers, thank you for all your help in pushing me further along the path to this doctoral degree. To Norbert Heuer, your mathematical insight has greatly advanced our understanding of the DPG method. To Tan Bui, thanks for your emphasis on mathematical rigor, and for your incredible understanding of both application areas and numerical methods and their mathematics. ICES is lucky to have you as a faculty member. To Jay Gopalakrishnan, thanks for our communications, and for keeping me updated with your own insights on DPG. Your interest and excitement in this area are contagious, and your students and colleagues all agree that you are one of the most kind and gracious professors they have ever worked with. To John Evans, congratulations on your new position – it was a pleasure to be able to write a paper with you, to hear your insights, and connect our two separate areas together.

For helpful discussions on both the mathematical and physical nature of my problem, I'm grateful to a host of folks at ICES. To Ivo Babuška, thanks for allowing me to host the ICES Forum with you for 3 years, and for sharing your wealth of knowledge on *hp* methods and singularities with me. To Rhys Ulerich, thanks for discussions and answers to all of the strangest and most esoteric questions concerning compilers, linking, and all things code-based in general. To Todd Oliver, thanks for the discussions on nondimensionalization and boundary conditions. Roy Stogner and Paul Baumann thanks for giving me some of PECOS' vast perspective on common problems in fluid dynamics simulations using finite elements. Paul Tsuji, thanks for the discussions on conditioning and Schur complement systems. Truman Ellis and Jamie Bramwell, thanks for listening to me talk on DPG and mathematics – when working on such a new topic, I am grateful for any colleagues who also understand it, with whom to share both successes and failures.

The Texas Advanced Computing Center (TACC) at The University of Texas at Austin has been very generous in providing HPC resources that have contributed to the research results reported

within this paper, as well as consultation for whenever I experienced issues on their machines. Thanks also to Chris Simmons and Karl Schulz for helping me figure out some of the tricks of the trade when it came to modern software engineering.

To my friends at ICES, thanks for making all the long hours spent in this building enjoyable. Fred Qiu, it was great having you both as a roommate and colleague. John Hawkins, thanks for your hospitality, and for the fun nights of bike building and conversation with you and Christa. Chris Mirabito, Paul Tsuji, Henry Chang, Omar Al Hinai, Prapti Neupane, and the rest of the cribbage crowd – may your fifteens be numerous. Jeff Hussman, Jesse Kelly, it was fun to have someone with whom I could share both my math and my music. Fred Nugen, it was great biking with you (you still owe me for those rims, by the way). To so many folks at ICES – Jessica, Lindley, Hamid, Vikram, Michele, Nick, James, and still so many more – thanks for all the good conversation and camaraderie over these last few years.

To my community at Vox Veniae, you have been a pillar of support for the last 5 years of my life. To list all the incredible people who have made an impact on my life would take almost the entire length of this acknowledgment section all over again.

Thanks to my family – my parents Henry and Wendy, and my brother David – thanks for the support throughout the last 5 years, and for always welcoming me home whenever I needed a break. I love you all so much.

I would be remiss if I did not acknowledge the financial support I received during the course of this project. This work was made possible by the CAM Fellowship at ICES and by the Department of Energy [National Nuclear Security Administration] under Award Number [DE-FC52-08NA28615].

A DPG method for convection-diffusion problems

Jesse L. Chan, Ph.D.

The University of Texas at Austin, 2013

Supervisors: Leszek Demkowicz
Robert Moser

Over the last three decades, CFD simulations have become commonplace as a tool in the engineering and design of high-speed aircraft. Experiments are often complemented by computational simulations, and CFD technologies have proved very useful in both the reduction of aircraft development cycles, and in the simulation of conditions difficult to reproduce experimentally. 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; in particular, reliable higher order and *hp*-adaptive methods for the Navier-Stokes equations that govern viscous compressible flow.

Solutions to the equations of viscous flow can display shocks and boundary layers, which are characterized by localized regions of rapid change and high gradients. The use of adaptive meshes is crucial in such settings — 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. An alternative to such is the construction of automatically adaptive schemes; such methods begin with a coarse mesh and refine based on the minimization of 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 more difficult in the context of higher order and *hp* methods [1].

Many of the above issues are tied to the notion of *robustness*, which we define loosely for CFD applications as the degradation of the quality of numerical solutions on a coarse mesh with respect to the Reynolds number, or nondimensional viscosity. For typical physical conditions of interest for the compressible Navier-Stokes equations, the Reynolds number dictates the scale of shock and boundary layer phenomena, and can be extremely high — on the order of 10^7 in a unit domain. For an under-resolved mesh, the Galerkin finite element method develops large oscillations which prevent convergence and pollute the solution.

The issue of robustness for finite element methods was addressed early on by Brooks and Hughes in the SUPG method [2], which introduced the idea of residual-based stabilization to combat such oscillations. Residual-based stabilizations can alternatively be viewed as modifying the standard finite element test space, and consequently the norm in which the finite element method converges. Demkowicz and Gopalakrishnan generalized this idea in 2009 by introducing the Discontinous Petrov-Galerkin (DPG) method with optimal test functions, where test functions are determined such that they minimize the discrete linear residual in a dual space. Under the ultra-weak variational formulation, these test functions can be computed locally to yield a symmetric, positive-definite system.

The main theoretical thrust of this research is to develop a DPG method that is provably robust for singular perturbation problems in CFD, but does not suffer from discretization error in the approximation of test functions [3, 4]. Such a method is developed for the prototypical singular perturbation problem of convection-diffusion, where it is demonstrated that the method does not

suffer from error in the approximation of test functions, and that the L^2 error is robustly bounded by the energy error in which DPG is optimal – in other words, as the energy error decreases, the L^2 error of the solution is guaranteed to decrease as well. The method is then extended to the linearized Navier-Stokes equations, and applied to the solution of the nonlinear compressible Navier-Stokes equations.

The numerical work in this dissertation has focused on the development of a 2D compressible flow code under the Camellia library, developed and maintained by Nathan Roberts at ICES [5]. In particular, we have developed a framework allowing for rapid implementation of problems and the easy application of higher order and *hp*-adaptive schemes based on a natural error representation function that stems from the DPG residual [6, 7].

Finally, the DPG method is applied to several convection diffusion problems which mimic difficult problems in compressible flow simulations, including problems exhibiting both boundary layers and singularities in stresses. A viscous Burgers' equation is solved as an extension of DPG to nonlinear problems, and the effectiveness of DPG as a numerical method for compressible flow is assessed with the application of DPG to two benchmark problems in supersonic flow. In particular, DPG is used to solve the Carter flat plate problem and the Holden compression corner problem over a range of Mach numbers and laminar Reynolds numbers using automatically adaptive schemes, beginning with very under-resolved/coarse initial meshes.

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 [1].

- **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 [8]. 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 [1].

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 $\Omega \subset \mathbb{R}^3$,

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

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 $x = 1$.

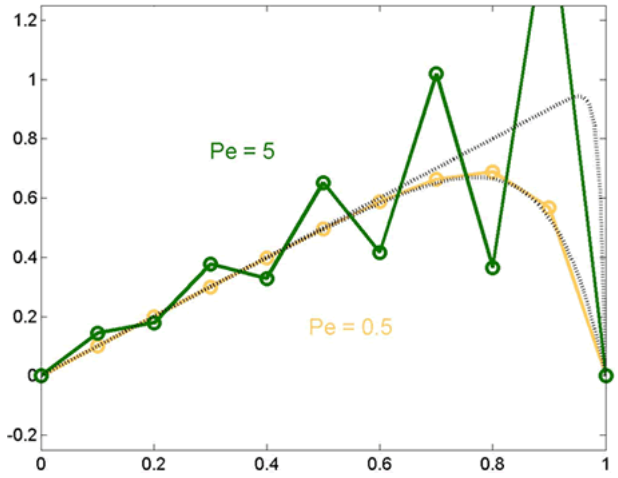


Figure 1.1: Oscillations in the 1D finite element solution of the convection-diffusion equation for small diffusion [9]. Standard finite volume and finite difference methods exhibit similar behavior.

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 [10]:

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

for $\|u\|_\epsilon^2 := \|u\|_{L^2}^2 + \epsilon \|u'\|_{L^2}^2$, 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]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.

1.2 Goal

From the perspective of the compressible Navier-Stokes equations, this loss of robustness is doubly problematic. Not only will any nonlinear solution suffer from similar unstable oscillations, but 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. Artificial viscosity and shock capturing methods have been used to suppress such oscillations and regularize the problem. While these methods will usually yield smooth and qualitatively resolved solutions, these methods are often overly diffusive, yielding results which are poor approximations of the true solution [11], though modern artificial viscosity and shock capturing schemes have improved greatly in recent years [12, 13]. We have taken an alternative approach in this work, avoiding artificial diffusion and shock capturing for the moment.

Our aim is to develop a stable, higher order scheme for the steady compressible laminar Navier-Stokes equations in transonic/supersonic regimes that is automatically adaptive beginning with very coarse meshes. This requires that both the method and the refinement scheme to perform adequately on coarse meshes with high Peclet numbers – in other words, that the adaptive method is robust in the diffusion parameter. We construct such a method in this work – in particular, we present a method for which automatic adaptivity can be applied to problems in compressible flow, beginning with very coarse meshes which do not reflect additional knowledge about the problem. 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 [14]. Harten, Enquist, Osher and Chakravarthy introduced the essentially non-oscillatory (ENO) scheme in 1987 [15], which was improved upon with the weighted essentially non-oscillatory (WENO) scheme in [16]. 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, which vanishes as the mesh size $h \to 0$.

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 [17].

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 [2]. 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 in 1D.

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

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

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 [2], as well as in an upcoming book by Hughes. The convection-diffusion equation can be written as follows:

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

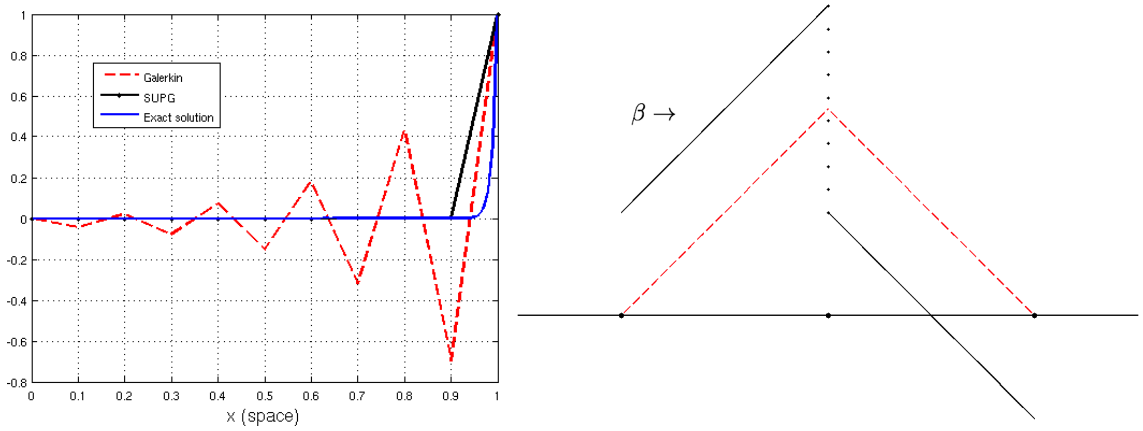
where $L_{\text{adv}} u \coloneqq \nabla \cdot (\beta u)$ is the first order advective operator, and $L_{\text{diff}} u \coloneqq \epsilon \Delta u$ is the second-order diffusive operator. Let us assume u to be a linear combination of piecewise-linear basis functions $\phi_i, i = 0, \ldots, N$ (then, 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

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

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"

[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.

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.



(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.

The idea of emphasizing the upwind portion of a test function is an older idea, introduced in 1977 by Zienkiewicz et al. in [18]. 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, the SUPG solution is very close to, but no longer the H_0^1 best approximation for 2D and 3D problems [19]. 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.

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 with $b(u - u_h, v) = 0$ for all $v \in V$). This addition of residual-based stabilization can be interpreted as a modification of the test functions. 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 3.1 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 [20]. These methods were later analyzed Raviart et al [21] and later by Johnson et al [22], who contributed a mathematical analysis of the original method of Reed and Hill, as well as by Cockburn and Shu [23], who solved the Euler equations by applying concept of Lax's numerical flux within the context of DG.

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 (and vice versa) through appropriate choices of this numerical flux [24]. 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[24].[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 [25, 7].

[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.

1.3.2.3 HDG

A more recent development in DG methods is the idea of *hybridized* DG (HDG), introduced by Cockburn, Gopalakrishnan and Lazarov [26]. 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* problem. Once the global problem is solved, interior degrees of freedom can be recovered in parallel through so-called *local* problems [27].

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 dissertation 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 results for the Burgers and compressible Navier-Stokes equations.

Chapter 2

Range of problems

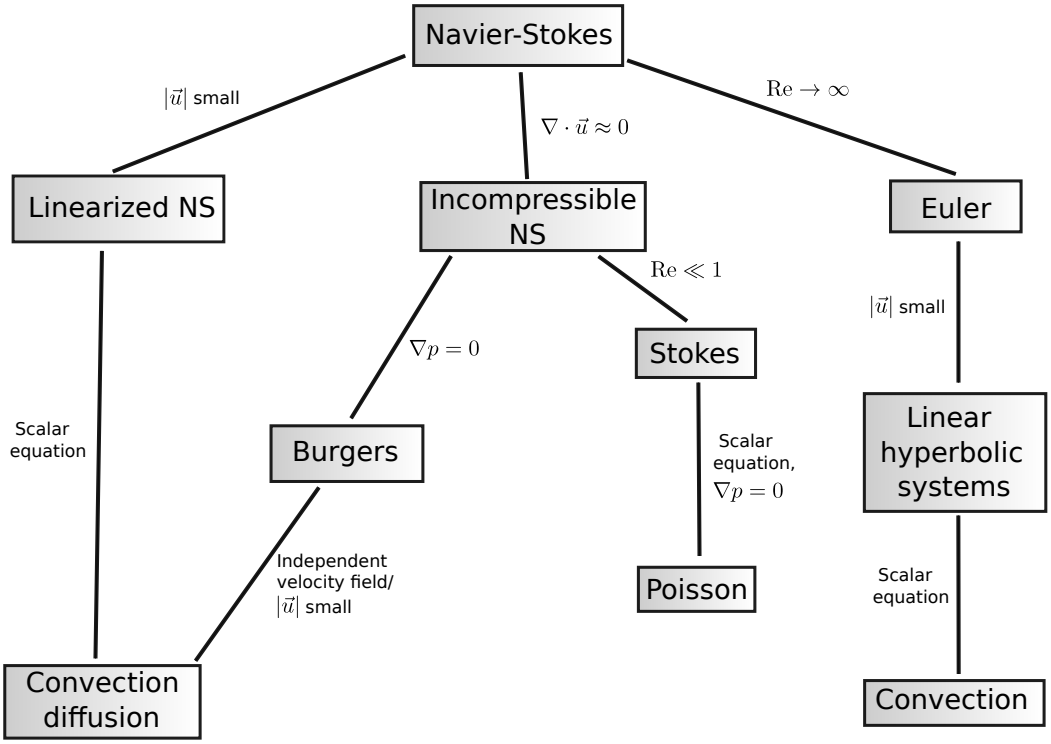


Figure 2.1: A diagram of common CFD problems and their simplifying assumptions.

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_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**

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \begin{bmatrix} \rho u_1 \\ \rho u_2 \end{bmatrix} = 0$$

- **Momentum conservation**

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

- **Energy conservation**

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

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 bulk viscosity can become important for compressible flows, we treat both coefficients separately. In

[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 [28] for one example.

general, μ and λ are functions of temperature. One method of modeling temperature dependence is through 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. [29] 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 [30], and the incompressible Navier-Stokes are covered in the upcoming dissertation of Nathan Roberts.

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.

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, 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.

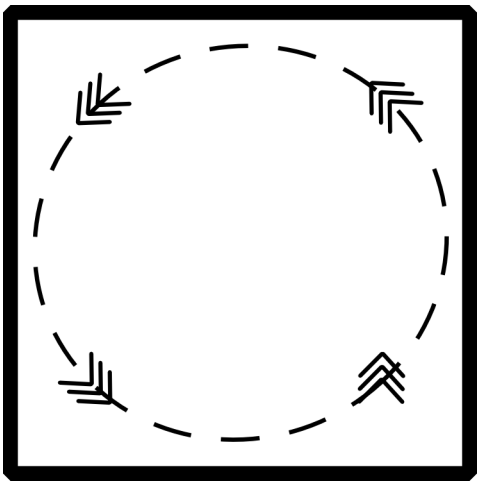


Figure 2.2: Setup for the vortex problem.

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 problems. The topic is expansive, and we direct the reader towards [12] 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 [31]. 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

3.1 Discontinuous Petrov-Galerkin methods with optimal test functions

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 [32]. The idea was continued with the SUPG method of Hughes, and the characteristic Petrov-Galerkin approach of Demkowicz and Oden [33], 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 [6]. 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 [34]. However, the method in [6] 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 dissertation, 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 subspaces. We are interested in the following problem

$$\begin{cases} \text{Given } l \in V', \text{ find } u_h \in U_h \text{ such that} \\ b(u_h, v_h) = l(v_h), \quad \forall v_h \in V_h, \end{cases} \tag{3.1}$$

where $b(\cdot,\cdot) : U \times V \to \mathbb{R}$ is a continuous bilinear form. U_h is chosen to be some trial space of approximating functions, but V_h is as of yet unspecified.

Throughout the dissertation, we suppose the variational problem (3.1) to be well-posed. In that case, 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 the dissertation.

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

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

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

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

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 $H(\mathrm{div})$-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 and a *localizable* norm $\|\cdot\|_V$[1] 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 still 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 [25, 6]. In this dissertation, we assume the error in approximating the optimal test functions is negligible, and refer to the work in [35] 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

[1] A localizable norm $\|v\|_{V(\Omega_h)}$ can be written in the form

$$\|v\|^2_{V(\Omega_h)} = \sum_{K \in \Omega_h} \|v\|^2_{V(K)},$$

where $\|v\|_{V(K)}$ is a norm over the element K.

"energy norm" — that is [36, 6, 37]

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

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 $\|u - u_h\|_{U,E} = \|Bu - Bu_h\|_{V'} = \left\|R_V^{-1}\left(l - Bu_h\right)\right\|_V$ can be determined by computing the *error representation function* $e \coloneqq R_V^{-1}\left(l - Bu_h\right)$ through

$$(e, \delta v)_V = b(u - u_h, \delta v) = l\left(\delta v\right) - b(u_h, \delta v)$$

and measuring its norm $\|e\|_V$. This is simply a consequence of the least-squares nature of DPG; the energy error is simply the norm of the residual in V'. Under the assumption of a localizable norm on V, we can compute the squared norm over the entire domain $\|e\|^2_{V(\Omega_h)}$ as the sum of individual element contributions $\sum_{K \in \Omega_h} \|e\|^2_{V(K)}$. We define $e_K^2 \coloneqq \|e\|^2_{V(K)}$ as a local error indicator with which we can drive adaptive mesh refinement.

Practically speaking, this implies that the DPG method is discretely stable on any mesh. In particular, DPG is unconditionally stable for higher order adaptive meshes, where discrete stability is often an issue.

3.2 Duality between trial and test norms (energy norm pairings)

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, Lemma 2.5 in [36]: assuming, for simplicity, that the bilinear form $b(u,v)$ is definite[2], given any norm $\|\cdot\|_U$ on the trial space U, for $\varphi \in U$, we can represent $\|\varphi\|_U$ via

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

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,

[2]By definite, we mean that

$$b(u,v) = 0, \forall v \in V \Rightarrow u = 0$$
$$b(u,v) = 0, \forall u \in U \Rightarrow v = 0,$$

which imply injectivity of the bilinear operator B and its transpose B', defined such that $\langle Bu, v\rangle = \langle u, B'v\rangle$. These conditions imply solvability of the variational problem.

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 dissertation, for simplicity of exposition, we shall call a pair of norms in U and V that induce each other as an *energy norm pairing*.

3.3 Discontinuous Petrov-Galerkin methods with the ultra-weak formulation

The name 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 [36]. Additionally, Cohen, Dahmen and Welper simultaneously extended the minimum residual concept behind DPG to general variational settings and avoid the use of discontinuous test functions by formulating the minimum residual method as a saddle-point problem [38].

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 introduced in Section 3.1 with the so-called "ultra-weak formulation" [25, 6, 7, 37, 39, 40]. 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 come 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.

We derive the ultra-weak variational formulation by multiplying our first order system $Au = f$ by a test function v. Integrating by parts over each individual element K, we have elementwise

$$(Au, v)_{L^2(K)} = (u, A_h^* v)_{L^2(K)} + \langle A_0 u, \gamma v \rangle_{\partial K} = (f, v)_{L^2(K)},$$

where $A_0 u$ is the trace term resulting from the integration by parts of A, and γv refers to the trace of v on the element boundary ∂K. We note that γv refers to the proper trace of the test function across an edge; for example, for scalar valued test functions v, γv is the boundary trace, while for vector valued test functions τ, $\gamma\tau$ corresponds to the normal trace.

If we assume $A_0 u$ is single-valued over a given edge e in the mesh skeleton Γ_h, we can sum up over all elements $K \in \Omega_h$ to get

$$\sum_{K\in\Omega_h} (Au, v)_{L^2(K)} = (u, A_h^* v)_{L^2(\Omega)} + \sum_{e\in\Gamma_h} \langle A_0 u, [\![\gamma v]\!]\rangle_e = (f, v)_{\Omega_h} ,$$

where $A_h^* v$ is the adjoint of A applied elementwise, and $[\![\gamma v]\!]$ is the jump of the proper trace of v. For simplicity of notation, we will refer to $[\![\gamma v]\!]$ as $[\![v]\!]$, and will refer to the duality pairing over all edges $\sum_{e\in\Gamma_h} \langle A_0 u, [\![\gamma v]\!]\rangle_e$ as the duality pairing over the mesh skeleton $\langle A_0 u, [\![v]\!]\rangle_{\Gamma_h}$. The ultra-weak formulation for $Au = f$ on Ω_h results from identifying the single-valued term $A_0 u$ as an additional unknown $\widehat{u}$ on Γ_h

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

where $\langle \cdot, \cdot \rangle_{\Gamma_h}$ is the duality pairing over 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. Boundary conditions are applied to the trace variable $\widehat{u}$.

For simplicity in writing, we will occasionally 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 no longer admits a trace on Γ_h. Consequently, we needed 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,

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

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 [41], 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 [42].

3.4 A canonical energy norm pairing for ultra-weak formulation

From the discussion in Section 3.2 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. In this section, we derive and discuss an important energy norm pairing which specifies the canonical norm in U and induces a test norm on V.

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

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

Using the framework developed in [36], one can show that $\left(\|(u,\widehat{u})\|_U, \|v\|_{V,U}\right)$ is an energy norm pairing in the sense discussed in Section 3.2. That is, the canonical norm $\|(u,\widehat{u})\|_U$ in U induces (generates) the norm $\|v\|_{V,U}$ in V.

The canonical norm $\|(u,\widehat{u})\|_U$ in U provides an optimal balance between the standard norms on the field u and the flux $\widehat{u}$ [37]. As a result, if the induced norm $\|v\|_{V,U}$ (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 $[\![v]\!]$. Since the jump terms couple elements together, the evaluation of the jump terms requires 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, since $v \in D(A_h^*)$, we can use as a test norm for v the broken graph norm:

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

This norm is a localization of $\|v\|_{V,U}$ to allow for the solution of optimal test functions on an element-by-element basis, and is considered to be the canonical norm on V. In the DPG literature [37], $\|v\|_{V,U}$ is known as the *optimal test norm*, while $\|v\|_V$ is known as the *quasi-optimal* or *graph test norm*.

Using variants of the graph test norm, numerical results show that the DPG method appears to provide a "pollution-free" method without phase error for the Helmholtz equation [37], and analysis of the pollution-free nature of DPG is currently under investigation. Similar results have also been obtained in the context of elasticity [39] and the linear Stokes equations [5]. On the

theoretical side, the graph test norm has been shown to yield a well-posed DPG methodology for the Poisson and convection-diffusion equations [41]. More recently, this theory has been generalized to show the well-posedness of DPG for the large class of PDEs of Friedrichs' type [42].

Chapter 4

The graph norm 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 (for a given test norm), and demonstrate its usefulness as a numerical method for solving singular-perturbed problems. In particular, we will examine three different choices of test norms on V – an ideal norm (which returns good results, but whose test functions are difficult to approximate), a robust norm (which is easy to approximate and computationally efficient to assemble but still returns good results over a range of ϵ), and finally, a coupled, robust test norm that borrows ideas from both the ideal and robust norm.[1]

4.1 DPG formulation for convection-diffusion

We consider the following model convection-diffusion problem on a domain $\Omega \subset \mathbb{R}^d$ with boundary $\partial\Omega \equiv \Gamma$

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

[1] This third test norm is motivated by observed numerical difficulties; the precise shortcomings of the previous robust test norm are not completely understood, though a possible explanation is offered for both the issues encountered by the robust norm and the success of the coupled robust test norm in overcoming these issues.

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 3.3 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 norm on this test space 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 $\beta_n \coloneqq \beta \cdot n$. On the inflow boundary, we apply the inflow boundary condition

$$u = u_{\mathrm{in}} \quad \text{on } \Gamma_-.$$

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

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

For DPG, we must also specify a test norm which defines the test space. Our focus will be on the graph test norm for convection-diffusion, which, under the ultra-weak variational formulation,

$$\|(v,\tau)\|^2_{V_{\text{graph}}} = \|\nabla_h \cdot \tau - \beta \cdot \nabla_h v\|^2_{L^2(\Omega)} + \left\|\epsilon^{-1}\tau - \nabla_h v\right\|^2_{L^2(\Omega)} + \|v\|^2_{L^2(\Omega)},$$

where ∇_h and $\nabla_h \cdot$ are understood to act elementwise.[2]

Though our focus is on the convection-diffusion equation, we will attempt to relate concepts to a more general framework and notation when possible. Overloading the notation $\widehat{u} \coloneqq \left(\widehat{u}, \widehat{f_n}\right)$, $u \coloneqq (u, \sigma)$, and $v \coloneqq (v, \tau)$, we can define the operator $A_h^* : V(\Omega_h) \to L^2(\Omega)$ through its action restricted to an individual element K

$$A_h^* v|_K = \left(\nabla \cdot \tau - \beta \cdot \nabla v, \epsilon^{-1}\tau - \nabla v\right) \quad \text{on } K \in \Omega_h.$$

We then have the abstract representation of both the ultra-weak variational formulation and the graph test norm as

$$b\left((u,\widehat{u}),v\right) \coloneqq \langle \widehat{u}, [\![v]\!] \rangle + (u, A_h^* v)_{L^2(\Omega)}$$

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

From this point onward, we will continue to overload our abstract notation in order to connect more general concepts with the concrete example of the convection-diffusion equation.

[2]Since $\left\|A_h^* v\right\|$ is not positive definite on its own, we typically add an L^2 term for all components of the test function; however, here the L^2 norm of τ is neglected as it is not required to preserve positive-definiteness of the norm.

4.1.1 L^2 optimality under the ultra-weak variational formulation

We will aim now to define energy settings and test/trial spaces under which the ultra-weak formulation produces L^2-optimal solutions for u, σ. Our approach focuses on the convection-diffusion equation, but we will generalize using more abstract notation when possible.

4.1.1.1 Test and trial spaces

We begin by defining the spaces

$$H_A = \left\{ (u,\sigma) \in H^1(\Omega) \times H(\mathrm{div};\Omega) : \left(\nabla \cdot (\beta u - \sigma), \frac{1}{\epsilon}\sigma - \nabla u \right) \in L^2(\Omega) \right\}$$
$$H_{A^*} = \left\{ (v,\tau) \in H^1(\Omega) \times H(\mathrm{div};\Omega) : \left(\nabla \cdot \tau - \beta \cdot \nabla v, \frac{1}{\epsilon}\tau + \nabla v \right) \in L^2(\Omega) \right\}$$

Note that in these definitions, we have chosen both trial and test functions from the fully conforming spaces $H(\mathrm{div};\Omega)$ and $H^1(\Omega)$ over Ω. Let us define the spaces $U = V = H^1(\Omega) \times H(\mathrm{div};\Omega)$ as the *conforming* spaces for which the inter-element jumps $\langle \widehat{u}, [\![\tau_n]\!] \rangle_{\Gamma_h^0}$ and $\left\langle \widehat{f}_n, [\![v]\!] \right\rangle_{\Gamma_h^0}$ both vanish. If we again overload notation by defining $u \coloneqq (u,\sigma)^T \in U$ and $v \coloneqq (v,\tau) \in V$, and define the operator A and its adjoint A^* through

$$Au \coloneqq \begin{bmatrix} \nabla \cdot (\beta u - \sigma) \\ \frac{1}{\epsilon}\sigma - \nabla u \end{bmatrix}, \quad A^* v \coloneqq \begin{bmatrix} \nabla \cdot \tau - \beta \cdot \nabla v \\ \frac{1}{\epsilon}\tau + \nabla v \end{bmatrix},$$

then we can recognize H_A and H_{A^*} as the *graph spaces* corresponding to the first-order system operator A and its adjoint A^*. We can now compactly characterize the spaces H_A and H_{A^*}

$$H_A = \left\{ u \in U, Au \in L^2(\Omega) \right\}, \quad H_{A^*} = \left\{ v \in V, A^* v \in L^2(\Omega) \right\}.$$

By choosing $(v,\tau) \in H_{A^*}$, we can eliminate the inter-element jumps in the ultra-weak variational formulation, such that we are left with

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

prior to the application of boundary conditions.

Specifying spaces for the trace variables $\widehat{f}_n$ and $\widehat{u}$ is a bit more involved. Note that H_A and H_{A^*} are dual to each other; denoting the trace space of H_A as $\widehat{H}_A = H^{1/2}(\Gamma) \times H^{-1/2}(\Gamma)$ and defining the traces (again overloading notation) through $\widehat{v} \coloneqq (v|_\Gamma, \tau_n|_\Gamma) = (\widehat{v}, \widehat{\tau})$, $\widehat{u} \coloneqq \left(\widehat{f}_n, \widehat{u}\right) \in \widehat{H}_A$, we are able to characterize the duality pairing over Γ as follows: we have

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

or, using abstract operator notation

$$\langle \widehat{u}, \widehat{v} \rangle = (Au, v) - (u, A^* v).$$

An interpretation of the above characterization would be that the trial variables $\widehat{f}_n$ and $\widehat{u}$ represent traces of functions $(\beta u - \sigma, u) \in H_A$. Our trial and test spaces can now be specified

$$(u, \sigma) \in L^2(\Omega), \left(\widehat{u}, \widehat{f}_n\right) \in H^{1/2}(\Gamma) \times H^{-1/2}(\Gamma), (v, \tau) \in H_{A^*}$$

or under our more general notation,

$$u \in L^2(\Omega), \widehat{u} \in \widehat{H}_A(\Gamma), v \in H_{A^*}$$

where $\widehat{H}_A(\Gamma) = \{ u|_\Gamma, u \in H_A \}$ consists of the boundary traces of functions in H_A. Thus, while we formally relax regularity constraints on (u, σ) by simply requiring $(u, \sigma) \in L^2(\Omega)$, we maintain regularity constraints through our choice of spaces for the trace variables.

4.1.1.2 Test space boundary conditions

Having replaced the ultra-weak formulation under a broken test space with the ultra-weak formulation using a globally conforming test space, we now aim to treat boundary conditions. Under

our model problem, boundary data u_0 is applied to $\widehat{u}$ on the entirety of Γ, such that our variational formulation becomes

$$\left\langle \widehat{f}_n, v \right\rangle_\Gamma + (u, \nabla \cdot \tau - \beta \cdot \nabla v) + \left(\sigma, \frac{1}{\epsilon}\tau + \nabla v \right) = (f, v) + \langle u_0, \tau_n \rangle_\Gamma \,.$$

We can restrict our test space[3] to $\tilde{H}_{A^*} \coloneqq \{(v, \tau) \in H_{A^*},\, v|_\Gamma = 0\} \subset H_{A^*}$, which reduces the formulation to

$$(u, \nabla \cdot \tau - \beta \cdot \nabla v) + \left(\sigma, \frac{1}{\epsilon}\tau + \nabla v \right) = (f, v) + \langle u_0, \tau_n \rangle_\Gamma \,.$$

If our trial space is now taken to be the discrete trial space U_h spanned by trial functions $\phi_i = 1, \ldots, N$, by choosing our discrete test space such that

$$\begin{aligned} \nabla \cdot \tau - \beta \cdot \nabla v &= u_i, \\ \frac{1}{\epsilon}\tau + \nabla v &= \sigma_i, \\ v &= 0 \quad \text{on } \Gamma, \end{aligned}$$

where u_i and σ_i are the u and σ components of the ith trial function ϕ_i, then our discrete variational problem for (u_h, σ_h) becomes

$$\begin{aligned} (u_h, u_i) + (\sigma_h, \sigma_i) &= (u_h, \nabla \cdot \tau - \beta \cdot \nabla v) + \left(\sigma_h, \frac{1}{\epsilon}\tau + \nabla v \right) \\ &= (f, v) + \langle u_0, \tau_n \rangle_\Gamma \\ &= (u, u_i) + (\sigma, \sigma_i) \end{aligned}$$

and the solutions $u_h, \sigma_h \in U_h$ to our discrete variational problem are exactly the best L^2-approximations to the u and σ.

[3]Formally speaking, we require only that $\left\langle \widehat{f}_n, v \right\rangle_\Gamma = 0, \forall \widehat{f}_n \in H^{-1/2}(\Gamma)$. However, since $v|_\Gamma \in H^{1/2}(\Gamma)$, and since the duality pairing $\langle \cdot, \cdot \rangle_{H^{-1/2}(\Gamma) \times H^{1/2}(\Gamma)}$ is definite, this condition is equivalent to $v|_\Gamma = 0$.

We can frame the above discussion concerning boundary conditions using a more abstract notation as well; let us define an boundary condition operator $C : \widehat{H}_A \to \widehat{H}_A$, such that boundary data is applied to the quantity $\langle C\widehat{u}, \widehat{v}\rangle$. For the convection-diffusion equation, $C\widehat{u} = (\widehat{u}, 0)$, and $\langle C\widehat{u}, \widehat{v}\rangle \coloneqq \langle \widehat{u}, \tau_n\rangle$. Due to the definiteness of the duality pairing $\langle \cdot, \cdot \rangle$, we have that $\langle C\widehat{u}, \widehat{v}\rangle = \langle \widehat{u}, C'\widehat{v}\rangle$, where C' is the conjugate of C with respect to the duality pairing. Next, we define the space $\tilde{H}_{A^*}$ as

$$\tilde{H}_{A^*} = \{v \in H_{A^*}, (I - C')\widehat{v} = 0\}.$$

We interpret $\tilde{H}_{A^*}$ as being the subspace of H_{A^*}, the graph space of conforming test functions, such that the remaining boundary terms vanish after imposition of boundary data.

Remark 1. *For non-homogeneous boundary conditions under standard finite element methods, boundary conditions are treated using lift and extension operators. In other words, given boundary data u_0 on some part of the boundary $\Gamma_0 \subset \Gamma$, then we decompose our solution u into $u = Eu_0 + \tilde{u}$, where $\tilde{u}$ comes from a so-called homogeneous space $\tilde{U} \coloneqq U/Eu_0$, and Eu_0 is a non-unique extension of the lift u_0 into the interior of the domain Ω. Under a Bubnov-Galerkin formulation, the test space is the same as the trial space, and we test with a homogeneous test space $\tilde{V}$ as well.*

We still utilize the same framework under the ultra-weak formulation: $\tilde{H}_{A^*}$ corresponds to the homogeneous test space, and we still utilize lifts and extension operators in dealing with the boundary data. However, an important distinction between standard formulations and the ultra-weak formulation is that, for broken test spaces, the lift extends not onto the domain Ω_h, but onto the internal skeleton Γ_h^0. Choosing globally conforming test spaces removes traces defined on the internal skeleton, and allows us to treat boundary conditions by only considering lifts defined on Γ.

The key step in achieving L^2 optimality is to choose test functions from a subspace of $\tilde{H}_{A^*}$:

by choosing for the discrete test space V_h

$$V_h \coloneqq \left\{ v_i \in \tilde{H}_{A^*}, A^* v_i = \phi_i \right\},$$

our discrete variational problem reduces to

$$(U_h, A^* V_i) = (U_h, \phi_i) = (f, v) + \langle u_0, \tau_n \rangle_\Gamma = (U, \phi_i), \quad i = 1, \ldots, N,$$

which we recognize as the L^2 projection of the solution U onto U_h. In other words, under the ultra-weak variational formulation and a conforming test space, the discrete test space that delivers the best L^2-approximation is made up of solutions to the adjoint equation, with the basis functions spanning the trial space acting as loads.

Remark 2. *We note that we do not use the space $\tilde{H}_{A^*}$ in practice to approximate test functions, as it would require additional logic differentiating between boundary and interior elements, as well as logic distinguishing between free boundary degrees of freedom and degrees of freedom on which boundary conditions are applied. The above discussion is mainly to motivate global features necessary for optimal test spaces.*

4.1.2 Globally conforming DPG test spaces

As noted above, L^2 optimality is achieved when the test space possesses certain *global* properties. An obvious question concerning DPG is how much non-local information can be gleaned from locally generated test spaces.[4]

First, we define the idea of *weakly* conforming spaces. For discrete trial variables $(u, \widehat{u}) \in U_h$,

[4]Global properties of the test norm are related to the adjoint equation, as shown in [3]. However, global properties of the test space have not been explored specifically in these prior works.

we denote a test function as weakly conforming if

$$b\left((0,\widehat{u}\right),v)=\langle\widehat{u},[\![v]\!]\rangle_{\Gamma_h^0}=0,\quad\forall\,(0,\widehat{u})\in U_h.$$

In other words, a test function is weakly conforming if its jumps are orthogonal to all trace functions in the discrete space U_h. We refer to the space of these test functions as $\tilde{V}(\Omega_h)=\{\tilde{v}\in V:\langle\widehat{u},[\![\tilde{v}]\!]\rangle_{\Gamma_h^0}=0,\forall\,(0,\widehat{u}_h)\in U_h\}$.

Let $V_{\text{opt}}=\{v_{\delta u}\in V:v_{\delta u}=R_V^{-1}B\delta u,\delta u\in U_h\}$ be the space of locally determined DPG test functions, and let

$$\tilde{V}_{\text{opt}}=\{\tilde{v}_{\delta u}\in\tilde{V}:\tilde{v}_{\delta u}=R_V^{-1}B\delta u,\delta u\in U_h\}$$

be the space of optimal test functions determined *globally* over the weakly conforming space $\tilde{V}$. In other words, the optimal test functions that span $\tilde{V}_{\text{opt}}$ are the result of the inversion of the Riesz operator *globally*, over the entire mesh (the idea is not new – the concept of globally optimal test functions was first introduced in [7] to prove mesh-independence). The following lemma concerning the test space spanned by the locally computed optimal test functions of DPG was proven first by Demkowicz and Gopalakrishnan in [43], which we reproduce briefly here.

Lemma 1. $\tilde{V}_{\text{opt}}\subseteq V_{\text{opt}}$.

Proof. We note first that, because $V_{\text{opt}}\subseteq V$, we can orthogonally decompose $V=V_{\text{opt}}\bigoplus V_{\text{opt}}^{\perp}$, where, for any $v_{\text{opt}}\in V_{\text{opt}}$, $\left(v_{\text{opt}},v^{\perp}\right)_V=0$ for all $v^{\perp}\in V_{\text{opt}}^{\perp}$. Let us now choose a globally conforming test function $\tilde{v}\in\tilde{V}_{\text{opt}}$. Since $\tilde{V}_{\text{opt}}\subseteq V$, we can decompose $\tilde{v}=v_{\text{opt}}+v_{\text{opt}}^{\perp}$. Demonstrating that $v_{\text{opt}}^{\perp}=0$ proves the lemma.

Let $v_{\widehat{u}}\in V_{\text{opt}}$ be an optimal test function corresponding to the flux variable $\widehat{u}$. We can use

the fact that $v_{\text{opt}}^{\perp} \in V$ to substitute it into the definition of $v_{\widehat{u}}$. By definition,

$$\left(v_{\widehat{u}}, v_{\text{opt}}^{\perp}\right)_{V(K)} = \left\langle \widehat{u}, v_{\text{opt}}^{\perp} \right\rangle_{\partial K} = 0,$$

where $(\cdot,\cdot)_{V(K)}$ denotes an element-wise inner product, and $\langle \cdot,\cdot \rangle_{\partial K}$ denotes the duality pairing between $\widehat{u}$ and v over the boundary ∂K. Summing up over all K, we have

$$\left(v_{\widehat{u}}, v_{\text{opt}}^{\perp}\right)_{V(\Omega_h)} = \left\langle \widehat{u}, [\![v_{\text{opt}}^{\perp}]\!] \right\rangle_{\Gamma_h^0} = 0.$$

Thus, we can conclude that $v_{\text{opt}}^{\perp} \in \tilde{V}$ is weakly conforming. Then, by definition of the weakly conforming optimal test space $\tilde{V}_{\text{opt}}$, for a conforming optimal test function $\tilde{v}_u$ corresponding to a field variable u, we have that

$$\left(\tilde{v}_u, v_{\text{opt}}^{\perp}\right)_{V(\Omega_h)} = \left(u, A_h^* v_{\text{opt}}^{\perp}\right)_{L^2(\Omega)} = \left(v_u, v_{\text{opt}}^{\perp}\right)_{V(K)} = \left(v_u, v_{\text{opt}}^{\perp}\right)_{V(\Omega_h)} = 0,$$

where v_u is a non-conforming locally determined test function. The above orthogonality conditions imply that, for the globally conforming test space $\tilde{V}_{\text{opt}}$,

$$0 = \left(\tilde{v}, v_{\text{opt}}^{\perp}\right)_V = \left(v_{\text{opt}} + v_{\text{opt}}^{\perp}, v_{\text{opt}}^{\perp}\right)_V = \left(v_{\text{opt}}^{\perp}, v_{\text{opt}}^{\perp}\right)_V = \left\| v_{\text{opt}}^{\perp} \right\|_V^2 .$$

□

We note that Lemma 1 still holds in the case where optimal test functions spanning V_{opt} are approximated using V_h, the enriched space, so long as V_h is a closed subspace of V. In other words, the space of globally optimal test functions approximated using a weakly conforming enriched space $\tilde{V}_h$ is contained within the approximate optimal test space $V_{\text{opt},h}$.

Lemma 1 has an immediate consequence concerning DPG solutions under weakly conforming test spaces.

Lemma 2. *Let $\tilde{u}$ be the field component of the DPG solution under a weakly globally conforming optimal test space, and let u be the field component of the standard DPG solution. Assuming both problems are uniquely solvable, $u = \tilde{u}$.*

Proof. This can be shown by taking the variational problem $b((u, \widehat{u}), v) = l(v)$ for u and $\widehat{u}$; since $\tilde{V}_{\text{opt}} \subseteq V_{\text{opt}}$, we can substitute in for v the weakly conforming optimal test functions spanning $\tilde{V}_{\text{opt}}$. Doing so reduces $b((u, \widehat{u}), v) = l(v)$ to the problem for $\tilde{u}$. □

4.1.3 DPG as a non-conforming method over the test space

While DPG optimal test functions under the ultra-weak variational formulation are determined locally, Lemma 1 demonstrates that these test spaces are in fact weakly-conforming approximations to globally determined test spaces. These results are summarized in Figure 4.1.

DPG: conforming local test spaces	**DPG: nonconforming global approximations**
Begin with broken test space $V(\Omega_h)$	Begin with conforming test space $V(\Omega)$
⇓	⇓
Exact locally conforming test functions	Exact globally conforming test functions
⇓	⇓
Use a locally conforming discretization	Weakly-conforming formulation
⇓	⇓
Non-conforming test space	Discretize weakly-conforming space
⇓	⇓
Contains weakly-conforming test functions	Weakly-conforming test functions

Figure 4.1: We can interpret DPG as constructing weakly-conforming approximations to globally conforming test spaces by beginning with either broken or globally conforming test spaces.

Furthermore, while Lemma 1 demonstrates that global properties are present in locally determined DPG test spaces, Lemma 2 emphasizes that the L^2 solution u depends solely on global, and not local, properties of the test space. This is due to the fact that we can eliminate the trace

component $\widehat{u}$ of the solution $(u, \widehat{u})$ by testing only with weakly conforming global test functions. Under this perspective, design of the enriched space V_h should focus not on resolving local but *global* features of test functions.

Finally, DPG can also be viewed as a non-conforming method over V under another perspective: work by Dahmen et al. in [38] uses the same functional setting as DPG; however, their starting point is to view the above problem into a saddle point system. Defining the error $e = R_V^{-1}(l - Bu) \in V'$, the variational problem with optimal test functions can be written as follows: solve for (e, u_h) such that

$$(e, \delta v)_V + b(u_h, \delta v) = l(\delta v), \quad \delta v \in V,$$

$$b(\delta u, e) = 0, \quad \delta u \in U_h.$$

The first equation defines the error representation function e as the Riesz inversion of the residual; the second equation defines orthogonality of the error in the energy inner product

$$\begin{cases} b(\delta u, e) = 0 \\ \\ \forall \delta u \in U_h \end{cases} \Longleftrightarrow \left\langle R_V^{-1}(l - Bu), B\delta u \right\rangle_{V \times V'} = 0 \Longleftrightarrow \left(R_V^{-1} B(u - u_h), R_V^{-1} B\delta u\right)_V = 0,$$

where the last condition is exactly the orthogonality of error $(B(u - u_h), B\delta u)_{V'} = 0$ with respect to the dual inner product.

If we use for $b(u, v)$ the ultra-weak variational formulation and use for $(\cdot, \cdot)_V$ a localizable inner product, we recover the DPG method. Under this saddle-point formulation, DPG takes the form

$$(e, \delta v)_{V(K)} + (u_h, A_h^* \delta v)_{L^2\Omega} + \langle \widehat{u}_h, v \rangle_{\Gamma_h} - l(\delta v) = 0, \quad \delta v \in V(K), \quad \forall K \in \Omega_h,$$

$$(\delta u, A_h^* e)_{L^2(K)} = 0, \quad (\delta u, 0) \in U_h, \quad \forall K \in \Omega_h,$$

$$\left\langle \widehat{\delta u}, [\![e]\!] \right\rangle_{\Gamma_h} = 0, \quad \left(0, \widehat{\delta u}\right) \in U_h.$$

If $\widehat{\delta u}$ comes from the space of polynomials of order p, then the above problem can be interpreted as a non-conforming DG method for e, where elements are coupled together by enforcing that inter-element jumps of the error representation $[\![e]\!]$ are orthogonal with respect to all polynomials up to order p defined on the element edge.

4.1.4 The graph test norm and L^2-optimal test functions

We can connect DPG's weakly-conforming test functions back to the L^2-optimal test spaces through a variant of the graph norm. Under a modification of the graph test norm

$$\|v\|^2_{\tilde{H}_{A^*}(\Omega)} = \|A^*v\|^2_{L^2(\Omega)} + \delta \|v\|^2_{L^2(\Omega)},$$

a regularizing L^2 term of magnitude δ is added to the seminorm term $\|A^*_h v\|^2_{L^2(\Omega)}$. In order to both guarantee positive-definiteness and to produce a localizable test norm, $\delta > 0$ is required. As this regularizing factor is removed in the limit $\delta \to 0$, however, we naturally recover a weakly-conforming approximation to the L^2-optimal test space.

Recall that a linear operator is continuously invertible only if it is bounded below; thus, for our problem $Au = f$ to have a solution, we require $\|Au\|_{L^2(\Omega)} > \gamma \|u\|_{L^2(\Omega)}$ for some constant $\gamma > 0$ and some arbitrary u. Classical theory gives that the adjoint is bounded below as well with the same constant

$$\|A^*v\|_{L^2(\Omega)} > \gamma \|v\|_{L^2(\Omega)}, \quad v \in \tilde{H}_{A^*}(\Omega).$$

If boundedness below holds, then $\|A^*v\|_{L^2(\Omega)}$ itself is a norm, and is equivalent to $\|A^*v\|^2_{L^2(\Omega)} + \delta \|v\|^2_{L^2(\Omega)}$ such that

$$\|A^*v\|_{L^2(\Omega)} \leq \|v\|_{\tilde{H}_{A^*}(\Omega)} \leq \left(1 + \frac{\delta}{\gamma}\right) \|A^*v\|_{L^2(\Omega)}.$$

Thus, as $\delta \to 0$, $\|v\|_{\tilde{H}_{A^*}(\Omega)}$ converges to $\|A^*v\|_{L^2(\Omega)}$. Under the test norm $\|A^*v\|_{L^2(\Omega)}$, globally

optimal test functions v_u satisfy the variational problem

$$(A^* v_u, A^* \delta v)_{V(\Omega)} = (u, A^* \delta v)_{L^2(\Omega)}, \quad \forall \delta v \in \tilde{H}_{A^*}(\Omega),$$

which is a least-squares formulation of the strong adjoint equation $A^* v_u = u$. Substituting this into the ultra-weak formulation with u replaced with $u_h \in U_h$ returns

$$(u - u_h, A^* v_{\delta u}) = (u - u_h, \delta u) = 0, \quad \delta u \in U_h$$

which we recognize as the condition under which u_h is the best L^2-projection of the solution.

In practice, DPG uses the variant on the broken graph test norm

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

and approximates optimal test functions using a weakly-conforming discretization. We can repeat the above analysis for weakly-conforming test spaces[5] to show that, as $\delta \to 0$, DPG optimal test functions satisfy the variational problem

$$(A_h^* v_u, A_h^* \delta v)_{V(\Omega_h)} = (u, A_h^* \delta v)_{L^2(\Omega)}, \quad \forall \delta v \in \tilde{V}(\Omega_h),$$

which is nothing more than a weakly conforming approximation[6] to the least squares problem for the L^2-optimal test space.

4.2 DPG test functions for the convection-diffusion equation

The weakly conforming properties of DPG test spaces proven in Lemma 1 motivate the question of how these global features manifest for specific problems. We explore this question in

[5]The difference between conforming and weakly-conforming approximations lies in the fact that the two generate different discrete boundedness-below constants γ_h. For a conforming method, it is known that $\gamma_h > \gamma$; for the weakly-conforming case, the relationship is less certain. An upcoming paper on wave propagation problems will aim to address this in more detail.

[6]We note that this property holds irregardless of boundary conditions: L^2-optimal test spaces satisfy certain boundary conditions that are built into the test space $\tilde{H}_{A^*}$, but a weak satisfaction of these boundary conditions by DPG test functions is automatically achieved by weak conformity.

more depth in context of the convection-diffusion equation for convection-dominated regimes

$$\nabla \cdot (\beta u - \epsilon \nabla u) = f, \quad \text{in } \Omega$$

$$u = u_0, \quad \text{on } \Gamma,$$

where $|\Omega| = O(1)$, and $\epsilon \ll 1$. Recall that the ultra-weak variational formulation for the convection-diffusion equation is given as

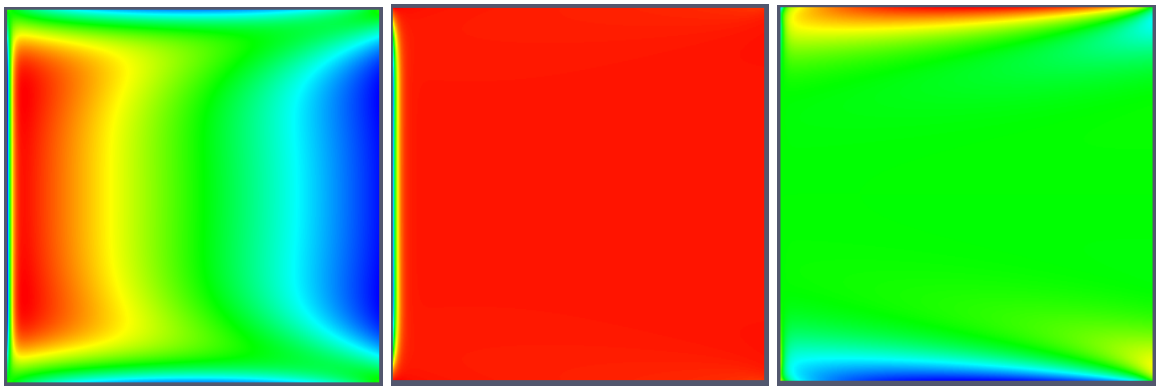
$$b\left(\left(u, \sigma, \widehat{u}, \widehat{f}_n\right), (v, \tau)\right) = \langle \widehat{u}, \tau_n \rangle_{\Gamma_h} + \left\langle \widehat{f}_n, v \right\rangle_{\Gamma_h} + (u, \nabla_h \cdot \tau - \beta \cdot \nabla_h v)_{L^2(\Omega)} + \left(\sigma, \frac{1}{\epsilon}\tau + \nabla_h v\right)_{L^2(\Omega)}.$$

4.2.1 Localization and boundary layers under the graph test norm

The graph test norm typically delivers near-optimal results - for problems of wave propagation, elasticity, and Stokes flow, DPG under the graph test norm delivers solutions nearly indistinguishable from the L^2 projections of the exact solution [37, 39, 30]. However, when applied to the convection-diffusion problem, the graph test norm performed very poorly in comparison.

This poor performance can be explained by the difficulty in approximating optimal test functions using our choice of enriched space V_h. It can be demonstrated that the optimal test functions generated under the graph test norm for convection-diffusion problems develop strong boundary layers of width ϵ. Figure 5.1 shows the result of 2D numerical experiments where a fine mesh was used to resolve an optimal test function resulting from the auxiliary problem (3.5), demonstrating the presence of strong boundary layers at the element inflow boundary ∂K_{in}. All figures of optimal test functions are produced using the FEniCS codebase [44].

Under the graph norm for convection-diffusion, the variational problem for test functions



(a) v (b) τ_x (c) τ_y

Figure 4.2: v and τ components of the 2D optimal test functions corresponding to the basis function $u = 1$ on the reference element for $\epsilon = 0.01$. The solution has been obtained using a fine 128×128 mesh of triangles, with $p = 3$.

over a single element is given as

$$\begin{aligned}((v,\tau),(\delta v,\delta\tau))_{V_{\text{graph}}(K)} = (u, \nabla_h \cdot \tau - \beta \cdot \nabla_h v)_K + \left(\sigma, \frac{1}{\epsilon}\tau + \nabla_h v\right)_{(K)} \\ + \langle \widehat{u}, \tau_n \rangle_{\partial K} + \left\langle \widehat{f}_n, v \right\rangle_{\partial K}\end{aligned}$$

where $((v,\tau),(\delta v,\delta\tau))_{V_{\text{graph}}(K)}$ is

$$\begin{aligned}((v,\tau),(\delta v,\delta\tau))_{V_{\text{graph}}(K)} = &(\nabla \cdot \tau - \beta \cdot \nabla v, \nabla \cdot \delta\tau - \beta \cdot \nabla \delta v)_{L^2(K)} \\ &+ \left(\epsilon^{-1}\tau - \nabla v, \epsilon^{-1}\delta\tau - \nabla\delta v\right)_{L^2(K)} + (v, \delta v)_{L^2(K)}\end{aligned}$$

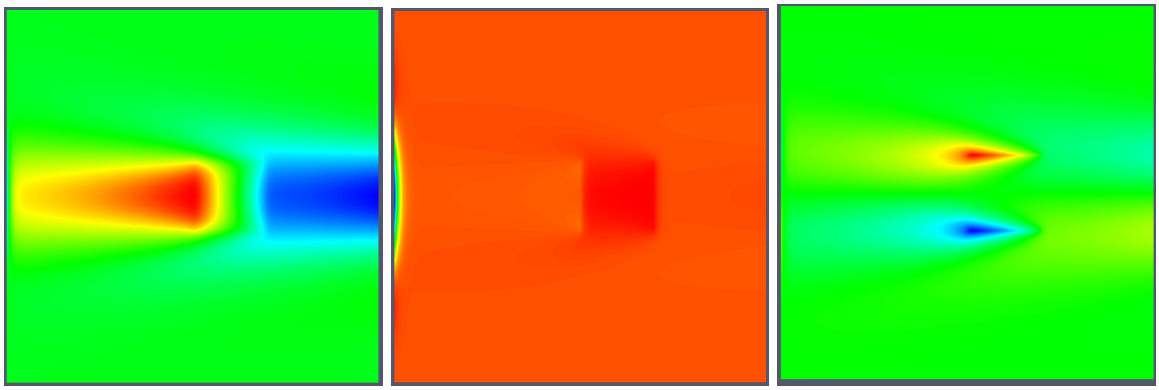
We can transform the above problem to the reference element; applying this simple scaling argument shows that, for elements of size h, we can expect a boundary layer of width h/ϵ relative to a unit domain.[7] In other words, the strength of the boundary layer is proportional to the element Peclet number $\text{Pe} = h/\epsilon$. For severely underresolved meshes where $h \gg \epsilon$, this makes

[7]This is assuming that the parameter ϵ dictates the width of expected boundary layers on a unit domain. The strong form of the above variational problem corresponds to a reaction-diffusion system, for which we expect this assumption to hold. Numerical experiments also appear to confirm that the boundary layer is of width $O(\epsilon)$.

the approximation of optimal test functions using a simple p-enriched space very difficult, though specially designed hp-Shishkin subgrid meshes have been used to resolve such test functions with some success in [40] (these are discussed further in Section 4.3.3). The construction of alternative test norms in [3] was motivated by the computational difficulty in applying the graph test norm to heavily convection-dominated regimes where $\epsilon \ll 1$.

Consider now globally optimal test functions under the test norm, where Problem (3.5) is solved using the weakly conforming space $\tilde{V}$. By the same scaling argument, strong boundary layers appear, but only at the global inflow boundary Γ_{in}. We illustrate this in Figure 4.3, where we use an $H^1 \times H(\text{div})$-conforming finite element space to approximate the globally conforming test function. We note that, by approximating optimal test functions using a conforming test space, our test functions no longer produce boundary layers over every single element. However, in return, our optimal test functions now contain non-local information - in particular, optimal test functions for trial functions with support only in the interior of the domain now contain boundary layers at the domain inflow boundary Γ_{in}.

In light of the non-local nature of globally optimal test functions, we might exploit the fact that, under DPG with the ultra-weak variational formulation, our locally determined test space naturally contains non-local information as well. By Lemma 1, we know that the globally (weakly) conforming optimal test space is contained within our locally determined optimal test space; by Lemma 2, we have that the DPG solutions under local and conforming optimal test spaces coincide up to L^2 field variables. In other words, the approximation of the field solution u is solely determined by the approximation of globally optimal test functions by the weakly conforming space enriched space $\tilde{V}_h$.



(a) v (b) τ_x (c) τ_y

Figure 4.3: v and τ components of the 2D optimal test functions corresponding to the piecewise constant $u = 1$ with support on a quad element defined on $[.5, .7] \times [.4, .6]$ for $\epsilon = 0.01$. Note the presence of nonlocal behavior in the form of the boundary layer at the inflow boundary.

4.3 Global effects in numerical experiments

We investigate now non-local effects present in the DPG optimal test spaces for convection-diffusion under the test norm. In particular, we focus on the resolution of boundary layers in optimal test functions for DPG and their effect on the *robustness* of the method with respect to ϵ. Under fully resolved optimal test functions under the graph test norm, we expect the DPG method to be robust in ϵ, and that any observed non-robustness can be attributed to approximation error in computed test functions.

4.3.1 Robustness

The difficulty encountered by most numerical and finite element methods for boundary layer solutions of convection-diffusion problems is a lack of *robustness* in the diffusion parameter ϵ; in other words, for a fixed resolution/number of degrees of freedom, as ϵ decreases, the finite element error degrades with respect to the best approximation error. This can be seen in typical error bounds for

finite element methods; if finite element error is appropriately measured in some norm $\|\cdot\|_U$, then

$$\frac{\|u - u_h\|_U}{\inf_{w_h \in U_h} \|u - w_h\|_U} = O(\epsilon^{-1}).$$

Under naive finite elements, which relies on the coercivity of the bilinear form to provide stability, the dependence of the above ratio on ϵ can be connected to the discrete coercivity constant, which (for appropriate assumptions on boundary conditions and β) is $O(\epsilon)$ with respect to the H^1 norm or seminorm [10]).[8]

4.3.2 Adaptivity and adjoint boundary layers

We adopt a modification of a problem first proposed by Eriksson and Johnson in [46] and later used in [3] to determine the robustness of DPG with respect to the diffusion parameter ϵ. 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. The use of adaptive quadrature was used to ensure accurate reporting of errors for solutions with boundary layers, and all computations have been done using the higher-order adaptive DPG codebase Camellia, built on the Sandia toolbox Trilinos [5].

For boundary conditions, we impose

$$u = u_0, \quad x = 0,$$

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

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

[8] These assumptions can be relaxed slightly in the presence of a first order term [45].

In this case, our exact solution is the series

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

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

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

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.

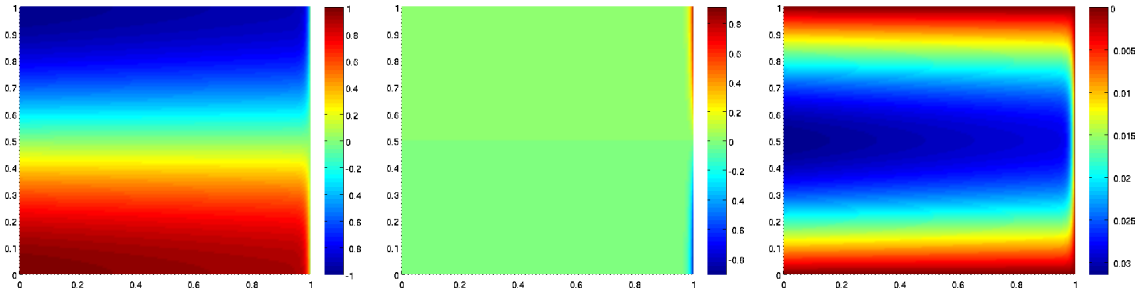


Figure 4.4: Exact solution for u, σ_x, and σ_y for $\epsilon = .01$, $C_1 = 1$, $C_n = 0$, $n \neq 1$

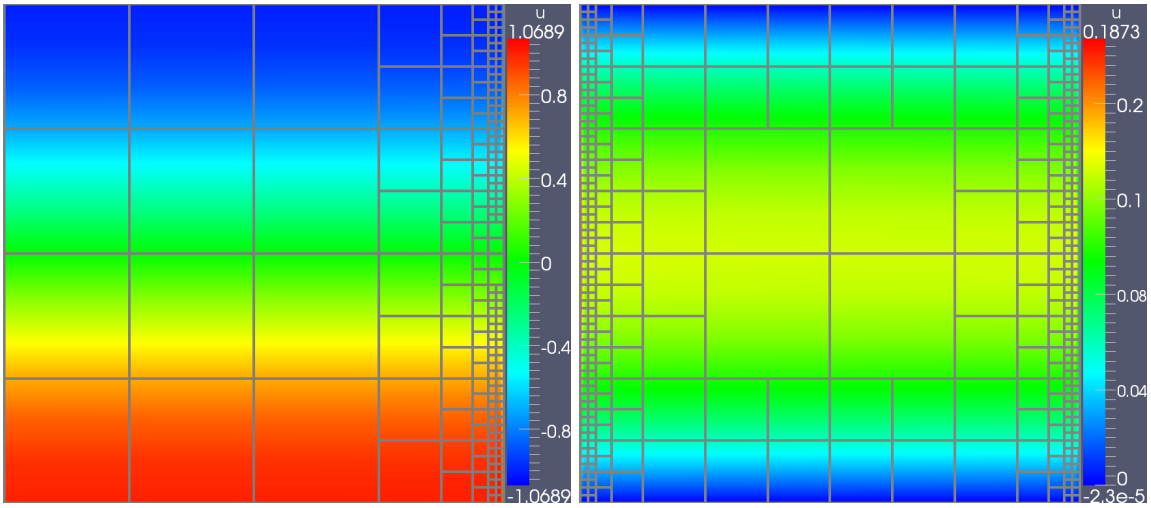
We begin first by recalling a phenomena observed early on in the application of DPG to convection-dominated diffusion problems. Under the functional setting of DPG, the choice of test norm defines the optimal test space, but also defines the norm in which error is measured. Early experiments in [7] by Demkowicz, Gopalakrishnan and Niemi demonstrated in computational experiments that, for naively chosen test norms, not only would the solution exhibit degeneration on

a fixed mesh as $\epsilon \to 0$, but under automatic adaptivity based on the error representation function, refined meshes would tend to exhibit strong refinements at the inflow boundary.[9]

Figure 4.5 shows two numerical solutions of the Erikkson-Johnson problem where the inflow profile $u_{\text{in}}(y) = y(1-y)$ along $x = 0$ is convected from left to right, terminating with a boundary layer at the outflow $x = 1$. The left figure shows the trial solution u under a test norm introduced in [3], which is shown to induce a DPG method whose solutions which do not degenerate as $\epsilon \to 0$. We speculate that the presence of refinements at the inflow under the naive test norm illuminates an interesting heuristic observation concerning DPG for convection-diffusion problems; if the form of your test norm neglects to account for boundary layers in optimal test functions, their effects will show up in the energy error, and the method will still seek to resolve the optimal test space through minimization of error via adaptive mesh refinement.

Motivated by the phenomena observed in Figure 4.5, we tested the effect of resolving the boundary layers in globally optimal test functions through the h-refinement of elements adjacent to the inflow boundary. Under the graph test norm, we expect such globally determined test functions to exhibit strong boundary layers, but only at the inflow boundary. Recalling Lemma 1, we have that the weakly conforming globally optimal test space is a proper subset of the direct sum of locally optimal test spaces over each element; thus, *a-priori* refinements of the mesh that anticipate the presence of an inflow boundary layer in computed test functions should allow for a better resolution of the global optimal test space and improve approximation properties. Additionally, since the solution is relatively smooth near $x = 0$, we do not expect additional mesh resolution near the inflow to significantly affect the best approximation error.

[9] Cohen, Dahmen and Welper reported similar results in [38] for a different variational formulation and test norm. The missing factor in choosing a well-behaved test norm appears to be the presence of the $O(1)$ streamline derivative term $\|\beta \nabla v\|_{L^2}$.



(a) Robust test norm (b) $H^1 \times H(\mathrm{div})$ test norm

Figure 4.5: An example of automatic mesh refinement under both a robust test norm (left) and a naively chosen non-robust test norm (right) for $\epsilon = .001$. The naively chosen test norm exhibits both degeneration of the solution and extraneous refinements at the inflow boundary. Both figures are produced after four automatic refinements after beginning on a uniform mesh of 4 quadratic elements with $\Delta p = 3$.

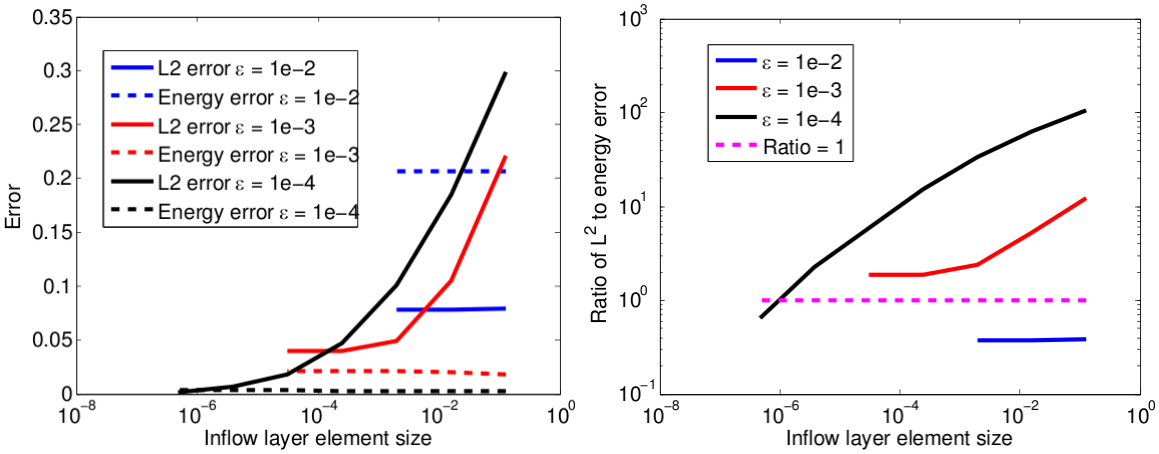


Figure 4.6: The effects of increased mesh resolution at the inflow boundary on L^2 and energy error (left) and their ratio (right). Experiments were done beginning on a uniform mesh of 8-by-8 quadratic elements, with $\Delta p = 3$.

Figure 4.6 demonstrates the effect of increased mesh resolution at the inflow boundary; under no additional inflow resolution, the ratio of L^2 to energy error grows as ϵ decreases, indicating a loss of robustness as discussed in Section 4.3.1. However, under increased inflow resolution, the ratio can be driven down to be $O(1)$. We expected initially for robustness to be restored under resolution of the diffusion scale; however, for $\epsilon \leq 1e-3$, achieving an $O(1)$ ratio required an order of magnitude finer resolution than the $h = O(\epsilon)$. The reasons for this may lie in the difference between boundary layers in optimal test functions for field and flux variables and optimal test functions for trace variables.

4.3.3 Under-resolution of boundary layers in optimal test functions

We note that globally optimal test functions produce strong boundary layers at the global inflow boundary. However, numerical experiments appear to indicate that the boundary layers in test functions for trace variables are stronger than layers in test functions for field and flux variables. Note that the auxiliary problem for test functions under the abstract graph test norm is

$$(v, \delta v)_{V_{\text{graph}}} = (A_h^* v, A_h^* \delta v)_{L^2(K)} + (v, \delta v)_{L^2(K)} = (u, A_h^* v)_{L^2(K)} + \langle \widehat{u}, v \rangle_{\partial K}$$

and induces a strong problem with a specific load. The corresponding strong form of the problem has boundary conditions $\gamma(A_h^* v) = \langle u, v \rangle - \langle \widehat{u}, v \rangle$, where $\gamma(A_h^* v)$ is the trace resulting from integration by parts of the operator A_h^*. For the convection-diffusion equation, this translates to the boundary conditions

$$\left(\frac{1}{\epsilon}\tau_n + \frac{\partial v}{\partial n}\right) - \beta_n \left(\nabla \cdot \tau - \beta \cdot \nabla v\right) = \widehat{f}_n - (\beta_n u - \sigma_n)$$

$$\nabla \cdot \tau - \beta \cdot \nabla v = \widehat{u} - u.$$

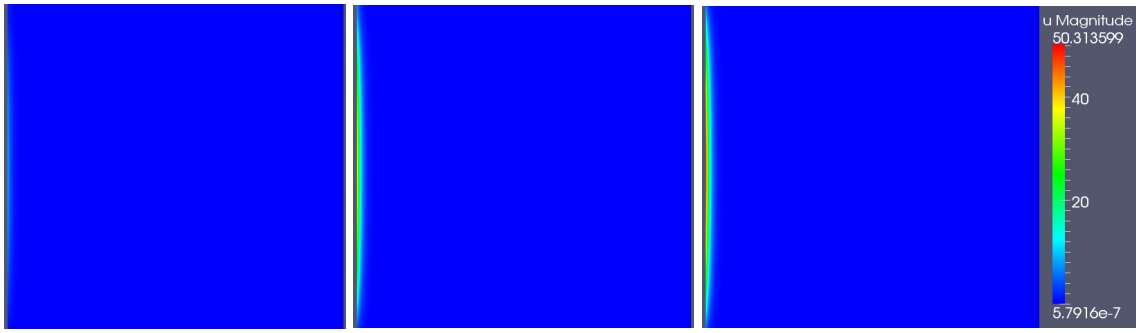
where u and σ are field variables, and $\widehat{f}_n$ and $\widehat{u}$ are the trace and flux variables for the convection-diffusion equation under the ultra-weak variational formulation. Reducing these two equations into

one, we have

$$\frac{1}{\epsilon}\tau_n + \frac{\partial v}{\partial n} = \widehat{f}_n - (\beta_n u - \sigma_n) + \beta_n (\widehat{u} - u)$$

The presence of the $\frac{1}{\epsilon}$ term in the boundary condition further increases the strength of the boundary layer when the auxiliary problem for a test function (3.5) is loaded with a non-zero $\widehat{u}$.

Figure 4.7 shows the magnitude of the τ component of optimal test functions for a trace loaded on the inflow boundary for quadratic meshes with increasing anisotropic resolution in the x-direction. Even between the over-resolved cases $h_x \approx \epsilon$, $h_x \approx 2^{-1}\epsilon$, and $h_x \approx 2^{-2}\epsilon$, the observed boundary layer in the optimal test functions continues to grow in magnitude (as evidenced by Table 4.3.3, implying that mesh resolution at the diffusion scale is still insufficient to resolve these features.



(a) 128×128 mesh (b) 256×128 mesh (c) 512×128 mesh

Figure 4.7: Magnitudes of the vector-valued component τ of the optimal test functions corresponding to the flux $\widehat{u} = y(1-y)$ on the boundary $x = 0$ for $\epsilon = 0.01$ over the reference element.

This phenomena was observed also in [47], where hp-Shishkin submeshes (discussed in [48]) were used to resolve optimal test functions locally over an element. In 1D, these meshes consist of a two small "needle" elements of size $h = p\epsilon$, where p is the polynomial order of the approximating functions (the 2D extension for quadrilateral elements in 2D is straightforward, and is given in detail

Mesh	Maximum value of $\lvert\tau_{\widehat{u}}\rvert$ for $\epsilon = .01$
128×128	8.6510571
256×128	38.794881
512×128	50.313599

Table 4.1: Maximum magnitudes of the τ component of the optimal test functions corresponding to the flux $\widehat{u} = y(1-y)$ on the boundary $x = 0$ for $\epsilon = 0.01$ over the reference element.

in [47]). Experiments using p-refined Shishkin subgrid meshes indicated that the relative error in the approximation of optimal test functions for traces $\widehat{u}$ on the inflow boundary did not converge to zero as the dimension of the enriched space V_h was increased (unlike the relative error for optimal test functions for field variables u, σ, and flux $\widehat{f}_n$), and it was concluded that the use of Shishkin meshes was insufficient to resolve the boundary layers present in optimal test functions for traces over inflow edges for sufficiently small ϵ.

The result of this under-resolution of optimal test functions having to do with traces on the inflow boundary is an underestimation of the inflow boundary term $\langle \widehat{u}, [\![v]\!] \rangle_{\Gamma_{\text{in}}}$, which leads to the underestimation of the solution in experiments [47], and explains the fine mesh resolution on the inflow boundary necessary to restore robustness to the DPG method in Figure 4.6.

Chapter 5

A robust DPG method for convection-diffusion

An obvious choice for the test norm would have been the quasi-optimal norm described in the previous chapter; 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 [3, 42, 30]. However, as was demonstrated, computations with the quasi-optimal test norm for convection-diffusion problems turn out to be quite problematic for small diffusion and coarse meshes.

In the application of DPG in [25, 6, 7, 37], 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.[1] 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

[1] 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.

delivering a best approximation, the energy error for a coarse mesh must be greater than or equal 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.

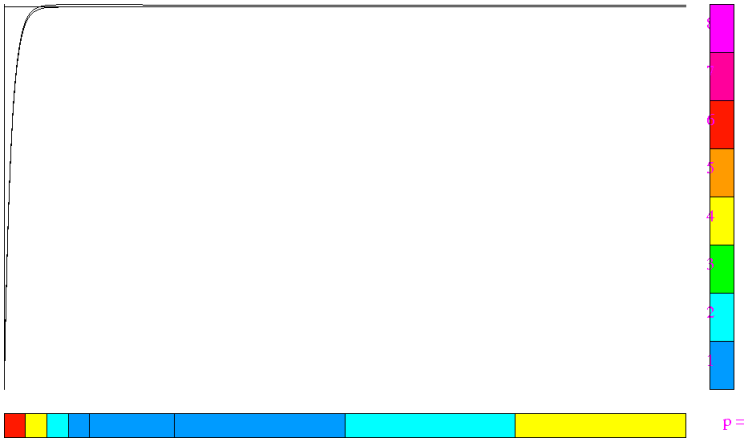


Figure 5.1: v and τ components of the 1D optimal test functions corresponding to the flux $\widehat{f}_n$ on the right-hand side of a unit element for $\epsilon = 0.01$. The solution has been obtained using automatic hp-adaptivity driven by the test norm with the error tolerance set at 1%.

5.1 A new inflow boundary condition

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 $\| (v,\tau) \|_V^2 = \|v\|^2_{H^1(\Omega_h)} + \|\tau\|^2_{H(\mathrm{div},\Omega_h)}$, 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 [7, 3].

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 5.2, 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 also adopt a new inflow boundary condition in this chapter. Previous work in [3] adopted Dirichlet boundary conditions everywhere on Γ. We employ the inflow condition of Hesthaven *et al.* [49], 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[2]; however, the latter

[2]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

induces a more "well-behaved" adjoint problem than the former, which, as we will discuss, affects the performance of DPG. Physically speaking, as the above boundary condition corresponds to the integration by parts of the conservation equation, a boundary condition on $\beta_n u - \sigma_n$ models a boundary condition on the conserved flux.

5.1.1 Norms on U

With the above boundary conditions at hand, we can specify norms on both trial and test spaces which we will use to perform a rigorous mathematical analysis of the DPG method under a given test norm. 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 3.3, suggests the following trial space (see [41, 42] for details):

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

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

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

$u=u_0$ on Γ_- for sufficiently small ϵ.

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

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

5.1.2 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

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

5.1.3 Analysis of a robust test norm

Under this new boundary condition, we adopt the 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 5.1.4. This work is intended to act as an extension of

work presented by Heuer and Demkowicz in [3]. The primary focus of the work 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.

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, by choosing $(v, \tau) \in H^1(\Omega) \times H(\mathrm{div}, \Omega)$ such that

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

and that boundary terms vanish, 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 $H(\text{div})$-conforming finite elements. A more detailed discussion on the reasoning behind the construction of test norms can be found in [3].

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

$$\left\| \left(u, \sigma, \widehat{u}, \widehat{f}_n \right) \right\|_U^2 \coloneqq \|u\|^2 + \|\sigma\|^2 + \|\widehat{u}\|^2 + \left\| \widehat{f}_n \right\|^2 . \tag{5.1}$$

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.

[3]Sharper estimates are attainable if these constants are allowed to vary over the mesh Ω_h. See Section 5.1.5 for a discussion.

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 $\widehat{f_n}$

$$\left\|\left(u,\sigma,\widehat{u},\widehat{f_n}\right)\right\|_{U,i}^2 \coloneqq \left(C_u^i\|u\|\right)^2 + \left(C_\sigma^i\|\sigma\|\right)^2 + \left(C_{\widehat{u}}^i\|\widehat{u}\|\right)^2 + \left(C_{\widehat{f}_n}^i\|\widehat{f}_n\|\right)^2, \quad i=1,2 \tag{5.2}$$

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 3.2 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 in Section 3.2 on the duality between test norms/energy norms, the norms (5.2) 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 5.1.3.1).

2. Derive adjoint estimates (Section 5.1.3.2).

5.1.3.1 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

$$\tau_i \cdot n = 0, \quad \text{on } \Gamma_- \tag{5.3}$$

$$v_i = 0, \quad \text{on } \Gamma_+ \tag{5.4}$$

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 5.1.3), 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 $\|\cdot\|_{U,2}$:

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

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 [41] and adapted to convection-diffusion in [41] and [3]. However, since $(\tau, v) \in H(\text{div}, \Omega) \times H^1(\Omega)$, 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

[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.

equations

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

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

$$\tau_1 \cdot n|_{\Gamma_-} = 0,$$

$$v_1|_{\Gamma_+} = 0.$$

and

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

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

$$\tau_2 \cdot n|_{\Gamma_-} = 0,$$

$$v_2|_{\Gamma_+} = 0.$$

More generally, we can analyze the adjoint equations

$$\epsilon^{-1}\tau + \nabla v = f \tag{5.5}$$

$$\nabla \cdot \tau - \beta \cdot \nabla v = g, \tag{5.6}$$

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 $\| (v_2, \tau_2) \|_V \lesssim \|f\|_{L^2}$.

5.1.3.2 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 $\beta, \nabla \cdot \beta = O(1)$ with respect to $|\Omega|$, the size of the domain. Additionally, we will assume that some or all of the following assumptions hold:[5]

$$\nabla \times \beta = 0, \quad 0 < C \leq |\beta|^2 + \frac{1}{2}\nabla \cdot \beta, \quad C = O(1), \tag{5.7}$$

$$\nabla\beta + \nabla\beta^T - \nabla \cdot \beta I = O(1), \tag{5.8}$$

$$\nabla \cdot \beta = 0. \tag{5.9}$$

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

- **Lemma 4**: *For β satisfying* (5.7) *and* (5.8)*, and $v_1 \in H^1(\Omega)$, satisfying equations* (5.5) *and* (5.6) *with $f = 0$, and with boundary conditions* (5.3) *and* (5.4),

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

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

- **Lemma 5**: *For β satisfying* (5.7)*, and $v \in H^1(\Omega)$ satisfying equations* (5.5) *and* (5.6) *and boundary conditions* (5.3) *and* (5.4)*, 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 $g = 0$.

[5]These assumptions correspond to convection fields which are divergence-free (5.9), curl-free (5.7), bounded away from zero (5.7), and of bounded variation (5.8). However, these are merely sufficient conditions; numerical experiments indicate that they may not be strictly necessary.

- **Lemma 6**: *For β satisfying* (5.7), (5.9), *and solutions $v_0 \in H^1(\Omega_h)$ and $\tau_0 \in H(\mathrm{div}, \Omega_h)$ of equations* (5.5) *and* (5.6) *with $f = g = 0$,*

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

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).

5.1.4 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

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

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

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

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 [50], the solution to the variational problem defining the optimal test function under the above test norm was computed. Figure 5.2 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

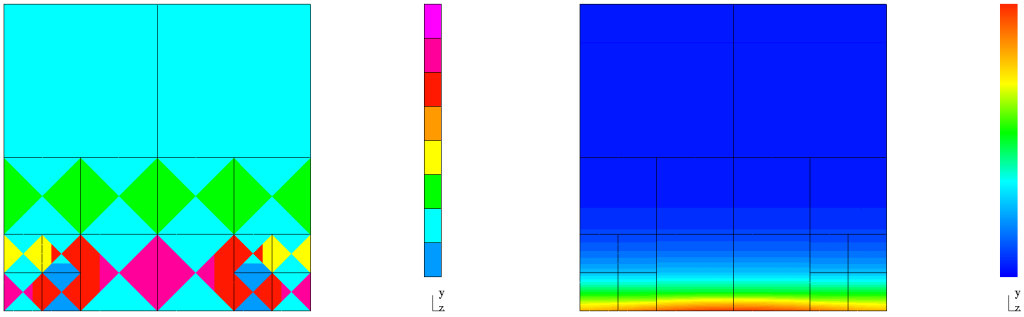


Figure 5.2: The v component of the optimal test function corresponding to flux $\widehat{u} = x(1-x)$ on the bottom side of a unit element for $\epsilon = 0.01$. The corresponding *hp*-mesh used to compute the solution is displayed to the left.

the optimal test functions, we follow [3] 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 now, we consider only isotropic refinements on quadrilateral elements in 2D.

We can now define $\|(v,\tau)\|_{V(K)}$, our test norm over a single element 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 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.

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

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

Lemma 3. *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 (5.7), (5.8), *and* (5.9), *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 3.2, 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 (5.5) and (5.6),

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

we can characterize the test norm for

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

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 5.1.3.1.

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 6 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 [41] gives us the Poincare inequality for discontinuous functions

  $$\|v_0\| \lesssim \|\nabla v_0\| + \|\,[\![v]\!]\,\|.$$

  Since $g=0$, $\|\nabla\cdot\tau_0\| = \|\beta\cdot\nabla v_0\| \lesssim \|\nabla v_0\|$, which we now have control over as well.

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

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

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

Noting that $\nabla \cdot \tau_1 = g + \beta \cdot \nabla v_1$ gives $\|\nabla \cdot \tau_1\| \lesssim \|g\|$ as well. Lemma 5 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 5 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 5 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 (5.5), (5.6).

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

$$\begin{aligned}
\|\epsilon C_\tau f\| &\leq C_\tau\|\tau\|+\epsilon C_\tau\|\nabla v\| \lesssim \|\,(\tau,v)\,\|_{V(\Omega_h)} \\
\|g\| &\leq \|\nabla\cdot\tau\|+\|\beta\cdot\nabla v\| \lesssim \|\,(\tau,v)\,\|_{V(\Omega_h)}
\end{aligned}$$

We estimate the supremum on the jumps of (τ,v) by following [3]; 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 [41], 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 $\left\| \left(u, \sigma, \widehat{u}, \widehat{f_n} \right) \right\|_E \lesssim \left\| \left(u, \sigma, \widehat{u}, \widehat{f_n} \right) \right\|_{U,2}$. □

5.1.6 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 5.1.5 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 5.1.3.2 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 5.1.5 hold in $\mathbb{R}^d$ for arbitrary d; however, we can show that for Dirichlet boundary conditions, the same results do not hold even in 1D. Consider the 1D analogue of the estimate given by Lemma 4. 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 4 in Section 5.1.3.2 is likewise reduced in 1D to the scalar equation

$$\epsilon v_1'' + \beta v_1' = -g \tag{5.10}$$

with $v_1 \in H_0^1((0,1))$. After multiplying this equation by $\beta v_1'$ and integrating 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 (5.10) exactly.

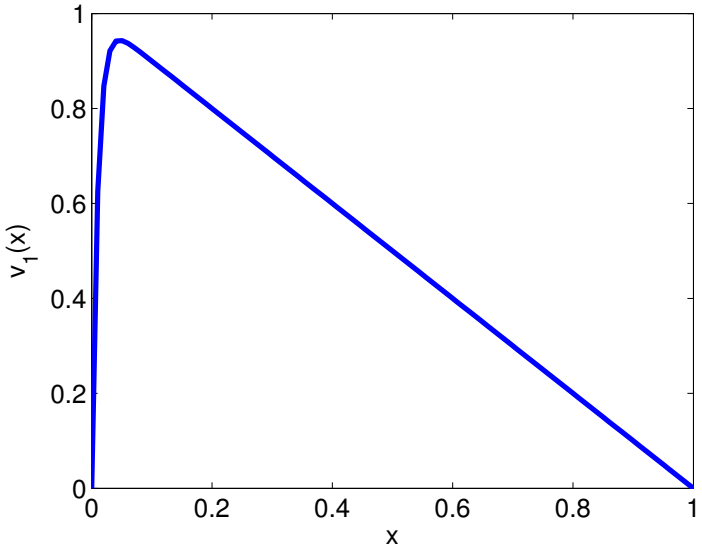


Figure 5.3: $v_1(x) = \frac{e^{-\frac{x}{\epsilon}}}{e^{\frac{1}{\epsilon}}-1}\left(e^{\frac{1}{\epsilon}}\left(e^{\frac{x}{\epsilon}}-1\right)+\left(e^{\frac{1}{\epsilon}}-1\right)e^{\frac{x}{\epsilon}}x\right)$, the solution to the adjoint equation for $f=0$ and constant β and load g for $\epsilon = .01$.

The solution v_1 is plotted in Figure 5.3, 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,[6] 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 [7], and indicate the same lack of robustness under the test norm derived in this work.[7]

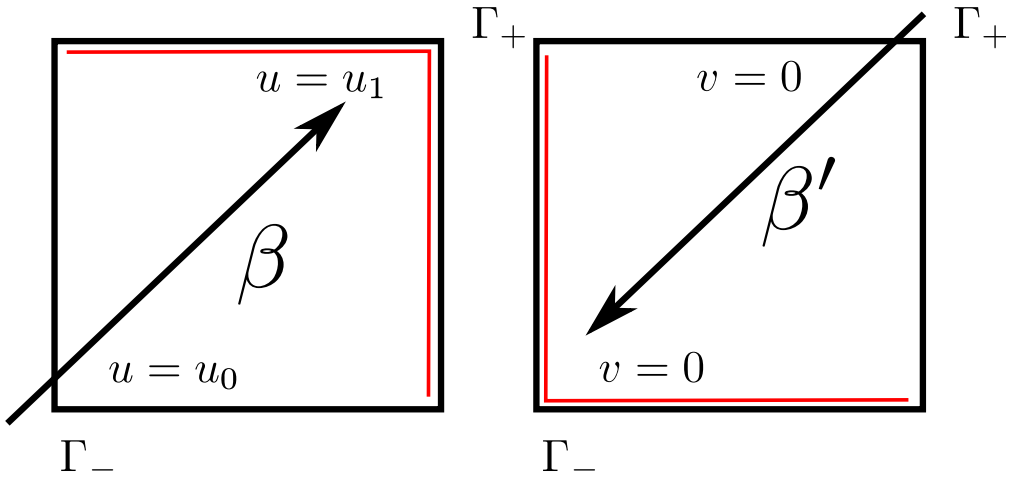
In higher dimensions, the adjoint problem is of the same form as the primal problem with

[6]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.

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

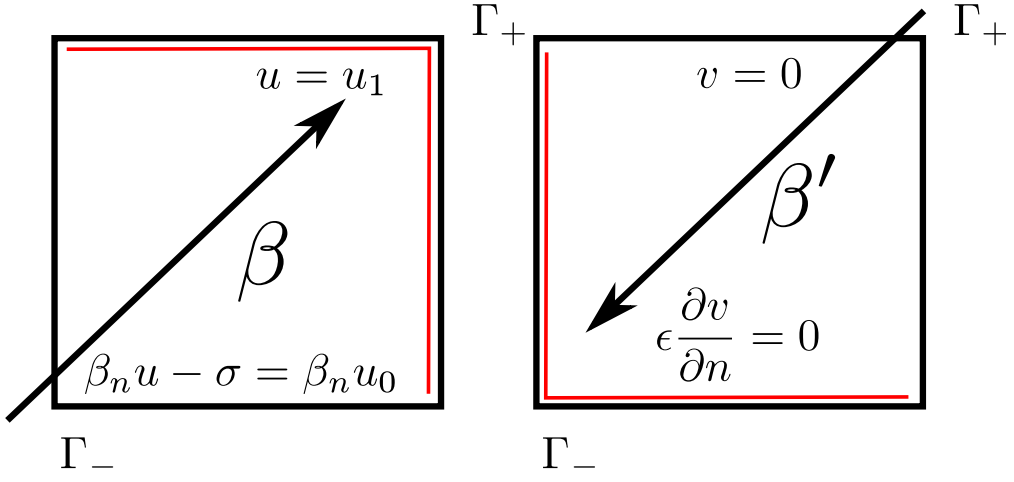
$$\|(\tau,v)\|_{V,w}^2 = \|v\| + \epsilon\|\nabla v\| + \|\beta\cdot\nabla v\|_{w+\epsilon} + \|\nabla\cdot\tau\|_{w+\epsilon} + \frac{1}{\epsilon}\|\tau\|_{w+\epsilon}$$

where $\|\cdot\|_{w+\epsilon}$ is a weighted L^2 norm, where the weight $w \in (0,1)$ is required to vanish on Γ_- and satisfy $\nabla w = O(1)$. The need for 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 5.4: 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 Dirichlet inflow condition, adjoint solutions can develop boundary layers at the outflow of the adjoint problem. Under the new inflow conditions, the wall-stop boundary condition is relaxed to a zero-stress condition at the adjoint outflow.

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 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 (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).[8] 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.

[8] We note that 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. Please refer to the earlier discussion in Section 4.3.2.

5.2 Numerical experiments

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 [3]; 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 $H(\mathrm{div};\Omega) \times H^1(\Omega)$ norm is used for the test norm $\| (\tau, v) \|_V$ (as was done in [6]), 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 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 [7] by dynamically rescaling the test norms based on element size, a practical remedy not covered yet by the present analysis.

5.2.1 Eriksson-Johnson model problem

To confirm our theoretical results, we use again the problem of Eriksson and Johnson [46] introduced in the previous chapter. This problem yields an exact solution with a boundary layer for the convection-diffusion problem with a forcing term independent of ϵ.

Unlike our previous experiments using the Eriksson-Johnson problem, which enforced Dirichlet boundary conditions on u on both inflow and outflow, we impose boundary conditions on $u = 0$

on Γ_+ and $\beta_n u - \sigma_n$ on Γ_-, which reduces to

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

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

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

The exact solution is the series

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

where $r_{1,2}$ are specified previously in Section 4.3.2, and 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 [5].

5.2.1.1 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 2 by 2 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 $\Delta p \geq 3$, and qualitatively the same for $\Delta p = 2$. h-refinements are executed using a greedy refinement algorithm, where the element energy error e_K^2 is computed for all elements K, and elements such that $e_K^2 \leq \alpha^2 \max_K e_K^2$ are refined. We make the arbitrary choice of taking $\alpha = .2$ for each of these experiments.

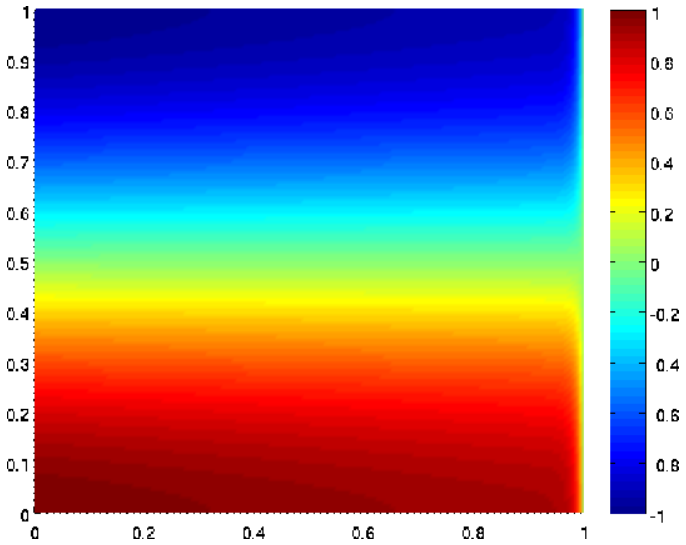


Figure 5.5: Solution for u for $\epsilon = .01$, $C_1 = 1$, $C_n = 0$, $n \neq 1$

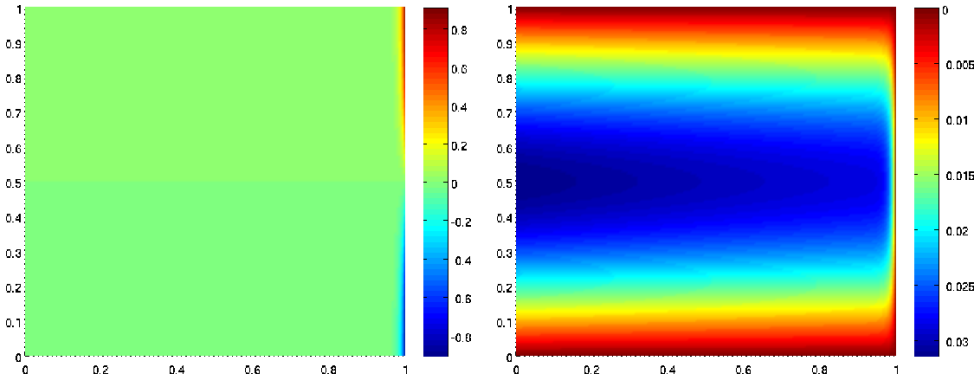


Figure 5.6: Solution for σ_x, and σ_y for $\epsilon = .01$, $C_1 = 1$, $C_n = 0$, $n \neq 1$

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 5.1.5 imply that, using the above test norm, $\|u - u_h\|_{L^2}/\|u - u_h\|_E \leq C$ independent of ϵ. Figure 5.8, which plots the ratio of L^2 to energy error, seems to imply that (at least for this model problem) $C = O(1)$. Additionally, while we do not have a robust lower bound ($\|u - u_h\|_{L^2}/\|u - u_h\|_E$ can approach 0 as $\epsilon \to 0$), our

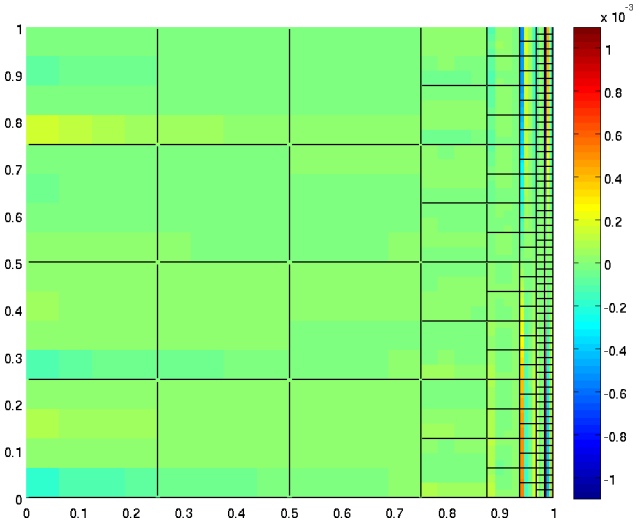


Figure 5.7: Adapted mesh and pointwise error for $\epsilon = .01$

numerical results appear to indicate the existence of an ϵ-independent lower bound.

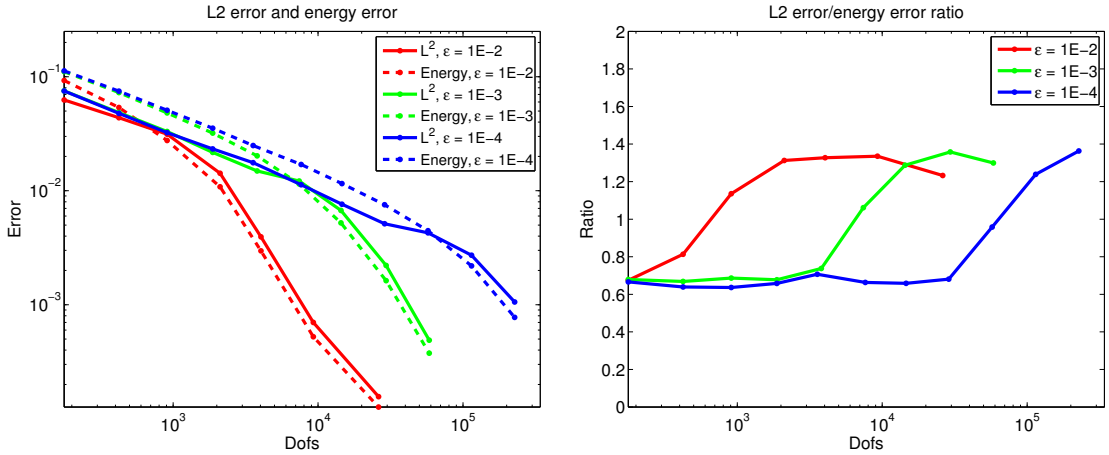


Figure 5.8: L^2 and energy errors, and their ratio for $\epsilon = .01$, $\epsilon = .001$, $\epsilon = .0001$

The effect of a mesh dependent scalings on the $\|v\|^2$ and $\|\tau\|^2$ terms in the test norm can be seen in the ratios of L^2 to energy error; as the mesh is refined, the constants in front of the L^2 terms for v and τ converge to stationary values (providing the full robustness implied by our adjoint

energy estimates), and the ratio of L^2 to energy error transitions from a smaller to a larger value. The transition point happens later for smaller ϵ, which we expect, since the transition of the ratio corresponds to the introduction of elements whose size is of order ϵ through mesh refinement. For this reason, the ratios of L^2 to energy error do not overlap perfectly with each other. We note that the introduction of anisotropic refinements appears to mitigate this effect slightly [3].

We examined how small ϵ needed to be in order to encounter roundoff effects as well. In [3], the smallest resolvable ϵ using only double precision arithmetic was $1e-4$. The solution of optimal test functions is now done using both pivoting and equilibration, improving conditioning. Roundoff effects still appear, but at smaller values of ϵ.

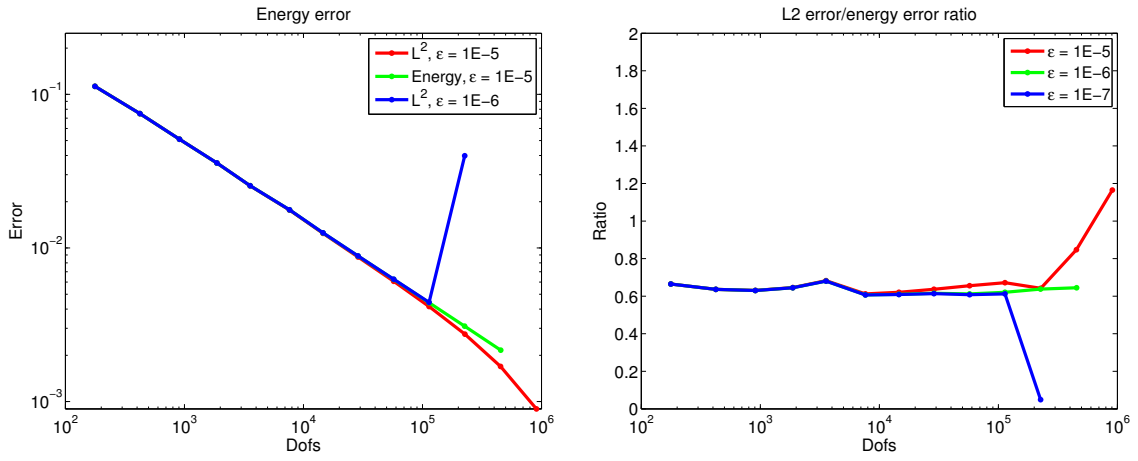


Figure 5.9: Energy error and L^2/energy error ratio for $\epsilon = 1e-5$, $\epsilon = 1e-6$, $\epsilon = 1e-7$. Non-monotonic behavior of the energy error indicates conditioning issues and roundoff effects.

Without anisotropic refinements, it still becomes computationally difficult to fully resolve the solution for 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 5.9 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, 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[9], 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.

5.2.1.2 Neglecting σ_n

In practice, we will not have prior knowledge of σ_n at the inflow, and will have to set $\beta_n u - \sigma_n = u_0$, ignoring the viscous contribution to the boundary condition. The hope is that for small ϵ, this omission will be negligible. Figure 5.10 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.

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

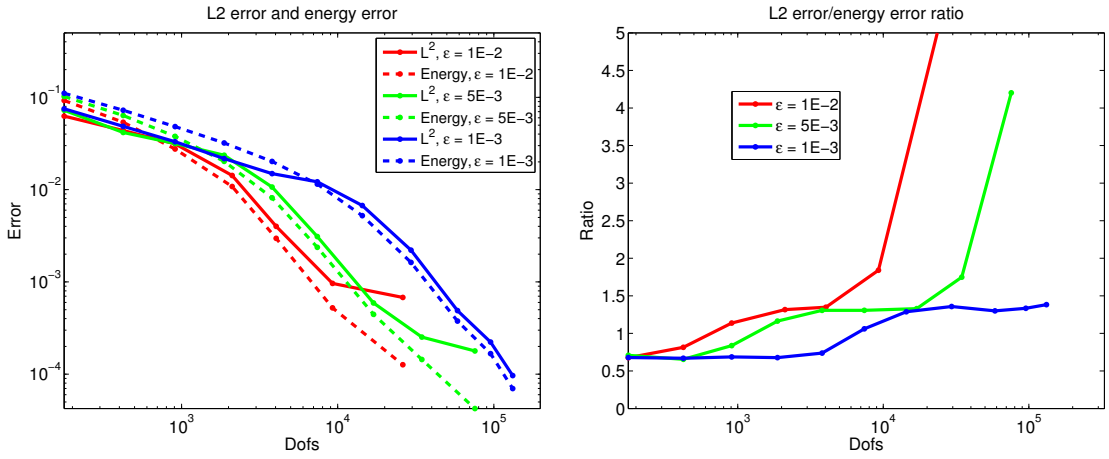


Figure 5.10: L^2 and energy errors and their ratio when neglecting σ_n at the inflow.

5.2.1.3 Discontinuous inflow data

We note also that an additional advantage of selecting this new boundary condition is a relaxation of regularity requirements: as $\widehat{f}_n \in H^{-1/2}(\Gamma_h)$, 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 5.11 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. However, the

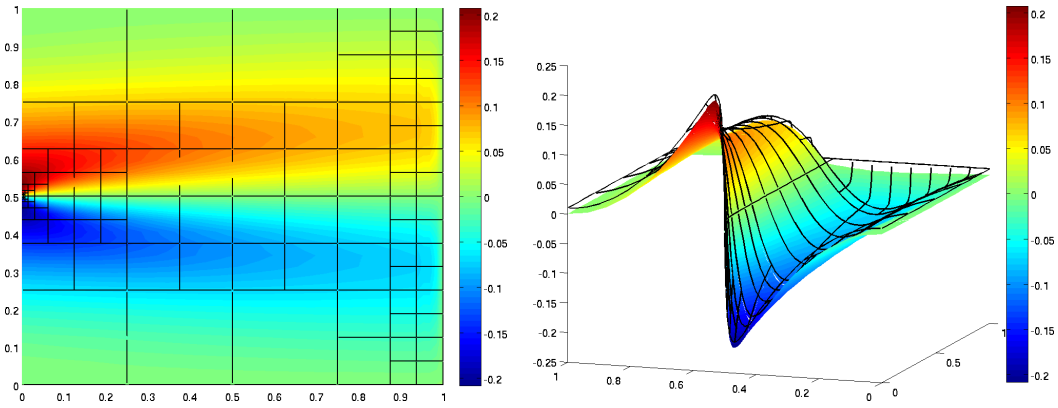


Figure 5.11: Solution variables u and $\widehat{u}$ with discontinuous inflow data u_0 for $\epsilon = .01$.

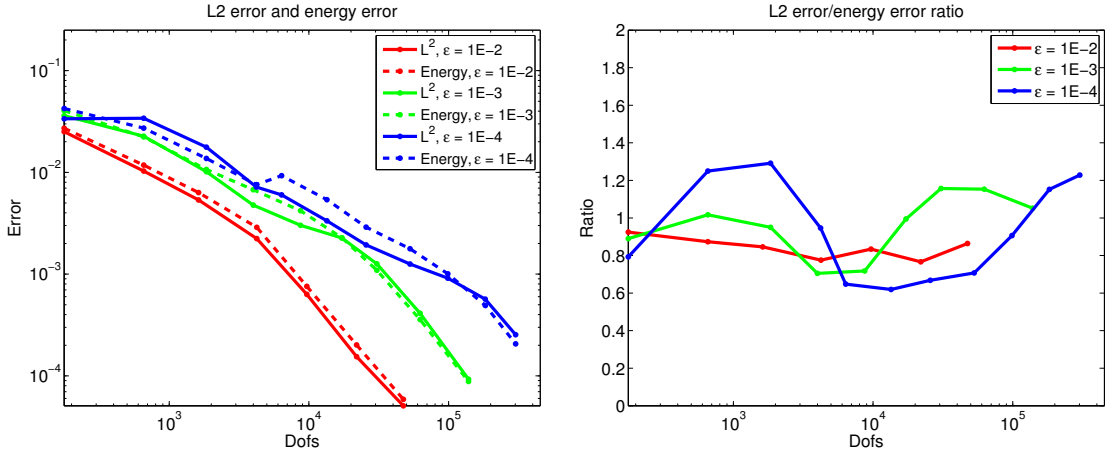


Figure 5.12: L^2 and energy errors, and their ratio for $\epsilon = .01$, $\epsilon = .001$, $\epsilon = .0001$, with discontinuous u_0 approximated by a Fourier expansion.

ratios still remain bounded as predicted by theory, and similarly to the previous problem with a smooth inflow condition, the ratios are close to 1 for ϵ varying over two orders of magnitude.

5.3 A coupled, robust test norm

In the course of our numerical experiments, we encountered unforeseen difficulties for certain problems under our current robust test norm. We illustrate in this section the observed issue using a second model problem with a singular solution, offer possible explanations for the phenomena observed, and propose a modification of the robust test norm presented previously, which we demonstrate eliminates the issues observed in numerical experiments.

5.3.1 A second model problem

We begin by first examining a different problem than convection-diffusion – we examine admissible solutions for the homogeneous Laplace's equation over the $y > 0$ half-plane under boundary conditions

$$u = 0 \text{ on } x > 0$$
$$\frac{\partial u}{\partial n} = 0 \text{ on } x < 0.$$

Let us consider the 2D case - a simple separation of variables argument in polar coordinates shows that the solution is of the form

$$u(r, \theta) = \sum_{n=0}^{\infty} R_n(r) \sin(\lambda_n \theta),$$

where $\lambda_n = n + \frac{1}{2}$, and $R_n(r) = C_{1,n} x^{\lambda_n} + C_{2,n} x^{-\lambda_n}$. By requiring $u(0, \theta) < \infty$, we have $R_n(r) = C_n x^{\lambda_n}$. We have now that solutions to this problem include u of the form

$$u = \sum_{n=0} x^{n+\frac{1}{2}} \sin\left(\left(n + \frac{1}{2}\right)\theta\right).$$

Note that, for the lowest-order term, the gradient of u displays a singularity at $r = 0$. It is well known that, for smooth boundary data, solutions to Laplace's equation can be decomposed into

the linear combination of smooth and singular contributions; the above analysis implies that, when boundary conditions change from Dirichlet to Neumann on the half-plane, the Laplace's equation will always develop a singularity in the stresses.

Consider now Laplace's equation $\triangle u = f$ on the box domain $\Omega = [0,1]^2$ with boundary conditions

$$u = 0 \text{ on } x > .5$$
$$\frac{\partial u}{\partial n} = 0 \text{ on } x < .5.$$

with forcing term $f = 1$. Extrapolating the results from the half-plane example to a finite domain,

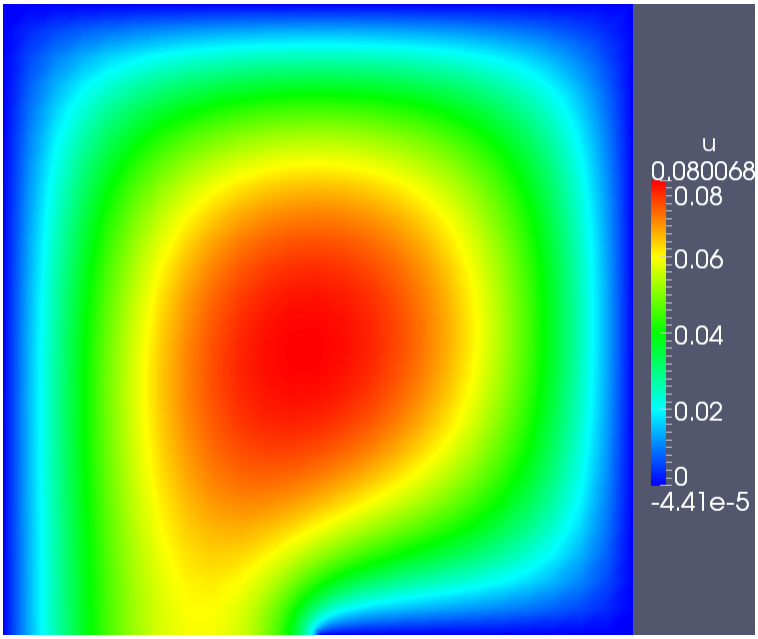


Figure 5.13: Solution of Laplace's equation on the unit quad with $f = 1$.

we expect the solution of Laplace's equation to be bounded, but to have a singularity in its gradient. Figures 5.13 and 5.14 are finite element solutions of the above problem under a quadratic h-refined

mesh. Figure 5.13 confirms that u is bounded, while Figure 5.14 confirms that singularities in the gradient appear at the point $(.5, 0)$, where the boundary condition changes from Neumann to Dirichlet.

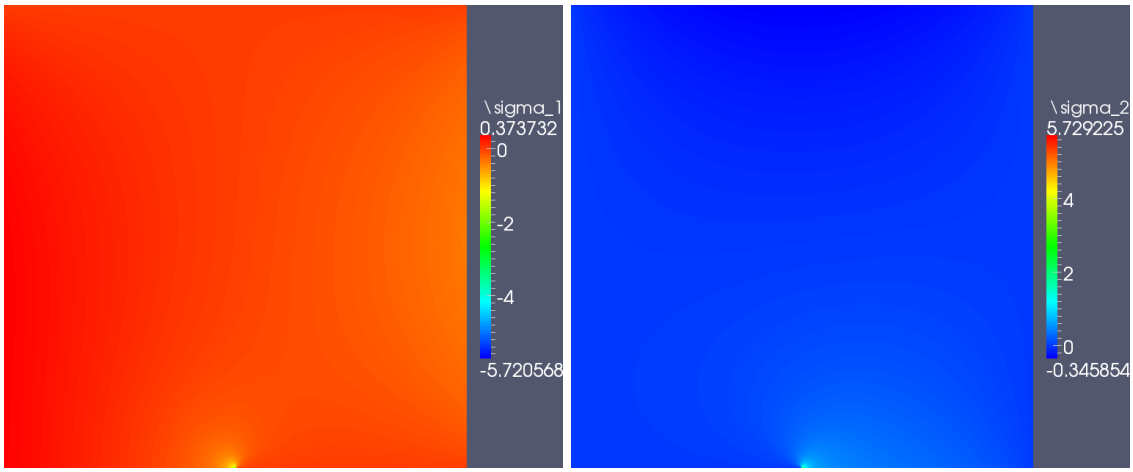


Figure 5.14: x and y components of the ∇u for u solving Laplace's equation with a change in boundary conditions. Both components develop singularities at the point where the boundary condition changes type.

We consider now the convection-diffusion problem, under a similar setup as before. We consider the domain $\Omega = [0, 1]^2$, with boundary conditions

$$\begin{aligned} u &= 0, \quad && \text{on } x = 0 \\ \frac{\partial u}{\partial n} &= 0, \quad && \text{on } x = 1, y = 1, \text{ and } y = 0, x < .5 \\ u &= 1, \quad && \text{on } .5 < x \leq 1. \end{aligned}$$

The problem is meant to simulate the transport of u over a domain with a "plate" boundary $x \in [.5, 1]$. For small ϵ, the problem develops a boundary layer over the plate, as well as a singularity at the plate tip $(x, y) = (.5, 0)$.[10] Unlike the Laplace example, we swap the Dirichlet boundary

[10]This problem is meant to mimic the Carter flat plate problem – a common early benchmark problem in viscous

condition at the outflow $x = 1$ with an outflow boundary condition.[11]

The above convection-diffusion problem is related back to the earlier Laplace/diffusion problem with a singularity – in most of the domain, convective effects dominate; however, if we localize the behavior of Laplace's equation to a circle of ϵ around $(.5, 0)$, then we again see a discontinuity in the stresses. Asymptotic expansion techniques indicate that singularities in solutions are determined primarily by the highest order differential operator present in the equation – in other words, the addition of a convective term to a scaled Laplacian (to recover the convection-diffusion equation) will not alter the presence of a singularity in the solution.

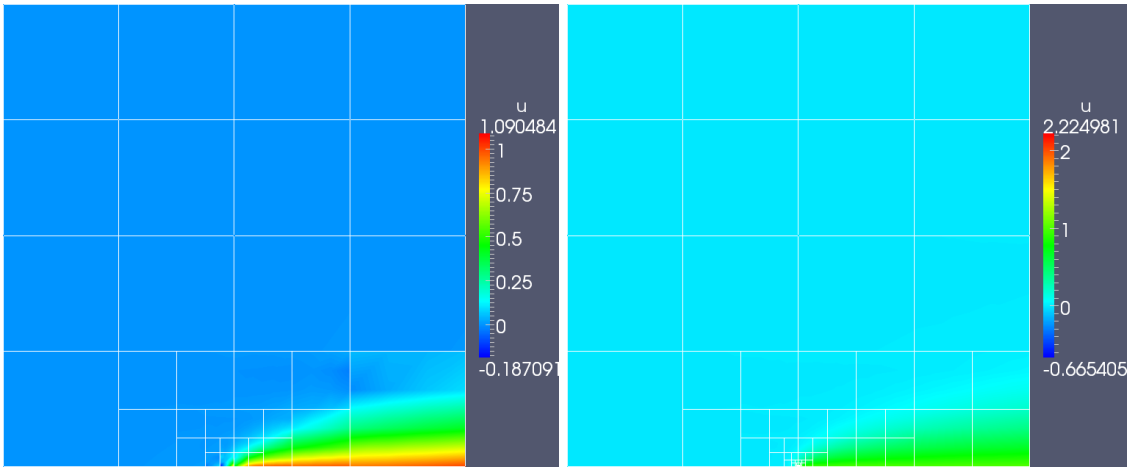


Figure 5.15: Solution u for $\epsilon = .01$ under the robust test norm. The solution oscillates strongly at the plate edge, growing in magnitude under additional refinements despite the absence of a singularity in u at that point.

Figures 5.15 and 5.16 demonstrate the behavior of the DPG method under the robust test

compressible flow problems – which can be shown to also exhibit a singularity in stress at the point $(.5, 0)$.

[11]The outflow "boundary condition" is simply the absence of an applied boundary condition, and is analyzed in more detail in [51]. This outflow condition appears to work well for convection-diffusion problems in the convective regime, and is the outflow condition we will use in our extension of DPG to a model problem in viscous compressible flow. Though the well-posedness of the problem under this boundary condition is questionable, we can still effectively illustrate the issues present under the robust test norm using this problem setup.

norm for the plate problem. The diffusion is taken to be fairly large ($\epsilon = 10^{-2}$), and automatic refinements are done until the element size h is at or below the diffusion scale. Due to the singular nature of the solution, refinements are clustered around $(.5, 0)$, and the order is set to be uniform with $p = 2$. While there should be no singularity in u, the magnitude of u grows as $h \to 0$, so long as $h \leq \epsilon$.

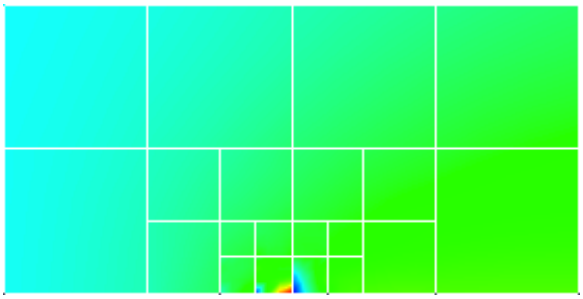


Figure 5.16: Zoomed solution u and adaptive mesh for $\epsilon = .01$ after over-resolution of the diffusion scale.

We note that the appearance of this non-physical singularity in u is allowed under the theory underlying the robust test norm; the error in the $L^2(\Omega)$-norm of the solution is guaranteed to be robustly bounded; however, the $L^2(\Omega)$ norm does allow for the presence of weak singularities (singularities of order $x^{-\frac{1}{2}}$). Apart from the oscillation of u at the singular point, the solution is well-behaved, and the stress $\sigma = \epsilon \nabla u$ is very well represented, as indicated in Figure 5.17.

5.3.2 A modification of the robust test norm

While oscillations of this sort in a solution near a singular point may be acceptable in certain simulations, it is a large problem for the methods in compressible flow simulations – physical constraints require several solution variables to remain positive throughout simulation.[12] We propose

[12] Apart from returning a non-physical solution, the violation of positivity constraints typically results in non-convergence of nonlinear solvers.

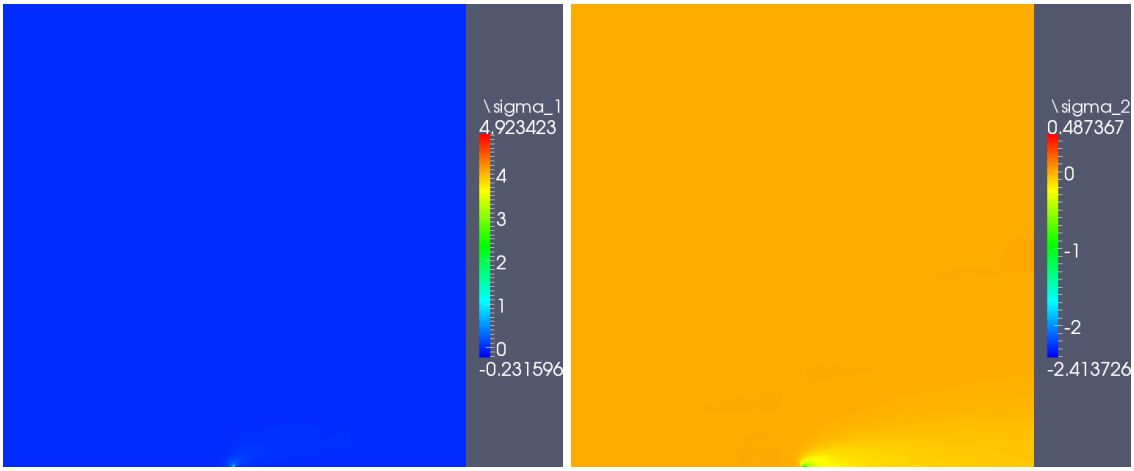


Figure 5.17: Viscous stresses for the plate problem.

a modification of the robust test norm that appears to remedy this issue, which we refer to as

$$\|(v,\tau)\|_V^2 \coloneqq \|v\|_{L^2(\Omega)}^2 + \epsilon\,\|\nabla v\|_{L^2(\Omega)}^2 + \|\beta\cdot\nabla v\|_{L^2(\Omega)}^2 + \|\nabla\cdot\tau - \beta\cdot\nabla v\|_{L^2(\Omega)}^2 + \|C_\tau\tau\|_{L^2(\Omega)}^2\,,$$

where C_τ is defined as before.[13] We note that, under the theory developed in Section 5.1.3, the above test norm is trivially provably robust using the same theory.[14]

While not rigorously understood, the author believes the issues related to the appearance of non-physical singularities to be related to the uncoupled nature of the test norm. Previous example problems exhibited boundary layers and sharp gradients in the stress σ, but not singularities, which contribute significantly more error. We expect that the oscillations observed in u are a sort of *pollution error*, where error in u is tied to error in σ. If we consider the ultra-weak variational

[13] We note that we have dropped the mesh-dependent scaling on $\|v\|_{L^2(\Omega)}$ from the robust norm; this is related to recent insights into the nature of DPG test spaces and explained in more detail in Appendix 2.1.

[14] This is due to the fact that $\|\nabla\cdot\tau - \beta\cdot\nabla v\|_{L^2(\Omega)}^2$ is robustly bounded by $\|\nabla\cdot\tau\|_{L^2(\Omega)}$ and $\|\beta\cdot\nabla v\|_{L^2(\Omega)}$. Alternatively, we can note that $\|\nabla\cdot\tau - \beta\cdot\nabla v\|_{L^2(\Omega)}^2 = \|g\|_{L^2(\Omega)}^2$, where g is a load of the adjoint problem related to robustness described in Section 5.1.3.

formulation for convection-diffusion

$$(u, \nabla \cdot \tau - \beta \cdot \nabla v)_{L^2(\Omega)} + \left(\sigma, \frac{1}{\epsilon}\tau + \nabla v\right)_{L^2(\Omega)} + \ldots,$$

we can see that it is a combination of test functions that corresponds to both u and σ. Recall from the previous section that, by choosing τ and v such that they satisfy the adjoint equation with forcing terms u and σ, we recover the best L^2 approximation. In other words, achieving optimality in the L^2 norm requires coupling between v and τ, which is achieved under the graph norm, but not the robust norm derived in the previous section. If coupling of the test terms delivers optimality in u and σ independently, we expect that decoupling v and τ from each other in a test norm from each other will have the effect of coupling error in σ to error in u, which would explain the spurious oscillations in u in the presence of singularities in σ.[15]

The drawback to using the above test norm is that the resulting local system for test functions is now completely coupled, whereas using the previous test norm, the system was block diagonal due to the decoupling in v and τ and could be constructed and inverted more efficiently. We hope to explore the difference between these two norms in more rigor and detail in the future; however, experiments in [53] indicate that this new norm performs equivalently or better than the robust test norm derived in the previous chapter for a large range of numerical examples.

Figure 5.18 shows the solution for $\epsilon = .01$, where the diffusion scale is both underresolved and resolved by h-adaptivity. In both cases, there are no additional oscillations near the plate tip – Figure 5.19 shows a zoomed image of the solution u at the point $(.5, 0)$. The stress is resolved similarly to the previous case; however, the solution u does not display spurious oscillations in either the underresolved or resolved cases. Figure 5.20 displays the same quantities, but for $\epsilon = 10^{-4}$, in

[15] Similar results have been observed in the Stokes equations, where error in u is coupled to the behavior of the pressure variable [52].

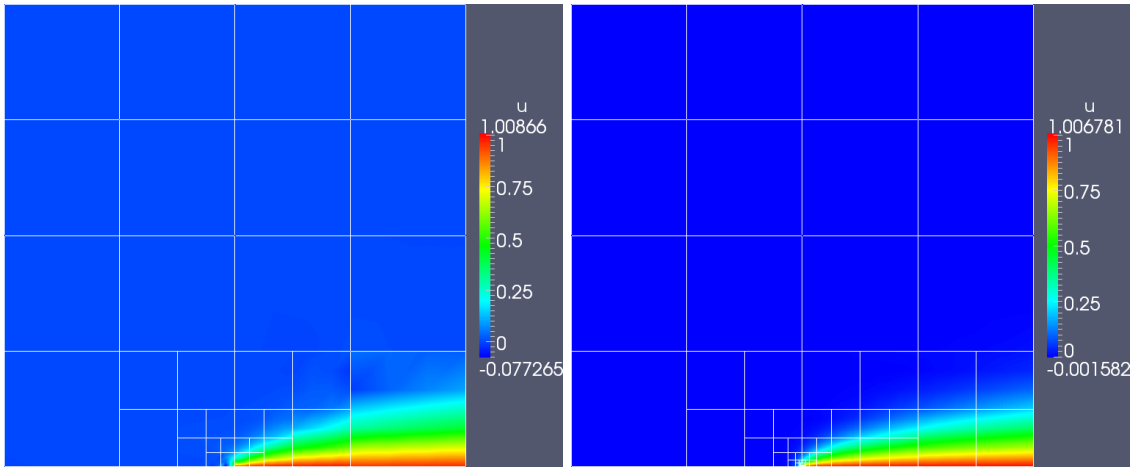


Figure 5.18: $\epsilon = 10^{-2}$ without h-resolving diffusion scale, and with h-resolution of diffusion scale.

order to demonstrate that the new test norm removes spurious oscillations in u (in the presence of singularities in σ) independently of ϵ.

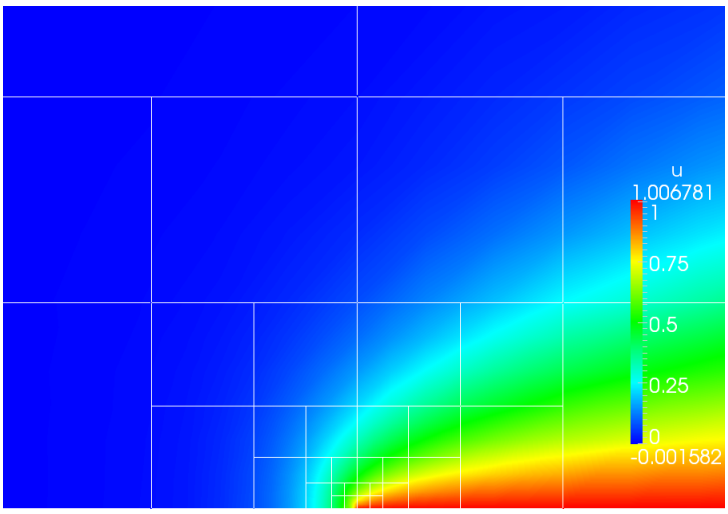


Figure 5.19: Zoom of solution u at the plate tip for $\epsilon = 10^{-2}$.

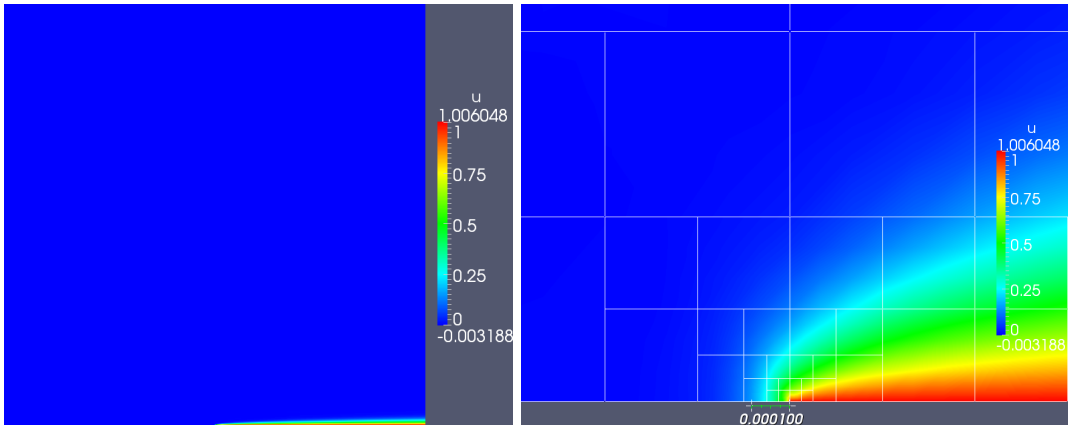


Figure 5.20: 14 refinements for $\epsilon = 10^{-4}$, min h is $O(10^{-5})$.

5.4 Anisotropic refinement

Isotropic adaptive mesh refinement has shown itself to be an effective way to resolve isolated solution features with large gradients, such as point singularities [54, 55]. However, for the resolution of phenomena such as shocks or boundary layers, anisotropic mesh refinement can resolve solution features for a much lower cost per degree-of-freedom, due to the fact that boundary layers in n-dimensions are primarily phenomena supported over $n-1$ dimensions.

As a least squares method, DPG already includes a natural error indicator with which to drive adaptive mesh refinement. To introduce anisotropic refinements, we need to introduce an anisotropy indicator in order to detect in which direction solution features are aligned. In general, a test norm can be expressed as the sum of normed quantities, both scalar and vector valued. If we restrict ourselves to quadrilateral elements for the moment, a general anisotropy indicator for DPG can be constructed by evaluating the $L^2(\Omega)$ norms of the individual components of vector valued terms in the test norm.

Under the robust test norm derived in this chapter for the convection-dominated diffusion

problem, we can define x and y error contributions over a single element

$$e_{x,K} = \epsilon \left\| \frac{\partial v}{\partial x} \right\|^2_{L^2(K)} + \|\tau_x\|^2_{L^2(K)}$$
$$e_{y,K} = \epsilon \left\| \frac{\partial v}{\partial y} \right\|^2_{L^2(K)} + \|\tau_y\|^2_{L^2(K)} .$$

We define the anisotropy indicator as the ratio $r_K = \frac{e_{x,K}}{e_{y,K}}$, and implement a simple refinement scheme following [7]. Given some anisotropic threshold ϵ_r, if $r_K > \epsilon_r$, then we can conclude that the error in the x direction is large compared to the y direction, and we anisotropically refine the element along the x-axis. Likewise, if $r_K < \frac{1}{\epsilon_r}$, this implies that the opposite is true, and we refine the element anisotropically along the y-axis.

We note that it is possible to compute the discrete system without needing much additional integration. Recall that if we let G be the symmetric positive-definite Gram matrix representing the inner product $(v, \delta v)_V$ on V_h, we solve for degrees of freedom c_e representing our error representation function e.

For both the graph and robust test norms, we can decompose the inner product that induces the test norm into

$$(v, \delta v)_V = \sum_i (v, \delta v)_{V,x_i} + (v, \delta v)_{V,\text{scalar}}$$

such that $(v, \delta v)_{V,x_i}$ is a seminorm containing the ith coordinate component of a vector-valued test term, and $(v, \delta v)_{V,\text{scalar}}$ is simply the non-vector portions of the test norm. For example, if we take the $H^1(\Omega)$ Sobolev norm

$$(v, \delta v)_V = (v, \delta v)_{L^2(\Omega)} + (\nabla v, \nabla \delta v)_{L^2(\Omega)}$$

then $(v, \delta v)_{V,x_i} = \left(\frac{\partial v}{\partial x_i}, \frac{\partial \delta v}{\partial x_i} \right)_{L^2(\Omega)}$, and $(v, \delta v)_{V,\text{scalar}} = (v, \delta v)_{L^2(\Omega)}$. Each bilinear term $(v, \delta v)_{V,x_i}$

induces a symmetric positive-semidefinite matrix G_{x_i}, such that

$$c_e^T G_{x_i} c_e = \|e\|_{V,x_i}^2 .$$

By storing G as the sum of G_{scalar} and G_{x_i}, we can then cheaply compute the anisotropic error indicators once we have the degrees of freedom corresponding to our error representation function.

Chapter 6

Extension to nonlinear problems and systems of equations

6.1 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:

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

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.

6.1.1 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 $u(t_i)$. For most applications, it is sufficient to approximate the nonlinear solution using a single Newton solve at each time step.

to initial guess[56, 57]. 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 [58]. 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.

6.1.2 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.

We note that, under convergence of any of the nonlinear iterations described above, the converged DPG solution does in fact satisfy the above minimization problem [58]. However, it is theoretically possible to accelerate convergence of a nonlinear iteration to the minimizer u_h, using the same minimum residual framework DPG is based upon.

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

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

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 future work.

6.1.3 DPG as a Gauss-Newton approximation

The above Hessian-based nonlinear DPG method essentially replaces a simple linearization of with a higher-order expansion of the residual $J(u_h)$. However, the application of DPG to the linearized equations actually yields a Gauss-Newton method for minimizing $J(u_h)$.

Recall that the discrete DPG method applied to a linear problem solves

$$B^T {R_V}^{-1} B u - B^T {R_V}^{-1} l = 0$$

where B is the rectangular matrix of size $m \times n$ resulting from the variational form $b(u,v)$ (where $m = \dim(U_h)$ and $n = \dim(V_h)$), and R_V is the Gram matrix of dimension $n \times n$ resulting from choice of norm $\|\cdot\|_V$ and inner product on V. If we have the nonlinear equation $B(u_h) - l$ and linearize it, we have $B_u \triangle u_h = r(u_h)$, where $r(u_h) = l - B(u_h)$ is the residual. The solution of the linearized equation by DPG minimizes then $\|r(u_h) - B_u \triangle u_h\|^2_{V'}$, which we can recognize as linear least squares equations from which the Gauss-Newton method is derived [59].

Initial experiments with Hessian-based nonlinear DPG under Burgers' equation indicate that eigenvalues of the stiffness matrix can become negative if the Hessian is applied. While the Gauss-Newton iteration may not converge as rapidly as the Hessian-based version of DPG, it does guarantee positive eigenvalues of the stiffness matrix, such that Δu_h is always a descent direction.

6.2 A viscous Burgers equation

We will illustrate the application of DPG to nonlinear problems using a viscous Burgers' equation on domain $\Omega = [0,1]^2 \in \mathbb{R}^2$ [58]

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

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 5.1.3 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 6.1 for $\epsilon=1e-4$.

For the solution of Burgers' equation, adaptivity began on a cubic 4×4 mesh. 9 iterations of adaptive mesh refinement were performed, and the resulting mesh and zoomed solution are displayed in Figure 6.2, demonstrating a fully resolved smooth transition of the solution into a smeared shock.

6.3 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.

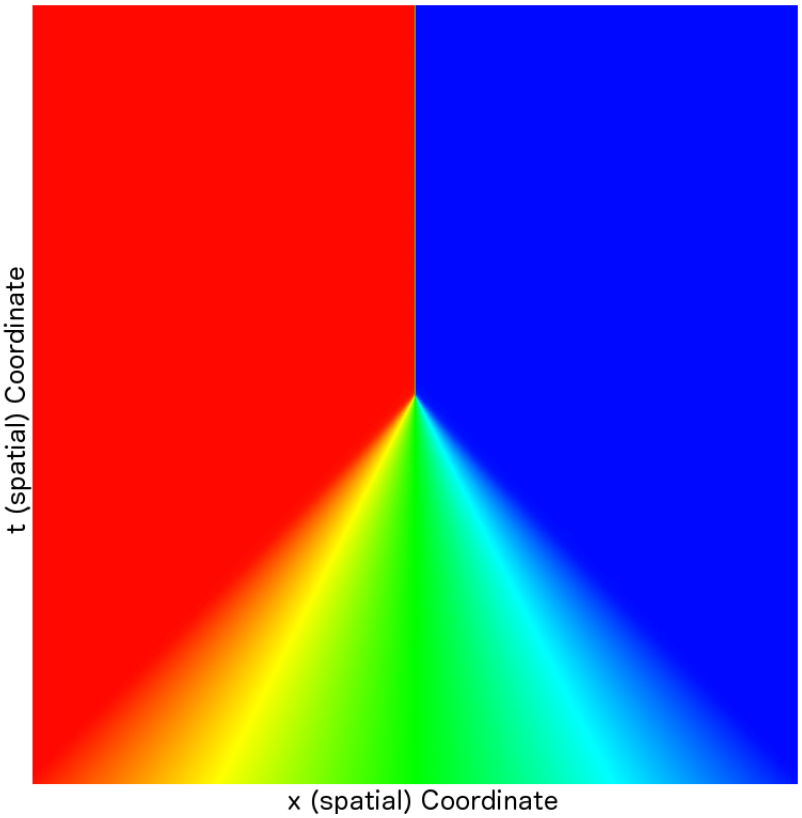


Figure 6.1: Shock solution for Burgers' equation with $\epsilon = 10^{-4}$.

- **Conservation equations**

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

  where $\boldsymbol{\sigma}$ is the stress tensor whose ijth term is σ_{ij}, and $\boldsymbol{u}$ is the vector $(u_1, u_2)^T$.

- **Newtonian fluid laws**

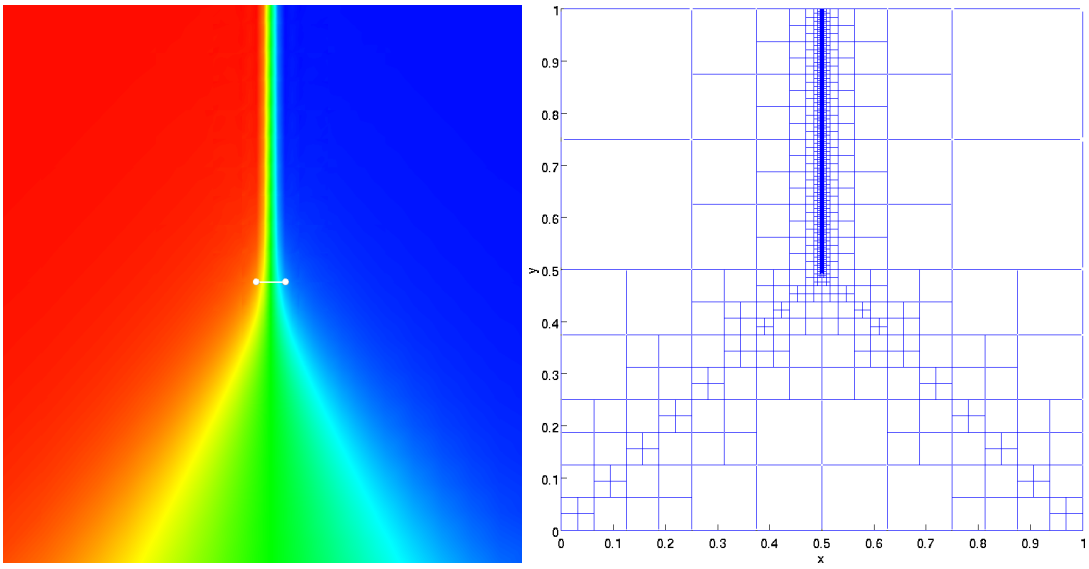


Figure 6.2: Adaptive mesh after 9 refinements, and zoom view at point (.5,.5) with shock formation and $1e-3$ width line for reference.

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

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

We also have

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

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

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

Thus our final form is

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

Notice that $\boldsymbol{\omega}$ is implicitly defined to be the antisymmetric 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 [3, 4].

- **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.

6.3.1 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

$$\begin{aligned}
\nabla \cdot \begin{bmatrix} \rho u_1 \\ \rho u_2 \end{bmatrix} &= 0 \\
\nabla \cdot \left(\begin{bmatrix} \rho u_1^2 + p \\ \rho u_1 u_2 \end{bmatrix} - \boldsymbol{\sigma}_1 \right) &= 0 \\
\nabla \cdot \left(\begin{bmatrix} \rho u_1 u_2 \\ \rho u_2^2 + p \end{bmatrix} - \boldsymbol{\sigma}_2 \right) &= 0 \\
\nabla \cdot \left(\begin{bmatrix} ((\rho e) + p) u_1 \\ ((\rho e) + p) u_2 \end{bmatrix} - \boldsymbol{\sigma}\mathbf{u} - \vec{q} \right) &= 0 \\
\frac{1}{2\mu}\boldsymbol{\sigma} - \frac{\lambda}{4\mu(\mu+\lambda)}\mathrm{tr}(\boldsymbol{\sigma})\boldsymbol{I} &= \nabla \mathbf{u} - \mathrm{Re}\,\boldsymbol{\omega} \\
\frac{1}{\kappa}\vec{q} &= \nabla T
\end{aligned}$$

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.

6.3.2 Linearization

As the equations for viscous compressible flow are nonlinear and cannot be solved exactly, we linearize the equations and adopt an iterative procedure for approximating the nonlinear solution.[3] We outline the linearization of both the conservation and stress laws in this section.

[3]We note that it is possible to linearize the strong form of the equations and then derive a linearized weak form, instead of linearizing the weak form of the nonlinear equations, which is done here. The two formulations are equivalent; however, extraneous linearization of the fluxes is avoided using the latter approach.

6.3.2.1 Conservation laws

The Navier-Stokes conservation laws can be written as

$$\begin{aligned}
\nabla \cdot \left[\begin{array}{c} \rho u_1 \\ \rho u_2 \end{array} \right] &= 0 \\
\nabla \cdot \left(\left[\begin{array}{c} \rho u_1^2 + p \\ \rho u_1 u_2 \end{array} \right] - \boldsymbol{\sigma}_1 \right) &= 0 \\
\nabla \cdot \left(\left[\begin{array}{c} \rho u_1 u_2 \\ \rho u_2^2 + p \end{array} \right] - \boldsymbol{\sigma}_2 \right) &= 0 \\
\nabla \cdot \left(\left[\begin{array}{c} ((\rho e) + p) u_1 \\ ((\rho e) + p) u_2 \end{array} \right] - \boldsymbol{\sigma}_1 \cdot \boldsymbol{u} - \boldsymbol{\sigma}_2 \cdot \boldsymbol{u} - \vec{q} \right) &= 0,
\end{aligned}$$

where $\boldsymbol{\sigma}_1$ is the ith column or row of $\boldsymbol{\sigma}$, or more generally, if we group our Eulerian and stress variables into the vector variables $\boldsymbol{U}$ and $\boldsymbol{\Sigma}$, respectively

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

where $F_i$and G_i are given as

$$\begin{aligned}
F_1 &= \left[\begin{array}{c} \rho u_1 \\ \rho u_2 \end{array} \right], \quad G_1 = 0 \\
F_2 &= \left[\begin{array}{c} \rho u_1^2 + p \\ \rho u_1 u_2 \end{array} \right], \quad G_2 = \boldsymbol{\sigma}_1 \\
F_3 &= \left[\begin{array}{c} \rho u_1 u_2 \\ \rho u_2^2 + p \end{array} \right], \quad G_3 = \boldsymbol{\sigma}_2 \\
F_4 &= \left[\begin{array}{c} ((\rho e) + p) u_1 \\ ((\rho e) + p) u_2 \end{array} \right], \quad G_4 = \boldsymbol{\sigma}_1 \cdot \boldsymbol{u} + \boldsymbol{\sigma}_2 \cdot \boldsymbol{u} + \vec{q}
\end{aligned}$$

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 functions

of the unknowns, 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.

6.3.2.2 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:

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

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.

6.3.3 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 5.1.3 and in [4] is defined elementwise on K as

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

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.

This test norm is extrapolated to the Navier-Stokes equations as follows: we denote the vector of H^1 test functions as $\boldsymbol{V} = \{v_1, v_2, v_3, v_4\}$, and similarly for $\boldsymbol{W} = \{\tau_1, \tau_2, \tau_3\}$. If $R_{\mathrm{Euler}}(\boldsymbol{U}, \boldsymbol{\Sigma})$ and $R_{\mathrm{visc}}(\boldsymbol{U}, \boldsymbol{\Sigma})$ are Eulerian and viscous nonlinear residuals, our formulation for the linearized Navier-Stokes equations can be written as

$$\nabla \cdot (A_{\mathrm{Euler}}\boldsymbol{\delta U} - A_{\mathrm{visc}}\boldsymbol{\delta \Sigma}) = R_{\mathrm{Euler}}(\boldsymbol{U}, \boldsymbol{\Sigma})$$
$$E_{\mathrm{visc}}\boldsymbol{\delta \Sigma} - \nabla \boldsymbol{\delta U} = R_{\mathrm{visc}}(\boldsymbol{U}, \boldsymbol{\Sigma})$$

with variational formulation

$$\left\langle \widehat{F}_n, \boldsymbol{V} \right\rangle_{\Gamma_h} + \left(\boldsymbol{\delta U}, \nabla \cdot \boldsymbol{W} - A^T_{\mathrm{Euler}} \nabla \boldsymbol{V}\right) + \left\langle \widehat{\boldsymbol{U}}, \boldsymbol{W} \cdot n \right\rangle_{\Gamma_h} + \left(\boldsymbol{\delta \Sigma}, E^T_{\mathrm{visc}}\boldsymbol{W} - A^T_{\mathrm{visc}} \nabla \boldsymbol{V}\right) =$$
$$\langle R_{\mathrm{Euler}}(\boldsymbol{U}, \boldsymbol{\Sigma}), \boldsymbol{V}\rangle + \langle R_{\mathrm{visc}}(\boldsymbol{U}, \boldsymbol{\Sigma}), \boldsymbol{W}\rangle$$

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

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

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, the test function for the conservation equation.[4]

6.3.4 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 $\widehat{T}$.

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 [3, 4].

[4]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.

6.4 Nonlinear solver

For our nonlinear solver, we use a pseudo-time stepping approach to iterate to a steady state solution, along with a greedy refinement scheme to eliminate spatial discretization error. We cover briefly the details of the pseudo-timestepping method in this section.

6.4.1 Pseudo-timestepping

The solution of the compressible Navier-Stokes equations can be quite challenging; as mentioned previously, the direct application of a Newton algorithm often will not converge, especially for high Reynolds numbers. Typically, a pseudo-timestepping algorithm is used in lieu of a full nonlinear Newton algorithm. The pseudo-timestep proceeds as follows: given the transient terms present in the conservation laws of compressible flow

$$\begin{aligned}
\frac{\partial \rho}{\partial t} + \dots \\
\frac{\partial (\rho u)}{\partial t} + \dots \\
\frac{\partial (\rho v)}{\partial t} + \dots \\
\frac{\partial (\rho e)}{\partial t} + \dots,
\end{aligned}$$

we discretize each time derivative using an implicit timestepping method. Though second order time discretizations have been shown to be effective [60, 61], we choose a first order backwards Euler discretization for simplicity. Due to the fact that we've chosen to solve the Navier-Stokes equations under primitive variables ρ, u, v, and T, the time terms are nonlinear as well, and must be linearized. After time discretization and linearization, we are left with a coupled system of reaction terms in

the problem for our solution update at every timestep

$$A_{\text{time}}\boldsymbol{\delta}U_i + \nabla\cdot(A_{\text{Euler}}\boldsymbol{\delta}U_i - A_{\text{visc}}\boldsymbol{\delta}\Sigma_i) = R_{\text{Euler}}(\boldsymbol{U}_i,\boldsymbol{\Sigma}_i) + R_{\text{time}}(\boldsymbol{U}_{i-1},\boldsymbol{U}_i)$$

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

where $R_{\text{time}}(\boldsymbol{U}_{i-1},\boldsymbol{U}_i)$ is the transient residual. Including the transient terms in our test norm as well, our final test norm is

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

Finally, convergence of the pseudo-timestepping method is determined by measuring the transient residual in the energy norm; in other words, we measure $\|e_{\text{time}}\|_V$, where

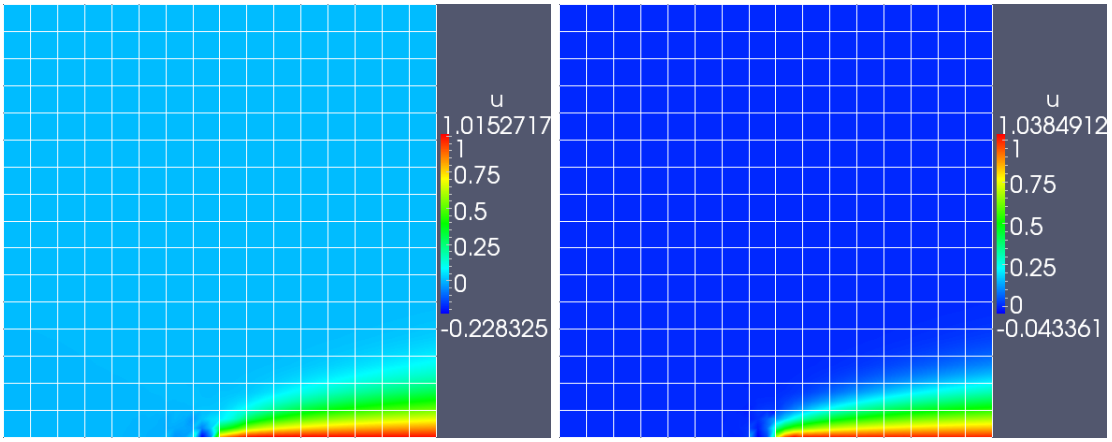
$$((\boldsymbol{V},\boldsymbol{W}),(\boldsymbol{dV},\boldsymbol{dW}))_V = (R_{\text{time}}(\boldsymbol{U}_{i-1},\boldsymbol{U}_i),\boldsymbol{V})_{L^2(\Omega)}, \quad \forall(\boldsymbol{dV},\boldsymbol{dW})\in V.$$

The energy norm thus provides a consistent measure in which convergence of the nonlinear iteration at each timestep, convergence of the pseudo-timestepping algorithm to steady state, and nonlinear residual can all be assessed.

6.4.1.1 Dependence of solution on dt

A surprising feature of pseudo-timestepping schemes for DPG is that, under the problem-dependent minimum-residual nature of the method, convergence to steady state can yield qualitatively slightly different solutions under different size timesteps. We illustrate this using the plate example for convection-diffusion, which we described first in Section 5.3.1. We consider the transient form of the conservation equation for convection-diffusion

$$\frac{\partial u}{\partial t} + \nabla\cdot(\beta u - \sigma) = f,$$



(a) $dt = .1$ (b) $dt = .01$

Figure 6.3: Comparison of pseudo-timestepping to steady state for a convection-diffusion problem under two different sizes of timestep.

where the stress law remains unchanged. Applying a backwards Euler time discretization, at each timestep, we will solve for the current solution u_i (as well as the current stress σ_i) given the previous timestep solution u_{i-1} under

$$\frac{u_i}{dt} + \nabla \cdot (\beta u_i - \sigma_i) = f + \frac{u_{i-1}}{dt}.$$

The conserved flux $\beta_n u - \sigma_n$ is set to freestream values on the inflow, and 0 on the top boundary $y = 1$ and half of the bottom boundary $x < .5, y = 0$. We set $u = 1$ for the boundary $x \in (.5, 1), y = 0$. For this example, $\beta = (1, 0)$ and $\epsilon = 10^{-3}$. This can be understood as the consequence of solving a "moving target" optimization problem – our variational formulation and test norm for this problem are

$$\left(u_i, \frac{1}{dt} v\right)_{L^2(\Omega)} + b\left(u_i, v\right) = l(v) + \left(u_{i-1}, \frac{1}{dt} v\right)_{L^2(\Omega)}$$
$$\left\|(\tau, v)\right\|_{V,dt}^2 = \frac{1}{dt} \left\|v\right\|_{L^2(\Omega)}^2 + \left\|v\right\|_V^2,$$

where $\|(\tau, v)\|_V$ is the coupled test norm introduced in Section 5.1.3, and $b(u, v)$ and $l(v)$ are the bilinear form and load for the steady state form of the convection diffusion equation. DPG solutions minimize the functional

$$J(u_i) = \sup_{v \in V} \frac{\left(u_i - u_{i-1}, \frac{1}{dt} v\right)_{L^2(\Omega)} + b(u_i, v) - l(v)}{\left(\frac{1}{dt} \|v\|^2_{L^2(\Omega)} + \|v\|^2_V\right)^{\frac{1}{2}}}$$

over a given mesh. As $u_{i-1} \to u_i$, which we expect to happen as the pseudo-timestepping algorithm converges to steady state, the minimized functional becomes

$$J(u_i) = \sup_{v \in V} \frac{b(u_i, v) - l(v)}{\left(\frac{1}{dt} \|v\|^2_{L^2(\Omega)} + \|v\|^2_V\right)^{\frac{1}{2}}}.$$

While the transient portion of the residual disappears, a factor of $\frac{1}{dt}$ is still present in the test norm.[5] Thus, we can expect that the nature of the steady state solution achieved through convergence of the pseudo-time algorithm can depend on the timestep dt. We observe the same phenomena for analogous problems in compressible flow as well.

6.4.1.2 Adaptive time thresholding

Adaptive timestepping (also known as pseudo-transient continuation) has been implemented successfully for problems in compressible flow [61]. Typical adaptive time-stepping schemes modify the time-step based on some notion of the transient residual R_i at timestep i, such that

$$dt_{i+k} = \left(\frac{R_{i+k}}{R_i}\right)^r dt_i,$$

[5] While the solution under smaller dt appears to give visually higher quality results, we stress that simply adding the term $\frac{1}{dt} \|v\|_{L^2(\Omega)}$ to the test norm under the steady state version of the convection-diffusion equations does not achieve the same effect. We are able to add this term without negative consequence due to the inclusion of the $\frac{u_i}{dt}$ term present in the variational formulation (see Chapter 5 and Appendix 2.2 for mathematical details). Numerical experiments indicate that including $\alpha \|v\|_{L^2(\Omega)}$ in the test norm, where $\alpha > \frac{1}{dt}$, converges much more slowly to a fine-mesh reference solution than if $\alpha = \frac{1}{dt}$.

where $r > 1$ dictates the rate of change of the timestep based on residual reduction, and k indicates an integer interval at which to modify the timestep. However, the minimum-residual nature of the DPG method and the "moving target" problem make the effectiveness of adaptive time-stepping schemes questionable.

For our current experiments, we implement instead an adaptive time thresholding, where we adaptive decrease our convergence criterion based on the spatial energy error. Recall that, under convergence of the pseudo-timestepping algorithm, the DPG energy error converges to the measure of the nonlinear residual in the dual norm. We set convergence criterion for the pseudo-timestepping algorithm to be such that

$$\|e_{\text{time}}\|_V < \max\{\epsilon_t, \epsilon_{t,k}\},$$

where $\epsilon_{t,k}$ is the tolerance at the kth refinement iteration, and $\epsilon_t < \epsilon_{t,k}$ is an absolute tolerance. We initialize $\epsilon_{t,k}$ to ϵ_t, then based on the energy error $\|u - u_h\|_E$, we set

$$\epsilon_{t,k} = \alpha_t \|u - u_h\|_E .$$

Since the linearized error at a single timestep is composed of a sum of the linearization error, transient residual, and nonlinear residual, if the transient residual and linearization error are small, we expect that the nonlinear residual at that point will be sufficient to be an effective error indicator with which to drive adaptive mesh refinement. In the following numerical experiments, α_t is set to .01.

The aim of this adaptive thresholding is to relax the convergence criteria for solution of the nonlinear system at each refinement step such that the same refinement pattern is achieved with or without the use of adaptive time thresholding. Numerical experiments seem to indicate that the same refinement pattern is produced with or without the implementation of this simple adaptive thresholding scheme, though wall-clock convergence times under adaptive thresholding are

faster. We hope to investigate both DPG-specific adaptive timestepping schemes and more advanced methods of balancing convergence criterion in the future.

6.4.2 Linear solver

A clear choice for a linear solver under the ultra-weak variational formulation is static condensation, or 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 $\dim A < \dim D$. This is due to the fact that, for the ultra-weak variational formulation (and for all HDG methods), the interior field degrees of freedom can be condensed out to yield a problem posed solely in terms of the coupled flux and trace degrees of freedom. 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 and trace degrees of freedom are solved for, the field degrees of freedom can be reconstructed locally. An additional advantage of the above approach is that the Schur complement maintains the same sparsity pattern implied by the connectivity of the globally coupled flux and trace degrees of freedom. Since the condensation process can be done locally, we can save memory by avoiding constructing the full stiffness matrix.

It has been shown that, unlike standard least-squares methods, DPG generates for the Poisson matrix a stiffness matrix with condition number $O(h^{-2})$ [35]. 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 iterative solvers in solving the condensed system. Initial experiments indicate that, at least for quasi-uniform and low-order meshes, both algebraic multigrid and preconditioned conjugate gradients are able to solve the condensed system fairly rapidly. We hope to experiment further with solvers for the condensed system, and to develop multigrid methods and preconditioners for adaptive and higher order meshes under DPG.

6.5 Test problems

We applied the DPG method to two test problems in compressible flow – flow over a flate plate, and flow over a compression ramp. While the physics of both problems are fairly simple, they nonetheless display several features (shocks, singularities, boundary layers) that are computationally difficult to resolve without adaptivity. Furthermore, the problems themselves are not usually solved without the aid of artificial or numerical diffusion and/or shock-capturing terms, which we eschew in our application of DPG to these model problems.

The numerical parameters used are as follows for:

- DPG parameters: $p = 2$ and $\Delta p = 2$ uniformly across the mesh.
- Adaptivity parameters: Energy threshold for refinements is $\alpha = .4$ for the Carter flat plate example and $\alpha = .5$ for the Holden ramp example.
- Time-stepping parameters: Initial timestep $\Delta t = .1$, and initial tolerance for transient residual $\epsilon_t = 1e - 7$.

Numerical experiments were run on a small cluster with 16 CPUs and 8GB memory, as well

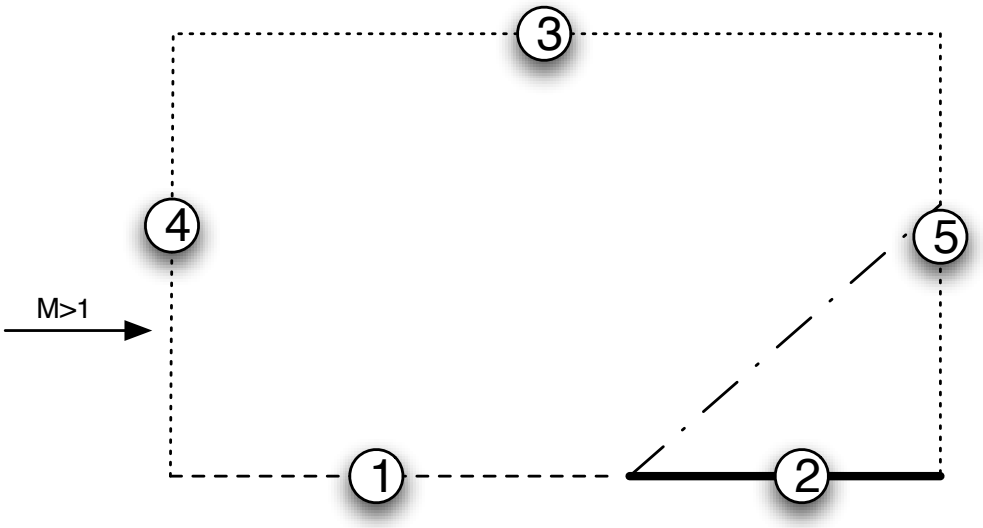


Figure 6.4: Carter flat plate problem.

as on the Lonestar machine at the Texas Advanced Computing Center (TACC), using a varying number of cores, from 48 to 120 depending on problem geometry and parameters. Elements were partitioned using the Zoltan library of Trilinos [62], using either a Hilbert space-filling curve or the REFTREE [63] algorithm.

6.5.1 Numerical experiments: Carter flat plate

The first problem of interest is the Carter flat plate problem[64]. An infinitesmally thin flat plate disrupts a free stream flow, causing a shock to form at the tip of the plate, and a boundary layer forms and widens along the length of the plate.

1. **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 $u_2 = 0$. They are

applied here to the bottom free-stream boundary.

2. **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 $\widehat{u}_1, \widehat{u}_2, \widehat{T}$.

3. **Symmetry boundary conditions** are applied also to the top free-stream boundary.

4. **Inflow boundary conditions:** free stream conditions are applied here to all four fluxes $\widehat{f}_{i,n}$.

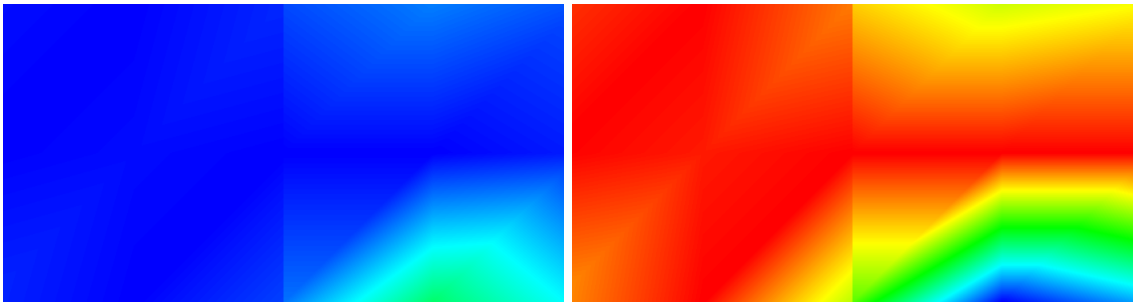
5. **Outflow boundary conditions:** the exact boundary conditions to enforce here are not universally agreed on. Many enforce $\frac{\partial u_1}{\partial n} = \frac{\partial u_2}{\partial n} = 0$ and $\frac{\partial T}{\partial n} = 0$, while others enforce an outflow boundary condition only in regions where the flow is subsonic.[65]. We adopt a "no boundary condition" outflow condition, first introduced in [66]. A mathematical analysis and explanation of this boundary condition for standard H^1 elements is given in [51].

We initialize our solution to

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

which, we also take as the freestream values for the above variables, and is consistent with what was done by Demkowicz, Oden, and Rachowicz in [60]. Stresses are set uniformly to zero. We take the computational domain to be $\Omega = [0, 2] \times [0, 1]$. 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.

We perform 10 steps of adaptive mesh refinement, beginning with a mesh of only two square elements. The solution on this coarsest mesh is given in Figure 6.5, and the final solutions after 10 refinement steps are given in Figure 6.6. For this specific example, we use only isotropic refinements.

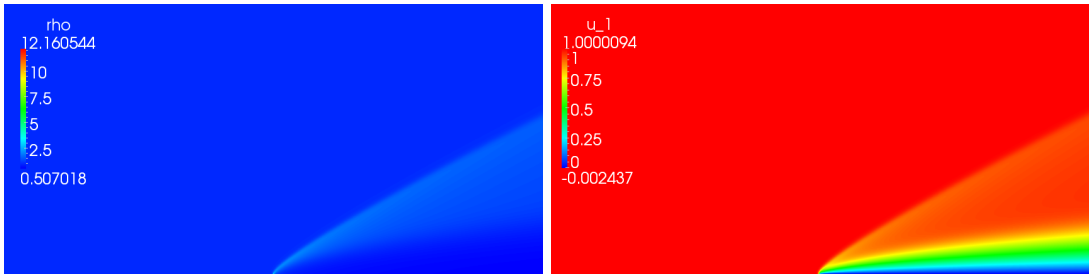


(a) T (b) u_1

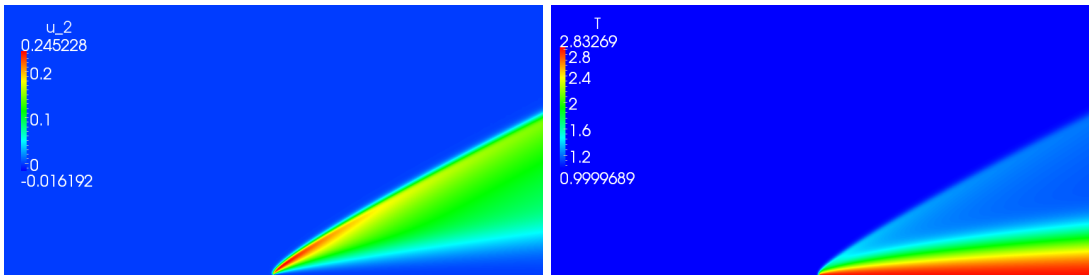
Figure 6.5: Converged solution on 2 cells.

Typical coarse meshes for adaptive CFD computations aim to resolve, at least to some degree, the features of the solution; often, physical features such as high gradients are used to drive refinement. For feature-based adaptivity to be effective, coarse mesh solutions must be of sufficiently high quality to resolve basic solution features. Often, artificial diffusion and shock capturing must also be applied in order to produce visually clean solutions on underresolved meshes. In contrast to this, the residual-based approach of DPG is able to place refinements accurately and efficiently despite the underresolution of solution features.

Figure 6.7 shows snapshots of the third and sixth steps of adaptive mesh refinement. The main contribution to energy error is at the plate tip – due to the change in boundary condition across the point $(.5, 0)$, the viscous stresses are singular at this point (this is analogous to the convection-diffusion plate example given in Section 5.3.1). Underresolution of the stresses at this point results in some pollution effects slightly upstream of the beginning of the plate; however, once $h \approx \mathrm{Re}^{-1}$ near the plate edge, these pollution effects disappear. We observe numerically that decreasing dt also limits how far upstream of the plate edge this pollution effect travels.



(a) ρ (b) u_1



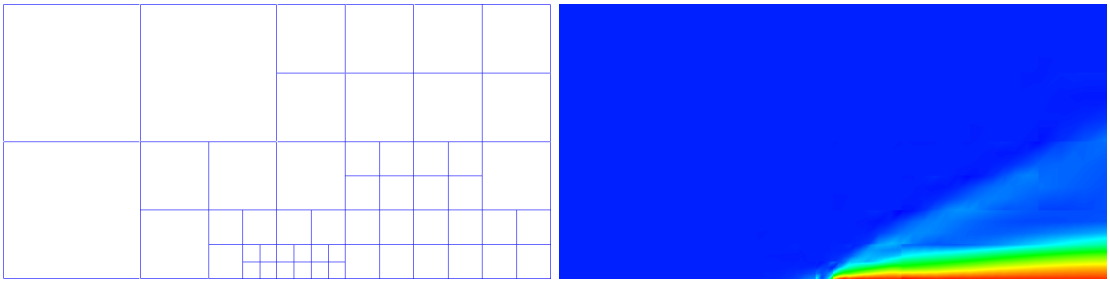
(c) u_2 (d) T

Figure 6.6: Solutions after twelve refinements for $p = 2$ and $\mathrm{Re} = 1000$, starting from a mesh of 2 elements.

We observe numerically that ρ also behaves singularly at the plate tip – due to the presence of this strong singularity, the coloring of the wide range of values for ρ in Figure 6.6 causes the solution to appears largely uniform, save for a flare up at the tip of the plate. To better visualize density, ρ is rescaled such that the features of the solution away from the singularity, as well as the final mesh after 10 refinement steps, are visible in Figure 6.13.

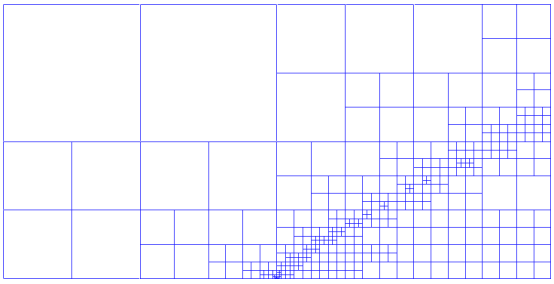
We can also zoom in on the plate tip to view the solution quality at the singular point. Figure 6.9 demonstrates that the solution remains smooth and well-resolved at the plate tip, despite the presence of a singularity in the viscous stresses.

We increased the Reynolds number to $10,000$ to assess the behavior of DPG for higher Reynolds numbers. However, we found it necessary to modify the method in several ways in order

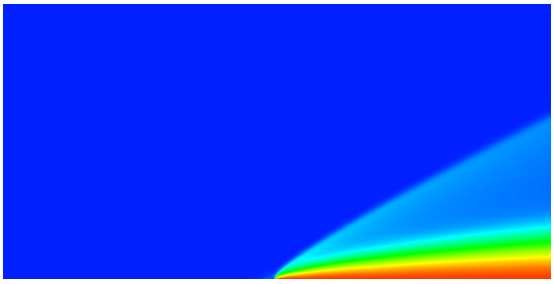


(a) Mesh, 4th refinement

(b) T, 4th refinement

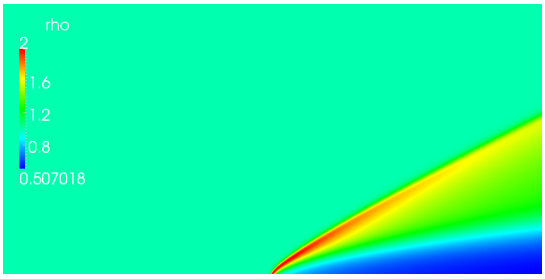


(c) Mesh, 8th refinement

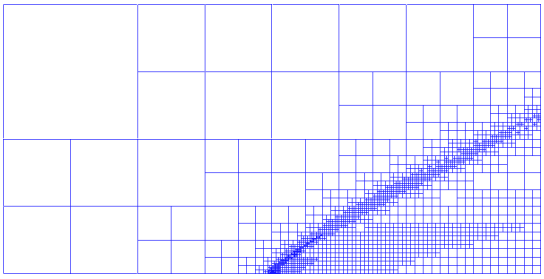


(d) T, 8th refinement

Figure 6.7: Snapshots of adaptive meshes and solutions for two different steps of adaptivity Re = 1000.

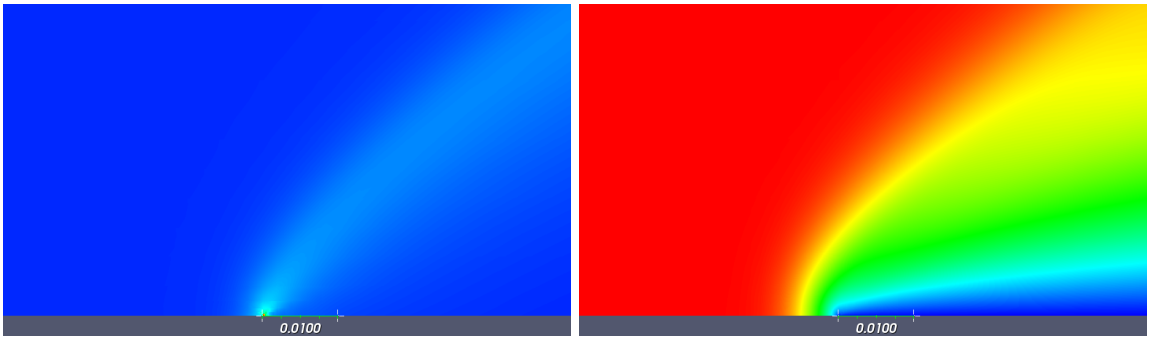


(a) ρ

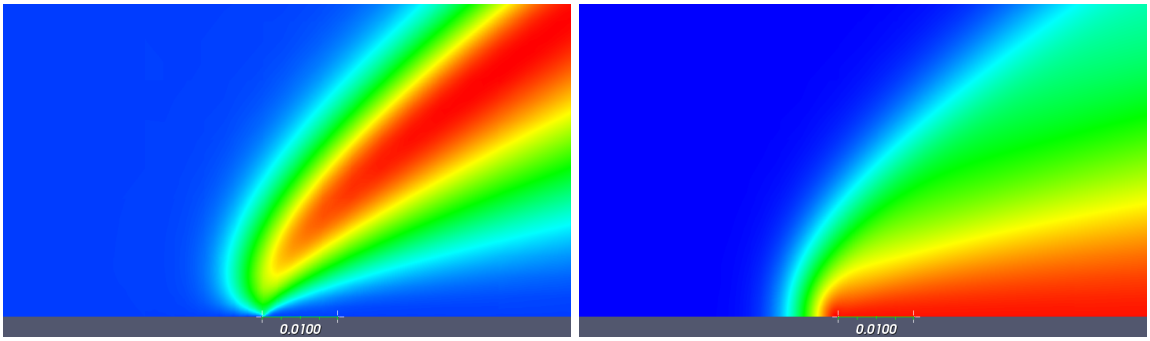


(b) Mesh

Figure 6.8: Rescaled solution for ρ in the range $[\rho_{\min}, 2]$ and adaptive mesh after 12 refinement steps.



(a) ρ (b) u_1



(c) u_2 (d) T

Figure 6.9: Zoom of solutions at the beginning of the plate for $p = 2$ and $\mathrm{Re} = 1000$.

to achieve satisfactory results.

First, we implemented a line search algorithm to enforce positivity of both temperature and density, which are physically defined to be positive quantities. Given updates $\triangle\rho$ and $\triangle T$, we update our previous solution by setting

$$\rho \coloneqq \rho + \alpha_{\text{line}} \triangle\rho$$

$$T \coloneqq T + \alpha_{\text{line}} \triangle T,$$

where α_{line} is chosen, for some $\delta > 0$, such that

$$\rho + \alpha_{\text{line}} \triangle\rho - \delta = 0$$

$$T + \alpha_{\text{line}} \triangle T - \delta = 0.$$

Since the addition of a line search can slow the convergence of a nonlinear algorithm, we incorporate also a Newton iteration at each timestep to effectively solve the nonlinear system at each timestep. We consider $\|\triangle U\|_E < \epsilon_{\text{Newton}}$ to be our condition for convergence of the Newton iteration, though we also limit the number of allowed Newton steps for computational efficiency. The full solver algorithm is given in Algorithm 2. For higher Reynolds numbers and highly refined meshes, the solution

Algorithm 2 Pseudo-timetepping adaptive algorithm with line search.

```
for number of refinement steps do
  while R_time > ε_t do
    k = 0
    while ‖△U‖_E > ε_Newton and k < maximum Newton steps do
      Solve for △U, determine α_line.
      U := U + α_line △U.
      k = k+1
    end while
    Increment timestep.
  end while
  Compute energy error and refine based on a greedy refinement algorithm.
  Set ε_t = α_t ‖u − u_h‖_E.
end for
```

update exhibits large oscillations, such that the solution at a timestep can become negative, under which the pseudo-timestepping algorithm can stall or even diverge[6]. However, we have observed

[6]We note that these large oscillations mimic experiences in 1D as well, where the linearized solution was shown to exhibit sharp gradients near shocks that did not disappear, even with additional mesh refinement, indicating that the presence of such overshoots and undershoots is a consequence of the linearization, as opposed to the stability of the discretization [67]. This is discussed in more detail in Section 6.5.3.

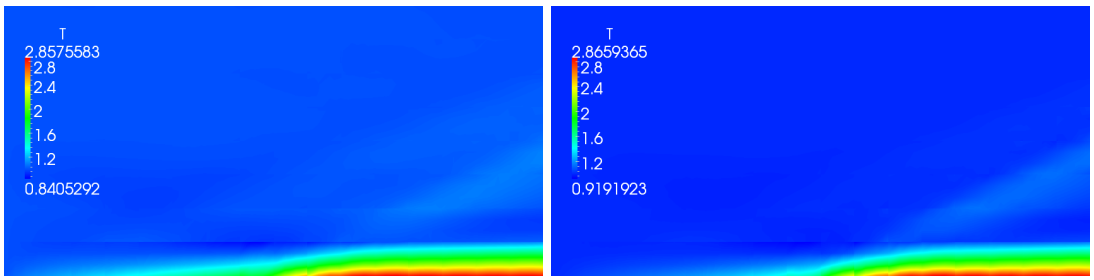
We note also that theory developed in Chapter 5 for the convection-diffusion problem assumes a smooth convection field. However, under linearization of the Burgers' and Navier-Stokes equations, the solution around which we linearize dictates the convection field, and can display large gradients. While we have not observed issues in the Burgers' equation related to this, we hope to revisit the analysis done in Chapter 5 and generalize it for convection fields with large gradients.

that requiring a strictly positive solution appears to be too restrictive a constraint for some meshes; the use of line search does not appear to be necessary for convergence of the pseudo-timestepping algorithm on coarser meshes (until the 10th refinement iteration, or $h < .001$).

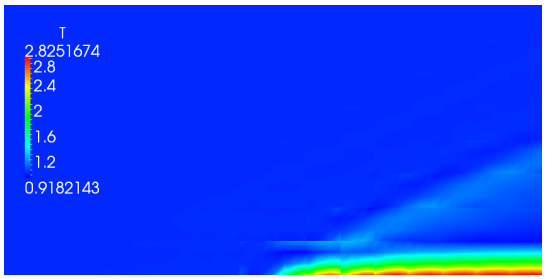
Finally, we implement an "effective" CFL number; though implicit time integration schemes are unconditionally stable (compared to explicit schemes), a CFL condition relating the size of the time increment to the (minimum) mesh size is often still used in practice to improve convergence speed and stability of the numerical scheme [68]. Our CFL number is chosen to be 64. We note that this CFL number is implemented for non-standard reasons. In fact, DPG is able to solve the steady-state system directly without the use of pseudo-timestepping (direct Newton iteration). However, as noted previously, the size of the timestep under which pseudo-timestepping converges greatly affects the qualitative nature of the solution. Figure 6.10 demonstrates the difference between convergence at large and small timesteps – for dt large relative to the mesh size, the solution experiences upstream "pollution" effects. Due to the upstream "pollution" present in the solution for $dt \geq 1$, the adaptive mesh refinement algorithm tends to add extraneous refinements on elements adjacent to the boundary $y = 0$, $x \in (0, 1)$. Decreasing the timestep alleviates this issue somewhat; however, an overly small timestep requires a large number of iterations to converge. The implementation of an effective CFL number aims to balance the size of the timestep with the mesh size.[7]

We note that, even with all of the above modifications, the pseudo-timetepping adaptive algorithm with line search would sometimes failed to converge below ϵ_t for $\mathrm{Re} \geq 10,000$. Figure 6.11

[7] An additional reason for the implementation of an effective CFL number is the conditioning of the local problem, which was discussed for the convection-dominated diffusion problem in [7]. The main problem concerning conditioning of local problems is the way that different test terms behave as a function of local element size. We illustrate this using the element Sobolev norm $\|v\|_{H^1(K)} = \|v\|_{L^2(\Omega)} + \|\nabla v\|_{L^2(\Omega)} - \|v\|_{L^2(\Omega)} = O(h^2)$ (where h is the element size), while $\|\nabla v\|_{L^2(\Omega)} = O(1)$. Thus, as $h \to 0$, the Sobolev norm over a single element approaches the Sobolev seminorm and loses positive definiteness, resulting in a highly ill-conditioned system to solve. The addition of a first-order pseudo-timestepping term allows us to increase the relative magnitude of $\|v\|_{L^2(\Omega)}$ with respect to $\|\nabla v\|_{L^2(\Omega)}$ and avoid conditioning issues while maintaining robustness of the method (see Appendix 2.2 for the proof of robustness).



(a) $dt = \infty$ (Direct Newton) (b) $dt = 1.0$

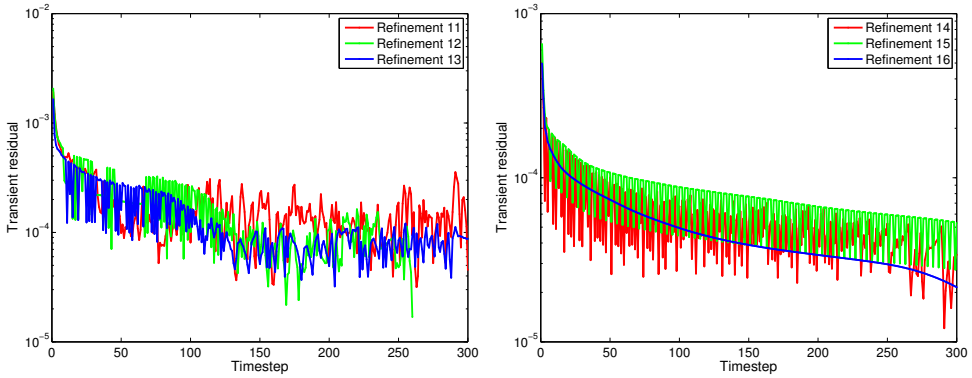


(c) $dt = .1$

Figure 6.10: Steady state solutions for $\mathrm{Re} = 10,000$ under three different timesteps on a 16×8 uniform mesh.

shows a plot of the transient residual after the 14th refinement step; while the residual initially decreases, it stalls at about $R_{\text{time}} \approx 10^{-4}$. An examination of the difference in the solutions between the final and 125th timesteps shows that the nonconvergence of the transient residual is due to oscillations in the solution (primarily in ρ) slightly upstream to the plate edge. Visually, the solution converges everywhere else, save for this area. Such behavior is also observed in [61], where, at the change in boundary conditions between the free stream and flat plate, an oscillation was observed in the solution which prevented convergence of his pseudo-time algorithm. His oscillations are of smaller magnitude ($O\left(10^{-6}\right)$ as opposed to $O\left(10^{-4}\right)$), which may be the result of several differences between the methods presented.[8] However, we note that, under additional refinements and increased

[8] Apart from the use of a standard Galerkin (continuous as opposed to discontinuous Galerkin) formulation, Kirk's approach differed from ours in the use of linear elements and the addition of artificial diffusion shock-capturing terms,



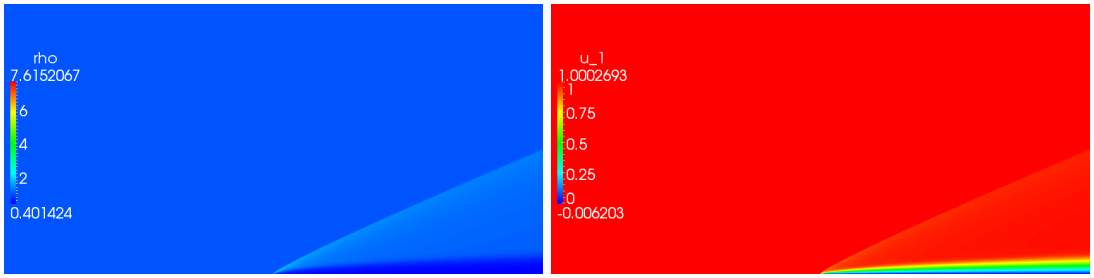
(a) Non-convergence (b) Recovery of convergence

Figure 6.11: Stalling and recovery of the pseudo-timestep iteration for Re = 10000.

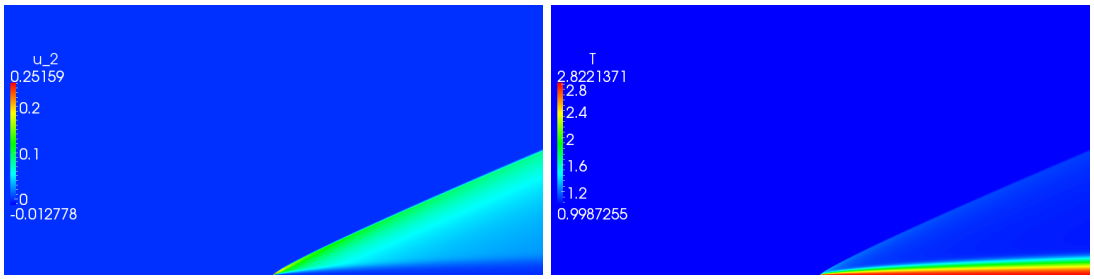
resolution of the solution, the transient residual once again decreases in a smooth monotonic fashion, as shown in Figure 6.11.

For additional computational efficiency, we also implemented an anisotropic refinement scheme. In 2D, the boundary layer is a primarily 1D phenomena, which we expect to be far better resolved by anisotropic refinement than by isotropic refinement. We experimented first using the scheme described in Section 5.4 as an anisotropy indicator; however, the anisotropic scheme appeared to be too conservative near the boundary layer, placing primarily isotropic refinements. We modified our scheme in two ways – first, we incorporated spatially variable thresholding. Typically, anisotropic refinement in the x direction is chosen if $e_{x,K} > \epsilon_r e_{y,K}$ (and vice versa for anisotropic refinement in the y-direction), where $e_{x_i,K}$ is the error in the x_i direction. We set $\epsilon_r = \epsilon_{r,K}$; in other words, we allow our anisotropic threshold to vary element-by-element, and decrease it from $\epsilon_r = 10$ to $\epsilon_{r,K} = 2.5$ for elements adjacent to the wall upon which the boundary layer forms. Additionally, since the boundary layer typically displays rapid variation in the y-direction (the direction orthog-

both of which could explain the lower point at which error stagnates. Reasons for this loss of convergence in the pre-asymptotic range should be explored further.



(a) ρ (b) u_1



(c) u_2 (d) T

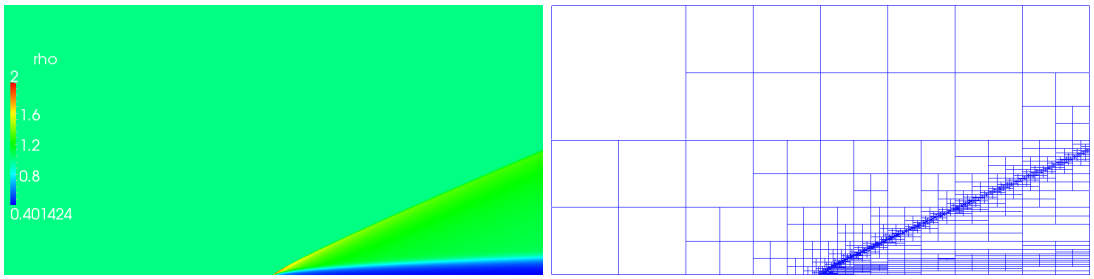
Figure 6.12: Solutions after 18 refinements for $p = 2$ and $\mathrm{Re} = 10000$.

onal to the wall), we relax the condition under which a vertically cut anisotropic refinement occurs to

$$2e_{y,K} > \epsilon_{r,K} e_{x,K}.$$

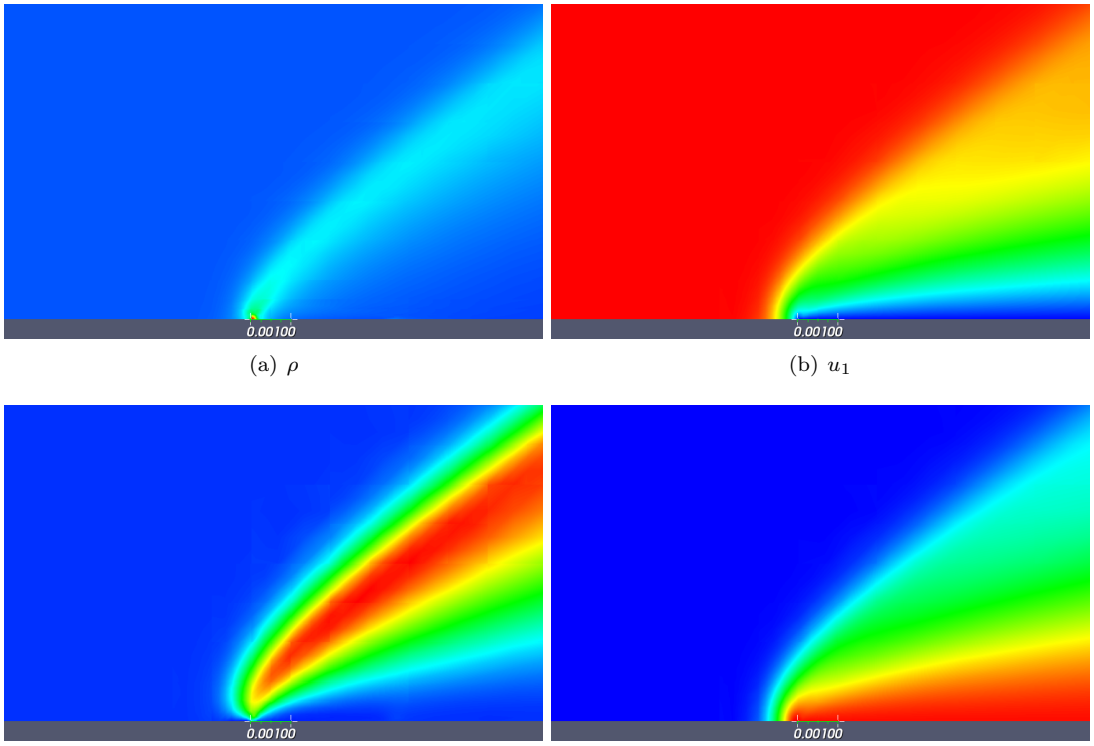
Despite the artificial modification of the anisotropic refinement scheme, the resulting meshes still resolve boundary layer solutions more efficiently than isotropic refinements. We hope to investigate reasons for the ineffectiveness of the pure anisotropic scheme for the compressible Navier-Stokes equations in future research.

We note that the resolution of the solution near the plate edge in Figure 6.14 for Reynolds number 10000 is qualitatively rougher than that for Reynolds 1000 in Figure 6.9; this is due to the greedy refinement algorithm emphasizing refinements most strongly at the singular point and not in shock resolution, as well as the fact that at a higher Reynolds number, the shock width is thinner



(a) ρ (b) Mesh

Figure 6.13: Rescaled solution for ρ in the range $[\rho_{\min}, 2]$ and adapted mesh.



(a) ρ (b) u_1

(c) u_2 (d) T

Figure 6.14: Zoom of solutions at the beginning of the plate for $p = 2$ and $\mathrm{Re} = 10000$.

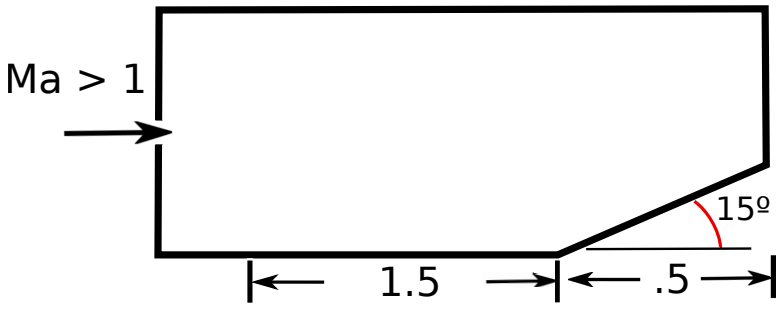


Figure 6.15: A modification of the Holden ramp/compression corner problem.

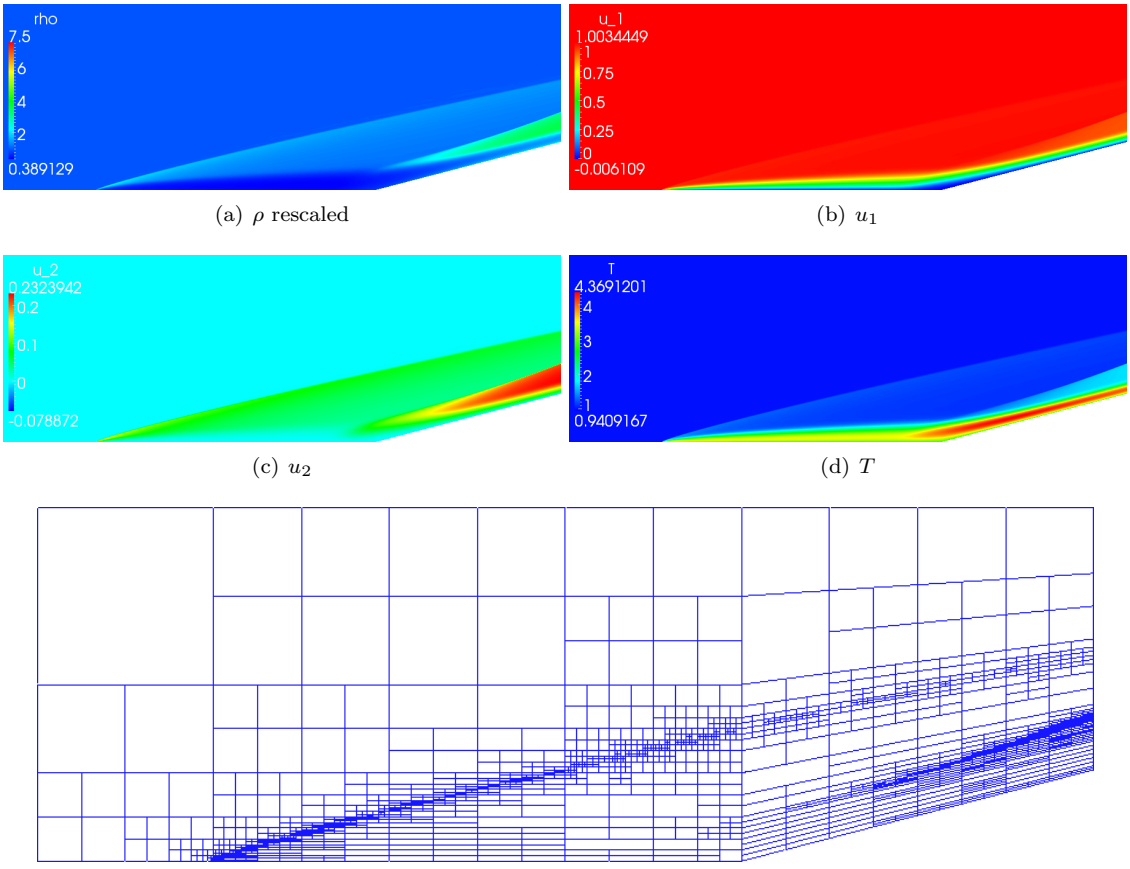
and more difficult to resolve. Further refinement steps improve resolution of the shock features.

6.5.2 Holden ramp problem

Our second problem is a modified version of the Holden ramp problem, which models supersonic/hypersonic flow over a compression corner (the geometry of which is given in Figure 6.15). Similarly to the Carter flat plate problem, a flat plate disrupts the flow and forms a weak shock at the plate tip due to viscous effects and no-slip boundary conditions. The boundary layer grows down the plate edge, deflecting upwards due to the presence of the compression corner. A stronger shock forms slightly upstream of the compression corner in order to deflect the incoming supersonic flow, and is a common test problem for adaptive finite element methods for compressible flow [69, 12, 61].

The original plate length is given to be .442, while the ramp length is given to be .269. We have modified the problem slightly in order to exactly represent the boundary conditions on a coarse mesh while keeping the ratio of plate length to ramp length roughly the same. Similarly to the Carter flat plate, we start out with a very coarse 2×3 mesh of 6 elements. We initialize our solution to the freestream values

$$\rho_\infty = 1, \qquad u_{1,\infty} = 1, \qquad u_{2,\infty} = 0, \qquad T_\infty = 1$$



(a) ρ rescaled

(b) u_1

(c) u_2

(d) T

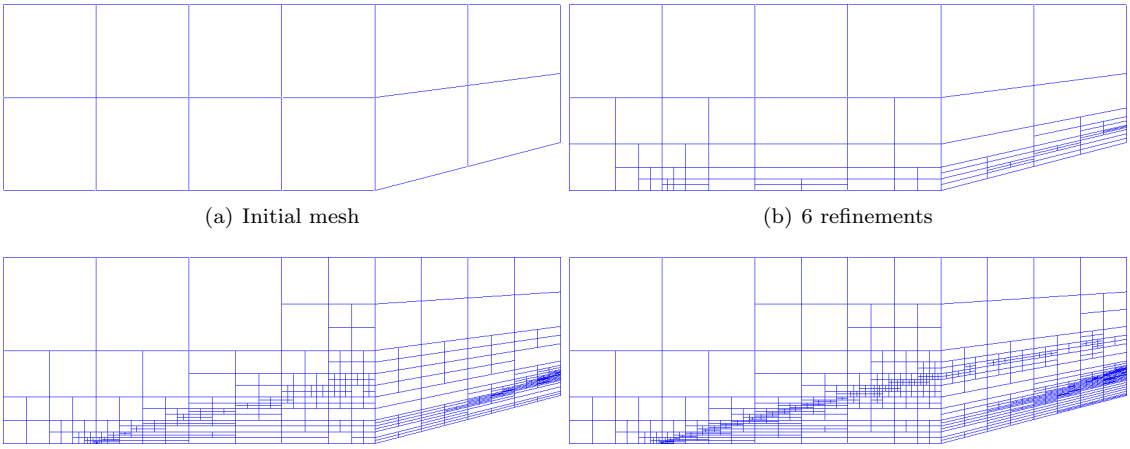
(e) Mesh after 24 refinements

Figure 6.16: Solutions and adaptive mesh for Ma = 6 and Re = 10000.

and again set stresses uniformly to zero.

We first solve under Mach 6 flow[9] and Reynolds number of 10,000, or a Reynolds number of 33,936 if measured with respect to the original plate length of .442. The wall temperature is set to $T_w = 2.8T_\infty$, identically to the flat plate problem. 24 mesh refinements were performed, resulting

[9]The increase in Mach number is to change the angle of the shock; under the current setup, Mach 3 flow produced a shock which reflected off the top boundary $y = 1$. We note that the effect of Mach number under our nondimensionalization of choice is to decrease the thermal diffusivity constant κ relative to the viscosities μ and λ.



(a) Initial mesh (b) 6 refinements

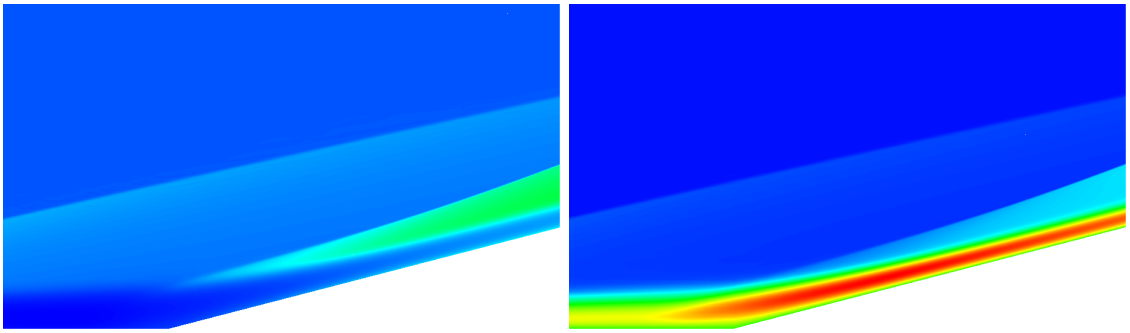
(c) 12 refinements (d) 18 refinements

Figure 6.17: Sequence of adaptive meshes for Ma = 6 and Re = 10000.

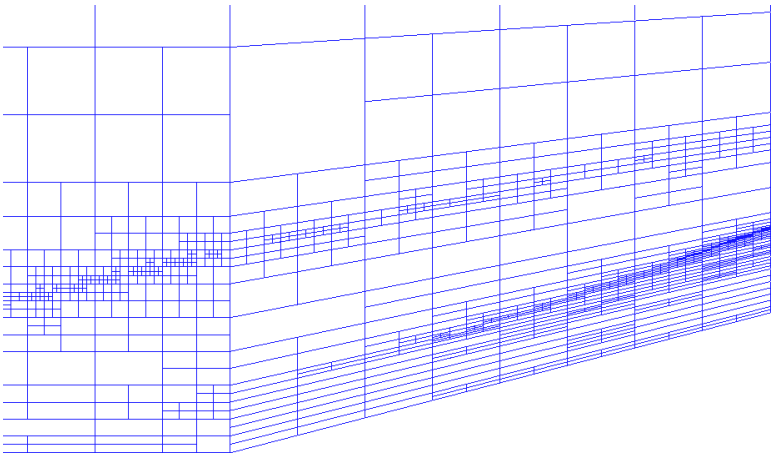
in a final mesh of 1858 elements. Figure 6.16 shows both solution values and final adaptive mesh. Due to a large maximum value of $\rho_{\max} = 14.1538$ at the plate tip, the resulting solution for ρ is scaled to better show qualitative features of the flow. The presence of the second shock deflecting the incoming supersonic flow at the ramp is clearly seen in both the solutions and the adaptive mesh refinements.

We can examine the sequence of adaptive meshes generated by the DPG method. Figure 6.17 shows the initial mesh, as well as the 6th, 12th, and 18th subsequent refinements generated automatically by the DPG method. Unlike the previous sequence of meshes generated by the flat plate example, refinements tend to be placed most heavily near the shock at the outflow ramp. By the 18th refinement, the adaptive mesh looks qualitatively very similar to the final mesh of 24 refinements, and further refinements focus on the resolution of the shocks originating at the plate edge and at the ramp outflow.

Figure 6.18 shows a zoom of the second stronger shock that develops near the ramp outflow.



(a) ρ rescaled (b) T

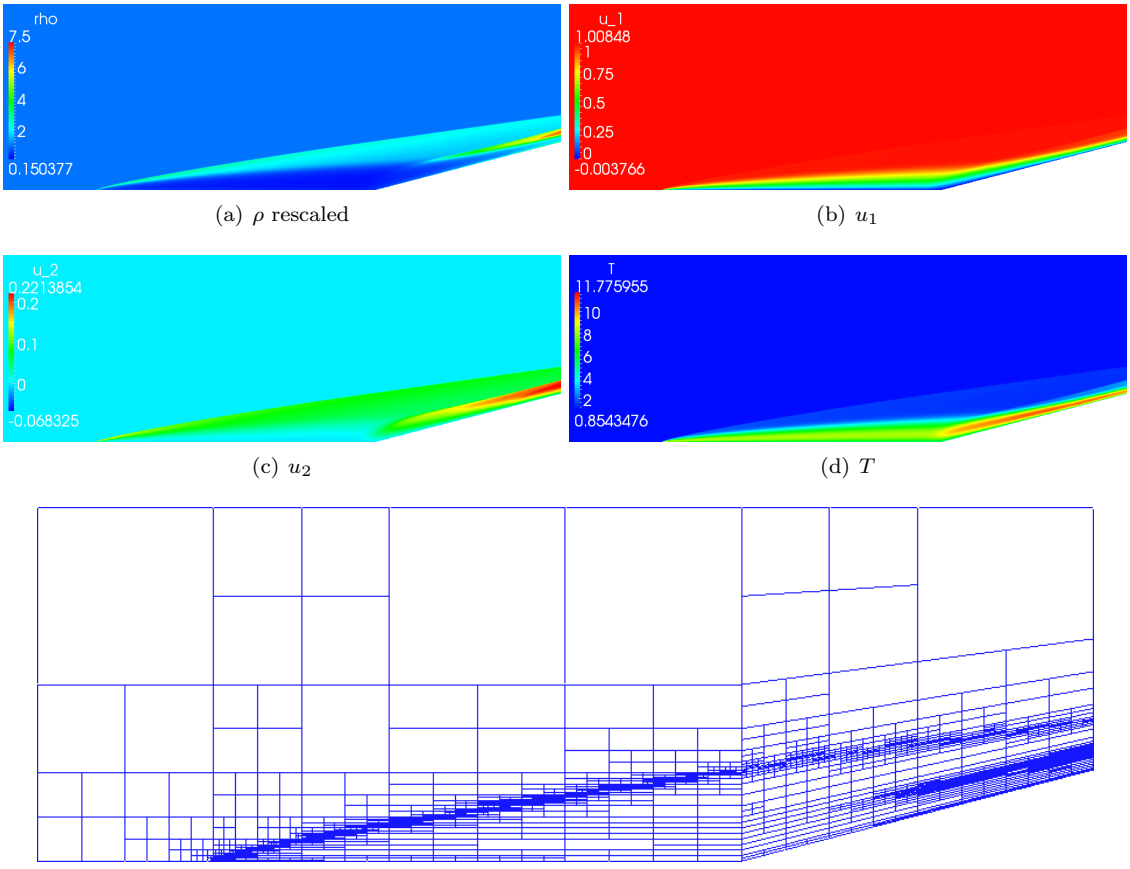


(c) Mesh

Figure 6.18: Zoom of solutions and adaptive mesh around shock.

Refinements are placed very heavily near the shock, which we expect due to the fact that a shock forms a stronger gradient than a boundary layer. We note that our residual-driven adaptivity scheme places mesh refinements in a very precise manner; refinements on ramp boundary are placed slightly more heavily upstream than downstream. We believe this is due to the fact that solution gradients are slightly higher in ρ at the upstream section of the ramp.

The second set of conditions under which we solve are under Mach 11.68 flow and Reynolds number of 16,4 42.4, or 55,800 if measured with respect to original plate length, with the wall



(a) ρ rescaled

(b) u_1

(c) u_2

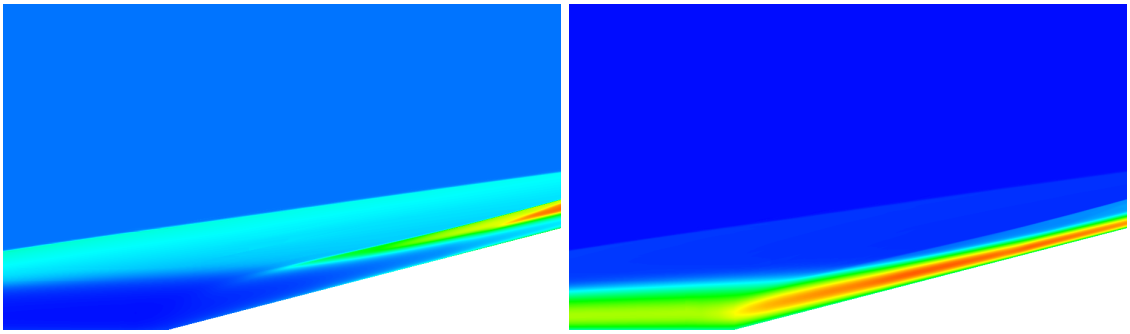
(d) T

(e) Mesh

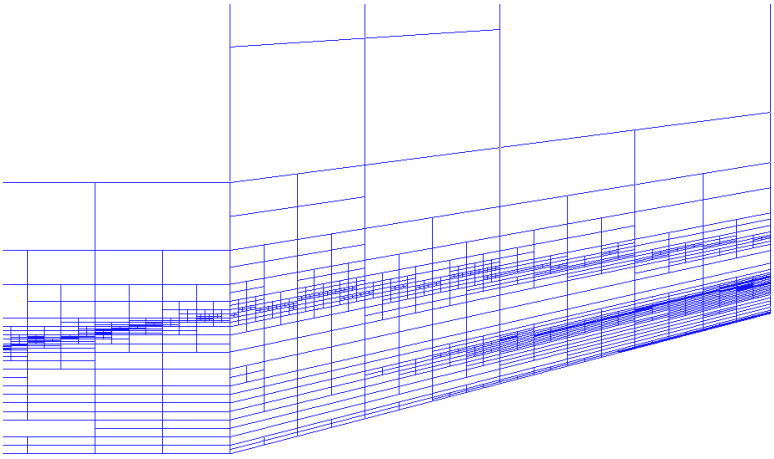
Figure 6.19: Solutions and adaptive mesh for Ma = 11.68 and Re = 16442.4.

temperature set to $T_w = 4.6T_\infty$. 24 automatic mesh refinements are performed under an energy threshold of .5, resulting in a final mesh of 2385 elements.

Figure 6.19 shows the resulting solution and final adaptive mesh. The increased Mach number changes again the angle of the weak shock resulting from the change in boundary conditions at the plate edge. Compared to the previous case of Mach 6 flow, where only 18 refinements steps were performed, 24 refinements were performed for the Mach 11.68 case, leading to a more highly resolved solution, especially near the plate tip. Due to a large maximum value of $\rho_{\max} = 50.1805$ at



(a) ρ rescaled (b) T



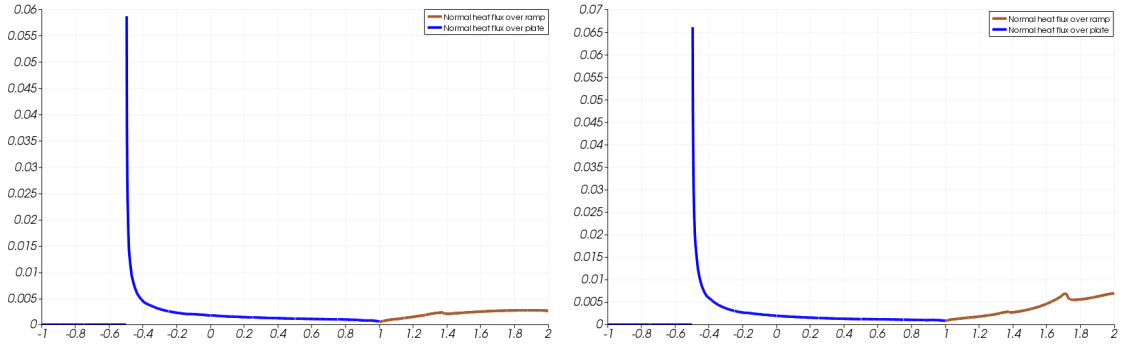
(c) Mesh

Figure 6.20: Zoom of solutions and adaptive mesh around shock.

the plate tip, the resulting solution for ρ is scaled to better show qualitative features of the flow.

Compared to the Mach 6 case, where 18 refinements were performed, the resolution of the mesh near the shock at the ramp outflow for Mach 11.68 flow and 18 refinements is qualitatively similar. Further refinement steps increase resolution of the mesh near the shock, as shown in Figure 6.20.

Finally, we plot the normal heat flux over the plate and ramp for both Holden problems in Figure 6.21. The normal heat flux over the flat plate is indicated by the blue line, while the dotted



(a) Mach 6, Reynolds 10,000 (b) Mach 11.68, Reynolds 16,442.4

Figure 6.21: Normal heat flux q_n over plate and ramp for the Holden problem. The blue line indicates q_n over the flat plate, while the brown line indicates q_n over the ramp.

brown line indicates the heat flux over the ramp, and both are plotted against the x-coordinate along the plate/ramp boundary. The normal heat flux is given by the normal trace of the conserved quantity in the energy equation

$$\widehat{f}_{4,n} = (\rho e + p)u_n - \boldsymbol{n} \cdot \boldsymbol{\sigma} \cdot \boldsymbol{u} + q_n.$$

Recognizing that $u_1 = u_2 = 0$ reduces the above to $\widehat{f}_{4,n} = q_n$.

The heat flux develops a strong singularity at the point $(-.5, 0)$, where the boundary condition changes from a Neumann/stress boundary condition to a Dirichlet/no-slip boundary condition.[10] Section 5.3.1 proves that the Laplace equation develops a singularity in stress at under any such change in boundary conditions, and the same phenomena is observed for a similar setup under the convection-diffusion equation (see also [53]).

[10]Figure 6.21 cuts off this singularity in order to show the qualitative behavior of q_n over the remainder of the boundary. The maximum visualized values in the singular portion of q_n are .35 for Mach 6 flow and 1.268 for Mach 11.68 flow.

6.5.3 Higher Reynolds numbers

For higher Reynolds number, the pseudo-timestepping, we have encountered difficulties which have made it difficult to converge to a steady-state solution, even under the addition of an effective CFL number and line search to enforce positivity of density ρ and temperature T. However, we believe these difficulties to be related to the nature of the equations, rather than the robustness of the method. We illustrate this with a simple example.

Let us consider the 1D steady state viscous Burgers' equation on $[-\infty, \infty]$

$$u\frac{\partial u}{\partial x} - \epsilon\frac{\partial^2 u}{\partial x^2} = 0.$$

The exact solution to this equation under boundary conditions

$$u(-\infty) = 1$$

$$u(\infty) = -1$$

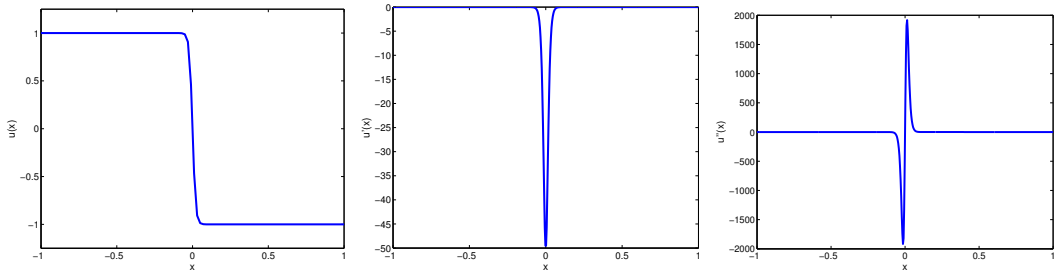
can be easily derived (see [12] for a simple derivation), and the solution and its first two derivatives are

$$u(x) = \frac{1 - e^{\frac{x}{\epsilon}}}{1 + e^{\frac{x}{\epsilon}}}, \qquad u'(x) = \frac{-2e^{\frac{x}{\epsilon}}}{\left(1 + e^{\frac{x}{\epsilon}}\right)^2 \epsilon}, \qquad u''(x) = \frac{-2e^{\frac{x}{\epsilon}}\left(e^{\frac{x}{\epsilon}} - 1\right)}{\left(1 + e^{\frac{x}{\epsilon}}\right)^3 \epsilon^2}$$

Figure 6.22 shows each of these functions for $\epsilon = .01$. From the form of $u'(x)$ and $u''(x)$, we know these oscillations will grow rapidly as ϵ decreases. Consider now the linearized Burgers' equation

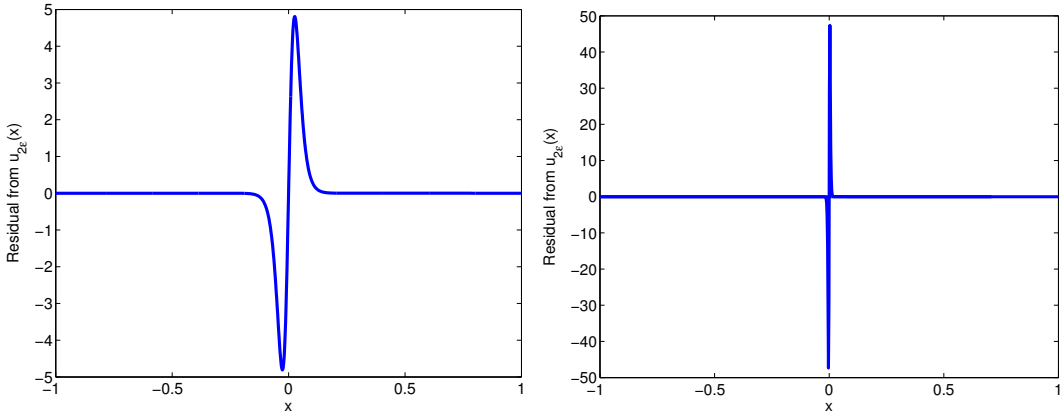
$$\frac{\partial u \Delta u}{\partial x} - \epsilon\frac{\partial^2 \Delta u}{\partial x^2} = -r(x)$$

where $r(x) = u\frac{\partial u}{\partial x} - \epsilon\frac{\partial^2 u}{\partial x^2}$ is the strong form of the nonlinear residual. Recall that for a pure Newton iteration, $u(x)$ is assumed to be known, and the linearized problem is driven by the residual. The solution $u(x)$ is updated $u \coloneqq u + \Delta u$, and is repeated until Δu and $r(x)$ are both approximately zero.



(a) $u(x)$ (b) $u'(x)$ (c) $u''(x)$

Figure 6.22: $u(x)$ and its derivative and second derivative for $\epsilon = .01$.



(a) $\epsilon = .01$ (b) $\epsilon = .001$

Figure 6.23: Residual for Burgers' equation with viscosity ϵ under the exact solution for 2ϵ.

Let $u_\epsilon(x)$ be the exact solution for a particular viscosity ϵ, and consider the setting of $u(x) = u_{\alpha\epsilon}(x)$, with $\alpha > 1$. In other words, the initial guess for the Newton iteration is taken to be the exact solution for the viscous Burgers' equation under a larger viscosity (a less sharp shock), a method known as continuation (specifically, continuation in viscosity ϵ).[11]

We plot $r(x)$ for $u(x) = u_{2\epsilon}(x)$ in Figure 6.23. While the form of the exact linearized solution

[11] We note that continuation in Reynolds number was attempted for the Navier-Stokes equations; however, the presence of large oscillations in the Newton update on highly adapted meshes caused the line search to return a near-zero step length, which stalled the nonlinear iteration prior to convergence to a steady state solution.

Δu is unknown, we see that the forcing term in the above equation develops oscillations that grow in magnitude and decrease in support as ϵ decreases. We have observed similar behaviors for discretized solutions to the Burgers' and Navier Stokes equations – the discrete linearized solution will develop gradients as well, though their magnitude will be limited by the resolution of the mesh. However, additional refinements near shocks will introduce additional oscillations, which are subsequently damped by additional Newton iterations.

In other words, not all oscillations are related to the method of discretization – the exact linearized solution itself contains large oscillations, which are picked up more and more strongly in a discrete scheme as the mesh resolution approaches the viscous length scale ϵ (see [70, 71] for additional numerical examples). We note that these large oscillations are not an issue for the viscous Burgers' equation, as there is no positivity constraint on u. However, as noted previously, the convergence of the nonlinear iteration for the compressible Navier-Stokes equations can stall or even diverge for sufficiently negative values of ρ and T, which happens easily in the presence of large oscillations. When a line search is implemented to enforce a strictly positive solution, the nonlinear iteration can stall, and numerical experiments have generated cases in which the line search length goes to zero, returning an under-converged solution.

There are several approaches to controlling the magnitude of oscillations in the linearized solution – a simple option is decreasing the size of the timestep; however, doing so can cause the number of iterations required for convergence to greatly increase. The application of artificial diffusion near sharp gradients is another way in which to control such oscillations; however, this results in a modified set of equations, and, though solutions with contemporary artificial viscosity methods can produce solutions with expected qualitative features, there is a wide array of choices for artificial diffusion, and it was not clear in the scope of this dissertation which artificial viscosity

method would be quantifiably superior, or most suitable for use with the DPG method.

Chapter 7

Conclusions and future direction

The goal of this work has been to explore the behavior of the Discontinuous Petrov-Galerkin (DPG) method as a method for the discretization and solution of convection-dominated diffusion problems, and produce both theory and numerical results using this method applied to model problems in this area. We begin in Chapter 1 by describing convection-diffusion problems in fluid dynamics and the issues faced by naive methods in the convective limit, with particular emphasis on the *robustness* of the method, and give a brief survey of numerical methods tailored for convection-dominated diffusion problems. In Chapter 2, we introduce the range of problems we are interested in addressing. Specifically, we are interested in the convection-diffusion class of singular perturbation problems in computational fluid dynamics, and discuss the compressible Navier-Stokes equations, as well as the simpler model problems of Burgers' equation and the linear convection-diffusion equation, upon which we develop our numerical method.

In Chapter 3, we introduce the DPG method for linear problems. The concept of problem-dependent optimal test functions is derived through equivalence with the minimization of a specific residual, and discontinuous test functions are introduced in order to localize the determination of such optimal test functions to a single element. The approximation of such test functions using a high order spectral method is discussed, and conclude the chapter by introducing the ultra-weak variational formulation with which the concept of optimal test functions is paired and its corresponding energy spaces.

In Chapter 4, we show how DPG's locally constructed test space can be interpreted as a non-conforming approximation of a weakly-conforming global test space under the ultra-weak formulation. Furthermore, the field solutions and trace solutions on the boundary Γ are shown to depend *only* upon properties of the non-conforming global test space; thus, when considering approximation error in test functions, resolution of global approximation error (as opposed to local approximation error) can be sufficient to produce a stable method. Additionally, global properties of test spaces are given under which the ultra-weak formulation delivers the best L^2-approximation to the exact solution. A connection is made to DPG through the graph test norm, which can be viewed as a regularization of the graph seminorm through the addition of an L^2 term of magnitude δ. As the regularization parameter $\delta \to 0$, DPG under this version of the graph test norm is shown to converge to a weakly-conforming approximation to the L^2-optimal test space. Finally, viewing DPG as an approximation to globally optimal test functions allows the construction of test spaces that focus on resolution of global features as opposed to local features. We illustrate this with the convection-diffusion problem, where we show that it is possible to restore robustness with respect to the diffusion parameter ϵ by neglecting the resolution of boundary layers on the element-local level and focusing on the resolution of global boundary layers in optimal test spaces.

In Chapter 5, we presented the analysis of a non-canonical test norm and its corresponding DPG energy norm for the convection-diffusion equation in the small-diffusion limit for solutions with strong boundary layers. Additionally, we have introduced a non-standard inflow boundary condition, and have explored the difference between between this and the standard Dirichlet inflow boundary condition. Both a definition and proof of robustness are given, and approximation of test functions is addressed. Numerical results are presented in order to verify the results derived in this paper. However, at least for our model problem, numerical experiments appear to demonstrate results that

are stronger than our proofs indicate, delivering solutions for u and σ that are extremely close to their best L^2 projections. Finally, the test norm is modified to address problems with singularities. A model problem using Laplace's equation is formulated to illustrate the presence of singular solutions to the convection-diffusion equation, and difficulties in control of singularities under the previously developed test norm are demonstrated. Numerical experiments demonstrate that the new test norm resolves previous issues, and is effective in controlling singularities in solutions of the convection-diffusion equation.

In the final chapter Chapter 6, we extend the methodology for linear problems to two model nonlinear problems. We describe several common methods for the solution of nonlinear equations, and describe the application of the DPG method to each of them. We demonstrate the effectiveness of DPG for nonlinear problems on a model Burgers problem with a shock solution, and then apply the DPG method to solving two model problems in supersonic/hypersonic compressible flow under different Reynolds numbers. In particular, we demonstrate for both the flat plate and compression ramp problem in supersonic/hypersonic in compressible flow that the DPG method is able to begin from a highly underresolved meshes (two elements for the Carter plate problem, and 12 elements for the Holden ramp problem), and through automatic adaptivity, is able to resolve physical features of the solution without the aid of artificial diffusivity or shock capturing terms. We believe this indicates both the robustness of the method on coarse grids and the effectiveness of the DPG error indicator for adaptive refinement.

In conclusion, we have examined carefully the application of the DPG method to linear convection-dominated diffusion problem, where Galerkin test functions are computed automatically based on a choice of basis functions and the variational formulation. We have introduced a novel variational formulation – the "ultra-weak" variational formulation – and have analyzed the nature

of test functions resulting from a "canonical" choice of test norm. After showing that such test functions display strong boundary layers, we concluded that resolution of the such test functions was infeasible, and developed a version of the DPG method for linear convection-dominated diffusion problems whose behavior does not degenerate as $\epsilon \to 0$. The end goal of such an analysis was to present a method which could adaptively solve a heavily convection-dominated diffusion problem despite beginning with a highly under-resolved initial mesh. We extrapolated such a method to a nonlinear Burgers equation and two model problems in viscous compressible flow and demonstrated its usefulness by using an automatic adaptivity scheme to fully capture features of the flow, starting with a mesh requiring no prior knowledge of the solution or physics of the simulation.

7.1 Accomplishments

In the theoretical scope of this dissertation, I have developed and proven the robustness of a Discontinuous Petrov-Galerkin method for convection-diffusion problems. In particular, I have introduced an alternative inflow boundary condition and demonstrated its regularizing affect on the adjoint problem, allowing for the use of a stronger test norm. Additionally, I have developed theory detailing the global nature of the DPG test space, and have shown that, for a specific series of test norms, the global DPG test space converges to a weakly-conforming approximation of the global test space under which the ultra-weak variational formulation yields the L^2-best approximation. Finally, I have extended the DPG framework to nonlinear problems, demonstrating the equivalence of the DPG method to a Gauss-Newton minimization scheme for the nonlinear residual.

In the numerical and computational scope of this dissertation, I have confirmed numerically the robustness of the DPG method for convection-diffusion problems in the convective limit under arbitrary high-order adaptive meshes. I have implemented an anisotropic refinement scheme to more

effectively capture lower-dimensional behavior of solutions of convection-diffusion problems, such as boundary layers. Finally, I have contributed to the development of the parallel *hp*-adaptive DPG codebase Camellia[5], under which the results in this dissertation were produced.

Finally, this dissertations includes the application of the DPG method to several model convection-diffusion problems. Convergence of the method is demonstrated under an exact solution to the scalar convection-diffusion problem, and the method is extrapolated to a nonlinear viscous Burgers' equation. Finally, the DPG method is extrapolated to systems of equations and used to solve the flat plate and compression ramp problems in supersonic/hypersonic compressible flow.

7.2 Future work

As is the case with any research, much work remains to be done. We outline here several areas of work which we hope to pursue in the future.

- **Nonlinear DPG** – as described in Section 6.1, there is a natural Hessian-based version of the DPG method which provides a second-order approximation to the nonlinear equation instead of the first-order one afforded by Newton-Raphson linearization. Unfortunately, under this Hessian-based version of DPG, the stiffness matrix may no longer be positive definite, which can lead to non-descent search directions. We hope to avoid such issues through the use of Newton-CG methods[59], which avoid negative search directions by terminating the CG iteration in the presence of negative curvature.

- **Navier-Stokes** – We have chosen the classical variables in which to cast the compressible Navier-Stokes equations; however, investigation of alternative sets of variables may have merit, as different choices of variables yield differing linearizations with their own advantages (for

example, all derivatives in time are linear with respect to the momentum variables, and the entropy variables of Hughes both symmetrize the Navier-Stokes equations and yield solutions obeying second law of thermodynamics for standard H^1 formulations [72]).

We also hope to investigate artificial viscosity methods as regularization for problems in viscous compressible flow. We present an analysis of the 1D Burgers' equation demonstrating that the exact solutions under Newton linearization contain large oscillations. While these oscillations are not the result of the stability of the spatial discretization, their presence can cause density and temperature to become non-physically negative, which can stall the convergence of the nonlinear solver. We hope to investigate artificial viscosity not as a stabilization mechanism of the discrete spatial discretization, but as a regularization of the strong problem with which to suppress the presence of large oscillations in the linearized solution.

Finally, though the method is inf-sup stable for arbitrary meshes, most of our experiments have focused on meshes of uniform p. We hope to implement a true hp-adaptive DPG method for the compressible Navier-Stokes equations in the future.

- **Alternative discretizations** – Recent works (see [38, 73, 74, 75]) have applied the same minimum residual methodology behind DPG to alternative discretizations, as well as the use of *continuous* test functions.[1] We hope to investigate the behavior of the minimum residual method under both continuous test functions and different variational formulations.

- **Alternative architectures** – For an efficient implementation of DPG, massively parallel low memory architectures are required. In this work, we have focused on an MPI implementation

[1] We note that the use of continuous test functions does not imply the computation of such test functions over the entire mesh; it is shown that the minimum residual method can be formulated instead as a saddle point problem, which is equivalent to computing optimal test functions. See [38, 75] for more details.

of DPG. However, future access to extremely large MPI-based clusters may be limited to those who can afford the cost of petascale – namely, government-sponsored projects and large engineering companies. We hope to experiment with the GPU implementation of DPG as a lower-cost, highly parallel HPC alternative.

Appendix

Appendix 1

Proof of lemmas/stability of the adjoint problem

We present now the proofs of the three lemmas used in Section 5.1.3.2 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 (5.3) and (5.4) 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 4. *Assume v satisfies* (1.1)*, with boundary conditions* (5.3) *and* (5.4)*, and β satisfies* (5.7)

and (5.8). *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 5 complete the estimate. □

Lemma 5. *Assume β satisfies* (5.7)*. Then, for v satisfying equation* (1.1) *with boundary conditions* (5.3) *and* (5.4) *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 [3], 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 [3]; 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 6. *Let β satisfy conditions* (5.7) *and* (5.9)*, and let $v \in H^1(\Omega_h)$, $\tau \in H(\mathrm{div}, \Omega_h)$ satisfy equations* (5.5) *and* (5.6) *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 [3]. 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 (5.7). 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 5 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. □

Appendix 2

Additional notes on convection-diffusion

2.1 Boundary layers in robust norm test functions and global/local test spaces

In Section 5.1.3, we introduced the test norm

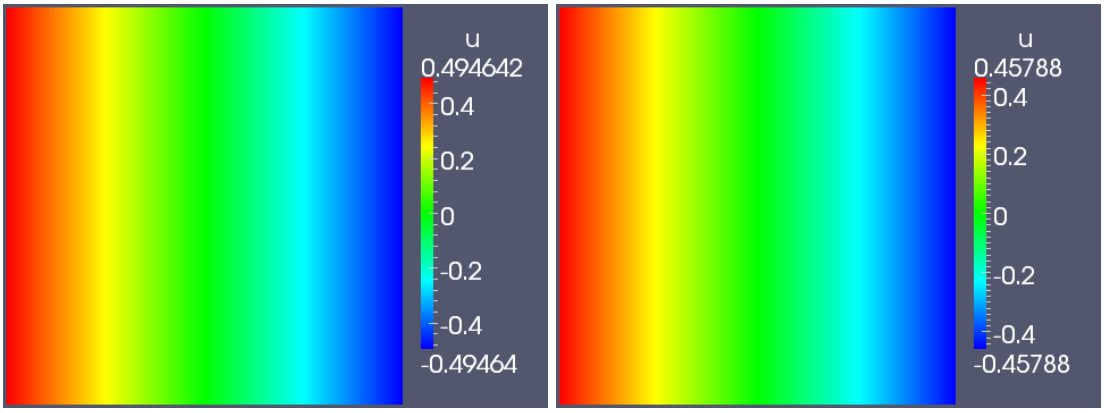
$$\|v,\tau\|_V^2 \coloneqq \alpha \|v\|_{L^2(\Omega)}^2 + \|\beta\nabla v\|_{L^2(\Omega)}^2 + \epsilon \|\nabla v\|_{L^2(\Omega)}^2 + \min\left\{\frac{1}{\epsilon}, \frac{1}{|K|}\right\} \|\tau\|_{L^2(\Omega)}^2 + \|\nabla\cdot\tau\|_{L^2(\Omega)}^2$$

where $\alpha \in [\epsilon, 1]$ was selected in such a way that optimal test functions over a single element did not contain boundary layers[1] which are difficult to approximate using our enriched space.

However, in [76], it was shown that the global test space made up of the union of local test spaces contains the test subspace of weakly conforming test functions that are the result of solving for optimal test functions globally using a weakly conforming enriched space (which we refer to as the *global test space*). Furthermore, the field solutions for the DPG method are dependent only upon the properties of the global test space.

In other words, the resolution of boundary layers that occur at element boundaries is not important unless these boundary layers appear in globally determined test functions too (boundary layers that appear on element boundaries but not at a global level can be considered a negligible side-effect of the "localization" of problems for optimal test functions).

[1] Boundary layers can appear in the cross-stream direction if α is not the same order as ϵ/h^2, where h is the element size. This is explained in more detail in [4]



(a) $\alpha = \epsilon$ (b) $\alpha = 1$

Figure 2.1: Optimal test functions for a $u = 1$.

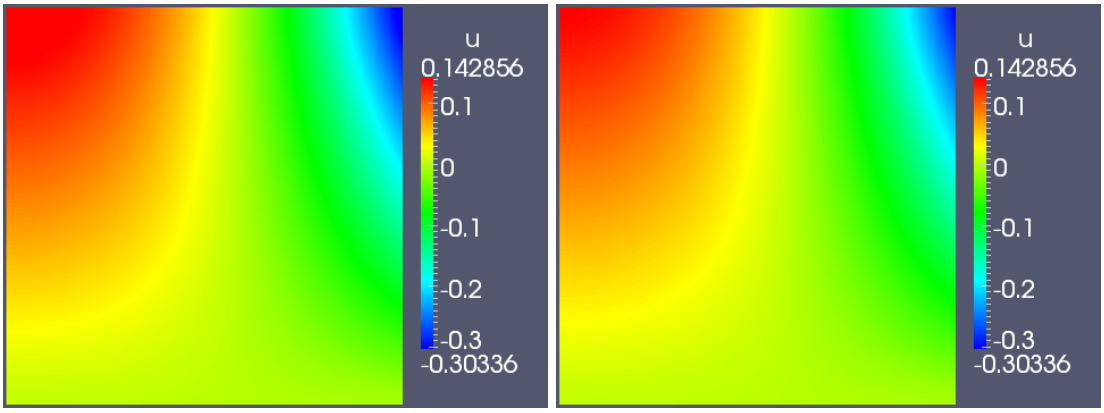
For these problems, we consider the Erikkson-Johnson model problem setup - the domain is a unit square in 2D, with $\beta = (1, 0)$. We computed global test functions corresponding to global field basis functions $1, xy, x(1-x)y(1-y)$ – a constant basis function, a bilinear basis function, and a quadratic bubble – as well as basis functions restricted to a small element in the middle of the domain for both $\alpha = \epsilon$ (where $\epsilon = .01$) and $\alpha = 1$. No boundary layers were observed in either case, and the test functions under both test norms are very similar.

2.2 Test norms for the convection-diffusion equation with first-order term

Very often, the convection-diffusion equation includes a first-order term, such that the form of the equation is

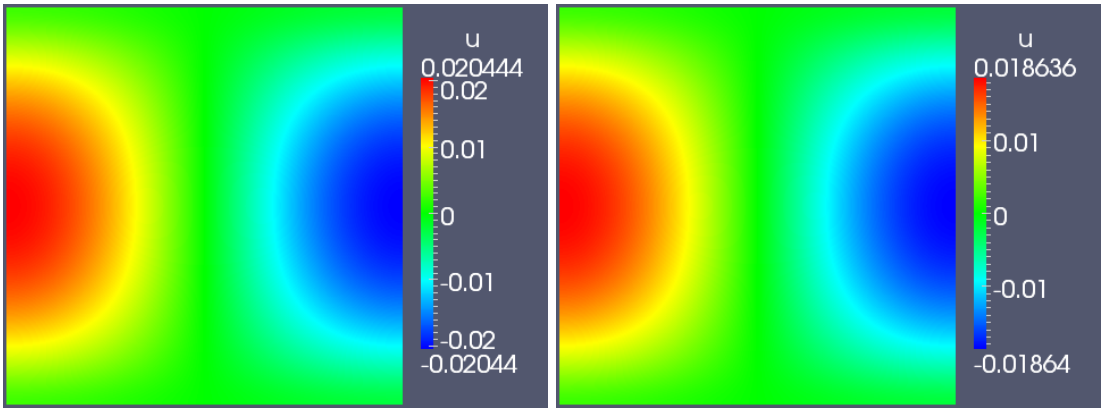
$$\nabla \cdot (\beta u - \epsilon \nabla u) + \alpha v = f$$

where α is some constant or function. This first-order term represents a reaction term, modeling production of the solvent u. Most commonly, however, this first order term appears in context of



(a) $\alpha = \epsilon$ (b) $\alpha = 1$

Figure 2.2: Optimal test functions for a $u = xy$.



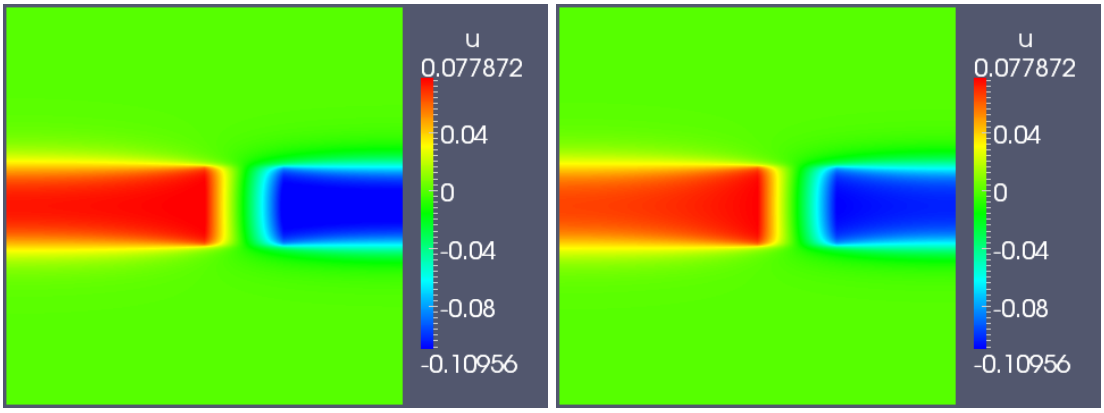
(a) $\alpha = \epsilon$ (b) $\alpha = 1$

Figure 2.3: Optimal test functions for a $u = x(1-x)y(1-y)$.

implicit time-stepping methods for the transient convection-diffusion equation

$$\frac{\partial u}{\partial t} + \nabla \cdot (\beta u - \epsilon \nabla u) = f.$$

For implicit time-stepping methods, we solve for the solution at the current timestep $u_{t_i} := u$ under some approximation of the time derivative: for example, implict Euler uses the first order



(a) $\alpha = \epsilon$ (b) $\alpha = 1$

Figure 2.4: Optimal test functions for a $u = 1|_K$, where $K = [.5, .7] \times [.4, .6]$.

approximation $\frac{\partial u}{\partial t} \approx \frac{u - u_{t_{i-1}}}{dt}$, introducing a reaction term u/dt

$$\frac{u}{dt} + \nabla \cdot (\beta u - \epsilon \nabla u) = f + \frac{u_{t_{i-1}}}{dt}.$$

Under this version of the convection-diffusion equation, we modify our test norm to match the magnitude of the first order term u/dt

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

Noting that, according to the previous numerical experiments, the magnitude of the $L^2(\Omega)$ term does not appear to create boundary layers in test functions for field variables. Thus, so long as the $\frac{1}{dt}$ is $O(h)$, the mesh size, we should expect (by a transformation to the unit element) optimal test functions to be locally resolvable using our enriched space.

Recall that proving robustness involves showing energy estimates on the following adjoint equation

$$\frac{1}{dt} v - \beta \cdot \nabla v - \epsilon \triangle v = u - \epsilon \nabla \cdot \sigma$$

where $u, \sigma \in L^2(\Omega)$ represent functions from the trial space.

Lemma 7. $\frac{1}{dt}\|v\|^2_{L^2(\Omega)} + \epsilon\|\nabla v\|^2_{L^2(\Omega)} \lesssim \|u\|^2_{L^2(\Omega)} + \|\sigma\|^2_{L^2(\Omega)}$.

Proof. Multiplying the equation by v and integrating gives

$$\int \frac{1}{dt}v^2 - \int \beta\cdot\nabla v v - \int \epsilon\triangle v v = \int uv - \int \epsilon\nabla\cdot\sigma v.$$

Integration by parts gives

$$\frac{1}{dt}\|v\|^2 + \int \frac{\nabla\cdot\beta}{2}v^2 - \int_\Gamma \frac{\beta_n}{2}v^2 + \epsilon\|\nabla v\|^2 - \epsilon\int_\Gamma v\frac{\partial v}{\partial n} = \int uv + \int \sigma\epsilon\nabla v - \epsilon\int_\Gamma \sigma_n v$$

Applying adjoint boundary conditions

$$v = 0, \quad \text{on } \Gamma_{\text{out}}$$
$$\frac{\partial v}{\partial n} = \sigma_n, \quad \text{on } \Gamma_{\text{in}}$$

reduces this to

$$\frac{1}{dt}\|v\|^2 + \int \frac{\nabla\cdot\beta}{2}v^2 + \int_{\Gamma_{\text{in}}} \frac{|\beta_n|}{2}v^2 + \epsilon\|\nabla v\|^2 = \int uv + \int \sigma\epsilon\nabla v.$$

An application of Young's inequality on the right hand side completes the proof. □

If we multiply by $\frac{1}{dt}v$ instead of v, we can derive a slightly different bound

Lemma 8. $\left\|\frac{1}{dt}v\right\|^2_{L^2(\Omega)} + \frac{\epsilon}{dt}\|\nabla v\|^2_{L^2(\Omega)} \lesssim \|u\|^2_{L^2(\Omega)} + \|\sigma\|^2_{L^2(\Omega)}$.

Proof. The proof is very similar to the above case. □

2.3 Error propagation in traces

In [3], both the graph norm and the constructed test norm (which we refer to as the *robust* test norm) are shown to satisfy the robust bound

$$\left(\|u\|^2_{L^2(\Omega)} + \|\sigma\|^2_{L^2(\Omega)} + \epsilon^2\|\widehat{u}\|_{H^{1/2}(\Gamma_h)} + \epsilon\left\|\widehat{f}_n\right\|_{H^{-1/2}(\Gamma_h)}\right)^{1/2} \lesssim \left\|\left(u,\sigma,\widehat{u},\widehat{f}_n\right)\right\|_E$$

where $\|\cdot\|_E$ is the energy norm in which DPG is optimal. The focus of this bound is the ϵ independence of the $L^2(\Omega)$ norms on u and σ; however, we have not addressed the issue of robustness of the traces $\widehat{u}$ and flux $\widehat{f}_n$ and how it manifests in practice.

We examine a test problem, with $\Omega = [0,1]^2$. Inflow conditions[2] are set such that

$$u \approx u - \epsilon \frac{\partial u}{\partial n} = 0, \quad x = 0, \quad y \in [0,1],$$

and wall boundary conditions are set to mimic a flat plate problem such that

$$u = 0, \quad y = 0, \quad x \in [.5,1].$$

Boundary conditions in the rest of the domain are set such that

$$\frac{\partial u}{\partial n} = 0, \quad y = 1,$$
$$\frac{\partial u}{\partial n} = 0, \quad y = 0, \quad x \in [0,.5].$$

We examine the traces along $y = 0$. Recall that traces are discretized as traces of H^1-conforming trial functions; thus, whereas we expect convergence in the $L^2(\Omega)$ norm for field variables, this is not true for the traces. In particular, we observe a Gibbs-type phenomena in the propagation of error for traces in Figure 2.5. Unlike the field variables, Gibbs-type overshoots and undershoots in the solution are not localized to a single element. This is due to the increased stencil size for traces, which are not locally supported over one element the way field and flux variables are.

2.4 Zero-mean scaling

In [3, 4], we proved the energy estimate

$$\|v\|^2_{L^2(\Omega)} \lesssim \|u\|^2_{L^2(\Omega)} + \|\sigma\|^2_{L^2(\Omega)}$$

[2] $u - \epsilon \frac{\partial u}{\partial n}$ approximates u for $\epsilon \ll 1$. These conditions are set such that we are able to use the robust test norm described in [4].

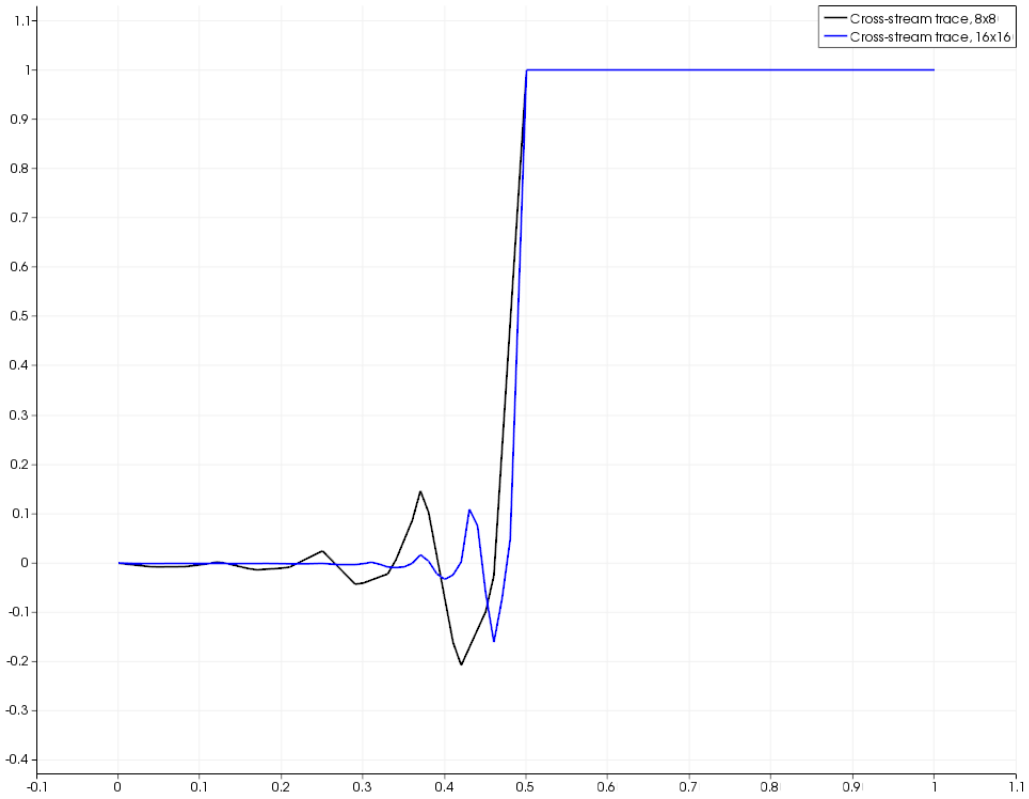


Figure 2.5: H^1-type propagation of error along $y = 0$.

which implied that including $\|v\|$ could be included in the test norm $\|v, \tau\|_V$ and produce a robust DPG method for convection-diffusion. However, it is also possible to replace the $L^2(\Omega)$ term with the first-order term

$$\frac{1}{h^2}\left|\int_K v\right|^2$$

which is a scaled measure of the average of v over an element. By Hölder's inequality, we have that

$$\left|\int_K v\right| \leq \left|\int_K v^2\right|^{\frac{1}{2}} \left|h^2\right|^{\frac{1}{2}},$$

implying that

$$\frac{1}{h^2}\left|\int_K v\right|^2 \leq \|v\|^2_{L^2(\Omega)}$$

For the locally conservative version of DPG, we were able to replace $\|v\|^2_{L^2(\Omega)}$ in the test norm with $\frac{1}{h^4}\left|\int_K v\right|^2$, because in context of local conservation (enforced by Lagrange multipliers), elementwise constants are already enforced to be in the test space, so the addition of the mean-squared term to the test norm is solely to enforce a scaling condition enforcing a zero-mean condition on the remainder of the test functions.

Bibliography

[1] 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.

[2] 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.

[3] L. Demkowicz and N. Heuer. Robust DPG method for convection-dominated diffusion problems. *SIAM J. Num. Anal*, 2013. accepted.

[4] J. Chan, N. Heuer, T. Bui Thanh, and L. Demkowicz. A robust DPG method for convection-dominated diffusion problems II: adjoint boundary conditions and mesh-dependent test norms. *Computers and Mathematics with Applications*, 2013. Accepted.

[5] 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.

[6] 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. see also ICES Report 9/16.

[7] L. Demkowicz, J. Gopalakrishnan, and A. Niemi. A class of discontinuous Petrov-Galerkin methods. Part III: Adaptivity. *Appl. Numer. Math.*, 62(4):396–427, April 2012. see also ICES Report 2010/1.

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

[9] J. Donea and A. Huerta. *Finite Element Methods for Flow Problems*. Wiley, 2003. Image from online exercises from online book.

[10] 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.

[11] P. M. Gresho and R. L. Lee. Don't suppress the wiggles - they're telling you something! *Computers and Fluids*, 9(2):223 – 253, 1981.

[12] 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.

[13] J.L. Guermond, R. Pasquetti, and B. Popov. Entropy viscosity method for nonlinear conservation laws. *Journal of Computational Physics*, 230(11):4248 – 4267, 2011. Special issue High Order Methods for CFD Problems.

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

[15] 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.

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

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

[18] 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.

[19] 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.

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

[21] P. Lasaint and P. A. Raviart. On a finite element method for solving the neutron transport equation. *Mathematical aspects of finite elements in partial differential equations*, pages 89–123, 1974. Proceedings of a symposium conducted by Math. Res. Center, Univ. of Wisconsin-Madison.

[22] C. Johnson and J. Pitkaranta. An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. *Math. Comp.*, (46):1–26, 1986.

[23] 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.

[24] 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.

[25] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. part I: The transport equation. *Computer Methods in Applied Mechanics and Engineering*, 199(23-24):1558 – 1572, 2010. see also ICES Report 2009-12.

[26] 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.

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

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

[29] 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, 2008.

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

[31] 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.

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

[33] 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.

[34] 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.

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

[36] 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. Accepted.

[37] 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.

[38] A. Cohen, W. Dahmen, and G. Welper. Adaptivity and variational stabilization for convection-diffusion equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 46(5):1247–1273, 2012.

[39] A. H. Niemi, J. A. Bramwell, and L. 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.

[40] A. Niemi, N. Collier, and V. Calo. Discontinuous Petrov-Galerkin method based on the optimal test space norm for one-dimensional transport problems. *Journal of Computational Science*, 2011. In press.

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

[42] T. Bui-Thanh, L. Demkowicz, and O. Ghattas. A unified discontinuous Petrov-Galerkin Method and its analysis for Friedrichs' systems. Technical Report 34, ICES, 2011. SIAM J. Num. Anal., revised version submitted.

[43] L. Demkowicz and J. Gopalakrishnan. An Overview of the DPG Method. Technical Report 13-02, ICES, January 2013.

[44] A. Logg, K. A. Mardal, G. N. Wells, et al. *Automated Solution of Differential Equations by the Finite Element Method*. Springer, 2012.

[45] M. Stynes. Convection-diffusion problems, SDFEM/SUPG and a priori meshes. *Int. J. Comput. Sci. Math.*, 1(2-4):412–431, January 2007.

[46] 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.

[47] V. Calo A. Niemi, N. Collier. Automatically Stable Discontinuous Petrov-Galerkin Methods for Stationary Transport Problems: Quasi-Optimal Test Space Norm. arXiv:1201.1847 [math.NA], submitted to CAMWA, January 2012.

[48] C. Schwab and M. Suri. The p and hp versions of the finite element method for problems with boundary layers. *Math. Comput.*, 65(216):1403–1429, October 1996.

[49] 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.

[50] 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.

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

[52] J. F. Gerbeau, C. Le Bris, and M Bercovier. Spurious velocities in the steady flow of an incompressible fluid subjected to external forces. *International Journal for Numerical Methods in Fluids*, 25(6):679–695, 1997.

[53] T. Ellis, J. Chan, N. Roberts, and L. Demkowicz. Element conservation properties in the DPG method. Technical report, ICES, July 2013. In preparation.

[54] I. Babuska and B.Q. Guo. Approximation properties of the h-p version of the finite element method. *Computer Methods in Applied Mechanics and Engineering*, 133(3-4):319 – 346, 1996.

[55] W. Rachowicz, J. T. Oden, and L. Demkowicz. Toward a universal h-p adaptive finite element strategy part 3. Design of h-p meshes. *Computer Methods in Applied Mechanics and Engineering*, 77(1-2):181 – 212, 1989.

[56] 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.

[57] M. D. Gunzburger. *Finite Element Methods for Viscous Incompressible Flows: A Guide to Theory, Practice, and Algorithms.* Computer Science and Scientific Computing. Elsevier Science, 1989.

[58] D. Moro-Ludeña, 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.

[59] J. Nocedal and S. J. Wright. *Numerical Optimization.* Springer, New York, 2nd edition, 2006.

[60] L. Demkowicz, J.T. Oden, and W. Rachowicz. A new finite element method for solving compressible Navier-Stokes equations based on an operator splitting method and h-p adaptivity. *Computer Methods in Applied Mechanics and Engineering*, 84(3):275 – 326, 1990.

[61] B. Kirk. *Adaptive Finite Element Simulation of Flow and Transport Applications on Parallel Computers.* PhD thesis in Computational and Applied Mathematics, The University of Texas at Austin, 2009.

[62] K. Devine, E. Boman, R. Heaphy, B. Hendrickson, and C. Vaughan. Zoltan data management services for parallel dynamic applications. *Computing in Science and Engineering*, 4(2):90–97, 2002.

[63] W.F. Mitchell. A refinement-tree based partitioning method for dynamic load balancing with adaptively refined grids. *Journal of Parallel and Distributed Computing*, 67:417–429, 2007.

[64] J. E. Carter. Numerical solutions of the supersonic, laminar flow over a two-dimensional compression corner. Technical Report TR R-385, NASA, July 1972.

[65] 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.

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

[67] J. Chan, L. Demkowicz, N. Roberts, and R. Moser. A new Discontinuous Petrov-Galerkin method with optimal test functions. part v: Solution of 1d Burgers and Navier-Stokes equations. Technical Report 10-25, ICES, June 2010.

[68] F. Shakib, T. J.R. Hughes, and Z. Johan. A new finite element formulation for computational fluid dynamics: X. the compressible Euler and Navier-Stokes equations. *Computer Methods in Applied Mechanics and Engineering*, 89(1 - 3):141 – 219, 1991. Second World Congress on Computational Mechanics.

[69] W. Rachowicz. An anisotropic h-adaptive finite element method for compressible Navier-Stokes equations. *Computer Methods in Applied Mechanics and Engineering*, 146(3-4):231 – 252, 1997.

[70] J. Chan, L. Demkowicz, and M. Shashkov. Space-time DPG for shock problems. Technical Report LA-UR 11-05511, LANL, September 2011.

[71] J. Chan, L. Demkowicz, R. Moser, and N. Roberts. A New Discontinuous Petrov-Galerkin Method with Optimal Test Functions. Part V: Solution of 1D Burgers' and Navier-Stokes Equations. Technical Report 10-25, ICES, June 2010.

[72] T.J.R. Hughes, L.P. Franca, and M. Mallet. A new finite element formulation for computational fluid dynamics: I. symmetric forms of the compressible Euler and Navier-Stokes equations and

the second law of thermodynamics. *Computer Methods in Applied Mechanics and Engineering*, 54(2):223 – 234, 1986.

[73] D. Broerson and R. Stevenson. A Petrov-Galerkin discretization with optimal test space of a mild-weak formulation of convection-diffusion equations in mixed form. Technical report, Korteweg-de Vries Institute for Mathematics, November 2012. Submitted.

[74] L. Demkowicz and J. Gopalakrishnan. A primal DPG method without a first order reformulation. *Computers and Mathematics with Applications*, 2013. To appear.

[75] J. Chan and J. Evans. A minimum residual finite element method for the convection-dominated diffusion equation. Technical Report 13-12, ICES, May 2013. Submitted to SIAM J. Comp. Sci.

[76] J. Chan, J. Gopalakrishnan, and L. Demkowicz. Global properties of DPG test spaces for convection-diffusion problems. Technical Report 13-05, ICES, February 2013.