UNIVERSITY OF CALIFORNIA SAN DIEGO

A Study on Discontinuous Petrov-Galerkin Finite Element Methods in Semi-linear Problems and Adaptivity

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Mathematics with a Specialization in Computational Science

by

Jor-el Thomas Caparas Briones

Committee in charge:

Professor Michael Holst, Chair
Professor Randolph Bank
Professor Julius Kuti
Professor Melvin Leok

2022

The Dissertation of Jor-el Thomas Caparas Briones is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

For the CotS. And for every version of Me, past, present, and future. We made it!

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Michael Holst for his support as the chair of my committee. Over the past eight years, he has been incredibly patient and understanding with all of my struggles throughout this writing process. I am of the belief that such patience exceeds that of saints. He has guided me gently throughout this whole process, passing down his wisdom not only in mathematics and the academic processes that I needed to navigate, but in life. I honestly could not have had a better professor or a better person as an advisor for my degree and I'm glad I wound up here.

I would also like to thank a number of people, without whom, I would not have gotten this far. The techs, the admin staff, and the doctors I worked with at the Jules Stein Eye Institute at UCLA have done more for me than they will ever know and pulled me out of a very dark place, and inspired me to pursue a PhD in the first place. My best friends, the Engineer (My-Quan), The Doctor Doctor (Eileen), and Astrid (Christina), have been unceasingly present and kept me afloat, lending me their wisdom, their ears, and even their presence from far away. Lysse and Laura pivoted me to a much healthier direction that rippled out through to my mental health, my PhD work, and beyond, and arguably I owe them both my life. I'd like to thank a dork (Shannon) for being a calming presence throughout the last year of this journey and a good friend to offer local stability in little ways that ended up meaning a lot. I'd like to thank the San Diego Pokemon Go community, who inspired me to travel across the land, searching far and wide, for mathematics, and more, and for being an unwitting source of comfort and support during the most grueling years. And I would like to thank the Girl's Chat (Love is Blind?) for being an incredible circle of support and encouragement, especially in those final moments where I nervously fretted over my defense.

The people I met over the past year, the people I have met up to that point, and the communities I ended up a part of, have showered me with so much love and encouragement, and have propelled me forward past the finish line and toward a bright future. If I were to list every contribution from every person, it would far exceed the length of my dissertation. It is an incredible feeling to be showered with so much support from so many, and so I give them my acknowledgment and eternal gratitude.

VITA

2013	Bachelor of Arts, University of California Los Angeles
2015–2022	Teaching Assistant, Department of Mathematics University of California San Diego
2016–2021	Teaching Assistant, Department of Computer Science and Engineering University of California San Diego
2016	Master of Arts, University of California San Diego
2022	Doctor of Philosophy, University of California San Diego

FIELDS OF STUDY

Major Field: Mathematics (Computational Science)

Studies in Mathematics
Professors Michael Holst and Randolph Bank

ABSTRACT OF THE DISSERTATION

A Study on Discontinuous Petrov-Galerkin Finite Element Methods in Semi-linear Problems and Adaptivity

by

Jor-el Thomas Caparas Briones

Doctor of Philosophy in Mathematics with a Specialization in Computational Science

University of California San Diego, 2022

Professor Michael Holst, Chair

In numerical analysis, finite element methods are a method of approximating solutions to differential equations on a domain. In such methods, the solution function is approximated by partitioning the domain into a mesh of elements, and testing candidate functions in a discrete trial space on that mesh against a discrete space of test functions. We explore certain classes of finite element methods called discontinuous Petrov-Galerkin (DPG) finite element methods, where the test space functions are allowed to be discontinuous across elements, and test spaces are selected specifically to optimize stability.

Because we are concerned with the accuracy of our approximation, we place focus on how the error behaves in DPG methods. We explore how DPG methods in semi-linear problems, as well as how DPG problems can interact with adaptive methods, a different framework for finite element methods. In addition, we establish some results about the error of DPG approximations, particularly the error using the subspace dual norms that arise from the construction of the test spaces.

Chapter 1

Introduction

We often model real world phenomena with differential equations, since behaviors of phenomena often change over time and space. In solving differential equations, we can simulate these phenomena in specific settings or over some period of time. In mathematics, there are often modeling equations whose exact solutions cannot be obtained analytically, even if it is known that solutions exist. In such situations, we often attempt to approximate solutions numerically, to be able to better understand the behavior of the physical phenomena that the equations themselves model. To that end, different numerical methods were developed to approximate solutions to these equations. In particular, this paper focuses on the specific aspects of one such method known as the finite element method.

1.1 Weak Formulation for Finite Element Methods

In the most general case, we consider the problem in which given some differential equation F on some domain Ω, contained in some space X, and some forcing function f. Our goal is to find a function $x \in X$ that solves the differential equation:

$$F(x) = f \tag{1.1}$$

Often, the original problem (1.1), known as the strong formulation, is intractable, or it cannot be solved analytically, so we look toward approximating the solution using the weak formulation of the problem. While we may not be able to find a solution directly, we can instead create some approximation by comparing the differential equation to some test space Y. So, we compare the residual $F(x) - f$ against some test space Y. In other words, we try to find a function $x \in X$ such that, for all $y \in Y$:

$$\langle F(x) - f, y \rangle = 0 \tag{1.2}$$

Under certain conditions, the solution x_w to the weak formulation (1.2) would coincide with the solution to the strong formulation (1.1). The weak form is an analogue to dot products in inner product spaces: If the inner product of some element $x = F(x_w) - f$ and every other element y in the space is 0, then x must itself be 0, and therefore $F(x) = f$. However, in general, the comparison of the residual $F(x) - f$ and every element of some space Y is not guaranteed to be an inner product. It could be the case that neither the test nor trial space is an inner product space, or that the norm of interest is not the inner product norm, or that the test and trial spaces themselves are not equal. Nevertheless, the idea is preserved for these methods.

In many applications, we may use integration to compare the residual to the test space. This involves first integrating the equation against functions y in a test space Y as follows:

$$\int F(x)y = \int fy \tag{1.3}$$

Our goal is then to find some function $x \in X$ for which the equation above holds for all $y \in Y$. For finite-dimensional spaces and under some other assumptions, the function that solves this system would also solve the strong formulation. Otherwise, the solution to this weak formulation (1.3) serves as an approximation to the solution to the original problem (1.1). In this paper, we are largely concerned with linear and semi-linear problems, in which the weak formulation can be reduced to forms that are linear in Y, and contain some part that is linear in X.

1.2 Discrete Weak Formulation and Convergence of FEM in the Linear Problem

For linear problems, we integrate the residual $F(x) - f$ against a test space Y, as in (1.3). The differential equation is then reduced to a bilinear form that is linear in both the X and Y parts, and a forcing function that acts as an element of the dual of the test space Y.Then, the weak formulation for the linear problem is as follows: For some bilinear form $a(\cdot, \cdot) : X \times Y \to \mathbb{R}$, and some forcing functional $f \in Y'$, we find some $x \in X$ such that

$$a(x, y) = f(y) \quad \forall y \in Y \tag{1.4}$$

Because the trial and test spaces themselves are often infinite-dimensional, we cannot solve even this weak form exactly through numerical methods. So instead, we approximate solutions using finite-dimensional subspaces of the trial and test spaces.

We have the following similar weak formulations using the finite-dimensional subspaces for our trial and test spaces. We have discrete subspaces X_h and Y_h of our trial space X and our test space Y, respectively. For some bilinear form $a(\cdot,\cdot) : X_h \times Y_h \to \mathbb{R}$, and some functional $f \in Y_h'$, we find some $x_h \in X_h$ such that

$$a(x_h, y_h) = f(y_h) \qquad \forall y_h \in Y_h \tag{1.5}$$

With the formulation in place, one must define the functions on the test and trial spaces to approximate the solution function. This is cumbersome on large, or strangely shaped domains, and there would not be an easy way to define said functions over arbitrary domains. In order to have a methodical approach to approximating the solution function we turn to finite element methods. In finite element methods, we break the problem domain Ω into a mesh of elements, onto which we define functions. In doing so we can have some kind of partition for a number of domains, and we can approach each domain with the same method.

We can, for example, split the domain into a uniform mesh of triangular elements, and define a basis of linear functions on each element. For each element K, we define a set of basis functions, often comparable for each element. Then, we attempt to find, using some algorithm built to solve the problem, the best approximation to the solution. The resulting function would approximate the whole solution, and would be the result of combining the approximations made on each element.

Because the discrete function spaces are finite-dimensional and the discrete weak formulation relies on a bilinear form, we can solve the weak formulation by solving a linear system. We take X_n and Y_n to be finite dimensional subspaces of dimension n of the trial space X and the test space Y, respectively. We let X_n have basis $\{\phi_i\}$ and Y_n have basis $\{\psi_i\}$. Then the problem (1.5), in which we find an approximation x_h, reduces to solving the well-posed matrix problem:

$$A\alpha = b, \quad A \in \mathbb{R}^{n\times n}, \quad \alpha, b \in \mathbb{R}^n, \quad x_h = \sum_{i=1}^{n} \alpha_i \phi_i(x).$$

We solve the system to find a vector of coefficients α_i for our basis functions ϕ_i, to construct our solution function x_h. The matrix A is known as a stiffness matrix, where $A_{ij} = a(\phi_i, \psi_j)$, while b corresponds to the forcing function f and $b_j = f(\psi_j)$. In practice, linear solvers would be used to solve the resulting linear system and generate the Galerkin solution function.

For classical finite element methods, the solution functions $x_{h,K}$ on each element must agree with the functions of adjacent elements along any shared boundaries or corners, thus the resulting approximate

solutions are continuous, which affords them the advantages of continuity. As will be discussed in chapter 2, it may be the case that the solution functions do not agree on the element boundaries, and in particular for discontinuous finite element methods.

Once the solution function is constructed, we use tools to estimate the error of the approximate solution function against the true solution for the weak formulation. If the error of the approximate solution function is too large, then the test and trial spaces are refined, either by refining the mesh into yet smaller elements, or enriching the function space in which the functions reside with more basis functions. The hope is that with a more refined function space or a finer mesh, parts of the solution that may not be easily approximated–such as discontinuities, rapid oscillations, or other non-smoothness–may be approximated more accurately. This process is repeated until the error is within some tolerance, and the result is an approximate solution to the weak form of the differential equation.

Classical finite element methods relied on the following assumptions about the problem:

(A1) The bilinear form is continuous, i.e. for some constant $M > 0$,

$$|a(x, y)| \leqslant M\|x\|\|y\|$$

(A2) The bilinear form is coercive, i.e., for some constant $m > 0$

$$|a(x, x)| \geqslant m\|x\|^2$$

(A3) For all $y_h \in Y_h$, we have for our approximate solution x_h and the exact solution x, that:

$$a(x_h, y_h) = a(x, y_h)$$

which results in Galerkin orthogonality, i.e.

$$a(x - x_h, y_h) = 0$$

In particular, the second assumption of coercivity (A2), along with the forcing function f being in Y', are sufficient conditions for the problem to be well posed, in that a solution that exists must be unique. We reprove this result below:

Theorem [Coercivity Implies Uniqueness]. If a bilinear form is coercive (A2) in the discrete weak

formulation above (1.5), then any Galerkin solution to the system must be unique.

Proof. Suppose x_1 and x_2 are both Galerkin solutions for the discrete weak formulation (1.5):

$$a(x_h, y_h) = f(y_h) \qquad \forall y_h \in Y_h$$

In other words, we assume that for all $y_h \in Y_h$

$$a(x_1, y_h) = a(x_2, y_h) = f(y_h)$$

Then we have, for all $y_h \in Y_h$:

$$a(x_1 - x_2, y_h) = a(x_1, y_h) - a(x_2, y_h) = f(y_h) - f(y_h) = 0$$

But $x_1 - x_2 \in Y_h$ since the discrete trial and test spaces are equal. And therefore, by coercivity (A2) and the above:

$$0 = |a(x_1 - x_2, x_1 - x_2)| \geqslant m\|x_1 - x_2\|^2 \implies x_1 = x_2 \quad a.e.$$

The uniqueness of the solution follows. □

The well-posedness of this problem, including the uniqueness of the solution, follows from the coercivity condition with the restriction on the forcing function f. this result is better known as the Lax-Milgram Theorem [11].

Also, from the assumptions (A1)–(A3), we establish a very important quasi-optimality result for finite element methods, known as Cea's Lemma. In this result, reproved below, we establish that the approximate solution found with the finite element method is within some scalar factor of the best approximate solution within the discrete space.

Lemma [Cea's Lemma]. Given the three assumptions (A1)-(A3) above, there exists some constant $C > 0$ such that, for the Galerkin solution x_h to the discrete weak formulation (1.5), and for the exact solution x for the weak formulation (1.4):

$$\|x - x_h\| \leqslant C \inf_{\xi_h \in X_h} \|x - \xi_h\|$$

Proof.

$$
\begin{aligned}
\|x - x_h\|^2 &\leqslant \frac{1}{m}|a(x - x_h, x - x_h)| && \text{(by coercivity (A2))} \\
&= \frac{1}{m}|a(x - x_h, x - \xi_h) + a(x - x_h, \xi_h - x_h)| && \text{(Due to linearity of the bilinear form)} \\
&= \frac{1}{m}|a(x - x_h, x - \xi_h)| && \text{(Due to Galerkin orthogonality (A3))} \\
&\leqslant \frac{M}{m}\|x - x_h\|\|x - \xi_h\| && \text{(by continuity of bilinear form (A1))} \\
&\implies \|x - x_h\| \leqslant \frac{M}{m}\|x - \xi_h\|
\end{aligned}
$$

Letting $C = \frac{M}{m}$, and taking the inf over all $\xi_h \in Y_h$, we establish the desired result. □

Or in other words, the Galerkin approximation found by this method is no worse than the best approximation multiplied by some factor, that is ideally small. If we refine the discrete subspaces so that the error between the weak solution and best approximation in the discrete subspace goes toward 0, then the error between the discrete weak Galerkin solution and the weak solution must also go to 0. Hence, if the weak solution x is approximable, and if we can refine the space indefinitely to the point that the best approximation error converges to 0, then we guarantee that the error from the Galerkin approximation may also converge to 0. We show these results with an example below.

The classical model problem in finite element methods is the Poisson equation. For some forcing function f, and over some domain Ω, we find some function $x \in X$ such that:

$$-\Delta x = f$$

Converting this to the weak formulation, we have, for some test space Y, and for all $y \in Y$:

$$\int_\Omega -y\Delta x = \int_\Omega fy$$

And integrating by parts and with appropriate boundary conditions on the boundary of the domain $\delta\Omega$, such as Dirichlet boundary conditions:

$$\int_\Omega \nabla x \cdot \nabla y = \int_\Omega fy$$

We take $f(y) = \int_\Omega fy$, and $a(x, y) = \int_\Omega \nabla x \cdot \nabla y$, which we call an energy inner product, which

induces an energy norm for the finite element method, with the bilinear form a, namely:

$$|||x||| = \sqrt{a(x,x)}$$

If X and Y were both $H^1(\Omega)$, we have that:

$$|a(x,x)| = \|\nabla x\|_{L^2}^2$$

And in particular, by the Poincare inequality and by definition of the H^1 norm, for some constants $C_1, C_2 > 0$:

$$C_1\|x\|_{H^1}^2 \leqslant \|\nabla x\|_{L^2}^2 \leqslant C_2\|x\|_{H^1}^2$$

In other words, for this specific case, the energy norm is equivalent to the H^1 norm, and error estimation in the energy norm can be seen as comparable to a error estimation in a standard error norm. In particular, for equivalent norms $\|\cdot\|_{N_1}$ and $\|\cdot\|_{N_2}$ we have that convergence of the error in one norm implies convergence of the error in the other:

$$\|x - x_h\|_{N_1} \to 0 \iff \|x - x_h\|_{N_2} \to 0$$

However, there are limitations to the classical method that have led others to alter the method by weakening the initial assumptions of the method. Solutions may belong to a space that is different than the test space, and are better approximated with a different space. If the solution is not an $H^1(\Omega)$ function, for example, then $H^1(\Omega)$ may not be an appropriate choice for the trial space, but it may be useful to exploit the fact that $H^1(\Omega)$ is an inner product space by using it as a test space. Even if one could in theory indefinitely refine the mesh to get an accurate approximation, the number of elements in a uniformly refined mesh increases exponentially. For practical uses, the computational power required to perform computations on the mesh is a limiting factor to how much the mesh can be refined, and therefore, how accurate a numerical solution would be. And in most cases, real world applications would rely on differential equations that result in a non-linear form, rather than a bilinear form, and addressing non-linear problems is a widely researched topic for finite element methods [13].

Convergence of Finite Element Methods

We are interested as well in the convergence of finite element methods, and in particular that the error of our approximation tends to 0 as we continue refining the mesh.

One way we can do this is to exploit orthogonality or quasi-orthogonality in a symmetric bilinear form. While bilinear forms are not necessarily symmetric, as they are in the example of the Poisson equation above, we consider the following. Should the bilinear form be symmetric, and if we had a classical finite element method, in which continuity, coercivity, and Galerkin orthogonality (A1)-(A3) are maintained, we can establish contraction if we also have the following:

$$\|x_h - x_{h+1}\|^2 \geqslant C\|x - x_h\|^2$$

For some constant $C \in (0, 1)$. In other words, the true error between the Galerkin solution and the regular weak solution is bounded above by some factor times the distance between two successive solutions.

Proof. By Galerkin Orthogonality (A3) and the symmetry of the bilinear form, we have:

$$\begin{aligned}
&\|\!|x - x_{h+1}|\!\|^2 + \|\!|x_{h+1} - x_h|\!\|^2 = a(x - x_{h+1}, x - x_{h+1}) + a(x_{h+1} - x_h, x_{h+1} - x_h) \\
&= a(x - x_{h+1}, x - x_{h+1}) + a(x_{h+1} - x_h, x_{h+1} - x_h) \\
&\quad + a(x - x_{h+1}, x_{h+1} - x_h) + a(x - x_{h+1}, x_{h+1} - x_h) \\
&= a(x - x_{h+1}, x - x_{h+1}) + a(x_{h+1} - x_h, x_{h+1} - x_h) \\
&\quad + a(x_{h+1} - x_h, x - x_{h+1}) + a(x - x_{h+1}, x_{h+1} - x_h) \\
&= a(x - x_h, x - x_h) = \|\!|x - x_h|\!\|^2
\end{aligned}$$

Then we have, from the extra bound on the error, and the above Pythagorean-esque identity:

$$\|\!|x - x_{h+1}|\!\|^2 = \|\!|x - x_h|\!\|^2 - \|\!|x_{h+1} - x_h|\!\|^2 \leqslant \|\!|x - x_h|\!\|^2 - C\|\!|x - x_h|\!\|^2 = (1 - C)\|\!|x - x_h|\!\|^2$$

Since $0 < 1 - C < 1$, the error in each successive iteration is smaller than the error of the previous iteration, by a factor of $\sqrt{1 - C}$ and the error contracts toward 0. $\square$

1.3 Variations on the classical method

As problems become more challenging, we turn to weakening our assumptions to broaden the scope of which classes of problems we can approach. For the rest of the chapter, we focus on doing so in a number of ways. To address cases where the trial space needs to be of broader scope than the test space, we turn to mixed methods, also known as Petrov-Galerkin methods, in which the trial and test space differ. To address cases in which indefinite uniform refinement becomes computationally expensive, we turn to adaptive finite element methods. For adaptive finite elements, the mesh is only refined in regions of the domain where the error is estimated to have a higher error, whereas regions where error is estimated to be low are only refined to maintain mesh regularity. And finally, to address cases in which the discrete weak formulation does not result in a bilinear form, we turn to nonlinear cases that semi-linear. In such cases, there is still a bilinear form, but the discrete weak formulation has some non-linear part that needs to be controlled.

Petrov-Galerkin Finite Element Methods

In one such variation, we find that solutions to the problem may not exist in our test space, so we expand our trial space to include more functions. In such methods, known as Petrov-Galerkin methods, the trial and test spaces differ, and we can no longer assume coercivity. Instead, we have a new assumption that offers a similar stability result,- known as the inf-sup condition. For our bilinear form a, in lieu of coercivity, we assume that a satisfies the continuous inf-sup condition and the discrete inf-sup condition, i.e.:

(A2-1) Continuous inf-sup condition: For the trial and test spaces X and Y, and for some constant $\beta > 0$,

$$0 < \beta \leqslant \inf_{x \in X \setminus 0} \sup_{y \in Y \setminus 0} \frac{a(x,y)}{\|x\|\|y\|}$$

(A2-2) Discrete inf-sup condition: For discrete trial and test spaces X_h and Y_H, and for some constant $\beta_h > 0$,

$$0 < \beta \leqslant \inf_{x_h \in X_h \setminus 0} \sup_{y_h \in Y_h \setminus 0} \frac{a(x_h,y_h)}{\|x_h\|\|y_h\|} = inf_{y_h \in Y_h \setminus 0} \sup_{x_h \in X_h \setminus 0} \frac{a(x_h,y_h)}{\|x_h\|\|y_h\|}$$

In particularly, we assume that the bilinear form satisfies both conditions (A2-1) and (A2-2) with a shared positive constant, which will be referred to as β. Much like the coercivity assumption in classic finite element methods, the inf-sup condition implies that any Petrov-Galerkin solution to the weak formulation must be unique. The result is proven below:

Theorem [Inf-Sup Implies Uniqueness]. If a bilinear form satisfies the discrete inf-sup condition (A2-2) in the discrete weak formulation (1.5) above, then any Galerkin solution to the discrete weak formulation must be unique.

Proof. Suppose x_1 and x_2 are both Galerkin solutions for the discrete weak formulation (1.5):

$$a(x_h, y_h) = f(y_h) \qquad \forall y_h \in Y_h$$

In other words, we assume that for all $y_h = in Y_h$

$$a(x_1, y_h) = a(x_2, y_h) = f(y_h)$$

Then we have, for all $y_h \in Y_h$:

$$a(x_1 - x_2, y_h) = a(x_1, y_h) - a(x_2, y_h) = f(y_h) - f(y_h) = 0$$

By our inf-sup condition (A2-2) and Galerkin orthogonality (A3), since $x_1 - x_2 \in X_h$, and if $x_1 - x_2 \neq 0$:

$$0 < \beta \leqslant \inf_{x_h \in X_h \backslash 0} \sup_{y_h \in Y_h \backslash 0} \frac{a(x_h, y_h)}{\|x_h\| \|y_h\|} \leqslant \sup_{y_h \in Y_h \backslash 0} \frac{a(x_1 - x_2, y_h)}{\|x_1 - x_2\| \|y_h\|} = 0$$

This is a contradiction, and hence it must be the case that $x_1 = x_2$. The uniqueness of the solution follows. □

Much like in the classical case, where the test and trial spaces are the same, it is of great interest to show that the same quasi-optimality result of Cea's Lemma holds, a generalization known as the Ladyzhenskaya-Babuška-Brezzi (LBB) Theorem. Rather than assume that the bilinear form is coercive, which is not defined for a bilinear form where the trial and test spaces are different, we assume that the continuous and discrete inf-sup conditions (A2-1) and (A2-2) hold.

Lemma [Ladyzhenskaya-Babuška-Brezzi (LBB) Theorem]. If the assumptions (A1), (A3), (A2-1), and (A2-2) are met, then the Galerkin solution to the discrete weak formulation (1.5) satisfies the following quasi-optimality estimate:

$$\|x - x_h\| \leqslant C \inf_{\xi_h \in X} \|x - \xi_h\|$$

for some constant $C > 0$. A short proof of this lemma is provided below:

Proof.

$$\begin{aligned}
\|x - x_h\| &\leqslant \|x - \xi_h\| + \|\xi_h - x_h\| \\
&\leqslant \|x - \xi_h\| + \frac{1}{\beta} \sup_{y_h \in Y_h, \|y_h\|_Y = 1} a(\xi_h - x_h, y_h) && \text{(Due to inf-sup stability (A2-2))} \\
&= \|x - \xi_h\| + \frac{1}{\beta} \sup_{y_h \in Y_h, \|y_h\|_Y = 1} a(\xi_h - x, y_h) && \text{(Due to Galerkin orthogonality (A3))} \\
&\leqslant \|x - \xi_h\| + \frac{M}{\beta} \|x - \xi_h\| && \text{(Due to continuity of bilinear form (A1))} \\
&= \left(1 + \frac{M}{\beta}\right) \|x - \xi_h\|
\end{aligned}$$

Since ξ_h was chosen arbitrarily, we can take the inf over all ξ_h and establish quasi-optimality:

$$\|x - x_h\| \leqslant \left(1 + \frac{M}{\beta}\right) \inf_{\xi_h \in X} \|x - \xi_h\|$$

□

In a paper by Jinchao Xu and Ludmil Zikataov , it was proven that if X is a Hilbert space, then the stability constant can be sharpened even further, and we can achieve [16]:

$$\|x - x_h\| \leqslant \frac{M}{\beta} \inf_{\xi_h \in X} \|x - \xi_h\|$$

Discrete Weak Formulation Semi-Linear Problem

Because many problems are nonlinear, there is great interest in solving nonlinear PDE's. One such solution is to linearize such differential equations, and then to solve the resulting discrete weak formulation with bilinear form. However, the linearization does not remove any of the issues posed by the nonlinear problem, and instead inherits them. It is therefore useful to consider nonlinear problems without linearization, and what problems can be solved when involving a nonlinear part. Given certain assumptions, it may be the case that arguments to address the non-linear parts of the problem can be addressed separately from linear parts, using useful properties of the non-linearity.

The weak formulation for the semi-linear problem is similar to the linear one, but includes a term that is not linear in X, but is linear in Y. That weak formulation is as follows: For some bilinear form $a(\cdot, \cdot) : X \times Y \to \mathbb{R}$, some nonlinear form $(b(\cdot), \cdot) : X \times Y \to \mathbb{R}$ that is only linear in Y, and some

functional $f \in Y'$, we find some $x \in X$ such that

$$(c(x), y) := a(x, y) + (b(x), y) = f(y) \qquad \forall y \in Y \tag{1.6}$$

Much like in the linear problem (1.4), we deal with finite-dimensional subspaces X_h and Y_h of the trial and test spaces, and discretize the formulation. The weak formulation for the discrete semi-linear problem is as follows. We have discrete subspaces X_h and Y_h of our trial space X and our test space Y, respectively. For some bilinear form $a(\cdot, \cdot) : X_h \times Y_h \to \mathbb{R}$, some nonlinear form $(b(\cdot), \cdot) : X_h \times Y_h \to \mathbb{R}$ that is only linear in Y_h, and some functional $f \in Y_h'$, we find some $x_h \in X_h$ such that

$$(c(x_h), y_h) = a(x_h, y_h) + (b(x_h), y_h) = f(y_h) \qquad \forall y_h \in Y \tag{1.7}$$

When dealing with semi-linear problems, it is often the case that nonlinear parts can be treated separately from the linear parts. If, for example, the nonlinear part obeyed some kind of controllable property, we can bound the error similarly to how we would bound it in the linear case.

In one such case, if the nonlinear part obeyed some sort of maximum principle, we can use some kind of a priori bounds to bound the error [5, 14]. First, as above, we can break the weak form into a linear and nonlinear part, with the linear part obeying the assumptions, as above. Then, using the a priori bounds, we can create error bounds comparable to those found in the linear part. An example is given below.

Problem Statement: Find x such that

$$\begin{aligned} -\Delta x + x^3 &= f \\ x &= 0 \quad \text{on } \delta\Omega \end{aligned}$$

Weak Formulation: Find $x \in H_0^1(\Omega)$ such that, for all test functions $y \in H_1$

$$\int_\Omega -y\Delta x + x^3 y = \int_\Omega \nabla x \nabla y + x^3 y = \int_\Omega fy$$

Because the problem is nonlinear, the resulting weak formulation does not lend itself easily to a bilinear form. Instead, we break down the function into a linear and nonlinear part:

$$\langle F(x), y\rangle = \langle L(x) + N(x), y\rangle = \langle \nabla x, \nabla y\rangle_{L^2(\Omega)} + \langle x^3, y\rangle_{L^2(\Omega)} = \langle f, y\rangle_{L^2(\Omega)}$$

The linear part being the standard bilinear form for the Poisson equation, and the nonlinear part being the L^2 inner product between the nonlinear term and the test function.

If we can show that the nonlinear term obeys some sort of Lipschitz property, i.e. $\langle x^3 - y^3, z \rangle_{L^2(\Omega)} \leqslant K\|x-y\|\|z\|$, then we can control the nonlinear term so that we satisfy a strengthened Cauchy inequality:

$$|a(x - x_1, x_2 - x_1)| \leqslant C\|x - x_1\|\|x_2 - x_1\|$$

for some $C > 0$ that we can make as small as we like. In doing so, we can establish orthogonality in the energy norm $\|\cdot\|$ induced by a.

Given x, x_1, x_2 solutions to the weak formulation and the PG approximation in X, X_1, and X_2, successive subspaces in a finite element refinement algorithm, respectively, we have:

$$\|x - x_1\|^2 = \|x - x_2\|^2 + \|x_1 - x_2\|^2 + a(x_1 - x_2, x - x_2) + a(x - x_2, x_1 - x_2)$$

If we could establish the above strengthened Cauchy inequality, we would have:

$$\begin{aligned}
\|x - x_1\|^2 &\geqslant \|x - x_2\|^2 + \|x_1 - x_2\|^2 - 2C\|x_1 - x_2\|\|x - x_2\| \\
&\geqslant \|x - x_2\|^2 + \|x_1 - x_2\|^2 - C(\|x_1 - x_2\| + \|x - x_2\|) \\
&= (1 - C)(\|x - x_2\|^2 + \|x_1 - x_2\|^2) \\
&\implies \Lambda\|x - x_1\|^2 \geqslant \|x - x_2\|^2 + \|x_1 - x_2\|^2, \\
&\text{where } \Lambda = \frac{1}{1 - C}
\end{aligned}$$

For this semi-linear example, contraction of the error and error indicator can be achieved, which is proven in [9] and restated here.

Theorem [Quasi-Orthogonality and Contraction of Error]. We define the following:

$$e_1 = \|x - x_1\|_X \quad e_2 = \|x - x_2\|_X \quad E_1 = \|x_2 - x_1\|_X$$

With corresponding error indicators:

$$\eta_1 = \eta_1(x_1) \approx e_1 \quad \eta_2 = \eta_2(x_2) \approx e_2$$

Let the following assumptions hold:

(Q1) (Quasi-Orthogonality) For some $\Lambda \geqslant 1$ for (Λ sufficiently close to 1),

$$e_{k+1}^2 + E_k^2 \leqslant \Lambda e_k^2 \implies e_{k+1}^2 \leqslant \Lambda e_k^2 - E_k^2$$

(Q2) (Upper Bound on Error using Error Indicator) $\exists C_1 > 0$ s.t.

$$e_k^2 \leqslant C_1 \eta_k^2$$

(Q3) (Error Indicator Reduction) $\exists C_2 > 0$ and $\omega \in (0,1)$ s.t.

$$\eta_{k+1}^2 \leqslant C_2 E_k^2 + (1-\omega)\eta_k^2$$

Let $X_1 \subset X_2 \subset X$ be a triple of Banach spaces, let $x \in X$, $x_1 \in X_1$, and $x_2 \in X_2$, with the errors defined as above. Let $\beta \in (0,1)$ be arbitrary, and assume the constant Λ in the quasi-orthogonality assumption satisfies the following bound:

$$1 \leqslant \Lambda < 1 + \frac{\beta\omega}{C_1 C_2}$$

With the constants as defined in the above assumptions. Then there exists $\gamma > 0$ and $\alpha < 1$ where:

$$e_2^2 + \gamma\eta_2^2 \leqslant \alpha^2(e_1^2 + \gamma\eta_1^2)$$

Adaptive Finite Element Methods (AFEM)

In another variation to the method, we address a problem in the computational intensity of the finite element algorithm. For a number of problems, it may be the case that certain parts of the domain would be well approximated at earlier refinements of the mesh, whereas other parts of the domain would take many iterations to arrive at a suitable level of accuracy. Each uniform refinement of a mesh would greatly increase the number of elements, and therefore the computational time.

Adaptive finite element methods are a variation of this method that aims to cut the computational time and resources by focusing mesh refinements on areas of the domain where error would be high, while not refining the mesh as readily where the error is already low.

These methods present their own issues that are not present in uniform refinement schemes. In particular, maintaining some degree of regularity over the mesh, and estimating the relative error of the individual elements of the mesh proves to be problematic.

If such issues can be resolved, the basic algorithm for adaptive methods is as follows:

(AFEM-1) Find the Petrov-Galerkin approximation x_h for the discrete subspace X_h.

(AFEM-2) Estimate the error η_k for each element K of the mesh, for this approximation x_h.

(AFEM-3) Mark some subset of elements to be refined using some kind of marking scheme, such as the Dorfler marking scheme [7]. This is the main difference between an adaptive scheme and a standard refinement scheme. Which elements are refined and which ones are not is dependent on the error estimators on the elements in the domain, and the algorithm adapts to address those errors.

(AFEM-4) Refine the mesh at the marked elements, and any surrounding elements as needed to maintain the desired regularity of the mesh.

Once the mesh is refined, we repeat the process, much like in classical finite element methods, until the approximated error is below some level of tolerance. Within this algorithm would be some kind of error estimator η that can be used to estimate the error on each element. Such error estimators would be useful if they possessed properties that made them comparable to the true error ε_h [15]. In particular, for error indicators η_h on elements K ($\eta_h(K)$), and the sum total of such errors on the whole domain, $\eta_h(\Omega)$, and for some constants $C_1, C_2 > 0$:

(EI-1) Error indicators form an upper bound on the true error.

$$\|x - x_h\|^2 \leqslant C_1 \eta_h(\Omega)^2$$

(EI-2) Each error indicator on an element K forms some kind of lower bound on the true error on that element.

$$\eta_h(K)^2 \leqslant C_2(\varepsilon_h(K)^2 + osc(\omega_h(K))^2$$

where $osc(\omega_h(K))^2$ is an oscillation term.

With these assumptions, the error indicators are roughly equivalent to the true error, in such a way that marking elements for refinement based on such error indicators can produce meaningful results.

1.4 Standard issues in FEM problems

While we can establish these results with the assumptions above, establishing the assumptions themselves proves to be difficult, particularly the stability conditions. In Petrov-Galerkin methods, for example, even if the continuous inf-sup condition is established, that does not immediately imply the discrete inf-sup condition is held, nor does it imply that each discrete subspace could share the same stability constant. For a problem unstable enough, it could be the case that the stability constant goes to 0 as we continue to refine our discrete subspace.

Establishing the stability assumptions at all is itself a challenge, and often relies on assumptions to be made on the mesh, such as angle conditions, or on the problem itself. And even if stability is established, we may not know what the constants themselves are. If, for example, the inf-sup or coercivity constants are small, and the continuity constant is large, then the Galerkin solution to the problem, while some constant factor away from the best approximation to the problem, may be very far from the best approximation. Even if the problem was well posed in such a scenario, Galerkin approximations to the solution may not converge to the weak solution as quickly as needed. Thus, we turn to methods that would enforce some sort of stability inherently, as closely to an optimal level of stability as possible.

Chapter 2

Discontinuous Petrov Galerkin Methods

We are interested in a problem that is well-posed, in that a unique solution to the problem exists, and that the solution's behavior depends continuously on boundary conditions. Thus the existence of a inf-sup constant or a coercivity constant is vital to showing the problem is well posed. Moreover, the quasi-optimality quasi-stability constant depends on the continuity and inf-sup or coercivity constants. Even if an inf-sup constant existed for both the whole space and the discrete trial and test spaces, if it is small, then the resulting quasi-stability constant would be large, and the Galerkin solution to the problem could potentially have a comparably large error to the best approximation in the trial space. Establishing the existence of an inf-sup or coercivity constant at all requires careful selection of test and trial spaces.

2.1 The Ideal DPG Method: Formulation 1 (Trial to Test Operator)

In order to address the problem of stability, we consider a different formulation of the problem. In particular, we select the discrete test spaces specifically to establish well-posedness and a better quasi-stability constant. We consider a mixed finite element method on a linear problem, with some Hilbert trial space X and some Hilbert test space Y. We are trying to find some $x \in X$ such that:

$$a(x, y) = f(y) \quad \forall y \in Y \tag{2.1}$$

A method was proposed by J. Gopalakrishnan and L. Demcowicz [6], in which we use a semi-norm, in which the Petrov-Galerkin approximation we find is one that is optimally efficient, with continuity and inf-sup constants equal to 1. The method is described below.

First, we define an energy norm on X as follows:

$$\|x\|_E = \sup_{\|y\|=1, y \in Y} |a(x,y)| \tag{2.2}$$

This energy norm, if the assumptions of the inf-sup condition and the continuity assumption hold, is equivalent to the associated norm on our trial space X. The result is proven below:

Theorem [Conditions for Equivalence of the Energy Norm]. If the bilinear form $a(\cdot,\cdot)$ follows the above assumptions with continuity constant $\|a\|$ and with inf-sup constant γ, then the energy norm as defined above is equivalent to the norm on X for which the assumptions hold

Proof. From the continuity assumption, we have:

$$\|x\|_E = \sup_{\|y\|_Y=1, y \in Y} |a(x,y)| \leqslant \sup_{\|y\|_Y=1, y \in Y} \|a\|\|x\|_X\|y\|_Y = \|a\|\|x\|_X$$

And from our inf-sup assumption, we have:

$$\begin{aligned}\|x\|_E = \sup_{\|y\|_Y=1, y \in Y} |a(x,y)| &= \sup_{\|y\|_Y=1, y \in Y} \|x\|_X \left| a\left(\frac{x}{\|x\|_X}, y\right)\right| \\ &\geqslant \|x\|_X \inf_{\|x\|_X=1, x \in X} \sup_{\|y\|_Y=1, y \in Y} |a(x,y)| \geqslant \gamma\|x\|_X\end{aligned}$$

So we have:

$$\gamma\|x\|_X \leqslant \|x\|_E \leqslant \|a\|\|x\|_X$$

□

By the Riesz representation theorem, since $a(\cdot,\cdot)$ is bounded and linear in the second component, we have that there is some operator $T : X \to Y'$ such that for each $x \in X$, there is some $Tx \in Y$ such that $a(x,y) = (Tx, y)$ for all $y \in Y$. We define a seminorm in X, which coincides exactly with the above sup norm [6], as follows:

$$|x| = \sqrt{(Tx, Tx)} = \|Tx\|_Y \tag{2.3}$$

The equivalence of the energy norm (2.2) and the trial-to-test seminorm (2.3) is shown below:

$$\|x\|_E = \sup_{\|y\|_Y=1, y\in Y} |a(x,y)| = \sup_{\|y\|_Y=1, y\in Y} |(Tx,y)_Y| \leqslant \sup_{\|y\|_Y=1, y\in Y} \|Tx\|_Y \|y\|_Y = \|Tx\|_Y$$

But for $y = \frac{Tx}{\|Tx\|_Y}$, we have:

$$|a(x,y)| = \left| \left(Tx, \frac{Tx}{\|Tx\|_Y} \right) \right| = \frac{\|Tx\|_Y^2}{\|Tx\|_Y} = \|Tx\|_Y$$

So the energy norm attains its sup at a specific value of y, and moreover, that sup is precisely the seminorm defined above. With this seminorm we have the following.

First, the continuity constant is 1, since:

$$a(x,y) = (Tx,y) \leqslant \|Tx\|_Y \|y\|_Y = |x| \|y\|_Y$$

Second, the inf-sup constant is also 1, since:

$$\begin{aligned} \inf_{x\in X, |x|=1} \sup_{y\in Y, \|y\|_Y=1} |a(x,y)| &= \inf_{x\in X, |x|=1} \sup_{y\in Y, \|y\|_Y=1} |(Tx,y)_Y| \\ &\geqslant \inf_{x\in X, |x|=1} \left| \frac{(Tx,Tx)_Y}{\|Tx\|_Y} \right| = \inf_{x\in X, |x|=1} \|Tx\|_Y = \inf_{x\in X, |x|=1} |x| = 1 \end{aligned}$$

Under this semi-norm, using the results from above, we get that the Petrov-Galerkin approximation x_h abides by:

$$|x - x_h| \leqslant 2 \inf_{\xi_h \in X_h} |x - \xi_h|$$

However, since X and Y are Hilbert spaces, we have from the result proven by Xu and Zikatanaov [16] that

$$|x - x_h| = \inf_{\xi_h \in X_h} |x - \xi_h|$$

In other words, the Galerkin solution to the discrete weak formulation is the best approximation in the discrete trial space, under the seminorm as defined above. For each discrete trial space X_h with basis $\{e_1, e_2, ..., e_n\}$, if the corresponding discrete test space had basis $\{Te_1, Te_2, ..., Te_n\}$, the test space basis is called a basis of optimal test functions, since the test functions guarantee the optimal inf-sup and

continuity constants are achieved, and the best approximation property is realized.

To that end, for a practical implementation of the method, Gopalakrishnan and Demcowicz have the following steps:

1. A mesh with a variational formulation and an underlying test space Y_h that allows for inter-element discontinuities is developed.

2. The trial subspace X_h is selected to have good approximation properties, as is the case for many finite element methods.

3. The optimal test functions are approximately computed on an element by element basis. This is accomplished by approximating the trial to test operator T on each element, and finding some $T_n : X_n \to \tilde{Y}_n$ such that for all $\tilde{y}_n \in \tilde{Y}_n$,

$$(T_n x_n, \tilde{y}_n) = b(x_n, \tilde{y}_n)$$

and T_n is injective on X_n

$\tilde{Y}_n$ in this case would be a space of computationally convenient discontinuous functions used to construct the optimal test functions. Such a space must itself be selected prior to constructing the test space basis, and it must have dimension at least as large as X_n to enforce the injectivity of T_n. If the trial subspace X_n has basis $\{e_1, e_2, ..., e_k\}$, then the corresponding discrete test space Y_n has vectors $\{T_n e_1, T_n e_2, ..., T_n e_k\} = \{t_1, t_2, ..., t_k\}$, which form a basis, since T_n is injective.

4. Finally, given the trial to test operator approximated in the subspace, the resulting system, which corresponds to a symmetric positive definite stiffness matrix, is solved. We prove that the system is symmetric positive definite below.

Theorem [Stiffness Matrix is SPD]. Using the injective operator T_n as described above, the resulting stiffness matrix formed with the bilinear form $b(\cdot, \cdot)$ is symmetric positive definite.

Proof. We have that the basis for the discrete trial space X_n is given by $\{e_1, ..., e_k\}$, while the basis for the discrete trial space Y_n is given by $\{T_n e_1, T_n e_2, ..., T_n e_k\} = \{t_1, t_2, ..., t_k\}$ The (i, j)th entry of the resulting stiffness matrix M is therefore given by $b(e_i, t_j)$. But we observe by the construction T_n:

$$\begin{aligned} M_{ij} = b(e_i, t_j) = (T_n e_i, t_j)_Y = (t_i, t_j)_Y \\ = (t_j, t_i)_Y = (T_n e_j, t_i)_Y = b(e_j, t_i) = M_{ji} \end{aligned}$$

So M is symmetric.

We used the fact that T_n is injective to conclude the vectors $\{t_1, t_2, ..., t_k\}$ formed a basis, and are therefore linearly independent. Given stiffness matrix M whose (i,j)th entry is given by $b(e_i, t_j)$, and some nonzero vector $v = (v_1, v_2, ..., v_n)^T$, we have:

$$
\begin{aligned}
v^T M v &= (v_1, v_2, ..., v_n) M (v_1, v_2, ..., v_n)^T \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} M_{ij} v_i v_j = \sum_{i=1}^{k}\sum_{j=1}^{k} b(e_i, t_j) v_i v_j = \sum_{i=1}^{k}\sum_{j=1}^{k} (T_n e_i, T_n e_j)_Y v_i v_j \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} (v_i t_i, v_j t_j)_Y = \sum_{j=1}^{k} \left(\sum_{i=1}^{k} v_i t_i, v_j t_j \right)_Y = \left(\sum_{i=1}^{k} v_i t_i, \sum_{j=1}^{k} v_j t_j \right)_Y \\
&\left(\text{Letting } t = \sum_{i=1}^{k} v_i t_i \right) \\
&= (t,t)_Y = \|t\|_Y^2 > 0
\end{aligned}
$$

The last inequality holds, since by the linear independence of the basis, t must be non-zero whenever v is. This shows that M is also positive definite. In particular, since $t = t = \sum_{i=1}^{k} v_i t_i$, this implies that $t = Bv$, where B is a matrix whose columns are the basis for Y_h, and $M = B^T B$.

□

It is especially useful that the DPG stiffness matrix is symmetric positive definite, as such matrices have convenient properties that make solving linear systems less computationally intensive.

2.2 DPG Method: Formulation 2 (Minimizing the Residual)

In practice, the trial to test operator T is not an explicitly defined operator, or even one that is easily approximated over an entire mesh. In order for the problem of finding optimal test functions to be computationally feasible, it is better to compute some approximation to this operator on each element separately. To that end, we allow the functions in the test space to be discontinuous between elements, so that the approximate operator T_n can be computed per the basis functions on each element of the test space independent of the other elements in that space.

Not much is known about this theoretical trial-to-test operator, so we look to another formulation of the problem to define the operator more explicitly, by following the method outlined in [6]. Consider the following, where we view the bilinear from and the forcing function as operators in Y'. (As before, $l(y_h) = \langle f, y_h \rangle$)

$$Bx_h(y_h) = l(y_h) \quad \forall y_h \in Y_h$$

where $B : X \to Y'$ is defined for each $x_h \in X_h$ as:

$$Bx_h(y_h) = \langle Bx_h, y_h \rangle = b(x_h, y_h) \quad \forall y_h \in Y_h$$

We try to minimize the following residual:

$$\arg\min_{x_h \in X_h} \frac{1}{2} \|Bx_h - l\|^2_{Y'}$$

Consider the Riesz Operator on Y, $R_Y : Y \to Y'$ defined for all $y \in Y$ as:

$$(y, \delta y) = R_Y y(\delta y) \quad \forall \delta y \in Y$$

Since the Riesz operator and its inverse is an isometry, we have:

$$\arg\min_{x_h \in X_h} \frac{1}{2} \|Bx_h - l\|^2_{Y'} = \arg\min_{x_h \in X_h} \frac{1}{2} \|R_Y^{-1}(Bx_h - l)\|^2_Y$$

Taking the Gâteaux derivative of $\frac{1}{2}\|R_Y^{-1}(Bx_h - l)\|^2_Y$ at x_h, for all $\delta x_h \in X_h$, and setting the result to 0, we have:

$$0 = \left(R_Y^{-1}(Bx_h - l), R_Y^{-1} B\delta x_h\right)_Y = (Bx_h - l)(R_Y^{-1} B\delta x_h)$$

We define $T = R_Y^{-1} B$ and $y_h = R_Y^{-1} B\delta x_h$ for each $\delta x_h \in X_h$. Minimizing the residual then amounts to finding x_h such that:

$$Bx_h(y_h) = l(y_h) \quad \forall y_h = R_Y^{-1} B\delta x_h = T\delta x_h \in Y_h$$

And this is exactly the original problem, but with more explicitly defined test functions using well-studied operators. Note that, for the remainder of the paper, we define the following operators. We define some operator $B : X \to Y'$, where for some $x \in X$ the trial space, and for all $y \in Y$ the test space, and for some weak formulation linear in y (and possibly in x):

$$Bx(y) = (b(x), y) \tag{2.4}$$

And then as above, we will refer to the operator $T : X \to Y$, the trial to test operator, using the inverse of the Riesz operator on Y and the above defined operator B (2.4):

$$T := R_Y^{-1} B \tag{2.5}$$

We also similarly define T_h as a discrete approximation to T which would be in practice the operator that is computed for this method. For each element K in the mesh, we similarly define $T_K := R_Y(K)^{-1}B$ and $T_{h,K} := R_{h,Y(K)}^{-1}B$ as the trial to test operators restricted to these elements, with corresponding restricted inverse Riesz operators. Thus, given these operators, we have an error estimator:

$$\eta_K = \frac{1}{2}\|R_{Y(K)}^{-1}(Bx_h - l)\|_{Y(K)}^2 \tag{2.6}$$

While the Riesz operator itself is a better known operator in mathematics, it is still the case that problems associated with using the trial to test operator are also problems with using the Riesz operator. It is difficult to compute the Riesz operator R_Y explicitly over the entire mesh, due to sheer compuational complexity. Literature has indicated that such an operator can be approximated across an entire mesh [6]. This however, requires a large amount of computational overhead. In one paper, a multigrid preconditioner for the Laplace operator is used to approximate the Riesz operator $R_V : H^{-1}(\Omega) \mapsto H_0^1(\Omega)$. This however required a multigrid mesh and global computation and other computational overhead [1, 2].

Instead, we attempt to approximate R_Y element by element, allowing for discontinuities between elements in the test space. In doing so, the element-wise Riesz operator becomes simpler to compute, since computations are localized rather than global across the entire domain. However, even if we go element by element, the Riesz operator is still an infinite-dimensional operator, and thus difficult to approximate. Thus each element my element operator must be given some sort of discretized approximation. By allowing for these discontinuities and discretizing this trial to test operator, we are able to generate a test subspace that would, under the seminorm

2.3 Existing DPG Approaches to Nonlinear problems: Linearization

A standard approach to nonlinear problems is to linearize the problem, as in with some derivative, and then to solve the linear problem.

A number of papers [10, 12, 13] approach nonlinear problems as follows. The algorithm begins with some guess for the solution x_0. Then the linearized problem is used to find some solution increment

function $\Delta x \in X$ such that:

$$b_{lin}[x_0](\Delta x, y) = f_{lin}[x_0](y) \quad \forall y \in Y$$

b_{lin} is the derivative of the nonlinear form, and $b_{lin}[x_0](\Delta x, y) = D_u b_{nl}[x_0](\Delta x, y)$. f_{lin} is the forcing term, which represents the nonlinear residual, and $f_{lin}[x_0](y) = f(y) - b_{nl}(x_0, y)$. Then, the solution x_0 is incremented by the optimal increment Δx_n^{opt}, which is found my minimizing the following residual, similar to minimizing the residual in the linear problem:

$$\Delta x_n^{opt} = \underset{\Delta x_n \in X_n}{\arg\min} \| B_{lin}[x_0] \Delta x_n = f_{lin}[x_0] \|_{Y'}$$

where $B_{lin}[x_0]$ is defined using the bilinear form $b_{lin}[x_0]$ and the Riesz representation theorem. Once the solution is incremented, if the test space norm is the graph norm, then the test space norm is also updated, and the process is repeated once the spaces are refined. In updating the test space norm in each iteration, discrete stability can be established for the linearized problem [10].

Linearizing a nonlinear problem is a common approach to nonlinear problems, but linearizations inherit the problems of their nonlinear counterparts. Thus it is of interest to approach nonlinear problems without linearization of the nonlinear parts, and instead treating the nonlinear parts separately. Through exploiting certain properties of certain classes of problems, we may be able to approach nonlinear problems without having to linearize them.

For example, in a paper by Carstensen, DPG is also applied to nonlinear problems, by exploiting problems in which the nonlinear part is differentiable. The problem now has two parts: a bilinear part $b(\cdot\,,\cdot)$ with $b : Y \times Y \to \mathbb{R}$, and a nonlinear, differentiable part $n(\cdot\,;\cdot)$ with $n : X \times Y \to \mathbb{R}$, which is linear in the second part [3]. The goal is to find some $(x_h, y_h) \in X_h \times Y_h$ such that:

$$\begin{aligned} b(y_h, \eta_h) + n(x_h; \eta_h) &= l(\eta_h) && \forall \eta_h \in Y_h \\ n'(x_h; \xi_h, y_h) &= 0 && \forall \xi_h \in X_h \end{aligned}$$

Like the linearized example above, this method does rely on the nonlinear part to be differentiable and uses that differentiability in finding the weak solution. However, unlike in the Linearized examples above, this method approaches the nonlinear problem by having a form that specifically includes some nonlinear part that is only linear in the second part.

Given some regular solution to the weak formulation x, for some $\epsilon > 0$ and for $x_h \in B(x, \epsilon)$:

(CAR-1) The inf-sup condition holds for the derivative, i.e.

$$\inf_{\xi_h \in X_h} \sup_{y_h \in Y_h} |n'(x_h; \xi_h, y_h)| \geqslant \beta$$

(CAR-2) We can establish quasi-stability in that:

$$\|x - x_h\|_X + \|y_h\|_Y \leqslant C(x, \epsilon) \inf_{\xi_h \in X_h} \|x - \xi_h\|_X$$

(CAR-3) We can establish equivalence of the error indicator (Residual norm) and the norm in X:

$$C_1 \|x - x_h\|_X \leqslant \|l - Bx_h\|_{Y'} \leqslant C_2 \|x - x_h\|$$

In particular, this approach to a nonlinear problem is such that the Galerkin Solution x_h produced by the method is required to be within some small $\epsilon > 0$ of the regular weak solution x, to guarantee the existence of the inf-sup constant for the nonlinear part. The theorem is restated below:

Theorem [Stability of non-linear solution (Carstensen)]. Given a regular solution x to $B(x) = F$, there exists an open ball $B(x, \epsilon) := \{\tilde{x} \in X : \|x - \tilde{x}\|_X < \epsilon\}$, of radius $\epsilon > 0$ around x such that, for all $\tilde{x}_h \in B(x, \epsilon) \cap X_h \subset D_h$, the following discrete inf-sup condition holds:

$$0 < \frac{\beta(x; X_h, Y_h)}{2\|\Pi_h\|} \leqslant \beta(\tilde{x}_h; X_h, Y_h) := \inf_{\xi_h \in S(X_h)} \sup_{\eta_h \in S(Y_h)} b'(\tilde{x}_h; \xi_h, \eta_h)$$

Much like the standard assumption that a mesh be sufficiently fine for approximations to converge, Carstensen's approach requires the solution itself to a priori achieve some level of accuracy for the DPG method to be well posed.

Similarly, Carstensen requires that x_h to be sufficiently close to the regular weak solution x for the standard quasi-stability result from linear DPG methods. That theorem is restated below:

Theorem [Quasi-optimality of Non-linear DPG (Carstensen)]. Given a solution x to $B(x) = F$, there exist constants $\epsilon > 0$ and $C(x, \epsilon) > 0$ such that any solution (x_h, y_h) to the nonlinear weak formulation above with $\|x - x_h\|_X < \epsilon$ satisfies:

$$\|x - x_h\|_X + \|y_h\|_Y \leqslant C(x\epsilon) \inf_{\xi_h \in X_h} \|x - \xi_h\|_X$$

2.4 Problems with DPG and Criticisms

When viewing DPG through the idealized DPG method, where we approximate a trial to test operator by choosing functions specifically, a number of issues arise. First, the seminorm (2.3), based on the trial to test operator $T : X \to Y$ determines the optimal convergence of the Galerkin approximation to the weak solution. However, this semi-norm itself is constructed based on the selection of functions for a test subspace Y_n, and, in the idealized method, the initial underlying subspace $\tilde{Y}_n$ from which the operator T is approximated. Demcowicz and Gopalakrishnan assert that in certain schema, the selection of the underlying subspace to select optimal test functions could be automatic with hp adaptivity, but leaves to question how a such an underlying subspace can be selected in other cases [4].

Even if an appropriate underlying subspace $\tilde{Y}_n$ is selected, there is the issue of the seminorm itself, which is exactly the sup norm on the discrete test subspace Y_n. Because the sup norm takes the sup of the bilinear form across all unit vectors in the discrete subspace, the sup norm of any x_n would depend on the selection of such a subspace. But this means that, because the mesh and underlying function spaces are repeatedly refined, the nature of the energy norm itself changes at every iteration of the method.

This brings into question how meaningful the convergence is, since the seminorm and sup norm are not consistent throughout. One could have a best approximation under some semi-norm, but that may be itself not comparable to the best approximation in a later iteration. It is true that the sup norm is itself equivalent to any norm on the trial space, but only when the continuity and inf-sup conditions hold for that sup norm. And because the subspaces from which the sup norm is defined will change at every iteration of the method, it may not be known that the inf-sup constants relating the sup norm to the trial space norm do not also change. Should the continuity constant hold for the continuous problem, the discrete problem inherits the same constant, but the same is not true of the inf-sup constant. Thus we return to the same problem that motivated DPG in the first place: stabilizing the method by guaranteeing discrete stability. We can guarantee discrete stability per iteration, but since the seminorms at each iteration are not the same as norms in any containing spaces, we do not know if the discrete stability is meaningful.

Outside of that, should a desirable underlying subspace be selected, the method still requires one approximate the operator T_n for each trial subspace X_n. The success of the ideal method depends on how well this operator is approximated. The method offsets this by allowing for element by element computations, but still requires a very careful selection of functions on some underlying subspace of the test space.

So, in order for the method to be meaningful, the error estimates at each iteration and error indicators for any adaptive methods must correspond in some way to some error norm on the whole trial space. Should the discrete sup norm error estimates not be equal to the sup norm on the whole space, then we aim for the errors to be equivalent or near equivalent to those errors. Moreover, we would like some way to quantify the difference between the discrete residual error norms, based on a discrete test subspace Y_h and its dual Y_h', and the standard residual error norm, based on the whole test space.

To that end, we look to the second formulation of DPG, where we consider refinements based on minimizing the residual of the discrete weak formulation. In that framework, when we are computing error, we are trying to find a function $\varepsilon_h \in Y_h$ such that, for our DPG approximation x_h, and for all $v_h \in Y_h$ [4]:

$$\langle \varepsilon_h, v_h \rangle = A(x_h, v_h) - L(v_h) \tag{2.7}$$

Or in other words, we find our error estimation function $\varepsilon_h = R_{Y_h}^{-1}(Ax_h - L)$, and then compute its discrete dual norm.

In doing so, the error converges to 0, with the optimal stability constants. However, much like in the idealized method, rather than being a standard error norm, the error norm for the method is one based on test functions specifically selected to generate the optimal convergence rate. Thus, while the method uses the Riesz operator, which is more studied and well understood than a so-called trial-to-test operator, a common criticism of the method is still that the convergence in this error norm does not necessarily imply that the approximate solutions converge to the exact solution at all. The error norms are themselves are sup values based on the subspace selection for Y_h.

If compared the the exact solution to the problem using a standard norm, such as H^1 or L^2, it may be the case that the approximate solution could be very inaccurate. It is therefore important to relate the DPG error norm with the standard norms and norms that remain consistent throughout each iteration, as such norms are the most commonly studied and regarded as the baseline for convergence arguments.

2.5 Addressing the Criticisms to DPG

The error estimation function for DPG is tied to a method that is stable but is generated specifically to converge, and depends on the discrete test function subspace. Even if the solution could be approximated using the norm that resulted from these functions, that may not necessarily imply that the

solution converges in a norm of interest. The error estimator $\|\varepsilon_h\|_{Y_h}$ should be itself equivalent to the error in a more standard norm that is not dependent on the functions in the space, and can therefore be applied to all discrete subspaces. Below is a set of assumptions that allow for the DPG error indicator to exhibit the same behavior as a standard error indicator in an adaptive algorithm, presented by Carstensen et al in [4].

We begin by making several assumptions on the linear problem, namely:

(B1) The bilinear form is continuous in the sense that

$$|a(x,y)| \leqslant M\|x\|_X\|y\|_Y \qquad \forall x \in X, y \in Y$$

Discrete continuity automatically follows from the continuity from the continuous problem.

(B2) The bilinear form satisfies the continuous inf-sup condition:

$$\inf_{\|y\|_Y=1} \sup_{\|x\|_X=1} |a(x,y)| = \inf_{\|x\|_X=1} \sup_{\|y\|_Y=1} |a(x,y)| \geqslant \gamma > 0$$

(B3) For discrete subspaces X_h, Y_h,There exists some linear operator $\Pi_h : Y \to Y_h$ such that for all $\xi_h \in X_h$ and for all $y \in Y$:

$$a(\xi_h, y) = a(\xi_h, \Pi y) \implies a(\xi_h, y - \Pi y) = 0$$

This operator is known as the Fortin operator, and its existence in the context of DPG is proven in [8]. We also see operators with this property in other Petrov-Galerkin contexts, referred to as a Galerkin projection operator [9].

The discrete inf-sup condition follows from the second and third condition above. A proof of this is below:

Theorem [Galerkin projection implies discrete inf-sup]. Given the continuous inf-sup condition assumption above and the existence of the projection operator Π, the following discrete inf-sup condition holds:

$$0 < C(\gamma, \|\Pi\|) \leqslant \inf_{\|x_h\|_{X_h}=1} \sup_{\|y_h\|_{Y_h}=1} |a(x_h, y_h)|$$

In other words, the discrete inf-sup condition holds for all discrete trial and test subspaces, with a constant that depends only on the continuous inf sup constant and the projection operator $\|\Pi\|$.

Proof. From the continuous inf-sup condition:

$$\begin{aligned}
\gamma &\leqslant \inf_{\|x\|_X=1} \sup_{\|y\|_Y=1} |a(x,y)| \leqslant \inf_{\|x_h\|_{X_h}=1} \sup_{\|y\|_Y=1} |a(x_h,y)| \\
&= \inf_{\|x_h\|_{X_h}=1} \sup_{\|y\|_Y=1} |a(x_h,\Pi y)| = \inf_{\|x_h\|_{X_h}=1} \sup_{\|y\|_Y=1} \frac{\|\Pi y\||a(x_h,\Pi y)|}{\|\Pi y\|} \\
&\leqslant \|\Pi\| \inf_{\|x_h\|_{X_h}=1} \sup_{\|y_h\|_{Y_h}=1} |a(x_h,y_h)| \\
&\implies 0 < \frac{\gamma}{\|\Pi\|} \leqslant \inf_{\|x_h\|_{X_h}=1} \sup_{\|y_h\|_{Y_h}=1} |a(x_h,y_h)|
\end{aligned}$$

□

It follows that, if the Fortin operator exists for each iteration of the algorithm, that we can establish discrete stability. If the operator norm is bounded below by some $\|\Pi\|$ across all iterations of the algorithm, then this discrete stability is held throughout each iteration. Since in that case there would be both discrete continuity and a discrete inf-sup condition, by Cea's lemma, we have the following quasi-optimality estimate:

$$\|x - x_h\|_{X_h} \leqslant \frac{M\|\Pi\|}{\gamma} \inf_{\xi_h \in X_h} \|x - \xi_h\|_X$$

So if we may choose discrete test spaces to allow for the existence of a Fortin operator, we can maintain discrete stability under the semi-norm and under the norm over the overall space. Moreover, if such an operator exists, by using similar proofs as in the previous chapter, we can establish that the problem is well-posed, even while using the DPG semi-norm.

We can also extend this equivalence argument if we use the error indicator function in the residual method as an element by element error indicator for adaptivity. Given the assumptions above, as well as the discrete inf-sup condition that follows, we show that the residual from the discrete dual norm is roughly equivalent to the norm in the trial space X, similar to standard equivalence arguments for a posteriori error for adaptive finite element methods.

First, we define the error estimation functions ε and ε_h as follows. $\varepsilon \in Y$ is the function such that, for the solution $x_h \in X_h$ to the discrete weak formulation:

$$\langle \varepsilon, y \rangle = a(x_h, y) - f(y) \quad \forall y \in Y \tag{2.8}$$

In particular, this residual error function norm coincides exactly with the residual error. In other words, $\|\varepsilon\|_Y = \|Ax_h - F\|_{Y'}$, which we prove below:

Lemma [Error function norm is equal to residual dual norm]. The norm of the residual error function defined above is equal to the residual dual norm:

$$\|\varepsilon\|_Y = \|Ax_h - F\|_{Y'}$$

Proof.

$$\|Ax_h - F\|_{Y'} = \sup_{y\in Y, \|y\|=1} a(x_h, y) - f(y) = \sup_{y\in Y, \|y\|=1} \langle \varepsilon, y\rangle \leqslant \sup_{y\in Y, \|y\|=1} \|\varepsilon\|_Y \|y\|_Y = \|\varepsilon\|_Y$$

But the sup is attained with $y = \frac{\varepsilon}{\|\varepsilon\|}$, so the two norms are equal. In particular, we proved that the residual dual norm is equivalent to the trial space norm if the bilinear form is continuous and abides by the inf-sup condition, so this error function norm would be equivalent to the error in the trial space.

□

Similarly, $\varepsilon_h \in Y_h$ is the function such that, for the solution $x_h \in X_h$ to the discrete weak formulation, as in [2.7]:

$$\langle \varepsilon_h, y_h\rangle = a(x_h, y_h) - f(y_h) \qquad \forall y_h \in Y_h$$

Both functions exist by the Riesz representation theorem. The norm of the latter function is used to approximate the residual error. We also define a difference function $\delta = \varepsilon - \varepsilon_h$. We notice that $\delta \perp Y_h$, since for any $y_h \in Y_h$:

$$\langle \delta, y_h\rangle = \langle \varepsilon - \varepsilon_h, y_h\rangle = \langle \varepsilon, y_h\rangle - \langle \varepsilon_h, y_h\rangle = [a(x_h, y_h) - f(y_h)] - [a(x_h, y_h) - f(y_h)] = 0$$

With these functions as defined above, we also define an error indicator function η below, restricted to each element:

$$\eta_K = \|L - Ax_h\|_{Y_h', K} = \|\varepsilon_h\|_{Y_h, K} \tag{2.9}$$

Because convergence in the DPG norm is dependent on the norm of the dual space or the discrete test spaces, rather than the continuous test space, it is important to prove that the discrete dual norm of

the residual error is comparable to the error on the discrete space. For the linear problem, Carstensen, Demcowicz, and Gopalakrishnan prove that the error indicator function would be equivalent to the error in the trial space norm, in the same was adaptive finite element algorithm error indicators are equivalent to the error in the trial space [4]. In adaptive finite element theory, it is the case that error indicators are desired to be equivalent in that it is bounded above and below by some scalar multiple of some standard error norm, with some controllable oscillation term that may affect the reliability of the error indicator.

Theorem [Linear DPG error indicator equivalence to standard error norm]. With the above assumptions, with $L \in Y'$, $x = A^{-1}L$, and $x_h \in X_h$, and with $\eta = \|L - Ax_h\|_{Y_h'} = \|\varepsilon_h\|_{Y_h}$ and $osc(L) = \|L \circ (I - \Pi)\|_{Y'}$ we can establish the following:

1. $\eta \leqslant \|A\|\|x - x_h\|_X$ (efficiency)
2. $\gamma^2\|x - x_h\|_X^2 \leqslant \eta^2 + (osc(L) + \|\Pi\|\eta)^2$ (reliability)
3. $osc(L) \leqslant \|A\|\|I - \Pi\| \min_{\xi_h \in X_h} \|x - \xi_h\|_X$ (controlling oscillation term)

Proof. Efficiency follows immediately below:

$$\eta = \|L - Ax_h\|_{Y_h'} = \|Ax - Ax_h\|_{Y_h'} = \|A(x - x_h)\|_{Y_h'} \leqslant \|A\|\|x - x_h\|_X$$

Reliability requires the use of the continuous inf-sup condition, the fact that $\delta = \varepsilon - \varepsilon_h \perp Y_h$, and the third assumption regarding the projection Π:

$$\begin{aligned}
\|\delta\|^2 = \langle \delta, \delta \rangle = \langle \delta, \delta - \Pi\delta \rangle = \langle \varepsilon, \delta - \Pi\delta \rangle + \langle \varepsilon_h, \Pi\delta \rangle \\
\leqslant \|L \circ (I - \Pi)\|\|\delta\| + \|\varepsilon_h\|\|\Pi\|\|\delta\| \\
\implies \|\delta\| \leqslant \|L \circ (I - \Pi)\| + \|\varepsilon_h\|\|\Pi\|
\end{aligned}$$

Reliability then follows by using the Pythagorean theorem, so that $\|\varepsilon\|^2\| = \|\delta\|^2 + \|\varepsilon_h\|^2$, and, with the inf-sup condition:

$$\gamma^2\|x - x_h\|^2 \leqslant \|\varepsilon\|^2 = \|\varepsilon_h\|^2 + \|\delta\|^2 \leqslant \|\varepsilon_h\|^2 + (\|L \circ (I - \Pi)\| + \|\varepsilon_h\|\|\Pi\|)^2$$

And finally, we control the oscillation term. For any $y \in Y$ with $\|y\| = 1$ and for any $\xi_h \in X_h$:

$$L \circ (I - \Pi)(y) = L(y - \Pi y) = a(x, y - \Pi y) = a(x - \xi_h, y - \Pi y)$$
$$\leqslant \|A\| \|I - \Pi\| \|x - \xi_h\|$$

The result follows from taking the minimum over all $\xi_h \in X_h$ $\square$

Because the functions are allowed to be discontinuous in DPG, given some mesh or triangulation of elements $\mathcal{T}$, with elements $K \in \mathcal{T}$, we can define a total error indicator by the sum of element-wise error indicators.

$$\eta = \sum_{K \in \mathcal{T}} \eta_K$$

And the above proof also follows for each element's error indicator. Because of this, these element-wise error indicators can be used as error indicators for an adaptive algorithm, with potential to combine adaptive finite element theory with DPG theory.

Chapter 3

Applications to DPG Methods

3.1 Setup to the Semi-Linear Approach for DPG

DPG methods for nonlinear problems often involve the linearization of the original nonlinear problem, or rely on approximated Galerkin solutions to be within a small neighborhood of the weak solution to converge [3, 13]. This often results in the common assumption that the initial mesh is sufficiently fine [3]. However, we are interested instead in dealing with approaching in non-linear problems without relying on linearization, or on Galerkin solutions to be specifically within some ball of the weak solution. Much of existing DPG theory, including the optimal efficiency constant, relies on the problem itself to be linear or linearized [6, 3, 10, 13]. Our aim for this section is to show that, for certain classes of semi-linear problems, we can also achieve some of the stability stability results of the linear problems.

The setup for the semi-linear approach is similar to the linear one. There is some trial space X used to approximate the solution function, and some test space Y against which the trial space functions are compared. The semi-linear form has two parts: a bilinear part $a(\cdot\,,\cdot)$ with $a : X \times Y \to \mathbb{R}$, and a nonlinear part $(b(\cdot)\,,\cdot)$ with $b : X \times Y \to \mathbb{R}$, which is linear in the second part, Y. Much like in the linear function, there is some forcing function $L \in Y'$. There is the weak formulation in which the goal is to find some $x \in X$, where:

$$(c(x), y) : a(x, y) + (b(x), y) = L(y) \qquad \forall y \in Y \tag{3.1}$$

Because the trial and test spaces are infinite dimensional, we approximate the solution using some finite subspaces $X_h \subset X$ and $Y_h \subset Y$, thus resulting in a discrete weak formulation. The goal for the discrete weak formulation is to find some $x_h \in X_h$ such that:

$$(c(x_h), y_h) := a(x_h, y_h) + (b(x_h), y_h) = L(y_h) \qquad \forall y_h \in Y_h \tag{3.2}$$

For the semi-linear for, we define corresponding operators $C : X \to Y'$, $A : X \to Y'$, and $B : X \to Y'$ so that for all $x \in X, y \in Y$, $Cx(y) = (c(x), y)$, $Ax(y) = a(x, y)$, and $Bx(y) = (b(x), y)$. In doing so we can develop analagous results using the residual of the semi-linear problem. And we also make the following assumptions to the linear part of the semi-linear problem and on the whole problem, similar to those made to approach the linear problem:

(C1) c is a bounded nonlinear form, with continuity constant $\|c\|$:

$$\|c\| = \sup_{x \in X \setminus 0} \sup_{y \in Y \setminus 0} \frac{(c(x), y)}{\|x\|\|y\|} < \infty$$

This implies that b is similarly bounded, with some norm $\|b\|$

(C2) For the linear part of the nonlinear problem, a satisfies the continuous inf-sup condition:

$$0 < \beta := \inf_{x \in X \setminus 0} \sup_{y \in Y \setminus 0} \frac{a(x, y)}{\|x\|\|y\|} = \inf_{y \in Y \setminus 0} \sup_{x \in X \setminus 0} \frac{a(x, y)}{\|x\|\|y\|}$$

(C3) $X_h \subseteq X$ and $Y_h \subseteq Y$ are closed subspaces that admit a bounded linear operator $\Pi : Y \to Y_h$ with operator norm $\|\Pi\| < \infty$ and the property $a(X_h, (I - \Pi)Y) = 0$, i.e.

$$\forall \xi_h \in X_h, y \in Y, \quad a(\xi_h, y) = a(\xi_h, \Pi y)$$

We also note that, due to the Riesz representation theorem, we can, even for the semi-linear problem, define the following functional: Because the semi-linear problem is still linear in Y, a Hilbert space, for each $x \in X$, there exists some unique $Cx \in Y$ such that, for all $y \in Y$:

$$\langle Cx, y \rangle = (c(x), y) = a(x, y) + (b(x), y) \tag{3.3}$$

Error Representation Function

We also define the error representation similarly to the linear problem, by using the Riesz representations of the residuals. To do this, te define the error estimation functions ε and ε_h as follows. $\varepsilon \in Y$ is the function such that, for the discrete Galerkin solution $x_h \in X_h$ to the discrete weak formulation:

$$\langle \varepsilon, v \rangle = a(x_h, y) + (b(x_h), y) - L(y) \qquad \forall y \in Y \tag{3.4}$$

Similarly, for the problem, we define a $\varepsilon_h \in Y_h$ as the function such that, for the solution $x_h \in X_h$ to the discrete weak formulation:

$$\langle \varepsilon_h, y_h \rangle = a(x_h, y_h) + (b(x_h), y_h) - L(y_h) \qquad \forall y_h \in Y_h \tag{3.5}$$

Both functions exist by the Riesz representation theorem. Because Y is typically infinite dimensional, we typically do not use ε to approximate the residual error, and instead use its discrete dual approximation $\varepsilon \in Y_h$. Much like in the linear case, that means the error representation function and its norm depends on our selection of discrete test space, as we are approximating its norm using the discrete subspace dual norm $\|\cdot\|_{Y_h'}$.

3.2 A Semi-linear DPG Problem Example

Let our domain Ω be a Lipschitz domain, and let $X = Y = H_0^1(\Omega)$. We allow our discrete test spaces Y_h to be piecewise discontinuous. We want to find some $x \in X$ such that:

$$a(x, y) + (b(x), y) = L(y) \quad \forall y \in Y = H_0^1(\Omega) \tag{3.6}$$

Or rather, for our discrete formulation, we want to find $x_h \in X_h \subset H_0^1(\Omega)$ such that:

$$a(x_h, y_h) + (b(x_h), y_h) = L(y_h) \quad \forall y_h \in Y_h \subset H_0^1(\Omega) \tag{3.7}$$

We also assume that the solution must abide by Dirichlet boundary conditions, i.e. $x = 0$ on $\delta\Omega$.

As is the case with many approaches to nonlinear problems, we will attempt to exploit properties of linearity to show that the problem is well posed and to develop error estimates for the problem. However, rather than just linearizing the nonlinear problem, and rather than restricting ourselves to a small ball about the weak solution, we instead turn toward approaches to semi-linear problems for which the nonlinear part exhibits exploitable properties that are like a linear problem, without being linear. By weakening our assumptions to allow for a nonlinear part, we extend the class of problems the method can account for, but establish comparable results to the linear problems. To that end, we also make any number of the following additional assumptions.

Semi-linear Problem Assumptions

(S1) Maximum Principle: Suppose that the weak solution u is bounded, in the sense that if u is real valued, $u_- < u(x) < u_+$

(S2) Critical Growth: Suppose the domain $\Omega \subset \mathbb{R}^d$. There exists some integer n such that for $1 \leqslant n < \frac{d+2}{d-2}$ (when $d \geqslant 3$) or $1 \leqslant n < \infty$ (when d = 2), that the semi-linear form abides by:

$$\left| c^{(n)}(\eta) \right| \leqslant K$$

for some constant $K < \infty$. In other words, the non-linear form has some number of bounded derivatives.

(S3) Strong Cauchy: For all x_1 and x_2 in the trial space x,

$$|(b(x - x_1), x_2 - x_1)| \leqslant C\|x - x_1\|\|x_2 - x_1\|$$

for some $C > 0$ that we can make as small as we like.

(S4) Strong Monotonicity: The non-linear form $b(\cdot, \cdot)$ is strongly monotone, i.e. for all $x, y \in H_0^1$:

$$(b(x) - b(y), x - y) \geqslant 0$$

(S5) Lipschitz Constant for the nonlinear form: For all $x, y \in X$, and for some constant $k_L > 0$

$$\|b(x) - b(y)\| \leqslant k_L\|x - y\|_X$$

We also include the continuity assumption and the existence of the projection operator for the nonlinear problem. In addition, we may make the following assumptions:

(C2-1) The residual operator, $Cx - L$, has a bounded derivative and its derivative has a bounded inverse for all $x \in H_0^1$

(C2-2) Coercivity: There is some $c > 0$ such that for all $x \in H_0^1$:

$$c\|x\|^2 \leqslant a(x, x)$$

(C2-3) Like in (S4), $b(\cdot,\cdot)$ is strongly monotone in $H_0^1(\Omega)$, i.e. for all $x, y \in H_0^1$:

$$(b(x) - b(y), x - y) \geqslant 0$$

(C2-4) The bounded operator B corresponds to a nonlinear term in the weak formulation (3.6) that is bounded by a polynomial that satisfies critical growth conditions (S2).

3.3 Semi-Linear DPG Stability and Optimality Results

From the above assumptions, we can establish a number of stability and optimality results. In particular, for these semi-linear problems, we would like to establish an equivalence with the residual norm of the error to the error norm in the trial space, as well as prove some quasi-optimality results.

Theorem [Semi-linear Case 1: Strongly Monotone, Lipschitz]. Suppose the semi-linear formulation is has the Lipschitz property (S5), with continuity constant k_L, non-linear part $(b(\cdot),\cdot)$ is strongly monotone (S4), and the linear part $a(\cdot,\cdot)$ is coercive with coercivity constant m (A2). Then we have that the residual error norm abides by the following stability estimates. For some constants $C_M > 0$ and $c_m > 0$, the norm of the error indicator function ε is equivalent to the error of the Galerkin approximation in the trial space in that:

$$c_m \|x - x_h\| \leqslant \|\varepsilon\|_Y \leqslant C_m \|x - x_h\|$$

Moreover, for some constant $C > 0$, the error of the Galerkin approximation has the following quasi-stability result:

$$\|x - x_h\| \leqslant C \inf_{\xi_h \in X_h} \|x - \xi_h\|$$

Proof.

Lipschitz Continuity (S5) of the non-linear form gives us an upper bound to the residual error norm, since we have:

$$\begin{aligned}\|\varepsilon\|_Y &= \sup_{y\in Y,\|y\|=1} (c(x_h),y) - f(y)\\ &= \sup_{y\in Y,\|y\|=1} (c(x_h)-c(x),y)\\ &\leqslant \sup_{y\in Y,\|y\|=1} k_L\|x_h-x\|\|y\| = k_L\|x-x_h\|\end{aligned}$$

We also have a lower bound:

$$\begin{aligned}\|\varepsilon\|_Y &= \sup_{y\in Y,\|y\|=1} (c(x_h),y) - f(y) = \sup_{y\in Y,\|y\|=1} \langle Cx_h - F, y\rangle\\ &= \sup_{y\in Y,\|y\|=1} \langle Cx_h - Cx, y\rangle \geqslant \left\langle Cx_h - Cx, \frac{x_h-x}{\|x_h-x\|}\right\rangle\\ &= \left(C(x_h)-C(x), \frac{x_h-x}{\|x_h-x\|}\right) \geqslant a\left(x_h-x, \frac{x_h-x}{\|x_h-x\|}\right) \geqslant m\|x-x_h\|\end{aligned}$$

So we have:

$$\begin{aligned}\|x-x_h\|^2 &\leqslant \frac{1}{m}a(x-x_h,x-x_h) \leqslant \frac{1}{m}(C(x)-C(x_h),x-x_h)\\ &= \frac{1}{m}\langle Cx-Cx_h, x-x_h\rangle = \frac{1}{m}\langle Cx-Cx_h, x-\xi_h\rangle\\ &= \frac{1}{m}\langle Cx-Cx_h, x-\xi_h\rangle \leqslant \frac{1}{m}\|Cx-Cx_h\|_Y\|x-\xi_h\|_Y\\ &\leqslant \frac{k_L}{m}\|x-x_h\|\|x-\xi_h\|\\ &\implies \|x-x_h\| \leqslant \frac{k_L}{m}\inf_{\xi_h\in X_h}\|x-\xi_h\|\end{aligned}$$

□

It should be noted that while we have this estimate when the non-linear part is globally Lipschitz, this would imply that the linear part is essentially bounded by some linear function, and thus grows slower than a linear function. This class of non-linear forms is limited and are very close to linear. It is therefore of interest to consider a larger class of non-linear functions, and achieve comparable error estimate results. To that end, we consider Sobolev spaces, which are well studied, as a setting for a specific class of nonlinear functions that follow the critical growth assumption.

Theorem [Critical growth assumptions imply local Lipschitz condition]. Let $\Omega \subset \mathbb{R}^n$ be a Lipschitz domain, with $n \geqslant 2$. Let $X = H^1(\Omega)$ be the trial space.

Suppose that the nonlinear form b follows the critical growth assumption (S2), with bounded

derivatives up to some $k > 0$. Suppose that $x \in X$ is the galerkin solution to the weak formulation (3.6), and has some L_∞ bound, and let x_h be some function in the discrete trial space X_h. Then for some constant $C > 0$, the following Locally Lipschitz condition holds:

$$(b(x) - b(x_h), y) \leqslant C\|\nabla x - \nabla x_h\|\|\nabla y\|$$

Proof.

Expanding $b(x_h)$ with a Taylor expansion, we have:

$$b(x_h) = b(x) + \sum_{i=1}^{k} \frac{b^{(i)}(x)}{i!}(x_h - x)^i$$

Since b follows critical growth assumptions, and x is L_∞, $b^{(n)}(x)$ is bounded for all $n < k$ and $b^{(k)}(\xi)$ is bounded for all $\xi \in X$. So we have, from the above, and for some constant $K_1 > 0$:

$$\begin{aligned}
\|b(x) - b(x_h)\| &= \|\sum_{i=1}^{k} \frac{b^{(i)}(x)}{i!}(x - x_h)^i\| \\
&\leqslant \sum_{i=1}^{k} \|\frac{b^{(i)}(x)}{i!}(x - x_h)^i\| \\
&\leqslant \sum_{i=1}^{k} K_1\|(x - x_h)^i\|_p = \sum_{i=1}^{k} K_1\|(x - x_h)\|_{ip}^i
\end{aligned}$$

By the Poincare-Sobolev inequality, there is some $K_2 > 0$ such that $\|(x-x_h)\|_{ip} \leqslant K_2\|\nabla(x-x_h)\|_2$. So we now have:

$$\begin{aligned}
\|b(x) - b(x_h)\| &\leqslant \sum_{i=1}^{k} K_1\|(x - x_h)\|_{ip}^i \leqslant \sum_{i=1}^{k} K_2\|\nabla(x - x_h)\|_2^i \\
&= K_2\left(\frac{1 - \|\nabla(x - x_h)\|_2^k}{1 - \|\nabla(x - x_h)\|_2}\right)\|\nabla(x - x_h)\|_2
\end{aligned}$$

We have that for some constant $R > 0$:

$$\frac{1 - \|\nabla(x - x_h)\|_2^k}{1 - \|\nabla(x - x_h)\|_2} \leqslant R$$

So it follows that for some $C > 0$:

$$\|b(x) - b(x_h)\| \leqslant C\|\nabla(x - x_h)\|_2$$

And the result follows. □

Theorem [Semi-linear Case 2: Non-linear part with critical growth]. Let $\Omega \subset \mathbb{R}^n$ be a Lipschitz domain, with $n \geqslant 2$. Let $X = Y = H_0^1(\Omega)$ be the trial space and test space.

Suppose that the nonlinear form b follows the critical growth assumption (S2), with bounded derivatives up to some $k > 0$. Suppose that $x \in X$ is the galerkin solution to the weak formulation (3.6), and has some L_∞ bound, and let x_h be some function in the discrete trial space X_h. Suppose also that the linear part $a(\cdot, \cdot)$ is coercive (C2-2) with coercivity constant m. Then we have that the residual error norm abides by the following stability estimates. For some constants $C_M > 0$ and $c_m > 0$, the norm of the error indicator function ε as defined by (3.3) is equivalent to the error of the Galerkin approximation in the trial space in that:

$$c_m\|x - x_h\| \leqslant \|\varepsilon\|_Y \leqslant C_m\|x - x_h\|$$

Moreover, for some constant $C > 0$, the error of the Galerkin approximation has the following quasi-stability result:

$$\|x - x_h\| \leqslant C \inf_{\xi_h \in X_h} \|x - \xi_h\|$$

Proof.

Critical growth assumptions for the non-linear form and L_∞ bounds on u imply that there exists some constant $k_G > 0$ such that:

$$(b(x) - b(x_h), y) \leqslant k_G\|\nabla x - \nabla x_h\|_2\|\nabla y\|_2$$

give us an upper bound to the residual error norm, since we have:

$$\begin{aligned}\|\varepsilon\|_Y &= \sup_{y\in Y,\|y\|=1}(c(x_h),y)-f(y))\\ &= \sup_{y\in Y,\|y\|=1}(c(x_h)-c(x),y)\\ &\leqslant \sup_{y\in Y,\|y\|=1} k_G\|\nabla(x_h-x)\|\|\nabla y\| \leqslant k_G\|x-x_h\|\end{aligned}$$

We also, with the same proof as in case 1, have a lower bound:

$$\begin{aligned}\|\varepsilon\|_Y &= \sup_{y\in Y,\|y\|=1}(c(x_h),y)-f(y))\\ &\geqslant a\left(x_h-x,\frac{x_h-x}{\|x_h-x\|}\right)\geqslant m\|x-x_h\|\end{aligned}$$

So we have:

$$\begin{aligned}\|x-x_h\|^2 &\leqslant \frac{1}{m}a(x-x_h,x-x_h)\\ &\leqslant \frac{1}{m}(c(x)-c(x_h),x-x_h)\\ &\leqslant \frac{1}{m}\|Cx-Cx_h\|_Y\|x-\xi_h\|_Y\\ &\leqslant \frac{k_G}{m}\|x-x_h\|\|x-\xi_h\|\\ &\implies \|x-x_h\|\leqslant \frac{k_G}{m}\inf_{\xi_h\in X_h}\|x-\xi_h\|\end{aligned}$$

So the results follow. □

3.4 Proof of Strang's Lemmas for DPG

Because the DPG method requires the approximation of the Riesz operator, as well as norms that are reliant on the discrete subspace upon which the discrete formulation is solved, it is useful to quantify the effect such approximations would have on the error of the approximation. To that end, we prove analogues to Strang's lemmas below, for linear DPG.

Theorem [Strang's First Lemma (linear DPG, classical)]. Suppose we have a coercive bilinear form $a(\cdot,\cdot)=\langle T(\cdot),\cdot\rangle$ as in the weak formulation and similarly coercive discretizations $a_h(\cdot,\cdot)=\langle T_h(\cdot),\cdot\rangle$. Suppose also that we have the continuous and discretized forcing functions f and f_h. Then for some constant $C>0$ independent of the mesh, the residual error in the approximation assumes the following

bound:

$$\|x - x_h\|_Y \leqslant C \left(\|T_h x_h - Tx\|_{Y_h'} + \inf_{\xi_h \in Y} \left(\|x - \xi_h\|_Y + \|(T - T_h)\xi_h\|_{Y_h'} \right) \right)$$

Proof.

We have, from the coercivity and continuity of the bilinear form a:

$$\begin{aligned}
&\alpha \|x_h - \xi_h\|_Y^2 \\
&\leqslant |a_h(x_h - \xi_h, x_h - \xi_h)| \\
&= |a_h(x_h - \xi_h, x_h - \xi_h) + a(x - \xi_h, x_h - \xi_h) - a(x - \xi_h, x_h - \xi_h)| \\
&= |a(x - \xi_h, x_h - \xi_h) + (a(\xi_h, x_h - \xi_h) - a_h(\xi_h, x_h - \xi_h)) \\
&\quad + (a_h(x_h, x_h - \xi_h) - a(x, x_h - \xi_h))| \\
&= |a(x - \xi_h, x_h - \xi_h) + \langle (T - T_h)\xi_h, x_h - \xi_h \rangle + \langle T_h x_h - Tx, x_h - \xi_h \rangle| \\
&\leqslant M \|x - \xi_h\| \|x_h - \xi_h\| + |\langle (T - T_h)\xi_h, x_h - \xi_h \rangle| + |\langle (T_h x_h - Tx, x_h - \xi_h \rangle| \\
\implies &\|x_h - \xi_h\| \leqslant \frac{M}{\alpha} \|x - \xi_h\| + \frac{1}{\alpha} \left| \left\langle (T - T_h)\xi_h, \frac{x_h - \xi_h}{\|x_h - \xi_h\|} \right\rangle \right| + \frac{1}{\alpha} \left| \left\langle (T_h x_h - Tx, \frac{x_h - \xi_h}{\|x_h - \xi_h\|} \right\rangle \right| \\
&\leqslant \frac{M}{\alpha} \|x - \xi_h\| + \frac{1}{\alpha} \sup_{y_h \in Y_h, \|y_h\| = 1} |\langle (T - T_h)\xi_h, y_h \rangle| + \frac{1}{\alpha} \sup_{y_h \in Y_h, \|y_h\| = 1} |\langle (T_h x_h - Tx, y_h \rangle| \\
&\leqslant \frac{M}{\alpha} \|x - \xi_h\| + \frac{1}{\alpha} \|(T - T_h)\xi_h\|_{Y_h'} + \frac{1}{\alpha} \|T_h x_h - Tx\|_{Y_h'}
\end{aligned}$$

Since, by the triangle inequality,

$$\|x - x_h\|_Y \leqslant \|x - \xi_h\|_Y + \|x_h - \xi_h\|_Y$$

We have the following:

$$
\begin{aligned}
\|x - x_h\|_Y &\leqslant \|x - \xi_h\|_Y + \|x_h - \xi_h\|_Y \\
&\leqslant \|x - \xi_h\|_Y + \frac{M}{\alpha}\|x - \xi_h\|_Y + \frac{1}{\alpha}\|(T - T_h)\xi_h\|_{Y_h'} + \frac{1}{\alpha}\|T_h x_h - Tx\|_{Y_h'} \\
&= \left(1 + \frac{M}{\alpha}\right)\|x - \xi_h\|_Y + \frac{1}{\alpha}\|(T - T_h)\xi_h\|_{Y_h'} + \frac{1}{\alpha}\|T_h x_h - Tx\|_{Y_h'} \\
\implies \|x - x_h\|_Y &\leqslant \frac{1}{\alpha}\|T_h x_h - Tx\|_{Y_h'} + \inf_{\xi_h \in Y_h}\left(\left(1 + \frac{M}{\alpha}\right)\|x - \xi_h\|_Y + \frac{1}{\alpha}\|(T - T_h)\xi_h\|_{Y_h'}\right)
\end{aligned}
$$

The conclusion follows.

□

Theorem [Strang's Second Lemma (linear DPG, classical)]. Suppose we have a coercive bilinar form $a(\cdot,\cdot) = \langle T(\cdot), \cdot\rangle$ as in the weak formulation and similarly coercive discretizations $a_h(\cdot,\cdot) = \langle T_h(\cdot), \cdot\rangle$. Suppose forcing function f is not approximated and is instead exact. Let u be the solution to the weak formulation, and u_h the discrete solution using some subspace of the trial space X_h, and some subspace of the test space Y_h Then the residual error in the approximation assumes the following bound:

$$
\|x - x_h\|_{Y_h} \leqslant C\left(\|T_h(x - x_h)\|_{Y_h'} + \inf_{\xi_h \in Y_h}\|x - \xi_h\|_{Y_h}\right)
$$

Proof. Because the bi-linear forms are coercive, we have:

$$
\begin{aligned}
\alpha\|x_h - \xi_h\|_{Y_h}^2 &\leqslant |a_h(x_h - \xi_h, x_h - \xi_h)| \\
&= |a_h(x_h - \xi_h, x_h - \xi_h) + a_h(x - x_h, x_h - \xi_h) - a_h(x - x_h, x_h - \xi_h)| \\
&= |a_h(x - \xi_h, x_h - \xi_h) + (a_h(x_h, x_h - \xi_h) - a_h(x_h, x_h - \xi_h))| + |a_h(x - x_h, x_h - \xi_h)| \\
&= |a_h(x - \xi_h, x_h - \xi_h)| + |\langle T_h(x - x_h), x_h - \xi_h\rangle| \\
&\leqslant M\|x - \xi_h\|\|x_h - \xi_h\| + |\langle (T_h(x - x_h), x_h - \xi_h\rangle| \\
\implies \|x_h - \xi_h\| &\leqslant \frac{M}{\alpha}\|x - \xi_h\| + \frac{1}{\alpha}\left|\left\langle (T_h(x - x_h), \frac{x_h - \xi_h}{\|x_h - \xi_h\|}\right\rangle\right| \\
&\leqslant \frac{M}{\alpha}\|x - \xi_h\| + \frac{1}{\alpha}\sup_{y_h \in Y_h, \|y_h\| = 1}|\langle (T_h(x - x_h), y_h\rangle| \\
&= \frac{M}{\alpha}\|x - \xi_h\| + \frac{1}{\alpha}\|T_h(x - x_h)\|_{Y_h'}
\end{aligned}
$$

Since, by the triangle inequality,

$$\|x - x_h\|_{Y_h} \leqslant \|x - \xi_h\|_{Y_h} + \|x_h - \xi_h\|_{Y_h}$$

We have the following:

$$\begin{aligned}
\|x - x_h\|_{Y_h} &\leqslant \|x - \xi_h\|_{Y_h} + \|x_h - \xi_h\|_{Y_h} \\
&\leqslant \|x - \xi_h\|_{Y_h} + \frac{M}{\alpha}\|x - \xi_h\| + \frac{1}{\alpha}\|T_h(x - x_h)\|_{Y_h'} \\
&= \left(1 + \frac{M}{\alpha}\right)\|x - \xi_h\|_{Y_h} + \frac{1}{\alpha}\|T_h(x - x_h)\|_{Y_h'} \\
\implies \|x - x_h\|_{Y_h} &\leqslant \frac{1}{\alpha}\|T_h(x - x_h)\|_{Y_h'} + \inf_{\xi_h \in Y_h}\left(\left(1 + \frac{M}{\alpha}\right)\|x - \xi_h\|_{Y_h}\right)
\end{aligned}$$

The conclusion follows.

□

For cases above, bilinear forms can only be coercive if the trial and test spaces are the same, so that the trial space functions can be in the second part of the bilinear form. The Petrov-Galerkin part of the simplest case comes from the fact that test subspaces are allowed to be discontinuous, and subspaces could be different but still have a well defined inner product. However, Petrov-Galerkin methods like DPG would typically have different function spaces for the trial and test functions, where a bilinear form related to an inner product would not be defined. In which case, we also consider Strang's lemmas for mixed methods, in which the trial and test spaces are different.

Theorem [Strang's First Lemma (linear DPG, mixed)]. Suppose we have a bilinear form $a(\cdot,\cdot) = \langle T(\cdot), \cdot\rangle$, as in the weak formulation 1.4, that fulfills the inf-sup condition (A2-1) and the discrete inf-sup condition (A2-2), with inf-sup constant β. Suppose we also have discretizations with the same inf-sup constants, $a_h(\cdot,\cdot) = \langle T_h(\cdot), \cdot\rangle$. Suppose also that we have the continuous and discretized forcing functions f and f_h. Then for some constant $C > 0$ independent of the mesh, the residual error in the approximation assumes the following bound:

$$\|x - x_h\|_X \leqslant C\left(\|T_h x_h - Tx\|_{Y_h'} + \inf_{\xi_h \in X_h}\left(\|x - \xi_h\|_X + \|(T - T_h)\xi_h\|_{Y_h'}\right)\right)$$

Proof.

We have, from the inf-sup condition and continuity of the bilinear form a:

$$
\begin{aligned}
\|x_h - \xi_h\|_X &\leqslant \frac{1}{\beta} \sup_{y\in Y_h, \|y_h\|_Y=1} a_h(x_h - \xi_h, y_h) \\
&= \frac{1}{\beta} \sup_{y\in Y_h, \|y_h\|_Y=1} a_h(x_h - \xi_h, y_h) + a(x - \xi_h, y_h) - a(x - \xi_h, y_h) \\
&= \frac{1}{\beta} \sup_{y\in Y_h, \|y_h\|_Y=1} a(x - \xi_h, y_h) + a(\xi_h, y_h) - a_h(\xi_h, y_h) + a(x, y_h) - a_h(\xi_h, y_h) \\
&\leqslant \frac{1}{\beta} \sup_{y\in Y_h, \|y_h\|_Y=1} |a(x - \xi_h, y_h)| + |\langle (T - T_h)\xi_h, y_h\rangle| + |\langle Tx - T_h\xi_h, y_h\rangle| \\
&\leqslant \frac{1}{\beta} \sup_{y\in Y_h, \|y_h\|_Y=1} M\|x - \xi_h\|_X\|y_h\|_Y + \|(T - T_h)\xi_h\|_{Y_h'} + \|Tx - T_h\xi_h\|_{Y_h'} \\
&= \frac{M}{\beta}\|x - \xi_h\|_X + \frac{1}{\beta}\|(T - T_h)\xi_h\|_{Y_h'} + \frac{1}{\beta}\|Tx - T_h\xi_h\|_{Y_h'}
\end{aligned}
$$

Since, by the triangle inequality,

$$
\|x - x_h\|_X \leqslant \|x - \xi_h\|_X + \|x_h - \xi_h\|_X
$$

We have the following:

$$
\begin{aligned}
\|x - x_h\|_X &\leqslant \|x - \xi_h\|_X + \|x_h - \xi_h\|_X \\
&\leqslant \|x - \xi_h\|_X + \frac{M}{\beta}\|x - \xi_h\|_X + \frac{1}{\beta}\|(T - T_h)\xi_h\|_{Y_h'} + \frac{1}{\beta}\|Tx - T_h\xi_h\|_{Y_h'} \\
&= \left(1 + \frac{M}{\beta}\right)\|x - \xi_h\|_X + \frac{1}{\beta}\|(T - T_h)\xi_h\|_{Y_h'} + \frac{1}{\beta}\|T_hx_h - Tx\|_{Y_h'} \\
\implies \|x - x_h\|_X &\leqslant \frac{1}{\beta}\|T_hx_h - Tx\|_{Y_h'} + \inf_{\xi_h\in X_h}\left(\left(1 + \frac{M}{\beta}\right)\|x - \xi_h\|_X + \frac{1}{\beta}\|(T - T_h)\xi_h\|_{Y_h'}\right)
\end{aligned}
$$

The conclusion follows.

□

Theorem [Strang's Second Lemma (linear DPG, mixed)]. Suppose we have a bilinear form, as in the weak formulation (1.4), $a(\cdot,\cdot) = \langle T(\cdot),\cdot\rangle$, that fulfills the inf-sup condition (A2-1) and the discrete inf-sup condition (A2-2), with inf-sup constant β. Suppose we also have discretizations with the same inf-sup constants, $a_h(\cdot,\cdot) = \langle T_h(\cdot),\cdot\rangle$. Suppose forcing function f is not approximated and is instead exact. Let u be the solution to the weak formulation, and u_h the discrete solution using some subspace of the

trial space X_h, and some subspace of the test space Y_h Then the residual error in the approximation assumes the following bound for some constant $C > 0$:

$$\|x - x_h\|_X \leqslant C \left(\|T_h(x - x_h)\|_{Y_h'} + \inf_{\xi_h \in X_h} \|x - \xi_h\|_X \right)$$

Proof.

We have, from the inf-sup condition and continuity of the bilinear form a:

$$\begin{aligned}
\|x_h - \xi_h\|_X &\leqslant \frac{1}{\beta} \sup_{y \in Y_h, \|y_h\|_Y = 1} a_h(x_h - \xi_h, y_h) \\
&= \frac{1}{\beta} \sup_{y \in Y_h, \|y_h\|_Y = 1} a_h(x_h - \xi_h, y_h) + a_h(x - x_h, y_h) - a_h(x - x_h, y_h) \\
&= \frac{1}{\beta} \sup_{y \in Y_h, \|y_h\|_Y = 1} a_h(x - \xi_h, y_h) - a_h(x - x_h, y_h) \\
&\leqslant \frac{1}{\beta} \sup_{y \in Y_h, \|y_h\|_Y = 1} |a(x - \xi_h, y_h)| + |\langle T_h(x - x_h), y_h \rangle| \\
&\leqslant \frac{1}{\beta} \sup_{y \in Y_h, \|y_h\|_Y = 1} M \|x - \xi_h\|_X \|y_h\|_Y + \|T_h(x - x_h)\|_{Y_h'} \\
&= \frac{M}{\beta} \|x - \xi_h\|_X + \frac{1}{\beta} \|T_h(x - x_h)\|_{Y_h'}
\end{aligned}$$

Since, by the triangle inequality,

$$\|x - x_h\|_X \leqslant \|x - \xi_h\|_X + \|x_h - \xi_h\|_X$$

We have the following:

$$\begin{aligned}
\|x - x_h\|_X &\leqslant \|x - \xi_h\|_X + \|x_h - \xi_h\|_X \\
&\leqslant \|x - \xi_h\|_X + \frac{M}{\beta} \|x - \xi_h\|_X + \frac{1}{\beta} \|T_h(x - x_h)\|_{Y_h'} \\
&= \left(1 + \frac{M}{\beta}\right) \|x - \xi_h\|_X + \frac{1}{\beta} \|T_h(x - x_h)\|_{Y_h'} \\
\implies \|x - x_h\|_X &\leqslant \frac{1}{\beta} \|T_h(x - x_h)\|_{Y_h'} + \left(1 + \frac{M}{\beta}\right) \inf_{\xi_h \in X_h} \|x - \xi_h\|_X
\end{aligned}$$

The conclusion follows. □

We then are also interested in how to characterise the effect of this approximation in the context of semi-linear problems under the conditions listed in the second section of this chapter. We prove those results below as well. An important difference is that while the Riesz operator may still exist, it is no longer a linear operator in this context.

Theorem [Strang's First Lemma (semi-linear DPG)]. Suppose we have a semi-linear form as in the weak formulation 3.1:

$$(c(\cdot),\cdot) = a(\cdot,\cdot) + (b(\cdot),\cdot) = \langle T(\cdot),\cdot\rangle$$

with a bilinear part $a(\cdot,\cdot)$ that is coercive as in (A2) and continuous as in (A1), and a non-linear part $(b(\cdot),\cdot)$ that is strongly monotone as in (S4) and abides by some Lipschitz condition (S5) or the condition proven with the discrete maximum and critical growth conditions (S1) and (S2), with some boundedness constant k. Suppose we similarly have discretizations $(c_h(\cdot),\cdot) = \langle T_h(\cdot),\cdot\rangle$, $a_h(\cdot,\cdot)$, and $(b_h(\cdot),\cdot)$. Suppose also that we have the continuous and discretized forcing functions f and f_h. Then for some constant $C > 0$ independent of the mesh, the residual error in the approximation assumes the following bound:

$$\|x - x_h\|_Y \leqslant C\left(\|T_h x_h - Tx\|_{Y_h'} + \inf_{\xi_h \in Y}\left(\|x-\xi_h\|_Y + \|(T-T_h)\xi_h\|_{Y_h'}\right)\right)$$

Proof.

We have, from the coercivity of the bilinear part a, and the strong monotonicity of the nonlinear part b:

$$
\begin{aligned}
\alpha\|x_h-\xi_h\|_Y^2 &\leqslant |a_h(x_h-\xi_h,x_h-\xi_h)| \leqslant |a_h(x_h-\xi_h,x_h-\xi_h)+(b_h(x_h)-b_h(\xi_h),x_h-\xi_h)|\\
&= |(c_h(x_h)-c_h(\xi_h),x_h-\xi_h)+(c(x)-c(\xi_h),x_h-\xi_h)-(c(x)-c(\xi_h),x_h-\xi_h)|\\
&= |(c(x)-c(\xi_h),x_h-\xi_h)+(c(\xi_h)-c_h(\xi_h),x_h-\xi_h)+(c_h(x_h)-c(x),x_h-\xi_h)|\\
&= |(c(x)-c(\xi_h),x_h-\xi_h)+\langle (T-T_h)\xi_h,x_h-\xi_h\rangle+\langle T_hx_h-Tx,x_h-\xi_h\rangle|\\
&\leqslant (M+k)\|x-\xi_h\|\|x_h-\xi_h\|+|\langle (T-T_h)\xi_h,x_h-\xi_h\rangle|+|\langle (T_hx_h-Tx,x_h-\xi_h\rangle|\\
\implies \|x_h-\xi_h\| &\leqslant \frac{M+k}{\alpha}\|x-\xi_h\|+\frac{1}{\alpha}\left|\left\langle (T-T_h)\xi_h,\frac{x_h-\xi_h}{\|x_h-\xi_h\|}\right\rangle\right|\\
&\quad+\frac{1}{\alpha}\left|\left\langle (T_hx_h-Tx,\frac{x_h-\xi_h}{\|x_h-\xi_h\|}\right\rangle\right|\\
&\leqslant \frac{M+k}{\alpha}\|x-\xi_h\|+\frac{1}{\alpha}\sup_{y_h\in Y_h,\|y_h\|=1}|\langle (T-T_h)\xi_h,y_h\rangle|+\frac{1}{\alpha}\sup_{y_h\in Y_h,\|y_h\|=1}|\langle (T_hx_h-Tx,y_h\rangle|\\
&\leqslant \frac{M+k}{\alpha}\|x-\xi_h\|+\frac{1}{\alpha}\|(T-T_h)\xi_h\|_{Y_h'}+\frac{1}{\alpha}\|T_hx_h-Tx\|_{Y_h'}
\end{aligned}
$$

Since, by the triangle inequality,

$$\|x-x_h\|_Y \leqslant \|x-\xi_h\|_Y+\|x_h-\xi_h\|_Y$$

We have the following:

$$
\begin{aligned}
\|x-x_h\|_Y &\leqslant \|x-\xi_h\|_Y+\|x_h-\xi_h\|_Y\\
&\leqslant \|x-\xi_h\|_Y+\frac{M+k}{\alpha}\|x-\xi_h\|_Y+\frac{1}{\alpha}\|(T-T_h)\xi_h\|_{Y_h'}+\frac{1}{\alpha}\|T_hx_h-Tx\|_{Y_h'}\\
&= \left(1+\frac{M+k}{\alpha}\right)\|x-\xi_h\|_Y+\frac{1}{\alpha}\|(T-T_h)\xi_h\|_{Y_h'}+\frac{1}{\alpha}\|T_hx_h-Tx\|_{Y_h'}\\
\implies \|x-x_h\|_Y &\leqslant \frac{1}{\alpha}\|T_hx_h-Tx\|_{Y_h'}+\inf_{\xi_h\in Y_h}\left(\left(1+\frac{M+k}{\alpha}\right)\|x-\xi_h\|_Y+\frac{1}{\alpha}\|(T-T_h)\xi_h\|_{Y_h'}\right)
\end{aligned}
$$

The conclusion follows.

□

Theorem [Strang's Second Lemma (semi-linear DPG)]. Suppose we have a semi-linear form as in the weak formulation 3.1:

$$(c(\cdot),\cdot)=a(\cdot,\cdot)+(b(\cdot),\cdot)=\langle T(\cdot),\cdot\rangle$$

with a bilinear part $a(\cdot,\cdot)$ that is coercive as in (A2) and continuous as in (A1), and a non-linear part $(b(\cdot),\cdot)$ that is strongly monotone as in (S4) and abides by some Lipschitz condition (S5) or the condition proven with the discrete maximum and critical growth conditions (S1) and (S2), with some boundedness constant k. Suppose we similarly have discretizations $(c_h(\cdot),\cdot) = \langle T_h(\cdot),\cdot\rangle$, $a_h(\cdot,\cdot)$, and $(b_h(\cdot),\cdot)$. Suppose forcing function f is not approximated and is instead exact. Let u be the solution to the weak formulation, and u_h the discrete solution using some subspace of the trial space X_h, and some subspace of the test space Y_h Then the residual error in the approximation assumes the following bound:

$$\|x - x_h\|_{Y_h} \leqslant C\left(\|T_h(x - x_h)\|_{Y_h'} + \inf_{\xi_h \in Y_h} \|x - \xi_h\|_{Y_h}\right)$$

Proof. We have, from the coercivity of the bilinear part a, and the strong monotonicity of the nonlinear part b:

$$\begin{aligned}
\alpha\|x_h - \xi_h\|_Y^2 &\leqslant |a_h(x_h - \xi_h, x_h - \xi_h)| \leqslant |a_h(x_h - \xi_h, x_h - \xi_h) + (b_h(x_h) - b_h(\xi_h), x_h - \xi_h)| \\
&= |(c_h(x_h) - c_h(\xi_h), x_h - \xi_h) + (c_h(x) - c_h(x_h), x_h - \xi_h) - (c_h(x) - c_h(x_h), x_h - \xi_h)| \\
&= |(c_h(x) - c_h(\xi_h), x_h - \xi_h) + (c_h(x_h) - c_h(x_h), x_h - \xi_h)| + |(c_h(x) - c_h(x_h), x_h - \xi_h)| \\
&= |(c_h(x) - c_h(\xi_h), x_h - \xi_h)| + |\langle T_h(x) - T_h(x_h), x_h - \xi_h\rangle| \\
&\leqslant (M + k)\|x - \xi_h\|\|x_h - \xi_h\| + |\langle (T_h(x) - T_h(x_h), x_h - \xi_h\rangle| \\
\implies \|x_h - \xi_h\| &\leqslant \frac{M+k}{\alpha}\|x - \xi_h\| + \frac{1}{\alpha}\left|\left\langle (T_h(x) - T_h(x_h), \frac{x_h - \xi_h}{\|x_h - \xi_h\|}\right\rangle\right| \\
&\leqslant \frac{M+k}{\alpha}\|x - \xi_h\| + \frac{1}{\alpha}\sup_{y_h \in Y_h, \|y_h\|=1} |\langle (T_h(x) - T_h(x_h), y_h\rangle| \\
&= \frac{M+k}{\alpha}\|x - \xi_h\| + \frac{1}{\alpha}\|T_h(x) - T_h(x_h)\|_{Y_h'}
\end{aligned}$$

Since, by the triangle inequality,

$$\|x - x_h\|_{Y_h} \leqslant \|x - \xi_h\|_{Y_h} + \|x_h - \xi_h\|_{Y_h}$$

We have the following:

$$
\begin{aligned}
\|x - x_h\|_{Y_h} &\leqslant \|x - \xi_h\|_{Y_h} + \|x_h - \xi_h\|_{Y_h} \\
&\leqslant \|x - \xi_h\|_{Y_h} + \frac{M+k}{\alpha}\|x - \xi_h\| + \frac{1}{\alpha}\|T_h(x) - T_h(x_h)\|_{Y_h'} \\
&= \left(1 + \frac{M+k}{\alpha}\right)\|x - \xi_h\|_{Y_h} + \frac{1}{\alpha}\|T_h(x) - T_h(x_h)\|_{Y_h'} \\
\implies \|x - x_h\|_{Y_h} &\leqslant \frac{1}{\alpha}\|T_h(x) - T_h(x_h)\|_{Y_h'} + \inf_{\xi_h \in Y_h}\left(\left(1 + \frac{M+k}{\alpha}\right)\|x - \xi_h\|_{Y_h}\right)
\end{aligned}
$$

The conclusion follows.

□

We are, moreover, interested in how the dual norm error in the test space Y differs from the discrete test space Y_h. We defined above the difference function, $\delta = \varepsilon - \varepsilon_h$, between the residual error ε and the approximate residual error ε_h. By the triangle inequality, we have:

$$\|\varepsilon\| \leqslant \|\delta\| + \|\varepsilon_h\|$$

Hence controlling the difference would keep the error estimation function ε_h comparable to the residual error. For the linear problem, we can show that the error estimator function ε_h is the best estimator for the residual error, in the following way.

Lemma [δ in the linear problem]. Given the error estimator functions ϵ and the approximate function ϵ_h, for fixed ε, δ is optimized by the error estimator function ε_h, in that:

$$\|\delta\| = \inf_{\eta_h \in Y_h}\|\varepsilon - \eta_h\|_Y$$

Proof. We use the fact that by construction, $\delta \in Y_h^\perp$, hence $\langle \delta, y_h \rangle = 0$ for any $y_h \in Y_h$. We have for any arbitrary $\eta_h \in Y_h$

$$
\begin{aligned}
&\|\delta\|_Y^2 = \langle \delta, \delta \rangle = \langle \delta, \varepsilon - \varepsilon_h \rangle = \langle \delta, \varepsilon - \eta_h \rangle \leqslant \|\delta\|_Y \|\varepsilon - \eta_h\|_Y \\
&\implies \|\delta\|_Y \leqslant \|\varepsilon - \eta_h\|_Y \quad \forall \eta_h \in Y_h \\
&\implies = \|\delta\| = \inf_{\eta_h \in Y_h}\|\varepsilon - \eta_h\|_Y
\end{aligned}
$$

□

In the DPG method, discrete test spaces are constructed to allow for the problem to have specific properties, such as an SPD stiffness matrix, or the existence of a projection operator. Because δ and the bounds above by Strang's lemma are affected by the subspace of Y_h, the selection of the discrete test space has a significant impact on the error bounds.

3.5 Adaptive Error Estimator for Semi-Linear Problems

Classical finite element methods would involve a mesh in which the basis functions and mesh were refined uniformly across all elements. This, however, is computationally intensive, as the number of elements may increase exponentially at each refinement iteration.

In particular, this amount of refinement would not be necessary for certain problems. If the solution to an equation was easily approximated except at a few specific points, refining the mesh in the easier to approximate areas would not improve the error of the approximated solution by much, while still increasing the problem size. Thus, it may be more useful to only refine areas of the mesh where the error is large.

This gives way to a posteriori error estimators, which estimate the error on each element after each refinement step. After the spaces are refined, the space is refined specifically at elements where the error estimator indicated high error.

Theorem [A Posteriori Error for Semi-linear DPG problems]. Given the assumptions above, we can establish similar equivalence arguments as in the linear case. i.e. we have:

1. Efficiency: $\eta \leqslant k_1 \|x - x_h\|_{H^1}$
2. Reliability: $\|x - x_h\|_{H^1} \leqslant C(c, A, B) \inf_{\xi_h \in X_h} \|x - \xi_h\|_{H^1}$
3. Control of the oscillation term: $\|(L - Bx_h) \cdot (I - \Pi)\| \leqslant C(c, A, B, I - \Pi) \inf_{\xi_h \in X_h} \|x - \xi_h\|_{H^1}$

Proof.

$$\eta = \|L - Cx_h\|_{Y_h'} \leqslant \|L - Cx_h\|_{H^{-1}} \leqslant k_1 \|x - x_h\|_{H^1}$$

Since the residual operator has a bounded derivative, efficiency follows from Taylor expansion, and k_1 above depends on the norm of the derivative operator in Ω

We again use the fact that $\delta = \varepsilon - \varepsilon_h \perp Y_h$, and the third assumption regarding the projection Π, and acquire the same result:

$$\|\delta\|_{H^1} \leqslant \|(L - Bx_h) \circ (I - \Pi)\|_{H^{-1}} + \|\varepsilon_h\|_{H^1} \|\Pi\|_{H^{-1}}$$

Since the residual operator has an invertible derivative and the inverse of that derivative is bounded:

$$\begin{aligned} k_2\|x - x_h\|^2 &\leqslant \|L - Cx_h\|^2_{H^{-1}} = \|\varepsilon\|^2_{H^1} = \|\delta\|^2_{H^1} + \|\varepsilon_h\|^2_{H^1} \\ &\leqslant \|\varepsilon_h\|^2_{H^1} + (\|(L - Bx_h) \circ (I - \Pi)\|_{H^{-1}} + \|\varepsilon_h\|_{H^1}\|\Pi\|_{H^{-1}})^2 \end{aligned}$$

Where k_2 depends on the norm of the inverse of the derivative

We use the coercivity, monotonicity, and critical growth assumptions to establish the following result:

$$\begin{aligned} c\|x - x_h\|^2_{H^1} &\leqslant a(x - x_h, x - x_h) \\ &= a(x - x_h, x - \xi_h) + a(x - x_h, \xi_h - x_h) \\ &= a(x - x_h, x - \xi_h) + (b(x) - b(x_h), x_h - \xi_h) \\ &= a(x - x_h, x - \xi_h) + (b(x) - b(x_h), x - \xi_h) \\ &\quad - (b(x) - b(x_h), x - x_h) \\ &\leqslant a(x - x_h, x - \xi_h) + (b(x) - b(x_h), x - \xi_h) \\ &\leqslant C(A, B)\|x - x_h\|\|x - \xi_h\|_{H^1} \end{aligned}$$

It follows that: $\|x - x_h\|_{H^1} \leqslant C(c, A, B) \inf_{\xi_h \in X_h} \|x - \xi_h\|_{H^1}$

Using the result above, we can bound $\|(L - Bx_h) \cdot (I - \Pi)\|$ by $C(c, A, B, I - \Pi) \inf_{\xi_h \in X_h} \|x - \xi_h\|_{H^1}$:

$$\begin{aligned} (L - Bx_h) \cdot (I - \Pi)(y) &= (Ax + Bx - Bx_h)(y - \Pi y) \\ &= (Ax - A\xi_h + Bx - Bx_h)(y - \Pi y) \\ &= (Ax - A\xi_h)(y - \Pi y) + (Bx - Bx_h)(y - \Pi y) \\ \implies \|(L - Bx_h) \cdot (I - \Pi)\| &\leqslant \|A\|\|x - \xi_h\|\|I - \Pi\| + k_3\|x - x_h\|\|I - \Pi\| \\ &\leqslant C(c, A, B, I - \Pi) \inf_{\xi_h \in X_h} \|x - \xi_h\|_{H^1} \end{aligned}$$

□

References

[1] J. H. Bramble, R. D. Lazarov, and J. E. Pasciak. A least-squares approach based on a discrete minus one inner product for first order systems. *Math. Comp.*, 66:935–955, 1997.

[2] J. H. Bramble and J. E. Pasciak. A new approximation technique for div-curl systems. *Math. Comp.*, 73:1739–1762, 2004.

[3] C. Carstensen, P. Bringmann, F. Hellwig, and P. Wriggers. Nonlinear discontinuous Petrov–Galerkin methods. *Num. Math.*, 139:529–561, 2018.

[4] C. Carstensen, L. Demkowicz, and J. Gopalakrishnan. A posteriori error control for dpg methods. *SIAM Journal on Numerical Analysis*, 52(3):1335–1353, 2014.

[5] P. G. Ciarlet and P. A. Raviart. Maximum principle and uniform convergence for the finite element method. *Computer Methods in Applied Mechanics and Engineering*, 2:17–31, 1973.

[6] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous petrov–galerkin methods. ii. optimal test functions. *Numer. Methods Partial Differential Eq.*, 27:70–105, 2011.

[7] W. Dorfler. A convergent adaptive algorithm for Poisson's equation. *SIAM J. Numer. Anal.*, 33:1106–1124, 1996.

[8] T. Führer. Ultraweak formulation of linear pdes in nondivergence form and dpg approximation. *Computers & Mathematics with Applications*, 95:67–84, 2021. Recent Advances in Least-Squares and Discontinuous Petrov–Galerkin Finite Element Methods.

[9] M. Holst, G. Tsogtgerel, and Y. Zhu. Local convergence of adaptive methods for nonlinear partial differential equations. Available as arXiv:1001.1382 [math.NA].

[10] B. Keith, P. Knechtges, N. Roberts, S. Elgeti, M. Behr, and L. Demkowicz. An ultraweak dpg method for viscoelastic fluids. *Journal of Non-Newtonian Fluid Mechanics*, 247:107–122, 2017.

[11] P. Lax and A. Milgram. Ix. parabolic equations. *Contributions to the Theory of Partial Differential Equations*, 33:167–190, 2016.

[12] N. V. Roberts, T. Bui-Thanh, and L. Demkowicz. The dpg method for the stokes problem. *Comput. Math. Appl.*, 67:966–995, 2014.

[13] N. V. Roberts, L. Demkowicz, and R. Moser. Galerkin methodology for adaptive solutions to the incompressible navier–stokes equations. *J. Comput. Phys.*, 301:456–483, 2015.

[14] V. Santos. On the strong maximum principle for some piece-wise linear finite element approximate problems of non-positive type. *Journal of the Faculty of Science, University of Tokyo: Mathematics*, 29:473, 1982.

[15] R. Verfürth. *A posteriori error estimation techniques for finite element methods.* OUP Oxford, 2013.

[16] J. Xu and L. Zikatanov. Some observations on Babuška and Brezzi theories. *Num. Math.*, 94(1):195–202, 2003.