OPTIMIZATION OF PLANE WAVE DIRECTIONS IN PLANE WAVE DISCONTINUOUS GALERKIN METHODS FOR THE HELMHOLTZ EQUATION

A Dissertation Presented to

the Faculty of the Department of Mathematics

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Akshay Agrawal

May 2016

OPTIMIZATION OF PLANE WAVE DIRECTIONS IN PLANE WAVE DISCONTINUOUS GALERKIN METHODS FOR THE HELMHOLTZ EQUATION

Akshay Agrawal

APPROVED:

Prof. Ronald H.W. Hoppe

Prof. Yuri A. Kuznetsov

Prof. Matthias Heinkenschloss

Assoc. Prof. Jingmei Qiu

Dean, College of Natural Sciences and Mathematics

Acknowledgements

This thesis work would not have been possible without the support and guidance from many individuals and I would like to thank them for their contributions. I want to start off by thanking my advisor Dr. Ronald Hoppe who has been a constant source of support over the last 4 years. He supported me financially and intellectually throughout my entire journey as a PhD candidate. His patience, and willingness to let me work at my own pace were essential to my success. His knowledge and insight into mathematics not only enabled me to complete this thesis but also helped me expand my mathematical horizons.

I would like to thank Dr. Matthias Heinkenschloss for proof reading my thesis and providing valuable feedback. I would also like to thank all the members of my defense committee for taking time out of their busy schedules to accommodate the schedule changes necessitated by the unpredictable Texas weather!

On a personal level, I would be remiss not to mention the huge role my father, Dr. Sanjeev Agrawal, has played in shaping me as a person and as a mathematician. He was the first Mathematical influence in my life and remains one of the biggest influences for me to this day. Since my childhood he nurtured the scientific curiosity and thought process, without which reaching this far in my academic career would have been impossible. I would also like to thank my mother Dolly Agrawal and sister Swati Agrawal for being a constant source of support through all these years. A very special thanks to my wife, Dr. Natasha Sharma, who has been a very valuable source of support and inspiration since the very beginning! She introduced me to the field of Finite Elements almost 4 years back and since

then her importance in all facets of my life has grown exponentially.

Finally I would like to thank all my friends here who have been my support structure throughout my graduate student life. All my seniors helped me settle in a foreign land - Pankaj, Anando, Aanchal, Ankita and Manisha. Nandini and Ananya were directly responsible for me not starving to death in my first year here! I want to thank Nishant, James, Aarti and others for always being there whenever I needed them and Parth for the countless FIFA sessions that helped me take my mind off gruelling programming sessions!

Without the support, encouragement, and help of all these people my journey through graduate school would have been a lot harder, if not impossible, and I want to express my sincerest gratitude to each and every one of them.

To my parents Dolly and Sanjeev Agrawal, my sister Swati Agrawal and my lovely wife Natasha Sharma

OPTIMIZATION OF PLANE WAVE DIRECTIONS IN PLANE WAVE DISCONTINUOUS GALERKIN METHODS FOR THE HELMHOLTZ EQUATION

An Abstract of a Dissertation Presented to
the Faculty of the Department of Mathematics
University of Houston

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

By
Akshay Agrawal
May 2016

Abstract

Recently, the use of special local test functions other than polynomials in Discontinuous Galerkin (DG) approaches has attracted a lot of attention and became known as DG-Trefftz methods. In particular, for the 2D Helmholtz equation, plane waves have been used in [13] to derive an Interior Penalty (IP) type Plane Wave DG (PWDG) method and to provide an a priori error analysis of its p-version with respect to equidistributed plane wave directions. However, the dependence on the distribution of the plane wave directions has not been studied. In this thesis, we study the dependence by formulating the choice of the directions as an optimal control problem with a tracking type objective functional and the variational formulation of the PWDG method as a constraint. The necessary optimality conditions are derived and numerically solved by a projected gradient method. Numerical results are given which illustrate the benefits of the approach.

Contents

List of Figures

CHAPTER 1

Introduction

Standard finite element discretizations of the Helmhlotz equation are inefficient at high frequencies. Due to numerical dispersion, the mesh must resolve the wavelength to increasing accuracy for large wavenumbers in order to prevent phase errors from building up over the domain and 'polluting' the Galerkin solution, see [5]. This effect is particularly problematic for low order methods.

For this reason, recently, the use of special local test functions other than polynomials in Discontinuous Galerkin (DG) approaches has attracted a lot of attention. These are known as the DG-Trefftz methods. In general a Trefftz method is a volume-oriented discretization scheme, for which all trial and test functions, when restricted to any element of a given mesh, are solutions of the PDE under consideration. For most of the Trefftz spaces used, continuity across interfaces separating mesh elements cannot be enforced strongly, as Trefftz functions are not as 'flexible' as piecewise polynomials. Therefore, often DG

formulations are used in conjunction with Trefftz spaces. A survey of Trefftz methods for Helmholtz equation can be found in [14].

In this thesis we concentrate on the Plane Wave DG (PWDG) method which uses plane wave basis functions, see [9, 10, 12, 13]. In [9] it was shown that the ultra weak variational formulation (UWVF) for Helmholtz equation, [3], is a special case of PWDG. In [3] O. Cessenat and B. Despres make the choice of plane wave directions as being uniformly distributed because in their experiments other choices did not lead to any significant improvements in the performance. Since then, the dependence of the performance of PWDG on the distribution of the plane wave directions has not been studied.

In this thesis we study this dependence by formulating the choice of the directions as an optimal control problem with a tracking type objective function and the variational formulation of the PWDG method as a constraint. We analyse the effects of optimally choosing the directions via two different examples.

This thesis is organised as follows: the second chapter gives an overview of the PWDG method used to numerically solve the Helmholtz problem.

In the third chapter, we introduce an objective functional and optimal control problem to study the dependence on the distribution of the plane wave directions. We derive the first order necessary optimality conditions for the stated optimal control problem. We then state the projected gradient method that is used to numerically solve the optimal control problem while detailing some of the calculations required therein.

In the fourth chapter, we present the numerical results for two different test problems which illustrate the benefits of the approach detailed in the third chapter.

Finally, our last chapter concludes the thesis with some possible future directions of research including further tests that can be done using our method.

CHAPTER 2

The Plane Wave Discontinuous Galerkin Method

For a bounded convex polygonal domain $\Omega \in \mathbb{R}^2$ with boundary $\Gamma = \partial\Omega$ we consider the Helmholtz equation

$$-\Delta u - \omega^2 u = 0 \qquad \text{in } \Omega, \tag{2.1a}$$

$$\boldsymbol{n} \cdot \nabla u + i\omega u = g \qquad \text{on } \Gamma = \partial\Omega. \tag{2.1b}$$

where $\omega > 0$ is the wavenumber, $g \in L^2(\Gamma)$ is a given function, and $\boldsymbol{n}$ denotes the exterior normal vector on Γ. We rewrite (2.1) as the first order system:

$$i\omega\boldsymbol{\sigma} - \nabla u = \mathbf{0} \qquad \text{in } \Omega, \tag{2.2a}$$

$$-\nabla \cdot \boldsymbol{\sigma} + i\omega u = 0 \qquad \text{in } \Omega, \tag{2.2b}$$

$$i\omega\boldsymbol{n} \cdot \boldsymbol{\sigma} + i\omega u = g \qquad \text{on } \Gamma. \tag{2.2c}$$

The variational formulation of (2.2) reads: Find $(\boldsymbol{\sigma}, u) \in \boldsymbol{H}(div, \Omega) \times H^1(\Omega)$ such that for all $(\boldsymbol{\tau}, v) \in \boldsymbol{H}(div, \Omega) \times H^1(\Omega)$ it holds

$$(i\omega\boldsymbol{\sigma}, \boldsymbol{\tau})_{0,\Omega} + (u, \nabla \cdot \boldsymbol{\tau})_{0,\Omega} = \langle u, \boldsymbol{n} \cdot \boldsymbol{\tau} \rangle_{H^{1/2}(\Gamma), H^{-1/2}(\Gamma)}, \tag{2.3a}$$

$$(\boldsymbol{\sigma}, \nabla v)_{0,\Omega} + (u, v)_{0,\Gamma} + (i\omega u, v)_{0,\Omega} = \left(\frac{1}{i\omega} g, v\right)_{0,\Gamma} \tag{2.3b}$$

where,

$$H^1(\Omega) = \left\{ f \in L^2(\Omega) \mid \partial_{x_i} f \in L^2(\Omega),\ i = 1, 2 \right\}$$

$$\boldsymbol{H}(div, \Omega) = \left\{ f \in L^2\left(\Omega; \mathbb{C}^2\right) \mid \operatorname{div}(f) \in L^2(\Omega) \right\}.$$

We consider a shape regular family of geometrically conforming, quasi-uniform simplicial triangulations $\mathcal{T}_h(\Omega)$ of the computational domain Ω. For $D \subset \overline{\Omega}$, we denote by $\mathcal{E}_h(D)$ the set of edges of the triangulation in D. For $T \in \mathcal{T}_h$, we refer to h_T as the diameter of T and set $h := \max\{h_T | T \in \mathcal{T}_h(\Omega)\}$. For $E \in \mathcal{E}_h\left(\overline{\Omega}\right)$, the length of E will be denoted by h_E. For functions $v \in \prod_{T \in \mathcal{T}_h(\Omega)} H^1(T)$ the trace of v on $E \in \mathcal{E}_h\left(\overline{\Omega}\right)$ may exhibit a jump across E. For $E \in \mathcal{E}_h\left(\overline{\Omega}\right)$ with $E = T_+ \cap T_-,\ T_\pm \in \mathcal{T}_h(\Omega)$ we define

$$\{v\}_E := \begin{cases} \frac{\left(v|_{T_+ \cap E} + v|_{T_- \cap E}\right)}{2} & , \quad E \in \mathcal{E}_h(\Omega) \\ v|_E & , \quad E \in \mathcal{E}_h(\Gamma) \end{cases}, \tag{2.4a}$$

$$[v]_E := \begin{cases} v|_{T_+ \cap E} - v|_{T_- \cap E} & , \quad E \in \mathcal{E}_h(\Omega) \\ v|_E & , \quad E \in \mathcal{E}_h(\Gamma) \end{cases}. \tag{2.4b}$$

For vector-valued functions we use an analogous notation.

We approximate (2.3a) and (2.3b) by introducing the following local spaces spanned by plane waves

$$\begin{aligned} V_p(T) &:= \left\{ v(\boldsymbol{x}) := \sum_{l=1}^{p} \alpha_l \exp(i\omega \boldsymbol{d}_l \cdot \boldsymbol{x}) \right\} \\ \boldsymbol{V}_p &:= V_p(T)^2 \end{aligned} \tag{2.5}$$

where $\alpha_l \in \mathbb{C}$ and $\boldsymbol{d}_l, 1 \leq l \leq p$, are p different unit directions

$$\boldsymbol{d}_l = (\cos(\theta_l), \sin(\theta_l))^T, \quad 1 \leq l \leq p = 2m + 1, \quad m \in \mathbb{N} \tag{2.6}$$

We set $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)^T$. Setting $\theta_{p+1} = \theta_1 + 2\pi$ we require that

$$\begin{aligned} \boldsymbol{\theta} \in \boldsymbol{K} &:= \left\{ \boldsymbol{\theta} \in [0, 2\pi)^p \mid \theta_{min} \leq \theta_{l+1} - \theta_l \leq \theta_{max},\ 1 \leq i \leq p \right\}, \\ \theta_{min} &:= \frac{2\pi\eta_1}{p}, \quad \theta_{max} := \frac{2\pi\eta_2}{p}, \quad 0 < \eta_1 < 1 < \eta_2 < 3/2. \end{aligned} \tag{2.7}$$

The associated global spaces are given by

$$\begin{aligned} V_h &:= \left\{ v_h \in L^2(\Omega) \mid v_h|_T \in V_p(T), T \in \mathcal{T}_h(\Omega) \right\}, \\ \boldsymbol{V}_h &:= \left\{ \boldsymbol{\tau}_h \in L^2(\Omega)^2 \mid \boldsymbol{\tau}_h|_T \in \boldsymbol{V}_p(T), T \in \mathcal{T}_h(\Omega) \right\}. \end{aligned} \tag{2.8}$$

Then,the PWDG approximation of (2.1a), (2.1b) amounts to computation of $(u_h, \boldsymbol{\sigma}_h) \in V_h \times \boldsymbol{V}_h$ such that for all $(v_h, \boldsymbol{\tau}_h) \in V_h \times \boldsymbol{V}_h$ it holds

$$\sum_{T \in \mathcal{T}_h(\Omega)} \left((i\omega \boldsymbol{\sigma}_h, \boldsymbol{\tau}_h)_{0,T} + (u_h, \nabla \cdot \boldsymbol{\tau}_h)_{0,T} \right) - \sum_{T \in \mathcal{T}_h(\Omega)} (\widehat{u}_h, \boldsymbol{n}_{\partial T} \cdot \boldsymbol{\tau}_h)_{0,\partial T} = 0, \tag{2.9a}$$

$$\sum_{T \in \mathcal{T}_h(\Omega)} \left((\boldsymbol{\sigma}_h, \nabla v_h)_{0,T} + (i\omega u_h, v_h)_{0,T} \right) - \sum_{T \in \mathcal{T}_h(\Omega)} (\boldsymbol{n}_{\partial T} \cdot \widehat{\boldsymbol{\sigma}}_h, v_h)_{0,\partial T} = 0. \tag{2.9b}$$

Here, the PWDG flux functions $\widehat{u}_h$ and $\widehat{\boldsymbol{\sigma}}_h$ are given by

$$\widehat{u}_h|_E := \begin{cases} \{u_h\}_E - \frac{\beta}{i\omega}[\nabla u_h]_E & , \quad E \in \mathcal{E}_h(\Omega) \\ u_h - \delta\left(\frac{1}{i\omega} \boldsymbol{n}_E \cdot \nabla u_h + u_h - \frac{1}{i\omega} g \right) & , \quad E \in \mathcal{E}_h(\Gamma) \end{cases}, \tag{2.10a}$$

$$\widehat{\boldsymbol{\sigma}}_h|_E := \begin{cases} \frac{1}{i\omega}\{\nabla u_h\}_E - \alpha[u_h]_E & , \quad E \in \mathcal{E}_h(\Omega) \\ \frac{1}{i\omega}\nabla u_h - (1-\delta)\left(\frac{1}{i\omega}\nabla u_h + \boldsymbol{n}_E u_h - \frac{1}{i\omega}\boldsymbol{n}_E g \right) & , \quad E \in \mathcal{E}_h(\Gamma) \end{cases}, \tag{2.10b}$$

where $\boldsymbol{n}_E$ is the exterior unit normal on E and $\alpha > 0,\ \beta > 0$ and $\delta \in (0,1)$ are flux parameters independent of $h,\ p$, and ω.

By choosing $\boldsymbol{\tau}_h = \nabla v_h$ in (2.9a), we can eliminate $\boldsymbol{\sigma}_h$ from (2.9a) and (2.9b),and obtain the following variational formulation of PWDG method:

Find $u_h \in V_h$ such that for all $v_h \in V_h$ it holds

$$\begin{aligned} &\sum_{T \in \mathcal{T}_h(\Omega)} \left((\nabla u_h, \nabla v_h)_{0,T} - \omega^2 (u_h, v_h)_{0,T} \right) - \\ &\sum_{T \in \mathcal{T}_h(\Omega)} \left((u_h - \widehat{u}_h, \boldsymbol{n}_{\partial T} \cdot \nabla v_h)_{0,\partial T} + i\omega (\boldsymbol{n}_{\partial T} \cdot \widehat{\boldsymbol{\sigma}}_h, v_h)_{0,\partial T} \right) = 0 \end{aligned} \tag{2.11}$$

Moreover, using Green's formula for the first term on the left-hand side in (2.11) and observing $\left(-\Delta - \omega^2 I\right) u_h|_T = 0,\ T \in \mathcal{T}_h(\Omega)$, we are led to a formulation of the PWDG method involving only integrals over the edges $E \in \mathcal{E}_h\left(\overline{\Omega}\right)$:

Find $u_h \in V_h$ such that

$$a_h(u_h, v_h) \ = \ l_h(v_h), \quad \forall\ v_h \in V_h, \tag{2.12}$$

where the sesquilinear form $a_h(\cdot,\cdot) : V_h \times V_h \to \mathbb{C}$ and the functional $l_h : V_h \to \mathbb{C}$ are given by

$$\begin{aligned} a_h(u_h, v_h) \ := \sum_{E\in\mathcal{E}_h(\Omega)} & \Big((\{u_h\}_E, \boldsymbol{n}_E \cdot [\nabla v_h]_E)_{0,E} + i\beta\omega^{-1} (\boldsymbol{n}_E \cdot [\nabla u_h]_e, \boldsymbol{n}_E \cdot [\nabla v_h]_E)_{0,E} \\ & - (\boldsymbol{n}_E \cdot \{\nabla u_h\}_E, [v_h]_E)_{0,E} + i\alpha\omega ([u_h]_E, [v_h]_E)_{0,E} \Big) + \\ \sum_{E\in\mathcal{E}_h(\Gamma)} & \Big((1-\delta)(u_h, \boldsymbol{n}_E \cdot \nabla v_h)_{0,E} + i\delta\omega^{-1} (\boldsymbol{n}_E \cdot \nabla u_h, \boldsymbol{n}_E \cdot \nabla v_h)_{0,E} \\ & - \delta (\boldsymbol{n}_E \cdot \nabla u_h, v_h)_{0,E} + i(1-\delta)\omega (u_h, v_h)_{0,E} \Big) \end{aligned} \tag{2.13a}$$

$$l_h(v_h) \ := \sum_{E\in\mathcal{E}_h(\Gamma)} \Big(i\delta\omega^{-1} (g, \boldsymbol{n}_E \cdot \nabla v_h)_{0,E} + (1-\delta)(g, v_h)_{0,E} \Big) \tag{2.13b}$$

As has been shown in [13], the variational equation (2.12) admits a unique solution $u_h \in V_h$. Moreover, if the solution u of (2.1a),(2.1b) satisfies $u \in H^{k+1}(\Omega),\ k \in \mathbb{N}$, and if the mesh width h of the triangulation $\mathcal{T}_h(\Omega)$ satisfies $\omega h \le \kappa$ for some $\kappa > 0$, then there exists a constant $C > 0$, independent of p and u, but depending on κ, such that the following a priori estimate holds true (cf. Theorem 3.14 in [13])

$$\|u - u_h\|_{0,\Omega} \le C\omega^{-1} \operatorname{diam}(\Omega)\, h^{k-1} \left(\frac{\log p}{p}\right)^{k-1/2} \|u\|_{k+1,\omega,\Omega} \quad , \tag{2.14}$$

where $\|\cdot\|_{k+1,\omega,\Omega}$ stands for the ω-weighted Sobolev norm

$$\|v\|_{k+1,\omega,\Omega} \ := \ \left(\sum_{j=0}^{k+1} \omega^{2(k+1-j)} |v|_{j,\Omega}^2 \right)^{1/2}, \quad v \in H^{k+1}(\Omega).$$

Setting $N := \mathrm{card}\left(\mathcal{T}_h\left(\Omega\right)\right)$ and $\boldsymbol{\theta} := \left(\theta_1, \cdots, \theta_p\right)^T$, the global PWDG space V_h is spanned by Np basis function

$$\begin{aligned} V_h &= \mathrm{span}\left(\phi_h^{(1)}, \cdots, \phi_h^{(Np)}\right), \\ \phi_h^{((k-1)p+l)} &:= \exp\left(i\omega\left(\cos\left(\theta_l\right), \sin\left(\theta_l\right)\right)^T \cdot \boldsymbol{x}\right)|_{T_k}, \quad 1 \le k \le N,\ 1 \le l \le p. \end{aligned} \tag{2.15}$$

Then, $u_h \in V_h$ can be written as

$$u_h = \sum_{j=1}^{Np} u_j \phi_h^{(j)}, \quad u_j \in \mathbb{C},\ 1 \le j \le Np. \tag{2.16}$$

Further, setting $\boldsymbol{y} := \left(y_1, \cdots, y_{Np}\right)^T \in \mathbb{C}^{Np}$ with $y_j := u_j,\ 1 \le j \le Np$, the PWDG approximation (2.12) represents a complex linear algebraic system

$$\boldsymbol{A}\left(\boldsymbol{\theta}\right)\boldsymbol{y} = \boldsymbol{b}\left(\boldsymbol{\theta}\right), \tag{2.17}$$

where the matrix $\boldsymbol{A}\left(\boldsymbol{\theta}\right) = \left(a_{kl}\left(\boldsymbol{\theta}\right)\right)_{k,l=1}^{Np} \in \mathbb{C}^{Np}$ and
the vector $\boldsymbol{b}\left(\boldsymbol{\theta}\right) = \left(b_1\left(\boldsymbol{\theta}\right), \cdots, \boldsymbol{b}_{Np}\left(\boldsymbol{\theta}\right)\right)^T \in \mathbb{C}^{Np}$ are given by

$$\begin{aligned} a_{kl}\left(\boldsymbol{\theta}\right) &:= a_h\left(\phi_h^{(l)}\left(\boldsymbol{\theta}\right), \phi_h^{(k)}\left(\boldsymbol{\theta}\right)\right), \qquad 1 \le k, l \le Np, \\ b_l\left(\boldsymbol{\theta}\right) &:= l_h\left(\phi_h^{(l)}\right), \qquad\qquad 1 \le l \le Np. \end{aligned} \tag{2.18}$$

CHAPTER 3

Optimization of Plane Wave Directions

The a priori estimate (2.14) for the L^2-norm of the global discretization error tells us how the error depends on the number of plane wave directions, p. It does not however provide any information on the appropriate choice of the directions $\boldsymbol{d}_l = (\cos(\theta_l), \sin(\theta_l))^T$, $1 \leq l \leq p$, except that they are supposed to satisfy assumption (2.7).

In fact, since

$$V_h = \operatorname{span}\left(\exp(i\omega \boldsymbol{d}_1 \cdot \boldsymbol{x})|_{T_1}, \cdots, \exp(i\omega \boldsymbol{d}_p \cdot \boldsymbol{x})|_{T_N}\right), \tag{3.1}$$

where $N := \operatorname{card}(\mathcal{T}_h(\Omega))$, the solution $u_h \in V_h$ of (2.12) depends on $\boldsymbol{\theta} := (\theta_1, \cdots, \theta_p)^T \in \boldsymbol{K}$ according to

$$u_h(\boldsymbol{\theta}) = \sum_{k=1}^{N} \sum_{l=1}^{p} u_{kl} \exp(i\omega \boldsymbol{d}_l \cdot \boldsymbol{x})|_{T_k}, \quad u_{kl} \in \mathbb{C} \tag{3.2}$$

We attempt to choose $\boldsymbol{\theta} \in \boldsymbol{K}$ such that with respect to the L^2-norm the solution $u_h(\boldsymbol{\theta})$ of (2.12) is as close as possible to a given desired state $u^d \in L^2(\Omega)$.

This can be formulated as the optimal control problem

$$\min_{u_h \in V_h,\ \boldsymbol{\theta} \in \boldsymbol{K}} J\left(u_h, \boldsymbol{\theta}\right) \ := \ \frac{1}{2} \| u_h\left(\boldsymbol{\theta}\right) - u^d \|_{0,\Omega}^2, \tag{3.3a}$$

subject to the PWDG constraint

$$a_h\left(u_h\left(\theta\right), v_h\left(\theta\right)\right) \ = \ l_h\left(v_h\left(\theta\right)\right), \quad v_h\left(\theta\right) \in V_h. \tag{3.3b}$$

Introducing the Hermitian Matrix $\boldsymbol{M}\left(\boldsymbol{\theta}\right) \ = \ \left(m_{kl}\left(\boldsymbol{\theta}\right)\right)_{k,l=1}^{Np} \in \mathbb{C}^{Np \times Np}$ and the vector $\boldsymbol{c}\left(\boldsymbol{\theta}\right) \ = \ \left(c_1\left(\boldsymbol{\theta}\right), \cdots, c_{Np}\left(\boldsymbol{\theta}\right)\right)^T$ according to

$$\begin{aligned} m_{kl}\left(\boldsymbol{\theta}\right) \ &:= \ \left(\phi_h^{(k)}, \phi_h^{(l)}\right)_{0,\Omega} \quad , \quad 1 \leq k, l \leq Np, \\ c_l\left(\boldsymbol{\theta}\right) \ &:= \ \left(u^d, \phi_h^{(l)}\right)_{0,\Omega} \quad , \quad 1 \leq l \leq Np, \end{aligned} \tag{3.4}$$

the algebraic formulation of (3.3a)-(3.3b) turns out to be

$$\min_{y \in \mathbb{C}^{Np},\ \boldsymbol{\theta} \in \boldsymbol{K}} J\left(\boldsymbol{y}, \boldsymbol{\theta}\right) \ := \ \frac{1}{2} \left\langle \boldsymbol{M}\left(\boldsymbol{\theta}\right) \boldsymbol{y}, \boldsymbol{y} \right\rangle - \mathrm{Re}\left(\left\langle \boldsymbol{c}\left(\boldsymbol{\theta}\right), \boldsymbol{y} \right\rangle\right) + \frac{1}{2} \left(u^d, u^d\right)_{0,\Omega}^2, \tag{3.5a}$$

subject to the state equation

$$e\left(\boldsymbol{y}, \boldsymbol{\theta}\right) \ := \ \boldsymbol{A}\left(\boldsymbol{\theta}\right) \boldsymbol{y} - \boldsymbol{b}\left(\boldsymbol{\theta}\right) \ = \ 0. \tag{3.5b}$$

We further denote by $\boldsymbol{G} \ : \ \boldsymbol{K} \to \mathbb{C}^{Np}$ the control-to-state map which assigns to the control $\boldsymbol{\theta} \in \boldsymbol{K}$ the unique solution $\boldsymbol{y} \in \mathbb{C}^{Np}$ of the state equation (3.5b) and by $J_{red} \ : \ \boldsymbol{K} \to \mathbb{R}$ the reduced objective functional

$$J_{red}\left(\boldsymbol{\theta}\right) \ := \ J\left(\boldsymbol{G}\left(\boldsymbol{\theta}\right), \boldsymbol{\theta}\right).$$

Then, the control-reduced formulation of the optimal control problem (3.5a)-(3.5b) reads as follows

$$\min_{\boldsymbol{\theta} \in \boldsymbol{K}} J_{red}\left(\boldsymbol{\theta}\right). \tag{3.6}$$

Theorem 3.1. *The optimal control problem (3.5a)-(3.5b) admits an optimal solution* $(\boldsymbol{y}^*, \boldsymbol{\theta}^*) \in \mathbb{C}^{Np} \times \boldsymbol{K}$.

Proof. Let $\left\{\boldsymbol{\theta}^{(n)}\right\}_{\mathbb{N}}$, $\boldsymbol{\theta}^{(n)} \in \boldsymbol{K}$, $n \in \mathbb{N}$, be a minimizing sequence, i.e.,it holds

$$J_{red}\left(\boldsymbol{\theta}^{(n)}\right) \to \min_{\boldsymbol{\theta} \in \boldsymbol{K}} J_{red}\left(\boldsymbol{\theta}\right) \text{ as } n \to \infty \tag{3.7}$$

Obviously, the sequence $\left\{\boldsymbol{\theta}^{(n)}\right\}_{\mathbb{N}}$ is bounded and hence, there exists a subsequence $\mathbb{N}' \subset \mathbb{N}$ and $\boldsymbol{\theta}^* \in \mathbb{R}^p$ such that

$$\boldsymbol{\theta}^{(n)} \to \boldsymbol{\theta}^*, \quad \mathbb{N}' \ni n \to \infty.$$

In view of the closedness of $\boldsymbol{K}$, we have $\boldsymbol{\theta}^* \in \boldsymbol{K}$. Moreover, due to the continuity of both the control-to-state map $\boldsymbol{G}$ and of the reduced objective functional J_{red} we deduce

$$\boldsymbol{G}\left(\boldsymbol{\theta}^{(n)}\right) \to \boldsymbol{G}\left(\boldsymbol{\theta}^*\right), \; J_{red}\left(\boldsymbol{\theta}^{(n)}\right) \to J_{red}\left(\boldsymbol{\theta}^*\right) \text{ as } \mathbb{N}' \ni n \to \infty.$$

Consequently, from (3.7) we have

$$J_{red}\left(\boldsymbol{\theta}^*\right) \; = \; \min_{\boldsymbol{\theta} \in \boldsymbol{K}} J_{red}\left(\boldsymbol{\theta}\right),$$

and with $\boldsymbol{y}^* \; := \; \boldsymbol{G}\left(\boldsymbol{\theta}^*\right)$ it follows that the pair $(\boldsymbol{y}^*, \boldsymbol{\theta}^*) \in \mathbb{C}^{Np} \times \boldsymbol{K}$ is an optimal solution of (3.5a)-(3.5b). □

Remark 3.2. *Since the control-to-state map* $\boldsymbol{G}$ *is a non-convex function of the control* $\boldsymbol{\theta}$*, we do not have uniqueness of the solution.*

3.1 First-Order Necessary Optimality Conditions

We will derive the first-order necessary optimality conditions for the optimal control problem (3.5a)-(3.5b) by the method of Lagrange multipliers which is justified if the linear independence constraint qualification holds true.

To this end, we note that the bound constraints on the control can be expressed as the inequalities $\boldsymbol{G}(\boldsymbol{\theta}) \leq \mathbf{0}$, where the mapping $\boldsymbol{g} = (g_1, g_2) \ : \ \mathbb{R}^p \to \mathbb{R}^p \times \mathbb{R}^p$ is defined by the means of

$$\begin{aligned} \boldsymbol{g_1}(\boldsymbol{\theta}) \ &:= \ (\theta_2 - \theta_1 - \theta_{max}, \cdots, \theta_{p+1} - \theta_p - \theta_{max}), \\ \boldsymbol{g_2}(\boldsymbol{\theta}) \ &:= \ (\theta_{min} - (\theta_2 - \theta_1), \cdots, \theta_{min} - (\theta_{p+1} - \theta_p)). \end{aligned} \tag{3.8}$$

For a local minimum $(\boldsymbol{y^*}, \boldsymbol{\theta^*}) \in \mathbb{C}^{Np} \times \boldsymbol{K}$ of (3.5a)-(3.5b), the active set is given by $A(\boldsymbol{\theta^*}) = A_1(\boldsymbol{\theta^*}) \cup A_2(\boldsymbol{\theta^*})$ where

$$A_1(\boldsymbol{\theta^*}) \ := \ \left\{q \in \{1, \cdots, p\} \mid \theta^*_{q+1} - \theta^*_q - \theta_{max} = 0\right\}, \tag{3.9a}$$

$$A_2(\boldsymbol{\theta^*}) \ := \ \left\{q \in \{1, \cdots, p\} \mid \theta_{min} - \left(\theta^*_{q+1} - \theta^*_q\right) = 0\right\} \tag{3.9b}$$

We refer to $I(\boldsymbol{\theta^*}) := \{1, \cdots, p\} \setminus A(\boldsymbol{\theta^*})$ as the inactive set. The linear independence constraint qualification requires linearization of $\left(e, (\boldsymbol{g_1})_{A_1(\boldsymbol{\theta^*})}, (\boldsymbol{g_2})_{A_2(\boldsymbol{\theta^*})}\right)$ at $(\boldsymbol{y^*}, \boldsymbol{\theta^*})$ to be surjective.

Theorem 3.3. *Let* $p_i^* \ := \ \mathrm{card}\left(A_i(\boldsymbol{\theta^*})\right), \ 1 \leq i \leq 2$. *The mapping*

$$\left(\nabla e(\boldsymbol{y^*}, \boldsymbol{\theta^*}), \nabla \boldsymbol{g}_{\mathbf{1}, A_1(\boldsymbol{\theta^*})}(\boldsymbol{\theta^*}), \nabla \boldsymbol{g}_{\mathbf{2}, A_2(\boldsymbol{\theta^*})}(\boldsymbol{\theta^*})\right) \ : \ \mathbb{C}^{Np} \times \mathbb{R}^p \to \mathbb{C}^{Np} \times \mathbb{R}^{p_1^*} \times \mathbb{R}^{p_2^*}$$

is surjective. In particular, for any $(\boldsymbol{r}, \boldsymbol{s}_1, \boldsymbol{s}_2) \in \mathbb{C}^{Np} \times \mathbb{R}^{p_1^*} \times \mathbb{R}^{p_2^*}$ *there exists a unique solution* $(\delta\boldsymbol{y}, \delta\boldsymbol{\theta}) \in \mathbb{C}^{Np} \times \mathbb{R}^p$ *of the equation*

$$\left(\nabla e(\boldsymbol{y^*}, \boldsymbol{\theta^*})(\delta\boldsymbol{y}, \delta\boldsymbol{\theta}), \nabla \boldsymbol{g}_{\mathbf{1}, A_1(\boldsymbol{\theta^*})}(\boldsymbol{\theta^*})\, \delta\boldsymbol{\theta}, \nabla \boldsymbol{g}_{\mathbf{2}, A_2(\boldsymbol{\theta^*})}(\boldsymbol{\theta^*})\, \delta\boldsymbol{\theta}\right) = (\boldsymbol{r}, \boldsymbol{s}_1, \boldsymbol{s}_2).$$

Proof. For $k \in A_1(\boldsymbol{\theta^*})$

$$\nabla g_{1,k'} = \begin{cases} -1 & , \quad k' = k \\ +1 & , \quad k' = k+1 \\ 0 & , \quad \text{otherwise} \end{cases}, \tag{3.10}$$

whereas for $k \in A_2(\boldsymbol{\theta}^*)$

$$\nabla g_{1,k'} = \begin{cases} +1 & , \quad k' = k \\ -1 & , \quad k' = k+1 \\ 0 & , \quad \text{otherwise} \end{cases} . \tag{3.11}$$

Since $I(\boldsymbol{\theta}^*) \neq \emptyset$, there exists $q \in \{1, \cdots, p\}$ such that $q \in I(\boldsymbol{\theta}^*)$. We renumber the controls according to $\widehat{\theta_k^*} := \theta_{q+k-1}^*$, $\widehat{\theta_{k+p}^*} := \widehat{\theta_k^*} + 2\pi$, $1 \leq k \leq p$, and set $(\delta\boldsymbol{\theta})_k = 0$ for $k \in I\left(\widehat{\boldsymbol{\theta}^*}\right)$. If $A\left(\widehat{\boldsymbol{\theta}^*}\right) = \emptyset$, there is nothing to show. If $A\left(\widehat{\boldsymbol{\theta}^*}\right) \neq \emptyset$, there exists

$$k_{min} := \min\left\{k \in \{2, \cdots, p\} | k \in A\left(\widehat{\boldsymbol{\theta}^*}\right)\right\}.$$

Moreover, in view of $p+1 \in I\left(\widehat{\boldsymbol{\theta}^*}\right)$, there also exists

$$k_{max} := \max\left\{k \in \{k_{min}+1, \cdots, p+1\} | k \in I\left(\widehat{\boldsymbol{\theta}^*}\right)\right\}.$$

In view of (3.10), (3.11), $(\delta\boldsymbol{\theta})_k$, $k_{min} \leq k \leq k_{max} - 1$, is the unique solution of the linear algebraic system with a regular upper triangular matrix. For the computation of $(\delta\boldsymbol{\theta})_k \in A\left(\widehat{\boldsymbol{\theta}^*}\right) \setminus \{k_{min}, \cdots, k_{max} - 1\}$ we proceed in the same way. On the other hand, the equation $\nabla e(\boldsymbol{y}^*, \boldsymbol{\theta}^*)(\delta\boldsymbol{y}, \delta\boldsymbol{\theta}) = \boldsymbol{r}$ can be equivalently written as

$$\boldsymbol{A}(\boldsymbol{\theta})\,\delta\boldsymbol{y} = \nabla_\theta\left(\boldsymbol{b}(\boldsymbol{\theta}^*) - \boldsymbol{A}(\boldsymbol{\theta}^*)\,\boldsymbol{y}^*\right)\delta\boldsymbol{\theta},$$

which has a unique solution $\delta\boldsymbol{y} \in \mathbb{C}^{Np}$. □

Due to Theorem 3.3, the necessary optimality conditions can be derived by the method of Lagrange multipliers.

Theorem 3.4. *Assume that* $(\boldsymbol{y}^*, \boldsymbol{\theta}^*) \in \mathbb{C}^{Np} \times \boldsymbol{K}$ *is an optimal solution of (3.5a)-(3.5b). Then,there exist an adjoint state* $\boldsymbol{p}^* \in \mathbb{C}^{Np}$ *and a multiplier* $\boldsymbol{\mu}^* = \left(\boldsymbol{\mu_1^*}, \boldsymbol{\mu_2^*}\right) \in \mathbb{R}_+^{2p}$, $\boldsymbol{\mu_i^*} = \left(\mu_{i,1}^*, \cdots, \mu_{i,p}^*\right)^T$, $1 \leq i \leq 2$, *such that the state equation, the adjoint state equation and the gradient equation*

$$\boldsymbol{A}\left(\boldsymbol{\theta}^*\right)\boldsymbol{y}^* - \boldsymbol{b}\left(\boldsymbol{\theta}^*\right) = 0,$$

$$\boldsymbol{A}^H\left(\boldsymbol{\theta}^*\right)\boldsymbol{p}^* + \boldsymbol{M}\left(\boldsymbol{\theta}^*\right)\boldsymbol{y}^* - \mathrm{Re}\left(\boldsymbol{c}\left(\boldsymbol{\theta}^*\right)\right) = 0,$$

$$\nabla_\theta J\left(\boldsymbol{y}^*, \boldsymbol{\theta}^*\right) + \mathrm{Re}\left(\left\langle \nabla_\theta\left(\boldsymbol{A}\left(\boldsymbol{\theta}^*\right)\boldsymbol{y}^* - \boldsymbol{b}\left(\boldsymbol{\theta}^*\right)\right), \boldsymbol{p}^*\right\rangle\right) + \nabla_\theta \boldsymbol{g_1}\left(\boldsymbol{\theta}^*\right)^T \boldsymbol{\mu_1^*} + \nabla_\theta \boldsymbol{g_2}\left(\boldsymbol{\theta}^*\right)^T \boldsymbol{\mu_2^*} = 0$$

are satisfied as well as the complementary conditions

$$g_{i,q}\left(\boldsymbol{\theta}^*\right) \leq 0,\ \mu_{i,q}^* \geq 0,\ g_{i,q}\left(\boldsymbol{\theta}^*\right)\mu_{i,q}^* = 0, \quad 1 \leq q \leq p,\ 1 \leq i \leq 2.$$

Proof. We introduce the Lagrangian $L : \mathbb{C}^{Np} \times \mathbb{R}^p \times \mathbb{C}^{Np} \times \mathbb{R}_+^{2p}$ according to

$$L\left(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{\mu}\right) := J\left(\boldsymbol{y}, \boldsymbol{\theta}\right) + \mathrm{Re}\left(\left\langle e\left(\boldsymbol{y}, \boldsymbol{\theta}\right), \boldsymbol{p}\right\rangle\right) + \boldsymbol{g_1}\left(\boldsymbol{\theta}\right)^T \boldsymbol{\mu_1} + \boldsymbol{g_2}\left(\boldsymbol{\theta}\right)^T \boldsymbol{\mu_2}.$$

Setting $\boldsymbol{x} := (\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{p})$ and $\boldsymbol{x}^* := (\boldsymbol{y}^*, \boldsymbol{\theta}^*, \boldsymbol{p}^*)$, the first order necessary optimal conditions are given by

$$\frac{\partial L}{\partial \boldsymbol{y}}\left(\boldsymbol{x}^*, \boldsymbol{\mu}^*\right) = 0, \quad \frac{\partial L}{\partial \boldsymbol{\theta}}\left(\boldsymbol{x}^*, \boldsymbol{\mu}^*\right) = 0, \quad \frac{\partial L}{\partial \boldsymbol{p}}\left(\boldsymbol{x}^*, \boldsymbol{\mu}^*\right) = 0, \tag{3.12a}$$

$$\frac{\partial L}{\partial \boldsymbol{\mu_i}}\left(\boldsymbol{x}^*, \boldsymbol{\mu}^*\right)^T \left(\boldsymbol{\nu_i} - \boldsymbol{\mu_i^*}\right) \leq 0, \quad \boldsymbol{\nu_i} \in \mathbb{R}_+^p,\ 1 \leq i \leq 2. \tag{3.12b}$$

The state equation, the adjoint state equation, and the gradient equation result from the third, first and second equation in (3.12a), whereas the complimentary conditions are a consequence of (3.12b) □

3.2 Projected Gradient Method

The projected gradient method is based on the formulation of the gradient equation as the variational inequality

$$-\nabla_\theta J\left(\boldsymbol{y}^*, \boldsymbol{\theta}^*\right) + \operatorname{Re}\left(\left\langle \nabla_\theta \left(\boldsymbol{b}^*\left(\boldsymbol{\theta}^*\right) - \boldsymbol{A}\left(\boldsymbol{\theta}^*\right)\boldsymbol{y}^*\right), \boldsymbol{p}^*\right\rangle\right) \in \partial I_{\boldsymbol{K}},$$

where $\partial I_{\boldsymbol{K}}$ is the indicator function of the constrained set $\boldsymbol{K}$.

The algorithm for the Projected Gradient Method is as follows:

Step 1: Choose an initial control $\boldsymbol{\theta}^{(0)} \in \boldsymbol{K}$ and a tolerance $TOL > 0$ and set $n = 0$

Step 2.1: Set $n = n + 1$ and compute $\boldsymbol{y}^{(n)} \in \mathbb{C}^{Np}$ and $\boldsymbol{p}^{(n)} \in \mathbb{C}^{Np}$ as the unique solutions of the state equation

$$\boldsymbol{A}\left(\boldsymbol{\theta}^{(n-1)}\right)\boldsymbol{y}^{(n)} = \boldsymbol{b}\left(\boldsymbol{\theta}^{(n-1)}\right)$$

and of adjoint state equation

$$\boldsymbol{A}^H\left(\boldsymbol{\theta}^{(n-1)}\right)\boldsymbol{p}^{(n)} = \operatorname{Re}\left(\boldsymbol{c}\left(\boldsymbol{\theta}^{(n-1)}\right)\right) - \boldsymbol{M}\left(\boldsymbol{\theta}^{(n-1)}\right)\boldsymbol{y}^{(n)}.$$

Step 2.2: Compute $\tilde{\boldsymbol{\theta}}^{(n)} \in \mathbb{R}^p$ according to

$$\tilde{\boldsymbol{\theta}}^{(n)} = \boldsymbol{\theta}^{(n-1)} - \kappa\left(\nabla_\theta J\left(\boldsymbol{y}^{(n)}, \boldsymbol{\theta}^{(n-1)}\right) + \nabla_\theta \operatorname{Re}\left(\left\langle \boldsymbol{A}\left(\boldsymbol{\theta}^{(n-1)}\right)\boldsymbol{y}^{(n)} - \boldsymbol{b}\left(\boldsymbol{\theta}^{(n-1)}\right), \boldsymbol{p}^{(n)}\right\rangle\right)\right),$$

where $\kappa > 0$ is Armijo line search parameter.

Step 2.3: Computer $\boldsymbol{\theta}^{(n)}$ as the projection of $\tilde{\boldsymbol{\theta}}^{(n)}$ onto the constraint set $\boldsymbol{K}$.

Step 2.4: If $n > 1$ and

$$\left|J\left(\boldsymbol{y}^{(n)}, \boldsymbol{\theta}^{(n)}\right) - J\left(\boldsymbol{y}^{(n-1)}, \boldsymbol{\theta}^{(n-1)}\right)\right| < TOL,$$

stop the algorithm.Otherwise, go to Step 2.1.

In the following section we present the calculations needed for the implementation of this method.

3.3 Some Important Calculations

Consider equation is Step 2.2 from Section 3.2. For the update formula we need to calculate the following quantity:

$$\nabla_\theta J\left(\boldsymbol{y},\boldsymbol{\theta}\right)+\nabla_\theta \operatorname{Re}\left(\left\langle \boldsymbol{A}\left(\boldsymbol{\theta}\right)\boldsymbol{y}-\boldsymbol{b}\left(\boldsymbol{\theta}\right),\boldsymbol{p}\right\rangle\right)$$

First we will calculate $\nabla_\theta J\left(\boldsymbol{y},\boldsymbol{\theta}\right)$. Here, from (3.5b) we know that $\boldsymbol{y}$ is the unique solution to

$$\boldsymbol{A}\left(\boldsymbol{\theta}\right)\boldsymbol{y}=\boldsymbol{b}\left(\boldsymbol{\theta}\right)$$

Also, from (3.4) and (3.5a) we know that

$$J\left(\boldsymbol{y},\boldsymbol{\theta}\right) \;=\; \frac{1}{2}\left\langle \boldsymbol{M}\left(\boldsymbol{\theta}\right)\boldsymbol{y},\boldsymbol{y}\right\rangle-\operatorname{Re}\left(\left\langle \boldsymbol{c}\left(\boldsymbol{\theta}\right),\boldsymbol{y}\right\rangle\right)+\frac{1}{2}\left(u^d,u^d\right)_{0,\Omega}^2 \tag{3.13}$$

where,

$$\begin{aligned} m_{kl}\left(\boldsymbol{\theta}\right) &:= \left(\phi_h^{(k)},\phi_h^{(l)}\right)_{0,\Omega} \quad , \quad 1\le k,l\le Np, \\ c_l\left(\boldsymbol{\theta}\right) &:= \left(u^d,\phi_h^{(l)}\right)_{0,\Omega} \quad , \quad 1\le l\le Np, \end{aligned} \tag{3.14}$$

and N is the total number of triangles in our triangulation $\mathcal{T}$ and p is the number of plane wave basis functions used. Let $y=\{\alpha_j\}_{j=1}^{Np}$.

Note that for any two given basis functions $\phi_h^{(k)}$ and $\phi_h^{(l)}$ either,

$$\mu\left(\operatorname{supp}\left(\phi_h^{(k)}\right)\cap\operatorname{supp}\left(\phi_h^{(l)}\right)\right)=0$$

or,

$$\operatorname{supp}\left(\phi_h^{(k)}\right)\cap\operatorname{supp}\left(\phi_h^{(l)}\right)=T\in\mathcal{T},$$

where μ is the 2-D Lebesgue measure.

Let $T_{k,l}$ be defined as

$$T_{k,l} \;:=\; \begin{cases} \emptyset & , \quad \text{if } \mu\left(\operatorname{supp}\left(\phi_h^{(k)}\right)\cap\operatorname{supp}\left(\phi_h^{(l)}\right)\right)=0 \\ \operatorname{supp}\left(\phi_h^{(k)}\right)\cap\operatorname{supp}\left(\phi_h^{(l)}\right) & , \quad \text{otherwise} \end{cases} \tag{3.15}$$

and set $T_l := \operatorname{supp}\left(\phi_h^{(l)}\right) \in \mathcal{T}$. Using this we can rewrite (3.14) as

$$\begin{aligned} m_{kl}(\boldsymbol{\theta}) &:= \int_{T_{k,l}} \exp(i\omega \boldsymbol{d}_k \cdot \boldsymbol{x})\, \overline{\exp(i\omega \boldsymbol{d}_l \cdot \boldsymbol{x})} d\boldsymbol{x} \quad , \quad 1 \le k, l \le Np, \\ c_l(\boldsymbol{\theta}) &:= \int_{T_l} u_d \overline{\exp(i\omega \boldsymbol{d}_l \cdot \boldsymbol{x})} d\boldsymbol{x} \qquad\qquad , \quad 1 \le l \le Np, \end{aligned} \tag{3.16}$$

where $\boldsymbol{d}_k = [\cos(\theta_k), \sin(\theta_k)]^T$.

By (3.13) we can see that

$$\nabla_\theta J(\boldsymbol{y}, \boldsymbol{\theta}) = \nabla_\theta \left(\frac{1}{2} \sum_{k,l=1}^{Np} m_{kl}(\boldsymbol{\theta}) \alpha_l \overline{\alpha_k} \right) - \nabla_\theta \left(\operatorname{Re} \sum_{k=1}^{Np} c_k(\boldsymbol{\theta}) \overline{\alpha_k} \right) \tag{3.17}$$

By differentiating equations in (3.16) with respect to θ_j we get

$$\begin{aligned} \frac{\partial}{\partial \theta_j} \left(\frac{1}{2} \sum_{k,l=1}^{Np} m_{kl}(\boldsymbol{\theta}) \alpha_l \overline{\alpha_k} \right) =& \frac{1}{2} \left(\sum_{l=1}^{8p} \alpha_j \overline{\alpha_l} \int_{T_{j,l}} \left(i\omega \boldsymbol{d}_j^* \cdot \boldsymbol{x} \right) exp(i\omega \boldsymbol{d}_j \cdot \boldsymbol{x}) \cdot \overline{exp(i\omega \boldsymbol{d}_l \cdot \boldsymbol{x})} d\boldsymbol{x} \right. \\ & \left. + \sum_{l=1}^{8p} \alpha_l \overline{\alpha_j} \int_{T_{j,l}} \left(-i\omega \boldsymbol{d}_j^* \cdot \boldsymbol{x} \right) exp(i\omega \boldsymbol{d}_l \cdot \boldsymbol{x}) \cdot \overline{exp(i\omega \boldsymbol{d}_j \cdot \boldsymbol{x})} d\boldsymbol{x} \right) \\ =& \frac{1}{2} \left(\sum_{l=1}^{8p} \alpha_j \overline{\alpha_l} \int_{T_{j,l}} \left(i\omega \boldsymbol{d}_j^* \cdot \boldsymbol{x} \right) exp(i\omega \boldsymbol{d}_j \cdot \boldsymbol{x}) \cdot \overline{exp(i\omega \boldsymbol{d}_l \cdot \boldsymbol{x})} d\boldsymbol{x} \right. \\ & \left. + \sum_{l=1}^{8p} \overline{\alpha_j \overline{\alpha_l} \int_{T_{j,l}} \left(i\omega \boldsymbol{d}_j^* \cdot \boldsymbol{x} \right) exp(i\omega \boldsymbol{d}_j \cdot \boldsymbol{x}) \cdot \overline{exp(i\omega \boldsymbol{d}_l \cdot \boldsymbol{x})} d\boldsymbol{x}} \right) \\ =& \operatorname{Re} \sum_{l=1}^{8p} \alpha_j \overline{\alpha_l} \int_{T_{j,l}} \left(i\omega \boldsymbol{d}_j^* \cdot \boldsymbol{x} \right) exp(i\omega \boldsymbol{d}_j \cdot \boldsymbol{x}) \cdot \overline{exp(i\omega \boldsymbol{d}_l \cdot \boldsymbol{x})} d\boldsymbol{x} \ , \end{aligned} \tag{3.18}$$

where $\boldsymbol{d^*}_j = [-\sin(\theta_j), \cos(\theta_j)]^T$ and

$$\begin{aligned}\frac{\partial}{\partial\theta_j}\left(\mathrm{Re}\sum_{k=1}^{8p} c_k(\boldsymbol{\theta})\overline{\alpha_k}\right) &= \mathrm{Re}\left(\frac{\partial}{\partial\theta_j}\sum_{k=1}^{8p} c_k(\boldsymbol{\theta})\overline{\alpha_k}\right)\\ &= \mathrm{Re}\left(\frac{\partial c_j(\boldsymbol{\theta})}{\partial\theta}\overline{\alpha_j}\right)\\ &= \mathrm{Re}\left(\overline{\alpha_j}\int_{\Omega}\left(-i\omega\boldsymbol{d}_j^*\cdot\boldsymbol{x}\right)u_d\overline{exp(i\omega\boldsymbol{d}_j\cdot\boldsymbol{x})}d\boldsymbol{x}\right)\end{aligned} \tag{3.19}$$

Now for $\nabla_\theta \mathrm{Re}\left(\langle \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{y} - \boldsymbol{b}(\boldsymbol{\theta}), \boldsymbol{p}\rangle\right)$ we have,

$$\begin{aligned}\frac{\partial}{\partial\theta_j}\mathrm{Re}\left(\langle \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{y} - \boldsymbol{b}(\boldsymbol{\theta}), \boldsymbol{p}\rangle\right) &= \mathrm{Re}\left(\frac{\partial}{\partial\theta_j}\langle \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{y} - \boldsymbol{b}(\boldsymbol{\theta}), \boldsymbol{p}\rangle\right)\\ &= \mathrm{Re}\left(\frac{\partial}{\partial\theta_j}\sum_{k=1}^{Np} p_k\overline{\left(\sum_{l=1}^{Np} a_{kl}(\boldsymbol{\theta})\alpha_l - b_k(\boldsymbol{\theta})\right)}\right)\\ &= \mathrm{Re}\left(\sum_{k=1}^{Np} p_k\overline{\left(\sum_{l=1}^{Np}\frac{\partial a_{kl}(\boldsymbol{\theta})}{\partial\theta_j}\alpha_l - \frac{\partial b_k(\boldsymbol{\theta})}{\partial\theta_j}\right)}\right)\end{aligned} \tag{3.20}$$

We can obtain formulas for $\frac{\partial a_{kl}(\boldsymbol{\theta})}{\partial\theta_j}$ and $\frac{\partial b_k(\boldsymbol{\theta})}{\partial\theta_j}$ by directly differentiating formulas in (2.18) using (2.13a),(2.13b).

Using (3.18)-(3.20) we can calculate the required terms for the update formula in Step 2.2 from Section 3.2.

CHAPTER 4

Numerical Results

This chapter is devoted to documentation of the numerical results that illustrate the effect of choosing the plane wave directions for PWDG method optimally.

We will consider two variants of the problem. First we will look at the Helmholtz equation in the convex domain $\Omega := (0,1) \times (-0.5, 0,5)$. For this domain we will consider two cases. In the first case the solution is continuous. In the second case the solution has a singularity. For the second case we consider the non-convex domain $\Omega := (-1,1)^2 \setminus (S_1 \cup S_2)$ where

$$S_1 = \text{conv}\left((0,0), (-0.25, +0.50), (-0.50, +0.50)\right)$$

$$S_2 = \text{conv}\left((0,0), (+0.25, -0.50), (+0.50, -0.50)\right)$$

4.1 Test Problem on Convex Domain

We consider a square domain $\Omega = (0,1) \times (-0.50, +0.50)$, partitioned by a mesh consisting of 8 triangles (see Figure 4.1, upper-left plot), so that $h = 1/\sqrt{2}$. We fix $\omega = 10$, such that the entire wavelength $\lambda = 2\pi/\omega \approx 0.628$ is completely contained in Ω.

We choose the inhomogeneous boundary condition (g in 2.1b) in such a way that the analytical solutions are the circular waves given, in polar coordinates $\boldsymbol{x} = (r\cos\varphi, r\sin\varphi)$ by

$$u(\boldsymbol{x}) = J_\xi(\omega r)\cos(\xi\varphi), \qquad \xi \geq 0;$$

here, J_ξ denotes the Bessel function of the first kind and order ξ. For $t \ll 1$, these functions behave like

$$J_\xi(t) = \frac{1}{\Gamma(\xi+1)}\left(\frac{t}{2}\right)^\xi$$

Thus, if $\xi \in \mathbb{N}$, u can be analytically extended to a Helmholtz solution in $\mathbb{R}^2$, while, if $\xi \notin \mathbb{N}$, its derivatives have a singularity at the origin. Then $u \in H^{\xi+1-\epsilon}(\Omega)$ for every $\epsilon > 0$, but $u \notin H^{\xi+1}(\Omega)$.

We consider the regular case $\xi = 1$ and singular cases $\xi = 2/3$ and $\xi = 3/2$. The profiles of the analytical solutions corresponding to these three cases are displayed in Figure 4.1, upper-right and lower plots.

We consider two choices of numerical fluxes: with constant parameters ($\alpha = \beta = \delta = 1/2$), or depending on p, h and ω ($\alpha = \beta^{-1} = \delta^{-1} = a_0 p/(\omega h \log p)$ with $a_0 = 10$. We consider $p = 3, 5, \cdots, 27$.

We also need to choose the starting control $\boldsymbol{\theta_0} = (\theta_1, \theta_2, \cdots, \theta_p)^T$. We consider two different choices: uniform distribution ($\theta_i = 2\pi(i-1)/p,\ 1 \leq i \leq p$), and random distribution, where each $\theta_i, 1 \leq i \leq p$ is chosen randomly from $[0, 2\pi)$.

Plots comparing the L^2 errors between the computed solutions and the analytical solutions

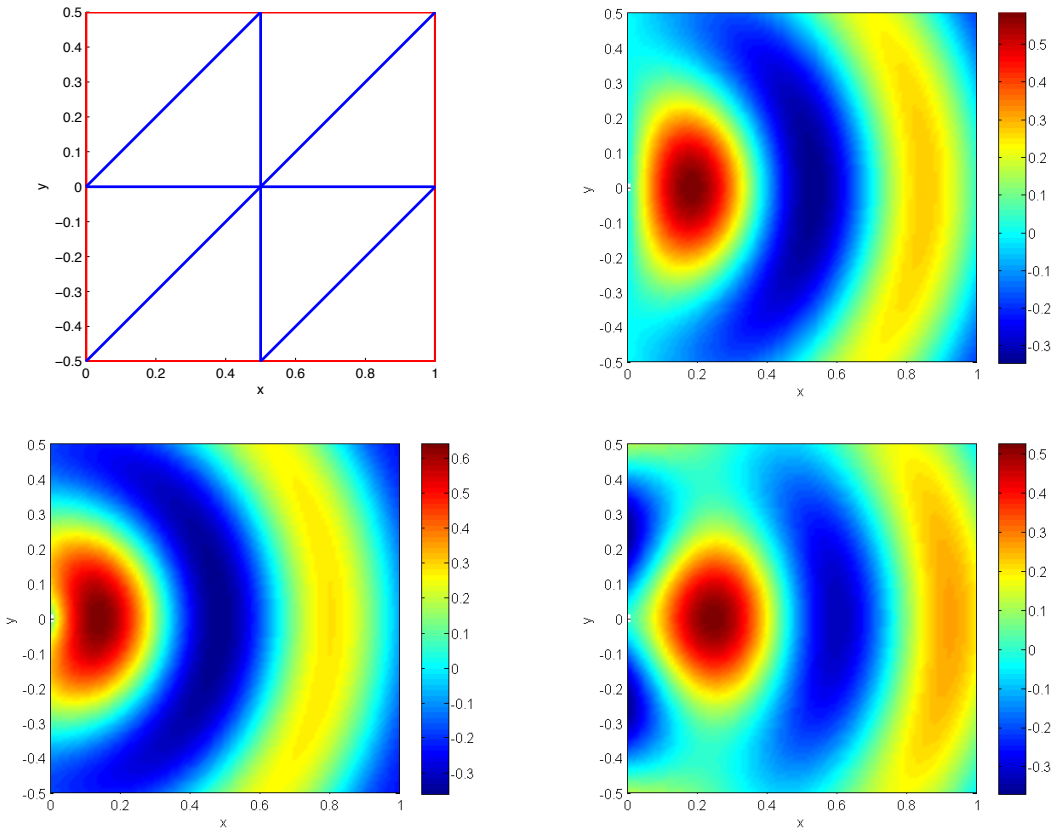

Figure 4.1: The mesh used for numerical experiments and the analytical solutions for $\xi = 1,\ 2/3,\ 3/2$. The colored bar on the right of each figure indicates the mapping between data values and colors.

for the two different choices of flux parameters ($\alpha = \beta = \delta = 1/2$) and ($\alpha = \beta^{-1} = \delta^{-1} = 10p/(\omega h \log p)$) are shown in Figures 4.2 and 4.5 respectively. Similarly for ($\alpha = \beta = \delta = 1/2$) and ($\alpha = \beta^{-1} = \delta^{-1} = 10p/(\omega h \log p)$), the starting control $\boldsymbol{\theta_0}$ and the optimal control obtained via the projected gradient method for particular choices of p are shown in Figures 4.3 and 4.5 respectively. In each figure, the plots corresponding to $\xi = 1,\ 2/3$, and $3/2$ are in first, second and third rows respectively. Also, the plots for uniformly distributed initial control are on the left and the plots for randomly distributed initial control are on the right.

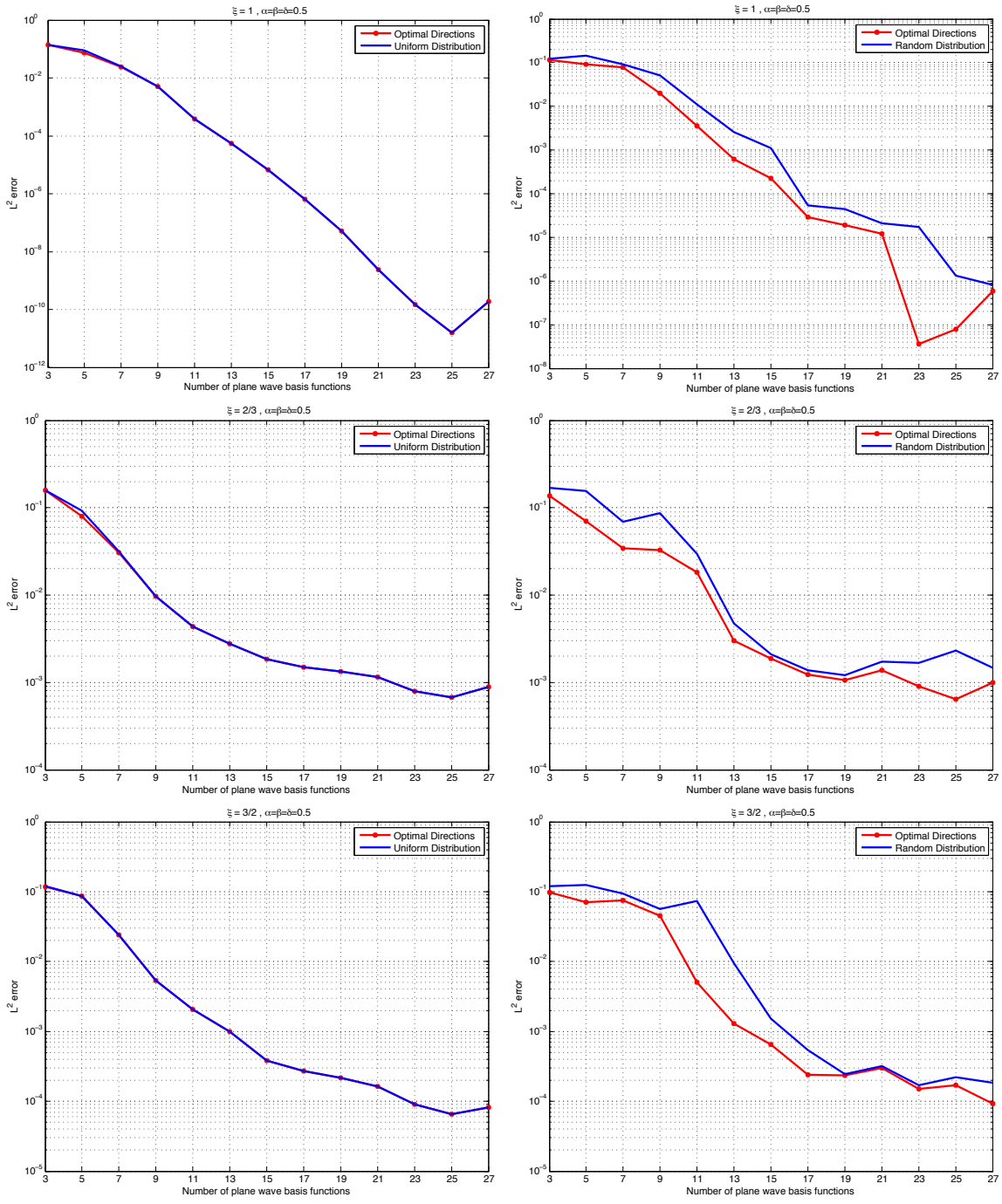

Figure 4.2: L^2 errors for $\alpha = \beta = \delta = 0.5$

Analysis of the results we get from the experiments show that in this example, if we consider the uniform distribution of the starting control $\boldsymbol{\theta_0} = (\theta_1, \theta_2, \cdots, \theta_p)^T$, where $\theta_i = 2\pi(i-1)/p,\ 1 \leq i \leq p$, optimization of the plane wave directions does not lead to a significant overall improvement in the L^2 error. We only obtain minor improvements in 10 out of 78 cases (13 values of p, 3 values of ξ, 2 choices of flux parameters α, β, δ) when starting with a uniform distribution of θ_0.

However, if we choose a random distribution of starting control $\boldsymbol{\theta_0}$, we observe significant improvements in the L^2 error in nearly all of the 78 test cases. In some cases we see an improvement of orders of magnitude.

These observations suggest that for this particular example, a uniform distribution of $\boldsymbol{\theta_0}$ is optimal or close to optimal for all tested cases. However, the improvements observed in case of a random distribution of $\boldsymbol{\theta_0}$ validates our belief that the optimal choice of plane wave directions in PWDG leads to a reduction in error between the computed and the analytical solutions.

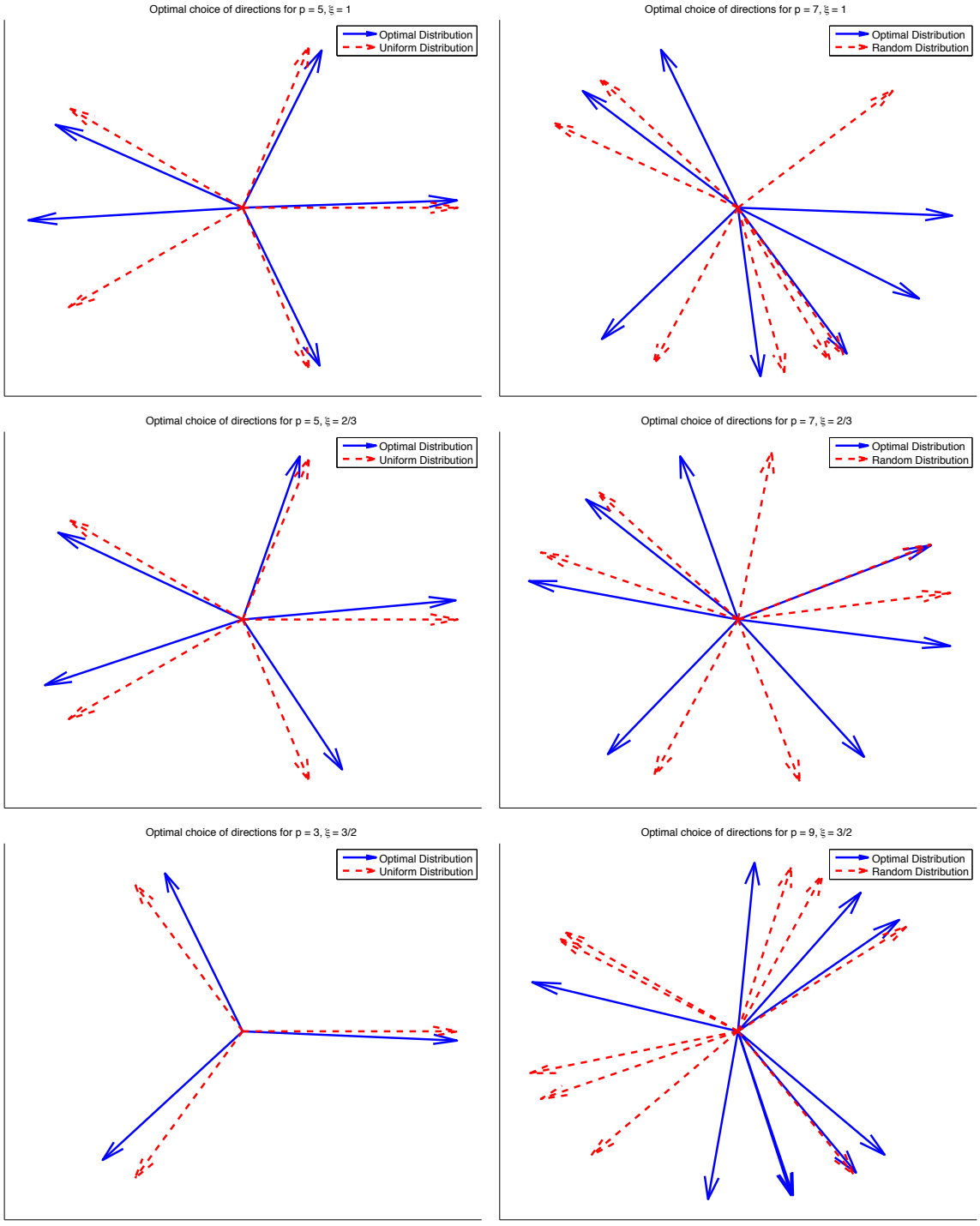

Figure 4.3: Starting and Optimal distributions of directions for $\alpha = \beta = \delta = 0.5$

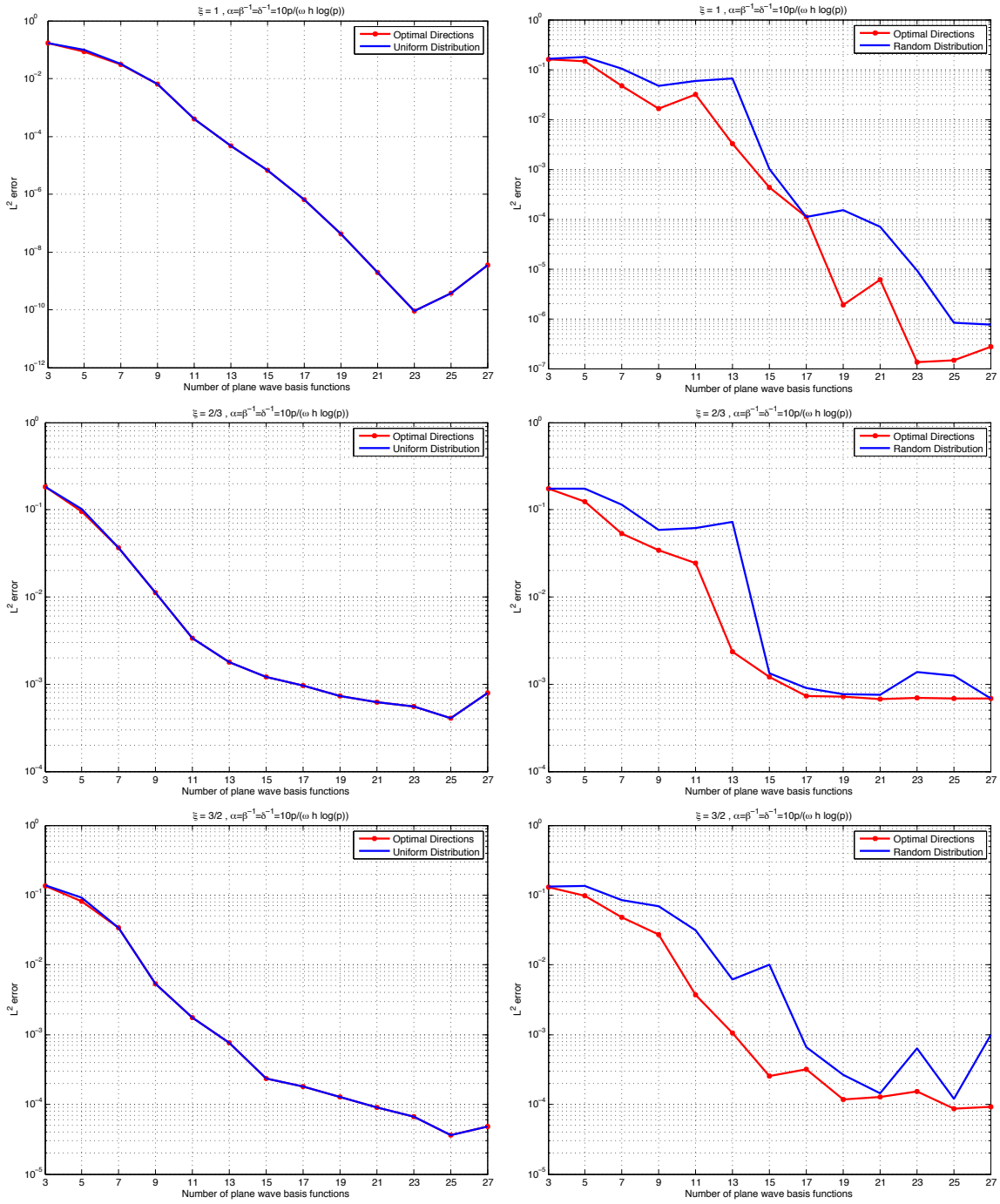

Figure 4.4: L^2 errors for $\alpha = \beta^{-1} = \delta^{-1} = 10p/(\omega h \log(p))$

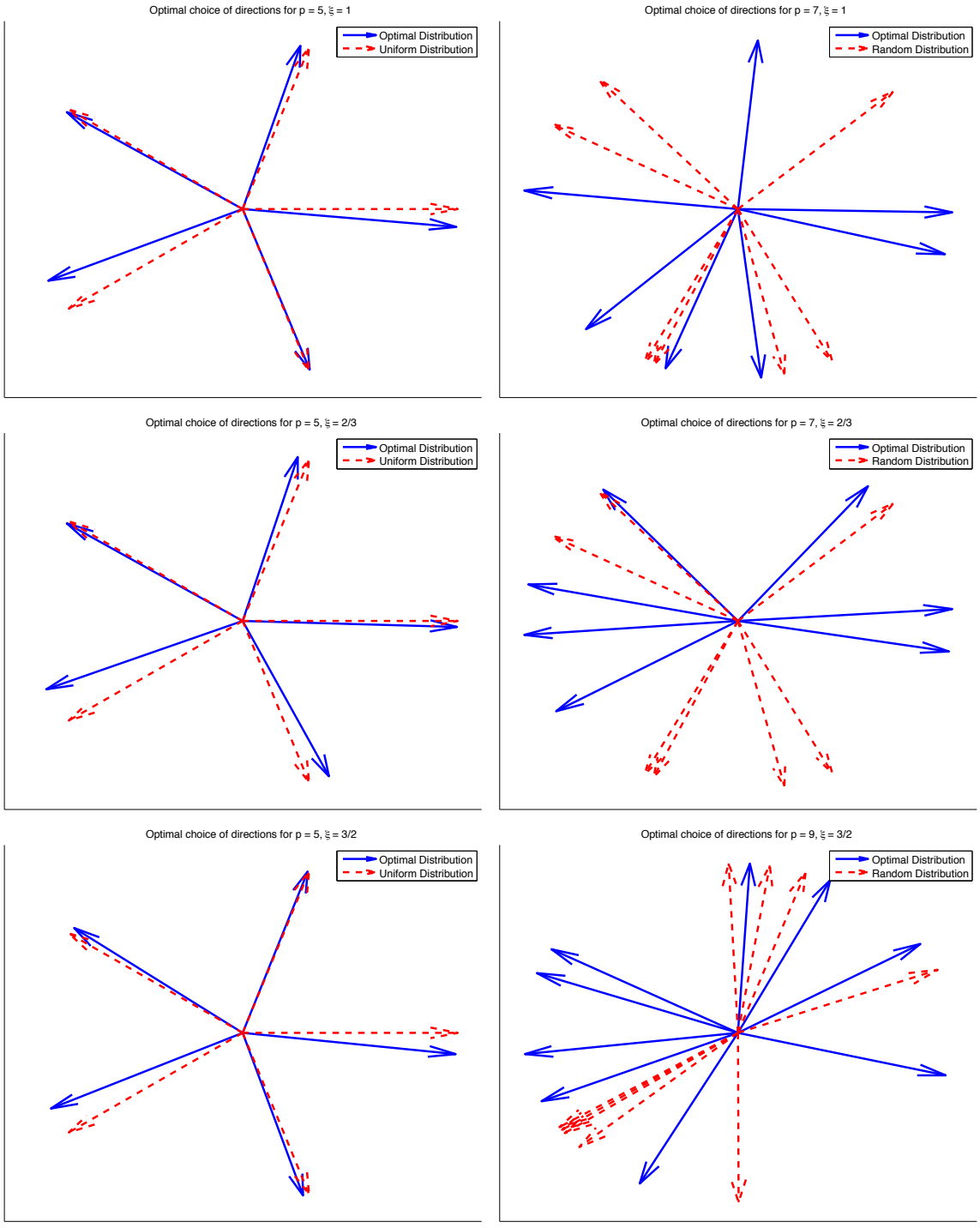

Figure 4.5: Starting and Optimal distributions of directions for $\alpha = \beta^{-1} = \delta^{-1} = 10p/(\omega h \log(p))$

4.2 Test Problem on Non-Convex Domain: Screen Problem

For this example, we choose $\Omega = (-1,1)^2 \setminus (S_1 \cup S_2)$ where

$$S_1 = \operatorname{conv}\left((0,0),(-0.25,+0.50),(-0.50,+0.50)\right)$$

$$S_2 = \operatorname{conv}\left((0,0),(+0.25,-0.50),(+0.50,-0.50)\right)$$

Let $\Gamma_R = \partial(-1,+1)^2$ and $\Gamma_D = \partial S_1 \cup \partial S_2$. Consider the problem

$$-\Delta u - \omega^2 u = 0 \qquad \text{in } \Omega, \tag{4.1a}$$

$$\boldsymbol{n} \cdot \nabla u + i\omega u = g \qquad \text{on } \Gamma_R, \tag{4.1b}$$

$$u = 0 \qquad \text{on } \Gamma_D \tag{4.1c}$$

which describes an acoustic wave with wave number $\omega > 0$ scattered at the sound-soft scatterer $\Omega_D = S_1 \cup S_2$ with boundary Γ_D.

Note that equation (4.1b) describes the non-homogeneous Robin boundary condition and (4.1c) describes the homogeneous Dirichlet boundary condition. To account for the additional Dirichlet boundary condition we have to modify the flux functions $\widehat{u_h}$ and $\widehat{\boldsymbol{\sigma}_h}$ from(2.10a)-(2.10b) as follows

$$\widehat{u}_h|_E := \begin{cases} \{u_h\}_E - \frac{\beta}{i\omega}[\nabla u_h]_E & , \quad E \in \mathcal{E}_h(\Omega) \\ u_h - \delta\left(\frac{1}{i\omega}\boldsymbol{n}_E \cdot \nabla u_h + u_h - \frac{1}{i\omega} g\right) & , \quad E \in \mathcal{E}_h(\Gamma_R) \\ u_h & , \quad E \in \mathcal{E}_h(\Gamma_D) \end{cases}, \tag{4.2a}$$

$$\widehat{\boldsymbol{\sigma}}_h|_E := \begin{cases} \frac{1}{i\omega}\{\nabla u_h\}_E - \alpha[u_h]_E & , \quad E \in \mathcal{E}_h(\Omega) \\ \frac{1}{i\omega}\nabla u_h - (1-\delta)\left(\frac{1}{i\omega}\nabla u_h + \boldsymbol{n}_E u_h - \frac{1}{i\omega}\boldsymbol{n}_E g\right) & , \quad E \in \mathcal{E}_h(\Gamma_R) \\ \frac{1}{i\omega}\nabla u_h - \kappa h_E^{-1}\left(\boldsymbol{n}_E u - \boldsymbol{n}_E g\right) & , \quad E \in \mathcal{E}_h(\Gamma_D) \end{cases}, \tag{4.2b}$$

Here $\kappa \approx (1+p)^2$ is a sufficiently large penalty parameter that allows us to enforce the Dirichlet boundary conditions on Γ_D.

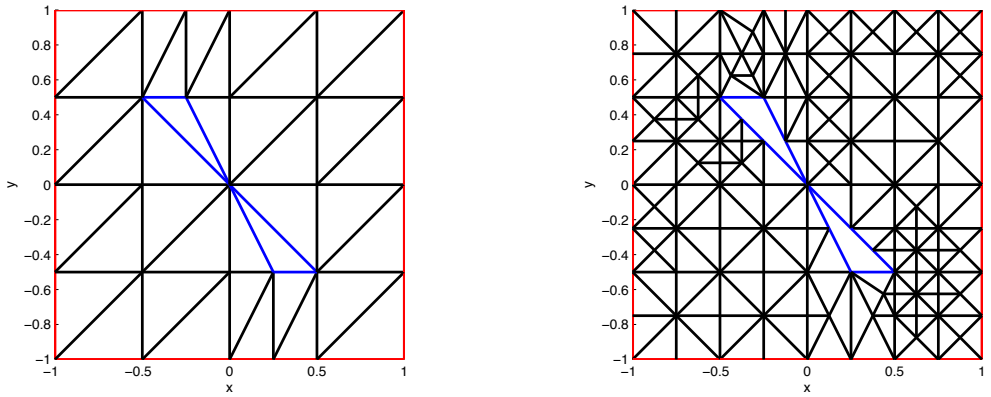

Figure 4.6: Starting mesh used in the Adaptive IPDG code (left) and final mesh obtained after 3 refinement steps (right)

For our test problem we consider $\omega = 15$ and a non-homogeneous Robin boundary value (g in 4.1b) as follows:

$$g = \omega \cos(y) + i\omega \sin(y).$$

Note that we can not calculate the analytical solution of (4.1a)-(4.1c). Therefore, we refer to [17] and [15], which use an Adaptive Interior Penalty Discontinuous Galerkin scheme to solve the same problem. We use its implementation and theory to obtain an approximation of the analytical solution of (4.1a)-(4.1c), which is needed for our optimal control method. We input the starting mesh (left plot in Figure 4.6) into the adaptive IPDG implementation from [17] and after 3 refinement steps we obtain the mesh (right plot in Figure 4.6) and the approximation to exact solution of (4.1a)-(4.1c) required for the optimal control algorithm. The profile for the approximate exact solution is displayed in Figure 4.7.

Note that we restrict the number of refinement steps to 3. This is required because further refinement leads to a finer mesh, which leads to stiffness matrix $\boldsymbol{A}(\boldsymbol{\theta})$ in (2.17) becoming ill-conditioned. We also restrict our choice of the number of plane wave directions p to $3, 5, 7$ and 9. Higher values of p lead to $\boldsymbol{A}(\boldsymbol{\theta})$ becoming ill-conditioned. We choose flux

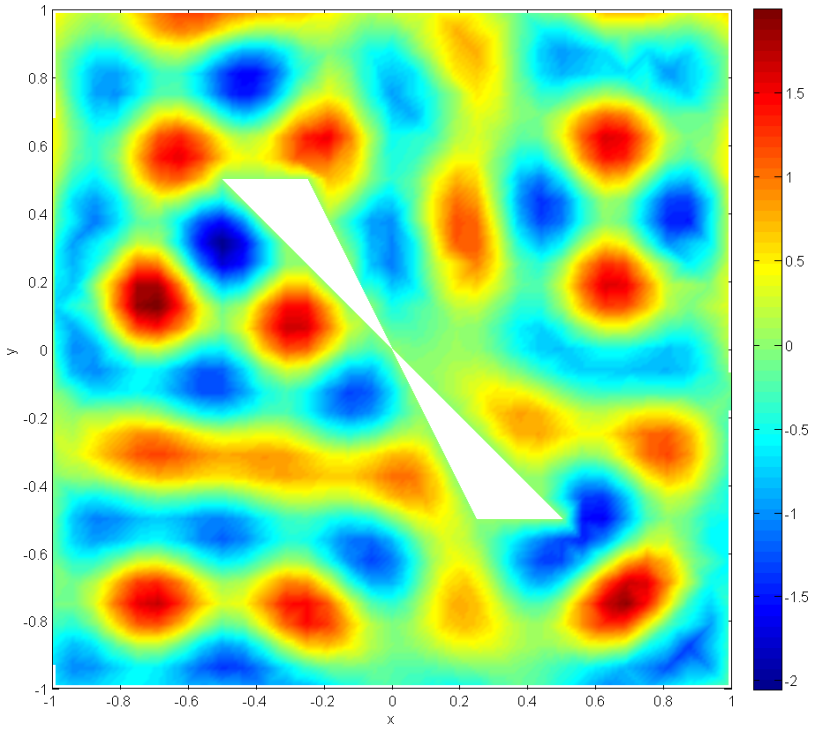

Figure 4.7: IPDG approximation to solution of (4.1a)-(4.1c). The colored bar on the right indicates the mapping between data values and colors.

parameters $\alpha = \beta = \delta = 1/2$.

For starting control $\boldsymbol{\theta_0} = (\theta_1, \theta_2, \cdots, \theta_p)^T$ we choose the uniform distribution, that is

$$\theta_i = 2\pi(i-1)/p, \ 1 \leq i \leq p.$$

Plot comparing the L^2 errors between the computed solution and the IPDG approximation of (4.1a)-(4.1c) is shown in Figure 4.8. The starting control $\boldsymbol{\theta_0}$ and the optimal control obtained via the projected gradient method for each choice of p are shown in Figure 4.9. In this example we observe that despite starting with a uniform distribution of starting control $\boldsymbol{\theta_0}$, optimization of plane wave directions leads to a significant reduction in the L^2 error in all of the tested cases. This is in contrast with the previous example where starting

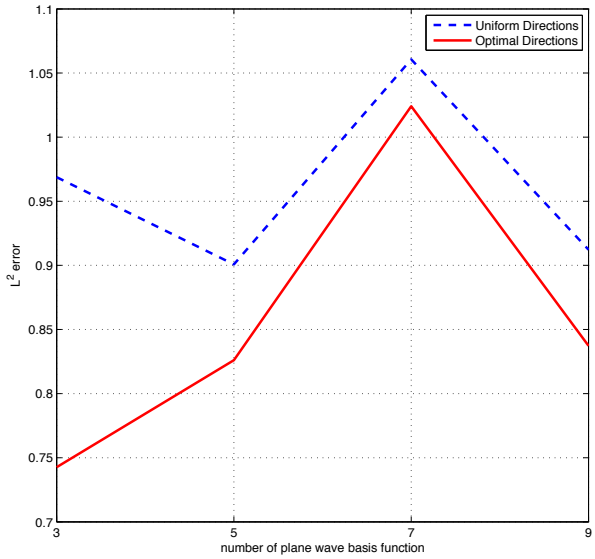

Figure 4.8: L^2 errors for Screen Problem

with a uniform distribution of starting control $\boldsymbol{\theta_0}$ resulted in insignificant reductions in L^2 error.

This observation can be explained by the fact that the analytical solutions in the first example were symmetric to some extent. However, the solution to the screen problem does not exhibit any form of symmetry. This suggests that while a uniform distribution of plane wave directions is close to optimal if the solution exhibits some symmetry, that is not the case when the solution is asymmetric. This again validates our our belief that optimal choice of plane wave directions leads to a reduction in the error between computed and actual solutions.

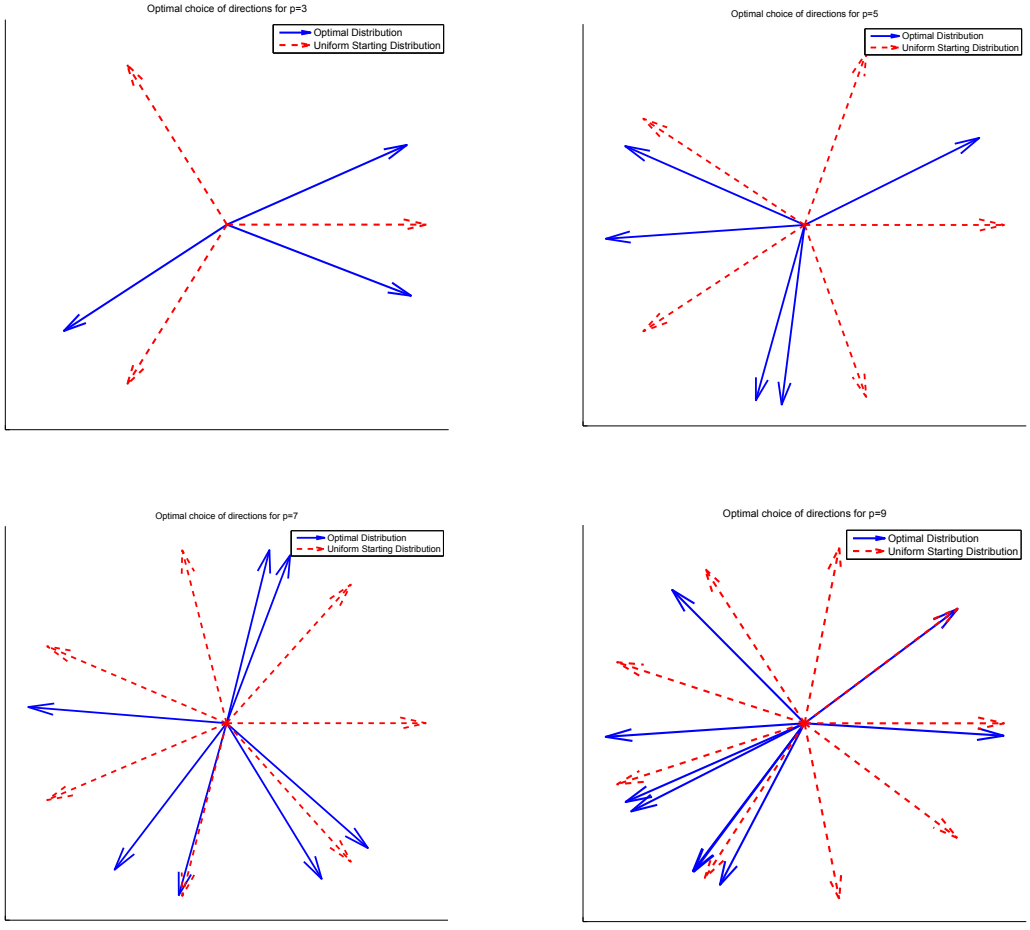

Figure 4.9: Starting and Optimal distributions of directions

CHAPTER 5

Conclusions and Future Work

We investigated the dependence on the plane wave directions chosen for the basis functions used in PWDG method used to solve the 2D Helmholtz equation. We studied this dependence by formulating the choice of the directions as an optimal control problem with a tracking type objective functional and the variational formulation of the PWDG method as a constraint. We proved that the necessary optimality conditions hold true. However, due to the problem being non-convex, we have multiple local minima. Thus, our optimal choice of directions depends on the initial value of the control that we use. We test out optimal control algorithm on two different examples.

The first example considers a Helmholtz equation with non-homogeneous Robin boundary conditions. This example is also considered in [13]. We observe that in this case the uniform distribution of directions is close to optimal. We assume that this is due to the symmetry of the solution. However, if we choose a random distribution of directions as our starting

control, as suggested in [3] by Cessenat and Despres, we do achieve significant reductions in error by optimizing the choice of plane wave directions. This illustrates the benefits of choosing the plane wave directions optimally. This is further validated by the second example in which we consider the screen problem which describes an acoustic wave being scattered by a sound-soft scatterer. In the second example, where we do have symmetry of the solution, we see significant gains by optimizing our choice of plane wave directions compared to the uniform distribution.

In the second example we had to greatly restrict our choice of p due to ill-conditioned systems being generated by the PWDG method. The development of an effective preconditioner remains an issue that needs to be addressed in future work. This would allow us to validate our work further for a wider range of problems. Another avenue for further investigation is to allow the directions to vary independently in each triangle of the triangulation. Currently we use the same directions in all triangles of the triangulation. We believe allowing the directions to vary independently will lead to significant reductions in error. The possibility of varying number of plane waves independently of the triangle can also be explored. Finally describing an hp-adaptive PWDG method for Helmholtz equation can be explored.

Bibliography

[1] Gustavo Benitez Alvarez, Abimael Fernando Dourado Loula, Eduardo Gomes Dutra do Carmo, and Fernando Alves Rochinha. A discontinuous finite element formulation for Helmholtz equation. *Computer Methods in Applied Mechanics and Engineering*, 195(33):4018–4035, 2006.

[2] Mohamed Amara, Rabia Djellouli, and Charbel Farhat. Convergence analysis of a discontinuous Galerkin method with plane waves and Lagrange multipliers for the solution of Helmholtz problems. *SIAM Journal on Numerical Analysis*, 47(2):1038–1066, 2009.

[3] Olivier Cessenat and Bruno Despres. Application of an ultra weak variational formulation of elliptic pdes to the two-dimensional Helmholtz problem. *SIAM journal on numerical analysis*, 35(1):255–299, 1998.

[4] Eric T Chung and Björn Engquist. Optimal discontinuous Galerkin methods for wave propagation. *SIAM Journal on Numerical Analysis*, 44(5):2131–2158, 2006.

[5] Arnaud Deraemaeker, Ivo Babuška, and Philippe Bouillard. Dispersion and pollution of the FEM solution for the Helmholtz equation in one, two and three dimensions. *International journal for numerical methods in engineering*, 46(4):471–499, 1999.

[6] Xiaobing Feng and Haijun Wu. Discontinuous Galerkin methods for the Helmholtz equation with large wave number. *SIAM Journal on Numerical Analysis*, 47(4):2872–2896, 2009.

[7] Xiaobing Feng and Haijun Wu. *hp*-discontinuous Galerkin methods for the Helmholtz equation with large wave number. *Mathematics of Computation*, 80(276):1997–2024, 2011.

[8] Gwénaël Gabard. Discontinuous Galerkin methods with plane waves for time-harmonic problems. *Journal of Computational Physics*, 225(2):1961–1984, 2007.

[9] C Gittelson, R Hiptmair, and I Perugia. Plane wave discontinuous galerkin methods. *Isaac Newton Institute Preprint Series*, 2007.

[10] Claude J Gittelson, Ralf Hiptmair, and Ilaria Perugia. Plane wave discontinuous Galerkin methods: Analysis of the h-version. *ESAIM: Mathematical Modelling and Numerical Analysis*, 43(02):297–331, 2009.

[11] Roland Griesmaier and Peter Monk. Error analysis for a hybridizable discontinuous Galerkin method for the Helmholtz equation. *Journal of Scientific Computing*, 49(3):291–310, 2011.

[12] R Hiptmair, A Moiola, and I Perugia. Plane wave discontinuous Galerkin methods: Exponential convergence of the hp-version. *Foundations of Computational Mathematics*, pages 1–39, 2015.

[13] Ralf Hiptmair, Andrea Moiola, and Ilaria Perugia. Plane wave discontinuous Galerkin methods for the 2D Helmholtz equation: Analysis of the p-version. *SIAM Journal on Numerical Analysis*, 49(1):264–284, 2011.

[14] Ralf Hiptmair, Andrea Moiola, and Ilaria Perugia. A survey of Trefftz methods for the Helmholtz equation. *arXiv preprint arXiv:1506.04521*, 2015.

[15] Ronald HW Hoppe and N Sharma. Convergence analysis of an adaptive interior penalty discontinuous Galerkin method for the Helmholtz equation. *IMA Journal of Numerical Analysis*, 2012.

[16] Kazufumi Ito and Karl Kunisch. *Lagrange multiplier approach to variational problems and applications*, volume 15. SIAM, 2008.

[17] Natasha S Sharma. *Convergence analysis of an adaptive Interior Penalty Discontinuous Galerkin method for the Helmholtz equation*. PhD thesis, University of Houston, 2011.

[18] Luc Tartar. *An introduction to Sobolev spaces and interpolation spaces*, volume 3. Springer, Berlin-Heidelberg-New York, 2007.

[19] Fredi Tröltzsch. *Optimal control of partial differential equations*, volume 112. 2010.