Title

Galerkin Methods for Solving the Helmholtz Problem in Scattering by Negative Materials with Applications for a Near-Resonance Regime

ISBN

9798297606555

Author

Latham, Benjamin Jarred

Publication Date

2025-07-24

GALERKIN METHODS FOR SOLVING THE HELMHOLTZ PROBLEM IN SCATTERING BY NEGATIVE MATERIALS WITH APPLICATIONS FOR A NEAR-RESONANCE REGIME

by

Benjamin Jarred Latham
Masters of Arts in Mathematics

A Dissertation

Submitted to the Graduate Faculty

of

University of California, Merced

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Merced, California

July
2025

This dissertation, submitted by Benjamin Jarred Latham in partial fulfillment of the requirements for the Degree of Doctor of Philosophy from University of California, Merced, has been read by the Faculty Advisory Committee under whom the work has been done and is hereby approved.

Professor Arnold Kim

Professor Chrysoula Tsogka

Professor Camille Carvalho

Professor Roummel Marcia

This dissertation is being submitted by the appointed advisory committee as having met all of the requirements of the School of Graduate Studies at the University of California Merced and is hereby approved.

PERMISSION

Title	GALERKIN METHODS FOR SOLVING THE HELMHOLTZ PROBLEM IN SCATTERING BY NEGATIVE MATERIALS WITH APPLICATIONS FOR A NEAR-RESONANCE REGIME
Department	Applied Mathematics
Degree	Doctor of Philosophy

In presenting this dissertation in partial fulfillment of the requirements for a graduate degree from University of California, Merced, I agree that the library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by the professor who supervised my dissertation work or, in his absence, by the Chairperson of the department or the dean of the School of Graduate Studies. It is understood that any copying or publication or other use of this dissertation or part thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of California in any scholarly use which may be made of any material in my dissertation.

Benjamin Jarred Latham
July 2025

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

LIST OF ACRONYMS

ABC	Absorbing Boundary Condition
BIE	Boundary Integral Equation
DG	Discontinuous Galerkin
DPG	Discontinuous Petrov–Galerkin
DtN	Dirichlet-to-Neumann
EPW	Evanescent Plane Wave
FEM	Finite Element Method
HDG	Hybridizable Discontinuous Galerkin
PML	Perfectly Matched Layer
PPW	Propagative Plane Wave
PW	Plane Wave
PWDG	Plane Wave Discontinuous Galerkin

ACKNOWLEDGEMENTS

I am deeply grateful to my advisor, Dr. Camille Carvalho, for her boundless enthusiasm, unwavering patience, and tireless mentorship, even late into the evening. I also thank Dr. Chrysoula Tsogka and Dr. Arnold Kim for serving on my committee, for welcoming me into their research group, and for their thoughtful guidance.

My heartfelt thanks go to Elsie Cortes for her friendship and companionship on many long journeys overseas. I'm likewise indebted to the rest of the Kim–Tsogka research group, the Imaging and Sensing Seminar, and the Applied Mathematics Department for their support. Special thanks to Dr. Symeon Papadimitropoulos for both his research advice and his friendship, and to Dr. Zoïs Moitier for his mentorship at the outset of my graduate studies.

I would like to acknowledge the Sacramento State Mathematics Department—especially Drs. Tracy Hamilton, Sayonita Ghosh Hajra, Gary Shannon, and Ed Shea for my education during my master's degree.

I would like to thank my family for their boundless support during the trying times of my doctorate, in particular my parents, Brenda and Eldon Latham. I would also like to thank Raquel Munoz, without whom I would have gone insane during the preparation of this document.

I would also like to thank the members of Crumbs-US for putting up with me - even when running code too late and waking up too early for meetings made me extra annoying.

Last, but by no means least, I would like to thank the following persons who have been among my closest supporters and confidantes: Dr. Majerle Reeves, Andrew Garcia, Hector Berrospe, Elaisha Major, Austin Leer, Sean Shearer, Shayla Young, Ian Shearer, Sierah Swindlehurst and Dimitri Phillips. Thank you for helping me build better worlds.

For my grandparents, who never gave up.

For Dimitri, whose kindness will never be forgotten.

ABSTRACT

Understanding light–matter interactions in plasmonic nanostructures and metamaterials is crucial for a wide range of applications from biochemical sensing and high resolution imaging to optical cloaking. This leads us to study Wave scattering by negative materials, which in turn leads to a Helmholtz equation with sign-changing coefficients, a regime in which classical coercivity of the variational formulation breaks down at resonance frequencies. This thesis develops and analyzes Galerkin methods to address the theoretical and computational challenges of this near-resonant plasmonic scattering problem. On the theoretical side, we characterize the near-resonance behavior by constructing the resonances and associated resonant modes for canonical geometries (e.g. a penetrable ball). As the material parameter ϵ_m approaches a critical value, the solution exhibits unbounded growth in its amplitude, with a loss of coercivity and can admit, for some wavenumbers, a plasmonic resonance. To give us well-posedness in the near region surrounding these pathological points, we re-introduce the technique of T-coercivity, developed in Bonnet-Ben Dhia et al. (2012, 2016, 2018) for this problem, where we may show we have Fredholm solvability. These theoretical results (including an adaption of the T-coercivity theory to the discrete setting) lay the groundwork for developing robust numerical schemes is the plasmonic regime.

On the computational side, we design and investigate several Galerkin strategies for this sign-changing Helmholtz problem, including a standard FEM method,

an enriched FEM (in 3D) and a Trefftz DG method (in 2D). The standard H^1 FEM is used as a baseline, but fails in the near-resonance case. In this near-resonant regime, the solution develops rapid oscillations and strongly localized fields (surface plasmons) that are difficult to capture with a low-order polynomial basis, often requiring prohibitively fine meshes. In 3D, We propose an enriched FEM, where we augment the usual FEM polynomial space with the analytical surface plasmon to encapsulate the dominant scattering mod. We derive a modified weak formulation in this extended function space and show that the resulting enriched system couples the standard H^1 solution with the added plasmonic basis function.

In parallel in 2D, we develop a Trefftz-DG method as an alternative Galerkin approach tailored to the Helmholtz operator. In this method, we abandon the polynomial basis and instead use Trefftz basis functions – specifically, propagative and evanescent plane waves – that exactly satisfy the local Helmholtz equation inside each element. This plane-wave enriched DG formulation more naturally embeds the oscillatory and decaying components of the solution into the approximation space. We extend the notion of T-coercivity to the discontinuous Galerkin setting to handle the sign-changing coefficient, and we prove a number of fundamental results for the Trefftz DG problem in the sign-changing case, leading to both a quasi-optimal convergence estimate for the Trefftz-DG scheme under an augmented norm (demonstrating stability of the method in the T-coercive case) and control of the L^2 norm by the method. Finally, we present numerical results for the Trefftz PWDG method, both for the sign-changing method in general, and for the near-resonance case in particular.

1. INTRODUCTION & MOTIVATION

In the field of modern optics, it is often of interest to study the propagation and scattering of light in nanostructures for the purposes of designing photovoltaic cells, ultra-sensitive sensors, and nano-antennas, developing high-resolution imaging techniques (including applications to Raman spectroscopy, medical imaging), investigating cloaking effects, and others Jiménez (2020); Goul (2017). In particular, the characterization and control of large field enhancements in those structures is key.

Our work focuses on the study of light interactions with plasmonic structures. These structures are made of metals or metamaterials spherical nanoparticles surrounded by common dielectric materials (silica, glass, air, etc.). At optical frequencies, noble metals such as gold and silver exhibit unusual optical properties such as a dielectric permittivity with a negative real part (and small dissipation effects). Metamaterials are constructed conglomerates of materials so that one can tune their effective permittivity (resulting for instance to a negative one). Due to this juxtaposition of *positive* materials and *negative* materials, the illumination of plasmonic structures gives rise of the so-called surface plasmons Maier (2007).

A surface plasmon is highly oscillatory (subwavelength phenomenon), and when excited, it corresponds to a resonant mode localized at the interface between a dielectric and a metal/metamaterial. It follows that surface plasmons create significantly large field enhancements that are useful in nanophotonics, but because of their high variance on an extremely small scale they can be difficult to capture, both experimentally and numerically.

Over the past decades, several works tackled plasmonic problems. In the context of plasmonic scattering resonances, studies based on variational methods and using FEM [Bonnet-Ben Dhia et al. (2013); Grieser (2014); Bonnet-Ben Dhia et al. (2016); Carvalho et al. (2017)] have been well developed in two dimensions, and Boundary Integral Equation (BIE) methods have provided several results in the quasi-static case Ammari et al. (2017); Helsing & Karlsson (2018); Bonnetier et al. (2019); Ammari, Chow, & Liu (2020); Ammari, Dabrowski, et al. (2020). As this thesis is concerned with the application of Galerkin methods, we also note that the broader family of Galerkin methods—including Discontinuous Galerkin (DG), Hybridizable Discontinuous Galerkin (HDG), and Discontinuous Petrov–Galerkin (DPG)—has attracted significant attention for plasmonic and metamaterial scattering. DG methods offer flexible non-conforming meshes and local *hp*–refinement, making them well suited to resolve the steep gradients at metal–dielectric interfaces Martín et al. (2023). DPG variants, with their built-in stability via optimal test norms, have been used to simulate anisotropic, sign-changing media such as transformation-optics cloaks Demkowicz & Li (2013), while HDG schemes efficiently handle nonlocal (hydrodynamic) Drude models relevant at the nanoscale Li et al. (2017). Complementing these, high-order *hp*–FEM combined with Perfectly Matched Layer (PML) has achieved exponential convergence for plasmonic particles in layered media Wang et al. (2013), and Trefftz–DG methods—using local plane-wave basis functions—have been shown to dramatically reduce dispersion error at high frequencies Hiptmair et al. (2011). Such approaches illustrate the diversity of Galerkin-family techniques available to tackle high-contrast, oscillatory plasmonic problems. The three dimensional case (and without the use of the quasi-static approximation) is more challenging: it becomes computationally demanding and several theoretical questions remain open (especially with tips and corners). The main challenges consist of capturing the fast subwavelength oscillations of surface plasmons as well as their excitation factor with potentially limited computational resources.

1.1. Derivation of the Scalar Helmholtz Equation

To model the electromagnetic scattering we start by considering Maxwell's equations and simplify to a scalar Helmholtz-like equation. (For details on the time-harmonic approach to Maxwell's equations we reference Kirsch & Hettlich (2015).) We consider a system consisting of a dielectric and a scattering material with piecewise constant permittivity $\epsilon(\mathbf{x})$. We assume a smooth and closed interface Γ between the scatterer and the dielectric. Let $\mathbf{E}$ and $\mathbf{H}$ denote the electric and magnetic fields, respectively. The displacement field $\mathbf{D}$ and the magnetic flux density $\mathbf{B}$ are related to $\mathbf{E}$ and $\mathbf{H}$ by:

$$\mathbf{D} = \epsilon(\mathbf{x})\mathbf{E}, \tag{1.1}$$

$$\mathbf{B} = \mu_0\mathbf{H}, \tag{1.2}$$

where $\epsilon(\mathbf{x})$ is the permittivity of the medium, and μ_0 is the permeability of free space.

Here we assume that $\epsilon(\mathbf{x})$ is piecewise constant within each medium, and that $\epsilon(\mathbf{x}) = \epsilon_0$ in free space.

Maxwell's equations in the absence of free charges and (as we are studying electromagnetic scattering) no external electric current are:

$$\nabla \times \mathbf{E} = -\mu_0 \frac{\partial \mathbf{H}}{\partial t}, \qquad \text{(Faraday's Law)} \tag{1.3}$$

$$\nabla \times \mathbf{H} = \epsilon(\mathbf{x}) \frac{\partial \mathbf{E}}{\partial t}, \qquad \text{(Ampère's Law)} \tag{1.4}$$

$$\nabla \cdot \mathbf{D} = 0, \qquad \text{(Gauss's Law for Electricity)} \tag{1.5}$$

$$\nabla \cdot \mathbf{B} = 0. \qquad \text{(Gauss's Law for Magnetism)} \tag{1.6}$$

We also assume the tangential and normal components of the electric and magnetic fields must be continuous across Γ:

$$[n \times \mathbf{E}]_\Gamma = 0, \tag{1.7}$$

$$[n \times \mathbf{H}]_\Gamma = 0, \tag{1.8}$$

$$[n \cdot \mathbf{D}]_\Gamma = 0, \tag{1.9}$$

$$[n \cdot \mathbf{B}]_\Gamma = 0. \tag{1.10}$$

where $[\cdot]_\Gamma$ denotes the jump of a quantity across the interface.

Furthermore, to ensure that the scattered electromagnetic field represents only outgoing waves, we impose the Silver–Müller radiation condition.

$$\mathbf{E} \times \hat{\mathbf{r}} - \eta_0 \mathbf{H} \to 0, \qquad \|\mathbf{x}\| \to \infty, \tag{1.11}$$

$$\mathbf{H} \times \hat{\mathbf{r}} + \frac{1}{\eta_0} \mathbf{E} \to 0, \qquad \|\mathbf{x}\| \to \infty, \tag{1.12}$$

where $\hat{\mathbf{r}} = \frac{\mathbf{x}}{\|\mathbf{x}\|}$ is the unit vector in the radial direction, and $\eta_0 = \sqrt{\mu_0/\epsilon_0}$ is the impedance of free space.

Assuming time-harmonic fields at frequency ω, we express the fields as:

$$\mathbf{E}(\mathbf{x}, t) = \mathbf{E}(\mathbf{x}) e^{-i\omega t}, \tag{1.13}$$

$$\mathbf{H}(\mathbf{x}, t) = \mathbf{H}(\mathbf{x}) e^{-i\omega t}. \tag{1.14}$$

Substituting into Maxwell's equations, we obtain:

$$\nabla \times \mathbf{E}(\mathbf{x}) = -i\omega\mu_0\mathbf{H}(\mathbf{x}), \tag{1.15}$$

$$\nabla \times \mathbf{H}(\mathbf{x}) = -i\omega\epsilon(\mathbf{x})\mathbf{E}(\mathbf{x}), \tag{1.16}$$

$$\nabla \cdot (\epsilon(\mathbf{x})\mathbf{E}(\mathbf{x})) = 0, \tag{1.17}$$

$$\nabla \cdot \mathbf{H}(\mathbf{x}) = 0. \tag{1.18}$$

Taking the curl of $\nabla \times \mathbf{E}$ and substituting $\nabla \times \mathbf{H}$:

$$\nabla \times (\nabla \times \mathbf{E}) = -i\omega\mu_0\nabla \times \mathbf{H} \tag{1.19}$$

$$= -i\omega\mu_0 \left(-i\omega\epsilon(\mathbf{x})\mathbf{E}\right) \tag{1.20}$$

$$= -\omega^2\mu_0\epsilon(\mathbf{x})\mathbf{E}. \tag{1.21}$$

Using the identity $\nabla \times (\nabla \times \mathbf{E}) = \nabla\left(\nabla \cdot \mathbf{E}\right) - \nabla^2\mathbf{E}$, we have:

$$\nabla\left(\nabla \cdot \mathbf{E}\right) - \nabla^2\mathbf{E} = -\omega^2\mu_0\epsilon(\mathbf{x})\mathbf{E}. \tag{1.22}$$

In regions where $\epsilon(\mathbf{x})$ is constant (within each material), we may use (1.17) to obtain:

$$\nabla \cdot \mathbf{E} = 0. \tag{1.23}$$

Therefore, the equation simplifies to:

$$\nabla^2\mathbf{E} + \omega^2\mu_0\epsilon(\mathbf{x})\mathbf{E} = 0. \tag{1.24}$$

within each domain. Given this is a Cartesian Vector-Helmholtz equation, we can decouple and solve components individually, leading to a scalar Helmholtz equation of the form:

$$\nabla^2 u + \omega^2 \mu_0 \epsilon(\mathbf{x}) u = 0. \tag{1.25}$$

As we are studying scattering, we will assume that the field $u(\mathbf{x})$ is a component of the electrical field, and that it can be split into two parts, the incident and scattered fields. Thus we write $u = u^{\text{inc}} + u^{\text{sca}}$.

Defining the wavenumber $k = \omega\sqrt{\mu_0 \epsilon_0}$ and the relative permittivity $\varepsilon_r(\mathbf{x}) = \epsilon(\mathbf{x})/\epsilon_0$, we have:

$$\nabla^2 u + k^2 \varepsilon_r(\mathbf{x}) u = 0. \tag{1.26}$$

Henceforth, we will drop the r subscript and refer to the relative permittivity only.

Since $\epsilon(\mathbf{x})$ is piecewise constant, we can write the equation in the form:

$$\nabla \cdot \left(\epsilon^{-1}(\mathbf{x}) \nabla u\right) + k^2 u = 0. \tag{1.27}$$

As we have this equation only inside each material, we need to describe the behaviour of our field at the interface Γ between the dielectric and the scattering media. We choose to derive the transmission conditions from the continuity of the tangential and normal components of the field across the interface, see eqs. (1.7) - (1.10). Here, as the field $u(\mathbf{x})$ represents a single component of the electromagnetic field, we can reduce transmission conditions to scalar forms.

$$[n \times \mathbf{E}]_\Gamma = 0 \implies [u]_\Gamma = 0, \tag{1.28}$$

$$[n \cdot \mathbf{D}]_\Gamma = 0, \quad \mathbf{D} = \epsilon(\mathbf{x})\mathbf{E} \implies \left[\epsilon^{-1}(\mathbf{x}) \partial_n u\right]_\Gamma = 0 \quad \text{on } \Gamma. \tag{1.29}$$

where $[u]_\Gamma$ denotes the jump of u across Γ, and $\partial_n u$ is the derivative normal to Γ.

Here, we take advantage of the scalar approximation of the Silver-Müller radiation condition to ensure u^{sca} represents outgoing waves at infinity. This is also commonly called the Sommerfeld radiation condition:

$$\lim_{r\to\infty} r^{(d-1)/2}\left(\frac{\partial u^{\text{sca}}}{\partial r} - iku^{\text{sca}}\right) = 0, \tag{1.30}$$

where d is the spatial dimension.

We will use this Helmholtz-like equation, gathered below, throughout the remainder of this document as our chief problem of interest.

$$\begin{aligned}
&\text{Find } u := u^{\text{inc}} + u^{\text{sca}} \in H^1_{\text{loc}}(\mathbb{R}^d) \text{ such that:}\\
&\quad \nabla\cdot(\epsilon^{-1}\nabla u) + k^2 u = 0, \quad \text{in } \mathbb{R}^d,\\
&\quad [u]_\Gamma = 0, \quad [\epsilon^{-1}\partial_n u]_\Gamma = 0 \quad \text{on } \Gamma,\\
&\lim_{r\to\infty} r^{(d-1)/2}\left(\frac{\partial u^{\text{sca}}}{\partial r} - iku^{\text{sca}}\right) = 0,\\
&k\in\mathbb{R}^+, \quad d\in\{2,3\}.
\end{aligned} \tag{1.31}$$

1.2. Document Structure

This thesis is organized into five chapters, each contributing progressively to the exploration and solution of the scalar Helmholtz problem in the context of scattering by negative materials, with particular emphasis on the near-resonance regime.

In Chapter 2 we focus on canonical (unit ball) analytical solutions for both 2D and 3D scattering problems. Leveraging the geometrical symmetry, explicit expressions for interior and exterior fields are obtained through separation of variables. This chapter also discusses resonances for these problems, presenting the means of computing the various families of resonances and providing samples of their characteristics.

Chapter 3 focuses on the Finite Element Method for the 3D scalar Helmholtz equation. The standard Galerkin treatment is discussed, along with the necessary modifications (T-coercivity, mesh symmetry) for the sign-changing problem. We provide numerical tests to illustrate convergence in both the standard and near-resonance case. Observations from these experiments identify critical limitations of standard FEM approaches in capturing rapid field oscillations near resonances. To overcome these challenges, the chapter subsequently introduces an enriched FEM scheme, augmenting the standard polynomial basis with analytically derived plasmonic modes.

In Chapter 4, we develop the use of the Plane Wave Discontinuous Galerkin method for the sign-changing Helmholtz problem in 2D. We present a full derivation of the weak formulation, and show well-posedness, continuity, quasi-optimality and convergence in the L^2 norm. A battery of numerical tests is additionally conducted to demonstrate practical results for this method, in both the general sign-changing case and with application in the near-resonance regime.

Finally, we outline ongoing work, conclusions and potential extensions to the problem are presented in Chapter 5.

2. SCATTERING BY A PENETRABLE BALL

2.1. Chapter 2 Introduction

In this chapter, with the goal of studying the aforementioned plasmonic behaviour, we will we focus on scattering by a penetrable ball, a canonical problem in which analytical solutions can be constructed. By considering a unit ball as the scatterer, we leverage the symmetry of the geometry to employ separation of variables and classical special functions (Bessel functions, Hankel functions, and spherical harmonics) to construct explicit expressions for both the interior and exterior fields. These will allow us to write true solutions for the purpose of numerical comparison, for example against the finite element methods presented in chapter 3 and the Trefftz methods presented in chapter 4.

Furthermore, the analytical formulation naturally leads to a resonance problem. By investigating the families of resonances and their associated modes, we uncover phenomena that are analogous to the surface plasmon. In particular, the emergence of resonances with vanishingly small imaginary parts (i.e., near-real resonances) signals the presence of highly oscillatory, surface-localized behavior which are analogous to that of the surface plasmon. We present visualization of these resonances and their paired modes.

2.2. Analytical Solutions

To that end, we wish to find cases for which analytical solutions can be constructed. We define:

Definition 2.2.1

$$H^1_{\text{loc}}(\mathbb{R}^d) = \{u \in L^2_{\text{loc}}(\mathbb{R}^d) \mid \nabla u \in L^2_{\text{loc}}(\mathbb{R}^d)\},$$
$$d \in \{2, 3\}.$$

where ∇u is the weak gradient of u.

Remark 2.2.1. $L^2_{\text{loc}}(\mathbb{R}^d)$ denotes the space of locally square-integrable functions, and although H^1_{loc} functions need not have classical traces on Γ, the transmission conditions are enforced in the sense of traces on each side of Γ.

As in chapter 1, we consider the following Helmholtz-like problem:

$$\begin{aligned}
&\text{Find } u := u^{\text{inc}} + u^{\text{sca}} \in H^1_{\text{loc}}(\mathbb{R}^d) \text{ such that:}\\
&\quad \nabla \cdot (\epsilon^{-1}\nabla u) + k^2 u = 0, \quad \text{in } \mathbb{R}^d,\\
&\quad [u]_\Gamma = 0, \quad [\epsilon^{-1}\partial_n u]_\Gamma = 0 \quad \text{on } \Gamma,\\
&\text{Sommerfeld Radiation condition:}\\
&\quad \begin{cases} \lim_{r\to\infty} \sqrt{r}\left(\frac{\partial u^{\text{sca}}}{\partial r} - iku^{\text{sca}}\right) = 0, & \text{if } d = 2,\\ \lim_{r\to\infty} r\left(\frac{\partial u^{\text{sca}}}{\partial r} - iku^{\text{sca}}\right) = 0, & \text{if } d = 3, \end{cases}\\
&k \in \mathbb{R}^+, \quad d \in \{2, 3\}.
\end{aligned} \tag{2.1}$$

Here, Γ is the closed, smooth boundary of the scatterer $\mathcal{D}$. We will henceforth refer to Γ as the interface, as it defines the region separating the scatterer from the surrounding homogeneous space. For our incident field, we will take a propagative plane wave incident upon our scatterer.

Notably, ϵ is a material parameter (analogous to the permittivity of the media). Here we use an "ideal model" - a non-lossy approximation with no dissipation, leading to a real valued ϵ. Therefore, we scale this parameter such that it is 1 exterior to the scatterer and real-valued (although not necessarily positive!) interior to the scatterer.

The potentially sign-changing nature of the problem now results in additional difficulties in solving it. While this problem is quite classical and known to be well-posed in positive materials, for negative scatterers, this is not the case. In particular, coercivity (and subsequently our guarantee of well-posedness) is lost. However, the tool of T-coercivity has been developed (Bonnet-Ben Dhia et al. (2016); Carvalho et al. (2017); Bonnet-Ben Dhia et al. (2018),) to handle this deficiency. We may have well-posedness of the problem for $u \in H^1_{\text{loc}}(\mathbb{R}^d)$ following an additional constraint on ϵ, which depends on the geometry of the scatterer $\mathcal{D}$. We define $\mathfrak{C} \subset \mathbb{R}$ as the critical region inside of which we do not have a guarantee of well-posedness.[Bonnet-Ben Dhia et al. (2018)] In 2D this interval is proven to be optimal (all values of $\epsilon \in \mathfrak{C}$ lead to ill-posed problems), but in 3D this remains an open question for non-smooth boundaries. However,in this thesis, we will primarily focus on smooth boundaries. In particular, we note that in the case where $\mathcal{D}$ is the unit ball, $\mathfrak{C} = \{-1\}$.
We then simply require:

$$\epsilon(x) = \begin{cases} 1, & x \in E := \mathbb{R}^d \setminus \overline{\mathcal{D}} \\ \epsilon_m \in \mathbb{R}^* \setminus \mathfrak{C}, & x \in \mathcal{D} \end{cases}$$

Of note, this well-posedness is only for the continuous case. In chapter 3 we will discuss numerical approaches, where there are some mesh constraints on the problem which can be very costly, and we will discuss alternative methods by which these constraints may be lifted in later chapters.
In order to pursue analytical solutions, we first assume that indeed, the scatterer $\mathcal{D}$ is a unit ball, as we may not have analytical solutions in other cases. Taking advantage of this geometry, we then seek to write the general solution using separation of variables.

For simplicity, and without loss of generality in regards to the angle of incidence of the plane wave, we will assume the incident field is as follows for the purposes of computing analytical solutions:

$$u^{\text{inc}} = \begin{cases} e^{iky}, & d=2 \\ e^{ikz}, & d=3 \end{cases} \tag{2.2}$$

2.2.1. Notation for the Construction of Analytical Solutions

Let d be the dimension under consideration, with $d \in \{2,3\}$. Throughout this document, we choose the following convention for angular notation:

$$\theta := \begin{cases} \text{polar angle}, & d=2 \\ \text{azimuthal angle}, & d=3 \end{cases}$$

$$\phi := \text{polar angle}, \quad d=3$$

We choose this convention to align with common mathematical notation, however it may differ in some respects with various works, particularly those with an emphasis on physics.

Additionally, we note that we will make use of the Bessel functions throughout these sections, which we will denote by $J_l(\cdot), Y_l(\cdot), I_l(\cdot), H_l^{(1)}(\cdot)$, the Bessel, modified Bessel and Hankel functions of the first kind of order l and $\mathbf{j}_l^{(1)}(\cdot), \mathbf{y}_l^{(1)}(\cdot), \mathbf{h}_l^{(1)}(\cdot)$ the corresponding spherical Bessel-type functions.

Furthermore, we make note of the Jacobi-Anger plane wave expansions, which allow us to express our incident field of choice in an expansion of orthonormal basis functions. (A Fourier series in 2D and a spherical harmonic expansion in 3D) [Olver et al.

(2010)]:

$$e^{ikr\cos(\theta)} = \sum_{l=-\infty}^{\infty} i^l J_l(kr) e^{il\theta}, \quad d = 2 \tag{2.3}$$

$$e^{ikr\cos(\phi)} = \sum_{l=0}^{\infty} i^l \sqrt{(2l+1)(4\pi)}\, \mathbf{j}_l(kr) Y_l^0(\phi), \quad d = 3 \tag{2.4}$$

Here, $Y_l^0(\cdot)$ are the spherical harmonics of order l.

2.2.2. Separation of Variables in Polar Coordinates (2D)

We will proceed to construct families of solutions for two dimensions first. To find analytical solutions using separation of variables, we begin by formulating the full transmission problem involving two Helmholtz equations—one inside the scatterer and one outside. We consider the scatterer $\mathcal{D}$ to be the unit disk in $\mathbb{R}^2$ centered at the origin. The boundary of the scatterer is denoted by $\Gamma = \partial\mathcal{D}$.

We aim to find a function $u = u^{\text{inc}} + u^{\text{sca}} \in H^1_{\text{loc}}(\mathbb{R}^2)$ satisfying the following equations:

$$\nabla^2 u + k^2 u = 0, \qquad \text{in } E = \mathbb{R}^2 \setminus \overline{\mathcal{D}}, \tag{2.5}$$

$$\nabla^2 u + \epsilon_m k^2 u = 0, \qquad \text{in } \mathcal{D}, \tag{2.6}$$

with transmission conditions at the interface Γ:

$$[u]_\Gamma = 0, \tag{2.7}$$

$$\left[\epsilon^{-1}\frac{\partial u}{\partial n}\right]_\Gamma = 0, \tag{2.8}$$

and the Sommerfeld radiation condition for the scattered field u^{sca}:

$$\lim_{r\to\infty} \sqrt{r}\left(\frac{\partial u^{\text{sca}}}{\partial r} - iku^{\text{sca}}\right) = 0. \tag{2.9}$$

We note that the operator of interest is now the Laplacian $\nabla^2 u$. The Laplacian in polar coordinates is given by:

$$\nabla^2 u = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} \tag{2.10}$$

Therefore, the governing equation in the region outside the scatterer $(r > 1)$ is:

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} + k^2 u = 0 \tag{2.11}$$

Similarly, inside the scatterer $(r < 1)$, the equation is:

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} + \epsilon_m k^2 u = 0 \tag{2.12}$$

We seek solutions of the form:

$$u(r,\theta) = R(r)\Theta(\theta) \tag{2.13}$$

As we seek nonzero solutions, we may substitute this into the equation and divide both sides by $R(r)\Theta(\theta)$ to obtain:

$$\frac{1}{R}\frac{d}{dr}\left(r\frac{dR}{dr}\right) + \left(k^2 r^2\right) = -\frac{1}{\Theta}\frac{d^2\Theta}{d\theta^2} \tag{2.14}$$

Since the left-hand side is a function of r only, and the right-hand side is a function of θ only, both sides must be equal to a constant, which we denote by l^2 with $l \in \mathbb{Z}$ (We choose this value of l to ensure that a valid Bessel equation is produced, see below). Thus, we have:

$$\frac{1}{\Theta}\frac{d^2\Theta}{d\theta^2} + l^2 = 0 \tag{2.15}$$

$$\frac{d^2\Theta}{d\theta^2} + l^2\Theta = 0 \tag{2.16}$$

with solutions:

$$\Theta(\theta) = A_l e^{il\theta} + B_l e^{-il\theta} \tag{2.17}$$

Since the solution must be periodic in θ with period 2π, l must be an integer, and we can write the angular part as:

$$\Theta(\theta) = e^{il\theta} \tag{2.18}$$

For the radial part, we have:

$$\frac{d}{dr}\left(r\frac{dR}{dr}\right) + \left(k^2r^2 - l^2\right)R = 0 \tag{2.19}$$

This is Bessel's differential equation. The general solution are therefore, the Bessel functions. Recall $u = u^{\text{inc}} + u^{\text{sca}}$. The incident field can be expanded in explicitly in terms of the Bessel functions via the Jacobi-Anger expansion (2.3). Therefore, we must simply expand u^{sca}, which must conform to the outgoing radiation condition (2.9). In particular, those Bessel functions which satisfy the outgoing radiation condition are $H_l^{(1)}(kr)$, the Hankel functions of the first kind, which are outgoing cylindrical waves. Therefore, we may write the general solution:

$$R(r) = \underbrace{i^l J_l(kr)}_{u^{\text{inc}}\text{ contribution}} + \underbrace{C_l H_l^{(1)}(kr)}_{u^{\text{sca}}\text{ contribution}} \tag{2.20}$$

Remark 2.2.2. We only consider the Hankel functions of the first kind $H_l^{(1)}(kr)$ to satisfy the Sommerfeld radiation condition in 2D; those of the second kind $H_l^{(2)}$ would correspond to incoming waves.

Therefore, the general solution outside the scatterer ($r > 1$) is the superposition of the product of these Hankel functions with the angular component $e^{il\theta}$:

$$u(r,\theta) = \sum_{l=-\infty}^{\infty} \left[i^l J_l(kr) + C_l H_l^{(1)}(kr) \right] e^{il\theta} \tag{2.21}$$

Similarly, inside the scatterer ($r < 1$), the radial equation becomes:

$$\frac{d}{dr}\left(r\frac{dR}{dr} \right) + \left(\epsilon_m k^2 r^2 - l^2 \right) R = 0 \tag{2.22}$$

Depending on the sign of ϵ_m, the nature of the equation changes. For positive materials, the equation remains a standard Bessel differential equation, and its finite solutions are Bessel functions of the first kind J_l. In the case of negative materials, the equation becomes a modified Bessel differential equation due to the negative coefficient of the r^2 term. The solutions are modified Bessel functions of the first kind I_l. We discard the Bessel functions of the second kind (Y_l and K_l) because they are singular at $r = 0$, and we require the solution to be finite at the origin.

To capture both cases, we introduce the notation:

$$G_{l,\epsilon}(z) = \begin{cases} J_l\left(\sqrt{\epsilon_m}\, z \right), & \text{if } \epsilon_m > 0 \\ I_l\left(\sqrt{|\epsilon_m|}\, z \right), & \text{if } \epsilon_m < 0 \end{cases} \tag{2.23}$$

Therefore, the solution inside the scatterer can be uniformly written as:

$$u(r,\theta) = \sum_{l=-\infty}^{\infty} B_l\, G_{l,\epsilon}(kr)\, e^{il\theta} \tag{2.24}$$

Outside the scatterer ($r > 1$), the solution remains:

$$u^{\text{sca}}(r,\theta) = \sum_{l=-\infty}^{\infty} A_l\, H_l^{(1)}(kr)\, e^{il\theta} \tag{2.25}$$

where $H_l^{(1)}$.

Applying the continuity conditions at $r = 1$:

$$u_{\text{outside}}(1,\theta) = u_{\text{inside}}(1,\theta) \tag{2.26}$$

$$\epsilon^{-1} \left.\frac{\partial u_{\text{outside}}}{\partial r}\right|_{r=1} = \epsilon_m^{-1} \left.\frac{\partial u_{\text{inside}}}{\partial r}\right|_{r=1} \tag{2.27}$$

Substituting the expressions for u and equating coefficients for each l, we obtain:

$$A_l\, H_l^{(1)}(k) - B_l\, G_{l,\epsilon}(k) = -i^l\, J_l(k) \tag{2.28}$$

$$A_l \left[H_l^{(1)}\right]'(k) - \epsilon_m^{-1} B_l \left[G_{l,\epsilon}\right]'(k) = -i^l \left[J_l\right]'(k) \tag{2.29}$$

These equations form a system for each l, which can be represented in matrix form:

$$\underbrace{\begin{bmatrix} H_l^{(1)}(k) & -G_{l,\epsilon}(k) \\ \left[H_l^{(1)}\right]'(k) & -\epsilon_m^{-1}\left[G_{l,\epsilon}\right]'(k) \end{bmatrix}}_{\mathcal{M}_l} \underbrace{\begin{bmatrix} A_l \\ B_l \end{bmatrix}}_{\mathcal{A}_l} = \underbrace{\begin{bmatrix} -i^l\, J_l(k) \\ -i^l \left[J_l\right]'(k) \end{bmatrix}}_{\mathcal{J}_l} \tag{2.30}$$

Here, $\mathcal{M}_l$ represents a mapping from the coefficients associated with the scattered field to those of the incident field and is therefore the inverse of the scattering matrix of order l [Bennaceur et al. (1999)]. By inverting the system (2.30) for each l up to some choice of truncation, we solve for the unknown coefficients A_l and B_l, thus obtaining our (truncated) analytical solution as presented in the expressions (2.21) and (2.24).

2.2.3. Separation of Variables in Spherical Coordinates (3D)

We will now perform the same task, but for the 3D case. As before, to find analytical solutions using separation of variables, we begin by formulating the full transmission problem involving two Helmholtz equations—one inside the scatterer and one outside. We consider the scatterer $\mathcal{D}$ to be the unit sphere in $\mathbb{R}^3$ centered at the origin. The boundary of the scatterer is denoted by $\Gamma = \partial\mathcal{D}$. We aim to find a function $u = u^{\text{inc}} + u^{\text{sca}} \in H^1_{\text{loc}}(\mathbb{R}^3)$ satisfying the following equations:

$$\nabla^2 u + k^2 u = 0, \qquad \text{in } E = \mathbb{R}^3 \setminus \overline{\mathcal{D}}, \tag{2.31}$$

$$\nabla^2 u + \epsilon_m k^2 u = 0, \qquad \text{in } \mathcal{D}, \tag{2.32}$$

with transmission conditions at the interface Γ:

$$[u]_\Gamma = 0, \tag{2.33}$$

$$\left[\epsilon^{-1}\frac{\partial u}{\partial n}\right]_\Gamma = 0, \tag{2.34}$$

and the Sommerfeld radiation condition for the scattered field u^{sca}:

$$\lim_{r\to\infty} r\left(\frac{\partial u^{\text{sca}}}{\partial r} - iku^{\text{sca}}\right) = 0. \tag{2.35}$$

We note that the operator of interest is now the Laplacian $\nabla^2 u$ in spherical coordinates. The Laplacian in spherical coordinates is given by:

$$\nabla^2 u = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial u}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 u}{\partial\phi^2} \tag{2.36}$$

In three dimensions, we consider the Helmholtz equation in spherical coordinates. The Laplacian in spherical coordinates is given by:

$$\nabla^2 u = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r}\right) + \frac{1}{r^2\sin\phi}\frac{\partial}{\partial\phi}\left(\sin\phi\frac{\partial u}{\partial\phi}\right) + \frac{1}{r^2\sin^2\phi}\frac{\partial^2 u}{\partial\theta^2} \tag{2.37}$$

The governing equation in the region outside the scatterer ($r > 1$) is:

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r}\right) + \frac{1}{r^2\sin\phi}\frac{\partial}{\partial\phi}\left(\sin\phi\frac{\partial u}{\partial\phi}\right) + \frac{1}{r^2\sin^2\phi}\frac{\partial^2 u}{\partial\theta^2} + k^2 u = 0 \tag{2.38}$$

Similarly, inside the scatterer ($r < 1$), the equation is:

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r}\right) + \frac{1}{r^2\sin\phi}\frac{\partial}{\partial\phi}\left(\sin\phi\frac{\partial u}{\partial\phi}\right) + \frac{1}{r^2\sin^2\phi}\frac{\partial^2 u}{\partial\theta^2} + \epsilon_m k^2 u = 0 \tag{2.39}$$

We seek solutions of the form:

$$u(r,\phi,\theta) = R(r)\Phi(\phi)\Theta(\theta) \tag{2.40}$$

Given that we seek nonzero solutions, we may begin by substituting this into the outside equation ($\epsilon = 1$) and dividing both sides by $R(r)\Phi(\phi)\Theta(\theta)$, we obtain:

$$\frac{1}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) + k^2r^2 = -\frac{1}{\Phi\sin\phi}\frac{d}{d\phi}\left(\sin\phi\frac{d\Phi}{d\phi}\right) - \frac{1}{\Theta\sin^2\phi}\frac{d^2\Theta}{d\theta^2} \tag{2.41}$$

Since the left-hand side is a function of r only, and the right-hand side is a function of ϕ and θ only, both sides must be equal to a separation constant, which we denote by $l(l+1)$, where l is an integer. While this choice may not seem obvious, it is necessary (as in the 2D case) to lead to viable Bessel equations. Thus, we have:

$$\frac{1}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) + k^2r^2 = l(l+1) \tag{2.42}$$

and

$$\frac{1}{\Phi \sin\phi} \frac{d}{d\phi} \left(\sin\phi \frac{d\Phi}{d\phi} \right) + \left(l(l+1) - \frac{1}{\Theta \sin^2\phi} \frac{d^2\Theta}{d\theta^2} \right) = 0 \tag{2.43}$$

Next, we separate the variables ϕ and θ by introducing another separation constant m^2 (Again, chosen in this way to guarantee solutions of the resulting equation):

$$\frac{1}{\Theta} \frac{d^2\Theta}{d\theta^2} = -m^2 \tag{2.44}$$

This leads to solutions of the form:

$$\Theta(\theta) = e^{im\theta} \tag{2.45}$$

with m being an integer to ensure periodicity in θ.

The equation for $\Phi(\phi)$ becomes:

$$\frac{1}{\sin\phi} \frac{d}{d\phi} \left(\sin\phi \frac{d\Phi}{d\phi} \right) + \left(l(l+1) - \frac{m^2}{\sin^2\phi} \right) \Phi = 0 \tag{2.46}$$

which is the associated Legendre equation. The solutions are the associated Legendre functions $P_l^m(\cos\phi)$. Olver et al. (2010)

Combining $\Phi(\phi)$ and $\Theta(\theta)$, we obtain the spherical harmonics:

$$Y_l^m(\phi, \theta) = N_l^m P_l^m(\cos\phi) e^{im\theta} \tag{2.47}$$

where N_l^m is a normalization constant.

The radial equation simplifies to:

$$\frac{d}{dr} \left(r^2 \frac{dR}{dr} \right) + \left(k^2 r^2 - l(l+1) \right) R = 0 \tag{2.48}$$

which is the spherical Bessel differential equation. The general solution is:

$$R(r) = C_l \mathbf{j}_l(kr) + D_l \mathbf{y}_l(kr) \tag{2.49}$$

where $\mathbf{j}_l$ and $\mathbf{y}_l$ are the spherical Bessel functions of the first and second kind, respectively.

Since $\mathbf{y}_l(kr)$ is singular at $r = 0$, and we require the solution to be finite at $r = 0$, we discard $\mathbf{y}_l$ inside the scatterer.

Similarly to 2D, outside the scatterer, we use spherical Hankel functions (which are linear combinations of the Bessel functions of first and second kind) to represent outgoing waves satisfying the radiation condition. Specifically, analagous to 2D, the spherical Hankel function of the first kind, $\mathbf{h}_l^{(1)}(kr)$, represents an outgoing spherical wave. (Again, no need for the second kind, see Remark 2.2.2)

Therefore, the general solution of the scattered field outside the scatterer ($r > 1$) is the superposition of the product of these outgoing spherical waves and spherical harmonics varying over l and m:

$$u^{\text{sca}}(r, \phi, \theta) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} A_{l,m} \mathbf{h}_l^{(1)}(kr) Y_l^m(\phi, \theta), \tag{2.50}$$

where the incident field contribution can be obtained from the Jacobi-Anger expansion (2.54).

Inside the scatterer ($r < 1$), while the angular components are the same, the radial equation becomes:

$$\frac{d}{dr}\left(r^2 \frac{dR}{dr}\right) + \left(\epsilon_m k^2 r^2 - l(l+1)\right) R = 0 \tag{2.51}$$

As before, depending on the sign of ϵ_m, the nature of the solution changes. When $\epsilon_m > 0$, the equation remains a standard spherical Bessel differential equation, and its finite solutions are spherical Bessel functions of the first kind $\mathbf{j}_l$. When $\epsilon_m < 0$, the equation becomes a modified spherical Bessel differential equation due to the negative coefficient of the r^2 term. The solutions are modified spherical Bessel functions of the first kind $\mathbf{i}_l$.

Again, we discard the spherical Bessel functions of the second kind because they are singular at $r = 0$.

To capture both cases, we introduce the notation:

$$\mathbf{g}_{l,\epsilon}(z) = \begin{cases} \mathbf{j}_l\left(\sqrt{\epsilon_m}\, z\right), & \text{if } \epsilon_m > 0 \\ \mathbf{i}_l\left(\sqrt{|\epsilon_m|}\, z\right), & \text{if } \epsilon_m < 0 \end{cases} \tag{2.52}$$

Therefore, the solution inside the scatterer can be uniformly written as:

$$u(r, \phi, \theta) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} B_{l,m}\, \mathbf{g}_{l,\epsilon}(kr)\, Y_l^m(\phi, \theta) \tag{2.53}$$

For our incident field $u^{\text{inc}} = e^{ikz}$, we utilize the spherical harmonic expansion (a special case of the Jacobi-Anger expansion) [Olver et al. (2010)]:

$$e^{ikz} = \sum_{l=0}^{\infty} i^l \sqrt{(2l+1)(4\pi)}\, \mathbf{j}_l(kr) Y_l^0(\phi, \theta) \tag{2.54}$$

Here, only the $m = 0$ terms appear due to the azimuthal symmetry of the incident plane wave along the z-axis.

Substituting the expressions for u^{inc}, u^{sca}, and u into the boundary conditions at $r = 1$:

$$[u]_{r=1} = 0 \tag{2.55}$$

$$[\epsilon^{-1} \partial_r u]_{r=1} = 0 \tag{2.56}$$

we equate the coefficients for each l (since the spherical harmonics form an orthogonal basis). This yields:

$$A_l \mathbf{h}_l^{(1)}(k) - B_l \mathbf{g}_{l,\epsilon}(k) = -i^l \sqrt{(2l+1)(4\pi)}\, \mathbf{j}_l(k) \tag{2.57}$$

$$A_l \left[\mathbf{h}_l^{(1)}\right]'(k) - \epsilon_m^{-1} B_l \left[\mathbf{g}_{l,\epsilon}\right]'(k) = -i^l \sqrt{(2l+1)(4\pi)}\, [\mathbf{j}_l]'(k) \tag{2.58}$$

These equations form a system for each l, which can be represented in matrix form:

$$\underbrace{\begin{bmatrix} \mathbf{h}_l^{(1)}(k) & -\mathbf{g}_{l,\epsilon}(k) \\ \left[\mathbf{h}_l^{(1)}\right]'(k) & -\epsilon_m^{-1}\left[\mathbf{g}_{l,\epsilon}\right]'(k) \end{bmatrix}}_{\mathcal{M}_l} \underbrace{\begin{bmatrix} A_l \\ B_l \end{bmatrix}}_{\mathcal{A}_l} = \underbrace{\begin{bmatrix} -i^l \sqrt{(2l+1)(4\pi)}\, \mathbf{j}_l(k) \\ -i^l \sqrt{(2l+1)(4\pi)}\, [\mathbf{j}_l]'(k) \end{bmatrix}}_{\mathcal{J}_l} \tag{2.59}$$

Here, $\mathcal{M}_l$ represents a mapping from the coefficients associated with the scattered field to those of the incident field and is therefore the inverse of the scattering matrix of order l [Bennaceur et al. (1999)]. By inverting the systems (2.30) and (2.59) for each l up to some choice of truncation, we solve for the unknown coefficients A_l and B_l, thus obtaining our (truncated) analytical solution as presented in the expressions for (2.24), (2.21), (2.53) and (2.50). For clarity we rewrite these below:

$$u(r,\theta) = \begin{cases} \displaystyle\sum_{l=-\infty}^{\infty} B_l\, G_{l,\epsilon}(kr)\, e^{il\theta}, & r < 1, \\ \displaystyle\sum_{l=-\infty}^{\infty} (A_l\, H_l^{(1)}(kr) + i^l J_l(kr))\, e^{il\theta}, & r \geq 1. \end{cases} \tag{2.60}$$

$$u(r,\theta,\phi) = \begin{cases} \displaystyle\sum_{l=0}^{\infty} B_l\, \mathbf{g}_{l,\epsilon}(kr)\, Y_l^0(\phi,\theta), & r < 1, \\ \displaystyle\sum_{l=0}^{\infty} (A_l\, \mathbf{h}_l^{(1)}(kr) + i^l \sqrt{(2l+1)(4\pi)}\, \mathbf{j}_l(kr)) Y_l^0(\phi,\theta), & r \geq 1. \end{cases} \tag{2.61}$$

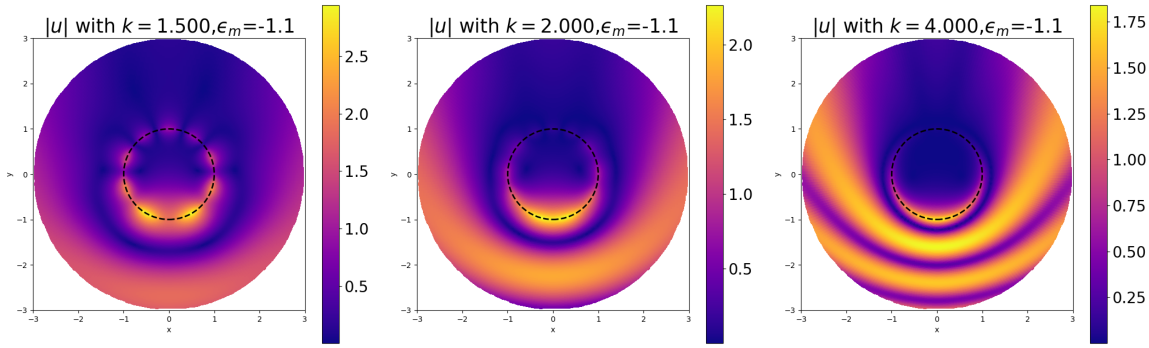


Figure 2.1. Absolute Analytic Solutions (2D) with $\epsilon_m = -1.1$ for various k

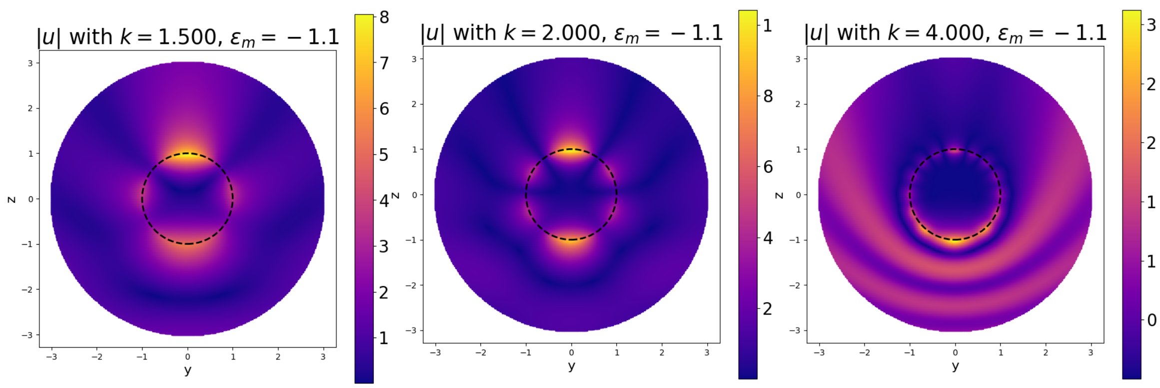
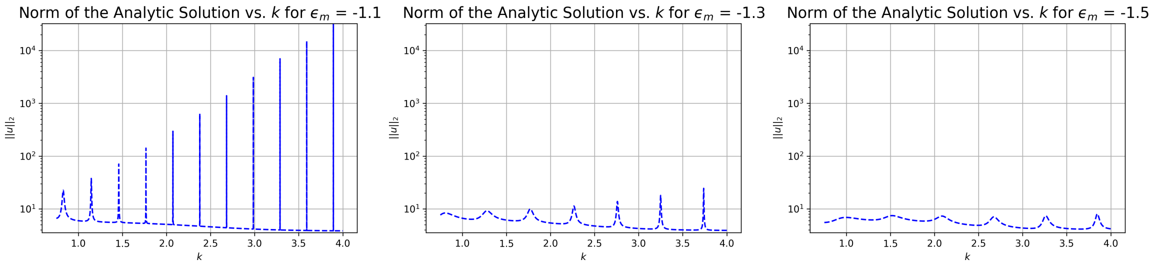


Figure 2.2. Absolute Analytic Solutions (3D) with $\epsilon_m = -1.3$ for various k.

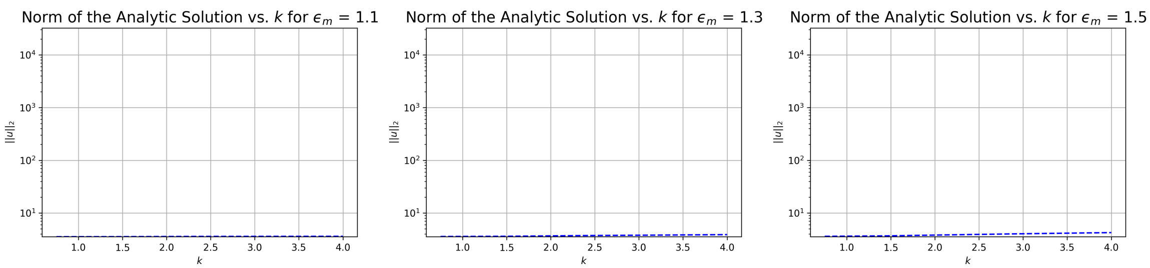
2.2.4. Analytical Solution Plots

With these analytic solutions in hand, we can truncate our series and domain and plot them on that domain. (Here we do so for a disk of radius 3)

Additionally, we plot 3D solutions in a disk-shaped slice in the $y - z$ plane: Additionally, we can easily use the orthogonality of the angular components to find the 2-norm of these solutions. Plotting the norms as functions of k demonstrates a "spiky" behavior for negative ϵ_m, causing the norms to increase rapidly and substantially for certain values of k. See Figure 2.3a. We note that these spikes are increasing linearly in a logarithmic plot, implying exponential growth of the norms at these values of k. These spikes are also more extreme at lower wave numbers k as $\epsilon_m \to 0$,



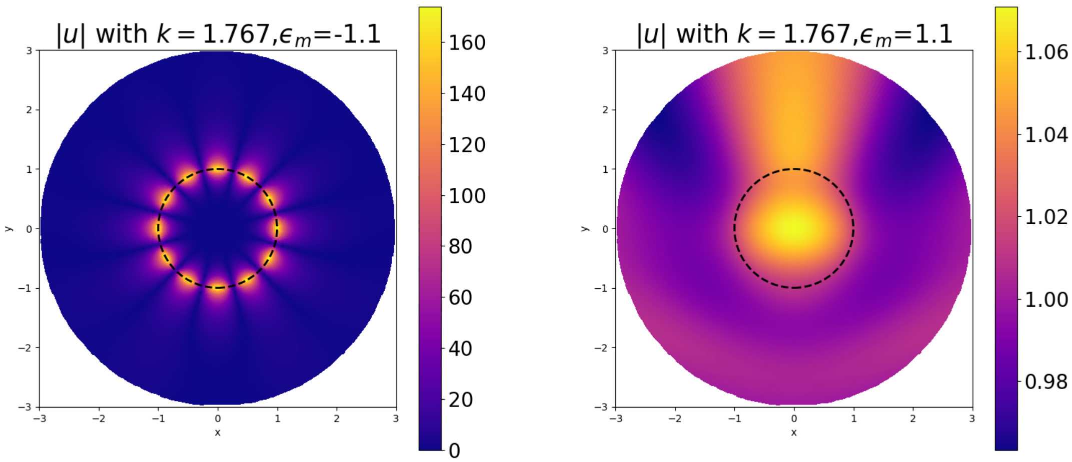
(a) $\epsilon_m \in \{-1.1, -1.3, -1.5\}$



(b) $\epsilon_m \in \{1.1, 1.3, 1.5\}$

Figure 2.3. Log plot of Norm of Analytic Solutions vs. k for varying ϵ_m values.

Naturally, this motivates us to locate where these spikes are and why they might occur. Firstly, we plot the behavior of the solution on the most extreme spike, and we see much higher-order behavior with a strong localization near the interface. See Figure 2.4a. Notably, both the spikes and the extreme behavior does not occur for positive ϵ_m, where our norm increases steadily with respect to k (thus our continued interest in "negative" materials - note the lack of large scale behaviour in the positive material, see Figure 2.4b). As we expect anomalous behaviour for our solution on or near resonances, this gives us our next point of inquiry.



(a) $\epsilon_m = -1.1$ (b) $\epsilon_m = 1.1$

Figure 2.4. Norm of Analytic Solution with $k = 1.766$ for $\epsilon_m = \pm 1.1$

2.3. Resonances

As we wish to investigate sudden larger-scale phenomena in our solutions, we now look for resonances, which requires solving the homogenous problem:

$$\begin{aligned}
&\text{Find } (\lambda, u) \in \mathbb{C} \times H^1_{\text{loc}}(\mathbb{R}^d) \text{ such that:} \\
&\quad \nabla \cdot (\epsilon^{-1} \nabla u) + \lambda^2 u = 0, \qquad \text{in } \mathbb{R}^d \\
&\quad [u]_\Gamma = 0, \quad [\epsilon^{-1} \partial_n u]_\Gamma = 0 \qquad \text{on } \Gamma \\
&\text{Sommerfeld Radiation condition:} \\
&\quad \begin{cases} \lim_{r\to\infty} \sqrt{r} \left(\frac{\partial u}{\partial r} - i\lambda u \right) = 0, & \text{if } d = 2, \\ \lim_{r\to\infty} r \left(\frac{\partial u}{\partial r} - i\lambda u \right) = 0, & \text{if } d = 3, \end{cases} \\
&\quad \lambda \in \mathbb{R}^+, \quad d = 2, 3
\end{aligned} \tag{2.62}$$

Resonances and resonant modes (λ, u) are solutions of the homogeneous problem, so we do not include an incident field. Of particular note, such solutions may not decay at infinity, e.g. outgoing cylindrical waves with a complex argument behave as:

[Olver et al. (2010)]

$$H_l^{(1)}(\lambda r) \sim \sqrt{\frac{2}{\pi|\lambda r|}} e^{i\left(\Re(\lambda r) - \frac{l\pi}{2} - \frac{\pi}{4}\right)} e^{-\Im(\lambda r)}, \quad r \gg 1 \tag{2.63}$$

and the exponential component associated to $\Im(\lambda r)$ is not guaranteed to decrease in the far-field.

We define the resolvent associated to this problem by

$$\mathcal{R}(\lambda) = (\nabla \cdot (\epsilon^{-1}\nabla) - \lambda^2)^{-1} \tag{2.64}$$

Given this, we can find the complex resonances:

Definition 2.3.1: Resonance

A complex number $\lambda \in \mathbb{C} \setminus (-\infty, 0]$ is called a *resonance* if it is a pole of the resolvent operator

$$\mathcal{R}(\lambda) = \left(\nabla \cdot \left(\epsilon^{-1}\nabla\right) - \lambda^2\right)^{-1}.$$

Equivalently, for some integer l, λ is a resonance if

$$\det \mathcal{M}_l(\lambda) = 0,$$

where $\mathcal{M}_l(\lambda)$ denotes the scattering matrix corresponding to the lth mode.

Specifically, we define the set of resonances in 2D as

$$\mathcal{R}(\epsilon_m) = \bigcup_{l \in \mathbb{Z}} \{\lambda \in \mathbb{C} \setminus (-\infty, 0] \mid \det(M_{l,d=2}(\lambda)) = 0\} \tag{2.65}$$

Here we reference the inverse of the scattering matrices (2.30) and (2.59). We recall the 2D matrix below:

$$\mathcal{M}_{l,d=2}(\lambda) = \begin{bmatrix} -H_l^{(1)}(\lambda) & G_{l,\epsilon}(\lambda) \\ -H_l^{(1)\prime}(\lambda) & G'_{l,\epsilon}(\lambda) \end{bmatrix} \tag{2.66}$$

In order to find these resonances, we must therefore find roots of the determinant of the family of matrices $\mathcal{M}_l$, corresponding to poles of the scattering matrices [Bennaceur et al. (1999)]. The determinant of any one of these matrices, for any given l, can be written:

$$\det(\mathcal{M}_{l,d=2})(\lambda) = -G'_{l,\epsilon}(\lambda)H_l^{(1)}(\lambda) + H_l^{(1)\prime}(\lambda)G_{l,\epsilon}(\lambda) \tag{2.67}$$

and we seek to solve

$$\det(\mathcal{M}_l)(\lambda) = 0 \tag{2.68}$$

A similar process can be done for the 3D problem, including defining the resonances as:

$$\mathcal{R}(\epsilon_m) = \bigcup_{l\in\mathbb{Z}^+} \{\lambda \in \mathbb{C} \setminus \mathbb{R}^- \mid \det(M_{l,d=3}(\lambda)) = 0\} \tag{2.69}$$

With associated matrices,

$$\mathcal{M}_{l,d=3} = \begin{bmatrix} -\mathbf{h}_l^{(1)}(\lambda) & g_{l,\epsilon}(\lambda) \\ -\mathbf{h}_l^{(1)\prime}(\lambda) & g'_{l,\epsilon}(\lambda) \end{bmatrix} \tag{2.70}$$

And determinants:

$$\det(\mathcal{M}_{l,d=3})(\lambda) = -\mathbf{g}'_{l,\epsilon}(\lambda)\mathbf{h}_l^{(1)}(\lambda) + \mathbf{h}_l^{(1)\prime}(\lambda)\mathbf{g}_{l,\epsilon}(\lambda) \tag{2.71}$$

In order to solve these problems, we compute these roots numerically in Python using the contour integration methods described in [Kravanja & Barel (2000)].

2.3.1. Resonant Modes

Given the homogeneous problem (2.62), our solutions (resonant modes) are similar to the scattering problem, as given in (2.24), (2.21), (2.53) and (2.50), with the exception that we do not include the component associated with the incident field. We impose a coefficient to ensure continuity of the solution across the interface, and as these are resonant modes, they will remain solutions up to any constant multiple. With this in mind, for some resonance λ_l, we can represent our solution up to a constant multiplier in 2d as follows:

$$u_l(r,\theta) = \begin{cases} I_l\big(\sqrt{|\epsilon_m|}\,\lambda_l\, r\big)\, e^{i\,l\,\theta}, & r < 1, \quad \epsilon_m < 0, \\ \dfrac{I_l\big(\sqrt{|\epsilon_m|}\,\lambda_l\big)}{H_l\big(\lambda_l\big)}\, H_l\big(\lambda_l\, r\big)\, e^{i\,l\,\theta}, & r \geq 1, \quad \epsilon_m < 0. \end{cases} \tag{2.72}$$

$$u_l(r,\theta) = \begin{cases} J_l\big(\sqrt{\epsilon_m}\,\lambda_l\, r\big)\, e^{i\,l\,\theta}, & r < 1, \quad \epsilon_m \geq 0, \\ \dfrac{J_l\big(\sqrt{\epsilon_m}\,\lambda_l\big)}{H_l\big(\lambda_l\big)}\, H_l\big(\lambda_l\, r\big)\, e^{i\,l\,\theta}, & r \geq 1, \quad \epsilon_m \geq 0. \end{cases} \tag{2.73}$$

And in 3d:

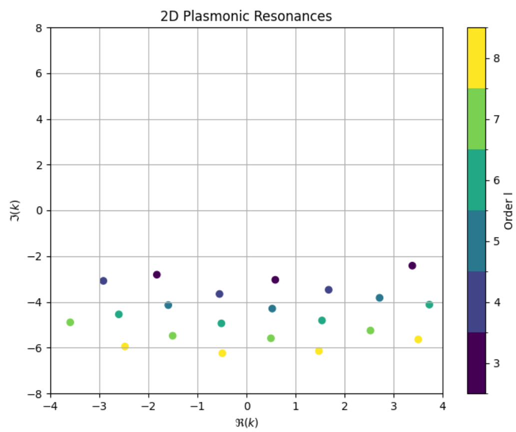
$$u_l(r,\phi) = \begin{cases} \mathbf{i}_l\big(\sqrt{|\epsilon_m|}\,\lambda_l\, r\big)\, Y_l^0(\phi), & r < 1, \quad \epsilon_m < 0, \\ \dfrac{\mathbf{i}_l\big(\sqrt{|\epsilon_m|}\,\lambda_l\big)}{\mathbf{h}_l\big(\lambda_l\big)}\, \mathbf{h}_l\big(\lambda_l\, r\big)\, Y_l^0(\phi), & r \geq 1, \quad \epsilon_m < 0, \end{cases} \tag{2.74}$$

$$u_l(r,\phi) = \begin{cases} \mathbf{j}_l\big(\sqrt{\epsilon_m}\,\lambda_l\, r\big)\, Y_l^0(\phi), & r < 1, \quad \epsilon_m \geq 0, \\ \dfrac{\mathbf{j}_l\big(\sqrt{\epsilon_m}\,\lambda_l\big)}{\mathbf{h}_l\big(\lambda_l\big)}\, \mathbf{h}_l\big(\lambda_l\, r\big)\, Y_l^0(\phi), & r \geq 1, \quad \epsilon_m \geq 0. \end{cases} \tag{2.75}$$

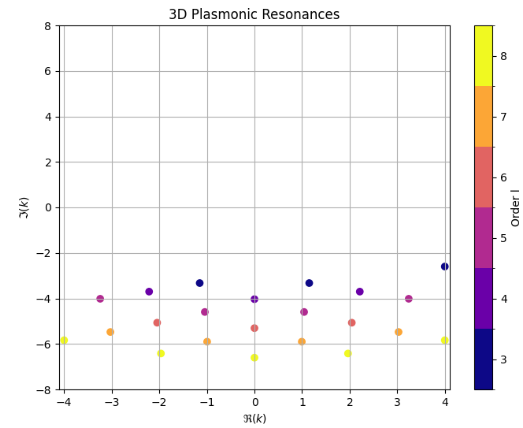
Note that for any such u_l, its complex complement $\bar{u}_l$ is also a resonant mode.

2.3.2. Plots of the Resonances and Associated Modes

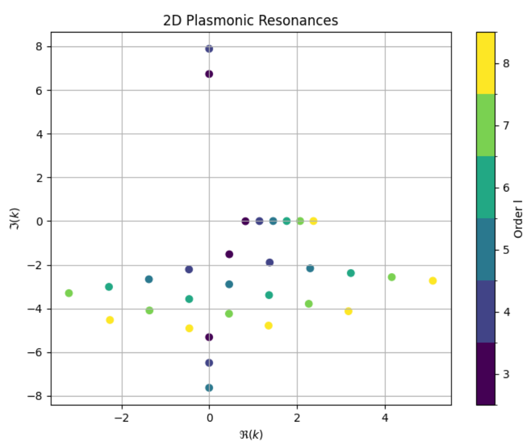
To demonstrate the behaviour of the resonances, we select $\epsilon_m = \pm 1.1$ and for $3 \leq l \leq 8$ we compute and plot the values of the resonances in the complex plane.



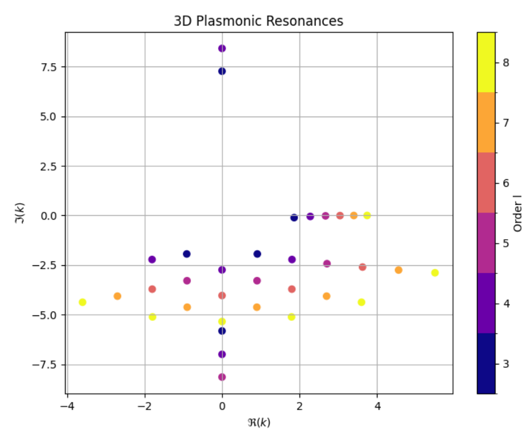
(a) 2D Resonances $\epsilon_m = 1.1$



(b) 3D Resonances $\epsilon_m = 1.1$



(c) 2D Resonances $\epsilon_m = -1.1$



(d) 3D Resonances $\epsilon_m = -1.1$

Figure 2.5. λ_l in $\mathbb{C}$ with $3 \leq l \leq 8$

We note the appearance of several different families among the resonances for negative scatterers. We seek in particular those which correspond to the phenomena of the surface plasmon. As we are practically interested in the values of $k \in \mathbb{R}$, we observe in closer detail the family of resonances which are close to the real axis.

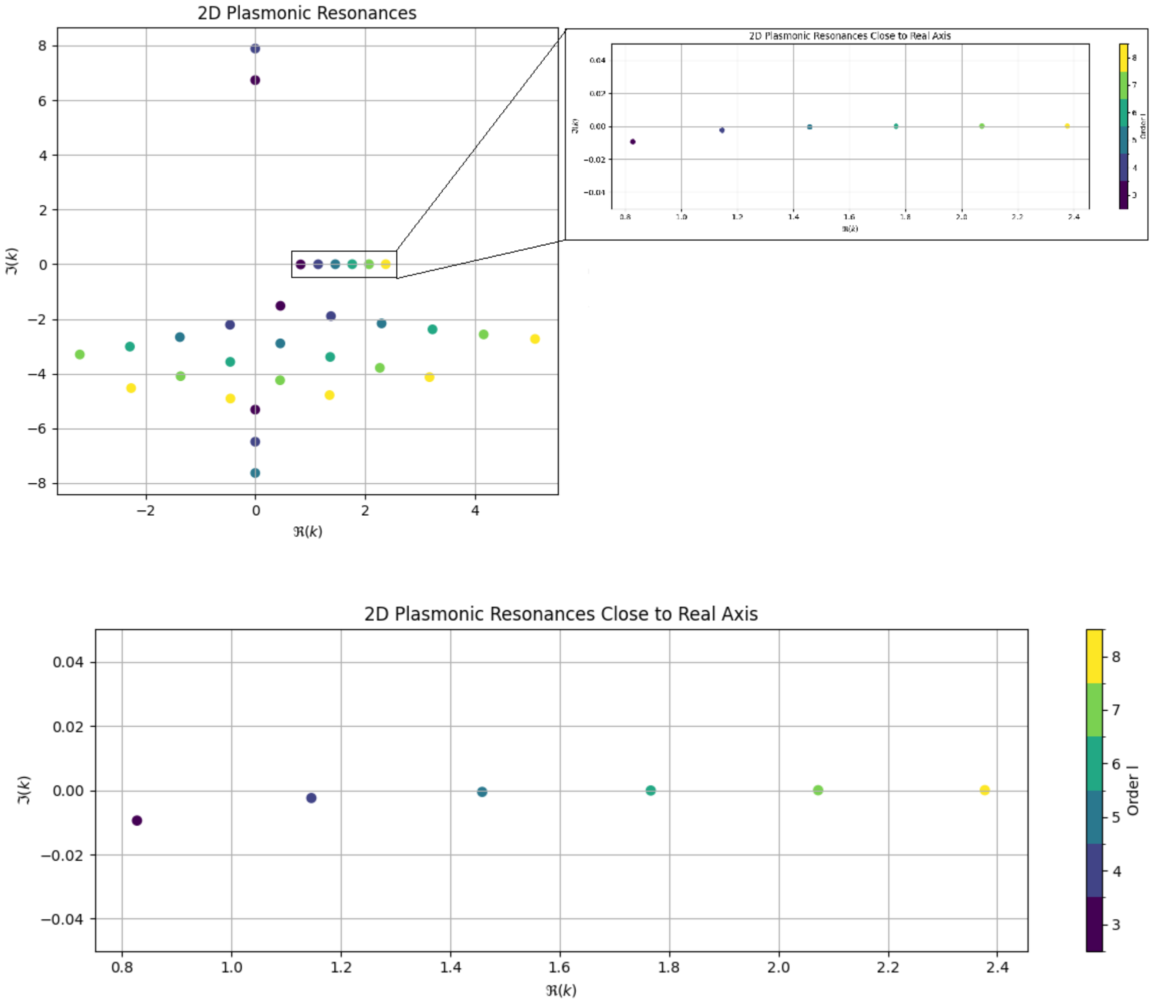


Figure 2.6. Family of λ_l with $\Im(\lambda) \ll 1$

These resonances have a vanishingly small imaginary component, thus we may approach them with real-valued wave numbers k for which $k = \Re(\lambda)$.

Given this, we can plot the resonant modes for various families of resonant modes in the negative scatterer case:

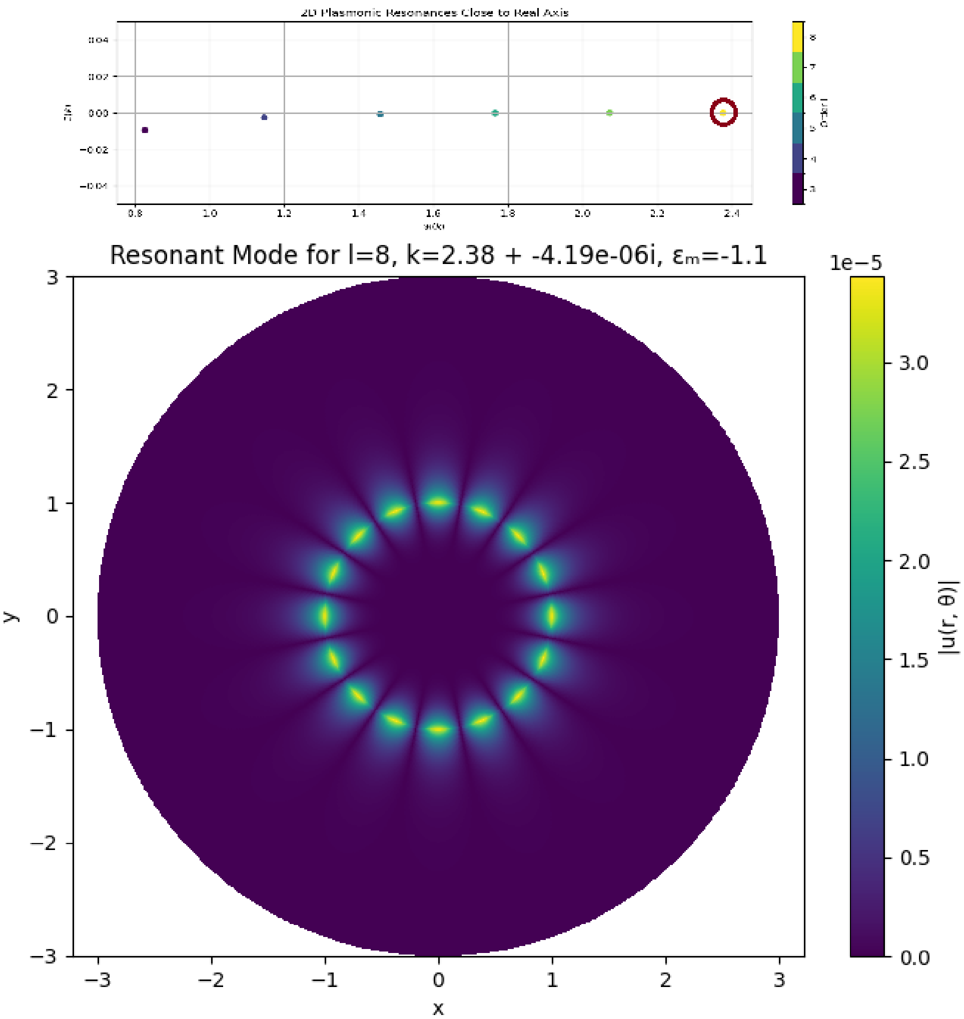


Figure 2.7. $|u_8|$ Resonant Mode with near-real k

We note that the associated resonant modes of the family close to the real axis display the oscillatory behaviour local to the surface of the scatterer we are seeking.

Plotting samples of other families (in this case, one resonance with substantial real and imaginary components and one which lies very close to the imaginary axis), we can see other behaviours. However, given $k \in \mathbb{R}$ as a restriction, we do not have wave numbers which lie near these resonances and imitate this behaviour.

We furthermore note that such behaviours do not occur in the case of positive ϵ, where the resonances are removed from the real axis and therefore cannot

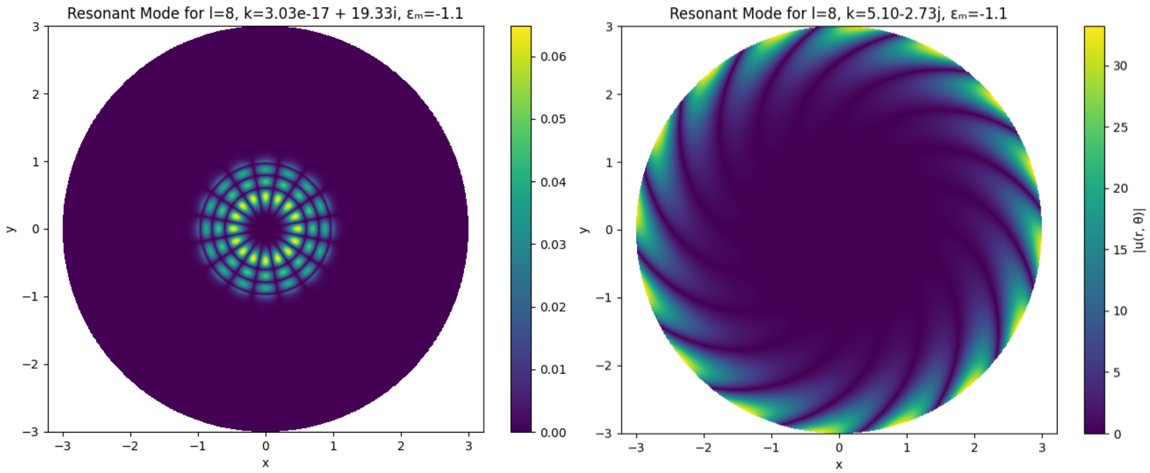


Figure 2.8. (a) $|u_8|$ 'Inner' Resonant Mode, (b) $|u_8|$ 'Outer' Resonant Mode

be accessed with $k \in \mathbb{R}$, as seen in Figure 2.5. Therefore, when investigating the near-resonance case, we must consider $\epsilon_m < 0$.

We may consider the behaviour of the analytical solution near these families. In Figure 2.9 we choose a 3D solution, fix $l = 10$ and plot the L^2 norm of the gradient of the analytical solution on a truncated spere about the scatterer for varying ϵ_m. We note that the associated L^2 norm grows rapidly as $\epsilon_m \to -1$. We discuss the reason for this phenomenon in greater detail in Theorem 4.2.1, but the L^2 norm of ∇u growing without bound signals that u is approaching the limit of H^1 regularity. Therefore as the solutions grow more extreme in gradient the will exhibit deteriorated convergence for h^1 conforming methods in the region near $\epsilon_m = -1$.

We may also consider in 3D, the case $0 > \epsilon_m > -1$. In this instance, we see the formation of resonances near the real axis, which does not occur for the 2D case. We refer to Figure 2.10 and note the presence of resonances on the real axis only in the 3D case. We do not investigate this intermediate regime further in the thesis, since our primary interest lies in the plasmonic regime $\epsilon_m < -1$, where surface-plasmon resonances arise in both 2D and 3D and admit a unified analytical and numerical

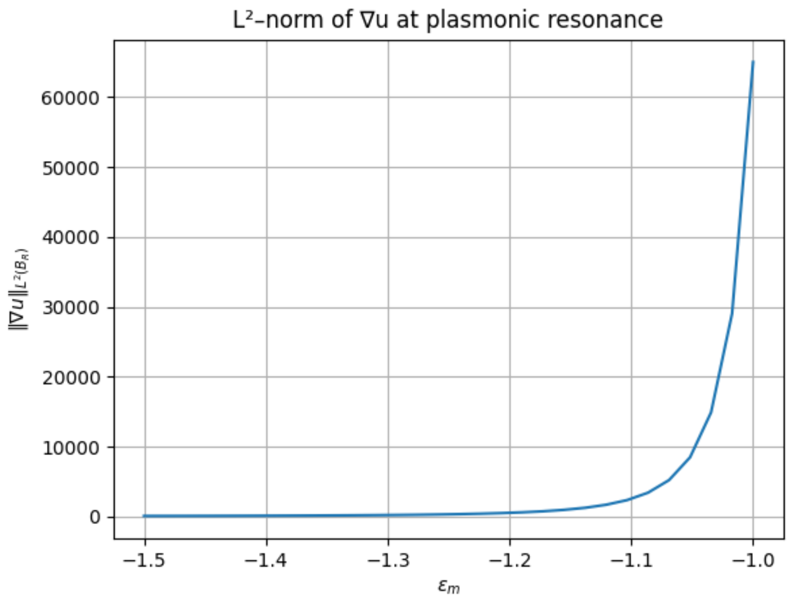
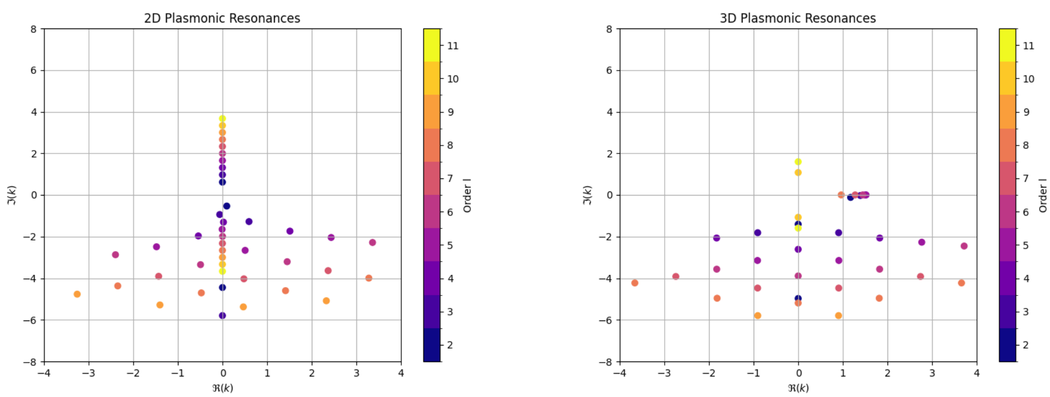


Figure 2.9. L^2 norm of ∇u on a truncated sphere (with order $l = 10$) as the material parameter ϵ_m varies.

treatment.

Having developed the analytical framework and characterized the resonances and resonant modes associated with scattering by a penetrable ball, we now turn our attention to the numerical approximation of these problems. In the next chapter, we introduce finite element methods (FEM) tailored to handle the challenges posed by plasmonic resonances. We will see how the insights gained from the analytical solutions guide the design of enriched numerical schemes that can better capture the highly localized oscillations and rapid decay inherent in plasmonic behavior.



(a) 2D Resonances $\epsilon_m = -0.9$ (b) 3D Resonances $\epsilon_m = -0.9$.

Figure 2.10. λ_l in $\mathbb{C}$ with $2 \leq l \leq 11$

3. FEM METHODS FOR PLASMONIC PROBLEMS

3.1. Chapter 3 Introduction

This chapter is dedicated to developing and analyzing two complementary finite element methodologies for solving a scattering problem in three dimensions where the material parameters may be negative. We focus primarily on the 3D case due to the vastly increased computational costs, which gives rise to the need for alternative methods for dealing with the plasmonic case. In the first part, we present the standard finite element method (FEM) formulation that employs a polynomial basis to approximate the solution of a transmission problem in a truncated domain. The formulation is derived by first constructing a weak form from the underlying differential equation, incorporating appropriate boundary conditions to enforce the Sommerfeld radiation condition. We discuss the discretization of this weak formulation using hierarchical polynomial basis functions defined on a symmetric mesh. Moreover, we explore conditions for well-posedness via T-coercivity arguments (giving us theoretical convergence of the FEM method) and demonstrate numerical convergence in regimes away from resonance.

However, when the scattering problem approaches a plasmonic resonance, the standard FEM encounters significant challenges in practice. Near resonance, the physical solution exhibits rapid oscillations and strong localization (i.e., the surface plasmon) that are difficult to capture with a basis consisting solely of low-order polynomials. In such regimes, even a highly refined mesh may require an impractically large number of degrees of freedom to resolve the fine-scale features of the solution.

Motivated by these limitations, the second part of the chapter introduces an enriched finite element framework. In this approach, the standard approximation space is augmented with the resonant plasmon mode—a function that captures the dominant high-frequency behavior associated with the resonance. By incorporating this mode explicitly, the enriched FEM can more accurately resolve the near-resonance behavior without the need for excessive mesh refinement. We derive the modified weak formulation in an extended space and obtain an augmented system that couples the standard H^1 component with the plasmonic enrichment. The resulting formulation not only improves the approximation quality for plasmonic problems, particularly in capturing the plasmonic behavior.

Throughout this chapter, we provide detailed derivations of the weak forms, a discussion of the associated boundary conditions, and numerical experiments that illustrate the performance of both the standard and enriched FEM approaches. The results highlight that while the standard FEM works reliably in non-resonant regimes, the enriched method is better able to capture the surface plasmon.

3.2. Galerkin Approach

3.2.1. FEM Intro

Recall we consider the following problem:

$$
\begin{aligned}
&\text{Find } u := u^{\text{inc}} + u^{\text{sca}} \in H^1_{\text{loc}}(\mathbb{R}^3) \text{ such that:}\\
&\quad \nabla \cdot (\epsilon^{-1} \nabla u) + k^2 u = 0, \quad \text{in } \mathbb{R}^3,\\
&\quad [u]_\Gamma = 0, \quad [\epsilon^{-1} \partial_n u]_\Gamma = 0 \quad \text{on } \Gamma,\\
&\lim_{r\to\infty} r\left(\frac{\partial u^{\text{sca}}}{\partial r} - iku^{\text{sca}}\right) = 0\\
&k \in \mathbb{R}^+.
\end{aligned}
\tag{3.1}
$$

We take $\mathfrak{C}$ to be the critical region inside of which we do not have a guarantee of well-posedness. Bonnet-Ben Dhia et al. (2018) The choice of this region may vary based on the shape of the material scatterer $\mathcal{D}$. We note that in the case where $\mathcal{D}$ is the unit sphere, $\mathfrak{C} = \{-1\}$. For further details, see Lemma 3.2.1. We then choose:

$$\epsilon(x) = \begin{cases} 1, & x \in E := \mathbb{R}^d \setminus \overline{\mathcal{D}} \\ \epsilon_m \in \mathbb{R}^* \setminus \mathfrak{C}, & x \in \mathcal{D} \end{cases}$$

We wish to approximate the solution of (3.1) using Finite Element Method (FEM) and compare with analytical results from section §2.2. As mentioned in the introduction for this chapter, we will be concerned primarily with the 3D problem (although a similar approach in 2D is not unworkable). With that aim, we artificially truncate the domain from $\mathbb{R}^3$ to a larger concentric radiation sphere Ω about $\mathcal{D}$, with radius R, and we use a mapping to impose the Sommerfeld radiation condition.

3.2.2. Weak form

By taking the L^2 product with an arbitrary test function $v \in \mathcal{V}(\Omega)$ and using the first Green's identity, the weak form of the problem set in Ω can be derived:

$$\begin{aligned} &\text{Find } u \in H^1(\Omega) \text{ such that, } \forall v \in \mathcal{V}(\Omega): \\ &-\int_\Omega \epsilon^{-1} \nabla u \nabla \overline{v} \; d\Omega + \int_{\partial\Omega} \epsilon^{-1} \partial_n u \overline{v} \; dS + \int_\Omega k^2 u \overline{v} \; d\Omega = 0 \\ \iff &-\int_\Omega \epsilon^{-1} \nabla u \nabla \overline{v} \; d\Omega + \int_{\partial\Omega} \epsilon^{-1} \partial_n u^{\text{sca}} \overline{v} \; dS + \int_\Omega k^2 u \overline{v} \; d\Omega = -\int_{\partial\Omega} \epsilon^{-1} \partial_n u^{\text{inc}} \overline{v} \; dS \end{aligned} \tag{3.2}$$

We now explore two potential boundary conditions to be applied for this problem. The first is a local, first order impedance boundary condition, defined by:

$$\partial_n u - iku = (-iku^{\text{inc}} + \partial_n u^{\text{inc}}) \quad \text{on } \partial\Omega$$

If we plug these terms into our weak form, and rearranging the terms to isolate those involving u on the left-hand side and the incident field contributions on the right-hand side, we obtain the weak formulation:

$$
\begin{aligned}
&-\int_{\Omega} \epsilon^{-1} \nabla u \cdot \nabla \overline{v}\, d\Omega + \int_{\Omega} k^2 u\, \overline{v}\, d\Omega + \int_{\partial\Omega} \epsilon^{-1} (ik\, u)\, \overline{v}\, dS \\
&= \int_{\partial\Omega} \epsilon^{-1} \left(ik\, u^{\mathrm{inc}} - \partial_n u^{\mathrm{inc}}\right) \overline{v}\, dS, \quad \forall v \in H^1(\Omega).
\end{aligned}
\tag{3.3}
$$

Alternatively, we might use a higher-order nonlocal approach, called the Dirichlet-to-Neumann (DtN) approach, (see J. Melenk & Sauter (2010)) wherein we expand the scattered field in terms of the spherical harmonics over the boundary $\partial\Omega$, and demand that the radial component be a spherical Hankel function of the first kind, corresponding to an outgoing spherical wave, in order to satisfy the radiation condition, yielding an expansion of the form:

$$
u^{\mathrm{sca}}(r, \theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} u^{\mathrm{sca}}_{l,m} \frac{1}{R} \frac{\mathbf{h}_l(kr)}{\mathbf{h}_l(kR)} Y_l^m(\theta, \phi), \tag{3.4}
$$

where the $u^{\mathrm{sca}}_{l,m}$ denotes the L^2 scalar product of u^{sca} with $\frac{1}{R} Y_l^m$ over the boundary $\partial\Omega$.

Taking the normal derivative of this expansion at this boundary gives us the operator $T_k(u^{\mathrm{sca}})$, defined below:

$$
T_k(u^{\mathrm{sca}})(r, \theta, \phi)\big|_{r=R} = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \frac{k}{R} u^{\mathrm{sca}}_{l,m} \frac{\mathbf{h}'_l(kR)}{\mathbf{h}_l(kR)} Y_l^m(\theta, \phi) \tag{3.5}
$$

We can then truncate this series and use it as our boundary condition by substituting for the partial derivative at the boundary in (3.2): $\partial_n u^{\mathrm{sca}} = T_k(u^{\mathrm{sca}}) \iff \partial_n u = T_k(u^{\mathrm{sca}}) + \partial_n(u^{\mathrm{inc}}) = T_k(u) - T_k(u^{\mathrm{inc}}) + \partial_n(u^{\mathrm{inc}})$. We finally obtain the weak form of the problem:

$$
\begin{aligned}
&\text{Find } u \in H^1(\Omega) \text{ such that:}\\
&-\int_\Omega \epsilon^{-1}\nabla u\cdot\nabla\overline{v}\,d\Omega + \int_\Omega k^2 u\,\overline{v}\,d\Omega + \int_{\partial\Omega}\epsilon^{-1}T_k(u)\,\overline{v}\,dS\\
&= -\int_{\partial\Omega}\epsilon^{-1}\partial_n u^{\mathrm{inc}}\,\overline{v}\,dS + \int_{\partial\Omega}\epsilon^{-1}T_k(u^{\mathrm{inc}})\,\overline{v}\,dS, \quad \forall v \in H^1(\Omega).
\end{aligned}
\tag{3.6}
$$

In order to include either approach (or other boundary conditions of this form), we define the boundary operator B_k by

$$
B_k(u) = \begin{cases} T_k(u), & \text{(using DtN)},\\ ik\,u, & \text{(using the impedance condition)}. \end{cases}
\tag{3.7}
$$

The different boundary conditions are better suited for different choices of mesh truncation and implementation. In cases where speed is desired, the ease of implementation and the speed of solving the resulting linear system makes the impedance boundary condition preferable. However, if one chooses a truncation of the domain near the interface Γ, then the impedance condition (being by design a first order boundary condition) will lose accuracy. Therefore, if trying to use, for example, higher order Finite Elements and reduce the number of meshed degrees of freedom outside of the scatterer, the DtN (which can be of arbitrarily high order depending on the choice of truncation made in the series (3.5)) can be much more accurate.

Following the choice of boundary condition, the weak formulation for the scattering problem can be written as

$$
\begin{aligned}
&\text{Find } u \in H^1(\Omega) \text{ such that:}\\
&-\int_\Omega \epsilon^{-1}\nabla u\cdot\nabla\overline{v}\,d\Omega + \int_\Omega k^2 u\,\overline{v}\,d\Omega + \int_{\partial\Omega}\epsilon^{-1}B_k(u)\,\overline{v}\,dS\\
&= -\int_{\partial\Omega}\epsilon^{-1}\partial_n u^{\mathrm{inc}}\,\overline{v}\,dS + \int_{\partial\Omega}\epsilon^{-1}B_k(u^{\mathrm{inc}})\,\overline{v}\,dS, \quad \forall v \in H^1(\Omega).
\end{aligned}
\tag{3.8}
$$

We discretize this problem, approximating our solution $u \in H^1(\Omega)$ by $u_h \in \mathcal{V}^p(\Omega_h)$, where $\mathcal{V}^p(\Omega_h)$ is the space of polynomial functions of up to order p on the triangular mesh Ω_h with meshwidth h. This problem is not necessarily well-posed for any value of ϵ, and FEM convergence is not guaranteed for any mesh, see Lemma 3.2.1. We define $\Omega \subset \mathbb{R}^d$ to be a finite closed geometry. We furthermore define $\mathcal{D} \subsetneq \Omega$ to be a finite scatterer contained inside this region with smooth, closed boundary Γ. We seek to prove the weak formulation and paired discrete problem given above in problem (3.8) is well-posed.

Lemma 3.2.1. (3.8) is well posed iff $\epsilon_m \neq -1$. Additionally, FEM convergence is guaranteed as long as the mesh is T-conforming. We follow the T-coercivity methods laid out in [Bonnet-Ben Dhia et al. (2012), Bonnet-Ben Dhia et al. (2018)].

Proof. Firstly, we note that (3.8) is classically well-posed for $\epsilon_m > 0$. We define a sesquilinear operator $a(u, v)$ to be T-coercive if, for some isomorphism $\mathbf{T}$:

$$\exists \alpha_{\mathbf{T}} > 0, \beta_{\mathbf{T}} > 0 \text{ s.t. } a(u, \mathbf{T}u) \geq \alpha_{\mathbf{T}} \|u\|_{H^1}^2 - \beta_{\mathbf{T}} \|u\|_{L^2}^2 \tag{3.9}$$

We then construct the continuous operators:

$$\mathbf{T}v = \begin{cases} v_1 & \text{in } \mathcal{D}, \\ -v_2 + 2\mathbf{R}v_1 & \text{in } \Omega \setminus \bar{\mathcal{D}}, \end{cases} \quad \text{and } \mathbf{T}'v = \begin{cases} v_1 - 2\mathbf{R}'v_2 & \text{in } \mathcal{D}, \\ -v_2 & \text{in } \Omega \setminus \bar{\mathcal{D}}, \end{cases} \tag{3.10}$$

Where $\mathbf{R}, \mathbf{R}'$ are operators with the property $\mathbf{R}v_1|_\Gamma = v_1|_\Gamma$, $\mathbf{R}'v_2|_\Gamma = v_2|_\Gamma$. In our case, for the case of a sphere or disk shaped $\mathcal{D}$, these are defined as symmetric reflections across the interface Γ, subject to a local cut-off function to ensure they remain onto Ω. For such a symmetric operator, per Bonnet-Ben Dhia et al. (2012), the continuous problem (3.1) is well posed for $\epsilon_m \notin \mathfrak{C} = [-\|\mathbf{R}'\|^2, -1/\|\mathbf{R}\|^2]$.

Remark 3.2.1. This is a strictly optimal in the case of 2D (the problem is well posed if and only if ϵ_m meets this requirement), however in 3D for non-smooth interfaces it remains an open question if there are any values of ϵ_m inside this region for which the problem is well posed, as the set $\mathfrak{C}$ depends on the smoothness of Γ. Therefore, for corners or non-convex shapes in 3D the optimal interval may shrink or lose clarity, see Bonnet-Ben Dhia et al. (2018).

Regardless, requiring ϵ_m to fall outside $\mathfrak{C}$ is sufficient to guarantee well-posedness for either case. In the simple case of the disk (or sphere) scatterer, locally the domains map onto each other with a simple radial symmetry, thus $\|\mathbf{R}\| = \|\mathbf{R}'\| = 1$. Therefore we require only that $\epsilon_m \neq -1$ to ensure T-coercivity and well-posedness of the continuous problem.

For the discrete problem, we consider the function space $\mathcal{V}^p(\Omega_h)$ the space of polynomial functions of up to order p on the mesh Ω_h. We note that for a generic mesh, $R(\mathcal{V}^p(\mathcal{D})) \neq \mathcal{V}^p(\Omega \backslash \overline{\mathcal{D}})$, as mesh cells across the interface may not be symmetric with each other. Therefore, our T-operators may not necessarily be valid acting on functions in $\mathcal{V}^p(\Omega_h)$. A simple remedy to inherit this property is to demand symmetrical mesh elements about the interface, so that $\mathbf{R}, \mathbf{R}'$ can be inherited from $H^1(\Omega)$. In other words, we define a T-conforming mesh for which the discrete problem has convergence to be one such that $\mathbf{T}(\mathcal{V}^p(\mathcal{D})) \subseteq \mathcal{V}^p(\mathcal{D})$. Additionally, due to the localization properties of this problem, we need only impose this constraint near the interface, see Bonnet-Ben Dhia et al. (2012). Following from Bonnet-Ben Dhia et al. (2018), we obtain the error estimate:

$$\|u - u_h\|_{L^2(\Omega)} \;\leq\; \frac{M}{\alpha_T} \inf_{v_h \in \mathcal{V}(\Omega_h)} \|u - v_h\|_{L^2(\Omega)},$$

where M is the continuity constant of the sesquilinear form $A_h(\cdot, \cdot)$. Moreover, as $\epsilon_m \to \mathfrak{C}$ the involution T becomes increasingly distorted and one shows (see [Bonnet-

Ben Dhia et al. (2012)]) that, for example in the case of a smooth ball scatterer, $\alpha_T = |\frac{1}{\epsilon_m} + 1|$, so that $\alpha_T \to 0$ as $\epsilon_m \to 1$. Hence $C = M/\alpha_T \to \infty$, and the error bound necessarily degenerates. While this example is valid for a sphere or disk, this same property holds for other contrasts as the geometry shifts and $\mathfrak{C}$ changes.

$\square$

Remark 3.2.2. We remark that we can explicitly see the explosion of the gradient (and therefore the necessary degeneration of FEM error on a fixed discretization) present as $\epsilon_m \to -1$ for a spherical scatterer in Figure 2.9.

Before presenting numerical results, we pause to summarize our theoretical expectations. First, while we have well-posedness in theory, the a priori bound from Lemma 3.2.1,

$$\|u - u_h\|_{L^2(\Omega)} \le \frac{M}{\alpha_T} \inf_{v_h \in \mathcal{T}(\Omega_h)} \|u - v_h\|_{L^2(\Omega)},$$

together with Céa's lemma for standard FEM (c.f. Ciarlet (1978)), shows that—provided ϵ_m stays away from the critical contrast set $\mathfrak{C}$—we recover the usual $\mathcal{O}(h^p)$ convergence (up to the pollution factor in k and error imposed by the boundary condition) familiar from classical theory Ihlenburg (1995). However, as $\epsilon_m \to -1$, the T-coercivity constant $\alpha_T \to 0$, the bound necessarily degenerates, and the true solution develops increasingly steep gradients (see Remark 3.2.2 and the analytic behaviour in §2.2).

With these insights in hand, §3.3 will verify off-resonance h-convergence for various k and p and demonstrate the breakdown of standard FEM as ϵ_m approaches the plasmonic resonance.

Remark 3.2.3. While our solution to the lack of T-coercivity in the discrete space by creating a mesh which is locally symmetric about the interface is simple in concept, it may not be simple in execution, and becomes more complicated to construct for less simple geometries. To address this problem, we provide codes for the construction of locally T-conforming sphere and disk-shaped meshes, see Latham (2025b).

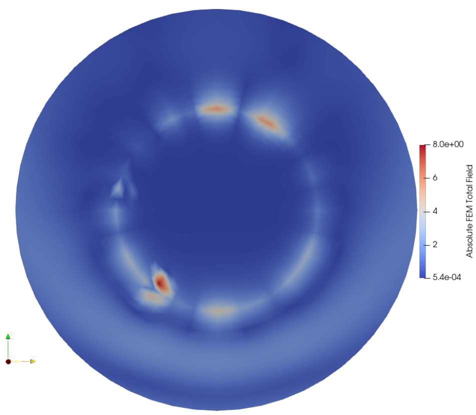
3.3. Numerical Results

3.3.1. Numerical Results Off-Resonance

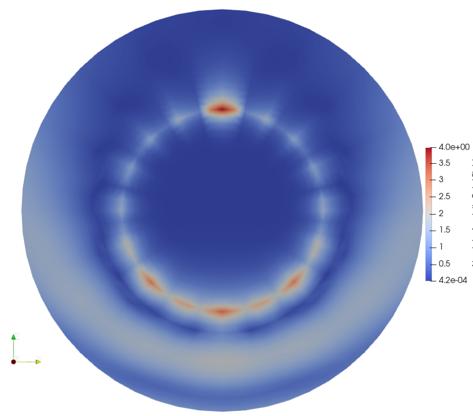
We discretize this weak form using a symmetric mesh and solve for u_h using a hierarchical polynomial basis. Here we make use of the FEM software NGSolve from which these functions are drawn. Firstly, we investigate a non-resonance sample case to test accuracy of the method. We take both order 2 and 3 hierarchical polynomial finite elements (see Figure 3.1 and Figure 3.2) as our basis, the wavenumber $k = 5$, and we use our first order impedance boundary condition, $B_k(u) = iku$. We set $\epsilon_m = -1.3$. The images below are slices of the 3D spherical domain in the $y - z$ plane.

We note in the relatively unrefined case of $p = 2$, there is some artifacting near the interface, but when we refine our order, we largely have agreement in both magnitude and behavior between the FEM and Analytic solutions, even for a low-order polynomial basis and relatively unrefined mesh. Additionally, for the purpose of visualization, our radius of truncation is quite small here ($R = 2$) and we expect to incur the bulk of our error from the first order impedance boundary condition.

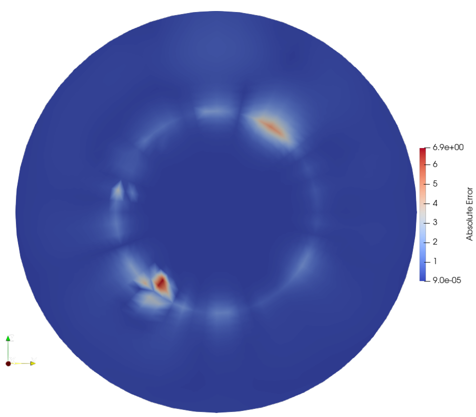
We then proceed to plot the error for more FEM experiments with a similar parameter set ($k = 5, \epsilon_m = -1.3$) and document the results in Figure 3.3a for a variety of meshwidths. Here we increase the radius of truncation to $R = 5$ so the pollution of error from the boundary condition is consequently much lower. We also decrease the order of FEM elements to make the computation more managable, so that $p = 2$ (we use order 2 finite elements) for these experiments. Consequently, we note that the FEM method converges (as expected) at a rate of at least order 2.



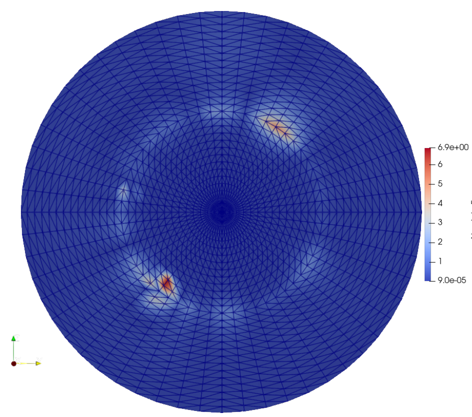
(a) FEM Absolute Total Field



(b) Analytic Absolute Total Field

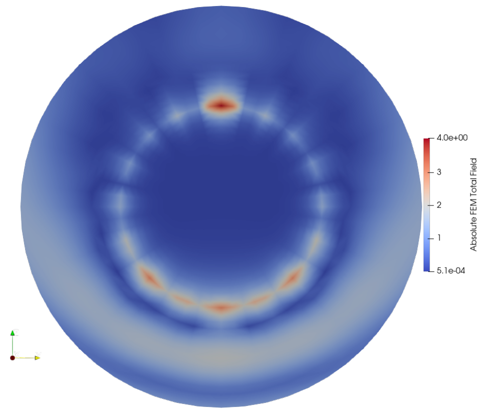


(c) Absolute Error

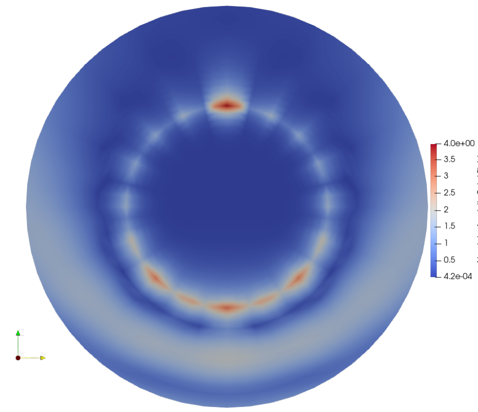


(d) Absolute Error with Mesh Displayed

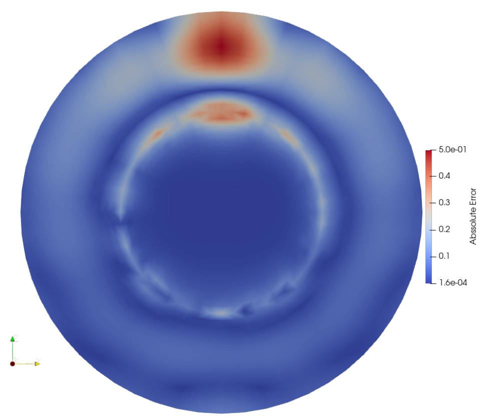
Figure 3.1. Results of FEM with $\epsilon_m = -1.3$, $k = 5$ using Impedance BC, order 2 basis functions and 76828 dofs ($h = 0.1$)



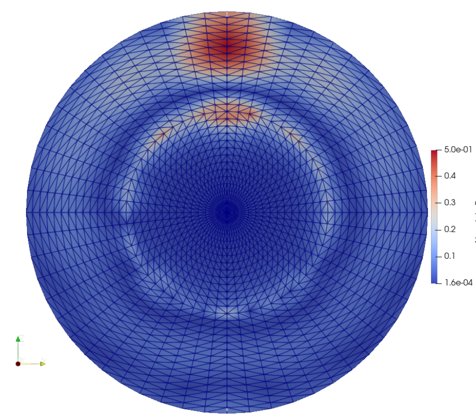
(a) FEM Absolute Total Field



(b) Analytic Absolute Total Field

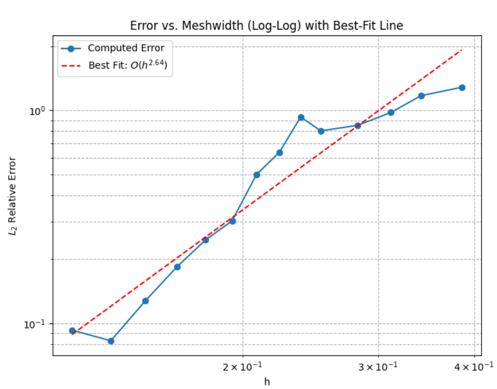


(c) Absolute Error

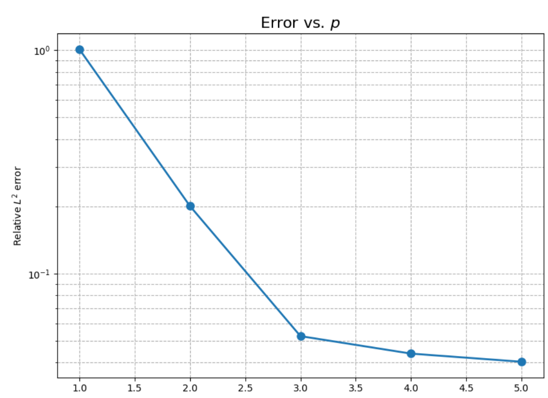


(d) Absolute Error with Mesh Displayed

Figure 3.2. Results of FEM with $\epsilon_m = -1.3$, $k = 5$ using Impedance BC, order 3 basis functions and 256287 dofs ($h = 0.1$)



(a) Relative L^2 error vs. meshwidth h for $\epsilon_m = -1.3$, $k = 5$

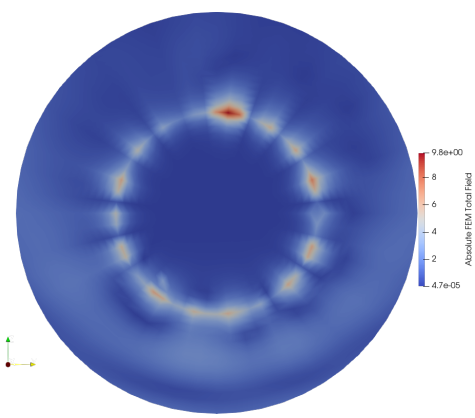


(b) Relative L^2 error vs. order p for $\epsilon_m = -1.3$, $k = 5$

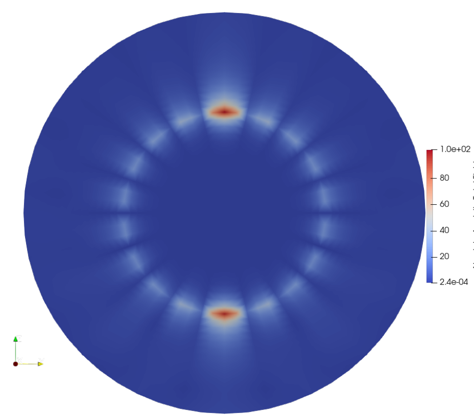
Figure 3.3. Convergence (in h left, p right) of standard FEM.

3.3.2. Numerical Results in the Plasmonic Regime

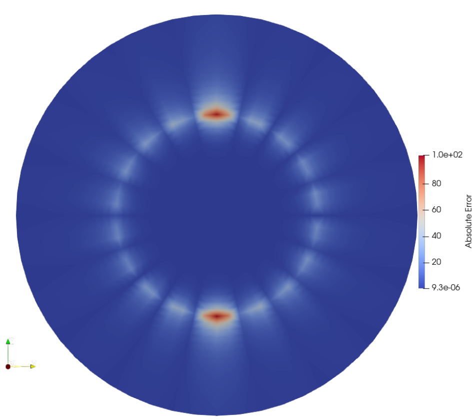
Having seen success for the standard negative material problem on a sphere, we now investigate a near-resonance case. Here, we choose $\epsilon_m = -1.3$, $k = 5.23380097703242$ which corresponds to $l^* = 14$. We compute this using the methods presented in §2.3. We again take orders both 2 and 3 hierarchical polynomial finite elements as our basis. For our boundary condition, we set $B_k(u) = iku$, the first order impedance condition. This choice of parameters corresponds precisely to the prior experiment (see Figure 3.2). The images presented below are slices of the 3D sphere in the $y - z$ plane.



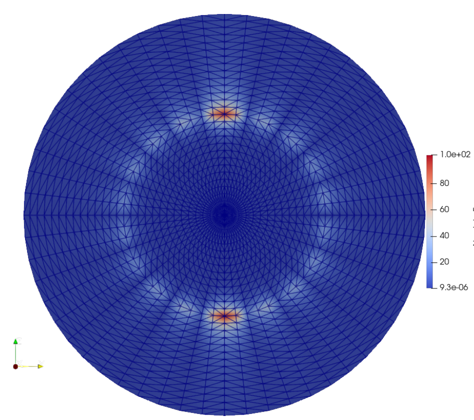
(a) FEM Absolute Total Field



(b) Analytic Absolute Total Field

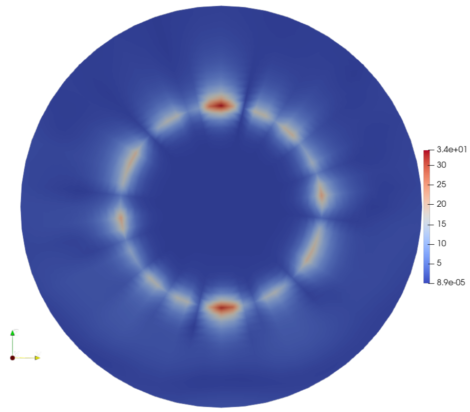


(c) Absolute Error

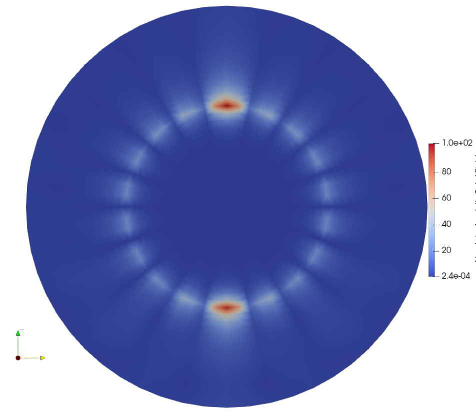


(d) Absolute Error with Mesh Displayed

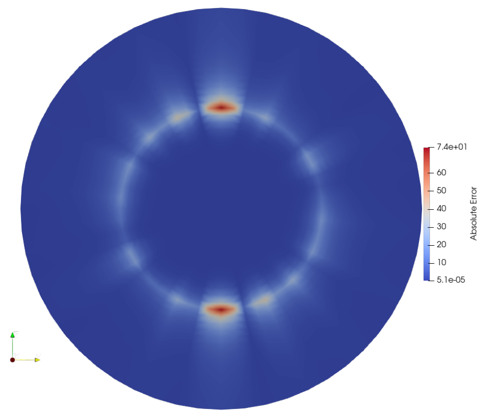
Figure 3.4. Results of FEM with $\epsilon_m = -1.3$, $k = 5.23380097703242$ using order 2 basis functions and 76828 dofs ($h = 0.1$)



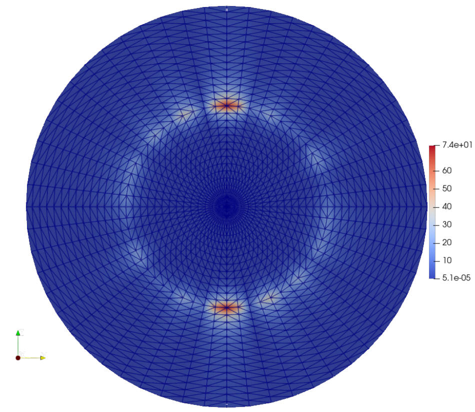
(a) FEM Absolute Total Field



(b) Analytic Absolute Total Field



(c) Absolute Error



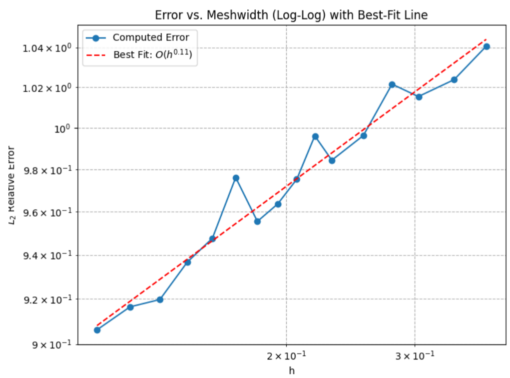
(d) Absolute Error with Mesh Displayed

Figure 3.5. Results of FEM with $\epsilon_m = -1.3$, $k = 5.23380097703242$ using order 3 basis functions and 256287 dofs ($h = 0.1$)

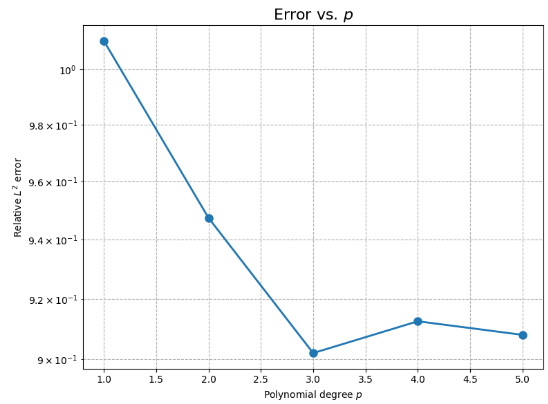
Despite increasing the number of basis functions, the error profile remains closely tied to the overall analytic solution—which exhibits much higher order behavior. In other words, even as the degrees of freedom escalate dramatically (from 76,828 for order 2 to 256,287 for order 3), the standard FEM continues to capture primarily the low-order components of the solution while missing the localized surface plasmon. Moreover, the inherent computational cost of additional refinements in three dimensions renders such strategies inefficient in this near-resonance regime.

To underscore the impact of the plasmonic behavior, we examine the relative L^2 error versus meshwidth h (see Figure 3.6a) and vs. polynomial order Figure 3.6b.

Here, the convergence rates are notably slower than in the off-resonance case, indicating that the refinement is effectively resolving only the lower-order features—not the rapid variations associated with the surface plasmon. This also ties in with the observed analytic gradient 2.9 - less regularity of the solution leads to slower convergence. These observations clearly demonstrate that simple spatial refinement is insufficient for accurately capturing the near-resonance dynamics.



(a) Relative L^2 error vs. meshwidth h for the near-resonance case $\epsilon_m = -1.3$, $k = 5.233$

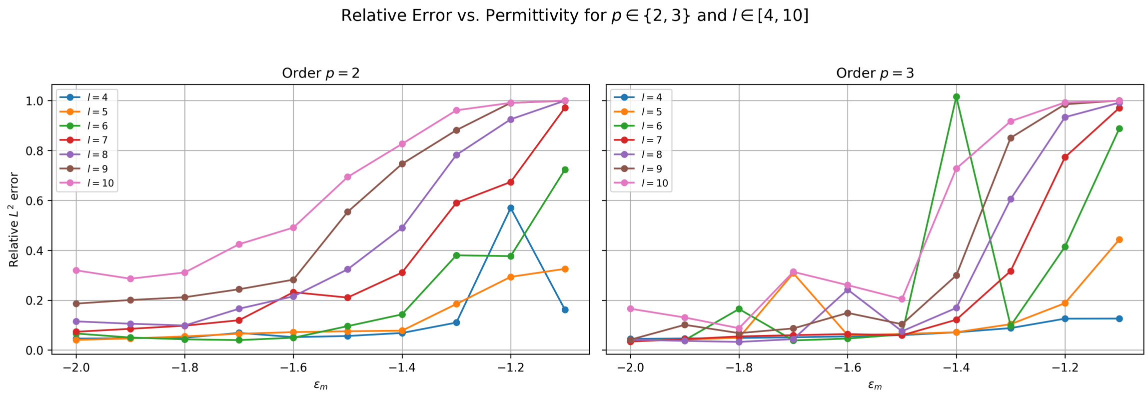


(b) Relative L^2 error vs. order p for the near-resonance case $\epsilon_m = -1.3$, $k = 5.233$

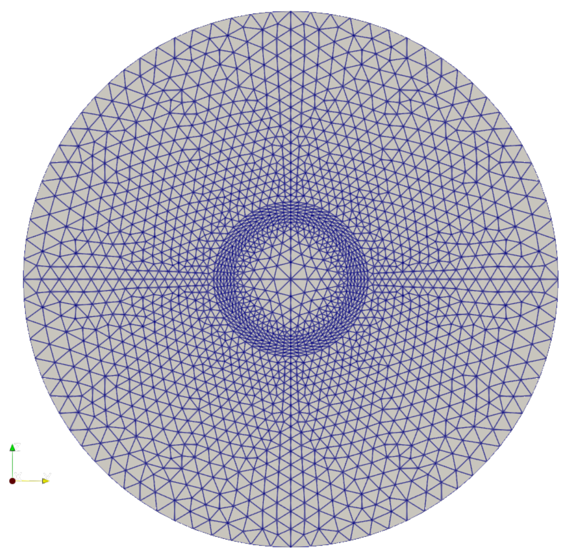
Figure 3.6. Convergence (in h left, p right) of standard FEM in the near-resonance regime.

To obtain a more thorough understanding of the near-resonance phenomena we are capable of resolving, we fix an integer mode l^* and vary the permittivity ϵ_m, approaching the critical value -1. For each ϵ_m, we select the corresponding resonance wavenumber $k = \Re(\lambda_{l^*})$. We then compute the relative L^2 errors obtained using the FEM discretization for polynomial orders $p = 2$ and $p = 3$, plotting these results in Figure Figure 3.7a.

We emphasize that for these computations, we employ graded meshes that are radially symmetric and locally T-conforming. These meshes are constructed using the code available via Latham (2025b).



(a) Relative L^2 errors vs. ϵ_m for fixed $k = \Re(\lambda_{l^*})$, for polynomial orders $p = 2$ (left) and $p = 3$ (right).



(b) Slice of the locally T-conforming spherical mesh used for the results in (a).

Figure 3.7. (a) Error curves for various ϵ_m on graded, radially symmetric, locally T-conforming meshes (see Lemma Lemma 3.2.1); (b) A representative mesh slice constructed with Latham (2025b).

We note that while for less extreme surface plasmons, we see relatively low L^2 error, however as $\epsilon_m \to -1$ or if a larger order resonance is chosen (e.g. larger l), we still see relative L^2 error near 1. For very extreme cases, therefore, there exists a regime for which L^2 error remains unacceptable high. This limitation strongly motivates the need for an alternative approach. In the following section, we introduce an enriched FEM framework that incorporates the resonant plasmon mode directly into the approximation space.

3.4. Enriched FEM Approach

To address the aforementioned underlying issue with our polynomial approach to the Helmholtz-like problem in the near-resonance regime, we will attempt to devise an approach based on FEM to account for the plasmons at the discrete level.

With our naive approach, that is to say using the FEM without accounting for the surface plasmon as we did above, the difficulty in computing these solutions appears numerically due to the rapid oscillating behavior and extremely fast decay of the field near the interface of the scatterer. In the case of multiple scatterers, creating a workable mesh and inverting the linear system (resulting from the discretization) to achieve even a remotely accurate result may be very expensive, perhaps on the order of millions of degrees of freedom, even with only local refinement. Thus, we seek to develop a method for solving these problems which can recover a more accurate result on a less refined mesh (a.k.a coarse discretization), relieving the intense computational costs associated with the problem.

For the remainder of this chapter, we assume that we are in a regime with excited plasmonic behaviour, and therefore, we are near some resonance λ_{l^*} in the sense that $k = \Re(\lambda_{l^*})$ for some fixed value of $l^* \in \mathbb{Z}$. Therefore, for simplicity, all variables previously dependent on the index l will now have the index correspond to the fixed l^* given by our choice of k. In particular, we define the surface plasmon u^{pla}

to be the resonant mode paired with λ_{l^*}. Furthermore, we will assume that $\epsilon_m < -1$, as demonstrated in the results of section §2.2.4, this is necessary for the plasmonic behavior we are interested in to appear.

$$u^{\text{pla}}(r,\phi) = \begin{cases} \mathbf{i}_{l^*}\left(\sqrt{|\epsilon_m|}\lambda_{l^*} r\right) Y_{l^*}^0(\phi), & \text{for } r < 1, \\ \dfrac{\mathbf{i}_{l^*}\left(\sqrt{|\epsilon_m|}\lambda_{l^*}\right)}{\mathbf{h}_{l^*}\left(\lambda_{l^*}\right)} \mathbf{h}_{l^*}\left(\lambda_{l^*} r\right) Y_{l^*}^0(\phi), & \text{for } r \geq 1. \end{cases} \tag{3.11}$$

3.4.1. Enrichment with the Resonant Mode

Typically, Enriched finite element methods extend the standard finite element approximation space by augmenting it with problem-specific basis functions that capture known local solution behaviors. In a standard partition-of-unity formulation, also known as the generalized or extended finite element method (XFEM), each element's shape functions is enriched with additional functions (e.g., singular solutions, plane waves, or boundary-layer profiles) that reflect the local analytical behavior of the governing equations. The theoretical motivation is that by embedding a priori knowledge of the solution into the trial space, one can achieve significantly higher accuracy without global mesh refinement. Indeed, these enriched spaces can incorporate known singular or oscillatory patterns of the solution, whereas a standard polynomial-based finite element method would require extremely fine meshes or high polynomial degree to resolve such features [Babuska & Melenk (1997); Belytschko et al. (2001)]. Inspired by these methods, we perform a global enrichment of the finite element space using the resonant mode, and take advantage of our near-resonant regime to develop an approximation of the solution.

Recall that as part of the standard FEM approach, we define the total field by the sum of the incident field and a scattered field:

$$u = u^{\text{inc}} + u^{\text{sca}} \tag{3.12}$$

And searched for a solution u^{sca} by discretizing $H^1(\mathbb{R}^3)$ with the FEM. Our plasmonic component, $u^{\text{pla}} \notin H^1(\mathbb{R}^3)$ (due to not satisfying the radiation condition, see (2.63)), and as shown by our prior results with the standard FEM, it is not well captured by basis functions spanning the finite dimensional approximation of this vector space. Therefore, rather than demanding our solution is entirely within $H^1(\mathbb{R}^3)$, we instead extend our problem to allow our solution to be in the following plasmonic space:

Definition 3.4.1: Plasmonic Space

We define the *plasmonic space* as

$$V^{\text{pla}}(\mathbb{R}^3) = H^1(\mathbb{R}^3) \oplus \text{span}\{u^{\text{pla}}\}, \tag{3.13}$$

where $H^1(\mathbb{R}^3)$ is the standard Sobolev space and u^{pla} denotes the associated plasmon.

This is the "plasmonic" space in which we seek to solve our problem. Our scattered field solution therefore, should behave as a linear combination of an H^1 function and the associated surface plasmon u^{pla}. In keeping with this, we will assume that we can write our solution as:

$$u = u^{\text{inc}} + \underbrace{u^{\text{reg}} + \alpha u^{\text{pla}}}_{u^{\text{sca}}}, \quad u^{\text{reg}} \in H^1(\mathbb{R}^3), \quad \alpha \in \mathbb{C} \tag{3.14}$$

where u^{reg} is the regular, low oscillation component contained in $H^1(\mathbb{R}^3)$ and α is the amplitude coefficient for u^{pla}.

In this enriched formulation, the boundary condition is imposed on the regular component of the scattered field through the boundary operator B_k. We recall:

$$B_k(u) = \begin{cases} T_k(u), & \text{(using the DtN approach)}, \\ \\ ik\,u, & \text{(using the first order impedance condition)}, \end{cases}$$

With this in mind, we may rewrite our original problem.
Now we will search for solutions in $V^{\mathrm{pla}}(\Omega)$:

$$\begin{aligned} &\text{Find } (\alpha, u^{\mathrm{reg}}) \in \mathbb{C} \times H^1(\Omega) \text{ such that } u = u^{\mathrm{inc}} + u^{\mathrm{reg}} + \alpha\, u^{\mathrm{pla}} \text{ satisfies:} \\ &\quad \nabla \cdot \left(\epsilon^{-1} \nabla u\right) + k^2 u = 0 \quad \text{in } \Omega, \\ &\quad [u]_\Gamma = 0, \quad \left[\epsilon^{-1} \partial_n u\right]_\Gamma = 0 \quad \text{on } \Gamma, \\ &\quad \left(\partial_n u - B_k(u)\right)\Big|_{\partial\Omega} = \left(\partial_n u^{\mathrm{inc}} - B_k(u^{\mathrm{inc}})\right)\Big|_{\partial\Omega} + \alpha \left(\partial_n u^{\mathrm{pla}} - B_k(u^{\mathrm{pla}})\right)\Big|_{\partial\Omega}, \end{aligned} \tag{3.15}$$

with the Sommerfeld radiation condition (here applied via the boundary operator (3.7)) applying to the regular component u^{reg}. We note that technically this is not the whole of the scattered field because the plasmon doesn't satisfy the explicit boundary condition, however it does approximately satisfy it for near field truncation, see the approximation give in (2.63). We remark here briefly that experiments with bounded resonances (resonances of the truncated system) for which this is exact do not yield substantial differences. Problem (3.15) can be rewritten as a problem in $(\alpha, u^{\mathrm{reg}})$ by taking advantage of the fact that u^{pla} is an eigenmode. First, we can rewrite $\nabla \cdot (\epsilon^{-1} \nabla u) + k^2 u = 0$ as:

$$\begin{aligned} &\nabla \cdot (\epsilon^{-1} \nabla u^{\mathrm{reg}}) + k^2 u^{\mathrm{reg}} + \alpha(\nabla \cdot (\epsilon^{-1} \nabla u^{\mathrm{pla}}) + k^2 u^{\mathrm{pla}}) = -(\nabla \cdot (\epsilon^{-1} \nabla u^{\mathrm{inc}}) + k^2 u^{\mathrm{inc}}) \\ &\Leftrightarrow \nabla \cdot (\epsilon^{-1} \nabla u^{\mathrm{reg}}) + k^2 u^{\mathrm{reg}} + \alpha(k^2 - \lambda^2) u^{\mathrm{pla}} = f^{\mathrm{inc}} \end{aligned} \tag{3.16}$$

with

$$f^{\text{inc}} = \begin{cases} (\frac{1}{\epsilon_m} - 1)k^2 u^{\text{inc}} & \mathcal{D} \\ 0 & \Omega \setminus \overline{\mathcal{D}} \end{cases} \tag{3.17}$$

Then (3.15) is equivalent to:

$$\begin{aligned} &\text{Find } (\alpha, u^{\text{reg}}) \in \mathbb{C} \times H^1(\Omega) \text{ such that:} \\ &\quad \nabla \cdot \left(\epsilon^{-1} \nabla u^{\text{reg}}\right) + k^2 u^{\text{reg}} + \alpha\,(k^2 - \lambda^2) u^{\text{pla}} = f^{\text{inc}} \quad \text{in } \Omega, \\ &\quad [u^{\text{reg}}] + [u^{\text{inc}}]_\Gamma = 0, \quad \left[\epsilon^{-1}\partial_n u^{\text{reg}}\right] + \left[\epsilon^{-1}\partial_n u^{\text{inc}}\right]_\Gamma = 0 \quad \text{on } \Gamma, \\ &\quad \left(\partial_n u^{\text{reg}} - B_k(u^{\text{reg}})\right)\Big|_{\partial\Omega} = 0. \end{aligned} \tag{3.18}$$

We now write the associated weak form (and we recall that $[u^{\text{pla}}]_\Gamma = [\epsilon^{-1}\partial_n u^{\text{pla}}]_\Gamma = 0$).

3.4.2. Augmented System

Continuing from equation (3.18), we now seek a weak form from which to derive a linear system to solve u^{reg} and α that we can discretize. In order to do this, we will draw test functions $v \in V^{\text{pla}}(\Omega)$, our plasmonic function space, leading to the weak form:

$$\int_\Omega \nabla\cdot(\epsilon^{-1}\nabla u^{\text{reg}})\overline{v}\, d\Omega + k^2 \int_\Omega u^{\text{reg}}\overline{v}\, d\Omega + \alpha \int_\Omega (k^2 - \lambda^2) u^{\text{pla}}\overline{v}\, d\Omega = \int_\Omega \overline{v} f^{\text{inc}} d\Omega, \quad \forall\, v \in V^{\text{pla}}(\Omega) \tag{3.19}$$

Focusing on the term

$$I = \int_\Omega \nabla \cdot \left(\epsilon^{-1}\nabla u^{\text{reg}}\right)\overline{v}\, d\Omega,$$

we first split the integration over the scatterer $\mathcal{D}$ (where $\epsilon = \epsilon_m$) and the exterior region $\Omega \setminus \overline{\mathcal{D}}$ (where $\epsilon = 1$):

$$I = \frac{1}{\epsilon_m} \int_{\mathcal{D}} \nabla \cdot \Big(\nabla u^{\mathrm{reg}}\Big) \overline{v} \, d\mathcal{D} + \int_{\Omega \setminus \overline{\mathcal{D}}} \nabla \cdot \Big(\nabla u^{\mathrm{reg}}\Big) \overline{v} \, d\Omega.$$

Applying Green's identity on each subdomain yields

$$\begin{aligned} \frac{1}{\epsilon_m} \int_{\mathcal{D}} \nabla \cdot \Big(\nabla u^{\mathrm{reg}}\Big) \overline{v} \, d\mathcal{D} &= -\frac{1}{\epsilon_m} \int_{\mathcal{D}} \nabla u^{\mathrm{reg}} \cdot \nabla \overline{v} \, d\mathcal{D} + \frac{1}{\epsilon_m} \int_{\Gamma} \partial_n u^{\mathrm{reg}} \, \overline{v} \, dS, \\ \int_{\Omega \setminus \overline{\mathcal{D}}} \nabla \cdot \Big(\nabla u^{\mathrm{reg}}\Big) \overline{v} \, d\Omega &= - \int_{\Omega \setminus \overline{\mathcal{D}}} \nabla u^{\mathrm{reg}} \cdot \nabla \overline{v} \, d\Omega + \int_{\partial(\Omega \setminus \overline{\mathcal{D}})} \partial_n u^{\mathrm{reg}} \, \overline{v} \, dS. \end{aligned}$$

Because the interface Γ is common to both subdomains, the corresponding boundary integrals cancel (thanks to the transmission conditions on u^{reg} and $\epsilon^{-1}\partial_n u^{\mathrm{reg}}$). Combining the above results, we obtain

$$I = - \int_{\Omega} \epsilon^{-1} \nabla u^{\mathrm{reg}} \cdot \nabla \overline{v} \, d\Omega + \int_{\partial\Omega} \partial_n u^{\mathrm{reg}} \, \overline{v} \, dS.$$

Finally, we enforce the boundary condition on $\partial\Omega$ via the operator B_k when using the first-order impedance condition). That is, we set

$$\partial_n u^{\mathrm{reg}} = B_k(u^{\mathrm{reg}}) \quad \text{on } \partial\Omega.$$

Thus, the expression becomes

$$\int_{\Omega} \nabla \cdot \Big(\epsilon^{-1} \nabla u^{\mathrm{reg}}\Big) \overline{v} \, d\Omega = - \int_{\Omega} \epsilon^{-1} \nabla u^{\mathrm{reg}} \cdot \nabla \overline{v} \, d\Omega + \int_{\partial\Omega} B_k(u^{\mathrm{reg}}) \, \overline{v} \, dS.$$

Note the cancellation of the interface terms, so the resulting weak form looks like:

$$
\begin{aligned}
&- \int_{\Omega} \epsilon^{-1} \nabla u^{\mathrm{reg}} \cdot \nabla \overline{v}\, d\Omega + \int_{\partial\Omega} \epsilon^{-1} B_k \Big(u^{\mathrm{reg}} + \alpha\, u^{\mathrm{pla}} \Big) \overline{v}\, dS \\
&+ k^2 \int_{\Omega} u^{\mathrm{reg}} \overline{v}\, d\Omega + \alpha\, (k^2 - \lambda^2) \int_{\Omega} u^{\mathrm{pla}} \overline{v}\, d\Omega \\
&= \int_{\Omega} \overline{v}\, f^{\mathrm{inc}}\, d\Omega, \quad \forall v \in V^{\mathrm{pla}}(\Omega).
\end{aligned}
\tag{3.20}
$$

Now let's consider the case where $v \in H^1(\Omega)$, which corresponds to the classical FEM framework. We rewrite (3.20) in (bi)linear forms as

$$
-\mathcal{K}_{\epsilon}(u^{\mathrm{reg}}, v) + \mathcal{B}(u^{\mathrm{reg}}, v) + k^2 \mathcal{M}(u^{\mathrm{reg}}, v) + \alpha \mathcal{M}(u^{\mathrm{pla}}, v) = f(v), \quad \forall v \in H^1(\Omega), \tag{3.21}
$$

where $-\mathcal{K}_{\epsilon}(u^{\mathrm{reg}}, v)$ is a bilinear form corresponding to a weighted stiffness matrix, $\mathcal{B}(u^{\mathrm{reg}}, v)$ corresponds to the surface mass matrix with the boundary operator, and $\mathcal{M}(u^{\mathrm{reg}}, v)$ is the bilinear form corresponding to the standard mass matrix. Note that $\mathcal{M}(u^{\mathrm{pla}}, v)$ is a coupling term that takes the enriched portion of the solution u^{pla} and projects it against the standard test function v. For the standard basis function, we will represent the right hand side terms with the linear form f. In order to obtain a system for the pair $(u^{\mathrm{reg}}, \alpha)$ we now consider $v = u^{\mathrm{pla}}$, which is to say, we must consider the contribution of the enrichment for the test function being multiplied against the trial function. Let us rewrite terms from (3.20) explicitly in that case. For the first term, proceeding as before (split intergral, use Green's identity and transmission conditions) we obtain

$$
\begin{aligned}
-\mathcal{K}_\epsilon(u^{\mathrm{reg}}, u^{\mathrm{pla}}) &= -\int_\Omega \epsilon^{-1} \nabla u^{\mathrm{reg}} \nabla \overline{u^{\mathrm{pla}}}\, d\Omega \\
&= -\int_{\mathcal{D}} \frac{1}{\epsilon_m} \nabla u^{\mathrm{reg}} \nabla \overline{u^{\mathrm{pla}}}\, d\mathcal{D} \;-\; \int_{\Omega \setminus \overline{\mathcal{D}}} \nabla u^{\mathrm{reg}} \nabla \overline{u^{\mathrm{pla}}}\, d\Omega \\
&= \int_{\mathcal{D}} u^{\mathrm{reg}}\, \nabla \cdot \left(\epsilon^{-1} \nabla \overline{u^{\mathrm{pla}}}\right) d\mathcal{D} \;-\; \int_\Gamma \frac{1}{\epsilon_m} u^{\mathrm{reg}}\, \partial_n \overline{u^{\mathrm{pla}}}\, dS \;+\; \int_{\Omega \setminus \overline{\mathcal{D}}} u^{\mathrm{reg}}\, \nabla \cdot \left(\epsilon^{-1} \nabla \overline{u^{\mathrm{pla}}}\right) d\Omega \\
&\qquad + \int_\Gamma u^{\mathrm{reg}}\, \partial_n \overline{u^{\mathrm{pla}}}\, dS \;-\; \int_{\partial\Omega} u^{\mathrm{reg}}\, \partial_n \overline{u^{\mathrm{pla}}}\, dS \\
&= \int_\Omega u^{\mathrm{reg}}\, \nabla \cdot \left(\epsilon^{-1} \nabla \overline{u^{\mathrm{pla}}}\right) d\Omega \;-\; \int_{\partial\Omega} u^{\mathrm{reg}}\, \partial_n \overline{u^{\mathrm{pla}}}\, dS \\
&= -\overline{\lambda}^2 \int_\Omega u^{\mathrm{reg}}\, \overline{u^{\mathrm{pla}}}\, d\Omega \;-\; \int_{\partial\Omega} u^{\mathrm{reg}}\, \partial_n \overline{u^{\mathrm{pla}}}\, dS \\
&= -\overline{\lambda}^2\, \mathcal{M}(u^{\mathrm{reg}}, u^{\mathrm{pla}}) \;-\; \mathcal{S}(u^{\mathrm{reg}}, u^{\mathrm{pla}}).
\end{aligned}
$$

In the last steps we used the fact that u^{pla} is a resonant mode associated to λ (satisfying (3.11)), and we define $\mathcal{S}(u^{\mathrm{reg}}, u^{\mathrm{pla}}) = \int_{\partial\Omega} u^{\mathrm{reg}} \partial_n \overline{u^{\mathrm{pla}}} dS$. We can explicitly compute

$$
\mathcal{M}(u^{\mathrm{pla}}, u^{\mathrm{pla}}) = \int_\Omega u^{\mathrm{pla}} \overline{u^{\mathrm{pla}}}\, d\Omega = \int_0^R r^2 |w(r)|^2 dr := C_1,
$$

where $w(r)$ is the radial component of u^{pla} as defined in (3.11), which we recall below for clarity:

$$
w(r) = \begin{cases} \frac{\mathbf{i}_{l^*}(\sqrt{-\epsilon_m}\lambda)}{\mathbf{h}_{l^*}(\lambda)} \mathbf{h}_{l^*}(\lambda r) & r \geq 1 \\ \mathbf{i}_{l^*}(\sqrt{-\epsilon_m}\lambda r) & r < 1 \end{cases} .
$$

Similarly, we write explicitly the last term, which leads to

$$
f(u^{\mathrm{pla}}) = \int_\Omega \overline{u^{\mathrm{pla}}} f^{\mathrm{inc}}\, d\Omega = k^2 (\frac{1}{\epsilon_m} - 1) i^{l^*} \sqrt{4\pi(2l^* + 1)} \int_0^1 r^2 \overline{w(r)} j_{l^*}(kr) dr := C_2.
$$

In the end, the weak form associated to $v = u^{\mathrm{pla}}$ can be written as:

$$
(k^2 - \overline{\lambda}^2) \mathcal{M}(u^{\mathrm{reg}}, u^{\mathrm{pla}}) + \mathcal{B}(u^{\mathrm{reg}}, u^{\mathrm{pla}}) - \mathcal{S}(u^{\mathrm{reg}}, u^{\mathrm{pla}}) + \alpha C_1 = C_2 \tag{3.22}
$$

Note that the boundary term is now incorporated via the operator B_k. For instance, if the DtN approach is chosen, then we obtain

$$\mathcal{B}(u^{\text{reg}}, u^{\text{pla}}) = \int_{\partial\Omega} \epsilon^{-1} T_k(u^{\text{reg}})\,\overline{u^{\text{pla}}}\, dS = k\,\frac{\mathbf{h}'_{l^*}(kR)}{\mathbf{h}_{l^*}(kR)}\,\overline{w(R)} \int_{\partial\Omega} u^{\text{reg}}\,\overline{Y^0_{l^*}}\, dS, \tag{3.23}$$

which follows from (3.5) and (3.11). In contrast, if the first-order impedance condition is used, then the corresponding boundary term becomes

$$\mathcal{B}(u^{\text{reg}}, u^{\text{pla}}) = \int_{\partial\Omega} \epsilon^{-1}(ik\,u^{\text{reg}})\,\overline{u^{\text{pla}}}\, dS = ik\overline{w(R)} \int_{\partial\Omega} u^{\text{reg}}\,\overline{Y^0_{l^*}}\, dS.$$

Gathering (3.21)-(3.22) we obtain the following system to solve

$$\begin{aligned} &-\mathcal{K}_\epsilon(u^{\text{reg}}, v) + \mathcal{B}(u^{\text{reg}}, v) + k^2\mathcal{M}(u^{\text{reg}}, v) + \alpha\mathcal{M}(u^{\text{pla}}, v) = f(v), \quad \forall v \in H^1(\Omega), \\ &(k^2 - \overline{\lambda}^2)\mathcal{M}(u^{\text{reg}}, u^{\text{pla}}) + \mathcal{B}(u^{\text{reg}}, u^{\text{pla}}) - \mathcal{S}(u^{\text{reg}}, u^{\text{pla}}) + \alpha C_1 = C_2 \end{aligned}$$

that we write in operator form as

$$\begin{bmatrix} -\mathcal{K}_\epsilon + \mathcal{B} + k^2\mathcal{M} & \mathcal{M}_p \\ (k^2 - \overline{\lambda}^2)\mathcal{M}_{r,p} + \mathcal{B}_p - \mathcal{S}_p & C_1 \end{bmatrix} \cdot \begin{bmatrix} u^{\text{reg}} \\ \alpha \end{bmatrix} = \begin{bmatrix} f \\ C_2 \end{bmatrix} \tag{3.24}$$

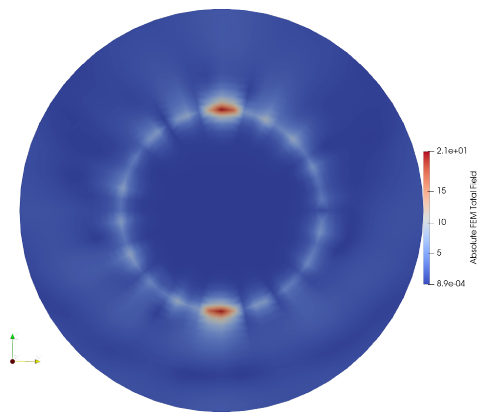
The goal is then to discretize (3.24) using finite elements.

Due to the 2×2 block nature of the enriched system (3.24), many sparse solvers (e.g. SciPy's `spsolve`) may struggle to directly invert this matrix when discretized, especially given that the new final row and column may be fairly dense. To avoid this, we employ a Schur complement approach, which is detailed in Appendix A.

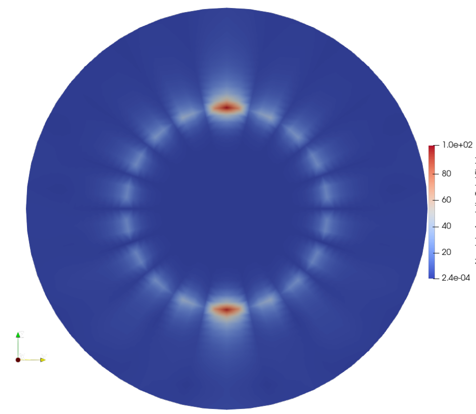
3.4.3. Numerical Results

Taking our problem (3.15), we again mesh Ω with mesh-width h (using symmetrical meshes) and choose as our basis functions hierarchical polynomials on those mesh elements of order p. As before, we choose $\epsilon_m = -1.3$, $k = 5.23380097703242$

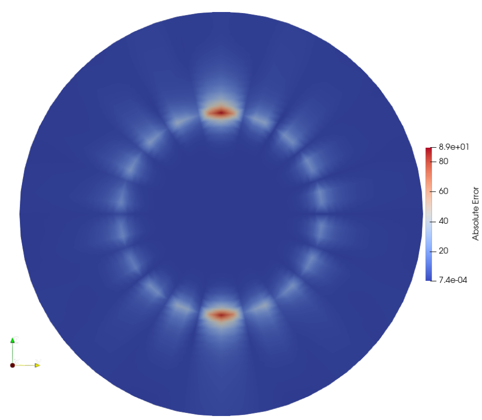
which corresponds to $l^* = 14$. We compute this using the methods presented in §2.3. We again take order 2 hierarchical polynomial finite elements as our basis. For our boundary condition, we set $B_k(u) = iku$, the first order impedance condition. This choice of parameters corresponds precisely to the prior experiments (see Figure 3.2 and Figure 3.8). The images presented below are slices of the 3D sphere in the $y-z$ plane.



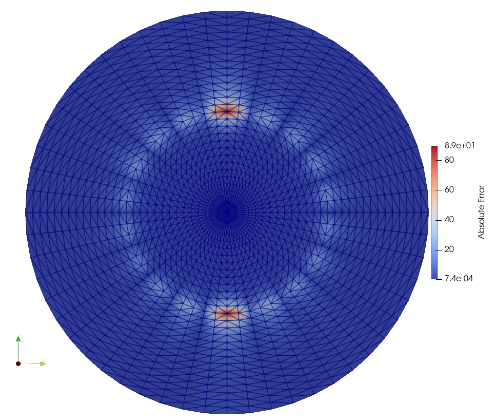
(a) FEM Absolute Total Field



(b) Analytic Absolute Total Field



(c) Absolute Error



(d) Absolute Error with Mesh Displayed

Figure 3.8. Results of Enriched FEM with $\epsilon_m = -1.3$, $k = 5.23380097703242$ using order 2 basis functions and 76829 dofs ($h = 0.1$)

We note that the enriched method well captures the behavior of the surface plasmon, but fails to capture the amplitude of the surface plasmon.

One common technique used to adjust for ill-conditioned matrices as produced by our method (when near-resonance), is the Tikhonov regularization. We explore such a method (and explain why it is ill-suited to the resonance problem) in Appendix B.

Instead, in the next chapter we abandon the purely continuous Galerkin framework and "enrich" our approximation spaces with Trefftz-type functions that exactly satisfy the local Helmholtz and modified Helmholtz equations on each element. By treating the near-interface region using a mixture of propagative plane waves (PPW) and evanescent plane waves (EPW) we hope to more naturally capture the true oscillatory and decaying behavior without resorting to ad hoc damping. This enriched discontinuous Galerkin approach will be developed in Chapter 4.

4. A TREFFTZ APPROACH

4.1. Chapter 4 Introduction

Trefftz methods are Galerkin approaches named after the seminal work by Erik Trefftz (Trefftz (1926)). The goal is to choose as basis elements functions which lie in the kernel of the governing operator. This has two benefits: Firstly, these functions can often be chosen to mimic the behavior of the solution, so they form better approximations at lower order than standard polynomial bases. Secondly, these methods often allow for a reduction in computational costs, due to the nature of the chosen functions. This manifests through repeated integration by parts to take advantage of the fact that these function lie within the kernel as well as analytical integration (as opposed to more expensive numerical integration) for inner products over elements in these methods.

In our case, we run into a need for both of these useful properties. As discussed in the prior chapter, we can often have non-polynomial behavior in the sign-changing problem, and our need for computation resources can also quickly grow out of hand. Therefore, we choose to investigate Trefftz methods to relieve these symptoms of the sign-changing problem. In particular, we are interested in plane wave-based methods, as these are functions which are Trefftz for the Helmholtz problem, and possess both oscillatory, and in some cases, evanescent, behavior.

In this chapter, we consider only the following 2 dimensional Helmholtz-like problem given below. Extensions to 3D are possible, but require greater challenges for implementation and theory. (In particular, 3D versions of the trace inequalities used within would need to be proven, and a recipe would need to be constructed for 'evenly spaced' plane waves, for details on one possible formulation, we direct readers to Galante et al. (2024).)

We define a Plane Wave (PW) as follows:

Definition 4.1.1: Plane Wave

$$\mathcal{P}(\mathbf{x}) = e^{ik_\epsilon \mathbf{d}\cdot\mathbf{x}},$$

$$\mathbf{x} = [x_1, x_2] \quad \text{is a position vector},$$

$$k_\epsilon = \begin{cases} k & \mathbf{x} \in \Omega \setminus \bar{\mathcal{D}} \\ \sqrt{\epsilon_m} k & \mathbf{x} \in \mathcal{D} \end{cases},$$

$$\mathbf{d} = (d_x, d_y) \in \mathbb{C}^2, \quad \mathbf{d}\cdot\mathbf{d} = 1, \quad \text{is a direction vector.}$$

Here, $k_\epsilon = k\sqrt{\epsilon}$ is the local wavenumber with ϵ the local permittivity. When $\Im(\mathbf{d}) \neq 0$, we introduce a decay parameter to the plane wave. We will call this an Evanescent Plane Wave (EPW). A plane wave with $\Im(\mathbf{d}) = 0$ we will call a Propagative Plane Wave (PPW).

We recall the problem at hand:

$$\begin{aligned}
&\text{Find } u := u^{\text{inc}} + u^{\text{sca}} \in H^1_{\text{loc}}(\mathbb{R}^2) \text{ such that:}\\
&\quad \nabla\cdot(\epsilon^{-1}\nabla u) + k^2 u = 0, \quad \text{in } \mathbb{R}^2,\\
&\quad [u]_\Gamma = 0, \quad [\epsilon^{-1}\partial_n u]_\Gamma = 0 \quad \text{on } \Gamma,\\
&\lim_{r\to\infty}\sqrt{r}\left(\frac{\partial u^{\text{sca}}}{\partial r} - iku^{\text{sca}}\right) = 0,\\
&k \in \mathbb{R}^+.
\end{aligned} \tag{4.1}$$

We take $\mathfrak{C}$ to be the critical region inside of which we do not have a guarantee of well-posedness. Bonnet-Ben Dhia et al. (2018) The choice of this region may vary based on the shape of the material scatterer $\mathcal{D}$. We note that in the case where $\mathcal{D}$ is the unit disk, $\mathfrak{C} = \{-1\}$. For further details, see Lemma 3.2.1. We then choose:

$$\epsilon(x) = \begin{cases} 1, & x \in E := \mathbb{R}^2 \setminus \overline{\mathcal{D}} \\ \epsilon_m \in \mathbb{R}^* \setminus \mathfrak{C}, & x \in \mathcal{D} \end{cases}$$

Chapter outline. In §4.2 we derive the Plane-Wave DG weak form for the sign-changing Helmholtz problem. In §4.2.2 proves discrete well-posedness via coercivity in the DG-seminorm. In §4.3 we establish some fundamental results - continuity, quasi-optimality and then a duality error bound in the DG and L^2 norms. §4.4 re-introduces both propagative and evanescent plane waves and discusses their practical sampling for our Trefftz basis. Finally, 4.5 details our numerical results studying this problem.

4.2. The PWDG Weak Form

The Plane Wave Discontinuous Galerkin (PWDG) method is a discontinuous Galerkin Trefftz method wherein we use as basis functions plane waves. (For an overview we suggest Hiptmair et al. (2016)) In this section, we will adapt this method for the sign-changing problem.

4.2.1. The Discrete Problem

Before we discuss our choice of plane waves in greater detail, we will first proceed with a standard DG Trefftz approach, but this time with application to our sign-changing problem. We desire to discretize the problem and solve in a similar manner to the previous FEM chapter §3.1.

In order to define a Trefftz DG weak form, we first need to truncate our domain as we did in the prior chapter. With that aim, we artificially truncate the domain from $\mathbb{R}^2$ to a larger concentric radiation disk Ω about $\mathcal{D}$, with radius R, and we use a first order impedance boundary condition to impose the Sommerfeld radiation condition:

$$B_k = \partial_n u - iku = (-iku^{\text{inc}} + \partial_n u^{\text{inc}}) \quad \text{on } \partial\Omega$$

We then create a conformal, triangular mesh of our domain $\Omega_h = \{K\}$ (with cells K) and an associated mesh skeleton Υ_h.

We define our mesh skeleton:

$$\Upsilon_h = \begin{cases} \Upsilon_h^I & \text{On interior faces} \\ \Upsilon_h^B & \text{On boundary faces} \end{cases}$$

Remark 4.2.1. We remark that in the case of a curved scatterer or artificial boundary, we cannot have a mesh of straight lines which is perfectly conformal to the physical problem. In this case, there is an approximation extant in the mesh. We do not consider cases with curved mesh elements to guarantee convexity of each mesh element (this hypothesis is necessary for stability estimates, see 4.3.3).

We further will show that, contrary to standard FEM, there will *not* be any mesh constrains with PWDG, giving this method an important advantage that we do not restrict ourselves to the symmetric mesh as in Lemma 3.2.1. We demonstrate the well-posedness of this discrete problem as a consequence of Theorem 4.2.2.

We begin by defining the ideal "broken" Trefftz space, which consists of functions that are piecewise (locally) in the kernel of the Helmholtz operator:

$$\mathcal{T}(\Omega_h) = \{v \in L^2(\Omega) : v|_K \in H^2(K)\, s.t.\, (\nabla \cdot \epsilon^{-1}\nabla + k^2)v = 0, \forall K \in \Omega_h\} \tag{4.2}$$

We now introduce a discrete Trefftz space for our Galerkin method:

Definition 4.2.1: Discrete Trefftz Subspace

Let $V^p(\Omega_h) \subset \mathcal{T}(\Omega_h)$ denote a finite-dimensional subspace of dimension p, composed of functions which, on each cell $K \in \Omega_h$, lie in the kernel of the local operator:

$$(\nabla \cdot (\epsilon^{-1}\nabla) + k^2)v = 0.$$

That is, for each $v \in V^p(\Omega_h)$, we have $v|_K \in H^2(K)$ and $(\nabla \cdot (\epsilon^{-1}\nabla) + k^2)v = 0$ on K. The concrete choice of basis functions for $V^p(\Omega_h)$ (in our case, plane waves see §4.4) will be specified later.

Here and throughout this chapter, dx and ds will be used to indicate surface and volume integrals, respectively. On each element $K \in \Omega_h$, we introduce the test function $v \in V^p(\Omega_h)$ and integrate by parts twice:

$$\int_K \overline{v}\big(\nabla \cdot (\epsilon^{-1}\nabla u) + k^2 u\big)\, dx = \int_K \overline{v}\, \nabla \cdot (\epsilon^{-1}\nabla u)\, dx + \int_K k^2 u\, \overline{v}\, dx,$$

Looking at the first term:

$$\int_K \overline{v}\, \nabla \cdot (\epsilon^{-1}\nabla u)\, dx = \int_{\partial K} \overline{v}\, (\epsilon^{-1}\nabla u \cdot n)\, ds - \int_K \nabla\overline{v} \cdot (\epsilon^{-1}\nabla u)\, dx,$$

the last term yields:

$$-\int_K \nabla\overline{v} \cdot (\epsilon^{-1}\nabla u)\, dx = -\Big[\int_{\partial K} u\, (\epsilon^{-1}\nabla\overline{v} \cdot n)\, ds - \int_K u\, \nabla \cdot (\epsilon^{-1}\nabla\overline{v})\, dx\Big],$$

Then:

$$\int_K \overline{v}\big(\nabla \cdot (\epsilon^{-1}\nabla u) + k^2 u\big)\, dx = \int_{\partial K} \big(\epsilon^{-1}\overline{v}\, \partial_n u - \epsilon^{-1}\partial_n\overline{v}\, u\big)\, ds + \int_K u\big(\nabla \cdot (\epsilon^{-1}\nabla\overline{v}) + k^2\overline{v}\big)\, dx.$$

By the Trefftz property $\nabla \cdot (\epsilon^{-1}\nabla\overline{v}) + k^2\overline{v} = 0$, the volume term vanishes, yielding the boundary-only elemental form:

$$\int_{\partial K}\Big(\epsilon^{-1}\overline{v}\,\partial_n u - \epsilon^{-1}\partial_n\overline{v}\,u\Big)\,ds = 0. \tag{4.3}$$

Notation. Throughout this chapter we use the following convention: Every appearance of $\|u\|_D$ for some domain D means $\|u\|_{L^2(D)}$. We additionally define:

$$A \lesssim B \quad \Longleftrightarrow \quad \exists\, C > 0 \text{ s.t. } A \leq C\,B.$$

We must also define the notation of jumps and averages. Let

$$\varphi : \Omega_h \;\to\; \mathbb{C}, \quad \boldsymbol{\tau} : \Omega_h \;\to\; \mathbb{C}^2$$

be piecewise-defined scalar and vector functions, respectively. On each interior face $e = \partial K^+ \cap \partial K^-$, let n^+ be the outward normal of K^+ and $n^- = -n^+$. On a boundary face we set $K^- = \emptyset$, so that $\varphi^- = \varphi^+ = \varphi$ and $n^- = n^+ = n$.

$$\begin{aligned} &\text{Interior face:} && \{\{\varphi\}\} = \tfrac{1}{2}\big(\varphi^+ + \varphi^-\big), && [\![\varphi]\!] = \varphi^+\,n^+ + \varphi^-\,n^-, \\ &\text{Boundary face:} && \{\{\varphi\}\} = \varphi, && [\![\varphi]\!] = \varphi\,n. \end{aligned}$$

$$\begin{aligned} &\text{Interior face:} && \{\{\boldsymbol{\tau}\}\} = \tfrac{1}{2}\big(\boldsymbol{\tau}^+ + \boldsymbol{\tau}^-\big), && [\![\boldsymbol{\tau}]\!] = \boldsymbol{\tau}^+\cdot n^+ \;+\; \boldsymbol{\tau}^-\cdot n^-, \\ &\text{Boundary face:} && \{\{\boldsymbol{\tau}\}\} = \boldsymbol{\tau}, && [\![\boldsymbol{\tau}]\!] = \boldsymbol{\tau}\cdot n. \end{aligned}$$

Summing over all cells converts (4.3) into integrals over the skeleton $\Upsilon_h = \Upsilon_h^I \cup \Upsilon_h^B$. Therefore, we note:

$$\sum_K \int_{\partial K} \epsilon^{-1}\overline{v}\,\partial_n u\,ds = \int_{\Upsilon_h^I} \epsilon^{-1}\partial_n u\,[\![\overline{v}]\!]\,ds + \int_{\Upsilon_h^B} \epsilon^{-1}\partial_n u\,\overline{v}\,ds,$$

and similarly for the $\partial_n\overline{v}\,u$ term. However, this leaves us with the contribution of u and its trace undefined (in a discrete sense) over the skeleton edges. To complete the Trefftz-DG formulation, we must specify how to evaluate the *numerical traces*—i.e., the approximations of the solution and flux at the interfaces between elements. These traces are critical for coupling the discontinuous local solutions across the mesh and enforcing the weak continuity conditions in a consistent and stable way. To do this, we replace the exact traces $u^{\pm}$ and $(\frac{1}{ik}\epsilon^{-1}\nabla u)^{\pm}$ on each face by (to be precisely defined) numerical fluxes

$$\hat{u} \quad \text{(the solution trace)}, \quad \hat{\sigma} \quad \text{(the flux trace)}.$$

Here, the "hat" denotes a face-based numerical flux: $\hat{u}$ approximates $u|_{\partial K}$, and $ik\hat{\sigma}$ approximates $\epsilon^{-1}\nabla u$ on the mesh skeleton.

By plugging in these fluxes, we obtain the discrete problem:

Problem 4.2.1. Find $u \in V^p(\Omega_h)$ such that

$$A_h(u,v) = f_h(v) \quad \forall v \in V^p(\Omega_h), \tag{4.4}$$

where the sesquilinear form $A_h(u,v)$ is defined by

$$A_h(u,v) = \int_{\Upsilon_h^I} \big((ik\hat{\sigma}\cdot n)[\![\overline{v}]\!] - [\![\epsilon^{-1}\nabla\overline{v}]\!]\;\hat{u}\big)\,ds + \int_{\Upsilon_h^B} \big((ik\hat{\sigma}\cdot n)\,\overline{v} - (\epsilon^{-1}\partial_n\overline{v})\,\hat{u}\big)\,ds, \tag{4.5}$$

with right-hand side

$$f_h(v) = \int_{\Upsilon_h^B} \left(-ik\, u^{\mathrm{inc}} + \partial_n u^{\mathrm{inc}}\right) \overline{v}\, ds. \tag{4.6}$$

We mus now define the fluxes we have chosen. Here we choose to adopt a *primal flux formulation*, introduced in Hiptmair et al. (2016); Moiola (2011), which has proven effective in Trefftz-DG schemes for the Helmholtz equation.

Let $\delta, \alpha, \beta \in \mathbb{R}^+$, with $0 \ll \alpha, \beta$.

Definition 4.2.2: Primal Fluxes

$$ik\hat{\sigma} = \begin{cases} \{\{\epsilon^{-1}\nabla u\}\} - \alpha ik[\![u]\!], & \text{in } \Upsilon_h^I \\ \nabla u - (1-\delta)(-\nabla u + ikun + B_k n), & \text{in } \Upsilon_h^B \end{cases}$$

$$\hat{u} = \begin{cases} \{\{u\}\} - \frac{1}{ik}\beta[\![\epsilon^{-1}\nabla u]\!], & \text{in } \Upsilon_h^I \\ u - \delta\left(\frac{1}{ik}\partial_n u - u - \frac{1}{ik}B_k\right), & \text{in } \Upsilon_h^B \end{cases}$$

On boundary faces, the terms involving δ enforce a consistent approximation of the first-order absorbing boundary condition, which approximates the Sommerfeld radiation condition:

$$B_k = \partial_n u - iku = -iku^{\mathrm{inc}} + \partial_n u^{\mathrm{inc}}.$$

Finally, we note that the fluxes are consistent: if u is a sufficiently regular solution of the continuous problem, then

$$\hat{u} = u, \qquad ik\hat{\sigma} = \epsilon^{-1}\nabla u,$$

For further details on the choice of these numerical fluxes, we direct readers to Appendix C.

Plugging the fluxes into the elemental form (4.5) and summing over all elements yields the following weak form:

Problem 4.2.2. Find $u \in V^p(\Omega_h)$ such that

$$A_h(u,v) = f_h(v) \quad \forall v \in V^p(\Omega_h), \tag{4.7}$$

where the sesquilinear form $A_h(u,v)$ is defined by

$$\begin{aligned} A_h(u,v) = &\int_{\Upsilon_h^I} \left(-[\![\overline{v}]\!] \cdot \{\{\epsilon^{-1}\nabla u\}\} + \alpha i k [\![\overline{v}]\!] \cdot [\![u]\!]\right)\, ds \\ &+ \int_{\Upsilon_h^I} \left([\![\epsilon^{-1}\nabla \overline{v}]\!] \cdot \{\{u\}\} + \frac{i}{k}\beta [\![\epsilon^{-1}\nabla \overline{v}]\!] \cdot [\![\epsilon^{-1}\nabla u]\!]\right)\, ds \\ &+ \int_{\Upsilon_h^B} \left((1-\delta) i k u \overline{v} - (2-\delta)\partial_n u\, \overline{v}\right)\, ds \\ &+ \int_{\Upsilon_h^B} \left((1+\delta) u\, \partial_n \overline{v} + \delta \frac{i}{k} \partial_n u\, \partial_n \overline{v}\right)\, ds, \end{aligned} \tag{4.8}$$

and the linear functional $f_h(v)$ is given by

$$f_h(v) = -\int_{\Upsilon_h^B} B_k \left(\frac{\delta}{ik}\partial_n \overline{v} + (1-\delta)\overline{v}\right)\, ds. \tag{4.9}$$

4.2.2. Well-posedness of the PWDG Problem

We now seek to show uniqueness and the existence of unique solutions to problem 4.2.2. We will accomplish this in two steps

1. Firstly, we will show $A_h(\cdot,\cdot)$ is associated to a (to be defined) norm $\|\cdot\|_{DG}$.

2. Secondly, we will show coercivity in this norm for $A_h(\cdot,\cdot)$ on the finite diminesional space $v \in V^p(\Omega_h)$.

To that aim, we being by introducing a semi-norm given by the weak form.

Definition 4.2.3: DG Semi-Norm

Define the DG semi-norm associated to the weak form, $|v|_{DG,h}$, by:

$$|v|_{DG,h}^2 = k^{-1}\|\beta^{1/2}[\![\epsilon^{-1}\nabla v]\!]\|_{\Upsilon_h^I}^2 + k\|\alpha^{1/2}[\![v]\!]\|_{\Upsilon_h^I}^2$$
$$+ k\|(1-\delta)^{1/2}v\|_{\Upsilon_h^B}^2 + k^{-1}\|\delta^{1/2}\partial_n v\|_{\Upsilon_h^B}^2$$

Lemma 4.2.1. Let $v \in V^p(\Omega_h)$. Then, $|v|_{DG} = \Im(A_h(v,v))$

Proof. We note that, by integration by parts and the Trefftz property of v,

$$0 = \sum_K \int_K (\nabla \epsilon^{-1}\nabla u + k^2 u)\bar{v}dx$$
$$= k^2(u,v)_\Omega - \int_\Omega \epsilon^{-1}\nabla u \nabla \bar{v}dx + \sum_K \int_{\partial K}(\partial_n u\bar{v})dx$$
$$= k^2(u,v)_\Omega - \int_\Omega \epsilon^{-1}\nabla u \nabla \bar{v}dx + \int_{\Upsilon_h^I}\{\{\epsilon^{-1}\nabla u\}\}[\![\bar{v}]\!]ds + \int_{\Upsilon_h^I}\{\{\bar{v}\}\}[\![\epsilon^{-1}\nabla u]\!]ds + \int_{\Upsilon_h^B}\partial_n uvds$$

Therefore,

$$\begin{aligned}
A_h(u,v) &= \int_{\Upsilon_h^I}\left(-[\![\overline{v}]\!]\cdot\{\{\epsilon^{-1}\nabla u\}\} + \alpha ik[\![\overline{v}]\!]\cdot[\![u]\!]\right)\,ds \\
&+ \int_{\Upsilon_h^I}\left([\![\epsilon^{-1}\nabla\overline{v}]\!]\cdot\{\{u\}\} + \frac{i}{k}\beta[\![\epsilon^{-1}\nabla\overline{v}]\!]\cdot[\![\epsilon^{-1}\nabla u]\!]\right)\,ds \\
&+ \int_{\Upsilon_h^B}\left((1-\delta)iku\overline{v} - (2-\delta)\partial_n u\,\overline{v}\right)\,ds \\
&+ \int_{\Upsilon_h^B}\left((1+\delta)u\,\partial_n\overline{v} + \delta\frac{i}{k}\partial_n u\,\partial_n\overline{v}\right)\,ds, \\
&+ k^2(u,v)_\Omega - \int_\Omega \epsilon^{-1}\nabla u\nabla\bar{v}dx \\
&+ \int_{\Upsilon_h^I}\{\{\epsilon^{-1}\nabla u\}\}[\![\bar{v}]\!]ds + \int_{\Upsilon_h^I}\{\{\bar{v}\}\}[\![\epsilon^{-1}\nabla u]\!]ds + \int_{\Upsilon_h^B}\partial_n uvds
\end{aligned}$$

We get the cancellation:

$$
\begin{aligned}
A_h(u,v) = & \cancel{\int_{\Upsilon_h^I} -[\![\overline{v}]\!] \cdot \{\{\epsilon^{-1}\nabla u\}\} ds} + \int_{\Upsilon_h^I} \alpha i k [\![\overline{v}]\!] \cdot [\![u]\!]\, ds \\
& + \int_{\Upsilon_h^I} \left([\![\epsilon^{-1}\nabla\overline{v}]\!] \cdot \{\{u\}\} + \frac{i}{k}\beta [\![\epsilon^{-1}\nabla\overline{v}]\!] \cdot [\![\epsilon^{-1}\nabla u]\!] \right) ds \\
& + \int_{\Upsilon_h^B} \left((1-\delta) i k u \overline{v} - (2-\delta)\partial_n u\, \overline{v}\right) ds \\
& + \int_{\Upsilon_h^B} \left((1+\delta) u\, \partial_n \overline{v} + \delta \frac{i}{k} \partial_n u\, \partial_n \overline{v} \right) ds, \\
& + k^2 (u,v)_\Omega - \int_\Omega \epsilon^{-1} \nabla u \nabla \bar{v} dx \\
& + \cancel{\int_{\Upsilon_h^I} \{\{\epsilon^{-1}\nabla u\}\} [\![\overline{v}]\!] ds} + \int_{\Upsilon_h^I} \{\{\bar{v}\}\} [\![\epsilon^{-1}\nabla u]\!] ds + \int_{\Upsilon_h^B} \partial_n u v ds
\end{aligned}
$$

We then consider $A_h(u,u)$ here:

$$
\begin{aligned}
A_h(u,u) = & 2\Re\Big(\int_{\Upsilon_h^I} [\![\epsilon^{-1}\nabla\overline{u}]\!] \cdot \{\{u\}\}\, ds + \int_{\Upsilon_h^I} \{\{\bar{u}\}\} [\![\epsilon^{-1}\nabla u]\!]\, ds\Big) \\
& + \int_{\Upsilon_h^I} \left(\alpha i k [\![\overline{u}]\!] \cdot [\![u]\!] + \frac{i}{k}\beta [\![\epsilon^{-1}\nabla\overline{u}]\!] \cdot [\![\epsilon^{-1}\nabla u]\!] \right) ds \\
& + \int_{\Upsilon_h^B} \left((1-\delta) i k u \overline{u} - (2-\delta)\partial_n u\, \overline{u}\right) ds \\
& + \int_{\Upsilon_h^B} \left((1+\delta) u\, \partial_n \overline{u} + \delta \frac{i}{k} \partial_n u\, \partial_n \overline{u} \right) ds, \\
& + k^2 (u,u)_\Omega - \int_\Omega \epsilon^{-1} \nabla u \nabla \bar{u} + \int_{\Upsilon_h^B} \partial_n u \bar{u} ds
\end{aligned}
$$

We can further rewrite:

$$
\begin{aligned}
A_h(u,u) = 2\Re(&\int_{\Upsilon_h^I} [\![\epsilon^{-1}\nabla\overline{u}]\!] \cdot \{\{u\}\}\, ds) \\
&+ \int_{\Upsilon_h^I} \left(\alpha ik[\![\overline{u}]\!] \cdot [\![u]\!] + \frac{i}{k}\beta[\![\epsilon^{-1}\nabla\overline{u}]\!] \cdot [\![\epsilon^{-1}\nabla u]\!]\right) ds \\
&+ \int_{\Upsilon_h^B} ((1-\delta)iku\overline{u} - (2-\delta)\partial_n u\,\overline{u})\ ds \\
&+ \int_{\Upsilon_h^B} \left((1+\delta)u\,\partial_n\overline{u} + \delta\frac{i}{k}\partial_n u\,\partial_n\overline{u}\right) ds, \\
&+ k^2(u,u)_\Omega - \int_\Omega \epsilon^{-1}\nabla u\nabla\bar{u}dx + \int_{\Upsilon_h^B} \partial_n uuds
\end{aligned}
$$

Now all terms are strictly real or strictly imaginary with the exception of the following "crossing terms" $Q(u)$:

$$
\begin{aligned}
Q(u) &= \int_{\Upsilon_h^B} -(2-\delta)\partial_n u\,\overline{u}\ + (1+\delta)u\,\partial_n\overline{u} + \partial_n u\bar{u} \\
&= \int_{\Upsilon_h^B} = \delta\partial_n u\,\overline{u}\,ds + \delta u\,\partial_n\overline{u} + -\partial_n u\,\overline{u}\,ds + u\,\partial_n\overline{u}\,ds \\
&= 2\delta\Re\left(\int_{\Upsilon_h^B} \partial_n u\overline{u}\,ds\right) + 2i\Im\left(\int_{\Upsilon_h^B} \overline{\partial_n u}u\,ds\right)
\end{aligned}
$$

Now, (as in equation 11 of Bonnet-Ben Dhia et al. (2016))

$$
\nabla\cdot(\epsilon^{-1}\nabla u) + k^2u = 0 \qquad\qquad u \in V^p(\Omega_h),
$$

Multiplying by $\overline{u}$ and integrating by parts gives:

$$
\int_\Omega \epsilon^{-1}|\nabla u|^2\,dx - k^2\int_\Omega |u|^2\,dx = \int_{\Upsilon_h^B} \partial_n u\,\overline{u}\,ds.
$$

Taking the imaginary part of both sides yields:

$$\Im\left(\int_{\Upsilon_h^B} \partial_n u\,\overline{u}\,ds\right) = 0.$$

Therefore, we can conclude that $\Im(Q(u)) = 0$ and furthermore that $\Im(A_h(u,u)) = |u|_{DG}$. $\square$

Lemma 4.2.2. $|\cdot|_{DG,h}$ is a norm in $\mathcal{T}(\Omega_h)$.

Proof. Let $v \in \mathcal{T}(\Omega_h)$. Firstly, we note that if $v = 0$, it follows that $|v|_{DG,h} = 0$. Therefore, to show this seminorm is a norm, we need only to show $|v|_{DG,h} = 0 \implies v = 0$. $|v|_{DG,h} = 0 \implies k\left\|(1-\delta)^{1/2}v\right\|^2_{\Upsilon_h^B} = 0 \implies v = 0$ on Υ_h^B and $k^{-1}\left\|\delta^{1/2}\partial_n v\right\|^2_{\Upsilon_h^B} = 0 \implies \partial_n v = 0$ on Υ_h^B.

Furthermore, given $v \in \mathcal{T}(\Omega_h)$, it is such that it satisfies the Helmholtz-like problem (4.1), thus we have $(\nabla\cdot(\epsilon^{-1}\nabla)+k^2)v = 0$ in Ω and $[v]_{\partial\mathcal{D}} = 0$, $[\epsilon^{-1}\partial_n v]_{\partial\mathcal{D}} = 0$ on $\partial\mathcal{D}$. If we can show this problem is Fredholm, the Holmgren uniqueness theorem will allow us to conclude $v = 0$.

In essence, therefore, we desire to show the following problem is Fredholm:

Find $u \in H_0^1(\Omega)$ such that

$$\begin{aligned}
&\left(\nabla\cdot\left(\epsilon^{-1}\nabla\right)+k^2\right)u = 0 \quad \text{in} \quad \Omega,\\
&[u]_{\partial\mathcal{D}} = 0, \quad [\epsilon^{-1}\partial_n u]_{\partial\mathcal{D}} = 0 \quad \text{on} \quad \partial\mathcal{D}\\
&\partial_n u = 0 \quad \text{on} \quad \partial\Omega.
\end{aligned}$$

This has associated weak form $a(u,v)+b(u,v) = 0$, with

$$a(u,v) = \int_\Omega \epsilon^{-1}\nabla u\cdot\overline{\nabla v}\,dx,$$

and

$$b(u,v) = -k^2 \int_\Omega u\bar{v}\, dx$$

a compact perturbation.

For $\epsilon > 0$, $a(\cdot,\cdot)$ is classically coercive, therefore the problem is Fredholm. Ciarlet (1978)

For $\epsilon < 0$, we recall the T-coercivity given by (3.2.1), and note that as a consequence of this proof, $a(\cdot, T\cdot)$ is T-coercive, and therefore this problem is Fredholm. This allows us to conclude via the Holmgren uniqueness theorem that $v = 0$, thus $|\cdot|_{DG,h}$ is a norm. □

Theorem 4.2.1. If $|v|_{DG,h}$ is a norm on $V^p(\Omega_h)$, then the weak form:
Find $u \in V^p(\Omega_h)$ such that $A_h(u,v) = f_h(v), \;\forall v \in V^p(\Omega_h)$ has a unique solution.

Proof. Assume $A_h(u,v) = 0$ for all $v \in V^p(\Omega_h)$ and that $|v|_{DG} = \|v\|_{DG}$ is a norm. Then:

- $A_h(u,u) = 0 \implies \Im(A_h(u,u)) = 0 \implies \|u\|_{DG,h} = 0 \implies u = 0.$
- For u_1, u_2 such that $A_h(u_1,v) = A_h(u_2,v)$ for all v, $A_h(u_1 - u_2, v) = 0 \implies \|u_1 - u_2\|_{DG,h} = 0 \implies u_1 - u_2 = 0$

Thus, uniqueness is guaranteed.

To show existence, we point out that since $V^p(\Omega_h)$ is finite-dimensional, injectivity of the linear map

$$u \;\mapsto\; (v \mapsto A_h(u,v)) \;:\; V^p(\Omega_h) \to (V^p(\Omega_h))'$$

already forces surjectivity. In other words, once we've shown $\Im(A_h(u,u)) = \|u\|^2_{DG,h}$ implies $u = 0$, the resulting square matrix is nonsingular, so *existence* follows automatically. □

As a consequence of this theorem, we note that $V^p(\Omega_h) \subset \mathcal{T}(\Omega_h)$ and therefore our discrete problem has a unique solution.

4.3. Error Analysis

Here we focus on a study of the general error agnostic of our choice of basis functions. Recalling the norm given in 4.2.3, we define the Augmented norm:

Definition 4.3.1: Augmented DG norm

For $v \in \mathcal{T}(\Omega_h)$, define the augmented DG norm by:

$$\|v\|^2_{DG+,h} := \|v\|_{DG,h} + k \left\|\beta^{-1/2}\{\{v\}\}\right\|^2_{\Upsilon_h^I} + k^{-1}\left\|\alpha^{-1/2}\{\{\epsilon^{-1}\nabla v\}\}\right\|^2_{\Upsilon_h^I} + k\left\|\delta^{-1/2}\, v\right\|^2_{\Upsilon_h^B}.$$

We remark that $V^p(\Omega_h)$ will inherit this norm as well. Now we refer to the quasi-optimality result of Moiola (see Proposition 4.3.5-6 Moiola (2011)):

Lemma 4.3.1. Continuity of A_h, cf. Moiola Prop. 4.3.5

For all $w_1, w_2 \in V^p(\Omega_h)$,

$$\left|A_h(w_1, w_2)\right| \lesssim \delta_{\max}||w_1||_{DG+,h}\ \ ||w_2||_{DG,h}.$$

For $\delta_{\max} = \max\{|1-\delta|, |2-\delta|, (1+\delta), 2\}$.

Proof. We split $A_h(w_1, w_2)$ into the eight contributions from interior (Υ_h^I) and boundary (Υ_h^B) faces and bound each with weighted Cauchy–Schwarz.

Interior terms:

$$
\begin{aligned}
I_1 &= \int_{\Upsilon_h^I} -[\![\overline{w}_2]\!] \cdot \{\{\epsilon^{-1}\nabla w_1\}\}\, ds, &\quad \big|I_1\big| &\le \|\beta^{1/2}[\![w_2]\!]\|_{\Upsilon_h^I} \left\|\beta^{-1/2}\{\{\epsilon^{-1}\nabla w_1\}\}\right\|_{\Upsilon_h^I},\\
I_2 &= \int_{\Upsilon_h^I} \alpha\, ik\, [\![\overline{w}_2]\!] \cdot [\![w_1]\!]\, ds, &\quad \big|I_2\big| &\le k\,\alpha\, \|[\![w_2]\!]\|_{\Upsilon_h^I} \|[\![w_1]\!]\|_{\Upsilon_h^I},\\
I_3 &= \int_{\Upsilon_h^I} [\![\epsilon^{-1}\nabla\overline{w}_2]\!] \cdot \{\{w_1\}\}\, ds, &\quad \big|I_3\big| &\le \|\beta^{1/2}\{\{w_1\}\}\|_{\Upsilon_h^I} \left\|\beta^{-1/2}[\![\epsilon^{-1}\nabla w_2]\!]\right\|_{\Upsilon_h^I},\\
I_4 &= \int_{\Upsilon_h^I} \frac{i}{k}\,\beta\, [\![\epsilon^{-1}\nabla\overline{w}_2]\!] \cdot [\![\epsilon^{-1}\nabla w_1]\!]\, ds, &\quad \big|I_4\big| &\le \frac{1}{k}\,\beta \left\|[\![\epsilon^{-1}\nabla w_2]\!]\right\|_{\Upsilon_h^I} \left\|[\![\epsilon^{-1}\nabla w_1]\!]\right\|_{\Upsilon_h^I}.
\end{aligned}
$$

And as by def Definition 4.2.3:

$$
\|\beta^{1/2}[\![\epsilon^{-1}\nabla w]\!]\|_{\Upsilon_h^I} \;\le\; k^{1/2}\; \|w\|_{DG,h}\,, \quad \|[\![w]\!]\|_{\Upsilon_h^I} \;\le\; (k\,\alpha)^{-1/2}\; \|w\|_{DG,h}\,,
$$

$$
\|\beta^{-1/2}\{\{w\}\}\|_{\Upsilon_h^I} \;\le\; k^{-1/2}\; \|w\|_{DG+,h}\,, \quad \|\alpha^{-1/2}\{\epsilon^{-1}\nabla w\}\|_{\Upsilon_h^I} \;\le\; k^{1/2}\; \|w\|_{DG+,h}\,.
$$

Each interior term I_j is a product of one of the first two norms with one of the last two, and hence for $j = 1, \dots, 4$

$$
|I_j| \;\lesssim\; \|w_1\|_{DG+,h} \;\; \|w_2\|_{DG,h}\,.
$$

Boundary terms:

$$
\begin{aligned}
B_1 &= \int_{\Upsilon_h^B} (1-\delta)\, ik\, w_1\, \overline{w}_2\, ds, &\quad \big|B_1\big| &\le k\,|1-\delta|\, \|w_1\|_{\Upsilon_h^B} \|w_2\|_{\Upsilon_h^B},\\
B_2 &= \int_{\Upsilon_h^B} -(2-\delta)\, \partial_n w_1\, \overline{w}_2\, ds, &\quad \big|B_2\big| &\le |2-\delta|\, \|\partial_n w_1\|_{\Upsilon_h^B} \|w_2\|_{\Upsilon_h^B},\\
B_3 &= \int_{\Upsilon_h^B} (1+\delta)\, w_1\, \partial_n\overline{w}_2\, ds, &\quad \big|B_3\big| &\le (1+\delta)\, \|w_1\|_{\Upsilon_h^B} \|\partial_n w_2\|_{\Upsilon_h^B},\\
B_4 &= \int_{\Upsilon_h^B} \delta\, \frac{i}{k}\, \partial_n w_1\, \partial_n\overline{w}_2\, ds, &\quad \big|B_4\big| &\le \frac{\delta}{k}\, \|\partial_n w_1\|_{\Upsilon_h^B} \|\partial_n w_2\|_{\Upsilon_h^B}.
\end{aligned}
$$

We define $\delta_{\max} = \max\{|1-\delta|, |2-\delta|, 1+\delta, \delta\}$, and each of $\|w\|_{\Upsilon_h^B}$, $\|\partial_n w\|_{\Upsilon_h^B}$ is controlled by $\|w\|_{DG,h}$ or $\|w\|_{DG+,h}$. Thus $\big|B_j\big| \lesssim \delta_{\max}\, \|\, |w_1\|_{DG+,h}\, |w_2|_{DG,h}$.

Putting all eight pieces together gives

$$\left|A_h(w_1, w_2)\right| \lesssim \sum_{j=1}^{4} |I_j| + \sum_{j=1}^{4} |B_j| \lesssim \delta_{\max} \|w_1\|_{DG+,h} \|w_2\|_{DG,h},$$

as claimed. $\square$

We now prove the quasi-optimality of this method in the augmented norm.

Lemma 4.3.2 (Quasi-optimality). Let u be the exact solution of the Helmholtz problem and $u_h \in V^p(\Omega_h)$ its Trefftz–DG approximation. Then:

$$||u - u_h||_{DG,h} \lesssim (1 + \delta_{\max}) \inf_{v_h \in V^p(\Omega_h)} ||u - v_h||_{DG+,h},$$

where $\delta_{\max} = \max\{|1 - \delta|, |2 - \delta|, (1 + \delta), 2\}$.

Proof. For any $v_h \in V^p(\Omega_h)$, the triangle inequality gives

$$||u - u_h||_{DG,h} \leq ||u - v_h||_{DG,h} + ||v_h - u_h||_{DG,h}. \tag{4.10}$$

By coercivity (Theorem 4.2.1) and Galerkin orthogonality one has

$$||v_h - u_h||^2_{DG,h} = \Im\left(A_h(v_h - u_h,\ v_h - u_h)\right) = \Im\left(A_h(u - v_h,\ u_h - v_h)\right).$$

From the quasi-optimality (Lemma 4.3.1)

$$\left|A_h(w_1, w_2)\right| \lesssim \delta_{\max} ||w_1||_{DG+,h} \ ||w_2||_{DG,h}$$

yields

$$||v_h - u_h||^2_{DG,h} \lesssim \delta_{\max} ||u - v_h||_{DG+,h} \ ||v_h - u_h||_{DG,h},$$

so

$$||v_h - u_h||_{DG,h} \lesssim \delta_{\max} ||u - v_h||_{DG+,h}.$$

Since $||u - v_h||_{DG,h} \leq ||u - v_h||_{DG+,h}$, combining with the triangle bound (4.10) gives

$$\|u - u_h\|_{DG,h} \lesssim (1 + \delta_{\max}) \|u - v_h\|_{DG+,h} \quad \forall\, v_h \in V^p(\Omega_h).$$

Taking the infimum over $v_h \in \mathcal{T}(\Omega_h)$ completes the proof. $\square$

We must now prove a technical result - stability estimates for the sign changing problem. We derive a similar result for the homogeneous Helmholtz problem given in Proposition 8.1.4 of J. M. Melenk (1995), but with the novelty of the sign-changing problem and using the T-coercivity machinery we introduced earlier in the thesis.

Lemma 4.3.3 (Stability Estimates for sign-changing Helmholtz). Let Ω be a bounded star-shaped (or convex) domain in $\mathbb{R}^2$ and let $\Phi \in L^2(\Omega)$. Assume we have a scatterer entirely contained within, and not touching the boundaries of Ω, $\mathcal{D} \subsetneq \Omega$ in which we may define

$$\epsilon(x) = \begin{cases} 1, & x \in E := \mathbb{R}^2 \setminus \overline{\mathcal{D}}, \\ \epsilon_m \in \mathbb{R}^* \setminus \mathfrak{C}, & x \in \mathcal{D}, \end{cases}$$

where $\mathfrak{C}$ is the geometry-dependent ill-posedness region for the T-coercive problem (4.1). We consider the adjoint problem:

$$\begin{aligned} &\text{Find } w \in H^1_{\text{loc}}(\mathbb{R}^2) \text{ such that:} \\ &\quad \nabla \cdot (\epsilon^{-1} \nabla w) + k^2 w = \Phi, \quad \text{in } \Omega, \\ &\quad [w]_{\partial\mathcal{D}} = 0, \quad [\epsilon^{-1} \partial_n w]_{\partial\mathcal{D}} = 0, \\ &\quad \partial_n w + ikw = 0 \quad \text{on } \partial\Omega, \\ &\quad k \in \mathbb{R}^+. \end{aligned} \tag{4.11}$$

Let C_1, C_2 be constants dependent on the diameter of Ω, the value of ϵ_m and the shape of the scatterer, but not dependent on k. Then the solution $w \in H^1_{\text{loc}}(\mathbb{R}^2)$ of (4.11) satisfies, for all $k > 0$,

$$\begin{aligned} k \left\| w \right\|_\Omega \; + \; \left\| \nabla w \right\|_\Omega \; &\leq \; C_1 \left\| \Phi \right\|_\Omega, \\ \left\| \nabla \cdot \epsilon^{-1} \nabla w \right\|_\Omega \; &\leq \; C_2 \left(1 + k \right) \left\| \Phi \right\|_\Omega. \end{aligned} \tag{4.12}$$

Proof. Assume, as in the statement of the proof, we have a domain Ω with an interior scatterer $\mathcal{D}$. We decompose:

$$\Omega = \mathcal{D} \cup \left(\Omega \setminus \overline{\mathcal{D}} \right)$$

with $\mathcal{D}$ and $\Omega \setminus \overline{\mathcal{D}}$ sharing the interface $\partial\mathcal{D} \cap \Omega$. Suppose that on $\mathcal{D}$ we have $\epsilon = \epsilon_m$ (constant) and on $\Omega \setminus \overline{\mathcal{D}}$ we have $\epsilon = 1$. Moreover, the transmission (jump) conditions

$$[w]_{\partial\mathcal{D}} = 0, \quad \left[\epsilon^{-1}\partial_n w\right]_{\partial\mathcal{D}} = 0$$

hold on $\partial\mathcal{D}$. We define subregions of Ω for brevity:

$$\Omega_1 = \mathcal{D}, \quad \Omega_2 = \Omega \setminus \overline{\mathcal{D}}. \tag{4.13}$$

We recall the problem 4.11 which has associated weak form:

$$\begin{aligned}
&\text{Find } w \in H^1_{\text{loc}}(\mathbb{R}^2) \text{ such that, for all test functions } v \in H^1_{\text{loc}}(\mathbb{R}^2),\\
&\mathfrak{A}(w, v) = \mathfrak{f}(v) \text{ given by:}\\
&-\int_\Omega \epsilon^{-1}\nabla w \cdot \overline{\nabla v}\, dx + k^2 \int_\Omega w\, \overline{v}\, dx - \int_{\partial\Omega} ik\, w\, \overline{v}\, ds = \int_\Omega \Phi\, \overline{v}\, dx.
\end{aligned} \tag{4.14}$$

We note that this weak form is a compact perturbation of a T-coercive form, and is therefore T-coercive - see 3.2.1. This implies the existence of a bounded involution $T : H^1(\Omega) \to H^1(\Omega)$ and an $\alpha_T, \beta_T > 0$, independent of k, such that

$$|\,(\mathfrak{A}\big(v,\, Tv\big)| \geq \alpha_T\, \|v\|^2_{H^1(\Omega)} - \beta_T ||v||^2_\Omega \quad \forall v \in H^1(\Omega).$$

We assume that T is local to some cutoff function about the interface which does not touch the boundary. Applying $Tw = v$ in the weak form, and considering the imaginary component, we get:

$$\Im(\mathfrak{A}(w, Tw)) = k\, \|w\|^2_{\partial\Omega} \;\; \leq \;\; |\int_\Omega \Phi\, \overline{w}\, dx,| \tag{4.15}$$

and

$$\alpha_T\, \|w\|^2_{H_1(\Omega)} - k^2\beta_T ||w||^2_\Omega \leq |\int_\Omega \Phi\, \overline{w}\, dx| \tag{4.16}$$

By Young's inequality, for any $1 \geq s > 0$,

$$\Big|\int_\Omega \Phi\, \overline{w}\, dx\Big| \;\leq\; \|\Phi\|_\Omega \, \|w\|_\Omega \;\leq\; \frac{1}{2s}\, \|\Phi\|^2_\Omega + \frac{s}{2}\, \|w\|^2_\Omega.$$

Hence from (4.15),

$$k\, \|w\|^2_{\partial\Omega} \;\leq\; \frac{1}{2s}\; \|\Phi\|^2_\Omega + \frac{s}{2}\, \|w\|^2_\Omega\, .$$

Now set

$$s \;=\; k\, \nu, \quad \text{with} \quad 0 < \nu \leq \frac{1}{k} \iff 0 < s \leq 1.$$

Multiplying the previous line by k gives

$$k^2\, \|w\|^2_{\partial\Omega} \;\leq\; \frac{1}{2\nu}\; \|\Phi\|^2_\Omega + \frac{\nu\, k^2}{2}\; \|w\|^2_\Omega\, . \tag{4.17}$$

Define $C_T = \max\{\left|\frac{\beta_T}{\alpha_T} - \frac{1}{\alpha\, T}\right|, 1\}$. We see (continuing from (4.16)):

$$\alpha_T\|\nabla w\|^2_\Omega + (\alpha_T - \beta_T)\|w\|^2_\Omega \leq \left|\int_\Omega \Phi\, \overline{w}\, dx\right|,$$
$$\|\nabla w\|^2_\Omega \leq \left|\tfrac{\beta_T}{\alpha_T} - \tfrac{1}{\alpha_T}\right| \|w\|^2_\Omega + \left|\int_\Omega \Phi\, \overline{w}\, dx\right|$$

Using Young's with $0 \leq \nu \leq 1$:

$$\leq C_T\Big(\|w\|^2_\Omega + \tfrac{1}{2\nu} k^{-2}\|\Phi\|^2_\Omega + \tfrac{\nu}{2}\, k^2\|w\|^2_\Omega\Big)$$
$$\|\nabla w\|^2_\Omega \leq C_T\Big((1 + \frac{\nu}{2}k^2)\|w\|^2_\Omega + \frac{1}{2\nu}(1 + k^{-2})\|\Phi\|^2_\Omega\Big). \tag{4.18}$$

Now we pick as a test function

$$v(\mathbf{x}) = \mathbf{x}\; \cdot\; \epsilon^{-1}\nabla\big(w(\mathbf{x})\big). \quad \text{(Here } \mathbf{x} = [x_1, x_2].)$$

We remark that this assumes sufficient regularity to define $\mathbf{x} \cdot \epsilon^{-1}\nabla w(\mathbf{x})$, which is only the case when $\mathbf{x}$ is normal to the scatter (when the scatterer is a disk). We assume that for this case. However, in the cases of a non-disk scatterer, it simply follows to split the weak form over the two subregions (4.13) and recombine using the transmission conditions, e.g. (4.24). We then compute the real portion of the weak form $\Re\big(\mathfrak{A}(w,\ \mathbf{x}\cdot\epsilon^{-1}\nabla w)\big)$.

Define

$$u(\mathbf{x}) = \epsilon^{-1}\nabla w(\mathbf{x}),$$

and consider the test function

$$v(\mathbf{x}) = \mathbf{x}\cdot u(\mathbf{x}) = \mathbf{x}\cdot\big(\epsilon^{-1}\nabla w(\mathbf{x})\big).$$

Then the first term in the weak form is

$$I := \int_\Omega \epsilon^{-1}\nabla w \ \cdot\ \overline{\nabla\big(\mathbf{x}\cdot u(\mathbf{x})\big)}\, dx. \tag{4.19}$$

By the chain rule,

$$\nabla\big(\mathbf{x}\cdot u(\mathbf{x})\big) = u(\mathbf{x}) + \mathbf{x}\cdot\nabla u(\mathbf{x}).$$

Thus,

$$I = \int_\Omega \epsilon^{-1}\nabla w \ \cdot\ \overline{u(\mathbf{x})}\, dx + \int_\Omega \epsilon^{-1}\nabla w \ \cdot\ \overline{\big(\mathbf{x}\cdot\nabla u(\mathbf{x})\big)}\, dx.$$

Since $\epsilon^{-1}\nabla w = u(\mathbf{x})$,

the first term becomes

$$I_1 = \int_\Omega \big|u(\mathbf{x})\big|^2\, dx. \tag{4.20}$$

For the second term, define

$$I_2 = \int_\Omega u(\mathbf{x}) \ \cdot \ \overline{\big(\mathbf{x} \cdot \nabla u(\mathbf{x})\big)}\, dx. \tag{4.21}$$

Taking the real part and applying the divergence identity (valid componentwise) to the vector field $\mathbf{x}\,|u(\mathbf{x})|^2$ (noting $\nabla \cdot \mathbf{x} = 2$) yields

$$2\,\Re\Big(I_2\Big) = \int_{\partial\Omega} \big(\mathbf{x} \cdot n(\mathbf{x})\big)|u(\mathbf{x})|^2\, ds - 2\int_\Omega |u(\mathbf{x})|^2\, dx.$$

Hence

$$\Re(I_2) = -\int_\Omega |u(\mathbf{x})|^2\, dx + \tfrac{1}{2}\int_{\partial\Omega} \big(\mathbf{x} \cdot n(\mathbf{x})\big)|u(\mathbf{x})|^2\, ds. \tag{4.22}$$

Adding $\Re(I_1)$ from (4.20) to $\Re(I_2)$ gives

$$\Re(I_1 + I_2) = \int_\Omega |u(\mathbf{x})|^2\, dx - \int_\Omega |u(\mathbf{x})|^2\, dx + \tfrac{1}{2}\int_{\partial\Omega} \big(\mathbf{x} \cdot n(\mathbf{x})\big)|u(\mathbf{x})|^2\, ds,$$

so that

$$\Re\Big(\int_\Omega \epsilon^{-1}\nabla w \ \cdot \ \overline{\nabla\big(\mathbf{x} \cdot \epsilon^{-1}\nabla w\big)}\, dx\Big) = \tfrac{1}{2}\int_{\partial\Omega} \big(\mathbf{x} \cdot n(\mathbf{x})\big)\epsilon^{-2}|\nabla w|^2\, ds. \tag{4.23}$$

We consider the mass term in the weak form

$$I_3 = -k^2 \int_\Omega w\, \overline{\Big(\mathbf{x} \cdot (\epsilon^{-1}\nabla w)\Big)}\, dx.$$

For brevity, define for $i = 1, 2$

$$J_i = \int_{\Omega_i} w\, \overline{\Big(\mathbf{x} \cdot \nabla w\Big)}\, dx,$$

with subregions of Ω defined in (4.13).

Since ϵ is piecewise constant, we split the integral:

$$I_3 = -k^2 \left[\int_{\Omega_1} w \, \overline{\left(\mathbf{x} \cdot (\epsilon_m^{-1} \nabla w) \right)} \, dx + \int_{\Omega_2} w \, \overline{\left(\mathbf{x} \cdot (\nabla w) \right)} \, dx \right]. \tag{4.24}$$

On each subdomain we can factor out the constant:

$$I_3 = -k^2 \left[\epsilon_m^{-1} \int_{\mathcal{D}} w \, \overline{\left(\mathbf{x} \cdot \nabla w \right)} \, dx + 1 \int_{\Omega \setminus \overline{\mathcal{D}}} w \, \overline{\left(\mathbf{x} \cdot \nabla w \right)} \, dx \right].$$

Then,

$$I_3 = -k^2 \left[\epsilon_m^{-1} J_1 + J_2 \right].$$

Now, we use the standard divergence identity for each subdomain. For a (possibly complex-valued) function w, consider the vector field

$$F(\mathbf{x}) = \mathbf{x} \, |w(\mathbf{x})|^2.$$

Then,

$$\nabla \cdot F(\mathbf{x}) = (\nabla \cdot \mathbf{x}) \, |w|^2 + \mathbf{x} \cdot \nabla(|w|^2) = 2 \, |w|^2 + 2 \, \Re\Big(w \, (\mathbf{x} \cdot \overline{\nabla w}) \Big),$$

since in $\mathbb{R}^2$ we have $\nabla \cdot \mathbf{x} = 2$. Integrate over Ω_i and apply the divergence theorem:

$$\int_{\Omega_i} \nabla \cdot F(\mathbf{x}) \, dx = \int_{\partial \Omega_i} (\mathbf{x} \cdot n_i) \, |w|^2 \, ds.$$

Thus,

$$2 \, \Re\Big(J_i \Big) = \int_{\partial \Omega_i} (\mathbf{x} \cdot n_i) \, |w|^2 \, ds - 2 \int_{\Omega_i} |w|^2 \, dx,$$

or equivalently,

$$\Re(J_i) = - \int_{\Omega_i} |w|^2 \, dx + \frac{1}{2} \int_{\partial \Omega_i} (\mathbf{x} \cdot n_i) \, |w|^2 \, ds. \tag{4.25}$$

Substituting (4.25) back into the expression for I_3, and taking the real part, we obtain

$$\Re(I_3) = k^2\left[\epsilon_m^{-1}\Big(\int_{\Omega_1} |w|^2\,dx - \frac{1}{2}\int_{\partial\Omega_1}(\mathbf{x}\cdot n_1)\,|w|^2\,ds\Big) + \Big(\int_{\Omega_2} |w|^2\,dx - \frac{1}{2}\int_{\partial\Omega_2}(\mathbf{x}\cdot n_2)\,|w|^2\,ds\Big)\right].$$

Here, the boundaries $\partial\Omega_1$ and $\partial\Omega_2$ consist of the external boundary $\partial\Omega$ and the internal interface $\partial\mathcal{D}$. The transmission conditions

$$[w]_{\partial\mathcal{D}} = 0, \quad [\epsilon^{-1}\partial_n w]_{\partial\mathcal{D}} = 0$$

ensure that the contributions from $\partial\mathcal{D}$ cancel when summing over the two subdomains. Therefore, we can write

$$\begin{aligned} \Re(I_3) &= k^2\Big[\epsilon_m^{-1}\int_{\Omega_1} |w|^2\,dx + \int_{\Omega_2} |w|^2\,dx\Big] - \frac{k^2}{2}\int_{\partial\Omega}(\mathbf{x}\cdot n)\,|w|^2\,ds. \\ \Re(I_3) &= k^2\int_{\Omega} \epsilon^{-1}(\mathbf{x})\,|w(\mathbf{x})|^2\,dx - \frac{k^2}{2}\int_{\partial\Omega}(\mathbf{x}\cdot n)\,|w|^2\,ds. \end{aligned} \tag{4.26}$$

Finally, combining the principal term (4.23) with the mass term (4.26), the real part of the weak form becomes the Rellich identity:

$$\begin{aligned} \Re\Big(\mathfrak{A}\big(w,\,\mathbf{x}\cdot\epsilon^{-1}\nabla w\big)\Big) &= \frac{1}{2}\int_{\partial\Omega}(\mathbf{x}\cdot n)\,|\nabla w|^2\,ds + k^2\int_{\Omega}\epsilon^{-1}|w|^2\,dx \\ &\quad - \frac{k^2}{2}\int_{\partial\Omega}(\mathbf{x}\cdot n)\,|w|^2\,ds + \Re\left(ik\int_{\partial\Omega} w\,\overline{\mathbf{x}\cdot\nabla w}\,ds\right) \\ &= \Re\left(\int_{\Omega}\Phi\,\overline{\mathbf{x}\cdot\epsilon^{-1}\nabla w}\,dx\right). \end{aligned} \tag{4.27}$$

$$\begin{aligned} & k^2\int_{\Omega}\epsilon^{-1}|w|^2\,dx \\ &= -\Re\left(ik\int_{\partial\Omega} w\,\overline{\mathbf{x}\cdot\nabla w}\,ds\right) + \frac{k^2}{2}\int_{\partial\Omega}(\mathbf{x}\cdot n)\,|w|^2\,ds \\ &- \frac{1}{2}\int_{\partial\Omega}(\mathbf{x}\cdot n)\,|\nabla w|^2\,ds + \Re\left(\int_{\Omega}\Phi\,\overline{\mathbf{x}\cdot\epsilon^{-1}\nabla w}\,dx\right). \end{aligned}$$

Using $\epsilon^{-1} = 1$ on $\partial\Omega$, $|\mathbf{x}| \leq \mathrm{diam}(\Omega)$, the Cauchy-Schwartz inequality and $\mathbf{x} \cdot n \geq \gamma$ for some $\gamma > 0$ (as a consequence of the star-shaped domain), we have:

$$-\frac{1}{2}\int_{\partial\Omega} (\mathbf{x} \cdot n)\, |\nabla w|^2 \, ds \leq -\frac{\gamma}{2}\ \|\nabla w\|^2_{\partial\Omega} \,.$$

Since $|\mathbf{x}| \leq \mathrm{diam}(\Omega)$, we note the inequality:

$$\frac{k^2}{2}\int_{\partial\Omega} (\mathbf{x} \cdot n)\, |w|^2 \, ds \ \leq\ \frac{k^2}{2}\,\mathrm{diam}(\Omega)\ \|w\|^2_{\partial\Omega} \,,$$

and

$$\left|\Re\big(ik\int_{\partial\Omega} w\, \overline{\mathbf{x} \cdot \nabla w}\, ds\big)\right| \ \leq\ k\, \mathrm{diam}(\Omega)\ \|w\|_{\partial\Omega}\ \|\nabla w\|_{\partial\Omega} \,.$$

Finally, the right-hand side volume term in (4.27) is controlled by:

$$\left|\Re\int_{\Omega} \Phi\, \overline{\mathbf{x} \cdot \epsilon^{-1}\nabla w}\, dx\right| \ \leq\ \max_{x\in\Omega} |\epsilon^{-1}(x)|\ \mathrm{diam}(\Omega)\ \|\Phi\|_{\Omega}\ \|\nabla w\|_{\Omega} \,.$$

Taking the absolute value of both sides then applying these three estimates:

$$\begin{aligned}
&|k^2\int_{\Omega} \epsilon^{-1}|w|^2\, dx| \\
&\leq \max_{x\in\Omega} |\epsilon^{-1}(x)|\ \mathrm{diam}(\Omega) \left|\tfrac{k^2}{2}\ \|w\|^2_{\partial\Omega} - \frac{\gamma}{2}\ \|\nabla w\|^2_{\partial\Omega} + k\ \|w\|_{\partial\Omega}\ \|\nabla w\|_{\partial\Omega} + \|\Phi\|_{\Omega}\ \|\nabla w\|_{\Omega}\right|
\end{aligned}$$

Then, define:

As $\gamma \leq \text{diam}(\Omega)$:

$$\nu' = \frac{\gamma}{\max_{x\in\Omega}(|\epsilon^{-1}(x)|)\text{diam}(\Omega)^2} \leq 1$$

Then, we see:

$$\frac{\nu' \max_{x\in\Omega}(|\epsilon^{-1}(x)|)\text{diam}(\Omega)^2}{2} - \frac{\gamma}{2} = 0 \tag{4.28}$$

We then apply Young's inequality with $\nu = \nu'$, so that we have:

$$k \left\| w \right\|_{\partial\Omega} \left\| \nabla w \right\|_{\partial\Omega} \leq \frac{1}{2\nu'} k^2 \left\| w \right\|_{\partial\Omega}^2 + \frac{\nu'}{2} \left\| \nabla w \right\|_{\partial\Omega}^2$$

Then we may group the boundary terms and note the coefficient of the boundary gradient term is 0 by (4.28), so we may drop this term from the right hand side. With C chosen large enough to absorb the ν constant and the diam):

$$|k^2 \int_\Omega \epsilon^{-1} |w|^2 dx| \leq C(k^2 \left\| w \right\|_{\partial\Omega}^2 + \left\| \Phi \right\|_\Omega \left\| \nabla w \right\|_\Omega)$$

Applying (4.17) on the boundary term and Young's on the volume term

$$\begin{aligned} &\leq C\Big(\tfrac{1}{2\nu} \left\| \Phi \right\|_\Omega^2 + \tfrac{\nu k^2}{2} \left\| w \right\|_\Omega^2 + \ \left\| \Phi \right\|_\Omega \left\| \nabla w \right\|_\Omega \Big) \\ &\leq C\big(\tfrac{1}{\nu} \left\| \Phi \right\|_\Omega^2 + \tfrac{\nu k^2}{2} \left\| w \right\|_\Omega^2 + \tfrac{\nu}{2} \left\| \nabla w \right\|_\Omega^2 \big) \end{aligned}$$

By using (4.18) to bound the gradient:

$$\leq C\Big[\tfrac{1}{\nu} \left\| \Phi \right\|_\Omega^2 \ + \ \frac{\nu k^2}{2} \left\| w \right\|_\Omega^2 \ + \ \frac{\nu C_T}{2} \Big((1 + \tfrac{\nu}{2} k^2) \left\| w \right\|_\Omega^2 + \tfrac{1+k^{-2}}{2\nu} \left\| \Phi \right\|_\Omega^2 \Big) \Big]$$

Now,

$$\min_{\mathbf{x}\in\Omega} |\epsilon^{-1}|k^2 \int_\Omega |w|^2 dx \leq |k^2 \int_\Omega \epsilon^{-1}|w|^2 dx|,$$

After multiplying both sides by $\min|\epsilon|$, up to absorbing constants, we have an inequality then of the form:

$$k^2 \int_\Omega |w|\,dx \leq C\big(\frac{1}{\nu}(1+k^{-2})\,\|\Phi\|_\Omega^2 + \nu k^2\,\|w\|_\Omega^2\big)$$

Then, given $\nu < 1$ we absorb $k^2\nu\,\|w\|_\Omega^2$ into the left hand side and get:

$$k^2 \int_\Omega |w|^2 dx \leq C\left(1+k^{-2}\right)\,\|\Phi\|_\Omega^2$$

Now, as we bound $k > k_{\min}$ away from 0, we can say:

$$k^2 \int_\Omega |w|^2 dx \leq C\left(1+k_{\min}^{-2}\right)\,\|\Phi\|_\Omega^2$$

absorbing the coefficient, we finally have:

$$k^2 \int_\Omega |w|^2 dx \leq C_1^2\,\|\Phi\|_\Omega^2$$

Taking the root of both sides, we have

$$k\,\|w\|_\Omega \leq C_1\,\|\Phi\|_\Omega \tag{4.29}$$

Here C_1 carries the same T-coercivity constant mentioned in Lemma 3.2.1, and therefore grows as $\epsilon_m \to \mathfrak{C}$. This gives us the result we were seeking in the first line of (4.12), for $k > 0$.

For any $v \in H^1(\Omega)$, its boundary trace $v|_{\partial\Omega}$ lies in $H^{1/2}(\partial\Omega)$, with

$$\|v\|_{H^{1/2}(\partial\Omega)} \;=\; \inf\big\{\|\mathfrak{V}\|_{H^1(\Omega)} : \mathfrak{V}|_{\partial\Omega} = v\big\},$$

and we shall use the shorthand $\|v\|_{1/2,\Omega} = \|v\|_{H^{1/2}(\partial\Omega)}$.

We also remark that following from our first result:

$$\|w\|_{\Omega} \;\leq\; \frac{C_1}{k}\;\|\Phi\|_{\Omega}\,. \tag{4.30}$$

Furthermore, we also have the standard Sobolev space trace bound, which we can combine with the prior result to show:

$$\|w\|_{1/2,\Omega} \leq C_{\mathrm{tr}}\;\|w\|_{H^1(\Omega)} \;\leq\; C_{\mathrm{tr}}\,\frac{C_1}{k}\;\|\Phi\|_{\Omega}\,. \tag{4.31}$$

For the second result in (4.12), we start from our Helmholtz-like equation and use the regularity we impose on w (subject to our transmission conditions across $\Gamma = \partial\mathcal{D}$) to obtain:

$$\big\|\nabla\cdot\epsilon^{-1}\nabla w\big\|_{\Omega} \lesssim C(\big\|\Phi + k^2 w\big\|_{\Omega} + \|\partial_n w\|_{1/2})$$

Next, split $\Phi + k^2 w$:

$$\|\nabla\cdot(\epsilon^{-1}\nabla w)\|_{\Omega} \;\lesssim\; C\Big(\|\Phi\|_{\Omega} + k^2\;\|w\|_{\Omega} + k\;\|w\|_{{}_{1/2,\Omega}}\Big)$$

We then apply (4.30) and absorb the boundary term via the Trace Theorem for Sobolev spaces - see (4.31).

$$\begin{aligned}
\left\|\nabla \cdot \epsilon^{-1} \nabla w\right\|_{\Omega} &\leq C\, \|\Phi\|_{\Omega} \left(1 + k^2 \frac{C_1}{k} + C_{\mathrm{tr}}\, C_1\right) \\
\left\|\nabla \cdot \epsilon^{-1} \nabla w\right\|_{\Omega} &\leq C_2 \,(1+k)\, \|\Phi\|_{\Omega}\,.
\end{aligned}$$

Thus (by standard absorption of the constants) we may conclude with the second result in our proof. □

Now, with these stability estimates in hand, the standard approach for convergence in the DG norm is through a duality argument, which holds here as well. We will show that the L^2 norm is controlled by the DG norm, and therefore that convergence guaranteed by the well-posedness of our problem in the DG can be meaningful in an L^2 sense. This style of argument is common - We refer to Lemma 4.3.7 of Moiola (2011) and Lemma 3.7 of Hiptmair et al. (2011). The main novelty here in the sign-changing setting is the application of Lemma 4.3.3, so that the constant here depends on the T-coercivity contrast present in that problem.

Theorem 4.3.1 (Duality Estimate of the L^2 norm). Assume the mesh is quasi-uniform ($h_K \approx h$ for all K) and k is bounded away from 0. Then there exist a constant $C > 0$ (depending only on the contrast and geometry, but *not* on h or k) such that for all $v \in \mathcal{T}(\Omega_h)$:

$$\|v\|_{\Omega} \;\leq\; C\left(1 + (k\,h)^{-1/2}\right) \|v\|_{DG,h},$$

Proof. We take as our weak form (4.7) with associated norm 4.2.3.

We consider the adjoint problem:

$$\begin{aligned} &\nabla \cdot (\epsilon^{-1} \nabla w) + k^2 w = \Phi \quad \text{on } \Omega \\ &\partial_n w + ikw = 0 \quad \text{on } \partial\Omega \end{aligned} \tag{4.32}$$

with $\Phi \in L^2(\Omega)$.

Let $v \in \mathcal{T}(\Omega_h)$. We take the conjugate of equation (4.32).

$$\nabla \cdot (\epsilon^{-1} \nabla \overline{w}) + k^2 \overline{w} = \overline{\Phi}$$

Multiplying this equation by v and performing integration by parts twice yields

$$\begin{aligned} &|(v, \Phi)|_\Omega = |\sum_K \int_{\partial K} \epsilon^{-1} \partial_n \overline{w} v \; ds - \int_{\partial K} \epsilon^{-1} \overline{w} \partial_n v ds| \\ &= |\int_{\Upsilon_h^I} \epsilon^{-1} \nabla \overline{w} [\![v]\!] - [\![\epsilon^{-1} \nabla v]\!] \overline{w} \; ds + \int_{\Upsilon_h^B} \overline{w}(-(n \cdot \nabla v) + ikv) \; ds| \\ &= |\int_{\Upsilon_h^I} \alpha \epsilon^{-1} (\nabla \overline{w}) \alpha^{-1} [\![v]\!] - \beta [\![\epsilon^{-1} \nabla v]\!] \beta^{-1} \overline{w} \; ds \\ &+ \int_{\Upsilon_h^B} \delta \overline{w} (-\delta^{-1}(n \cdot \nabla v) + \delta^{-1} ikv) \; ds| \end{aligned}$$

Using the Cauchy Schwartz and Triangle Inequalities:

$$\begin{aligned} &\leq \sum_{\ell \in \Upsilon_h^I} k^{1/2} \left\| \alpha^{1/2} [\![v]\!] \right\|_\ell k^{-1/2} \left\| \alpha^{-1/2} \epsilon^{-1} \nabla w \right\|_\ell + k^{-1/2} \left\| \beta^{1/2} [\![\epsilon^{-1} \nabla v]\!] \right\|_\ell k^{1/2} \left\| \beta^{-1/2} w \right\|_\ell \\ &+ \sum_{\ell \in \Upsilon_h^B} k^{-1/2} \left\| \delta^{1/2} \nabla v \right\|_\ell k^{1/2} \left\| \delta^{-1/2} w \right\|_\ell + k^{1/2} \left\| (1-\delta)^{1/2} v \right\|_\ell k^{1/2} \left\| (1-\delta)^{-1/2} w \right\|_\ell \\ &\leq \sum_{\ell \in \Upsilon_h^I} k^{1/2} \left\| \alpha^{1/2} [\![v]\!] \right\|_\ell k^{-1/2} \left\| \alpha^{-1/2} \epsilon^{-1} \nabla w \right\|_\ell + k^{-1/2} \left\| \beta^{1/2} [\![\epsilon^{-1} \nabla v]\!] \right\|_\ell k^{1/2} \left\| \beta^{-1/2} w \right\|_\ell \\ &+ \sum_{\ell \in \Upsilon_h^B} k^{-1/2} \left\| \delta^{1/2} \nabla v \right\|_\ell k^{1/2} \left\| \delta^{-1/2} w \right\|_\ell + k^{1/2} \left\| (1-\delta)^{1/2} v \right\|_\ell k^{1/2} \left\| (1-\delta)^{-1/2} w \right\|_\ell \end{aligned}$$

Again using the Cauchy-Schwartz Inequality:

$$
\begin{aligned}
&\leq \Big[\sum_{\ell \in \Upsilon_h^I} k^{-1} \left\| \beta^{1/2} [\![\epsilon^{-1} \nabla v]\!] \right\|_{\Upsilon_h^I}^2 + k \left\| \alpha^{1/2} [\![v]\!] \right\|_{\Upsilon_h^I}^2 + \sum_{\ell \in \Upsilon_h^B} k^{-1} \left\| \delta^{1/2} \partial_n v \right\|_{\Upsilon_h^B}^2 \\
&+ k \left\| (1-\delta)^{1/2} v \right\|_{\Upsilon_h^B}^2 \Big]^{1/2} \Big[\sum_{\ell \in \Upsilon_h^I} k \left\| \beta^{-1/2} w \right\|_\ell^2 + k^{-1} \left\| \alpha^{-1/2} \epsilon^{-1} \nabla w \right\|_\ell^2 \\
&+ \sum_{\ell \in \Upsilon_h^B} k \left\| (1-\delta)^{-1/2} w \right\|_\ell^2 \Big]^{1/2} \\
&= \|v\|_{DG} \, (\mathcal{G}_\epsilon)^{1/2}
\end{aligned}
$$

Define $\gamma = \beta$ on interior faces Υ_h^I and $\gamma = 1 - \delta$ on boundary faces Υ_h^B. Then,

$$
\mathcal{G}_\epsilon \leq \sum_K k \left\| \gamma^{-1/2} w \right\|_{\partial K}^2 + k^{-1} \left\| \alpha^{-1/2} \epsilon^{-1} \nabla w \right\|_{\partial K}^2
$$

We take note of the following trace inequality (see Theorem 1.6.6 of Brenner & Scott (2007)):

$$
\|u\|_{\partial K}^2 \ \leq \ C_{\mathrm{tr}} \|u\|_K \Big(h_K^{-1} \|u\|_K \ + \ \|\nabla u\|_K \Big).
$$

This allows us to write:

$$
\begin{aligned}
\mathcal{G}_\epsilon &= \sum_K \Big[k \|\gamma^{-1/2} w\|_{\partial K}^2 + k^{-1} \|\alpha^{-1/2} \, \epsilon^{-1} \nabla w\|_{\partial K}^2 \Big] \\
&\leq C_{\mathrm{tr}} \sum_K \Big(k h^{-1} \|w\|_K^2 + k \|w\|_k \|\nabla w\|_K + k^{-1} h^{-1} \|\epsilon^{-1} \nabla w\|_K^2 \\
&+ \ k^{-1} h^{-1} \|\epsilon^{-1} \nabla w\|_K^2 + k^{-1} \|\nabla w\|_K \ \|\nabla \cdot (\epsilon^{-1} \nabla w)\|_K \Big) \\
&\lesssim C \, (kh)^{-1} \ \|\Phi\|_\Omega^2 + C \ \|\Phi\|_\Omega^2 + C \, (kh)^{-1} \ \|\Phi\|_\Omega^2 + C \left(\tfrac{1}{k} + 1 \right) \ \|\Phi\|_\Omega^2 \\
\mathcal{G}_\epsilon &\lesssim C_T(\Omega) (1 + (kh)^{-1}) \, \|\Phi\|_\Omega^2
\end{aligned}
$$

Of note, we assume that $1+\frac{1}{k}$ is easily bounded above by C without k dependence due to our assumption that k is bounded away from 0. Absorbing constants as appropriate and dividing our original weak form by $\|\Phi\|_{\Omega}^2$, then taking the sqaure root of both sides, we get:

$$\frac{|(v,\Phi)|_{\Omega}}{\|\Phi\|_{\Omega}} \leq C(1+(kh)^{-1/2})\,\|v\|_{DG}$$

Here, Φ was an arbitrary function in $L^2(\Omega)$. This inequality holds for all such functions, therefore we may take the supremum of all such functions and arrive at our conclusion:

$$\|v\|_{\Omega} \leq C(1+(kh)^{-1/2})\,\|v\|_{DG} \tag{4.33}$$

□

Corollary 4.3.1. Under the assumptions of the previous theorem, let u be the exact solution of (4.7) and $u_h \in \mathcal{T}(\Omega_h)$ the corresponding DG approximation. Then

$$\|\,u-u_h\|_{\Omega)} \;\leq\; C\left(1+(k\,h)^{-1/2}\right)\|\,u-u_h\|_{DG,h}.$$

Proof. In the duality bound of the previous theorem we had for all $v \in \mathcal{T}(\Omega_h)$

$$\|v\|_{\Omega)} \;\leq\; C\left(1+(k\,h)^{-1/2}\right)\|v\|_{DG,h}.$$

Simply set $v = u - u_h$. Since $u - u_h \in \mathcal{T}(\Omega_h)$, this immediately gives the stated estimate. □

Remark 4.3.1. While it may appear benign, we point out the dependence of C on our geometry in 4.3.1 and the values of ϵ_m means as $\epsilon_m \to \mathfrak{C}$ (in the case of a disk-shaped scatterer, $\mathfrak{C} = -1$), we will see a large increase in C due to the T-coercivity property of the continuous problem. In other words, the constant C from Lemma 4.3.3 will go to 0, just as in Theorem 4.2.1. This will cause the error bound to degenerate, and in particular in a near-resonance case (for more details see §2.3) we will see large errors unbounded by this proof. A similar phenomenon in 3D FEM can be seen in Figure 3.7am and we present the same results in 2D with the PWDG in figure 4.14.

One might want to have sharper estimates in the error by taking advantage of known basis functions (e.g. plane waves). However, the theory of best approximation of Helmholtz solutions by plane waves does not admit a ready expansion to the varying material problem we have posed here. Indeed, the presence of near-resonant modes in and of themselves spoil any guarantee of steady convergence in all cases for the sign-changing problem.

4.4. Evanescent Plane Waves

we now turn to the concrete construction of $V^p(\Omega_h)$ using plane waves. In practice in the sign-changing regime we often enrich the standard propagative waves with evanescent modes to capture sub-wavelength behavior. We are interested in plane wave basis functions for two reasons. Firstly, as shown below, they lie in the kernel of the Helmholtz operator, and therefore serve as Trefftz functions. Secondly, it is classical Herglotz theory that any Helmholtz solution in a ball admits a continuous superposition of real-direction plane waves. Colton & Kress (2013) However, in the case of ϵ_m negative, by Theorem 4.2.1 we still have a T-coercive sign-changing problem, and we must adapt our basis correspondingly.

For $\mathbf{x} \in \mathbb{R}^2 \setminus \overline{\Omega}$, the propagative plane wave is a solution to this problem, not considering the boundary condition. That is to say,

$$(\Delta + k^2)\, e^{ik\mathbf{d}\cdot\mathbf{x}} = (-k^2\mathbf{d}\cdot\mathbf{d} + k^2)\, e^{ik\mathbf{d}\cdot\mathbf{x}} = 0, \tag{4.34}$$

Interior to the scatterer, we have a modified interior Helmholtz equation. We note that

$$(\nabla \cdot (\epsilon^{-1}\nabla) + k^2)u = 0 \quad \mathbf{x} \text{ in } \Omega$$
$$\Leftrightarrow (\nabla \cdot (\epsilon_m^{-1}\nabla) + k^2)u = 0 \quad \mathbf{x} \text{ in } \Omega$$
$$\Leftrightarrow (\Delta - (k\sqrt{-\epsilon_m})^2)u = 0 \quad \mathbf{x} \text{ in } \Omega$$

Consequently, one obtains interior solutions of the form $e^{ik\sqrt{\epsilon_m}\mathbf{d}\cdot\mathbf{x}}$. Now as $\sqrt{\epsilon_m}$ is a pure imaginary number, if $\mathbf{d} \in \mathbb{R}^2$ (in other words, not complex) this is not a plane wave at all and is instead an exponential function decaying in the chosen direction. Therefore, we cannot construct a continuous plane wave-type solution to our problem using only propagating plane waves. Thus, in order to create a Trefftz plane wave method, we must explore the usage of Evanescent Plane Waves (EPWs).

We further note our interest in evanescent plane waves by recalling that under certain circumstances, rapid-scale evanescent behavior can occur in this problem as illustrated in Figure 2.7 and Figure 3.5 where strong sub-wavelength decay appears near the interface. We define $(\theta, \zeta) \in [0, 2\pi) \times \mathbb{R}$ to be our real and decay propogation angles respectively. Then, we define our choice of direction vector $\mathbf{d} = [\cos(\theta + i\zeta), \sin(\theta + i\zeta)]$ controls the direction of decay for an evanescent plane wave (EPW). To that end, we define:

$$\mathbf{d}_r = [\cos(\theta), \sin(\theta)]$$
$$\mathbf{d}^{\perp} = [-\sin(\theta), \cos(\theta)]$$

And we recall the complex sine and cosine identities:

$$\mathbf{d} = [\cos(\theta)\cosh(\zeta) - i\sin(\theta)\sinh(\zeta), \sin(\theta)\cosh(\zeta) + i\cos(\theta)\sinh(\zeta)]$$

It then follows if $\mathbf{x} = [x_1, x_2]$:

$$
\begin{aligned}
e^{ik\sqrt{\epsilon}\mathbf{d}\cdot\mathbf{x}} &= e^{ik\sqrt{\epsilon}((\cos(\theta)\cosh(\zeta)x_1+\sin(\theta)\cosh(\zeta)x_2)+i(\sin(\theta)\sinh(\zeta)x_1-\cos(\theta)\sinh(\zeta)x_2))} \\
&= e^{ik\sqrt{\epsilon}\cosh(\zeta)(\mathbf{d}_r\cdot\mathbf{x})}\cdot e^{-k\sqrt{\epsilon}\sinh(\zeta)(\mathbf{d}^{\perp}\cdot\mathbf{x})}
\end{aligned}
\tag{4.35}
$$

From this we can see we have a plane wave with wavenumber $k\sqrt{\epsilon}\cosh(\zeta)$ with decay given by $k\sqrt{\epsilon}\sinh(\zeta)$ in the perpendicular direction.

We can define our specific choice then of the Trefftz plane wave space V^p as follows:

Definition 4.4.1: PWDG Function Space

$$PW^p_{k_m}(\mathbb{R}^2) = \{v \in C^\infty(\mathbb{R}^2) : v = \sum_{j=1}^{p} \varphi_j e^{ik_m\mathbf{d}_j\cdot\mathbf{x}}, \varphi_j \in \mathbb{C}\ \forall j\}$$

$$\mathbf{d}\cdot\mathbf{d} = 1$$

$$k_m(r) = \begin{cases} k & r > 1 \\ \sqrt{\epsilon_m}k & r \leq 1 \end{cases}$$

$$V^p(\Omega_h) = \{v \in L^2(\Omega) : v|_K \in PW^p_{k_m}(\mathbb{R}^2) \quad \forall K \in \Omega_h\}$$

From a numerical standpoint, large sets of purely real-direction PWs exhibit severe ill-conditioning: expansion coefficients blow up exponentially and the basis functions become nearly linearly dependent. Gittelson et al. (2009); Hiptmair et al. (2011) In contrast, Parolin, Huybrechs and Moiola proved that every 2D Helmholtz field in a disk admits an EPW expansion with uniformly bounded coefficients. Parolin et al. (2023) By augmenting our PWDG space with a judiciously sampled set of evanescent directions (θ, ζ), we therefore hope to recover the high approximation power of Trefftz methods while maintaining the numerical stability granted by the EPWs.

In practice, we have several approaches to choose our propagation angles (θ, ζ). In most approaches, we simply choose θ to be evenly spaced in $[0, 2\pi)$. The simplest choice is to select a small range of ζ evenly spread around 0.

Another choice, useful for the near-resonance problem and inspired by §3.4.1, is to use a least-squares method to fit a set of plane waves to the decay of the surface plasmon in a small radial band about the interface. We know this decay analytically from §2.3. Such a fitting would not be exactly, but might allow for the capture of the general character of the resonance in the EPW basis.

Another choice is given by choosing a distribution of ζ: We pick $\theta_j = 2\pi(j-1)/M$, $j = 1, \ldots, M$, uniformly in $[0, 2\pi)$, and we draw our decay parameters ζ_k from the cumulative distribution function Ξ_P of Parolin et al. Parolin et al. (2023); Galante et al. (2024). Concretely, for a target total of N plane waves set

$$M = \lceil \sqrt{N} \rceil, \qquad P = \lfloor N/8 \rfloor,$$

and then

$$\Xi_P(\zeta_k) \;=\; \frac{k - \frac{1}{2}}{M}, \quad k = 1, \ldots, M, \qquad \zeta_k = \Xi_P^{-1}\Big(\tfrac{k-\frac{1}{2}}{M}\Big).$$

The function $\Xi_P \colon \mathbb{R} \to [0, 1]$ admits the explicit formula

$$\Xi_P(\zeta) = \tfrac{1}{2} + \tfrac{1}{2}\,\mathrm{sign}(\zeta)\left[\tfrac{1}{2} - \tfrac{1}{2P+1}\left(\frac{\Gamma\big(\frac{1}{2}, |k|R\,|\zeta|\big)}{2\,\Gamma\big(\frac{1}{2}, |k|R\big)} + \sum_{p=1}^{P} \mu_p(\zeta)\right)\right], \tag{4.36}$$

where

$$\mu_p(\zeta) = \frac{(|k|R)^{2p}\,\Gamma\big(\frac{1}{2} - 2p,\ |k|R\,|\zeta|\big) + (|k|R)^{-2p}\,\Gamma\big(\frac{1}{2} + 2p,\ |k|R\,|\zeta|\big)}{(|k|R)^{2p}\,\Gamma\big(\frac{1}{2} - 2p,\ |k|R\big) + (|k|R)^{-2p}\,\Gamma\big(\frac{1}{2} + 2p,\ |k|R\big)},$$

and $\Gamma(s, x)$ is the upper incomplete Gamma-function. In practice we invert (4.36) using a simple bisection scheme, exploiting the oddness of $\Xi_P(\zeta) - \frac{1}{2}$. This combines the oscillatory approximation power of plane waves with the uniform-bound stability of Parolin et al.'s evanescent expansion. For the derivation of the stability of these bases, we again direct readers to Parolin et al. (2023); Galante et al. (2024).

However, this comes at a risk - while these are chosen to be theoretically stable in the sense of converging to a Helmholtz solution in finite coefficients, practically they can often introduce ill-conditioning to the system for large ζ (leading to basis functions which are near zero on some skeleton edges). This encourages us to implement a floor/ceiling to remove ζ values which are too large or too small for the sake of preserving independence of the

Regardless of how they are chose, The resulting set of evanescent-propagative directions

$$\{(\theta_j, \zeta_k) : j, k = 1, \ldots, M\}$$

yields the Trefftz-DG trial space

$$\{ e^{i\,k\sqrt{\epsilon_m}\,\mathbf{d}(\theta_j,\zeta_k)\cdot\mathbf{x}} \}_{j,k=1}^{M} \subset PW_{k_m}^{p}(\mathbb{R}^2),$$

.

4.5. Numerical Results

4.5.1. On the choice of parameters

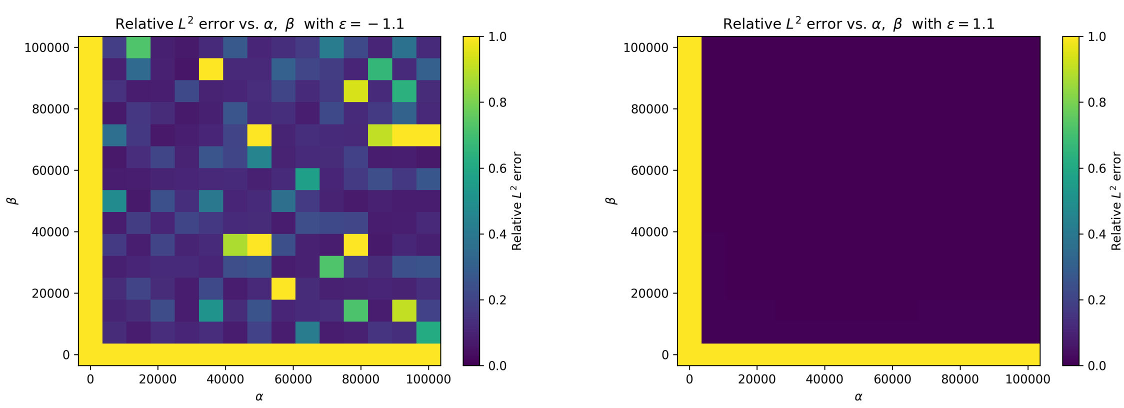
Remark 4.5.1. Numerical results used in this chapter are available publicly on Github, see Latham (2025a,b).

As discussed in Appendix C, the penalty parameters α, β are very important for the behavior of our numerical method. Often in the literature $\mathcal{O}(1)$ constants are used as penalty parameters in such IP methods. See Moiola (2011); Lehrenfeld & Stocker (2023). However, in our sign-changing problem, such parameters are insufficient. To illustrate this, we begin by taking the problem as given in (4.1) and discretizing using the weak form (4.7), then allowing α and β to vary between $[10^{-1}, 10^{6}]$

for the parameter set: $\epsilon_m \in \{-1.1, 1.1\}$, with a disk-shaped scatterer of radius 1. For simplicity, we use p PPWs as our basis functions Though this low–order discretization prevents extreme accuracy, the qualitative trends in error over the (α, β) plane remain clear. We note that we do not impose such consideration for the boundary parameter δ , as typically for both sign-changing and non-sign-changing problems we wish to have large δ for stability. Additionally, the boundary is usually chosen to be truncated several meshlengths away from the scatterer (at least) to ensure that the boundary condition is a good approximation, so the behavior of δ does not substantially change in the sign-changing problem.

In general, so long as we don't incur floating point error, large α and β tend to more diagonally dominate the system, causing better conditioning in general. Indeed, for the strongly continuous case with $\epsilon_m > 0$, which is handled well by FEM, we see that large constant values of these parameters tend to provide better error. See Figure 4.1b.

However, in the case of negative ϵ_m, the behavior is markedly different. See Figure 4.1a. Here, we need a non-uniform approach to obtain better L^2 error in regards to our problem.



(a) $\epsilon_m = -1.1$

(b) $\epsilon_m = 1.1$

Figure 4.1. L^2 error for varying constant α, β on a uniform mesh of $h = 0.15$, for (a) the sign–changing case and (b) the positive–ϵ case.

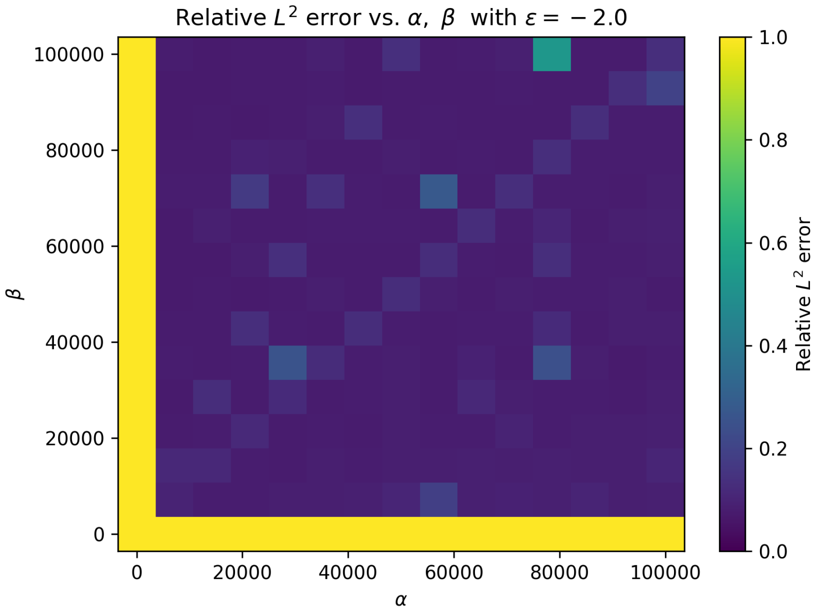


Figure 4.2. L^2 error for varying constant α, β on a uniform mesh of $h = 0.15$, for $\epsilon_m = -2$

We note that this instability in terms of our flux is *not* independent of ϵ_m, indeed, if we attempt the experiment with $\epsilon_m = -2$ (see Figure 4.2), we notice that we are approaching results similar to the positive ϵ_m case. This is due to the rapidly changing gradient present in the negative ϵ_m case as $\epsilon_m \to \mathfrak{C}$, see Figure 2.9 in the section on analytical results.

Refining the problem in p and taking a longer view of the parameters involved in this problem gives us the result in Figure 4.3. We see that the quadrant of the α, β plane defined by $10^3 \leq \alpha \leq \beta$ seems to be the region wherein we return good L^2 errors.

We propose the following recipe for fluxes in the sign changing case based on the empirical evidence in our study:

$$\alpha = C_\alpha(\frac{2p^2}{kh}), \quad \beta = C_\beta(\frac{2p^2}{kh}), \tag{4.37}$$

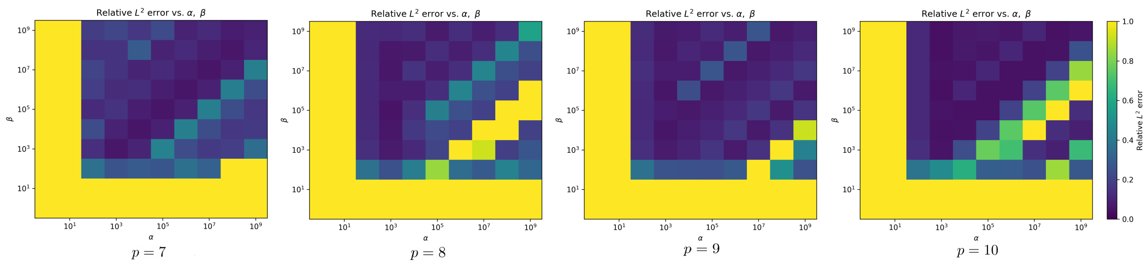


Figure 4.3. L^2 error for varying constant α, β and varying p on a uniform mesh of $h = 0.15$, for $\epsilon_m = -1.1$. See fig Figure 4.4 for a zoomed view of $p = 10$

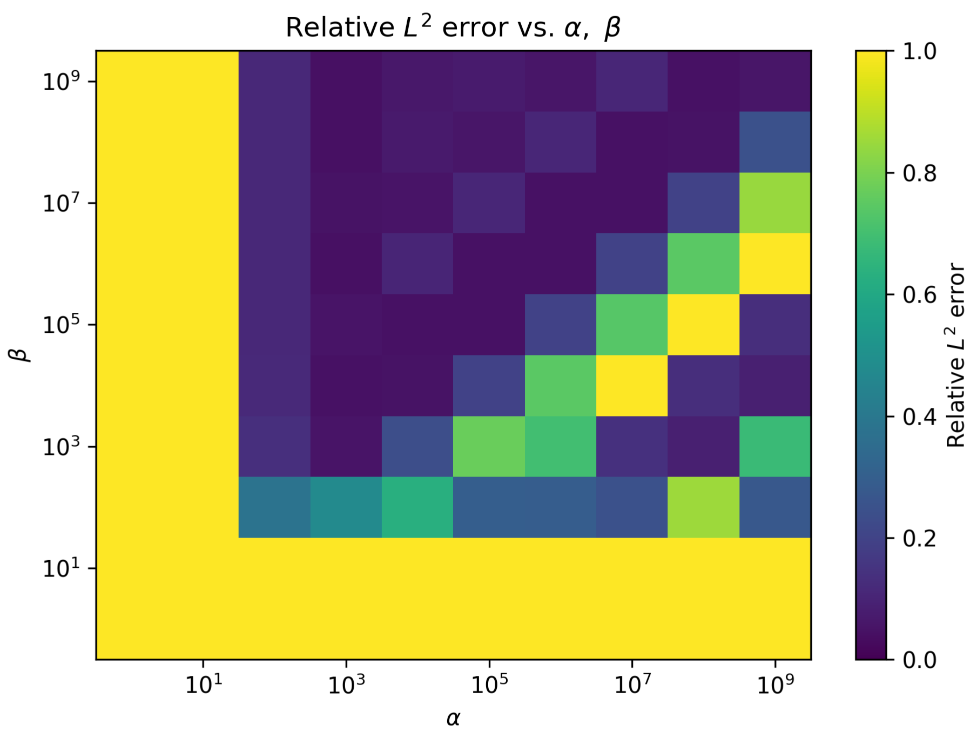


Figure 4.4. L^2 error for varying constant α, β and $p = 10$ on a uniform mesh of $h = 0.15$, for $\epsilon_m = -1.1$.

where C_α is $\mathcal{O}(1)$ and $C_\beta \geq C_\alpha$. We remark that in our sample cases, given above, this assumption gives $\beta \geq \alpha \gtrsim 10^3$, which is the sector with the best error behavior.

We further remark that here, we set p to be the number of equidistant real directions chosen, not necessarily the total number of plane waves (in the case of EPWs).

We seek to further investigate the near-contrast case $\epsilon_m = -1.1$. One might propose that, due to the known analytic behavior of the gradient being steeper interior to the scatterer, an adjustment to β as a piecewise function varying in material might make sense. However, we see little gain in practice. Here, setting $k = 4, h = 0.1, p = 8, R = 4$ we try varying β interior to the scatterer:

$$\beta = \begin{cases} \beta_{in} & \mathbf{x} \in \mathcal{D} \\ 10^5 & \mathbf{x} \in \Omega \setminus \bar{\mathcal{D}} \end{cases}$$

But as we see in the results (see Figure 4.5), even changing β_{in} over a large range of values has little impact on the error, so long as β satisfies the recipe we proposed above.

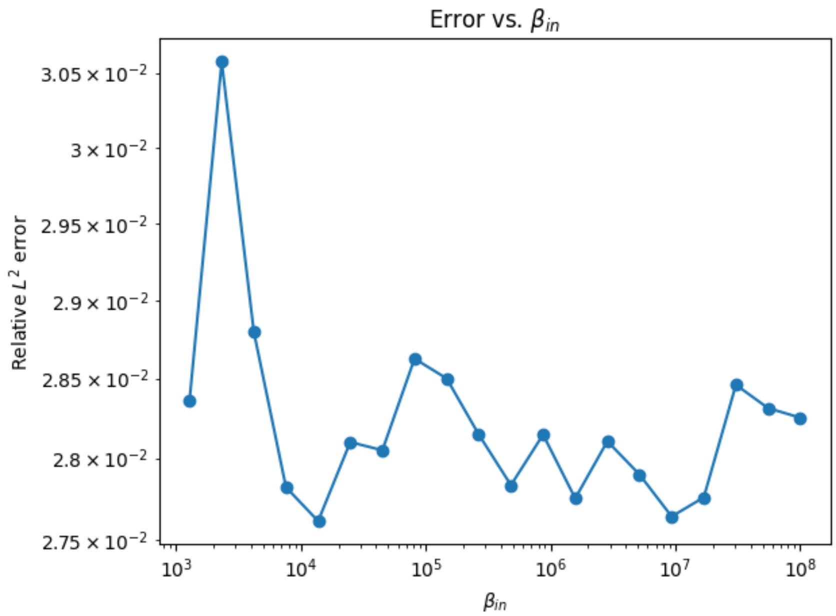


Figure 4.5. L^2 error vs. various $\beta|_{\mathcal{D}}$ for $k = 4, h = 0.1, p = 8, R = 4$

4.5.2. Convergence Results and FEM comparison

We now wish to consider a comparison between an FEM method in 2D and the proposed Trefftz method for a variety of settings. We will also investigate the role of various choices of plane wave on accuracy. Our principle focus will be on a circular scatterer for which we have an analytic solution (see §2.2).

We begin by considering the influence of the error induced by our choice of a first-order absorbing boundary condition.

We have:

$$u^{\text{sca}}(r, \theta) = \sum_{n \in \mathbb{Z}} a_n \, H_n^{(1)}(kr) e^{in\theta}, \quad r \geq 1.$$

On a circle of radius R, the (exact) DtN operator yields:

$$\partial_r u^{\text{sca}}(R, \theta) = \sum_{n \in \mathbb{Z}} a_n \, k \, \frac{H_n^{(1)'}(kR)}{H_n^{(1)}(kR)} H_n^{(1)}(kR) e^{in\theta}.$$

For large kR, we have (from the NIST asymptotic expansion for large argument Bessel functions [Olver et al. (2010)]):

$$k\frac{H_n^{(1)'}(kR)}{H_n^{(1)}(kR)} = ik + \mathcal{O}\big((kR)^{-1}\big), \tag{4.38}$$

so for each fixed mode n,

$$\partial_r u^{\mathrm{sca}}(R,\theta) - ik\, u^{\mathrm{sca}}(R,\theta) = \sum_{n\in\mathbb{Z}} a_n\, \mathcal{O}\big((kR)^{-1}\big)\, e^{in\theta}.$$

Experiment 4.5.1 (PWDG error plateaus for varied R). For fixed k, therefore, we expect the error to decay linearly in R, the radius of our numerical truncation. We indeed see this result confirmed in Figure 4.6.

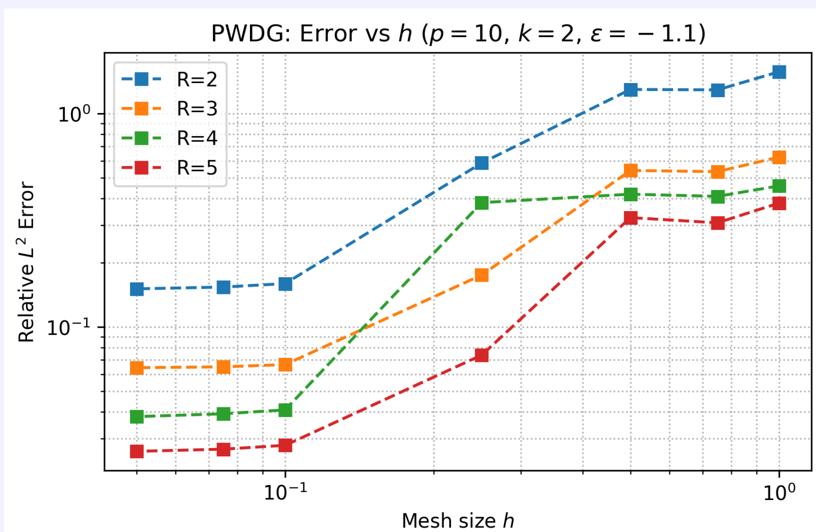
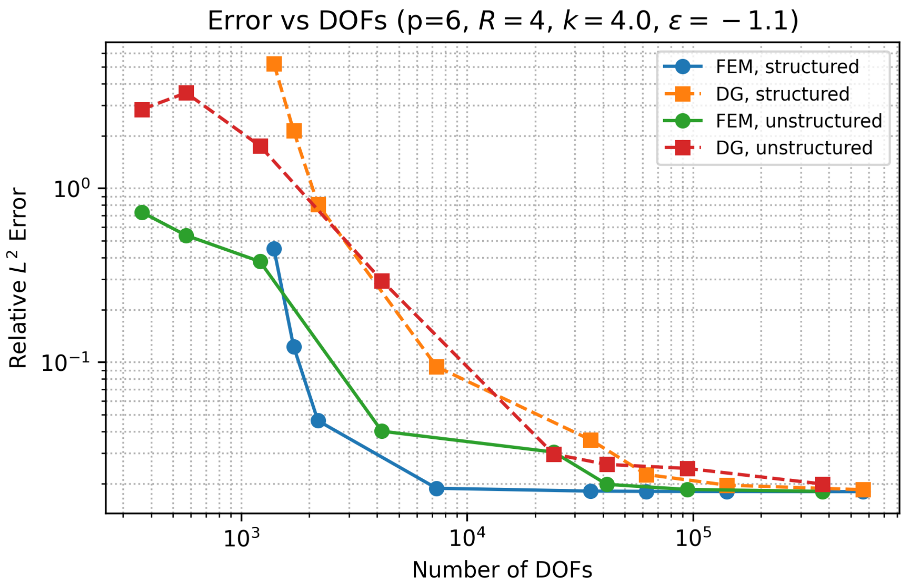
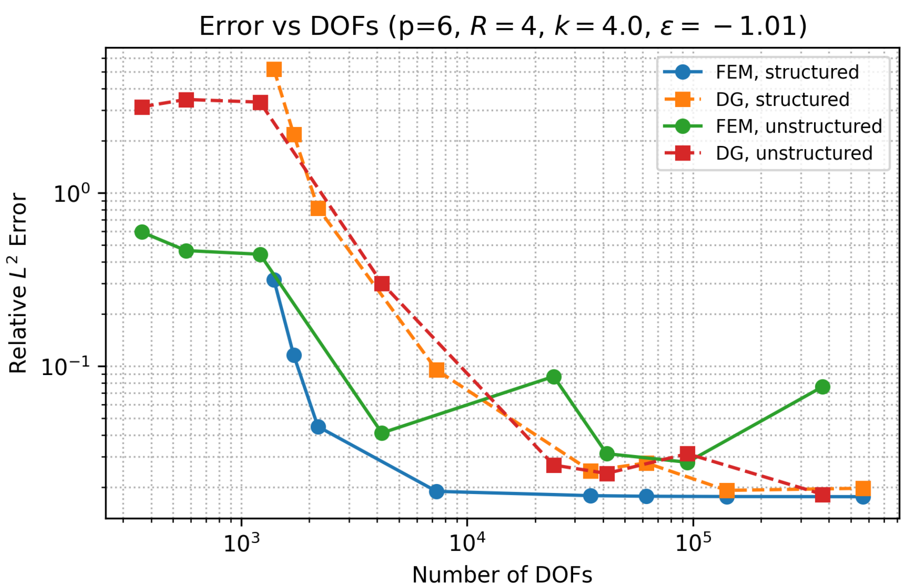


Figure 4.6. PWDG error for various radii R for $k = 2$, $\epsilon_m = -1.1, p = 10$

We now turn to the required structure of mesh for these experiments. As we discovered earlier, see Lemma 3.2.1, to guarantee convergence of the discrete problem when using the continuous Galerkin framework, we require a symmetric mesh. However, this was not found to be a necessity in section §4.2.2 Indeed, the DG norm converges regardless of the chosen mesh, as demonstrated in the below 4.7. However, in the case of the FEM, this is not the case. Indeed we can see as $\epsilon_m \to -1$, the FEM solution begins to behave poorly, not converging to the error plateau from our boundary condition. This is one of the principle advantages here to using the PWDG method - we are free of this mesh constraint.



(a) $\epsilon_m = -1.1$

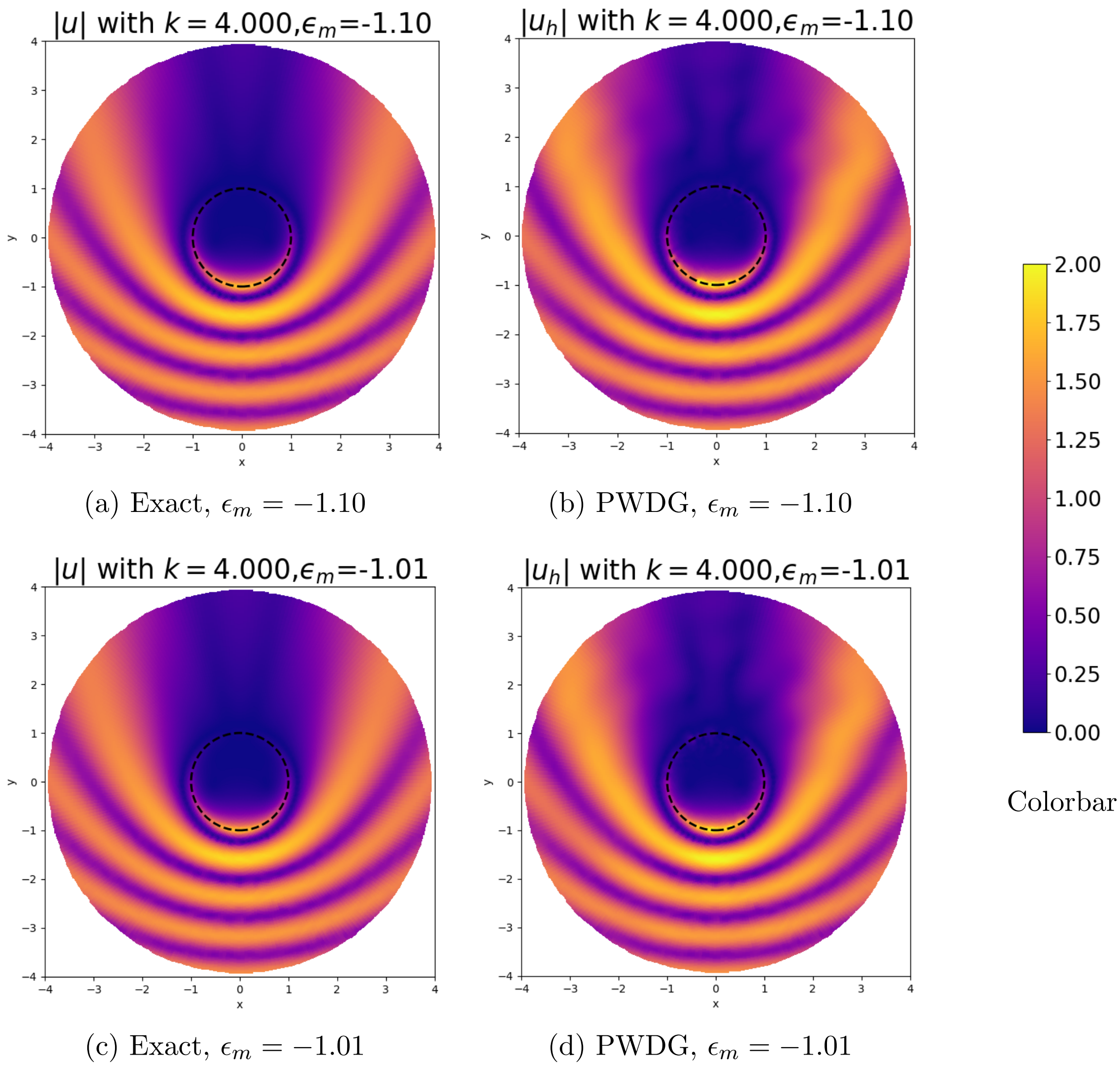


(b) $\epsilon_m = -1.01$

Figure 4.7. Relative L^2-error vs. total degrees of freedom for FEM (order-2 hierarchical polynomials) on structured/unstructured meshes and Trefftz PWDG (6 EPWs) on the disk ($R = 4$, $k = 4$). (a) $\epsilon_m = -1.10$. (b) $\epsilon_m = -1.01$. As $\epsilon_m \to -1$ the FEM on unstructured meshes stalls, while the PWDG remains robust.

Remark 4.5.2. We remark that we have a large number of degrees of freedom present in these experiments, and that substantially smaller values of h tend to be impractical, due to the demands on domain size from the first order Absorbing Boundary Condition (ABC). This might be addressed (at the cost of implementation and additional conditioning of the linear system) by using a higher order ABC or a Dirichlet-to-Neumann type boundary condition, this has been explored for scattteres in a waveguide in Monk et al. (2025).

To get a concrete feel for what the solution looks like on that coarse grid, Figure 4.8 shows side-by-side snapshots of the exact field $|u|$ and the Trefftz approximation $|u_h|$ at $h = 0.25$. The top row of figures corresponds to $\epsilon_m = -1.10$, the bottom row to $\epsilon_m = -1.01$.



(a) Exact, $\epsilon_m = -1.10$

(b) PWDG, $\epsilon_m = -1.10$

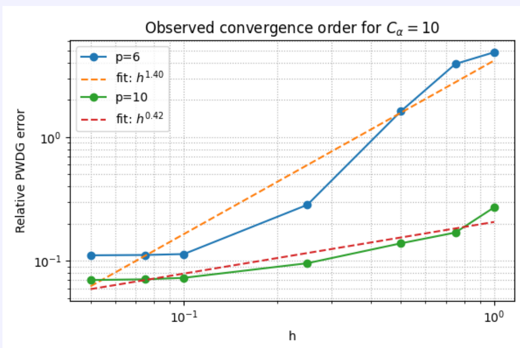
(c) Exact, $\epsilon_m = -1.01$

(d) PWDG, $\epsilon_m = -1.01$

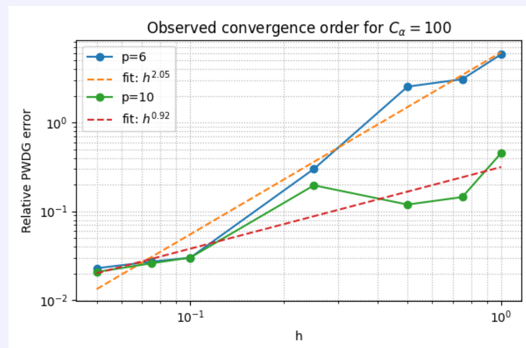
Colorbar

Figure 4.8. Solutions on disk ($R = 4$, $k = 4$, $h = 0.25$) . Top row: $\epsilon_m = -1.1$. Bottom row: $\epsilon_m = -1.01$.

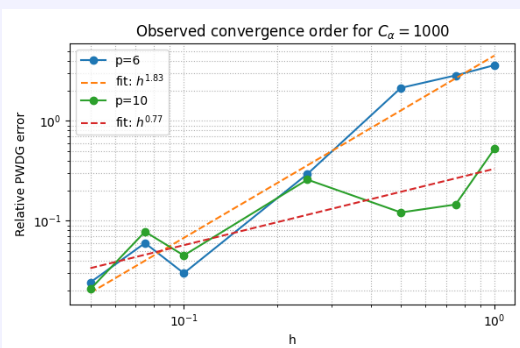
Experiment 4.5.2 (PWDG convergence rates for $k = 4$). One might ask, for varied values of p, can we observe good convergence rates? We take this same experiment and try to determine good convergence rates. Here we set $\alpha = \beta = C_\alpha \frac{p^2}{kh}$ and see how the convergence rates are affected for this experiment. See Figure 4.9.



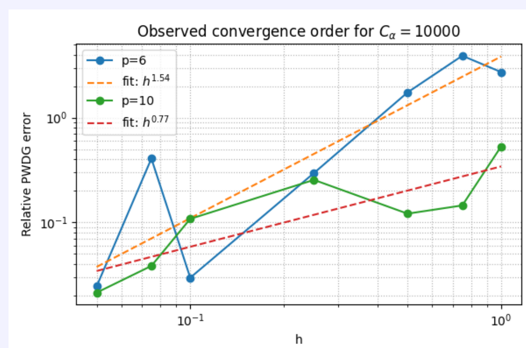
(a) $k = 4, C_\alpha = 10$



(b) $k = 4, C_\alpha = 100$



(c) $k = 4, C_\alpha = 1000$

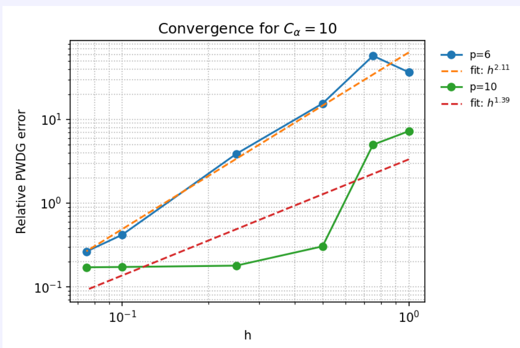


(d) $k = 4, C_\alpha = 10000$

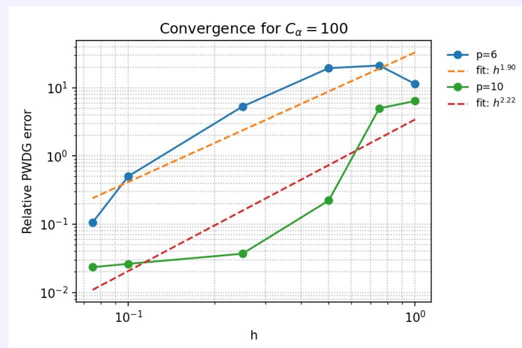
Figure 4.9. Observed convergence order of the PWDG method with $k = 4, \epsilon_m = -1.1, R = 4$ (blue: $p = 6$, green: $p = 10$) for four different penalty–scaling constants C_α. Each subplot shows a log–log fit $\sim h^r$ (dashed lines) against the raw errors.

We observe that as we increase our penalty parameter in this case, we gain additional accuracy for very small h, but at the cost of stability, causing 'peaks' of error. Additionally, we note that we do not see large order gains from increasing the value of p here, but instead we we see a downward shift, especially for small values of h. However, we note that the PWDG can swiftly converge to the error plateau for higher p in this case, especially for the less continuous case $C_\alpha = 10$.

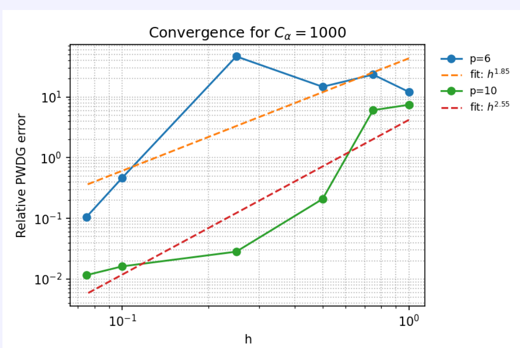
Experiment 4.5.3 (Convergence study for PPWs with $k = 8$). One reasonable question then, is given a higher-frequency problem, can we see a greater disparity between lower and higher order choices for number of PPWs? Indeed, in 4.10, we do this experiment and we see that as k increases to 8, we see just this phenomenon.



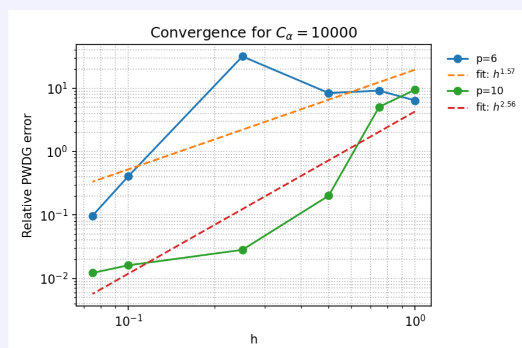
(a) $k = 8, C_\alpha = 10$



(b) $k = 8, C_\alpha = 100$



(c) $k = 8, C_\alpha = 1000$



(d) $k = 8, C_\alpha = 10000$

Figure 4.10. Observed convergence order of the PWDG method with $k = 8, \epsilon_m = -1.1, R = 8$ (blue: $p = 6$, green: $p = 10$) for four different penalty–scaling constants C_α. Each subplot shows a log–log fit $\sim h^r$ (dashed lines) against the raw errors.

We are interested to compare the high-frequency regime, to see how effective the PWDG is compared to the standard FEM in this regime, to see how the pollution error for higher k has an effect on the two methods.

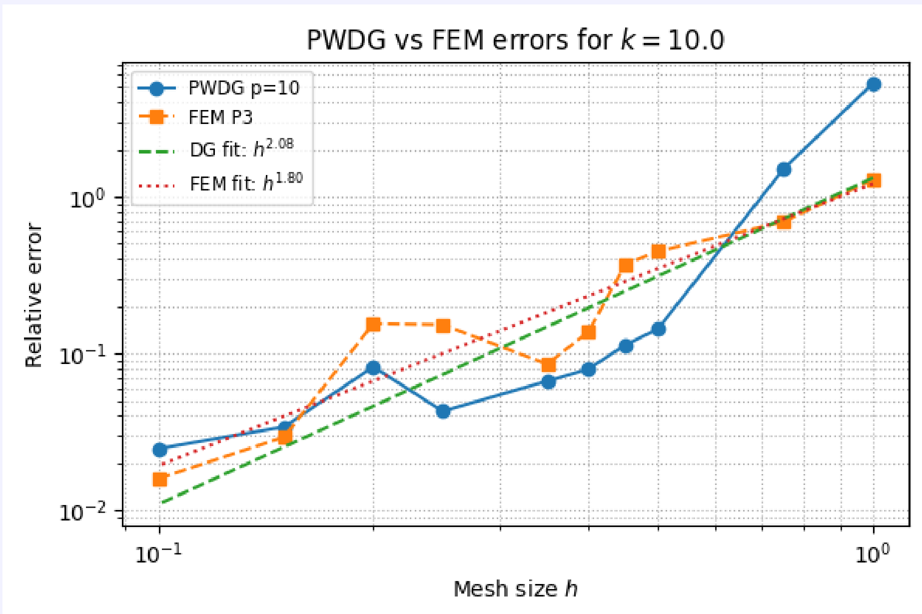


Figure 4.11. Observed convergence order of the PWDG method with $C_\alpha = 100, k = 10, \epsilon_m = -1.1, R = 8, p = 10$ and order 3 for FEM. Included are a log–log fit $\sim h^r$ (dashed lines) against the raw errors.

Experiment 4.5.4. Here we note several different regions of behavior. In large h, both methods fail, and for small h, both approach the same error plateau. However, for intermediate values of h, we see that the PWDG seems to mildly outperform the FEM. This implies that for settings where it is not achievable to have such refined h (as is common in 3D settings with multiple spheres for example) the PWDG may be a more attractive method to use in such intermediate cases.

Experiment 4.5.5. Of note, we cannot increase the value of p in these experiments substantially further than 10, as doing so tends to lead to ill-conditioning, as the plane waves grow more linearly dependent on each other - see Figure 4.12.

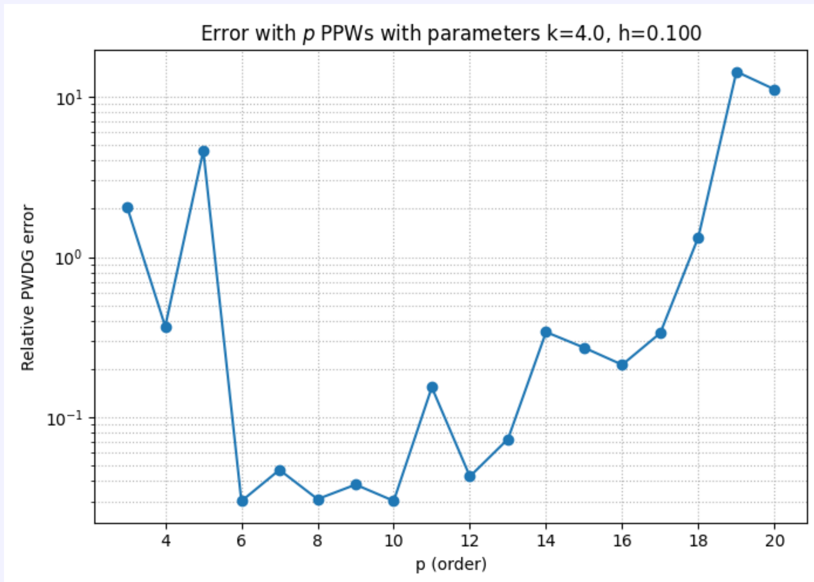
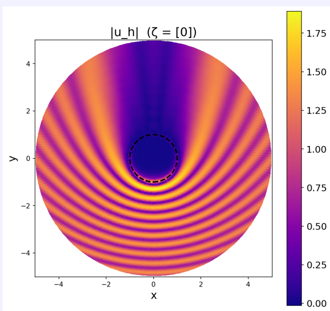


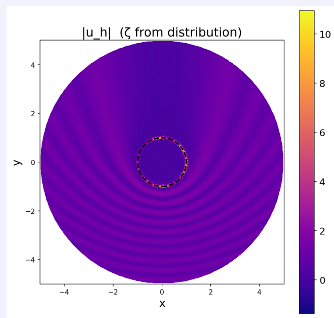
Figure 4.12. Relative PWDG error as a function of plane-wave order p for $k = 4.0$ and $h = 0.1$. For $p \gtrsim 10$ the error spikes sharply, indicating the onset of ill-conditioning as the plane-wave basis becomes nearly linearly dependent.

Evanescent Plane Waves Thus far, we have used only PPWs in our experiments. We will not reconsider the prior convergence results for EPWs, as we will not see large variation, but they are appropriate for individual use in solving the near resonance problem, as here we see much more rapid decay (as opposed to the smoother scattering problems given above). Indeed, we show below the difference in using EPWs in the smooth problem in the below experiment.

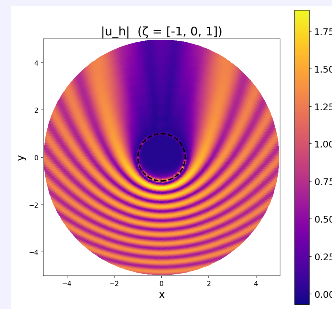
Experiment 4.5.6 (EPWs in a smooth solution). Here, we set $k = 8$, $h = 0.1, p = 10$ and we plot the solution using the PWDG for a variety of different sets of EPWs (Evanescent Plane Waves), as well as a solution which only uses PPWs. As a reminder, we note that the defining characteristic of the EPW is its decay rate ζ, and that a PPW is just a special case of $\zeta = 0$. This is defined in greater detail at §4.4.



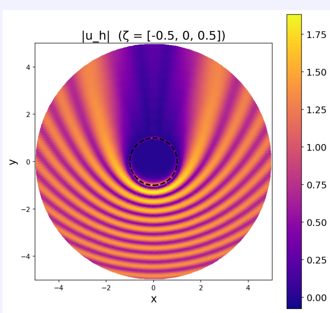
(a) $\zeta = 0$ (PPW only)



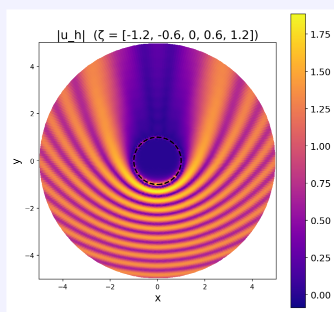
(b) Parolin distribution



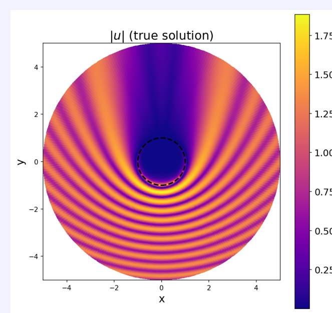
(c) $\zeta = [-1, 0, 1]$



(d) $\zeta = [-0.5, 0, 0.5]$



(e) $\zeta = [-1.2, -0.6, 0, 0.6, 1.2]$



(f) True solution $|u|$

Figure 4.13. Amplitude $|u_h|$ for $k = 8$, $h = 0.1$, $p = 10$, comparing PPW, various EPW sets, and the exact solution.

We note that we don't see a large variance in behavior (and this obviously holds true for error as well) for the spreads of plane waves with the exception of those chosen by distribution. Here, the ζ given by the distribution discussed in §4.4 is approximately (rounding to the nearest tenth) $[-2, -1.3, 0, 1.3, 2]$. This large decay is not native to the problem's solution, so introducing it primes the system with false peaks.

4.5.3. Near-resonance Experiments

We now seek to tie this method to our previous inquiry into near resonance problems, see [§2.3, §3.4.1]. In this regime, we consider a disk of radius 1. We choose $k = \Re(\lambda_{l^*})$ for some fixed l^* so that our problem corresponds to the near-resonance regime. We begin by performing a similar experiment to Figure 3.7a to study the behaviour of the Trefftz PWDG method as we take $\epsilon_m \to -1$. We recall from Theorem 4.3.1 that we expect our constant of error to behave with respect to the coercivity constant from T-coercivity, see 3.2.1.

Experiment 4.5.7 (PWDG error for near-resonance regime with PPWs as $\epsilon_m \to -1$). We choose $k = \Re(\lambda_{l^*})$ for some fixed l^* so that our problem corresponds to the near-resonance regime and take $\epsilon_m \to -1$, and indeed we see the error behaves as expected, which is to say poorly behaved and growing worse as we approach the region of ill-posedness.

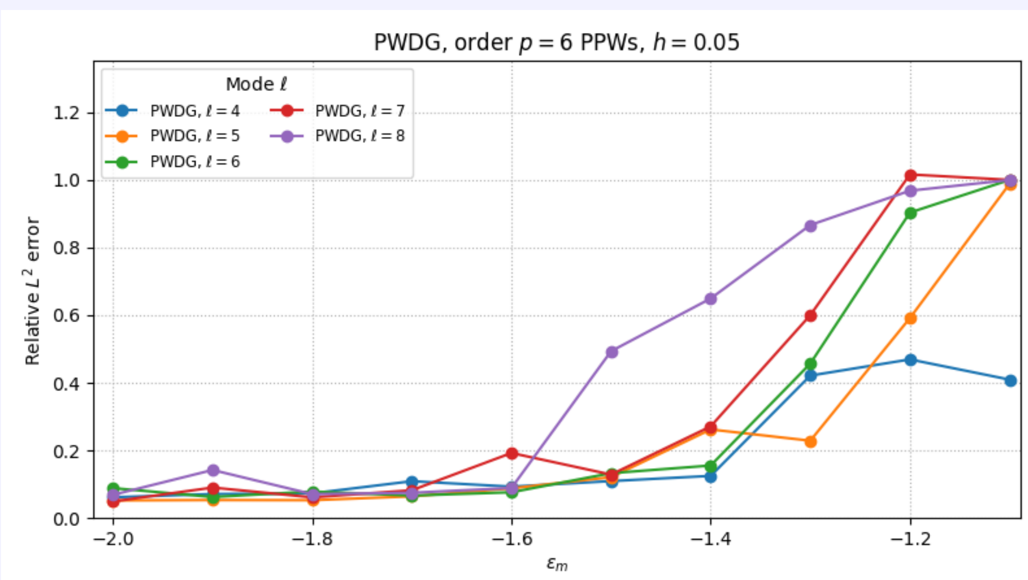


Figure 4.14. Relative L^2 errors vs. ϵ_m for fixed $k = \Re(\lambda_{l^*})$ with 6 PPWs used as basis functions.

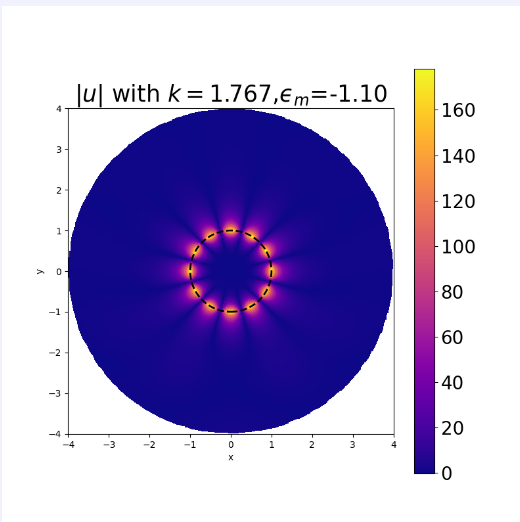
Remark 4.5.3. In the following experiment 4.5.8, we choose a relatively low-frequency surface plasmon. We note the relative improvement of the PWDG when used with EPWs (which are used in a layer about the interface only to capture the surface plasmon), however there are several important caveats.

1. Firstly, while the behaviour is better captured, the L^2 error for this experiment is not actually substantially improved, it hovers near 1 regardless. We can capture the behavior of the plasmon somewhat, but there is a slight phase drift, and the offset causes large errors to propogate.

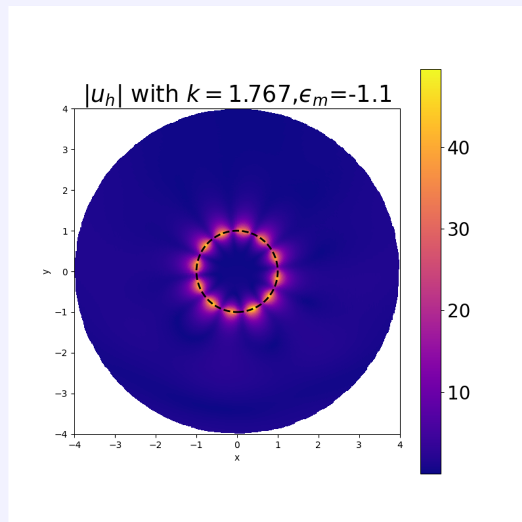
2. Secondly, this relies on ζ being chosen specifically to cause the decay to mimic that of the surface plasmon. This trial and error approach, while not unheard of for PWDG methods (see Gittelson (2008)), can be expensive and may take more than one solve to choose the right parameters, due to the conditioning of the matrix.

3. Thirdly, this method has stability concerns. We still have the T-coercivity property per Theorem 4.3.1, which is inherent in the continuous problem, and perforce it will be present in the discrete problem as the spectrum of the discrete problem converges to that of the continuous problem. This means for k chosen near resonance, we will naturally have a poorly conditioned linear system. This ill-conditioning will only be exacerbated by highly evanescent basis functions or large penalty parameters. This ill-conditioning will make the problem sensitive to its parameters and may interfere with some solvers.

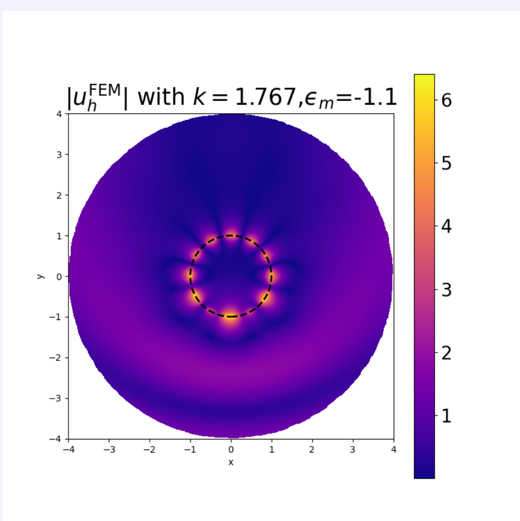
Experiment 4.5.8 (Near-resonance E-PWDG vs. FEM). In the below experiment, we take a small spread of $\zeta = [-1.45, -0.8, 0, 0.8, 1.45]$ and use EPWs as basis elements in a thin layer about the interface. We see a marked response from the system due to the evanescent decay of those elements, but we also note a phase drift in the result. The relative L^2 error is still quite large - for the PWDG with EPWs is .9791 and the FEM error is .9991. See Remark 4.5.3.



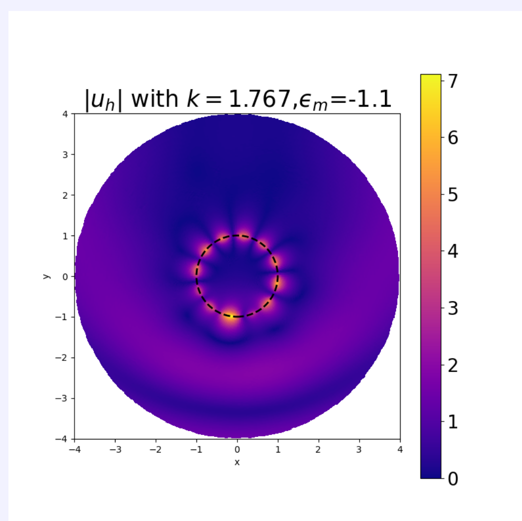
(a) Analytical Solution



(b) PWDG with EPWs



(c) FEM



(d) PWDG with PPWs

Figure 4.15. Near-resonance experiment with $l = 6, k = 1.7666093821137867$, $p = 10$, $h = 0.05$, $C_\alpha = 1000$. $\zeta = [- -1.45, -0.8, 0, 0.8, 1.45]$ chosen *a posteriori* and applied locally about the interface.

Summary of Numerical Findings. In light of the experiments above, we conclude:

1. *Flux scaling.* The choice $\alpha = C_\alpha \frac{2p^2}{kh}$, $\beta \geq C_\alpha \frac{2p^2}{kh}$ with consistently lies in the low-error quadrant of the (α, β)-plane.

2. *Mesh robustness.* Trefftz-DG converges equally on structured and unstructured grids, whereas standard FEM stalls on unstructured meshes as $\varepsilon_m \to -1$.

3. *High-frequency performance.* In the $k = 8, 10$ regime, PWDG mildly outperforms FEM for moderate h, and attains its error plateau more quickly as $h \to 0$.

4. *Near-resonance.* For ε_m approaching the critical value, E-PWDG can recover plasmonic decay features that FEM cannot but ill-conditioning and phase-drift remain challenging.

5. EXTENSIONS & CONCLUSION

5.1. Conclusions

In this dissertation, we have investigated the scalar Helmholtz equation describing wave scattering by negative-index materials, with particular attention given to the challenging near-resonance regime. We have discussed analytical characterizations as well as numerical methods, providing an analysis of the mathematical and numerical challenges posed by sign-changing coefficients and plasmonic resonances.

We began with Chapter 2 by deriving explicit analytical solutions for scattering by a canonical geometry, namely a penetrable spherical scatterer, leveraging symmetry through separation of variables. This approach allowed precise characterization of resonances and associated resonant modes, revealing their highly localized and oscillatory nature. Particularly notable was the clear identification of near-real resonances indicative of plasmonic modes, see 2.3. Such modes displayed significantly amplified localized fields at the interface, giving us insight into the wavenumbers needed to excite plasmonic behaviors.

Numerically, we first considered in Chapter 3 standard Finite Element Methods to approximate solutions to the Helmholtz problem in 3D sections. Our analysis showed that standard polynomial-based FEM requires specific symmetric meshes, and faces severe limitations for the near-resonance case. Mesh refinements and higher-order polynomial bases were insufficient to resolve this, requiring prohibitively fine discretizations. In an attempt to overcome these limitations, we introduced an enriched FEM method which incorporated the analytically derived resonant mode into the FEM, hoping to capture the behaviour of the plasmon by using *a priori* knowledge of the resonance. The results were mixed. While we saw the oscillations of the plasmon captured by the enrichment, they ultimately failed to resolve the amplitude due to the ill-conditioning of the system and the poor approximation of the amplitude of the surface plasmon by the polynomial basis.

In 2D in Chapter 4, we introduce and derive the PWDG method for the sign-changing problem. We provide theoretical results, including well-posedness, quasi-optimality, and convergence in the L^2 norm. Our numerical results revealed several different settings in which the PWDG outpaces a standard FEM approach in the sign-changing problem, notably being better for non-symmetric meshes and competitive for some mesh sizes h in the higher-frequency regime. In the near-resonance case, we have shown some improvements in individual experiments, but we have also shown that we are still restricted by the underlying T-coercivity of the continuous problem. Future work for this will explore automated selection of ζ-distributions and more robust boundary conditions to further harness the Trefftz framework in near-resonance regimes.

5.2. Other Future Research Directions

While the methods developed in this thesis address the fundamental near-resonance scattering by a single ball metamaterial, there remain many other avenues to extend and apply these techniques. We outline several key future research directions and extensions:

Multiple Scatterers, General 3D Geometries and other equations: A natural next step is to extend our methods to configurations with multiple scatterers and smooth 3D domains, moving beyond the canonical single-ball geometry studied here. In practical nanophotonics, plasmonic nanoparticles often appear in clusters or arrays, where inter-particle coupling can induce new collective resonances (plasmonic coupling). Grieser (2014) Handling multiple such metamaterial scatterers presents increased challenges: the mesh generation becomes more complex, and the linear systems grow substantially in size (potentially millions of unknowns for even a few scatterers). The PWDG method we propose, if extended to 3D (as has been done for the non-sign-changing problem, see Moiola (2011); Hiptmair et al. (2016)) could be suited for multiple scattering, since it does not require highly structured meshes in the way FEM does. Additionally, it may have some computational advantages for 3D, as the computation of volumic integrals in these methods can be quite expensive, and the Trefftz method naturally avoids these. Indeed, recent work has applied DG integral-equation hybrids to light scattering in complex nanoparticle assemblies, indicating the feasibility of DG for multi-particle configurations. Martín et al. (2023)

For this reason, extending our Trefftz method to 3D and tackling a more robust physical problem, by including either the more realistic Drude's model for permitivity and the more complex full Maxwell's equations are goals we'd be interested to extend in this project. This is conceptually viable, but would require an expanded analytical framework and codebase.

Alternative Trefftz Bases and Hybrid Approaches: Our PWDG implementation used plane waves (propagating and evanescent) as basis functions. There is ample room to explore alternative or optimized Trefftz bases to further improve accuracy and efficiency. One idea is to develop a hybrid basis that that mixes Trefftz functions with standard polynomials or other function sets. For example, one could imagine a local plane-wave Trefftz near the interface, paired with polynomial FEM in the far fields where we seek only to capture what is effectively low-order smooth scattering.

Furthermore, there may be other Trefftz function families beyond plane waves worth exploring. For some examples of potential basis families, Bessel/Hankel functions and more broadly the popular generalized harmonic polynomials. Each Trefftz basis has its own approximation properties; by experimenting with the choice of basis functions we might better address dispersion error (although this would likely need to be paired with a study of the approximation properties of such functions for the sign-changing problem, in itself a nontrivial task). Adaptive selection of basis functions is another promising area: one could envision an *hp*-adaptive Trefftz method that not only refines the mesh h but also adaptively chooses the number and type of Trefftz functions per element based on error indicators. This would allocate evanescent waves where rapid decays are needed and more propagating waves where the solution is oscillatory. For some works on adaptive PWDG methods, we refer to Gittelson (2008); Hiptmair et al. (2016). Another intriguing direction in this same vein is the development of embedded Trefftz methods, which aim to integrate Trefftz trial functions into more traditional finite element or discontinuous Galerkin frameworks. Rather than replacing the entire approximation space with Trefftz functions (as in a pure PWDG), an embedded approach would insert Trefftz subspaces into a larger scheme. Recent work Lehrenfeld & Stocker (2023) introduces an embedded Trefftz DG method, wherein certain solution components are constrained to lie in a Trefftz space to improve efficiency. This might be practical for a large-scale application with, for example, multiple spherical scatterers.

Other DG Methods: The DPG (Discontinuous Petrov Galerkin) approach is known for its stability achieved by using optimal test spaces. This shares the same idea of the ultra-weak variational formulation as the inspiration for the PWDG, although in application it is quite different. In essence, a DPG scheme chooses test functions (in an enriched infinite-dimensional space, solved approximately) such that the discrete bilinear form is automatically stable. DPG variants have been successfully applied to simulate transformation optics cloaking, which involves anisotropic, negative-index materials Demkowicz & Li (2013). We foresee that a similar DPG strategy could be applied to the plasmonic resonance problem: by carefully tailoring the test norm (test and trial functions are different in a PDG method), one might circumvent the need for T-coercivity constraints altogether, since stability is enforced by the Petrov–Galerkin framework. This might allow for sharper estimates in the near-resonance problem. Of course, this is a hypothesis which further research would be needed to validate. Additionally, DPG provides a built-in *a posteriori* error estimate from the residual, which could guide adaptive refinement near critical interfaces. While this increases the local computational cost, the payoff is a method that is provably stable, and may even be able to handle better the near-resonance case using built-in stability. Despite this compelling advantage, the main detractor from this route is the increased costs, which clash with our interest in larger systems of multiple scatterers.

The other principle avenue of investigation for non-Trefftz DG methods, which leans more on the computationally efficient side (but away from the analytical advantages which we might need to leverage for near-resonance cases) is the hybrid discontinuous Galerkin methods (HDG) with static condensation. Traditionally, these methods often have super-convergence properties, but we are skeptical that they can tackle the near-resonance case to have better approximation properties than the PWDG. Still, they have the advantage of efficiency, so paired with highly graded meshes they are likely competitive with standard FEM.

REFERENCES

Ammari, H., Chow, Y., & Liu, H. (2020). Quantum ergodicity and localization of plasmon resonances. *arXiv preprint*. Retrieved from `http://arxiv.org/abs/2003.03696` (arXiv:2003.03696)

Ammari, H., Dabrowski, A., Fitzpatrick, B., & Millien, P. (2020). Perturbation of the scattering resonances of an open cavity by small particles. part i: The transverse magnetic polarization case. *Zeitschrift für angewandte Mathematik und Physik*, *71*(4), 102. doi: 10.1007/s00033-020-01324-6

Ammari, H., Millien, P., Ruiz, M., & Zhang, H. (2017). Mathematical analysis of plasmonic nanoparticles: The scalar case. *Archive for Rational Mechanics and Analysis*, *224*, 597–658. doi: 10.1007/s00205-017-1084-5

Arnold, D. N., Brezzi, F., Cockburn, B., & Marini, D. (2002). Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM Journal on Numerical Analysis*, *39*(5), 1749–1779. doi: 10.1137/S0036142901384162

Babuska, I., & Melenk, J. M. (1997). The partition of unity method. *International Journal for Numerical Methods in Engineering*, *40*(4), 727–758. doi: 10.1002/(SICI)1097-0207(199707)40:4⟨727::AID-NME453⟩3.0.CO;2-#

Belytschko, T., Gracie, R., & Ventura, G. (2001). A review of extended/generalized finite element methods for material modeling. *International Journal for Numerical Methods in Engineering*, *50*(9), 2145–2162. doi: 10.1002/nme.1009

Bennaceur, K., Dobaczewski, J., & Ploszajczak, M. (1999). Continuum effects for the mean-field and pairing properties of weakly bound nuclei. *Physical Review C*, 7–8. doi: 10.1103/physrevc.60.034308

Bonnet-Ben Dhia, A.-S., Carvalho, C., Chesnel, L., & Ciarlet, P. J. (2016). On the use of perfectly matched layers at corners for scattering problems with sign-changing coefficients. *Journal of Computational Physics*, *322*, 224–247.

Bonnet-Ben Dhia, A.-S., Carvalho, C., & Ciarlet, P. J. (2018). Mesh requirements for finite element approximation of problems with sign-changing coefficients. *Numerische Mathematik*, *138*, 801–838.

Bonnet-Ben Dhia, A.-S., Chesnel, L., & Ciarlet, P. J. (2012). T-coercivity for scalar interface problems between dielectrics and metamaterials. *ESAIM: Mathematical Modelling and Numerical Analysis*, *46*(6), 1363–1387. doi: 10.1051/m2an/2012006

Bonnet-Ben Dhia, A.-S., Chesnel, L., & Claeys, X. (2013). Radiation condition for a non-smooth interface between a dielectric and a metamaterial. *Mathematical Models and Methods in Applied Sciences*, *23*(9), 1629–1662. doi: 10.1142/s0218202513500188

Bonnetier, E., Dapogny, C., Triki, F., & Zhang, H. (2019). The plasmonic resonances of a bowtie antenna. *Analysis in Theory and Applications*, *35*(1), 85–116. doi: https://doi.org/10.4208/ata.OA-0011

Brenner, S. C., & Scott, L. R. (2007). *Mathematical theory of finite element methods* (3rd ed., Vol. 15). New York: Springer-Verlag.

Carvalho, C., Chesnel, L., & Ciarlet, P. J. (2017). Eigenvalue problems with sign-changing coefficients. *Comptes Rendus Mathematique*, *355*(6), 671–675. doi: 10.1016/j.crma.2017.05.002

Ciarlet, P. (1978). *The finite element method for elliptic problems* (Vol. 4). Amsterdam: North-Holland.

Colton, D., & Kress, R. (2013). *Inverse acoustic and electromagnetic scattering theory* (3rd ed.). Springer.

Demkowicz, L., & Li, J. (2013). Numerical simulations of cloaking problems using a dpg method. *Computational Mechanics*, *51*, 661–672. doi: 10.1007/s00466-012-0744-4

Galante, N., Moiola, A., & Parolin, E. (2024). *Stable approximation of helmholtz solutions in the ball using evanescent plane waves.* arXiv:2401.04016. (Preprint)

Gittelson, C. J. (2008). *Plane wave discontinuous galerkin methods* (Unpublished master's thesis). ETH Zürich, Zürich, Switzerland.

Gittelson, C. J., Hiptmair, R., & Perugia, I. (2009). Plane wave dg methods: analysis of the h-version. *ESAIM: Mathematical Modelling and Numerical Analysis*, *43*, 297–331.

Goul, R. e. a. (2017). Quantitative analysis of surface-enhanced raman spectroscopy of rhodamine 6g using a composite graphene and plasmonic au nanoparticle substrate. *Carbon*, *111*, 386–392. doi: 10.1016/j.carbon.2016.10.019

Grieser, D. (2014). The plasmonic eigenvalue problem. *Reviews in Mathematical Physics*, *26*(3), 1450005. doi: 10.1142/s0129055x14500056

Hansen, P. C. (1998). Rank-deficient and discrete ill-posed problems: Numerical aspects of linear inversion.

Helsing, J., & Karlsson, A. (2018). On a helmholtz transmission problem in planar domains with corners. *Journal of Computational Physics*, *371*, 315–332. doi: 10.1016/j.jcp.2018.05.044

Hiptmair, R., Moiola, A., & Perugia, I. (2011). Plane wave dg methods for the 2d helmholtz equation: analysis of the p-version. *SIAM Journal on Numerical Analysis*, *49*(2), 264–284.

Hiptmair, R., Moiola, A., & Perugia, I. (2016). A survey of trefftz methods for the helmholtz equation. *Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations*, 237–279. doi: 10.1007/978-3-319-41640-3_8

Ihlenburg, F. (1995). *Finite element analysis of acoustic scattering* (Vol. 132). New York: Springer-Verlag.

Jiménez, P. e. a. (2020). Surface-enhanced raman spectroscopy: Benefits, trade-offs and future developments. *Chem. Sci.*, *11*, 4563.

Kirsch, A., & Hettlich, F. (2015). *The mathematical theory of time-harmonic maxwell's equations: Expansion-, integral-, and variational methods.* Springer Cham. doi: 10.1007/978-3-319-11086-8

Kravanja, P., & Barel, M. (2000). *Zeros of analytic functions.* Berlin, Heidelberg: Springer.

Latham, B. J. (2025a). *BenJLatham/Trefftz_EPW: A code for solving a trefftz method using evanescent plane waves.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.15602383` (Source code available at `https://github.com/BenJLatham/Trefftz_EPW`) doi: 10.5281/zenodo.15602383

Latham, B. J. (2025b, March). *T-Conf: Symmetric sphere and disk transfinite mesh generator.* Retrieved from `https://github.com/BenJLatham/T-Conforming-Balls` doi: 10.5281/zenodo.15376402

Lehrenfeld, C., & Stocker, P. (2023). Embedded Trefftz discontinuous Galerkin methods. *International Journal for Numerical Methods in Engineering*, *124*(17), 3637–3661. doi: 10.1002/nme.7258

Li, L., Lanteri, S., Mortensen, N. A., & Wubs, M. (2017). A hybridizable discontinuous galerkin method for solving nonlocal optical response models. *Computer Physics Communications*, *219*, 99–107. doi: 10.1016/j.cpc.2017.05.012

Maier, S. (2007). *Plasmonics: Fundamentals and applications.* New York, NY: Springer. doi: 10.1007/0-387-37825-1

Martín, V. F., Solís, D. M., Jericó, D., Landesa, L., Obelleiro, F., & Taboada, J. M. (2023). Discontinuous galerkin integral equation method for light scattering from complex nanoparticle assemblies. *Optics Express*, *31*(2), 1034–1048. doi: 10.1364/OE.478414

Melenk, J., & Sauter, S. (2010). Convergence analysis for finite element discretizations of the helmholtz equation with dirichlet-to-neumann boundary conditions. *Mathematics of Computation*, *79*(272), 1871–1914.

Melenk, J. M. (1995). *On generalized finiteelement methods* (PhD thesis). University of Maryland, College Park, College Park, MD. (ProQuest Dissertations & Theses No. 9539704)

Melenk, J. M., & Sauter, S. A. (2011). Wavenumber-explicit convergence analysis for galerkin discretizations of the helmholtz equation. *SIAM Journal on Numerical Analysis*, *49*(3), 1210–1243. doi: 10.1137/100791492

Moiola, A. (2011). Trefftz-discontinuous galerkin methods for time-harmonic wave problems.
(Seminar for Applied Mathematics, PhD Thesis.)

Monk, P., Pena, M., & Selgas, V. (2025, May). Trefftz discontinuous galerkin approximation of an acoustic waveguide. arXiv:2408.14833v2 [math.NA].

Olver, F. W. J., et al. (Eds.). (2010). *Nist digital library of mathematical functions.* Retrieved from `http://dlmf.nist.gov/10.60.E7` (Equation 10.60.17)

Paige, C. C., & Saunders, M. A. (1982). LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software*, *8*(1), 43–71.

Parolin, E., Huybrechs, D., & Moiola, A. (2023). Stable approximation of helmholtz solutions in the disk by evanescent plane waves. *ESAIM: Mathematical Modelling and Numerical Analysis*.

Tikhonov, A. N., & Arsenin, V. Y. (1977). *Solution of ill-posed problems.* Washington, D.C.: Winston & Sons.

Trefftz, E. (1926). Ein gegenstück zum ritzschen verfahren. In *Verhandlungen des 2. internationalen kongresses für technische mechanik* (pp. 131–137). Zürich.

Wang, M., Schmidt, K., Alparslan, A., & Hafner, C. (2013). Hp-fem and pml analysis of plasmonic particles in layered media. *Progress In Electromagnetics Research*, *142*, 523–544. doi: 10.2528/PIER13022106

APPENDIX A: SCHUR-COMPLEMENT REDUCTION OF THE ENRICHED SYSTEM

Due to the 2×2 block nature of the enriched system (3.24), many sparse solvers (e.g. SciPy's `spsolve`) may struggle to directly invert this matrix when discretized, especially given that the new final row and column may be fairly dense. To avoid this, we employ a Schur complement approach. Starting from the block system (3.24),

$$\begin{bmatrix} -\mathcal{K}_\epsilon + \mathcal{B} + k^2\,\mathcal{M} & \mathcal{M}_p \\ (k^2 - \overline{\lambda}^2)\,\mathcal{M}_{r,p} + \mathcal{B}_p - \mathcal{S}_p & C_1 \end{bmatrix} \begin{bmatrix} u^{\mathrm{reg}} \\ \alpha \end{bmatrix} = \begin{bmatrix} f \\ C_2 \end{bmatrix},$$

to be concise, we denote $A \;=\; -\mathcal{K}_\epsilon + \mathcal{B} + k^2\,\mathcal{M}$,

The Schur complement of the top-left block A is

$$S \;=\; C_1 \;-\; \Big((k^2 - \overline{\lambda}^2)\,\mathcal{M}_{r,p} + \mathcal{B}_p - \mathcal{S}_p\Big)\,\big(-\mathcal{K}_\epsilon + \mathcal{B} + k^2\,\mathcal{M}\big)^{-1}\,\mathcal{M}_p.$$

Hence we solve in three steps:

(i) Solve the sparse A-system twice to obtain two intermediate fields

$$t = \big(-\mathcal{K}_\epsilon + \mathcal{B} + k^2\,\mathcal{M}\big)^{-1} f, \qquad s = \big(-\mathcal{K}_\epsilon + \mathcal{B} + k^2\,\mathcal{M}\big)^{-1}\mathcal{M}_p.$$

(ii) Compute the scalar amplitude α by

$$\alpha = \frac{C_2 \;-\; \Big((k^2 - \overline{\lambda}^2)\,\mathcal{M}_{r,p} + \mathcal{B}_p - \mathcal{S}_p\Big)\,t}{C_1 \;-\; \Big((k^2 - \overline{\lambda}^2)\,\mathcal{M}_{r,p} + \mathcal{B}_p - \mathcal{S}_p\Big)\,s}.$$

(iii) Recover the regular field by back-substitution

$$u^{\mathrm{reg}} = t \; - \; \alpha\, s.$$

Because the only factorization we ever need is of the sparse block $-\mathcal{K}_\epsilon + \mathcal{B} + k^2\,\mathcal{M}$, we preserve full sparsity - this is very useful in particular for the purpose of, for example, preconditioning the sparse matrix.

APPENDIX B: TIKHONOV REGULARIZATION

As §3.3.2 (and wavenumber explicit analyses such as J. M. Melenk & Sauter (2011)) show, when the incident wavenumber k approaches a plasmonic resonance λ_l, the standard polynomial FEM discretization of our truncated-sphere transmission problem develops near-zero eigenvalues. The resulting ill-conditioning renders h-refinement ineffective: no reasonable mesh will resolve the strongly localized surface plasmon. Although the enriched-FEM of §3.4.1 recovers the correct modal shape by augmenting the basis with the resonant mode, it still requires an impractically large number of degrees of freedom to match the true amplitude.

In other words, the discrete system inherits the continuous spectrum's vanishing gap as $k \approx \lambda_l$, causing stagnation of iterative solvers and the breakdown of the usual convergence rates (see Figure 3.6a). One technique commonly used to address such a pathological problem is Tikhonov regularization. (Tikhonov & Arsenin (1977); Hansen (1998))

Assume we have a discrete linear system

$$Ax = b,$$

The Tikhonov regularization method works by shifting the spectrum of the system matrix away from zero, reducing the condition number and stabilizing the numerical inversion. We modify the minimization problem by adding an L_2 penalty characterized by γ:

$$\min_x \left\{ \|Ax - b\|_2^2 + \gamma \|x\|_2^2 \right\} \Leftrightarrow \min_x \left\{ \|Ax - b\|_2^2 + \gamma x^T M x \right\} \tag{B.1}$$

where $\gamma > 0$ is the regularization parameter and M is the mass matrix, defined by the FEM basis functions ϕ_i:

$$M_{i,j} = \int_{\Omega} \phi_i(x)\, \phi_j(x)\, dx$$

Taking the derivative with respect to x and setting it to zero leads to the Tikhonov-regularized system:

$$\big(A^T A + \gamma M\big)\, x = A^T b. \tag{B.2}$$

Directly forming and factorizing the normal equations

$$(A^T A + \gamma M)\, x = A^T b$$

is both memory- and time-intensive—$A^T A$ is much denser than A, and its condition number is roughly $k(A)^2$. A more efficient, "matrix-free" approach is to recast Tikhonov regularization as the augmented least-squares problem

$$\min_x \left\| \begin{bmatrix} A \\ \sqrt{\gamma}\, I \end{bmatrix} x - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|_2^2,$$

and to solve it using the scipy LSQR algorithm [Paige & Saunders (1982), Hansen (1998)]. LSQR only requires matrix–vector products with A and A^T, never explicitly forming $A^T A$. We note that in In SciPy's LSQR implementation, the parameter `damp` forms the augmented least-squares problem

$$\min_x \left\| \begin{bmatrix} A \\ \mathrm{damp}\, I \end{bmatrix} x - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|_2^2.$$

Writing out its normal equations gives

$$\left(A^T A + \mathrm{damp}^2\, I\right) x = A^T b.$$

Comparing with the Tikhonov system (B.2), in which the penalty term is $\gamma \, \|x\|_2^2$, shows that

$$\mathrm{damp}^2 = \gamma \quad \Longrightarrow \quad \mathrm{damp} = \sqrt{\gamma}.$$

Thus by passing `damp = sqrt(gamma)` to LSQR one exactly realizes the shift $A^T A \mapsto A^T A + \gamma I$, preserving sparsity and avoiding explicit formation of $A^T A$.

In practice, the choice of the regularization parameter γ is critical. A common strategy is to sample γ on a logarithmic grid (e.g. $\gamma = 10^{-8}, 10^{-7}, \ldots, 10^{-1}$) and monitor both the relative residual $\|Ax_\gamma - b\|_2 / \|b\|_2$ and the solution seminorm $\|x_\gamma\|_M = \sqrt{x_\gamma^T M x_\gamma}$. In our numerical tests we will sweep γ over $[10^{-8}, 10^{-1}]$, identify the corner, and compare against the discrepancy-principle choice to see which yields the best recovery.

Figure B.1 shows that, in the off-resonant regime, the relative error remains low for $\gamma \lesssim 10^{-4}$ but increases sharply once the penalty dominates. By contrast, the near-resonant curve stays essentially flat near unity for all γ, indicating that a Tikhonov shift cannot recover the localized plasmonic peak. This underscores that, although regularization does improve conditioning, it alone cannot resolve the high-frequency surface mode without an enriched basis. In other words, when we expose our problem to this shift, we are in essence solving the wrong problem - one away from our surface plasmon.

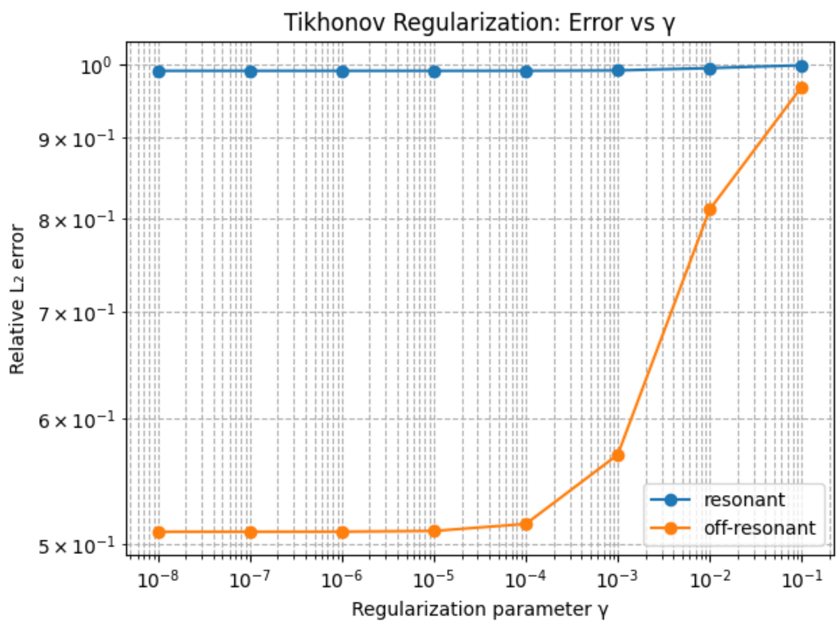


Figure B.1. Relative L^2 error of the Tikhonov-regularized LSQR solution as a function of the regularization parameter γ, for both the near-resonant test ($k \approx 5.2338$, blue) and the off-resonant baseline ($k = 5$, orange). Both axes are logarithmic.

APPENDIX C: ON THE DERIVATION OF PRIMAL NUMERICAL FLUXES FOR PWDG

In this appendix, we present a derivation of the so-called primal numerical fluxes used in Trefftz discontinuous Galerkin (DG) methods, particularly those adapted for sign-changing Helmholtz-type problems. These fluxes play a critical role in ensuring consistency, stability, and convergence of the numerical method. Our derivation follows from first principles, and draws on ideas from classical DG theory as well as Trefftz-specific developments in Hiptmair et al. (2016); Moiola (2011); Gittelson et al. (2009). We restrict ourselves to the following 2D problem throughout.

Consider the following problem:

$$\begin{aligned}
&\text{Find } u := u^{\text{inc}} + u^{\text{sca}} \in H^1_{\text{loc}}(\mathbb{R}^2) \text{ such that:}\\
&\quad \nabla \cdot (\epsilon^{-1}\nabla u) + k^2 u = 0, \quad \text{in } \mathbb{R}^2,\\
&\quad [u]_{\Upsilon_h^I} = 0, \quad [\epsilon^{-1}\partial_n u]_{\Upsilon_h^I} = 0 \quad \text{on } \Upsilon_h^I,\\
&\lim_{r\to\infty} \sqrt{r}\left(\frac{\partial u^{\text{sca}}}{\partial r} - iku^{\text{sca}}\right) = 0,\\
&k \in \mathbb{R}^+.
\end{aligned} \tag{C.1}$$

With the associated elementary form:

Problem C.0.1. Find $u \in V^p(\Omega_h)$ such that

$$A_h(u,v) = f_h(v) \quad \forall v \in V^p(\Omega_h),$$

with

$$\begin{aligned}
A_h(u,v) = &\int_{\Upsilon_h^I} \Big((ik\,\hat{\sigma}\cdot n)\,[\![\overline{v}]\!] \;-\; [\![\epsilon^{-1}\nabla\overline{v}]\!]\,\hat{u}\Big)\,ds\\
&+ \int_{\Upsilon_h^B} \Big((ik\,\hat{\sigma}\cdot n)\,\overline{v} \;-\; (\epsilon^{-1}\partial_n\overline{v})\,\hat{u}\Big)\,ds,
\end{aligned}$$

and

$$f_h(v) = \int_{\Upsilon_h^B} \left(-ik\, u^{\text{inc}} + \partial_n u^{\text{inc}}\right) \overline{v}\, ds.$$

For more details, we refer to the derivation present in §4.2.

In the Trefftz-DG method for the sign-changing Helmholtz equation, finding the numerical fluxes translates into defining, for each mesh skeleton face, two numerical quantities - a numerical trace of the solution, denoted $\hat{u}$ and a numerical flux, denoted $\hat{\sigma}$, which approximates the (possibly discontinuous) physical flux $\epsilon^{-1}\nabla u$.

These numerical quantities are inserted into the bilinear form to replace the (ambiguous) traces that arise from integration by parts over the discontinuous finite element space.

We begin by rewriting the second-order Helmholtz equation with sign-changing coefficients in first-order form. The original PDE

$$\nabla \cdot (\epsilon^{-1}\nabla u) + k^2 u = 0$$

can be equivalently rewritten as the first-order system:

$$\sigma = \epsilon^{-1}\nabla u, \tag{C.2}$$

$$\nabla \cdot \sigma + k^2 u = 0. \tag{C.3}$$

This first-order system is defined element-wise and must be supplemented with suitable transmission and boundary conditions.

Across internal interfaces Υ_h^I (between adjacent elements) the following transmission conditions must hold:

$$[u]_{\Upsilon_h^I} = 0, \quad [\sigma \cdot n]_{\Upsilon_h^I} = 0, \tag{C.4}$$

where $[\cdot]$ denotes the jump across an interface:

$$[u] := u^+ - u^-, \qquad [\sigma \cdot n] := \sigma^+ \cdot n^+ + \sigma^- \cdot n^-,$$

with $n^\pm$ denoting the outward normal from the respective adjacent element.

On the outer artificial boundary $\partial\Omega$, we impose a first-order absorbing (impedance) boundary condition that approximates the Sommerfeld radiation condition:

$$\partial_n u - iku = B_k, \tag{C.5}$$

where $B_k := \partial_n u^{\text{inc}} - iku^{\text{inc}}$ is the incident field contribution. In terms of the flux variable, this becomes:

$$\sigma \cdot n - iku = B_k. \tag{C.6}$$

Therefore, the complete formulation of the problem consists of the equations:

$$\sigma = \epsilon^{-1}\nabla u \quad \text{in } \Omega, \tag{C.2}$$

$$\nabla \cdot \sigma + k^2 u = 0 \quad \text{in } \Omega, \tag{C.3}$$

$$[u] = 0, \quad [\sigma \cdot n] = 0 \quad \text{on internal interfaces } \Upsilon_h^I, \tag{C.4}$$

$$\sigma \cdot n - iku = B_k \quad \text{on } \partial\Omega. \tag{C.6}$$

To motivate the form of the primal fluxes, consider the basic transmission interface between two subdomains where u is smooth. Across an interior edge or face F, one expects continuity of u and of the scaled normal derivative $\epsilon^{-1}\partial_n u$. However, in the DG setting, continuity across elements is no longer enforced strongly; instead, the transmission conditions (C.4) and boundary condition (C.6) must be imposed weakly.

Let $F \in \Upsilon_h^I$ be an interior face shared between two elements. The numerical traces we choose to enforce these conditions are defined by:

$$ik\,\hat{\sigma} = \{\{\sigma\}\} - \alpha\, ik\, [\![u]\!], \tag{C.7}$$

$$\hat{u} = \{\{u\}\} - \frac{1}{ik}\beta\, [\![\sigma]\!], \tag{C.8}$$

where $\{\{\cdot\}\}$ denotes the average of interior and exterior traces across the face, and $[\![\cdot]\!]$ denotes the jump as in (C.4). The constants $\alpha, \beta > 0$ are user-defined stabilization parameters.

This choice penalizes discontinuities in the solution u (via the α-term) and introduces a correction to the numerical solution trace $\hat{u}$ to account for the flux imbalance. Both fluxes are consistent: if u and σ are continuous across the face, then $\hat{u} = u$, $\hat{\sigma} = \sigma$.

Similarly on the boundary, let $F \in \Upsilon_h^B$ lie on the outer boundary $\partial\Omega$, where the impedance condition (C.6) is imposed. We define on each F:

$$ik\,\hat{\sigma} = \sigma - (1-\delta)\left(-\sigma + iku\, n + B_k\, n\right), \tag{C.9}$$

$$\hat{u} = u - \delta\left(\frac{1}{ik}\sigma\cdot n - u - \frac{1}{ik}B_k\right), \tag{C.10}$$

where δ is a boundary flux parameter. These expressions arise from a desire to continue to impose consistency as if the exact impedance condition holds, then again $\hat{u} = u$ and $\hat{\sigma} = \sigma$.

Here we do not suggest choices for the actual numerical values of α, β and δ as these are discussed in the numerical results section §4.5.1. However, the penalty parameters α and β serve a dual role in the primal flux formulation: they enforce stability by penalizing discontinuities, and they influence the amount of inter-element coupling. As $\alpha, \beta \to 0$, the penalty terms vanish, and the numerical scheme becomes under-constrained at the interfaces.

Conversely, as $\alpha, \beta \to \infty$, the scheme increasingly penalizes discontinuities, effectively enforcing strong continuity in u and $\epsilon^{-1}\nabla u \cdot n$, respectively. Therefore, in the asymptotic limit of very large α and β, the method begins to strongly resemble a conformal method.

This behavior mirrors that of interior penalty (IP) methods in other DG frameworks and we refer readers to the study on the intersection of IP and DG methods for various fluxes given in Arnold et al. (2002). The important point remains that choosing appropriately values of α and β is essential to constructing a method which is not too stiff and provides solutions neither spurious or trivial.

CURRICULUM VITAE

Benjamin J. Latham

2795 Wood Duck Drive, Los Banos, CA 93635

(209) 710-7634

`blatham@ucmerced.edu`

EDUCATION

Current Position *University of California Merced* (2019–2025)

Doctoral Candidate, Department of Applied Mathematics, Expected Graduation: July 2025

Thesis: Galerkin Methods for Solving the Helmholtz Problem in Scattering by Negative Materials with Applications for a Near-Resonance Regime

Areas of Interest:

- Numerical Wave Propagation, Plasmonics & Light-Scattering
- Finite Element Methods for Complex Media
- Trefftz Methods for Helmholtz-like Problems

All required coursework complete.

Prior Education *Sacramento State University*

MA in Mathematics, 2019

Culminating experience: Master's comprehensive examination in Analysis and Abstract Algebra

Research Topics: Applying Algebraic Topology to discover the simplicial structure of text documents

BA in Mathematics, 2017

Areas of Focus: Real, Complex, and Functional Analysis; Topology; Abstract Algebra

Honors: Cum Laude, Dean's Honor List

RELATED EXPERIENCE

Current Research Projects Research Advisor: C. Carvalho; Committee Members: C. Tsogka, A. Kim

- Plasmonic Scattering by Gold Nanoparticles using FEM, Analytic and Asymptotic Solutions to PDEs on a Sphere
- Trefftz DG Methods for Solving Helmholtz-like Problems Related to Surface Plasmonics on Spheres & Disks

Seminars UC Merced Applied Math Seminar; UC Merced Sensing and Imaging Seminar

7 research presentations given in UC Merced Sensing and Imaging Seminar (waves portion)

Prior Research *Research Assistant*, Spring 2019

Worked with Dr. Sayonita Ghosh Hajra (Sacramento State) on text analysis using simplicial homology and algebraic topology; coded in R and assisted with R package development.

INTERNATIONAL CONFERENCES & TALKS

- Benjamin Latham, Arnold Kim, Camille Carvalho. "Capturing Plasmonic Behaviors in Light Scattering by Spheres Using Finite Element Methods and Asymptotic Quadrature." WAVES 2022 Conference.

- Benjamin Latham. “Capturing Plasmonic Behaviors on Spheres Using Enriched Finite Element Techniques.” Minisymposium MS226: Mesh Constraints, SIAM SEAS 2023, Amsterdam.

- Benjamin Latham, Camille Carvalho. “Plane Wave DG Method for Solving the Complex Media Helmholtz Equation on a Disk.” WAVES 2024 Conference.

PUBLICATIONS

- Benjamin Latham, Arnold Kim, Camille Carvalho. “Capturing Plasmonic Behaviors in Light Scattering by Spheres Using Finite Element Methods and Asymptotic Quadrature.” Available at `https://par.nsf.gov/servlets/purl/10341726`.

- Benjamin Latham, Camille Carvalho. “Plane Wave DG Method for Solving the Complex Media Helmholtz Equation on a Disk.” WAVES 2024 Proceedings.

- Benjamin Latham, Camille Carvalho. “Scattering by Ideal Plasmonic Nanoparticles: an Evanescent/Propagative Plane Wave Discontinuous Galerkin (EPPWDG) Approach.” (In preparation)

MEMBERSHIPS

- Pi Mu Epsilon National Math Honor Society (2017–2019)

- SIAM (Society for Industrial and Applied Mathematics) (2019–Present)

TEACHING EXPERIENCE

Upper Division *UC Merced*

Teaching Assistant, Complex Analysis (Math 122), 2020–2021

Led discussions; graded exams and homework; administered grades.

Lower Division *UC Merced*

Teaching Assistant, Calculus (Math 021), 2019–2020

Led discussions; graded exams; administered grades.

San Juan Unified School District

Visiting Teacher, 2017–2019: Classroom organization; led discussions; specialized instruction.

Math Instructional Specialist, 2015–2016: Tutored struggling students; presented lectures; graded assignments.

FELLOWSHIPS & GRANTS

- NSF Grant DMS-2009366
- Applied Mathematics Travel Fellowship

SKILLS Programming: Python (very fluent); Matlab (fluent); R & C++ (competent)

Research Skills: High-Performance Computing (Pinnacles Cluster at UC Merced); GitHub (and Bitbucket)

Mathematical Training: Numerical & Functional Analysis; PDEs; Boundary Integral Equations; Scientific Computing