Spring 5-10-2016

# Discrete Stability of DPG Methods

Ammar Harb
*Portland State University*

# Discrete Stability of DPG Methods

by

Ammar Harb

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy
in
Mathematical Sciences

Dissertation Committee:
Jay Gopalakrishnan, Chair
Gerardo Lafferriere
Bin Jiang
Jeffrey Ovall
Christopher Butenhoff

Portland State University
2016

**Abstract**

This dissertation presents a duality theorem of the Aubin-Nitsche type for discontinuous Petrov Galerkin (DPG) methods. This explains the numerically observed higher convergence rates in weaker norms. Considering the specific example of the mild-weak (or primal) DPG method for the Laplace equation, two further results are obtained. First, for triangular meshes, the DPG method continues to be solvable even when the test space degree is reduced, provided it is odd. Second, a non-conforming method of analysis is developed to explain the numerically observed convergence rates for a test space of reduced degree. Finally, for rectangular meshes, the test space is reduced, yet the convergence is recovered regardless of parity.

I dedicate this dissertation to my mother and my father who passed away on 12/8/2015 and 8/12/2008, respectively.

**Acknowledgments**

First and foremost, I would like to thank my advisor, Prof. Jay Gopalakrishnan for his excellent guidance, patience, caring, , and creating a very pleasant environment in which to discuss and work.

I would like to express my gratitude to my committee members (Professors Gerardo Lafferriere, Bin Jiang, Jeffrey Ovall, Christopher Butenhoff) who have agreed graciously to serve on my PhD committee, spent time thinking about my work, and provided invaluable feedback.

I would like to thank Mike Karkulik, postdoc at PSU, and Tathagata Goswami, PhD student at PSU, for helping me in debugging my work and for helpful discussions.

I would like to thank my colleagues Nicole Olivares, Paulina Seplveda, and Timaeus Bouma. Their insights, commentary and encouragement, shared in many conversations, were and continue to be a great asset to my scientific life.

I am deeply grateful for the constant support by my family and friends, especially my older brother, Amer, and my wife, Muna. Without them, I would never have been able to finish my dissertation. Thank you.

Finally, This work was partially supported by the AFOSR under grant FA9550-12-1-0484 and by the NSF under grant DMS-1318916.

**Table of Contents**

**Abstract** i

**Dedication** ii

**Acknowledgments** iii

**List of Tables** vi

**List of Figures** vii

**Chapter 1: Introduction** 1
- 1.1 Motivation . . . 2
- 1.2 Literature Review . . . 3
  - 1.2.1 The DG Methods . . . 4
  - 1.2.2 The HDG Methods . . . 6
  - 1.2.3 The DPG Method . . . 7
  - 1.2.4 Scope . . . 8

**Chapter 2: The DPG Method** 10
- 2.1 General results . . . 11
  - 2.1.1 The Error Analysis of the Practical DPG Method . . . 14
  - 2.1.2 DPG Method as a Mixed Method . . . 15
  - 2.1.3 Weakly conforming test space . . . 17
  - 2.1.4 Injectivity . . . 19
- 2.2 Application to the Poisson Equation . . . 21
  - 2.2.1 The DPG Approximation . . . 24
- 2.3 Numerical Results . . . 26
- 2.4 DPG Convergence Rates for Singularities . . . 30

**Chapter 3: Reduced Degree DPG Methods Based on Parity** **35**
3.1 Introduction . . . 36
3.2 Explaining the even-odd separation . . . 37
3.3 Numerical Results . . . 44

**Chapter 4: Nonconforming Analysis** **47**
4.1 Introduction . . . 48
4.2 Case 3: A nonconforming analysis . . . 48
4.3 Numerical Analysis . . . 52

**Chapter 5: Duality Argument** **55**
5.1 Introduction . . . 56
5.2 General Settings . . . 56
5.3 Case 1: Application of the duality argument . . . 59
5.4 Application to the Helmholtz Equation with Impedance Boundary Condition . . . 63

**Chapter 6: The Convergence Rates of The DPG Method with Rectangular Meshes** **69**
6.1 Introduction . . . 70
6.2 Application to the Poisson Equation . . . 70
6.3 Error Analysis . . . 73
6.4 Reduced-Degree Test Spaces . . . 76
6.5 Numerical Results . . . 85

**Chapter 7: Future Work and Conclusion** **94**
7.1 Accomplishments . . . 96
7.2 Future work . . . 97

**References** **99**

## List of Tables


Table 2.1 Case 1: $(k_u, k_q, k_v) = (k, k-1, k+1)$ . . . . . 32
Table 2.2 Case 1: $(k_u, k_q, k_v) = (k, k-1, k+1)$ . . . . . 33
Table 2.3 Case 1: $(k_u, k_q, k_v) = (k, k-1, k+1)$ . . . . . 34
Table 2.4 Case 1: $(k_u, k_q, k_v) = (k, k-1, k+1)$ . . . . . 34

Table 3.1 Summary of numerically observed convergence rates . . . . . 37

Table 4.1 Summary of numerically observed convergence rates . . . . . 48

Table 6.1 Summary of the convergence rates . . . . . 73
Table 6.2 $H^1(\Omega)$ and $L^2(\Omega)$ convergence for the case 7 . . . . . 92
Table 6.3 $H^1(\Omega)$ and $L^2(\Omega)$ convergence for the case 8 . . . . . 93
Table 6.4 Summary of the convergence rates . . . . . 93

## List of Figures


Figure 2.1 Case 1 $H^1-$Error . . . . . 28
Figure 2.2 Case 1 $L^2-$Error . . . . . 29
Figure 2.3 Singularity (k=1, n=16) . . . . . 32
Figure 2.4 Singularity (k=10, n=2) . . . . . 33

Figure 3.1 Case 2 $H^1-$Error . . . . . 45
Figure 3.2 Case 2 $L^2-$Error . . . . . 46

Figure 4.1 Case 3 $H^1-$Error . . . . . 53
Figure 4.2 Case 3 $L^2-$Error . . . . . 54

Figure 6.1 Case 4 $H^1-$Error . . . . . 87
Figure 6.2 Case 5 $H^1-$Error . . . . . 88
Figure 6.3 Case 4 $L^2-$Error . . . . . 89
Figure 6.4 Case 5 $L^2-$Error . . . . . 90
Figure 6.5 Case 6 $L^2-$Error . . . . . 91

# Chapter 1

# Introduction

## 1.1 Motivation

The discontinuous Petrov-Galerkin (DPG) method was introduced in 2009 by Demkowicz and Gopalakrishnan [16, 17, 21]. It started with analyzing spectral methods for the simplest 1D convection problem. The DPG method guarantees the best approximation property in the so-called energy (dual or residual) norm. This best approximation is obtained by calculating optimal test functions which realize the supremum of the inf-sup condition. Using spaces of the Discontinuous Galerkin (DG) method, the calculation of the optimal test functions can be done locally for each element of the mesh. Like a hybrid method, the DPG method also has interface variables.

Our goal is to reduce the cost of the method as much as possible, while maintaining the convergence rate of the method. Part of this work focuses on reducing the order of the test spaces. This has many advantages. First, consider the left hand side matrix of the linear system arising from the DPG method. As we shall see later, its assembly requires computation of the Gram matrix of the test space. Even though this matrix is block diagonal, it is of some practical interest to reduce the block size, especially when operating near the limit of memory bandwidth in multi-core architectures. Second, consider the right hand side computation. In cases where load terms are expensive to evaluate, reduction of test space degree brings significant computational savings. Finally, the third and the most compelling reason that prompted us to investigate this issue, is that there are practical limits on the degree of polynomials one can use in most finite element software. We prefer to hit this practical limiting degree with the trial space, rather than with the test space, because it is the approximation properties of the trial space

that determines the final solution quality.

As an example to illustrate the main points, we will use the Poisson equation with Dirichlet boundary condition,

$$-\Delta u = f \qquad \text{on } \Omega, \tag{1.1a}$$

$$u = 0 \qquad \text{on } \partial\Omega. \tag{1.1b}$$

There are two DPG methods for the Poisson equation. One is based on an ultra-weak formulation [18] (where constitutive and conservation equations are both integrated by parts) while the other is based on the so-called mild-weak, or primal formulation, developed in [8, 19] (where only the conservation equation is integrated by parts). The example which motivates our study is the latter.

## 1.2 Literature Review

Finite element methods (FEM) have been proven to be extremely useful in the numerical approximation of the self-adjoint elliptic partial differential equations solutions [22]. It can be applied to very general and complicated geometries of interest. In addition, there are tools for their error analysis, which depends on the variational interpretation of the FEM as a minimization problem over finite-dimensional spaces. However, the use of (classical) FEM for the numerical solution of hyperbolic problems and other strongly non-self-adjoint PDE problems is, generally speaking, not satisfactory. These problems do not arise naturally in a variational setting. Indeed, the use of FEM for such problems has been a subject of research in the 1970s and 1980s and for most of the 1990s. Instead, finite volume

methods (FVM) have been predominantly used in industrial software packages for the numerical solution of hyperbolic systems, especially in the area of Computational Fluid Dynamics [22]. This motivated the introduction of a new class of FEM, namely, the discontinuous Galerkin (DG) FEM.

### 1.2.1 The DG Methods

In 1971 Reed and Hill [32] proposed the DG method for the numerical solution of the nuclear transport PDE problem. They combine features of the finite element and the finite volume methods (FVM) and have been successfully applied to hyperbolic, elliptic, parabolic and mixed form problems arising from a wide range of applications. These methods were later analyzed by LeSaint and Raviart [28] (error estimate for tensor product mesh) and by Johnson and Pitkaranta [27]. In the area of elliptic problems, Nitsche's work on weak imposition of essential boundary conditions [30] for (classical) FEM, allowed for finite element solution spaces that do not satisfy the essential boundary conditions.

Few years later, Baker [4] proposed the first modern discontinuous Galerkin method for elliptic problems, later followed by Wheeler [33], Arnold [1] and others. An application of the penalty method to the finite element method is analyzed by Babuska [3]. The DG methods were applied to many partial differential equations. A good review can be found in the reference [14] by Cockburn et al.

DG methods exhibit attractive properties for the numerical approximation of problems of hyperbolic or nearly-hyperbolic type, compared to both classical FEM and FVM. Indeed, in contrast with classical FEM, but together with FVM, DG methods are, by construction, locally (or nearly locally) conservative with respect to the state variable; moreover, they exhibit enhanced stability properties in the

vicinity of sharp gradients (e.g., boundary or interior layers) and/or discontinuities which are often present in the analytical solution of convection/transport dominated PDE problems. Additionally, DG methods offer advantages in the context of automatic local mesh and order adaptivity, such as increased flexibility in the mesh design (irregular grids are admissible) and the freedom to choose the elemental polynomial degrees without the need to enforce any conformity requirements. The implementation of genuinely (locally varying) high-order reconstruction techniques for FVM still remains a computationally difficult task, particularly on general unstructured hybrid grids. Therefore, DG methods emerge as a very attractive class of arbitrary order methods for the numerical solution of various classes of PDE problems where classical FEM are not applicable and FVM produce typically low order approximations. Finally, The DG methods are attracting the interest of many scientists because:

- They enforce the equations in an element-by-element fashion through a Galerkin formulation which can give rise to locally conservative methods.
- They can handle any type of mesh, element shape and basis functions.
- They are ideally suited for $hp-$adaptivity.
- They have a built-in stabilization mechanism which does not degrade their (high-order) accuracy.
- They can be applied to a wide variety of partial differential equations.

However, the DG methods (for second-order elliptic equations) have been criticized because:

- For the same mesh and the same polynomial degree, the number of globally coupled degrees of freedom of the DG methods is much bigger than those of the standard FEM. Moreover, the orders of convergence of both the vector and scalar variables are also the same.

- For the same mesh and the same index, the number of globally coupled degrees of freedom of the DG methods are much bigger than those of the hybridized version of the RT and BDM methods. Moreover, the orders of convergence of both the vector and the local average of the scalar variables are smaller by one.

### 1.2.2 The HDG Methods

Discontinuous function spaces are used in the DG methods which creates a major disadvantage: unlike in continuous Galerkin (CG) methods (standard FEM), degrees of freedom are not shared between elements. As a consequence, the number of unknowns is substantially higher compared to a CG discretization. Especially for implicit time discretization this imposes large memory requirements, and potentially leads to increased time-to-solution. In order to avoid these disadvantages, a technique called hybridization may be utilized (see [12]), resulting in hybridized discontinuous Galerkin (HDG) methods introduced by Cockburn, Gopalakrishnan and Lazarov [13]. Here, the globally coupled unknowns have support on the mesh skeleton, i.e. the element interfaces(numerical traces and numerical fluxes resulted from integration by parts), only. This reduces the size of the global system and coincidentally improves the sparsity pattern.

The HDG method is created using shape functions with support only on element

edges. These are interface or hybrid variables. The value of the solution inside element interiors can be recovered from interface variables. The HDG methods is considered as one of the stabilized methods. Stabilization techniques are employed through the choice of the HDG numerical flux. The HDG methods are obtained by discretizing characterizations of the exact solution written in terms of many local problems, one for each element of the mesh, with suitably chosen data, and in terms of a single global problem that actually determines them. This permits efficient implementation since they inherit the above-mentioned structure of the exact solution. This is what renders them efficiently implementable, especially within the framework of $hp-$adaptive methods, as is typical of DG methods.

The way in which they are defined allows them to be, in some instances, more accurate than already existing DG methods. In fact, in some cases when standard DG methods do not converge, HDG methods do. The HDG methods can be used for steady-state problems and for time-dependent problems when implicit time-marching methods are used.

HDG methods use a characterization of the exact solution in terms of solutions of local problems and transmission conditions. Also it uses discontinuous approximations for both the solution inside each element and its trace on the element boundary. The local solvers is defined by using a Galerkin method to weakly enforce the equations on each element. It defines a global problem by weakly imposing the transmission conditions.

### 1.2.3 The DPG Method

The Discontinuous Petrov-Galerkin (DPG) was introduced by Demkowicz and Gopalakrishnan [16,17,21]. Petrov-Galerkin methods generalize the Galerkin method

(Bubnov-Galerkin method) by allowing distinct trial and test spaces. The trial space is where the solution is sought. The test space is used to enforce the equations. In DPG method, one computes "optimal" test functions and they are called "optimal" since the supremum in the inf-sup condition is attained at them. With such optimal test functions, the discrete variational form of the problem inherits stability from the continuous variational form of it. The choice of the inner product on the test space affects the computation of the optimal test functions, so it is important in the method. Computing those optimal test functions is an extra step in the DPG method compared to other finite element methods. In addition, these test spaces can have discontinuities on the boundaries of the mesh elements, which makes the calculations of the optimal test functions local and inexpensive. By the DPG method, we get optimal solution in an energy norm and it produces a symmetric positive definite stiffness matrix.

There are several advantages of the DPG method over other finite element method: (a) stability is guaranteed, (b) its stiffness matrix is always symmetric and positive definite so we can use iterative solvers, and (c) it has an error representation function that can be used for a-posteriori error estimation and adaptivity.

It is considered as a hybrid DG method and can be thought of as a least-square method in nonstandard norms, or as Petrov-Galerkin methods with special test spaces, or as a nonstandard mixed method.

### 1.2.4 Scope

This dissertation will proceed in five main parts. We will begin by introducing the abstract Discontinuous Petrov-Galerkin (DPG) method as a mixed method for linear problems highlighting some important properties of the method. Our next

step will be presenting the new reduced degree DPG method for odd and even degrees. Then, we will analyze the DPG method by nonconforming analysis using Strang lemma. Next, the duality argument for the DPG method will be presented. Finally, We will study the convergence rates of the DPG method for rectangular meshes and show how we can get the same convergence rate with a reduced test space.

# Chapter 2

# The DPG Method

In this chapter, we introduce the DPG method. First, we begin with the general settings of the method. Then, we introduce the main example in this study, namely, the Poisson Equation with Dirichlet boundary conditions.

## 2.1 General results

We start with a linear variational problem, and we want to approximate $x \in X$ satisfying

$$b(x, y) = \ell(y) \qquad \forall y \in Y. \tag{2.1}$$

Here $X$ is called a trial space and $Y$ is called a test space. We assume that $X$ and $Y$ are Hilbert spaces, b is a bilinear and continuous form on $X \times Y$ ($b(\cdot, \cdot) : X \times Y \to \mathbb{R}$) and $\ell$ is a linear continuous form on $Y$, i.e. an element of the dual space $Y^*$. Here and throughout $(.,.)_Y$ denotes the inner product in Y with the corresponding norm $\|y\|_Y$.

Suppose $X_0$ and $\hat{X}$ are Hilbert spaces over $\mathbb{R}$. Solutions are sought in the "trial space" $X = X_0 \times \hat{X}$ and have an "interior" component in $X_0$ and an "interface" component in $\hat{X}$. Suppose there are continuous bilinear forms $\hat{b}(\cdot, \cdot) : \hat{X} \times Y \to \mathbb{R}$ and $b_0(\cdot, \cdot) : X_0 \times Y \to \mathbb{R}$ be set by

$$b(\,(w, \hat{w}), y\,) = b_0(w, y) + \hat{b}(\hat{w}, y), \tag{2.2}$$

for all $(w, \hat{w}) \in X$ and $y \in Y$.

Given any $\ell \in Y^*$ we are interested in approximating an $x \equiv (x_0, \hat{x}) \in X$ satisfying equation 2.1.

**Definition 2.1.** *Given any trial space $X_h$, we define it optimal test space for the*

*continuous bilinear form* $b(.,.) : X \times Y \to \mathbb{R}$ *by*

$$Y_h^{opt} = T(X_h)$$

*where* $T : X \to Y$ *(the trial-to-test operator) be defined by*

$$(Tw, y)_Y = b(w, y), \quad \forall y \in Y,\, w \in X. \tag{2.3}$$

Equation 2.3 uniquely defines a $Tw$ for any given $w \in X$, by Riesz representation theorem. We call $Tw$ the "optimal" test function of w, because it solves an optimization problem, as we see next.

**Proposition 2.2.** *(Optimizer). For any* $w \in X$*, the maximum of*

$$f_w(y) = \frac{|b(w, y)|}{\|y\|_Y}$$

*over all nonzero* $y \in Y$ *is attained at* $y = Tw$*.*

*Proof.* By duality in Hilbert spaces,

$$\sup_{0 \neq y \in Y} f_z(y) = \sup_{0 \neq y \in Y} \frac{|(Tz, y)_Y|}{\|y\|_Y} = \|Tz\|_Y$$

and $f_z(Tz) = \|Tz\|_Y$. □

The DPG approximation to $x \equiv (x_0, \hat{x}) \in X$ , lies in a finite dimensional trial subspace $X_h \subseteq X$ (where $h$ denotes a parameter determining the finite dimension).

Let $X_{h,0} \subseteq X_0$ and $\hat{X}_h \subseteq \hat{X}$ be finite-dimensional subspaces and let $X_h = X_{h,0} \times \hat{X}_h$. So the "ideal" DPG method reads: find $x_h \in X_h$ satisfying

$$b(x_h, y) = \ell(y), \qquad \forall y \in Y_h^{opt} = T(X_h). \tag{2.4}$$

Since the trial space is different from the test space in general, this is a Petrov-Galerkin approximation, which is well-posed [17, 18]. The main difficulty of the ideal method is that in order to compute $x_h$ (see [25]) , one needs a basis for $Y_h^{opt}$, which must be obtained by applying $T$. This is infeasible, as seen from 2.3, if $Y$ is infinite dimensional, unless a solution to 2.3 can be written out in closed form. In certain one-dimensional problems, and in some multi-dimensional problems, like the transport equation, the application of $T$ can be exactly written out in closed form (see [16, 17]). But for the vast majority of interesting problems, this is not possible. To overcome the above-mentioned difficulty, we need to define $Y^r$ which is a finite-dimensional subspace of $Y$ and approximate the operator $T$ by $T^r$, where $T^r : X \to Y^r$ be defined by

$$(T^r w, y)_Y = b(w, y), \quad \forall y \in Y^r,\, w \in X_h. \tag{2.5}$$

Instead of solving the "ideal" DPG method 2.4 in a closed form, we will solve the so-called "practical" DPG method(see [25]). For (2.1), the "practical" DPG method computes $x_h \equiv (x_{h,0}, \hat{x}_h)$ in $X_h$ satisfying

$$b(x_h, y) = \ell(y), \qquad \forall y \in Y_h^r = T^r(X_h). \tag{2.6}$$

**Remark 2.3.** *The process of introducing the interface space $\hat{X}_h$ is called hybridization. $\hat{X}_h$ is a space of new unknowns of interelement fluxes and traces. It localizes the action of the operator $T^r$ with the use of a space $Y$ that contains functions discontinuous across mesh element interfaces. This then implies that the Gram matrix becomes block diagonal, with one block per mesh element (since $Y^r$ may now be chosen to be a DG subspace). The application of $T^r$ is thus reduced to an easy block diagonal inversion.*

### 2.1.1 The Error Analysis of the Practical DPG Method

In this section, we will show the discrete stability of the practical DPG method. So we need to prove the discrete inf-sup condition over the space $Y^r$, which can be done easily by proving the existence of a Fortin operator into $Y^r$. A fundamental quasioptimality result for the practical DPG methods is stated in Theorem 2.6 below. It holds under these assumptions.

**Assumption 2.4.** *Suppose $\{z \in X : \ b(z,y) = 0, \ \forall y \in Y\} = \{0\}$ and suppose there exist $C_1, C_2 > 0$ such that*

$$C_1\|y\|_Y \leqslant \sup_{0\neq z\in X} \frac{|b(z,y)|}{\|z\|_X} \leqslant C_2\|y\|_Y \qquad \forall y \in Y. \tag{2.7}$$

**Assumption 2.5.** *There is a linear operator $\Pi : Y \to Y^r$ and a $C_\Pi > 0$ such that*

*for all* $w_h \in X_h$ *and all* $v \in Y$*,*

$$b(w_h, v - \Pi v) = 0, \qquad \text{and} \qquad \|\Pi v\|_Y \leqslant C_\Pi \|v\|_Y.$$

**Theorem 2.6** (see [25])**.** *Suppose Assumptions 2.4 and 2.5 hold. Then the DPG method* (2.6) *is uniquely solvable for* $x_h$ *and*

$$\|x - x_h\|_X \leqslant \frac{C_2 C_\Pi}{C_1} \inf_{z_h \in X_h} \|x - z_h\|_X$$

*where* $x$ *is the unique exact solution of* (2.1)*.*

### 2.1.2 DPG Method as a Mixed Method

Another well-known result, motivated by [15], is an equivalence of the DPG method with a mixed Bubnov-Galerkin formulation, so it is easily implementable in codes without support for Petrov-Galerkin forms. Instead of identifying the second argument in the formulation 2.6 as the optimal test function, we identify the first argument as the error representation function. To state it, we first define the error representation function: let $\varepsilon^r$ be the unique element of $Y^r$ satisfying

$$(\varepsilon^r, y)_Y = \ell(y) - b(x_h, y), \forall y \in Y^r. \tag{2.8}$$

**Theorem 2.7.** *The following are equivalent statements:*

*i)* $x_h \in X_h$ *solves the DPG method* (2.6)*.*

*ii)* $x_h \in X_h$ *and* $\varepsilon^r \in Y^r$ *solve the mixed formulation*

$$(\varepsilon^r, y)_Y + b(x_h, y) = \ell(y) \qquad \forall y \in Y^r, \tag{2.9a}$$

$$b(z_h, \varepsilon^r) = 0 \qquad \forall z_h \in X_h. \tag{2.9b}$$

Its simple proof is omitted (see e.g. [23]).

Thus, the method comes with a "built-in" a-posteriori error estimation or, more precisely, a-posteriori error evaluation. This is useful for $hp-$adaptivity. So, no need to code an error estimator for driving adaptivity in DPG methods.

**Remark 2.8.** *The norm of* $\varepsilon^r$ *is bounded by the error: Choosing* $y = \varepsilon^r$ *in* (2.8)*, we obtain*

$$\|\varepsilon^r\|_Y^2 = (\varepsilon^r, \varepsilon^r)_Y = \ell(\varepsilon^r) - b(x_h, \varepsilon^r) = b(x - x_h, \varepsilon^r).$$

*Hence, by Assumption 2.4,*

$$\|\varepsilon^r\|_Y \leqslant C_2 \|x - x_h\|_X. \tag{2.10}$$

*This theme is further developed in [10], where* $\|\varepsilon^r\|_Y$ *is established to be both a reliable and an efficient error estimator.*

### 2.1.3 Weakly conforming test space

In this section we discuss an alternate interpretation of the DPG method based on the concept of global optimal test functions (see [20]).

In order to study a relation between the approximate local and global test functions, we will consider the abstract notation for the bilinear form 2.2. The variable $\hat{w}$ denotes the unknown abstract trace defined on the internal skeleton only. For global-conforming test functions, the second term vanishes. To see that, we define a weakly conforming test space as follows,

$$Y_0^r = \{y \in Y^r : \ \hat{b}(\hat{w}_h, y) = 0, \ \ \forall \hat{w}_h \in \hat{X}_h\} \tag{2.11}$$

and let $T_0^r : X_0 \to Y_0^r$ be defined by $(T_0^r w, y)_Y = b_0(w, y)$ for all $y \in Y_0^r$. In the examples we have in mind, $Y^r$ is a discontinuous Galerkin (DG) space, and $Y_0^r$ is a subspace with weak interelement continuity constraints, i.e., a weakly conforming space. In such cases, the application of the operator $T_0^r$ requires a global inversion. We then compare these two DPG methods:

$$\text{Find } (x_{h,0}, \hat{x}_h) \in X_h : \quad b(\,(x_{h,0}, \hat{x}_h), y\,) = \ell(y) \qquad \forall y \in Y_h^r \equiv T^r(X_h). \tag{2.12a}$$

$$\text{Find } x_{h,0} \in X_{h,0} : \quad b_0(x_{h,0}, y) = \ell(y) \qquad \forall y \in Y_{h,0}^r \equiv T_0^r(X_{h,0}). \tag{2.12b}$$

The first is the same as (2.6), the standard DPG method. We view (2.12a) as a "hybridized" form of the second method (2.12b), and the next theorem shows in what sense they are equivalent. The method (2.12b) is not the preferred for implementation due to the expense of applying $T_0^r$, but we will use it later for error analysis.

**Theorem 2.9.** *The test spaces satisfy* $Y^r_{h,0} \subset Y^r_h$. *Hence, if* $(x_{h,0}, \hat{x}_h) \in X_h$ *solves* (2.12a), *then* $x_{h,0}$ *solves* (2.12b).

*Proof.* Let $Y^r_\perp$ be the $Y$-orthogonal complement of $Y^r_h$ in $Y^r$. Then we have the orthogonal decomposition

$$Y^r = Y^r_h + Y^r_\perp \tag{2.13}$$

where $Y^r_\perp$ is the $Y$-orthogonal complement of $Y^r_h$ in $Y^r$. Let $y_0 \in Y^r_{h,0}$. Apply (2.13) to decompose $y_0 = y_h + y_\perp$, with $y_h \in Y^r_h$ and $y_\perp \in Y^r_\perp$.

First, we claim that $y_\perp \in Y^r_0$. This is because

$$\hat{b}(\hat{w}_h, y_\perp) = (T^r(0, \hat{w}_h), y_\perp)_Y = 0 \qquad \forall \hat{w}_h \in \hat{X}_h.$$

The last identity followed from the orthogonality of $y_\perp$ to $T^r(X_h)$.

Next, we claim that $y_\perp = 0$. It suffices to prove that $(y_0, y_\perp)_Y = 0$ since $(y_0, y_\perp)_Y = \|y_\perp\|^2_Y$. Since $y_0 \in Y^r_{h,0}$, there is a $w_h \in X_{h,0}$ such that $y_0 = T^r_0 w_h$. Then,

$$\begin{aligned} (y_0, y_\perp)_Y = (T^r_0 w_h, y_\perp)_Y = b_0(w_h, y_\perp) \qquad & \text{as } y_\perp \in Y^r_0 \\ = (T^r(w_h, 0), y_\perp)_Y = 0 \qquad & \text{as } T^r(X_h) \perp y_\perp. \end{aligned}$$

Finally, since $y_\perp = 0$, we have $y_0 = y_h + 0 \in Y^r_h$. Thus $Y^r_{h,0} \subset Y^r_h$. The second statement of the theorem is now obvious by choosing $y \in Y^r_{h,0}$ in (2.12a). $\square$

We conclude that the practical DPG method may be interpreted simply as a localization of the corresponding global PG methodology (see [20]). From the analysis point of view, it looks like one can deemphasize convergence of traces (and fluxes) and focus on studying the convergence of the interior variables only. In particular, a discretization of traces with discontinuous elements is non-conforming from the point of view of the DPG method but it is perfectly OK from the point of view of the global PG method and non-conforming discretization of optimal test functions. As the approximation of optimal test functions in non-conforming, one has to account for both approximation and consistency errors using the Second Strang Lemma.

### 2.1.4 Injectivity

The bilinear forms $b(.,.)$ and $\hat{b}(.,.)$ generate operators. There is a relation between these operators, namely, the injectivity of one of them implies the injectivity of the other under some assumptions. By the injectivity, we guarantee the unique solvability of the method.

Let $B_h : X_h \to (Y^r)^*$ be the operator generated by the form $b(\cdot,\cdot)$, i.e.,

$$(B_h w_h)(y) = b(w_h, y), \qquad \forall w_h \in X_h,\ y \in Y^r.$$

Similarly, let $\hat{B}_h : \hat{X}_h \to (Y^r)^*$ be defined by

$$(\hat{B}_h \hat{z}_h)(y) = \hat{b}(\hat{z}_h, y), \qquad \forall \hat{z}_h \in \hat{X}_h,\ y \in Y^r. \tag{2.14}$$

The injectivity of $B_h$ yields the unique solvability of the DPG method

**Assumption 2.10.** *Suppose*

*a)* $X_{h,0} \subseteq Y^r$,

*b)* $\hat{b}(\hat{z}_h, z_0) = 0$ *for all* $\hat{z}_h \in \hat{X}_h$ *and* $z_0 \in X_{h,0}$, *and*

*c) any* $z_0 \in X_{h,0}$ *satisfying* $b_0(z_0, z_0) = 0$ *must be zero.*

**Theorem 2.11.** *If* $B_h$ *is injective, then* $\hat{B}_h$ *is injective, and the DPG method* (2.6) *is uniquely solvable. Conversely, if* $\hat{B}_h$ *is injective, then* $B_h$ *is injective, provided Assumption 2.10 holds.*

*Proof.* Suppose $B_h$ is injective. The injectivity of $\hat{B}_h$ is obvious from $\hat{B}_h \hat{w}_h = B_h(0, \hat{w}_h)$. We also claim that $T^r$ is injective: Indeed, if $w_h \in X_h$ satisfies $T^r w_h = 0$, then $0 = (T^r w_h, y)_Y = b(w_h, y) = (B_h w_h)(y)$ for all $y \in Y^r$, so $w_h = 0$. The injectivity of $T^r$ implies that $\dim(Y_h^r) = \dim(X_h)$, so the DPG method (2.6) yields a square system. Moreover, since (2.6) is the same as

$$(T^r x_h, T^r w_h)_Y = \ell(T^r w_h) \qquad \forall\, w_h \in X_h,$$

the injectivity of $T^r$ also implies that there is a unique solution $x_h$ in $X_h$.

Now suppose $\hat{B}_h$ is injective. To prove that $B_h$ is injective, consider a $(w_0, \hat{w}) \in$

$X_h$ satisfying $B_h(w_0, \hat{w}) = 0$. Then

$$\begin{aligned} 0 &= (B_h(w_0, \hat{w}))(w_0) && \text{by Assumption 2.10(a)} \\ &= b\left((w_0, \hat{w}), w_0\right) = b_0(w_0, w_0) + \hat{b}(\hat{w}, w_0) \\ &= b_0(w_0, w_0), && \text{by Assumption 2.10(b).} \end{aligned}$$

Therefore, by Assumption 2.10(c), $w_0 = 0$. It only remains to show that $\hat{w} = 0$. But $(\hat{B}_h\hat{w})(y) = \hat{b}(\hat{w}, y) = b(\,(0, \hat{w}), y) = (B_h(w_0, \hat{w}))(y) = 0$ for all $y \in Y^r$. Hence the injectivity of $\hat{B}_h$ implies $\hat{w} = 0$. □

## 2.2 Application to the Poisson Equation

Suppose $\Omega$ is a bounded open polygon in $\mathbb{R}^2$ with Lipschitz boundary, meshed by $\Omega_h$, a geometrically conforming shape regular finite element mesh of triangles. Let $h = \max_{K\in\Omega_h} \operatorname{diam} K$. Let $\partial\Omega_h$ denote the collection of all element boundaries $\partial K$ for all elements $K$ in $\Omega_h$. We assume that $\partial K$ is Lipschitz for all $K \in \Omega_h$, so that we may use trace theorems on each element, but the shape of the elements is otherwise arbitrary. We now study the DPG approximation to the Dirichlet problem

$$-\Delta u = f \qquad \text{on } \Omega, \tag{2.15a}$$

$$u = 0 \qquad \text{on } \partial\Omega. \tag{2.15b}$$

All functions are real-valued in this section.

There are two DPG methods for the Laplace's equation. One is based on an ultra-weak formulation [18] (where constitutive and conservation equations are both integrated by parts) while the other is based on the so-called mild-weak, or primal formulation, developed in [8, 19] (where only the conservation equation is integrated by parts). The latter is the one which we are going to use in this dissertation. We obtain a weak formulation (as in [19]) by multiplying 2.15 by a test function and integrate by parts over each element $K$ then sum up to get the so-called the *primal* formulation, as follows:

Find $(u, \hat{q}_n) \in X = X_0 \times \hat{X}$ satisfying

$$(\nabla u, \nabla v)_{\Omega_h} - \langle \hat{q}_n, v \rangle_{\partial\Omega_h} = (f, v)_{\Omega_h} \quad \forall v \in Y \tag{2.16}$$

where the spaces and the notations used are defined below using the framework of section 2.1, as follows,

$$\begin{aligned}
\text{Set } X_0 = H_0^1(\Omega), \qquad & \hat{X} = H^{-1/2}(\partial\Omega_h), \\
Y = H^1(\Omega_h), \qquad & \text{where} \\
H^1(\Omega_h) = \{v :\ & v|_K \in H^1(K),\ \forall K \in \Omega_h\}, \\
H^{-1/2}(\partial\Omega_h) = \{\eta \in \prod_K & H^{-1/2}(\partial K) :\ \exists\, r \in H(\text{div}, \Omega) \text{ such that} \\
& \eta|_{\partial K} = r \cdot n|_{\partial K}, \quad \forall\, K \in \Omega_h\},
\end{aligned}$$

where $n$ denotes the unit outward normals on the boundary of mesh elements. The

space $H^{-1/2}(\partial\Omega_h)$ is normed, as in [31], by

$$\|\hat{r}_n\|_{H^{-1/2}(\partial\Omega_h)} = \inf\left\{\|r\|_{H(\mathrm{div},\Omega)} : \ \forall r \in H(\mathrm{div},\Omega) \text{ such that } \hat{r}_n|_{\partial K} = r\cdot n|_{\partial K}\right\}. \tag{2.17}$$

The "broken" Sobolev space $H^1(\Omega_h)$ is normed by

$$\|v\|^2_{H^1(\Omega_h)} = (v,v)_{\Omega_h} + (\nabla v, \nabla v)_{\Omega_h}, \tag{2.18}$$

Throughout the derivatives are always calculated element by element, and

$$(r,s)_{\Omega_h} = \sum_{K\in\Omega_h} (r,s)_K, \qquad \langle \ell, w\rangle_{\partial\Omega_h} = \sum_{K\in\Omega_h} \langle \ell, w\rangle_{1/2,\partial K}.$$

where $(\cdot,\cdot)_K$ denotes the $L^2(K)$-inner product and $\langle \ell, \cdot\rangle_{1/2,\partial K}$ denotes the action of a functional $\ell$ in $H^{-1/2}(\partial K)$.

The bilinear and linear forms of the weak formulation are set by

$$b_0(u,v) = (\nabla u, \nabla v)_{\Omega_h}, \qquad \hat{b}(\hat{q}_n, v) = -\langle \hat{q}_n, v\rangle_{\partial\Omega_h}, \qquad \ell(v) = (f,v)_{\Omega_h}.$$

This completes the definition of all the notations that appeared in 2.16.

**Remark 2.12.** *The variational formulation 2.16 is similar to the one in ( [7],p.141), which is called primal hybrid method. There, it is used as motivation to introduce hybrid and non-conforming methods. An important difference between the*

*two formulations is that, the formulation in [7] is standard Galerkin (or Bubnov-Galerkin), while 2.16 is a Petrov-Galerkin formulation since $X \neq Y$.*

The well-posedness of 2.16 is proven in [19], i.e. assumption 2.4 was verified for this formulation there. We will denote the exact solution of the resulting weak formulation (2.1) by $(u, \hat{q}_n) \in X$. Note that $\hat{q}_n|_{\partial K} = \partial_n u|_{\partial K}$ for all $K \in \Omega_h$.

### 2.2.1 The DPG Approximation

Consider applying the method on a two-dimensional domain $\Omega$ meshed by a geometrically conforming finite element mesh of triangles of mesh size $h$. The method produces an approximation $u_h$ to the solution $u$ of the Laplace's equation in the interior of the mesh elements, as well as an approximation to the flux $q$ on the element interfaces. The first is a polynomial of degree at most $k_u$ on each mesh element and the second is a polynomial of degree at most $k_q$ on each mesh edge. The method uses test functions $v$ that are polynomials of degree at most $k_v$ on each mesh element.

To complete the specification of the method, it only remains to set the discrete spaces. Let $P_k(D)$ denote the set of polynomials of degree at most $k$ on the domain $D$ (with the understanding that the set is empty when $k < 0$). Let $P_k(\Omega_h) = \{v : v|_K \in P_k(K) \text{ for all } K \in \Omega_h\}$ and let $P_k(\partial\Omega_h)$ denote the set of functions $v$ on $\partial\Omega_h$ having the property $v|_E \in P_k(E)$ for all edges of $\partial K$ and for all

$K \in \Omega_h$. In [19], the finite dimensional spaces are set as follows: $k \geqslant 0$,

Case 1

$$\begin{aligned} X_{h,0} &= P_k(\Omega_h) \cap X_0 && \Rightarrow k_u = k \\ \hat{X}_h &= P_{k-1}(\partial\Omega_h) \cap \hat{X} && \Rightarrow k_q = k-1 \\ Y^r &= P_{k+1}(\Omega_h) && \Rightarrow k_v = k+1 \end{aligned}$$

**Remark 2.13.** *We call this as case 1, as we are going to take into consideration more cases and analyze them focusing on the covergence rates of the DPG method for each case.*

The discrete solution in each of these cases is denoted by $(u_h, \hat{q}_{n,h}) \in X_h$. Assumption 2.5 was verified in [19] using the above-mentioned finite dimensional spaces.This then led to [19, Theorem 4.1], which states that

$$\|u-u_h\|_{H^1(\Omega)}+\|\hat{q}_n-\hat{q}_{n,h}\|_{H^{-1/2}(\partial\Omega_h)} \leqslant C \inf_{(w_h,\hat{r}_{n,h})\in X_h} \big(\|u-w_h\|_{H^1(\Omega)}+\|\hat{q}_n-\hat{r}_{n,h}\|_{H^{-1/2}(\partial\Omega_h)}\big).$$

Here and henceforth, $C$ denotes a generic constant independent of the size of the triangles in $\Omega_h$ (but dependent on mesh shape regularity), whose value at different occurrences may vary. As explained in previous papers (see e.g., [18]), applications of the Bramble-Hilbert Lemma in the Lagrange and Raviart-Thomas spaces show

that

$$\inf_{w_h \in P_l(\Omega_h) \cap X_0} \|u - w_h\|_{H^1(\Omega)} \leqslant C h^l |u|_{H^{l+1}(\Omega)}, \qquad \forall l \geqslant 0, \tag{2.19a}$$

$$\inf_{\hat{r}_{n,h} \in P_{m-1}(\partial\Omega_h) \cap \hat{X}} \|\hat{q}_n - \hat{r}_{n,h}\|_{H^{-1/2}(\partial\Omega_h)} \leqslant C h^m \left( |u|_{H^{m+1}(\Omega)} + |f|_{H^m(\Omega)} \right), \quad \forall m \geqslant 1. \tag{2.19b}$$

Therefore,

$$\|u - u_h\|_{H^1(\Omega)} + \|\hat{q}_n - \hat{q}_{n,h}\|_{H^{-1/2}(\partial\Omega_h)} \leqslant C h^k \left( |u|_{H^{k+1}(\Omega)} + |f|_{H^k(\Omega)} \right). \tag{2.20}$$

So the convergence rate of $\|u - u_h\|_{H^1(\Omega)}$ is $O(h^k)$.

## 2.3 Numerical Results

In this section, we report results from a numerical experiment. The presented DPG method for the Laplace equation was used to solve the Dirichlet problem with $\Omega$ set to the unit square. The function $f$ was chosen so that the exact solution is

$$u = sin(\pi x) sin(\pi y) \tag{2.21}$$

We construct an $n \times n$ uniform mesh by dividing $\Omega$ into $n^2$ congruent squares and further subdividing each square into two triangles by connecting the diagonal of positive slope. Its mesh size is $h = \sqrt{2}/n$. The method is applied on a sequence of such meshes with geometrically increasing $n$. The implementation of the method

is done using FEniCS.

The convergence rates for case 1 in $H^1-$norm and in $L^2-$norm are shown in the figures below. The slopes report the rate of convergence in $L^2(\Omega)$, approximately calculated using two successive error values by $\log_2(\|u - u_h\|_{L^2(\Omega)}/\|u - u_{h/2}\|_{L^2(\Omega)})$. The $H^1(\Omega)$-convergence rate is computed similarly. We observe from the figures that the $L^2(\Omega)$-rate is one order higher than the $H^1(\Omega)$-rate, as expected from Theorem 5.5.

Figure 2.1: Case 1 $H^1-$Error

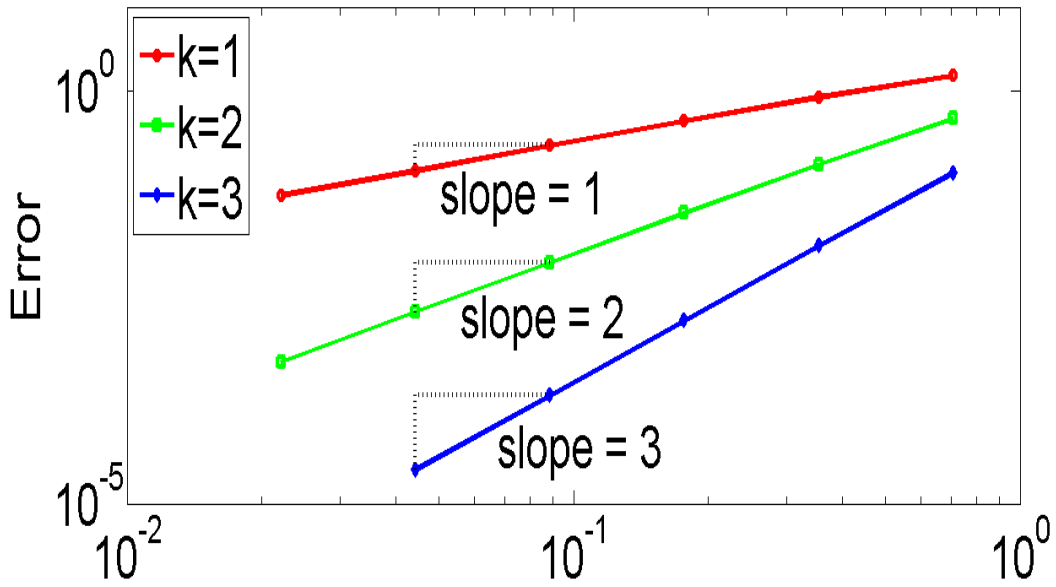
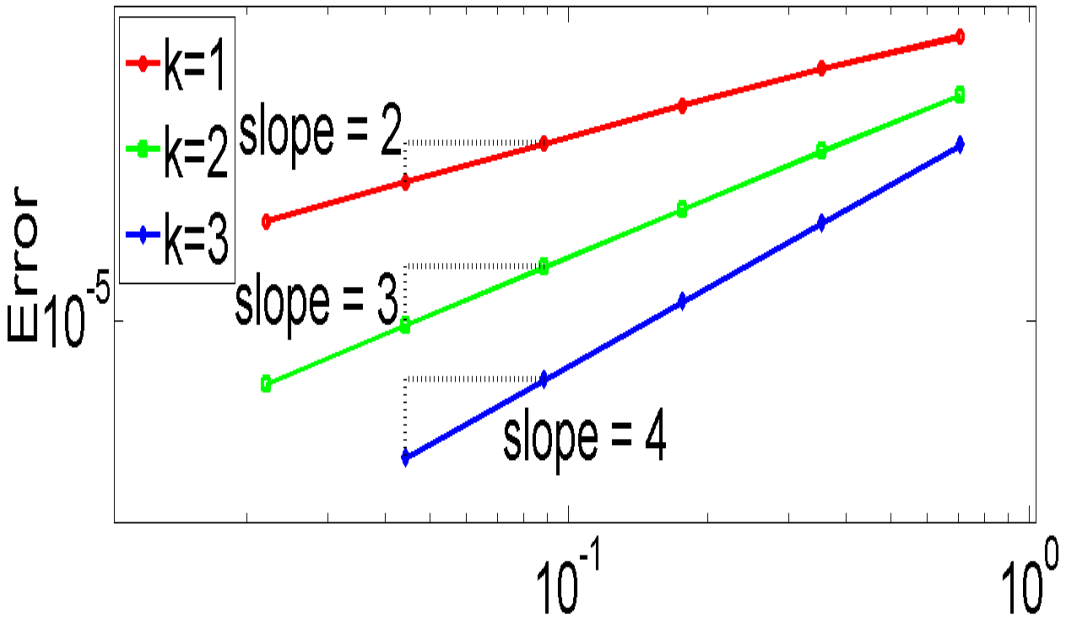


Figure 2.2: Case 1 $L^2-$Error



## 2.4 DPG Convergence Rates for Singularities

In this section, we will present an example where convergence rates do not improve even with increasing the polynomial degrees. We encounter this situation with singularities as in the following example.

**Example 2.14.**

$$-\Delta u = f \qquad \text{on } \Omega = [0,1]^2, \tag{2.22a}$$

$$u = 0 \qquad \text{on } \partial\Omega. \tag{2.22b}$$

Where The function f was chosen so that the exact solution is

$$u = x(1-x)y(1-y)\log(x^2+y^2) \tag{2.23}$$

We notice that there is a singularity at the origin.

We construct an $n \times n$ uniform mesh by dividing $\Omega$ into $n^2$ congruent squares and further subdividing each square into two triangles by connecting the diagonal of positive slope. Its mesh size is $h = \sqrt{2}/n$. The method is applied on a sequence of such meshes with geometrically increasing $n$. The implementation of the method is done using NGSolve. In table 2.4, the last column reports the rate of convergence in $L^2(\Omega)$, approximately calculated using two successive rows by $\log_2(\|u-u_h\|_{L^2(\Omega)}/\|u-u_{h/2}\|_{L^2(\Omega)})$. The $H^1(\Omega)$-convergence rate is computed similarly. We observe from the table that the $L^2(\Omega)$-rate is one order higher than the $H^1(\Omega)$-rate, as expected from Theorem 5.5. In addition, the convergence rate does not improve when $k > 3$ due to the regularity limit, since we approximate the

singular solution 2.23 which is in $H^{s+1}(\Omega)$ for $s < 3$, this observation in accordance with 2.19. We consider here case 1 from section 2.2.1 and the numerical results are shown in table 2.4.

Figure 2.3: Singularity (k=1, n=16)

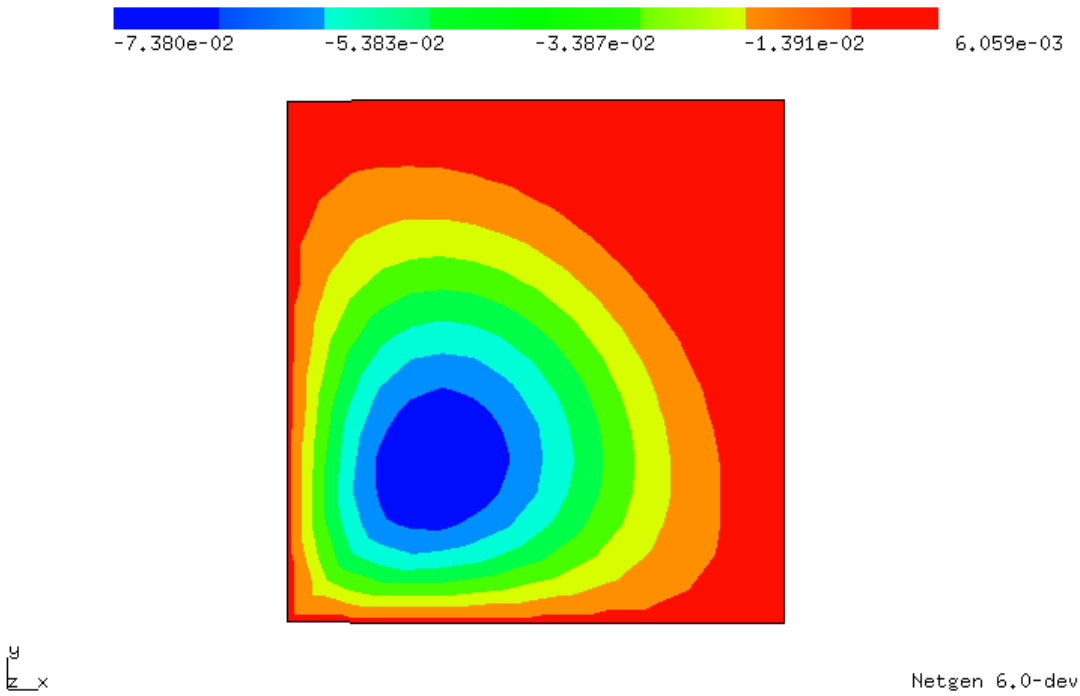



Table 2.1: Case 1: $(k_u, k_q, k_v) = (k, k-1, k+1)$

| $n$ | $\|u-u_h\|_{H^1(\Omega)}$ | rate | $\|u-u_h\|_{L^2(\Omega)}$ | rate |
|---|---|---|---|---|
| | | $k=1$ | | |
| 2 | 1.24E-01 | 1.02 | 1.67E-02 | 1.79 |
| 4 | 6.15E-02 | 0.85 | 4.80E-03 | 1.67 |
| 8 | 3.41E-02 | 0.98 | 1.51E-03 | 1.94 |
| 16 | 1.73E-02 | 1.00 | 3.93E-04 | 1.98 |
| 32 | 8.68E-03 | 1.00 | 9.98E-05 | 2.00 |
| 64 | 4.34E-03 | | 2.50E-05 | |
| | | $k=2$ | | |
| 2 | 7.19E-02 | 1.85 | 1.53E-03 | 2.57 |
| 4 | 1.99E-02 | 1.91 | 2.57E-04 | 2.97 |
| 8 | 5.30E-03 | 1.93 | 3.28E-05 | 3.00 |
| 16 | 1.39E-03 | 1.96 | 4.10E-06 | 3.02 |
| 32 | 3.57E-04 | 1.98 | 5.06E-07 | 3.02 |
| 64 | 9.05E-05 | | 6.24E-08 | |

Figure 2.4: Singularity (k=10, n=2)

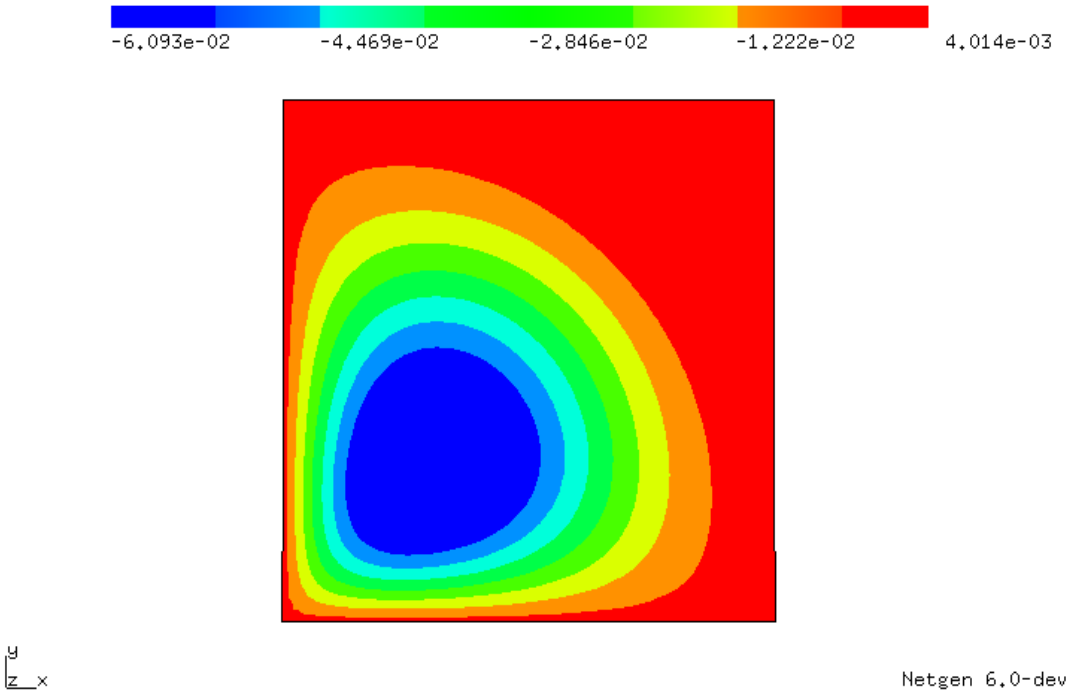



Table 2.2: Case 1: $(k_u, k_q, k_v) = (k, k-1, k+1)$

| $n$ | $\|u-u_h\|_{H^1(\Omega)}$ | rate | $\|u-u_h\|_{L^2(\Omega)}$ | rate |
|---|---|---|---|---|
| | | $k=3$ | | |
| 2 | 1.19E-02 | 2.25 | 7.18E-04 | 3.49 |
| 4 | 2.51E-03 | 2.61 | 6.38E-05 | 3.59 |
| 8 | 4.13E-04 | 2.70 | 5.29E-06 | 3.66 |
| 16 | 6.35E-05 | 2.74 | 4.18E-07 | 3.72 |
| 32 | 9.48E-06 | 2.78 | 3.17E-08 | 3.76 |
| 64 | 1.38E-06 | | 2.34E-09 | |
| | | $k=4$ | | |
| 2 | 5.20E-03 | 2.65 | 2.40E-05 | 3.75 |
| 4 | 8.26E-04 | 2.88 | 1.78E-06 | 3.88 |
| 8 | 1.12E-04 | 2.93 | 1.21E-07 | 3.90 |
| 16 | 1.48E-05 | 2.95 | 8.09E-09 | 3.93 |
| 32 | 1.90E-06 | 2.97 | 5.30E-10 | 3.96 |
| 64 | 2.43E-07 | | 3.42E-11 | |

Table 2.3: Case 1: $(k_u, k_q, k_v) = (k, k-1, k+1)$

| $n$ | $\|u-u_h\|_{H^1(\Omega)}$ | rate | $\|u-u_h\|_{L^2(\Omega)}$ | rate |
|---|---|---|---|---|
| $k=5$ | | | | |
| 2 | 1.61E-03 | 2.71 | 3.95E-05 | 4.07 |
| 4 | 2.46E-04 | 2.93 | 2.36E-06 | 3.97 |
| 8 | 3.23E-05 | 2.96 | 1.50E-07 | 3.98 |
| 16 | 4.14E-06 | 2.98 | 9.55E-09 | 3.98 |
| 32 | 5.25E-07 | 2.99 | 6.03E-10 | 3.99 |
| 64 | 6.61E-08 | | 3.80E-11 | |
| $k=6$ | | | | |
| 2 | 6.66E-04 | 3.03 | 3.42E-06 | 3.87 |
| 4 | 8.13E-05 | 3.02 | 2.33E-07 | 3.99 |
| 8 | 1.00E-05 | 3.01 | 1.47E-08 | 3.99 |
| 16 | 1.25E-06 | 3.00 | 9.23E-10 | 4.00 |
| 32 | 1.56E-07 | 3.00 | 5.79E-11 | 4.00 |
| 64 | 1.95E-08 | | 3.62E-12 | |

Table 2.4: Case 1: $(k_u, k_q, k_v) = (k, k-1, k+1)$

| $n$ | $\|u-u_h\|_{H^1(\Omega)}$ | rate | $\|u-u_h\|_{L^2(\Omega)}$ | rate |
|---|---|---|---|---|
| $k=10$ | | | | |
| 2 | 8.30E-05 | 2.96 | 2.99E-07 | 4.03 |
| 4 | 1.07E-05 | 2.98 | 1.83E-08 | 4.01 |
| 8 | 1.35E-06 | 2.99 | 1.13E-09 | 4.01 |
| 16 | 1.70E-07 | 2.99 | 7.04E-11 | 4.00 |
| 32 | 2.13E-08 | | 4.39E-12 | |

# Chapter 3

# Reduced Degree DPG Methods Based on Parity

## 3.1 Introduction

The purpose of this chapter is to provide a theoretical explanation for some numerically observed convergence rates of the discontinuous Petrov-Galerkin (DPG) method (case 2). While some aspects of the theory that follows are general, we will use the Laplace equation throughout as the example to illustrate the main points.

Considering the notation used in section 2.2.1, we want to reduce the polynomial degree of the test space, in order to get a cheaper method with the same convergence rates. It is the interplay between the convergence rates and the degrees $k_u, k_q, k_v$ that we intend to study.

We will study here the following case: $k \geqslant 1$, ($k$ is odd)

Case 2

$$
\begin{aligned}
X_{h,0} &= P_{k-1}(\Omega_h) \cap X_0 && \Rightarrow k_u = k-1 \\
\hat{X}_h &= P_{k-1}(\partial\Omega_h) \cap \hat{X} && \Rightarrow k_q = k-1 \\
Y^r &= P_k(\Omega_h) && \Rightarrow k_v = k
\end{aligned}
$$

Case 1 in section 2.2.1 is the standard DPG setting for which error estimates in the energy norm are proven in [19] . Case 2 is motivated by a desire to reduce the test space degree.

Our numerical experience with a few examples with smooth solutions is summarized in Table (3.1) . We observed that Case 2 is not always stable: It yielded singular stiffness matrices for some even $k$. However, when $k$ is odd, it converged, albeit at one order less than the standard DPG case displayed in the first row. In

addition, we observed that the convergence rate in $L^2(\Omega)$, in both cases, is one order higher than in $H^1(\Omega)$.

We explain the higher convergence rate in $L^2(\Omega)$ by developing a duality argument for DPG methods (analyzed in chapter 5). The duality theory is general and can be applied beyond the Laplace example as in section 5.4 where it has been applied for the Helmholtz equation. In this chapter, We give a complete theoretical explanation for the even-odd behavior, including negative results by counterexamples for even $k$, and a proof of a positive result for odd $k$.

## 3.2 Explaining the even-odd separation

We must first check if the DPG system is solvable for case 2. For this, Theorem 2.11 is useful. Clearly, Assumption 2.10 holds – in fact, it holds for all cases: items (a) and (b) are obvious, while (c) follows by the Poincaré inequality. Hence, applying Theorem 2.11, we conclude that the DPG method in Case 2 is uniquely solvable if and only if $\hat{B}_h$ is injective.

**Example 3.1.** *We begin with a negative result showing that $\hat{B}_h$ is not injective when $k = 2$. On a mesh consisting of a single element in the xy-plane, namely the unit triangle with vertices $a_0 = (0,0), a_1 = (1,0)$ and $a_2 = (0,1)$, we choose a basis for $\hat{X}_h$: Letting $e_i$ denote the edge opposite to $a_i$ and $1_{e_i}$ denote the indicator*

Table 3.1: Summary of numerically observed convergence rates

| | $h$-convergence rates of $u_h$ | |
|---|---|---|
| | in $H^1(\Omega)$ | in $L^2(\Omega)$ |
| Case 1 | $k$ | $k+1$ |
| Case 2 ($k$ odd) | $k-1$ | $k$ |

*function of $e_i$, the basis is $(1_{e_2}, x|_{e_2}, 1_{e_1}, y|_{e_1}, 1_{e_0}/\sqrt{2}, x|_{e_0}/\sqrt{2})$. For the test space $Y^r$, we choose the polynomial basis ($1$, $x$, $y$, $x^2$, $xy$, $y^2$). The stiffness matrix of the operator $\hat{B}_h$ with respect to these bases is*

$$\begin{pmatrix} 1 & 1/2 & 1 & 1/2 & 1 & 1/2 \\ 1/2 & 1/3 & 0 & 0 & 1/2 & 1/3 \\ 0 & 0 & 1/2 & 1/3 & 1/2 & 1/6 \\ 1/3 & 1/4 & 0 & 0 & 1/3 & 1/4 \\ 0 & 0 & 0 & 0 & 1/6 & 1/12 \\ 0 & 0 & 1/3 & 1/4 & 1/3 & 1/12 \end{pmatrix},$$

*whose determinant is zero. Hence the DPG method is not uniquely solvable in this example.*

**Example 3.2.** *We will present another negative result showing that $\hat{B}_h$ is not injective when $k = 4$. On a mesh consisting of the unit triangle only, we choose a basis for $\hat{X}_h$ (using the notation in example 3.1): the basis is*

$$(1_{e_2}, x|_{e_2}, x^2|_{e_2}, x^3|_{e_2}, 1_{e_1}, y|_{e_1}, y^2|_{e_1}, y^3|_{e_1}, 1_{e_0}/\sqrt{2}, x|_{e_0}/\sqrt{2}, x^2|_{e_0}/\sqrt{2}, x^3|_{e_0}/\sqrt{2}).$$

*For the trial space $Y^r$, we choose the polynomial basis*

$$(1\,, x\,, y\,, yx\,, x^2\,, y^2\,, x^2y\,, y^2x\,, y^2x^2\,, x^3\,, y^3\,, x^3y\,, y^3x\,, x^4\,, y^4)$$

*Let $B$ be the stiffness matrix of the operator $\hat{B}_h$ with respect to these bases, we find that*

$$det(B^T * B) = 0.$$

*Hence the DPG method is not uniquely solvable in this example.*

We now show that for odd $k$, the situation is better.

**Lemma 3.3.** *Let $K$ be a triangle and $k \geqslant 1$ be an odd integer. Any $w$ in $P_k(K)$ satisfying*

$$\int_E w\, q\, ds = 0 \qquad \forall\, q \in P_{k-1}(E),\ \forall \text{ edges } E \subset \partial K, \tag{3.1a}$$

$$\int_K w\, r\, dx = 0 \qquad \forall\, r \in P_{k-3}(K),\ \text{if } k \geqslant 3, \tag{3.1b}$$

*must vanish on $K$.*

*Proof.* Equation (3.1a) implies that $w|_E$ must be a scaled Legendre polynomial of degree exactly $k$ on $E$. Since $k$ is odd, this implies that the values of $w$ at the endpoints of each edge must have opposite signs. This is impossible unless $w$ vanishes on $\partial K$. But if $w|_{\partial K} = 0$, then $w \equiv 0$ if $k = 1$. If $k \geqslant 3$, then $w = \lambda_1\lambda_2\lambda_3 s_{k-3}$, for some $s_{k-3} \in P_{k-3}(K)$ where $\lambda_i$ is the $i$th barycentric coordinate. Then (3.1b) implies $w \equiv 0$ on $K$. □

**Theorem 3.4.** *In Case 2, for odd $k \geqslant 3$, these statements hold:*

*i) The DPG method is uniquely solvable.*

*ii) The solution $(u_h, \hat{q}_{n,h})$ of the DPG method satisfies*

$$\|u-u_h\|_{H^1(\Omega)}+\|\hat{q}_n-\hat{q}_{n,h}\|_{H^{-1/2}(\partial\Omega_h)} \leqslant Ch^{k-1}\left(|u|_{H^k(\Omega)}+|f|_{H^{k-1}(\Omega)}\right). \quad (3.2)$$

*iii) If $\Omega$ is convex, then*

$$\|u-u_h\|_{L^2(\Omega)} \leqslant Ch^{k}\left(|u|_{H^k(\Omega)}+|f|_{H^{k-1}(\Omega)}\right). \quad (3.3)$$

*Proof.* By Theorem 2.6, if we verify Assumption 2.5, then the DPG method is uniquely solvable.

To do so, we first claim that there exists a $C_{\Pi} > 0$ and a unique $\Pi v \in P_k(K)$ for any $v \in H^1(K)$, such that

$$\int_E (v-\Pi v)q\,ds = 0 \qquad \forall\, q \in P_{k-1}(E),\ \forall \text{ edges } E \subset \partial K, \quad (3.4a)$$

$$\int_K (v-\Pi v)r\,dx = 0 \qquad \forall\, r \in P_{k-3}(K) \quad (3.4b)$$

$$\|\Pi v\|_{H^1(K)} \leqslant C_{\Pi}\|v\|_{H^1(K)} \qquad \forall\, v \in H^1(K). \quad (3.4c)$$

It is easy to see that (3.4a)–(3.4b) forms a square system for $\Pi$, so existence of $\Pi v$ follows from uniqueness. But uniqueness is already proved by Lemma 3.3. The estimate (3.4c) will be proved by the following simple scaling argument:

Define the bilinear form $a(.,.) : Y^r \times \bar{X}_h \to \mathbb{R}$, where $\bar{X}_h = P_{k-1}(E) \times P_{k-3}(K)$.

The equations (3.4a) and (3.4b) can be written in the form,

$$a(\Pi v,(q,r)) = F(q,r) \quad \forall (q,r) \in P_{k-1}(E) \times P_{k-3}(K) \tag{3.5}$$

where

$$a(\Pi v,(q,r)) = \int_E \Pi v q \, ds + \int_K \Pi v r \, dx$$

and

$$F(q,r) = \int_E v q \, ds + \int_K v r \, dx$$

Suppose that $\{t_i\}_{i=1}^N = \{(q_i,r_i)\}_{i=1}^N$ is a basis of the space $P_{k-1}(E) \times P_{k-3}(K)$, and $\{e_j\}_{j=1}^M$ is a basis of $P_k(K)$. So $\Pi v = \sum_{j=1}^M p_j e_j$, where $p_j$'s are constant, which implies,

$$\sum_j p_j a(e_j,t_i) = F(t_i) \quad \forall i = 1,2,...,N \tag{3.6}$$

which can be written as a system of equations

$$AP = b$$

where $A_{ij} = a(e_j,t_i)$ which is a square matrix, $P = [p_1,p_2,...,p_M]^T$ and $b = [F(t_1),F(t_2),...,F(t_N)]^T$. which implies,

$$P = A^{-1}b$$

so

$$\|P\|_\infty \leqslant \left\|A^{-1}\right\|_\infty \|b\|_\infty$$

On the reference triangle $\hat{K}$ (with vertices $(0,0)$,$(1,0)$, and $(0,1)$), we get

$$\|\Pi v\|_{H^1(\hat{K})} = \left\|\sum_j p_j e_j\right\|_{H^1(\hat{K})} \leqslant \max\{p_1, p_2, ..., p_M\} \left\|\sum_j e_j\right\|_{H^1(\hat{K})} \leqslant C \|P\|_\infty$$

Where $C$ is a constant depends on $e_j$'s.

$$\leqslant C \left\|A^{-1}\right\|_\infty \|b\|_\infty \leqslant C \max_{r_i,q_i} |F(ti)| = C \max_{r_i,q_i} (\int_E v r_i + \int_{\hat{K}} v q_i)$$

where $C = C(e_j) \|A^{-1}\|_\infty$. The constant C might be different from step to step,

$$\leqslant C(\max_E \|v\|_{L^2(E)} + \|v\|_{L^2(\hat{K})}) \leqslant C \|v\|_{H^1(\hat{K})}$$

The last inequality was obtained by the trace theorem. So we have shown the following on the reference element,

$$\left\|\widehat{\Pi} v\right\|^2_{L^2(\hat{K})} + \left|\widehat{\Pi} v\right|^2_{H^1(\hat{K})} \leqslant C(\|\hat{v}\|^2_{L^2(\hat{K})} + |\hat{v}|^2_{H^1(\hat{K})}) \tag{3.7}$$

Let $G \in C^1(\Omega)$ such that $K = G(\hat{K})$ and is given by $G(\hat{x}) = h_K \hat{x} + \overline{b}$, where $h_K$ is the diameter of $K$ and $\overline{b} \in \mathbb{R}^2$. We suppose that $DG(\hat{x})$, the Jacobian matrix is invertible for any $\hat{x}$ and that $G$ is globally invertible on $K$. We then

have $DG^{-1}(x) = (DG(\hat{x}))^{-1}$. If $\hat{v}(\hat{x})$ is a function on $\hat{K}$, we define $v(x)$ on $K$ by $v = \hat{v} \circ G^{-1}$. By the scaling equation (in two dimensions,equation 2.37 of [7]), $|\hat{v}|_{m,\hat{K}} = h^{m-1} |v|_{m,K}$, where $|.|_{m,K}$ is the seminorm of $H^m(K)$, we get

$$h^{-2} \|\Pi v\|^2_{L^2(K)} + |\Pi v|^2_{H^1(K)} \leqslant C(h^{-2} \|v\|^2_{L^2(K)} + |v|^2_{H^1(K)})$$

Applying the last inequality to $v - \bar{v}$ (where $\bar{v}$ is average value of v over $K$) to get,

$$h^{-2} \|\Pi(v - \bar{v})\|^2_{L^2(K)} + |\Pi(v - \bar{v})|^2_{H^1(K)} \leqslant C(h^{-2} \|v - \bar{v}\|^2_{L^2(K)} + |v - \bar{v}|^2_{H^1(K)})$$

By Friedrichs' inequality we get (Note that $\Pi(v - \bar{v}) = \Pi v - \bar{v}$),

$$h^{-2} \|\Pi v - \bar{v}\|^2_{L^2(K)} + |\Pi v|^2_{H^1(K)} \leqslant C(|v|^2_{H^1(K)} + |v|^2_{H^1(K)}) \leqslant C |v|^2_{H^1(K)}$$

It implies the following two inequalities,

$$\|\Pi v - \bar{v}\|_{L^2(K)} \leqslant Ch |v|_{H^1(K)}$$

$$|\Pi v|_{H^1(K)} \leqslant C |v|_{H^1(K)} \tag{3.8}$$

Now,

$$\|\Pi v\|_{L^2(K)} \leqslant \|\bar{v}\|_{L^2(K)} + \|\Pi v - \bar{v}\|_{L^2(K)} \leqslant \|v\|_{L^2(K)} + Ch\,|v|_{H^1(K)} \leqslant (1+Ch)\,\|v\|_{H^1(K)} \tag{3.9}$$

Adding the inequalities (3.8) and (3.9) to get,

$$\|\Pi v\|_{H^1(K)} \leqslant C_\Pi\,\|v\|_{H^1(K)}$$

where $C_\Pi = \max\{1 + Ch, C\}$.

The energy error estimate (3.2) now follows from Theorem 2.6 and (2.19). The $L^2$ error estimate (3.3) follows from Theorem 5.2: The required verification of Assumption 5.1 proceeds as in the proof of Theorem 5.5 – the only difference is in the degrees of approximation spaces in the first two infimums in (5.23), a difference that is inconsequential for the rest of the arguments. □

Theorem 3.4 explains all entries in the second row of table (3.1) . The convergence rate in (3.2) is suboptimal and limited by the low degree of $u_h$. This motivates the next case.

### 3.3 Numerical Results

Using the exact solution 2.21 and the same mesh as in section 2.3, The convergence rates for case 2 in $H^1-$norm and in $L^2-$norm are shown in the figures below.

Figure 3.1: Case 2 $H^1-$Error

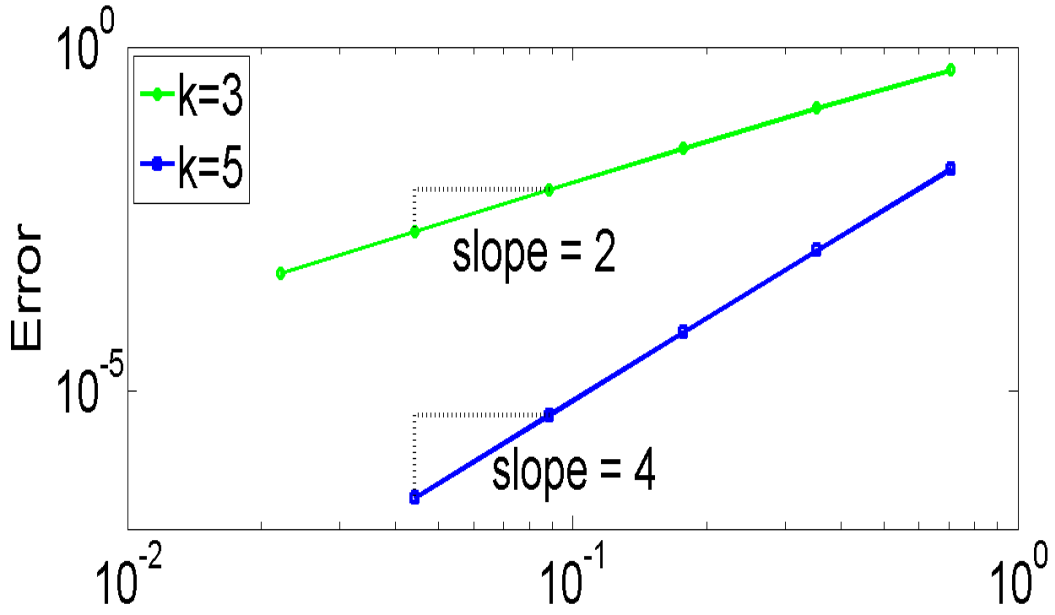
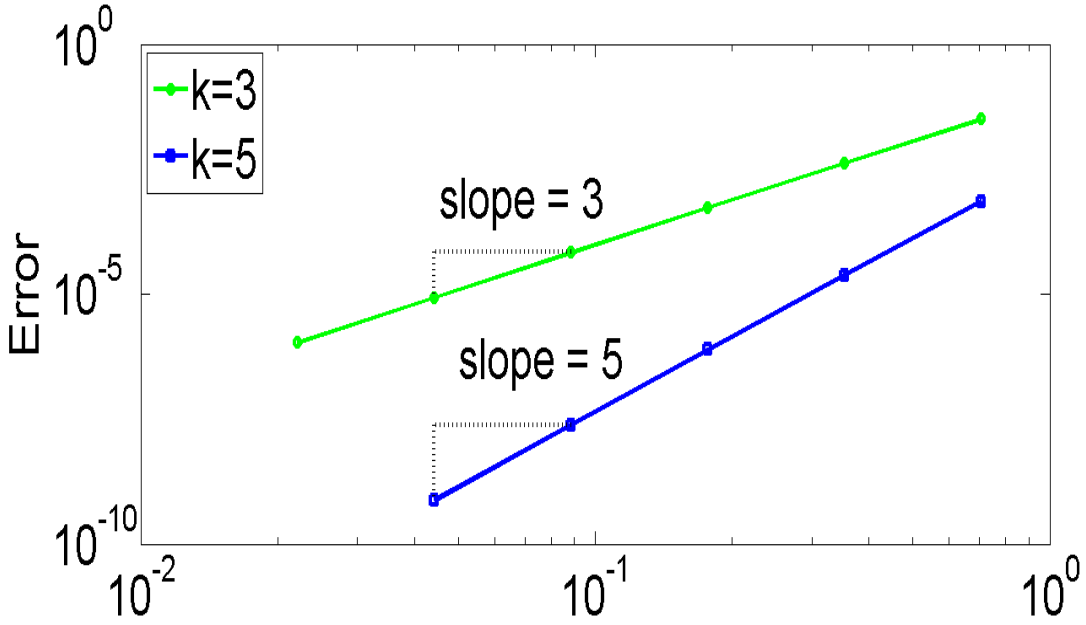


Figure 3.2: Case 2 $L^2-$Error



# Chapter 4

# Nonconforming Analysis

## 4.1 Introduction

In this chapter, we will use case 2 in chapter 3 to introduce a more interesting case, case 3. The only difference between Case 2 and Case 3 is that the degree of $u_h$ is increased by one. Keeping $k$ odd for Case 3, we find that the original DPG convergence rates can be recovered, in spite of using a smaller $k_v$. Similarly, we observed that the convergence rate in $L^2(\Omega)$, in all cases, is one order higher than in $H^1(\Omega)$.

The table below summarizes the three cases under study and table (4.1) provides the convergence rates in the $H^1$-norm and the $L^2$-norm for the three cases.

| | $k_u$ | $k_q$ | $k_v$ |
|---|---|---|---|
| Case 1: | $k$ | $k-1$ | $k+1$, |
| Case 2: | $k-1$ | $k-1$ | $k$, |
| Case 3: | $k$ | $k-1$ | $k$. |

## 4.2 Case 3: A nonconforming analysis

We analyze Case 3 using a technique of analysis different from the previous subsection, appealing to Theorem 2.9 and the second Strang lemma (see e.g. [11]) in the analyses of nonconforming methods.

Table 4.1: Summary of numerically observed convergence rates

| | $h$-convergence rates of $u_h$ | |
|---|---|---|
| | in $H^1(\Omega)$ | in $L^2(\Omega)$ |
| Case 1 | $k$ | $k+1$ |
| Case 2 ($k$ odd) | $k-1$ | $k$ |
| Case 3 ($k$ odd) | $k$ | $k+1$ |

**Theorem 4.1.** *In Case 3, for odd $k \geqslant 1$, these statements hold:*

*i)* *$\hat{B}_h$ is injective and the DPG method is uniquely solvable.*

*ii)* *The $u_h$-component of the solution satisfies*

$$\|u - u_h\|_{H^1(\Omega)} \leqslant Ch^k \left(|u|_{H^{k+1}(\Omega)} + |f|_{H^k(\Omega)}\right). \tag{4.1}$$

*iii)* *If $\Omega$ is convex, then*

$$\|u - u_h\|_{L^2(\Omega)} \leqslant Ch^{k+1} \left(|u|_{H^{k+1}(\Omega)} + |f|_{H^k(\Omega)}\right). \tag{4.2}$$

*Proof.* First, observe that if $k \geqslant 3$, then by the unisolvency of the DPG method in Case 2, namely Theorem 3.4 (i), its $B_h$ is injective, which implies by Theorem 2.11 that $\hat{B}_h$ of Case 2 is injective. But since the flux $(\hat{X}_h)$ and test spaces $(Y^r)$ of Case 3 are identical to that of Case 2, both cases have the same $\hat{B}_h$. Hence $\hat{B}_h$ of Case 3 is injective and consequently by Theorem 2.11, $B_h$ of Case 3 is injective. Thus we have proved the first statement of the theorem for $k \geqslant 3$. For $k = 1$, if $(\hat{B}_h \hat{r}_{n,h})(w) = \langle \hat{r}_{n,h}, w \rangle_{\partial\Omega_h} = 0$ for all $w \in Y^r$, then

$$\int_{\partial K} w\, \hat{r}_{n,h}\, ds = 0, \qquad \forall\, w \in P_k(K).$$

The matrix of this system (for $\hat{r}_{n,h}$) is the transpose of the matrix of (3.1) (for $w$),

which is invertible by Lemma 3.3. Hence $\hat{r}_{n,h} = 0$, i.e., $\hat{B}_h$ is injective when $k = 1$.

Next we prove (4.1). Recall that $Y_0^r$ is defined in (2.11) and $Y_{h,0}^r$ in (2.12b). By Theorem 2.9, $u_h \in X_{h,0}$ satisfies (2.12b), i.e.,

$$b_0(u_h, y) = (f, y)_\Omega, \qquad \forall y \in Y_{h,0}^r. \tag{4.3}$$

We proceed by viewing this as a nonconforming Petrov-Galerkin discretization of

$$b_0(u, y) = (f, y)_\Omega, \qquad \forall y \in H_0^1(\Omega)$$

and bounding the consistency error in an argument akin to the second Strang lemma. Let $C_p$ denote the constant, derived from Poincaré inequality, such that $\|w\|_{H^1(\Omega)} \leqslant C_p \|\operatorname{grad} w\|_{L^2(\Omega)}$ for all $w \in H_0^1(\Omega)$. Then, for any $w_h \in X_{h,0}$

$$\begin{aligned}
\|u_h - w_h\|_{H^1(\Omega)} &\leqslant C_p \sup_{z_h \in X_{h,0}} \frac{(\operatorname{grad}(u_h - w_h), \operatorname{grad} z_h)_\Omega}{\|\operatorname{grad} z_h\|_{L^2(\Omega)}} \leqslant C_p^2 \sup_{z_h \in X_{h,0}} \frac{b_0(u_h - w_h, z_h)}{\|z_h\|_{H^1(\Omega)}} \\
&\leqslant C_p^2 \sup_{y \in Y_0^r} \frac{b_0(u_h - w_h, y)}{\|y\|_Y} = C_p^2 \|T_0^r(u_h - w_h)\|_Y = C_p^2 \sup_{y \in Y_{h,0}^r} \frac{b_0(u_h - w_h, y)}{\|y\|_Y} \\
&= C_p^2 \sup_{y \in Y_{h,0}^r} \frac{b_0(u_h - u, y) + b_0(u - w_h, y)}{\|y\|_Y} \\
&= C_p^2 \sup_{y \in Y_{h,0}^r} \frac{(f, y)_\Omega - b_0(u, y) + b_0(u - w_h, y)}{\|y\|_Y},
\end{aligned} \tag{4.4}$$

where we have used (6.26). Since $b((u, \hat{q}_n), y) = (f, y)_\Omega$ for all $y \in Y$, the term representing the consistency error in (6.27) can be written as $(f, y)_\Omega - b_0(u, y) =$

$\hat{b}(\hat{q}_n, y)$. By the definition of $Y_0^r$ (see (2.11)), we also have $\hat{b}(\hat{q}_n, y) = \hat{b}(\hat{q}_n - \hat{r}_{n,h}, y)$ for any $\hat{r}_{n,h} \in \hat{X}_h$ and $y \in Y_0^r$. Therefore,

$$\|u_h - w_h\|_{H^1(\Omega)} \leqslant C_p^2 \sup_{y \in Y_{h,0}^r} \frac{b((u - w_h, \hat{q}_n - \hat{r}_{n,h}), y)}{\|y\|_Y} \leqslant C_p^2 C_2 \left( \|\hat{q}_n - \hat{r}_{n,h}\|_{\hat{X}} + \|u - w_h\|_{H^1(\Omega)} \right).$$

Since $\hat{r}_{n,h}$ and $\hat{q}_n$ are element-by-element traces of an $r_h$ in $R_{k-1}$ and $q = \operatorname{grad} u$, respectively,

$$\|\hat{r}_{n,h} - \hat{q}_n\|_{\hat{X}} \leqslant \|r_h - \operatorname{grad} u\|_{H(\operatorname{div},\Omega)},$$

so

$$\|u_h - w_h\|_{H^1(\Omega)} \leqslant C \left( \inf_{r_h \in R_{k-1}} \|r_h - \operatorname{grad} u\|_{H(\operatorname{div},\Omega)} + \|u - w_h\|_{H^1(\Omega)} \right).$$

Finally, by the triangle inequality,

$$\begin{aligned} \|u - u_h\|_{H^1(\Omega)} &\leqslant \|u - w_h\|_{H^1(\Omega)} + \|u_h - w_h\|_{H^1(\Omega)} \\ &\leqslant C \left( \|u - w_h\|_{H^1(\Omega)} + h^k (|u|_{H^{k+1}(\Omega)} + |f|_{H^k(\Omega)}) \right) \end{aligned}$$

for any $w_h \in X_{h,0}$. Choosing $w_h$ to be an appropriate interpolant, the proof of (4.1) is finished.

The final estimate (4.2) is proved by verifying Assumption 5.1 (along the lines of the proof of theorem 5.5) and applying theorem 5.2. □

The final row of table 4.1 is now completely explained by theorem 4.1.

## 4.3 Numerical Analysis

Using the exact solution 2.21 and the same mesh in section 2.3, The convergence rates for case 3 in $H^1-$norm and in $L^2-$norm are shown in the figures.

Figure 4.1: Case 3 $H^1-$Error

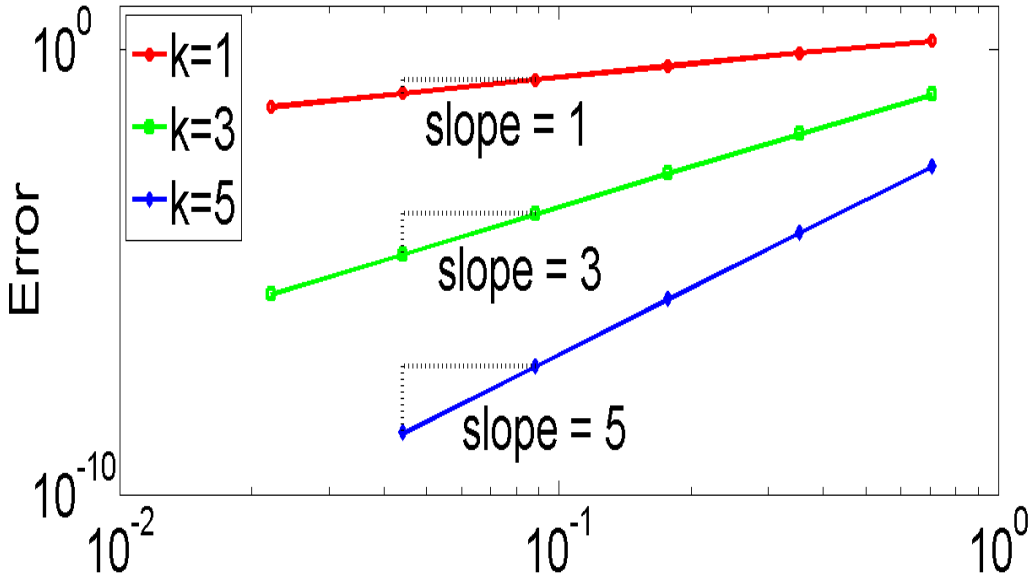
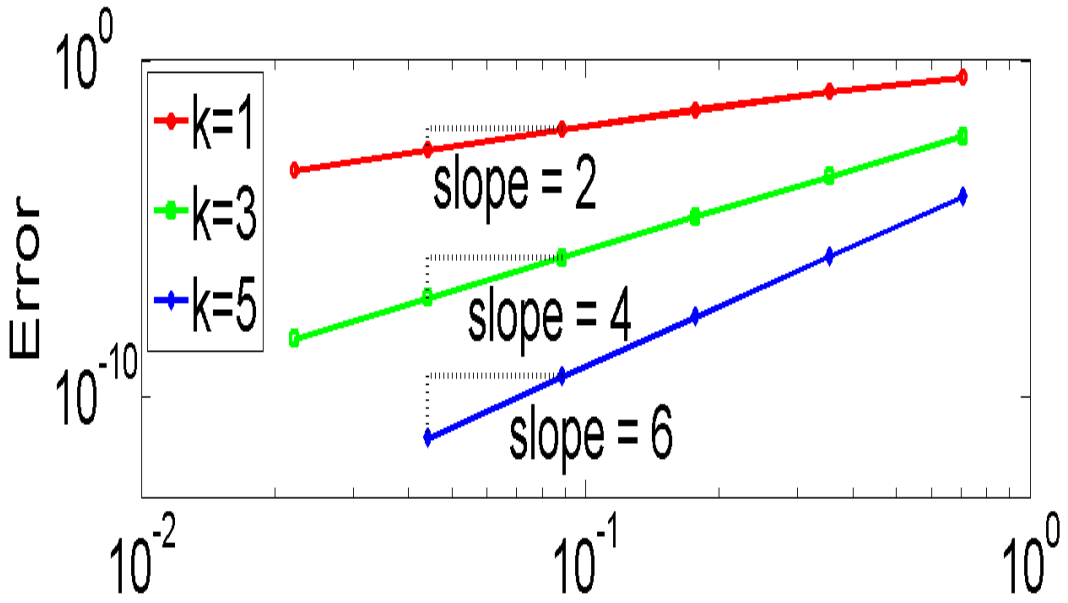


Figure 4.2: Case 3 $L^2-$Error



# Chapter 5

# Duality Argument

## 5.1 Introduction

The duality argument is a technique used by Aubin and Nitsche (see [2, 29, 30]) to derive a priori error estimates in the $L^2$-norm for Bubnov-Galerkin methods. The approach is also known as Aubin-Nitsche trick.

For the DPG methods, we can perform an analogue of the Aubin-Nitsche duality argument. First we need to write the DPG method as a Bubnov-Galerkin method, which is feasible using theorem 2.7.

In addition, there are many ingredients for the duality argument which need to be established for the DPG methods, such a regularity theorem, a Bramble-Hilbert lemma, and a Galerkin Orthogonality property.

## 5.2 General Settings

By virtue of Theorem 2.7, we may rewrite the DPG method (2.6) as follows: Find $x_{h,0} \in X_{0,h},\ \hat{x}_h \in \hat{X}_h$, and $\varepsilon^r \in Y^r$ solving

$$
\begin{aligned}
b_0(w, \varepsilon^r) &= 0 && \forall w \in X_{0,h}, && \text{(5.1a)}\\
\hat{b}(\hat{w}, \varepsilon^r) &= 0 && \forall \hat{w} \in \hat{X}_h, && \text{(5.1b)}\\
b_0(x_{h,0}, y) + \hat{b}(\hat{x}_h, y) + (\varepsilon^r, y)_Y &= \ell(y), && \forall y \in Y^r. && \text{(5.1c)}
\end{aligned}
$$

Defining

$$
a(z, \hat{z}, v | w, \hat{w}, y) = \overline{b_0(w, v)} + \overline{\hat{b}(\hat{w}, v)} + b_0(z, y) + \hat{b}(\hat{z}, y) + (v, y)_Y,
$$

the mixed system (5.1) can then be rewritten as

$$a(x_{h,0}, \hat{x}_h, \varepsilon^r | w, \hat{w}, y) = \ell(y), \qquad \forall w \in X_{0,h}, \hat{w} \in \hat{X}_h, y \in Y^r,$$

where the complex conjugate on the first two terms make the form $a$ sesquilinear. Now, observe that with $\varepsilon = 0$, the exact solution $(x_0, \hat{x}, \varepsilon) \in X_0 \times \hat{X} \times Y$ satisfies the same equation for all $w \in X_0, \hat{w} \in \hat{X}, y \in Y$. Hence, we have a 'Galerkin orthogonality' relation

$$a(x_0 - x_{h,0}, \hat{x} - \hat{x}_h, \varepsilon - \varepsilon^r | w, \hat{w}, y) = 0, \tag{5.2}$$

for all $w \in X_{0,h}, \hat{w} \in \hat{X}_h, y \in Y^r$. Note also that

$$\begin{aligned} |a(z, \hat{z}, v | w, \hat{w}, y)| &\leqslant C_2 \|(z, \hat{z})\|_X \|y\|_Y + C_2 \|(w, \hat{w})\|_X \|v\|_Y + \|v\|_Y \|y\|_Y \\ &\leqslant \left(C_2^2 \|(z, \hat{z})\|_X^2 + 2\|v\|_Y^2\right)^{1/2} \left(C_2^2 \|(w, \hat{w})\|_X^2 + 2\|y\|_Y^2\right)^{1/2} \\ &\leqslant \|a\| \, \|(z, \hat{z}, v)\|_{X_0 \times \hat{X} \times Y} \|(w, \hat{w}, y)\|_{X_0 \times \hat{X} \times Y} \end{aligned}$$

where $\|a\|$ is a constant not larger than $\max(C_2^2, 2)$. Under the following assumption, we can extend the Aubin-Nitsche technique [29] to DPG methods, as seen in the next theorem.

**Assumption 5.1.** *Suppose $L$ and $Z$ are Hilbert spaces such that the embeddings $Z \subseteq X_0 \times \hat{X} \times Y$ and $X_0 \subseteq L$ are continuous. Assume that there is a $C_3(h) > 0$*

*such that for any $g \in L$, there is a $U(g) \in Z$ satisfying*

$$a(w, \hat{w}, y | U(g)) = (w, g)_L \tag{5.3}$$

*for all $(w, \hat{w}, y) \in X_0 \times \hat{X} \times Y$ and*

$$\inf_{W \in X_{0,h} \times \hat{X}_h \times Y^r} \|U(g) - W\|_{X_0 \times \hat{X} \times Y} \leqslant C_3(h) \|g\|_L. \tag{5.4}$$

**Theorem 5.2.** *Suppose Assumption 5.1 holds. Then,*

$$\|x - x_{h,0}\|_L \leqslant C_3(h) \|a\| \|(x, \hat{x}, \varepsilon) - (x_{h,0}, \hat{x}_h, \varepsilon^r)\|_{X_0 \times \hat{X} \times Y}.$$

*Proof.* Setting $g = w = x - x_{h,0}$, $\hat{w} = \hat{x} - \hat{x}_h$, and $y = \varepsilon - \varepsilon^r$ in (5.3),

$$\begin{aligned}
\|x - x_{h,0}\|_L^2 &= a(x - x_{h,0}, \hat{x} - \hat{x}_h, \varepsilon - \varepsilon^r | U(x - x_{h,0})) \\
&= a(x - x_{h,0}, \hat{x} - \hat{x}_h, \varepsilon - \varepsilon^r | U(x - x_{h,0}) - W), \qquad \text{by (5.2)}, \\
&\leqslant \|a\| \|(x - x_{h,0}, \hat{x} - \hat{x}_h, \varepsilon - \varepsilon^r)\|_{X_0 \times \hat{X} \times Y} \|U(x - x_{h,0}) - W\|_{X_0 \times \hat{X} \times Y}
\end{aligned}$$

for any $W \in X_{0,h} \times \hat{X}_h \times Y^r$. Hence (5.4) completes the proof. □

**Remark 5.3.** *Let $A : X_0 \times \hat{X} \times Y \to (X_0 \times \hat{X} \times Y)^*$ be the operator generated by $a(\cdot, \cdot)$, i.e., $(A(z, \hat{z}, v))(w, \hat{w}, y) = a(z, \hat{z}, v | w, \hat{w}, y)$ for all $(z, \hat{z}, v), (w, \hat{w}, y) \in X_0 \times \hat{X} \times Y$. If Assumption 2.4 holds, then $A$ is a bijection. (This follows from*

*the Babuška-Brezzi theory [7], applied to the mixed system* (2.9)*: the "inf-sup condition" follows from* (2.7)*, and the "coercivity in the kernel condition" is trivial.) Hence, the dual operator of $A$ is also a bijection whereby we conclude that* (5.3) *has a unique solution $U(g)$.*

**Remark 5.4.** *All results of this section hold for spaces over the real field $\mathbb{R}$ – one only needs to replace $\mathbb{C}$ by $\mathbb{R}$, sesquilinear by bilinear, and conjugate-linear by linear to obtain the corresponding statements for real valued function spaces. The DPG method for the Helmholtz equation [24] provides an example where sesquilinear forms over $\mathbb{C}$ are used. For simplicity, in the remaining sections we will restrict ourselves to real-valued functions.*

### 5.3 Case 1: Application of the duality argument

**Theorem 5.5.** *Suppose $\Omega$ is convex. Then, for Case 1,*

$$\|u - u_h\|_{L^2(\Omega)} \leqslant Ch^{k+1} \left( |u|_{H^{k+1}(\Omega)} + |f|_{H^k(\Omega)} \right).$$

*Proof.* Set

$$Z_1 = H^2(\Omega) \cap X_0, \qquad L = L^2(\Omega),$$

$$Z_2 = H^2(\Omega) \cap Y, \qquad Z = Z_1 \times \hat{X} \times Z_2.$$

To verify Assumption 5.1, let $g \in L$. By Remark 5.3, there is a unique $U(g) \equiv (z, \hat{z}_n, d) \in X_0 \times \hat{X} \times Y$ solving (5.3). Writing out (5.3) in component form,

$$(d, y)_Y + (\nabla z, \nabla y)_{\Omega_h} - \langle \hat{z}_n, y \rangle_{\partial\Omega_h} = 0, \qquad \forall y \in Y, \tag{5.5a}$$

$$(\nabla d, \nabla w)_{\Omega_h} = (g, w)_{\Omega_h} \qquad \forall w \in X_0, \tag{5.5b}$$

$$\langle d, \hat{w}_n \rangle_{\partial\Omega_h} = 0 \qquad \forall \hat{w}_n \in \hat{X}. \tag{5.5c}$$

We need to understand the regularity of solutions of (5.5). Considering the $d$ component first, we claim that (5.5c) implies $d \in H_0^1(\Omega)$: Indeed the distributional gradient $\nabla d$ acting on a test function $\phi \in \mathcal{D}(\Omega)^2$ satisfies $(\nabla d)(\phi) = -(d, \operatorname{div} \phi)_{\Omega_h} = (\nabla d, \phi)_{\Omega_h} - \langle d, \phi \cdot n \rangle_{\partial\Omega_h}$ and the last term vanishes by (5.5c), so the distributional gradient is in $L^2(\Omega)^2$. It is also easy to see that the trace of $d$ vanishes on $\partial\Omega$. Then, (5.5b) implies that $-\Delta d = g$. Next, consider $z \in H_0^1(\Omega)$. Equation (5.5a) with $y \in H_0^1(\Omega)$ yields $(\nabla z, \nabla y)_{\Omega_h} = -(d, y)_{\Omega_h} - (\nabla d, \nabla y)_{\Omega_h} = -(d, y)_{\Omega_h} + (\Delta d, y)_{\Omega_h}$ which implies $\Delta z = d + g$. Finally, using the equations for $z$ and $d$ in (5.5a) and integrating by parts, we find $\langle \hat{z}_n, y \rangle_{\partial\Omega_h} = \langle n \cdot \nabla(d + z), y \rangle_{\partial\Omega_h}$.

Summarizing, the classical form of (5.5) is

$$-\Delta d = g, \qquad \text{on } \Omega, \tag{5.6a}$$

$$d = 0, \qquad \text{on } \partial\Omega, \tag{5.6b}$$

$$\Delta z = d + g, \qquad \text{on } \Omega, \tag{5.6c}$$

$$z = 0, \qquad \text{on } \partial\Omega, \tag{5.6d}$$

$$\hat{z}_n = n \cdot \nabla(d + z), \qquad \text{on } \partial K,\ \forall K \in \Omega_h. \tag{5.6e}$$

Thus, by full regularity of the Dirichlet problem on a convex domain [26], $d$ and $z$ are in $H^2(\Omega)$, and moreover,

$$\begin{aligned}
\|d\|_{Z_2} &\leqslant C\|g\|_L, \\
\|z\|_{Z_1} &\leqslant C\left(\|d\|_L + \|g\|_L\right) \leqslant C\|g\|_L, \\
\|\hat{z}\|_{\hat{X}} &\leqslant \|\nabla(d+z)\|_{H(\mathrm{div},\Omega)} \\
&= \|\nabla(d+z)\|_L + \|\Delta(d+z)\|_L \\
&= \|\nabla(d+z)\|_L + \|d\|_L \qquad \text{by (5.6)}, \\
&\leqslant C\|g\|_L.
\end{aligned}$$

Hence

$$\|(z, \hat{z}, d)\|_Z \leqslant C\|g\|_L. \tag{5.7}$$

To complete the verification of Assumption 5.1, we now only need to bound some approximation errors. By the Bramble-Hilbert lemma,

$$\begin{aligned}
&\inf_{W\in X_{0,h}\times\hat{X}_h\times Y^r}\|U(g)-W\|^2_{X_0\times\hat{X}\times Y}\\
&\quad=\inf_{w_h\in P_k(\Omega_h)\cap X_0}\|z-w_h\|^2_{H^1(\Omega)}+\inf_{v_h\in P_{k+1}(\partial\Omega_h)}\|d-v_h\|^2_{H^1(\Omega_h)}+\inf_{\hat{w}_h\in P_{k-1}(\partial\Omega_h)\cap\hat{X}}\|\hat{z}-\hat{w}_h\|^2_{\hat{X}} \qquad (5.8)\\
&\quad\leqslant Ch^2\left(|d|^2_{H^2(\Omega)}+|z|^2_{H^2(\Omega)}\right)+\inf_{r_h\in R_{k-1}}\|\nabla(d+z)-r_h\|^2_{H(\mathrm{div},\Omega)}
\end{aligned}$$

where $R_{k-1}$ is the Raviart-Thomas subspace [31] of $H(\mathrm{div},\Omega)$ consisting of all vector functions which when restricted to an element takes the form $xp_1+p_2$ for some $p_1\in P_{k-1}(K)$ and some $p_2\in P_{k-1}(K)^2$. Let $\Pi^h_{\mathrm{RT}}$ denote the Raviart-Thomas projection into $R_{k-1}$. By its well-known commutativity property with the $L^2$-projection $\Pi^h_{k-1}$ onto $P_{k-1}(\Omega_h)$, we have

$$\begin{aligned}
\inf_{r_h\in R_{k-1}}\|\nabla(d+z)-r_h\|_{H(\mathrm{div},\Omega)}&\leqslant\|(I-\Pi^h_{\mathrm{RT}})\nabla(d+z)\|_{H(\mathrm{div},\Omega)}\\
&\leqslant\|(I-\Pi^h_{\mathrm{RT}})\nabla(d+z)\|_L+\|(I-\Pi^h_{k-1})\Delta(d+z)\|_L\\
&\leqslant\|(I-\Pi^h_{\mathrm{RT}})\nabla(d+z)\|_L+\|(I-\Pi^h_{k-1})d\|_L,\qquad\text{by (5.6)},\\
&\leqslant Ch|d+z|_{H^2(\Omega)}+Ch|d|_{H^1(\Omega)},
\end{aligned}$$

where we used the Bramble-Hilbert lemma again in the final step. Hence using the

regularity estimate (5.7),

$$\inf_{W \in X_{0,h} \times \hat{X}_h \times Y^r} \|U(g) - W\|_{X_0 \times \hat{X} \times Y} \leqslant Ch \|g\|_L,$$

thus verifying Assumption 5.1. Now, applying Theorem 5.2,

$$\|u - u_h\|_{L^2(\Omega)} \leqslant Ch \left( \|u - u_h\|_{H^1(\Omega)} + \|\hat{q}_n - \hat{q}_{n,h}\|_{H^{-1/2}(\partial\Omega_h)} + \|\varepsilon - \varepsilon^r\|_{H^1(\Omega_h)} \right)$$

where $\varepsilon = 0$ and $\varepsilon^r$ is as in (2.8). This implies, by virtue of (2.10) in Remark 2.8,

$$\|u - u_h\|_{L^2(\Omega)} \leqslant Ch \left( \|u - u_h\|_{H^1(\Omega)} + \|\hat{q}_n - \hat{q}_{n,h}\|_{H^{-1/2}(\partial\Omega_h)} \right)$$

so the proof is finished using (2.20). □

## 5.4 Application to the Helmholtz Equation with Impedance Boundary Condition

We consider the following boundary value problem for the Helmholtz equation

$$-\Delta u - k^2 u = f_1 \qquad \text{on } \Omega, \tag{5.9a}$$

$$\nabla u \cdot n + iku = f_2 \qquad \text{on } \partial\Omega. \tag{5.9b}$$

where $f_1 \in L^2(\Omega)$ and $f_2 \in L^2(\partial\Omega)$ with positive wave number $k > 0$. The weak formulation problem of 5.9 is: Find $(u, \hat{q}_n) \in X = X_0 \times \hat{X}$ satisfying

$$(\nabla u, \nabla v)_{\Omega_h} + ik\langle u, v\rangle_{\partial\Omega} - k^2(u, v)_{\Omega_h} - \langle \hat{q}_n, v\rangle_{\partial\Omega_h \setminus \partial\Omega} = (f_1, v)_{\Omega_h} + \langle f_2, v\rangle_{\partial\Omega}, \quad \forall v \in Y \tag{5.10}$$

Set,

$$\begin{aligned} &X = H^1(\Omega), && \hat{X} = H_0^{-1/2}(\partial\Omega_h), && Y = H^1(\Omega_h), \\ &Z_1 = H^2(\Omega) \cap Y, && Z_2 = H^2(\Omega) \cap X, && L = L^2(\Omega), \\ &Z = Z_1 \times Z_2 \times \hat{X}. \end{aligned}$$

To verify assumption 5.1, let $g \in L$. Write $U(g)$ in component form as $(d, u, \hat{q}_n)$. Then 5.3 becomes

$$a(v, w, \hat{w} | d, u, \hat{q}_n) = (w, g)_L$$

which gives using 5.1: Find $(d, u, \hat{q}_n) \in Y \times X \times \hat{X}$ satisfying

$$(d, v)_Y + (\nabla u, \nabla v)_{\Omega_h} + ik\langle u, v\rangle_{\partial\Omega} - k^2(u, v)_{\Omega_h} - \langle \hat{q}_n, v\rangle_{\partial\Omega_h \setminus \partial\Omega} = 0, \qquad \forall v \in Y, \tag{5.11a}$$

$$(\nabla d, \nabla w)_{\Omega_h} + ik\langle d, w\rangle_{\partial\Omega} - k^2(d, w)_{\Omega_h} = (g, w)_{\Omega_h} \qquad \forall w \in X, \tag{5.11b}$$

$$\langle d, \hat{w}\rangle_{\partial\Omega_h} = 0 \qquad \forall \hat{w} \in \hat{X}. \tag{5.11c}$$

Equation 5.11c implies that $d \in H^1(\Omega)$. Therefore, integration by parts for equation 5.11b over $\Omega$ to get

$$-(\Delta d, w)_\Omega + \langle \nabla d \cdot n, w \rangle_{\partial\Omega} + ik\langle d, w \rangle_{\partial\Omega} - k^2(d, w)_\Omega = (g, w)_\Omega$$

which implies,

$$\Delta d + k^2 d = -g, \text{ in } \Omega, \tag{5.12a}$$

$$\nabla d \cdot n + ikd = 0 \text{ on } \partial\Omega, \tag{5.12b}$$

Equation 5.11a with $v \in H_0^1(\Omega)$ yields

$$(d, v)_Y + (\nabla u, \nabla v)_{\Omega_h} - k^2(u, v)_{\Omega_h} = 0, \tag{5.13a}$$

$$-(\Delta u, v)_{\Omega_h} - k^2(u, v)_{\Omega_h} = -(d, v)_{\Omega_h} + (\Delta d, v)_{\Omega_h} \tag{5.13b}$$

Where the last equation is obtained by integrating by parts and the definition of the $H^1-$norm, this implies

$$\Delta u + k^2 u = d - \Delta d \tag{5.14}$$

$$= d + g + k^2 d, \text{ by } 5.12a \tag{5.15}$$

$$= (k^2 + 1)d + g \tag{5.16}$$

Equation 5.11a gives

$$\langle \hat{q}_n, v\rangle_{\partial\Omega_h\setminus\partial\Omega} = (d,v)_{\Omega_h} + (\nabla d, \nabla v)_{\Omega_h} + (\nabla u, \nabla v)_{\Omega_h} + ik\langle u, v\rangle_{\partial\Omega} - k^2(u,v)_{\Omega_h}$$

Integration by parts and equation 5.14 yield

$$\langle \hat{q}_n, v\rangle_{\partial\Omega_h\setminus\partial\Omega} - ik\langle u, v\rangle_{\partial\Omega} = \langle \nabla d\cdot n, v\rangle_{\partial\Omega_h} + \langle \nabla u\cdot n, v\rangle_{\Omega_h}$$

which results in the following two equations

$$\langle \hat{q}_n, v\rangle_{\partial\Omega_h\setminus\partial\Omega} = \langle \nabla d\cdot n, v\rangle_{\partial\Omega_h\setminus\partial\Omega} + \langle \nabla u\cdot n, v\rangle_{\partial\Omega_h\setminus\partial\Omega}, \tag{5.17a}$$

$$-ik\langle u, v\rangle_{\partial\Omega} = \langle \nabla d\cdot n, v\rangle_{\partial\Omega} \qquad + \langle \nabla u\cdot n, v\rangle_{\partial\Omega} \tag{5.17b}$$

Equation 5.17a implies

$$\hat{q}_n = \nabla d\cdot n + \nabla u\cdot n, \text{ on } \partial\Omega_h\setminus\partial\Omega \tag{5.18}$$

Equations 5.17b and 5.12b imply

$$\nabla u\cdot n + iku = ikd, \text{ on } \partial\Omega \tag{5.19}$$

The following is a summary of the boundary value equations we've got,

$$\Delta d + k^2 d = -g \qquad \text{on } \Omega, \tag{5.20a}$$

$$\nabla d\cdot n + ikd = 0 \qquad \text{on } \partial\Omega. \tag{5.20b}$$

$$\Delta u + k^2 u = (k^2+1)d + g \qquad \text{on } \Omega, \tag{5.21a}$$

$$\nabla u \cdot n + iku = ikd \qquad \text{on } \partial\Omega. \tag{5.21b}$$

Thus, by full regularity on a convex domain $d, u \in H^2(\Omega)$, and moreover,

$$\begin{aligned}
\|d\|_{Z_1} &\leqslant C\|g\|_L, \\
\|u\|_{Z_2} &\leqslant C\left(\|d\|_L + \|g\|_L\right) \leqslant C\|g\|_L, \\
\|\hat{q}_n\|_{\hat{X}} &\leqslant \|\nabla(d+u)\|_{H(\mathrm{div},\Omega)} \\
&\leqslant C\|g\|_L.
\end{aligned}$$

Hence

$$\|(d, u, \hat{q}_n)\|_Z \leqslant C\|g\|_L. \tag{5.22}$$

Assume that $Y_h, X_h$, and $\hat{X}_h$ are finite dimensional subspaces of $Y, X$, and $\hat{X}$, respectively. by the Bramble-Hilbert lemma and standard approximation theory, for $W = (v_h, w_h, \hat{w}_h) \in Y_h \times X_h \times \hat{X}_h$, we have

$$\begin{aligned}
&\|U(g) - W\|^2_{Y\times X\times\hat{X}} \\
&\quad = \|d - v_h\|^2_{H^1(\Omega_h)} + \|u - w_h\|^2_{H^1(\Omega)} + \|\hat{q}_n - \hat{w}_h\|^2_{H^{-1/2}(\partial\Omega_h)} \\
&\quad \leqslant Ch^2\left(|d|^2_{H^2(\Omega)} + |u|^2_{H^2(\Omega)}\right) + \|(I-\Pi)\nabla(d+u)\|^2_{H(\mathrm{div},\Omega)}
\end{aligned} \tag{5.23}$$

where $\Pi$ is the Raviart-Thomas projection. By its commutativity property with

the $L-$projection $P$,

$$\begin{aligned}
&\|(I-\Pi)\nabla(d+u)\|^2_{H(\mathrm{div},\Omega)} \\
&\quad\leqslant \|(I-\Pi)\nabla(d+u)\|^2_L + \|(I-P)\Delta(d+u)\|^2_L \\
&\quad\leqslant \|(I-\Pi)\nabla(d+u)\|^2_L + C(\|(I-P)d)\|^2_L + \|(I-P)u)\|^2_L) \\
&\quad\leqslant Ch\left(|d+u|^2_{H^2(\Omega)} + |d|^2_{H^1(\Omega)} + |u|^2_{H^1(\Omega)}\right)
\end{aligned} \tag{5.24}$$

Where the last inequality is obtained by the Bramble-Hilbert lemma, which implies

$$\|U(g)-W\|^2_{Y\times X\times\hat{X}} \leqslant Ch\|g\|_L \tag{5.25}$$

So assumption 5.1 is verified. Therefore, by theorem 5.2, the interior $x_h$ converges (in the $L^2$-norm) one order faster than the numerical traces.

# Chapter 6

# The Convergence Rates of The DPG Method with Rectangular Meshes

## 6.1 Introduction

The goal of this chapter is to study the convergence rates of the DPG method for rectangular meshes. The Poisson equation with Dirichlet boundary condition is the primary example to study the convergence rates here. In order to show the discrete stability of the practical DPG method, we need to prove the existence of a Fortin operator as in the assumption 2.5. We will use the Fortin operator which was proven to exist in [9] to introduce some of the cases which we intend to study in the chapter.

Our goal is to improve the convergence rates of the DPG method. To achieve this, we will use some techniques from [31] to introduce new cases in section 6.4.

We need first to introduce the problem and the rectangular finite dimensional spaces before defining the cases which we are going to analyze as we have done in the previous chapters.

## 6.2 Application to the Poisson Equation

Suppose $\Omega$ is a bounded open polygon in $\mathbb{R}^2$ with Lipschitz boundary, meshed by $\Omega_h$, a geometrically conforming shape regular finite element mesh of quarilaterals. Let $\partial\Omega_h$ denote the collection of all element boundaries $\partial K$ for all elements $K$ in $\Omega_h$, where $h$ is the mesh size. We now study the DPG approximation to the Dirichlet problem

$$-\Delta u = f \qquad \text{on } \Omega, \tag{6.1a}$$

$$u = 0 \qquad \text{on } \partial\Omega. \tag{6.1b}$$

All functions are real-valued in this section.

The weak formulation reads:

Find $(u, \hat{q}_n) \in X = X_0 \times \hat{X}$ satisfying

$$(\nabla u, \nabla v)_{\Omega_h} - \langle \hat{q}_n, v \rangle_{\partial\Omega_h} = (f, v)_{\Omega_h} \quad \forall v \in Y \tag{6.2}$$

where

$$b_0(u, v) = (\nabla u, \nabla v)_{\Omega_h}, \qquad \hat{b}(\hat{q}_n, v) = -\langle \hat{q}_n, v \rangle_{\partial\Omega_h}, \qquad \ell(v) = (f, v)_{\Omega}.$$

and

$$\begin{aligned} X_0 &= H_0^1(\Omega), \\ \hat{X} &= H^{-1/2}(\partial\Omega_h), \\ Y &= H^1(\Omega_h), \end{aligned}$$

Assumption 2.4 was verified for this formulation in [19]. So the exact solution of the resulting weak formulation (2.1) is denoted by $(u, \hat{q}_n) \in X$. Note that $\hat{q}_n|_{\partial K} = \partial_n u|_{\partial K}$ for all $K \in \Omega_h$.

Now, we want to set the discrete spaces. But before doing that, we want to define some spaces in order to determine what discrete spaces we are going to use.

Let $\hat{K}$ be the unit square $[0, 1] \times [0, 1]$ in the $(\xi, \eta)$-plane with vertices $a_1 = (0, 0)$, $a_2 = (1, 0)$, $a_3 = (1, 1)$, $a_4 = (0, 1)$. For any integer $k > 1$, we define $Q_{k,k}(\hat{K})$ to be the space of polynomials of degree at most $k$ in each variable separately on

$\hat{K}$ which can be written in the form

$$p(\xi,\eta) = \sum_{0\leqslant i,j\leqslant k} c_{ij}\xi^i\eta^j.$$

For the numerical fluxes and traces we need local polynomial spaces defined on the boundary $\partial K$ as

$$P_r(\partial K) = \{P \in L^2(\partial K), P|_E \in P_r(E) \text{ for all edges } E \text{ of } K\},$$

$$\widetilde{P_r}(\partial K) = P_r(\partial K) \cap C(\partial K)$$

where $P_r(E)$ stands for polynomials of degree $r$ on $E$ and $C(\partial K)$ stands for the space continuous functions on $\partial K$. With the above-defined finite dimensional space, we introduce the following cases to solve 6.1 using the DPG method. We want to study these cases, and compare their convergence rates. That is, for any integer $k \geqslant 1$, we set

| Case 4 | Case 5 | Case 6 |
|---|---|---|
| $X_{h,0} = Q_{k,k}(\Omega_h) \cap X_0$ | $X_{h,0} = Q_{k,k}(\Omega_h) \cap X_0$ | $X_{h,0} = Q_{k+1,k+1}(\Omega_h) \cap X_0,$ |
| $\hat{X}_h = P_k(\partial\Omega_h) \cap \hat{X}$ | $\hat{X}_h = P_{k-1}(\partial\Omega_h) \cap \hat{X}$ | $\hat{X}_h = P_k(\partial\Omega_h) \cap \hat{X},$ |
| $Y^r = Q_{k+2,k+2}(\Omega_h)$ | $Y^r = Q_{k+2,k+2}(\Omega_h)$ | $Y^r = Q_{k+2,k+2}(\Omega_h).$ |

We introduce case 4 with exactly the same spaces used in the construction of the $\Pi$-operator which was constructed in [9]. By using the Bramble-Hilbert Lemma, we observe that the convergence rates in the $H^1$-norm (when the primal variable is

approximated) is one degree lower than that of the flux variable $\hat{q}_n$, which motivates us to introduce case 5 as the convergence rates for the two variables are the same. Also, case 5 is cheaper than case 4, yet both have the same convergence rate. as summarized in the table.

Case 6 is introduced since the convergence rates match for the two variables. We have proven theoretically that the convergence rate is one degree higher compared to the cases 4 and 5. All convergence rates are summarized in the table.

## 6.3 Error Analysis

In this section, we want to analyze cases 4 and 5. We will show the well-posedness of these cases by proving that the assumption 2.5 is satisfied depending on the Fortin operator of [9].

**Lemma 6.1.** *Let $B(K)$ be defined as*

$$B(K) = \{z \in Q_{k+2,k+2}(K) : z \text{ is zero at the vertices of } K\}.$$

Table 6.1: Summary of the convergence rates

| | $h$-convergence rates of $u_h$ | |
|---|---|---|
| | in $H^1(\Omega)$ | in $L^2(\Omega)$ |
| Case 4 | $k$ | $k+1$ |
| Case 5 | $k$ | $k+1$ |
| Case 6 | $k+1$ | $k+2$ |

*Then there exists a projector $R_K^0$ onto $B(K) \subset H^1(K)$ such that*

$$(R_K^0 z, v)_K = (z, v)_K \qquad \forall v \in Q_{k,k}(K), \tag{6.3}$$

$$\left\langle R_K^0 z, w \right\rangle_{\partial K} = \langle z, w \rangle_{\partial K} \qquad \forall w \in P_k(\partial K), \tag{6.4}$$

$$h_K^{-1} \left\| R_K^0 z \right\|_{L^2(K)} + \left| R_K^0 z \right|_{H^1(K)} \leqslant C(h_K^{-1} \left\| z \right\|_{L^2(K)} + |z|_{H^1(K)}) \qquad \forall z \in H^1(K). \tag{6.5}$$

*Proof.* The proof can be found in ( [9], Lemma 4.2). □

We can now construct a projector into the enriched finite element space such that the $H^1$-norm is bounded by an $h$-independent number. This is the content of the following Lemma.

**Lemma 6.2.** *There is a projector $R_K$ from $H^1(K)$ into $Q_{k+2,k+2}(K)$ such that*

$$(R_K z, v)_K = (z, v)_K \qquad \forall v \in Q_{k,k}(K), \tag{6.6}$$

$$\langle R_K z, w \rangle_{\partial K} = \langle z, w \rangle_{\partial K} \qquad \forall w \in P_k(\partial K), \tag{6.7}$$

$$\left\| R_K z \right\|_{H^1(K)} \leqslant C \left\| z \right\|_{H^1(K)} \qquad \forall z \in H^1(K). \tag{6.8}$$

*Proof.* The proof can be found in ( [9], Lemma 4.3). □

The next theorem shows the unique solvability for cases 4 and 5 and the convergence rate for each case.

**Theorem 6.3.** *In Cases 4 and 5, these statements hold:*

*i) The DPG method is uniquely solvable.*

*ii) The solution* $(u_h, \hat{q}_{n,h})$ *of the DPG method satisfies*

$$\|u - u_h\|_{H^1(\Omega)} + \|\hat{q}_n - \hat{q}_{n,h}\|_{H^{-1/2}(\partial\Omega_h)} \leqslant Ch^k \left(|u|_{H^{k+1}(\Omega)} + |f|_{H^k(\Omega)}\right). \quad (6.9)$$

*iii) If* $\Omega$ *is convex, then*

$$\|u - u_h\|_{L^2(\Omega)} \leqslant Ch^{k+1} \left(|u|_{H^{k+1}(\Omega)} + |f|_{H^k(\Omega)}\right). \quad (6.10)$$

*Proof.* By theorem 2.6, if we verify assumption 2.5, then the DPG method is uniquely solvable. But this assumption was verified by lemmas 6.1 and 6.2.

The proof of 6.9 is straight forward by 2.19. Also, the final estimate is proved by verifying Assumption 5.1 (along the lines of the proof of theorem 5.5) and applying theorem 5.2. □

Another way to match the convergence rates of the primal variable $u$ and the flux variable $\hat{q}$ is to introduce case 6. The next theorem analyzes case 6.

**Theorem 6.4.** *In Case 6, for* $k \geqslant 1$*, these statements hold:*

*i)* $\hat{B}_h$ *is injective and the DPG method is uniquely solvable.*

*ii) The $u_h$-component of the solution satisfies*

$$\|u - u_h\|_{H^1(\Omega)} \leqslant Ch^{k+1} \left( |u|_{H^{k+2}(\Omega)} + |f|_{H^{k+1}(\Omega)} \right). \tag{6.11}$$

*Proof.* First, observe that if $k \geqslant 1$, then by the unisolvency of the DPG method in Case 4, its $B_h$ is injective, which implies by Theorem 2.11 that $\hat{B}_h$ of Case 4 is injective. But since the flux $(\hat{X}_h)$ and test spaces $(Y^r)$ of Case 6 are identical to that of Case 4, both cases have the same $\hat{B}_h$. Hence $\hat{B}_h$ of Case 6 is injective and consequently by Theorem 2.11, $B_h$ of Case 6 is injective. Thus we have proved the first statement of the theorem for $k \geqslant 1$.

The the second part of the theorem can be proven in the same way as the proof of theorem 6.7.

□

## 6.4 Reduced-Degree Test Spaces

In this section, we are going to introduce more interesting cases. These are cheaper than the cases 4 and 6, respectively, yet they provide the same convergence rate in the primal variable. We need first two lemmas, then we can introduce the discrete spaces which are used in cases 7 and 8.

**Lemma 6.5** (see [31])**.** *Assume that $\hat{K}$ is the unit square. Let $k \geqslant 1$ be an integer.*

*Then the space of functions* $\mu \in P_{k-1}(\partial\hat{K})$ *such that*

$$\forall v \in \widetilde{P_k}(\partial\hat{K}), \int_{\partial\hat{K}} \mu v \, d\gamma = 0 \tag{6.12}$$

*is one dimensional.*

*Proof.* Since $\hat{K}$ is a unit square, we have $\widetilde{P_k}(\partial\hat{K}) = Q_{k,k}(\hat{K})|_{\partial\hat{K}}$ and $dim(P_{k-1}(\partial\hat{K})) = 4k$, $dim(\widetilde{P_k}(\partial\hat{K})) = 4k$.

For all $v \in \widetilde{P_k}(\partial\hat{K})$, we get $\mu v \in P_{2k-1}(\partial\hat{K})$ so that the integral $\int_{\partial\hat{K}} \mu v \, d\gamma$ can be computed exactly in terms of the values of the function $\mu v$ at $(k+1)$ *Gauss-Lobatto quadrature points* of each side of $\partial\hat{K}$. Denote by $\{a_1, a_5, \ldots, a_{k+3}, a_2\}$ (resp. $\{a_2, a_{k+4}, \ldots, a_{2k+2}, a_3\}$, $\{a_3, a_{2k+3}, \ldots, a_{3k+1}, a_4\}$, $\{a_4, a_{3k+2}, \ldots, a_{4k}, a_1\}$) the set of $(k+1)$ Gauss-Lobatto points of the side $[a_1, a_2]$ (resp. $[a_2, a_3]$, $[a_3, a_4]$, $[a_4, a_1]$). Clearly, for each $i = 1, \ldots, 4k$, there exists a unique function $v_i \in \widetilde{P_k}(\partial\hat{K})$ such that

$$v_i(a_j) = \delta_{ij}, \qquad 1 \leqslant j \leqslant 4k.$$

Then, replacing $v$ by $v_i$, $1 \leqslant i \leqslant 4k$, in 6.12 gives

$$\mu(a_i) = 0, \;\; 5 \leqslant i \leqslant 4k \tag{6.13}$$

and

$$\begin{cases} \mu_{12}(a_1) + \mu_{14}(a_1) &= 0, \\ \mu_{21}(a_2) + \mu_{23}(a_2) &= 0, \\ \mu_{32}(a_3) + \mu_{34}(a_3) &= 0, \\ \mu_{41}(a_4) + \mu_{43}(a_4) &= 0, \end{cases} \tag{6.14}$$

where $\mu_{ij} = \mu_{ji}$ is the restriction of $\mu$ to the $[a_i, a_j]$. Let

$$0 = \xi_0 < \xi_1 < \cdots < \xi_{k-1} < \xi_k = 1$$

be the $(k+1)$ Gauss-Lobatto abscissea for $[0,1]$; we introduce the homogeneous polynomial of degree $k-1$ in the variables $\xi$ and $\eta$

$$p_{k-1}(\xi, \eta) = \prod_{i=1}^{k-1} (\eta_i \xi - \xi_i \eta) \tag{6.15}$$

where $\eta_i = 1 - \xi_i$, $1 \leqslant i \leqslant k-1$. Since $\eta_i = \xi_{k-i}$, $1 \leqslant i \leqslant k-1$ we get

$$p_{k-1}(\xi, \eta) = (-1)^{k-1} p_{k-1}(\eta, \xi).$$

Thus, conditions 6.13 mean

$$\begin{cases} \mu_{12} = c_{12} p_{k-1}(\xi, 1-\xi), \\ \mu_{23} = c_{23} p_{k-1}(\eta, 1-\eta), \\ \mu_{34} = c_{34} p_{k-1}(1-\xi, \xi), \\ \mu_{41} = c_{41} p_{k-1}(1-\eta, \eta), \end{cases} \tag{6.16}$$

Using 6.16, conditions 6.14 become

$$\begin{cases} c_{12} p_{k-1}(\xi, 1-\xi) + c_{14} p_{k-1}(\eta, 1-\eta) &= 0, \\ c_{21} p_{k-1}(1-\xi, \xi) + c_{23} p_{k-1}(\eta, 1-\eta) &= 0, \\ c_{32} p_{k-1}(1-\eta, \eta) + c_{34} p_{k-1}(1-\xi, \xi) &= 0, \\ c_{41} p_{k-1}(1-\eta, \eta) + c_{43} p_{k-1}(\xi, 1-\xi) &= 0, \end{cases} \tag{6.17}$$

which implies,

$$\begin{cases} c_{12} + c_{14} = 0, \\ (-1)^{k-1} c_{12} + c_{23} = 0, \\ (-1)^{k-1} c_{23} + (-1)^{k-1} c_{43} = 0, \\ c_{43} + (-1)^{k-1} c_{14} = 0, \end{cases} \tag{6.18}$$

Which implies,

$$\begin{cases} c_{12} = -c_{23} = c_{34} = -c_{41} = c & \text{when k is odd}, \\ c_{12} = c_{23} = c_{34} = c_{41} = c & \text{when k is even}. \end{cases} \tag{6.19}$$

The system 6.18 has nontrivial solutions for all k, and the space of $\mu_{k-1}(\partial\hat{K})$ is one-dimensional.

□

In [31], they have constructed a suitable hybrid element by introducing a new space as in the following lemma.

**Lemma 6.6** (see [31]). *Let $k \geqslant 1$ be an integer. Define $\hat{P}_k(\hat{K})$ to be the space of polynomials spanned by $Q_{k,k}(\hat{K})$ and the function*

$$v_0(\xi,\eta) = \begin{cases} H(\xi,\eta)[(\xi(1-\xi))^{(k-1)/2} + (\eta(1-\eta))^{(k-1)/2}] & , k \text{ is odd} \\ H(\xi,\eta)(2\xi-1)(2\eta-1)[(\xi(1-\xi))^{(k-2)/2} + (\eta(1-\eta))^{(k-2)/2}] & , k \text{ is even} \end{cases} \tag{6.20}$$

*where $H(\xi,\eta) = \xi(1-\xi) - \eta(1-\eta)$. Then the pair of spaces $(\hat{P}_k(\hat{K}), P_{k-1}(\partial\hat{K}))$ satisfies*

$$\left\{\mu_{k-1}(\partial\hat{K}); \forall v \in \hat{P}_k(\hat{K}), \int_{\partial\hat{K}} \mu v \, d\gamma = 0\right\} = \{0\} \tag{6.21}$$

*Proof.* Let $\mu$ be a function of $P_{k-1}(\partial\hat{K})$ such that

$$\forall v \in \hat{P}_k(\hat{K}), \quad \int_{\partial\hat{K}} \mu v \, d\gamma = 0$$

By Lemma 6.5, $\mu$ can be written in the form 6.16. Hence, it is sufficient to prove that

$$\sum_{i=1}^{4} \int_{[a_i,a_{i+1}]} p_{k-1} v_0 \, d\gamma \neq 0 \quad (a_1 = a_5)$$

Since these four integrals are equal, we have only to check that

$$\int_{[a_1,a_2]} p_{k-1}(\xi, 1-\xi)v_0 \, d\xi \neq 0. \tag{6.22}$$

When $k$ is odd, the left-hand side of 6.22 can be written as

$$\int_{[a_1,a_2]} q_{k-1}(\xi)r(\xi)\xi(1-\xi) \, d\xi, \tag{6.23}$$

where

$$q_{k-1} = p_{k-1}(\xi, 1-\xi), \quad r(\xi) = (\xi(1-\xi))^{(k-1)/2}$$

Since the roots of $q_{k-1}$ are the Gauss-Lobatto abscissae $\xi_1, \ldots, \xi_{k-1}$, the polynomial $q_{k-1}$ is orthogonal to polynomials of degree $\leqslant k-2$ with respect to the weight function $\xi(1-\xi)$. Now, $r$ is a polynomial of degree $k-1$ so that

$$\int_0^1 q_{k-1}r(\xi)\xi(1-\xi)d\xi \neq 0.$$

Otherwise, $q_{k-1}$ would be orthogonal to all polynomials of degree $\leqslant k-1$ which is clearly impossible, so $c_{12} = c_{23} = c_{34} = c_{41} = c = 0$, which implies that $\mu = 0$. A similar argument for even k yields that $\mu = 0$ □

Using lemma 6.6, we introduce cases 7 and 8 as follows

| Case 7 | Case 8 |
|---|---|
| $X_{h,0} = Q_{k,k}(\Omega_h) \cap X_0$ | $X_{h,0} = Q_{k+1,k+1}(\Omega_h) \cap X_0$ |
| $\hat{X}_h = P_k(\partial\Omega_h) \cap \hat{X}$ | $\hat{X}_h = P_k(\partial\Omega_h) \cap \hat{X}$ |
| $Y^r = \hat{P}_{k+1}(\Omega_h)$ | $Y^r = \hat{P}_{k+1}(\Omega_h)$ |

Case 7 is comparable with case 4, as they have the same trial finite dimensional subspaces $X_{h,0}$ and $\hat{X}_h$. But $Y^r$ of case 7 has less degrees of freedom than those of case 4. Yet both cases have the same convergence rate. By the Bramble-Hilbert Lemma, the convergence rates in case 7 do not match, which motivates us to introduce case 8 in order to get the same convergence rates in the $H^1$-norm.

In next theorems, we will prove the unique solvability of the DPG method for cases 7 and 8. In addition, the convergence rate of cases 7 and 8 will be proven theoretically, which matches what we have observed numerically, to be of orders $k$ and $k+1$, respectively.

**Theorem 6.7.** *In Cases 7 and 8, for $k \geqslant 1$, these statements hold:*

*i) $\hat{B}_h$ is injective and the DPG method is uniquely solvable.*

*ii) In case 7, the $u_h$-component of the solution satisfies*

$$\|u - u_h\|_{H^1(\Omega)} \leqslant Ch^k \left(|u|_{H^{k+1}(\Omega)} + |f|_{H^k(\Omega)}\right). \tag{6.24}$$

*iii) In case 8, the $u_h$-component of the solution satisfies*

$$\|u - u_h\|_{H^1(\Omega)} \leqslant Ch^{k+1} \left( |u|_{H^{k+2}(\Omega)} + |f|_{H^{k+1}(\Omega)} \right). \tag{6.25}$$

*Proof.* Lemma 6.6 implies that the operator $\hat{B}_h$ is injective for case 4. In addition, theorem 2.11 implies the injectivity of the operator $B_h$, since the three assumptions are obviously satisfied in this case.

Next we prove (6.24). It is similar to the proof of theorem (3.5) of [6]. Recall that $Y_0^r$ is defined in (2.11) and $Y_{h,0}^r$ in (2.12b). By Theorem 2.9, $u_h \in X_{h,0}$ satisfies (2.12b), i.e.,

$$b_0(u_h, y) = (f, y)_\Omega, \qquad \forall y \in Y_{h,0}^r. \tag{6.26}$$

We proceed by viewing this as a nonconforming Petrov-Galerkin discretization of

$$b_0(u, y) = (f, y)_\Omega, \qquad \forall y \in H_0^1(\Omega)$$

and bounding the consistency error in an argument akin to the second Strang lemma. Let $C_p$ denote the constant, derived from Poincaré inequality, such that

$\|w\|_{H^1(\Omega)} \leqslant C_p \|\operatorname{grad} w\|_{L^2(\Omega)}$ for all $w \in H_0^1(\Omega)$. Then, for any $w_h \in X_{h,0}$

$$\begin{aligned}
\|u_h - w_h\|_{H^1(\Omega)} &\leqslant C_p \sup_{z_h \in X_{h,0}} \frac{(\operatorname{grad}(u_h - w_h), \operatorname{grad} z_h)_\Omega}{\|\operatorname{grad} z_h\|_{L^2(\Omega)}} \leqslant C_p^2 \sup_{z_h \in X_{h,0}} \frac{b_0(u_h - w_h, z_h)}{\|z_h\|_{H^1(\Omega)}} \\
&\leqslant C_p^2 \sup_{y \in Y_0^r} \frac{b_0(u_h - w_h, y)}{\|y\|_Y} = C_p^2 \|T_0^r(u_h - w_h)\|_Y = C_p^2 \sup_{y \in Y_{h,0}^r} \frac{b_0(u_h - w_h, y)}{\|y\|_Y} \\
&= C_p^2 \sup_{y \in Y_{h,0}^r} \frac{b_0(u_h - u, y) + b_0(u - w_h, y)}{\|y\|_Y} \\
&= C_p^2 \sup_{y \in Y_{h,0}^r} \frac{(f, y)_\Omega - b_0(u, y) + b_0(u - w_h, y)}{\|y\|_Y}, \qquad (6.27)
\end{aligned}$$

where we have used (6.26). Since $b((u, \hat{q}_n), y) = (f, y)_\Omega$ for all $y \in Y$, the term representing the consistency error in (6.27) can be written as $(f, y)_\Omega - b_0(u, y) = \hat{b}(\hat{q}_n, y)$. By the definition of $Y_0^r$ (see (2.11)), we also have $\hat{b}(\hat{q}_n, y) = \hat{b}(\hat{q}_n - \hat{r}_{n,h}, y)$ for any $\hat{r}_{n,h} \in \hat{X}_h$ and $y \in Y_0^r$. Therefore,

$$\|u_h - w_h\|_{H^1(\Omega)} \leqslant C_p^2 \sup_{y \in Y_{h,0}^r} \frac{b((u - w_h, \hat{q}_n - \hat{r}_{n,h}), y)}{\|y\|_Y} \leqslant C_p^2 C_2 \left( \|\hat{q}_n - \hat{r}_{n,h}\|_{\hat{X}} + \|u - w_h\|_{H^1(\Omega)} \right).$$

Since $\hat{r}_{n,h}$ and $\hat{q}_n$ are element-by-element traces of an $r_h$ in $R_k$ and $q = \operatorname{grad} u$, respectively, where $R_k$ is the Raviart-Thomas subspace [31] of $H(\operatorname{div}, \Omega)$ consisting of all vector functions which when restricted to an element takes the form $xp_1 + p_2$

for some $p_1 \in P_k(K)$ and some $p_2 \in P_k(K)^2$.

$$\begin{aligned}\|\hat{r}_{n,h} - \hat{q}_n\|_{\hat{X}} &\leqslant \|r_h - \operatorname{grad} u\|_{H(\operatorname{div},\Omega)} \\ &\leqslant Ch^{k+1}\left(|u|_{H^{k+1}(\Omega)} + |f|_{H^k(\Omega)}\right)\end{aligned}$$

where the last inequality is obtained from [1]. So

$$\|u_h - w_h\|_{H^1(\Omega)} \leqslant C\left(\inf_{r_h \in R_k} \|r_h - \operatorname{grad} u\|_{H(\operatorname{div},\Omega)} + \|u - w_h\|_{H^1(\Omega)}\right).$$

Finally, by the triangle inequality,

$$\begin{aligned}\|u - u_h\|_{H^1(\Omega)} &\leqslant \|u - w_h\|_{H^1(\Omega)} + \|u_h - w_h\|_{H^1(\Omega)} \\ &\leqslant C\left(\|u - w_h\|_{H^1(\Omega)} + h^{k+1}(|u|_{H^{k+1}(\Omega)} + |f|_{H^k(\Omega)})\right)\end{aligned}$$

for any $w_h \in X_{h,0}$. Choosing $w_h$ to be an appropriate interpolant, the proof of (6.24) is finished. Finally, the proof of (6.25) is similar.

□

## 6.5 Numerical Results

All the numerical results in this chapter obtained by solving the Poisson equation with Dirichlet boundary condition using the DPG method. The domain $\Omega$ set to

the unit square. The function $f$ was chosen so that the exact solution is

$$u = sin(\pi x)sin(\pi y) \tag{6.28}$$

We construct an $n \times n$ uniform mesh by dividing $\Omega$ into $n^2$ congruent squares. Its mesh size is $h = \sqrt{2}/n$. The method is applied on a sequence of such meshes with geometrically increasing $n$. The implementation of the method is done using Python and the NGSolve Finite Element Library and the mesh has been generated by the Netgen Mesh Generator.

The convergence rates for cases 4, 5, and 6 are shown in the figures below.

Figure 6.1: Case 4 $H^1-$Error

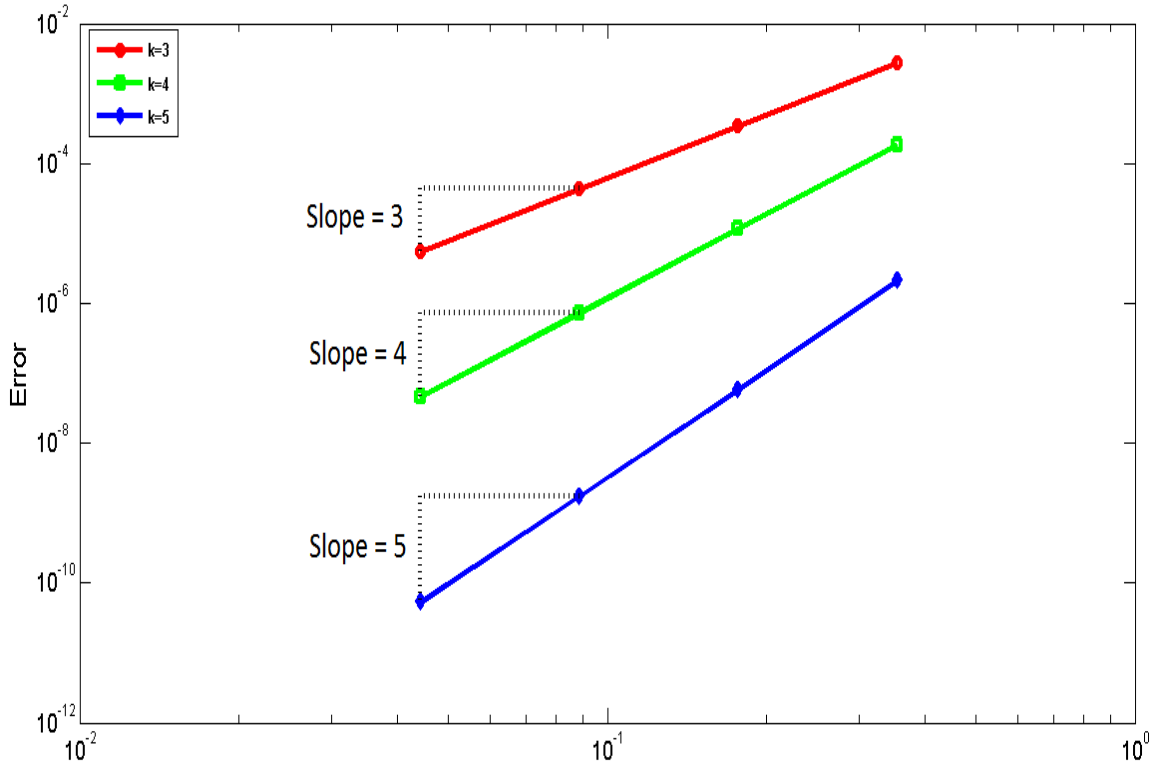


Figure 6.2: Case 5 $H^1-$Error

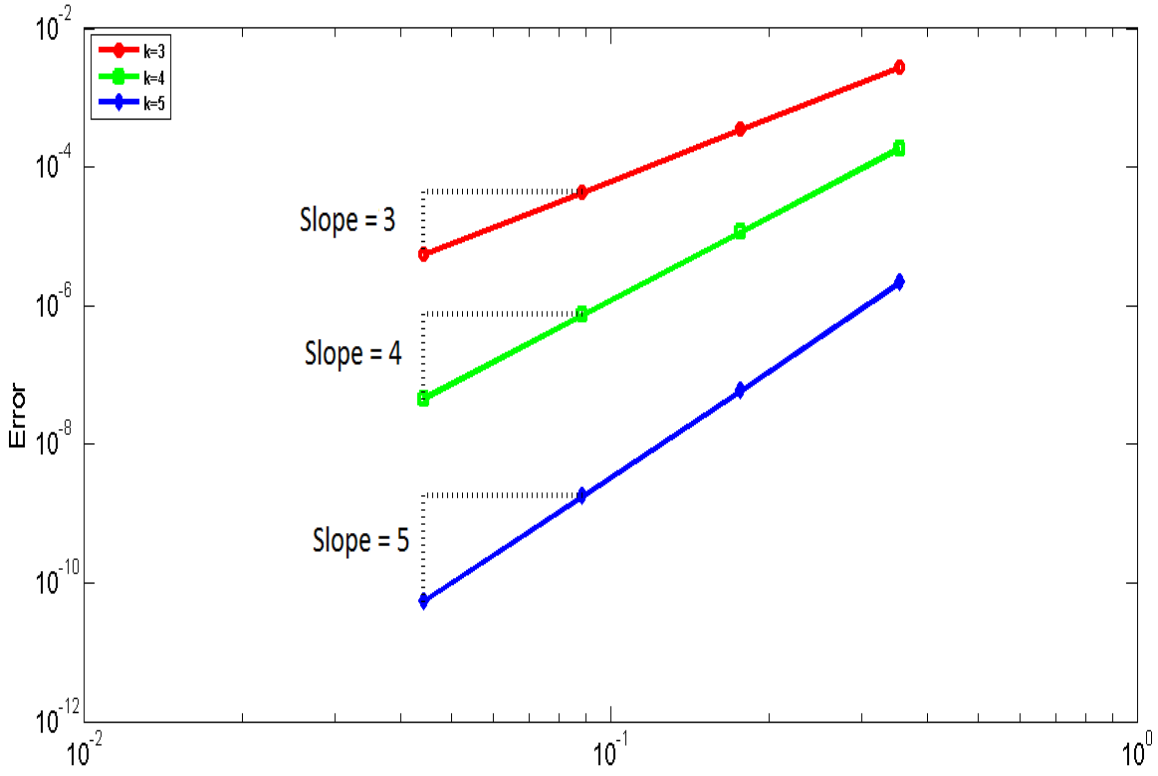


Figure 6.3: Case 4 $L^2-$Error

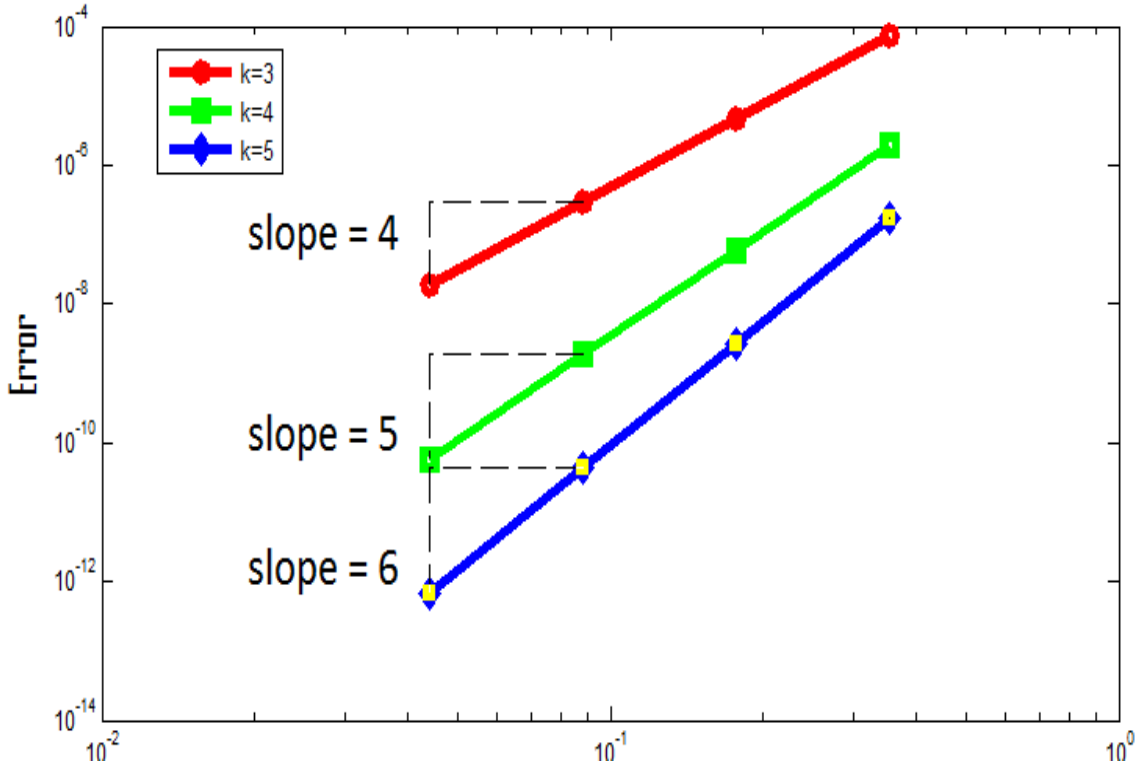


Figure 6.4: Case 5 $L^2-$Error

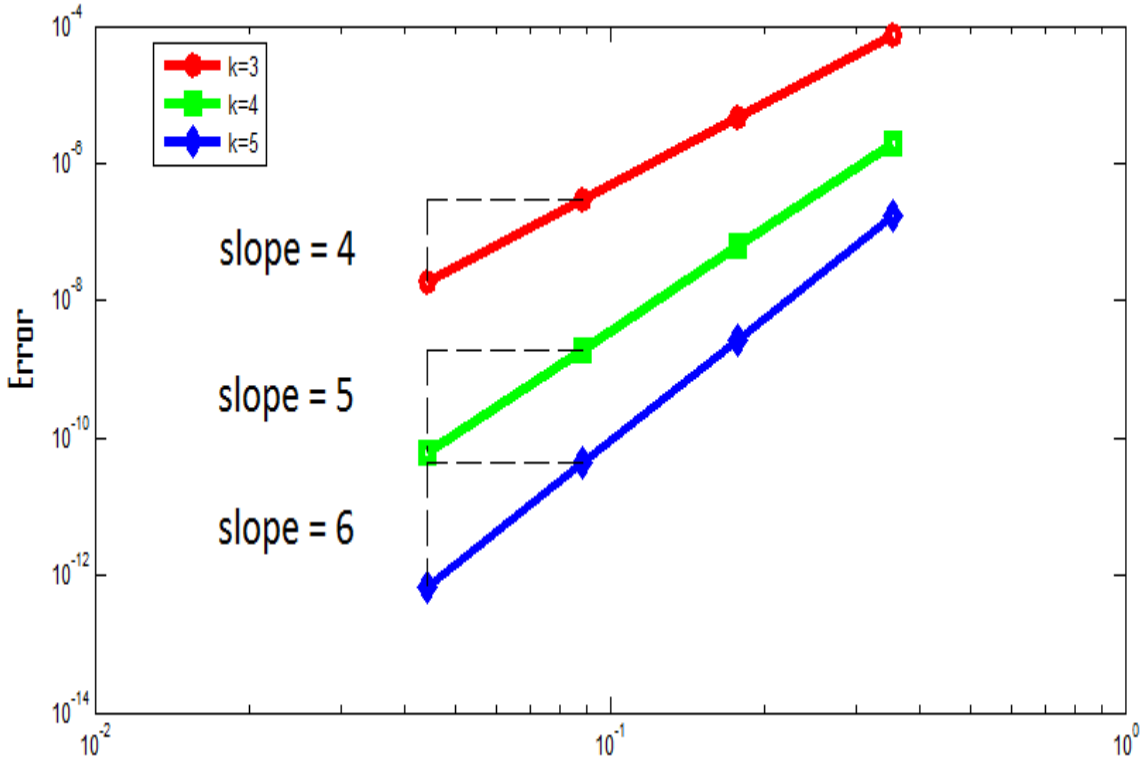
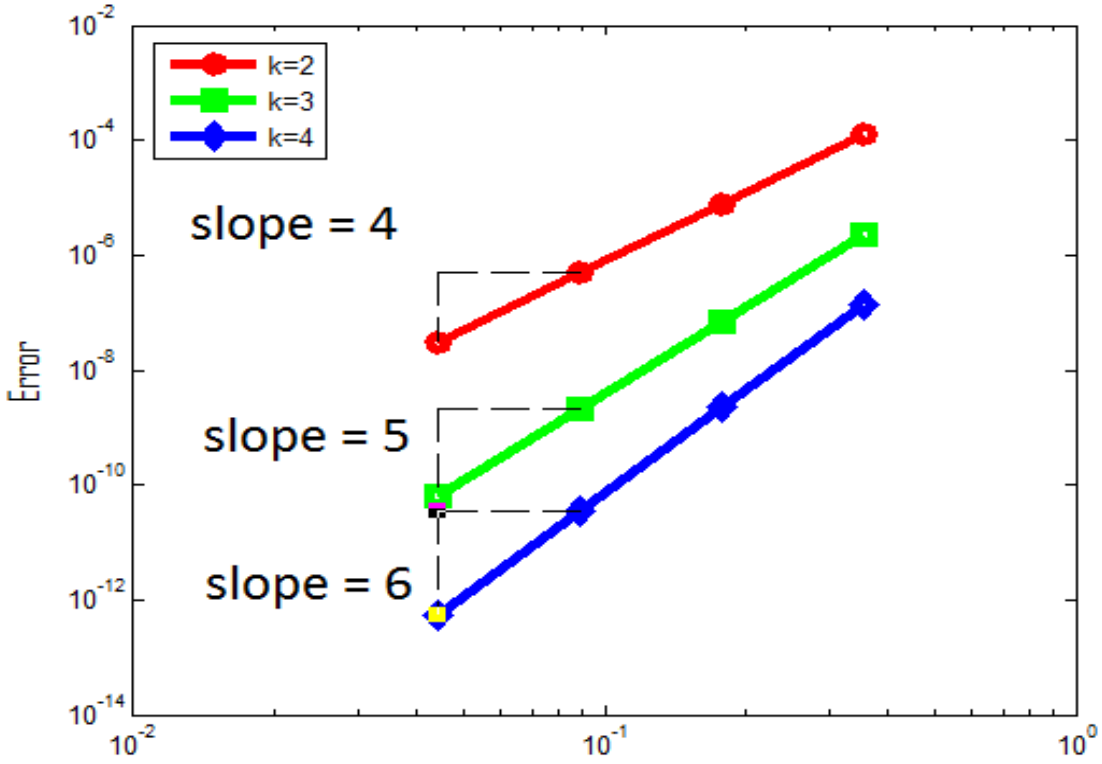


Figure 6.5: Case 6 $L^2-$Error



Some numerical results for cases 7 and 8 are shown in the following tables. It is worth mentioning that the $Y^r$ space is not standard in cases 7 and 8, so it is not implemented in the finite elements libraries. We have implemented it using a C++ code and we have used it in the NGSolve finite element library to get these numerical results.

Table 6.2: $H^1(\Omega)$ and $L^2(\Omega)$ convergence for the case 7

| $n$ | $\|u-u_h\|_{H^1(\Omega)}$ | rate | $\|u-u_h\|_{L^2(\Omega)}$ | rate |
|---|---|---|---|---|
| | | $k=2$ | | |
| 4 | 2.51E-02 | 1.93 | 4.69E-03 | 1.88 |
| 8 | 6.58E-03 | 1.87 | 1.27E-03 | 2.01 |
| 16 | 1.80E-03 | 1.99 | 3.16E-04 | 2.00 |
| 32 | 4.54E-04 | 2.13 | 7.87E-05 | 2.00 |
| 64 | 1.03E-04 | | 1.97E-05 | |

Table 6.3: $H^1(\Omega)$ and $L^2(\Omega)$ convergence for the case 8

| $n$ | $\|u-u_h\|_{H^1(\Omega)}$ | rate | $\|u-u_h\|_{L^2(\Omega)}$ | rate |
|---|---|---|---|---|
| | | $k=2$ | | |
| 4 | 5.11E-03 | 3.54 | 7.03E-05 | 4.02 |
| 8 | 4.40E-04 | 3.22 | 4.34E-06 | 3.97 |
| 16 | 4.71E-05 | 3.08 | 2.77E-07 | 3.99 |
| 32 | 5.59E-06 | 3.02 | 1.74E-08 | 4.00 |
| 64 | 6.87E-07 | | 1.09E-09 | |

The table below summarizes the cases covered in this chapter with their convergence rates in the $H^1$-norm and the $L^2$-norm.

Table 6.4: Summary of the convergence rates

| | $h$-convergence rates of $u_h$ | |
|---|---|---|
| | in $H^1(\Omega)$ | in $L^2(\Omega)$ |
| Case 4 | $k$ | $k+1$ |
| Case 5 | $k$ | $k+1$ |
| Case 6 | $k+1$ | $k+2$ |
| Case 7 | $k$ | $k$ |
| Case 8 | $k+1$ | $k+2$ |

# Chapter 7

# Future Work and Conclusion

The goal of this work has been to explore the convergence rates of the Discontinuous Petrov-Galerkin (DPG) method with triangular and rectangular meshes.

We began in Chapter 1 by introducing the DPG method as a Discontinuous Galerkin (DG) method and as a Hybrid Discontinuous Galerkin (HDG) method, and we give a brief literature review of each category of these finite element methods.

In Chapter 2, we introduced the general settings of the DPG method. Additionally, we present how the DPG method can be seen as a mixed method which makes it a standard finite element method where the implementation of the method easier. Another advantage of the mixed formulation is a built-in error representation function which is useful for adaptivity. Finally, we present the Poisson equation with Dirichlet boundary condition as a model problem to illustrate the results of this study. Specially, we use the primal weak formulation where only the conservation equation is integrated by parts.

In Chapter 3, we introduced the reduced degree DPG method for triangular meshes. The goal is to study the impact on the convergence rates of the DPG method. The polynomial degree of the finite dimensional test subspace has been decreased, and as a result, we have observed a parity in the behavior. We present the different behavior of the method for even and odd polynomial degrees. Furthermore, we explain this behavior by introducing counter examples for the even-degree case showing that the DPG method is not uniquely solvable. In contrast, we construct a Fortin operator for the odd-degree case which is required to prove the stability on the discrete level.

In Chapter 4, we showed another technique for error analysis, namely, the nonconforming analysis using Strang Lemma. The construction of a Fortin operator

is infeasible for some cases which motivated us to use global optimal test functions by using a weakly conforming test space. With this technique, we analyzed the error just for the primal variable, while the trace variable is vanished due to the use of the weakly conforming test space. The achievement is that we were able to reduce the test space degree and yet recover the convergence rate which is obtained previously by other researchers.

In Chapter 5, we presented a duality argument version for the DPG method. The theory is applicable for the DPG in general. It interprets the higher convergence rate in weaker norms. We showed how this theory is applied to the Poisson equation and the Helmholtz equation.

Finally, in chapter 6, we presented the construction of a reduced finite dimensional test subspace over rectangular meshes. As we have seen for the triangular meshes, sometimes constructing a Fortin operator is not possible. So we used a technique introduced by Raviart and Thomas [31], to construct a finite element space over rectangles which implies the unique solvability of the DPG method. Furthermore, we were able to recover the convergence rate of the standard DPG method (the one obtained by constructing a Fortin operator).

In conclusion, we have examined carefully the convergence rate of the DPG method over different type of meshes. We introduced new cases where the test space is reduced without losing the convergence rates obtained earlier by other researcher.

## 7.1 Accomplishments

In the theoretical scope of this dissertation, we have developed and proven a duality argument theory for the DPG methods. In particular, we have applied the

duality argument to two examples, namely, the Poisson equation with Dirichlet boundary condition and the Helmholtz Equation with impedance boundary condition. Furthermore, with the duality argument of the DPG method, we are able to measure the error in the $L^2$-norm and explain theoretically the one-order higher convergence rate than the one observed in the $H^1$-norm. Additionally, we have explained theoretically the parity in the behavior of the DPG method for some reduced test spaces. Also, we have applied the nonconforming analysis and the Strang lemma to the DPG method and gotten error estimates in $H^1$-norm. Finally, we used different techniques to construct reduced finite dimensional spaces for the DPG method with various type of meshes.

In the numerical and computational scope of this dissertation, we recovered the convergence rates of the DPG method despite using reduced test spaces. Convergence of the method is demonstrated under an exact solution to the Poisson equation with Dirichlet boundary condition problem. Those convergence rates obtained confirm the theoretical results.

## 7.2 Future work

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

- **Three Dimensional Meshes**

  We want to extend this study to partial differential equations over three dimensional domains. We might start with the Poisson equation with Dirichlet boundary condition and show the stability on the discrete level, which is achieved by constructing a Fortin operator over the unit cube. Then, we

may work on reducing the test space and come up with DPG methods which are cheaper than the one we aim to obtain by the Fortin operator.

- **Reduced Test Space for The DPG Method Applied to Various Problems**

  The DPG method has been applied to many Partial Differential Equations and exhibited its robustness. We plan to start from there and try to minimize the cost of the DPG method by constructing test spaces with reduced degrees, in order to recover the convergence rates obtained earlier.

# REFERENCES


[1] D.N. Arnold: An interior penalty finite element method with discontinuous elements. SIAM J. Numer. Anal. 19, 1982, 742760.

[2] J.-P. Aubin: Behavior of the error of the approximate solutions of boundary value problems for linear elliptic operators by Galerkins and finite difference methods, Ann. Scuola Norm. Sup. Pisa (3) 21 (1967), 599637.

[3] I. Babuska: The finite element method with penalty. Math. Comp. 27, 1973, 221228.

[4] G.A. Baker: Finite element methods for elliptic equations using nonconforming elements. Math. Comp. 31, 1977, 4559.

[5] Biagi, C. J., M. A. Uman, J. Gopalakrishnan, J. D. Hill, V. A. Rakov, T. Ngin, and D. M. Jordan (2011), Determination of the electric field intensity and space charge density versus height prior to triggered lightning, J. Geophys. Res., 116, D15201, doi:10.1029/2011JD015710.

[6] T. Bouma, J. Gopalakrishnan, A. Harb. Convergence rates of the DPG method with reduced test space degree. *Comput. Math. Appl.*, 68 (2014) 1550-1561.

[7] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*. Number 15 in Springer Series in Computational Mathematics. Springer-Verlag, New York, 1991.

[8] D. Broersen and R. Stevenson. A Petrov-Galerkin discretization with optimal test space of a mild-weak formulation of convection-diffusion equations in mixed form. *Preprint*, 2013.

[9] V. M. Calo, N. O. Collier, and A. H. Niemi. Analysis of the discontinuous Petrov-Galerkin method with optimal test functions for the Reissner-Mindlin plate bending model, Preprint: arXiv:1301.6149, (2013).

[10] C. Carstensen, L. Demkowicz, and J. Gopalakrishnan. *A posteriori* error control for DPG methods. *Preprint*, 2013.

[11] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*. North-Holland Publishing Company, Amsterdam, 1978.

[12] B. Cockburn, J. Gopalakrishnan, A Characterization of Hybridized Mixed Methods for Second Order Elliptic Problems, SIAM Journal on Numerical Analysis 42 (1) (2004) 283301.

[13] B. Cockburn, J. Gopalakrishnan, R. Lazarov, Unified Hybridization of Discontinuous Galerkin, Mixed, and Continuous Galerkin Methods for Second Order Elliptic Problems, SIAM Journal on Numerical Analysis 47 (2) (2009) 13191365.

[14] B. Cockburn, G.E. Karniadakis, and C.-W. Shu (eds.): Discontinuous Galerkin Methods. Theory, computation and applications. Papers from the

1st International Symposium held in Newport, RI, May 2426, 1999. Springer-Verlag, Berlin, 2000.

[15] W. Dahmen, C. Huang, C. Schwab, and G. Welper. Adaptive Petrov-Galerkin methods for first order transport equations. *SIAM J Numer. Anal.*, 50(5):2420–2445, 2012.

[16] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. Part I: The transport equation. Comput. Methods Appl. Mech. Engrg., (23-24):15581572, 2010. see also ICES Report 2009-12.

[17] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. Part II: Optimal test functions. Numer. Meth. Part. D. E., 27:70105, 2011. see also ICES Report 9/16.

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

[19] L. Demkowicz and J. Gopalakrishnan. A primal DPG method without a first-order reformulation. *Computers and Mathematics with Applications*, 66(6):1058–1064, 2013.

[20] L. Demkowicz and J. Gopalakrishnan. An overview of the DPG method. Technical report, ICES, 2013.

[21] L. Demkowicz, J. Gopalakrishnan, and A. Niemi, A class of discontinuous Petrov-Galerkin methods. Part III: Adaptivity, Applied Numerical Mathematics, 62 (2012), pp. 396-427.

[22] E.H. Georgoulis, Discontinuous Galerkin Methods for Linear Problems: An Introduction, In E. H. Georgoulis, A. Iske, and J. Levesley (eds.), Approximation Algorithms for Complex Systems, Springer Proceedings in Mathematics, Vol. 3, Springer-Verlag, Berlin, 2011

[23] J. Gopalakrishnan. Five lectures on DPG methods. Available as arXiV preprint 1306.0557, 2013.

[24] J. Gopalakrishnan, I. Muga, and N. Olivares. Dispersive and dissipative errors in the DPG method with scaled norms for the Helmholtz equation. J. Sci. Comput., 36 (2014), pp. A20A39.

[25] J. Gopalakrishnan and W. Qiu. An analysis of the practical DPG method. *Math. Comp*, electronically appeared, doi: 10.1090/S0025-5718-2013-02721-4, 2013.

[26] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. Number 24 in Monographs and Studies in Mathematics. Pitman Advanced Publishing Program, Marshfield, Massachusetts, 1985.

[27] C. Johnson and J. Pitkaranta: An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. Math. Comp. 46, 1986, 126.

[28] P. Lesaint and P.-A. Raviart: On a finite element method for solving the neutron transport equation. In: Mathematical Aspects of Finite Elements in Partial Differential Equations, Math. Res. Center, Univ. of Wisconsin-Madison, Academic Press, New York, 1974, 89123.

[29] J. Nitsche. Ein Kriterium für die Quasi-Optimalität des Ritzschen Verfahrens. *Numer. Math.*, 11:346–348, 1968.

[30] J. Nitsche: Uber ein Variationsprinzip zur L osung von Dirichlet-Problemen bei Verwendung von Teilraumen, die keinen Randbedingungen unterworfen sind. Abh. Math. Sem. Uni. Hamburg 36, 1971, 915.

[31] P.-A. Raviart and J. M. Thomas. Primal hybrid finite element methods for 2nd order elliptic equations. *Math. Comp.*, 31(138):391–413, 1977.

[32] W.H. Reed and T.R. Hill: Triangular Mesh Methods for the Neutron Transport Equation. Technical Report LA-UR-73-479, Los Alamos Scientific Laboratory, 1973.

[33] M.F. Wheeler: An elliptic collocation-finite element method with interior penalties. SIAM J. Numer. Anal. 15, 1978, 152161.